TESIS DE LA UNIVERSIDAD DE ZARAGOZA
2026
66

Lorenzo Mur Labadia

Learning visual models for egocentric perception

Director/es
Martínez Cantín, Rubén
Guerrero Campo, Josechu

ISSN 2254-7606

Prensas de la Universidad
Universidad Zaragoza

Universidad de Zaragoza
Servicio de Publicaciones

ISSN 2254-7606

Universidad Zaragoza
1542

Tesis Doctoral

LEARNING VISUAL MODELS FOR EGOCENTRIC PERCEPTION

Autor

Lorenzo Mur Labadia

Director/es

Martínez Cantín, Rubén
Guerrero Campo, Josechu

UNIVERSIDAD DE ZARAGOZA
Escuela de Doctorado

2026

Universidad Zaragoza

PhD Thesis

Learning visual models for egocentric perception

Author

Lorenzo Mur-Labadia

Supervisors

Jose J. Guerrero

Ruben Martinez-Cantin

Doctoral Program in Systems Engineering and Computer Science
Doctoral School (EDUZ)
2025

Acknowledgments

First and foremost, my sincere gratitude goes to my two supervisors, Josechu and Ruben, for giving me the opportunity to start my career as a researcher, and for their unwavering support, encouragement, and constructive feedback throughout this project. Their expertise and insights have been instrumental in shaping the direction of this work. I would also like to thank Ana Cris, Jesús, Alejandro, and Eduardo M. for their help during our collaborations.

Of course, I am very grateful to all my labmates (Sergio, Víctor, Juanjo, Julio, Javi R., Javi P., Tomás, Edu and Julia) for their support, camaraderie, and thought-provoking discussions. Special thanks to David, Carlos, and Maria, co-authors of several works presented in this thesis, for their hard work and for trusting me as a partner in their research.

I would also like to express my profound gratitude to Prof. Farinella and Antonino F. for providing me the opportunity to collaborate with such talented people. Thanks to everyone on the Catania team—pizza has never been the same. Grazie!!

A todos mis amigos que han vivido conmigo (Juan, Raúl, Jaime, Alejandro, José, Ignacio, Alfredo, Jorge y Mario); por hacer la vuelta a Zaragoza cada finde más llevadera. A mi segunda familia el Club de Amigos, y en especial a los monitores que confiaron en la loca idea de hacer un musical. A Baltasar, Salinas, Nieves, Alfredo, Selma, Izarbe y Josan Montull por vuestro apoyo, interés y entrega en el tiempo libre.

Quiero dar profundamente las gracias a mis padres y mis hermanas Isabel y Maria, por quererme tanto y tan bien y apoyarme en todas mis decisiones. Gracias a papi por enseñarme el sentido de la responsabilidad y demostrar que no importan los kilómetros para ver a la familia, y a mami por creer siempre en mi y darme todo el amor del mundo. A toda mi familia, en especial a mi abuela Mª Luisa, quien lleva luchando durante 15 años contra la degeneración macular y por quien me interesó el tema de los dispositivos visuales asistenciales. En los momentos que me sentía sin fuerzas, tu ejemplo ha sido siempre la mejor inspiración.

Por último, gracias a Angelita por acompañarme a todos sitios, ser la mejor compañera de aventuras y mi mayor fan.

Abstract

Egocentric vision holds great potential to revolutionize human-machine interaction by enabling systems to perceive and interpret the world from the user's perspective. In first-person videos, the actor moves continuously within a dynamic environment, requiring models capable of inferring user intentions, detecting objects and their functionalities, and reasoning about the relevance of the surrounding 3D scene. The goal of this thesis is to advance egocentric perception by developing visual models of objects, affordances, and environments, while integrating egocentric vision with multi-modal representations.

As a first step, we model the world as a collection of functional objects, where objects afford distinct interactions. We propose learning to accurately segment object parts according to their associated affordance and to quantify the uncertainty of these predictions. While detecting object affordances provides valuable insights into object functionalities, it is insufficient for comprehensive scene understanding, where objects are embedded within a broader physical space: the environment. To this end, we introduce a multi-label affordance mapping that links activity-centric zones to spatial locations, demonstrating its utility for embodied tasks such as task-oriented navigation. To further enhance temporal robustness, we propose a fusion strategy that leverages the predictive distribution of a Bayesian neural network. Finally, to better capture the dynamics of egocentric videos and enable a richer semantic understanding, we represent environments with implicit functions through a decomposed neural radiance field.

In the next part, we combine object and affordance models to improve the anticipation of the short-term object interaction. We introduce end-to-end architectures that extend classical object detectors for anticipation, and explore strategies to ground interaction predictions in past human behavior via environmental affordances and interaction hotspots.

Finally, we expand egocentric perception through the integration of multi-modal representations. First, we align egocentric video with natural language by localizing the temporal boundaries of activities within long, untrimmed videos. Second, we bridge first- and third-person perspectives by reformulating the cross-view segmentation as an object mask matching task, enabling effective alignment of object representations across viewpoints.

Overall, the methods proposed in this thesis achieve state-of-the-art results across a broad range of egocentric perception tasks. We hope that this work encourage future research in first-person visual perception.

Resumen

La visión egocéntrica posee un gran potencial para revolucionar la interacción humano-máquina, al permitir que los sistemas perciban e interpreten el mundo desde la perspectiva del usuario. En los vídeos en primera persona, el actor se desplaza de manera continua dentro de un entorno dinámico, lo que exige el desarrollo de modelos capaces de predecir las intenciones del usuario, detectar objetos y sus funcionalidades, y razonar sobre la relevancia de la escena tridimensional circundante. El objetivo de esta tesis es avanzar en la percepción egocéntrica mediante el desarrollo de modelos visuales de objetos, affordances y entornos, integrando la visión en primera persona con representaciones multimodales.

Como primer paso, se modela el mundo como una colección de objetos funcionales, en los cuales cada objeto posibilita distintas interacciones. Se propone aprender a segmentar con precisión las partes de los objetos en función de las affordances asociadas, así como cuantificar la incertidumbre de dichas predicciones. Si bien la detección de affordances proporciona información valiosa sobre la funcionalidad de los objetos, no resulta suficiente para una comprensión completa de la escena, dado que los objetos están integrados en un espacio físico más amplio: el entorno. Con este fin, se introduce un mapa de affordances multi-etiqueta que vincula zonas centradas en la actividad con ubicaciones espaciales, demostrando su utilidad en la navegación orientada a tareas específicas. Para reforzar la robustez temporal, se propone una estrategia de fusión que aprovecha la distribución predictiva de una red neuronal bayesiana. Finalmente, con el propósito de capturar de manera más efectiva la dinámica de los vídeos egocéntricos y conseguir una mayor comprensión semántica, se representa el entorno mediante funciones implícitas a través de un campo neuronal de radiancia descompuesto.

En la siguiente parte del trabajo, se combinan los modelos de objetos y affordances para mejorar la anticipación de las interacciones a corto plazo. Se introducen arquitecturas end-to-end que extienden los detectores de objetos clásicos para la anticipación, y se exploran estrategias para fundamentar las predicciones de anticipación en el comportamiento humano previo, utilizando para ello las affordances del entorno y las zonas de interacción preferente (interaction hotspots).

Por último, se amplía la percepción egocéntrica mediante la integración de representaciones multimodales. En primer lugar, se alinea el vídeo en primera persona con el lenguaje natural, localizando los límites temporales de las actividades en vídeos extensos. En segundo lugar, se conectan las perspectivas en primera y en tercera persona, reformulando la segmentación entre vistas como una tarea de correspondencia de máscaras de objetos, lo que permite una alineación eficaz de las representaciones de

objetos entre diferentes puntos de vista.

En conjunto, los métodos propuestos en esta tesis alcanzan el estado del arte en una amplia gama de tareas de percepción egocéntrica. Se espera que este trabajo sirva de inspiración para futuras investigaciones en el campo de la percepción visual en primera persona.

Index

Figures List

Table List

Chapter 1

Introduction

Egocentric vision shows an immense potential to revolutionize human-machine interaction by enabling systems to perceive and interpret the world from the user's perspective. Unlike traditional third-person computer vision, which observes the world from an external and static viewpoint, egocentric vision captures dynamic, intention-guided, and contextually rich user experiences. In first-person videos, the actor actively moves within a static scene, and sporadically interacts with a variety of objects; creating a tight integration between objects, actions, and the surrounding environment. This shift in perspective introduces both new opportunities and unique challenges, paving the way for novel technologies and emerging markets.

An inspiring application of egocentric vision is an assistive technology for visually impaired individuals [8, 9], as Figure 1.1 shows. In this case, the egocentric AI device provides context-aware, real-time support during everyday activities—such as locating misplaced objects, anticipating potential user errors, or guiding rehabilitation exercises. Beyond basic assistance, egocentric AI has the potential to enable expert-level guidance during skill acquisition. For example, Augmented Reality (AR) devices could project the correct body posture while learning football from an expert coach, guide a novice surgeon through a procedure with step-by-step instructions from a world-class expert, or assist with furniture assembly by detecting and correcting procedural mistakes. Furthermore, understanding human behavior from a first-person perspective will enable robots to act more autonomously and naturally. By modeling user intentions and the sequential structure of long-term goals, robots can proactively offer real-time assistance or prepare the environment to align with the user's objectives.

In the last decade, the advent of deep learning has substantially advanced visual perception. Early progress in convolutional neural networks [10] led to substantial improvements in image classification [11], object detection [12], and semantic segmentation [13]. Subsequently, the introduction of transformer-based architectures [14, 15] further improved performance by capturing global context and long-range dependencies

Figure 1.1: **Egocentric vision for visual assistive devices.** During the every-day activities, EGO-Granny provides context-aware assistance in real-time.

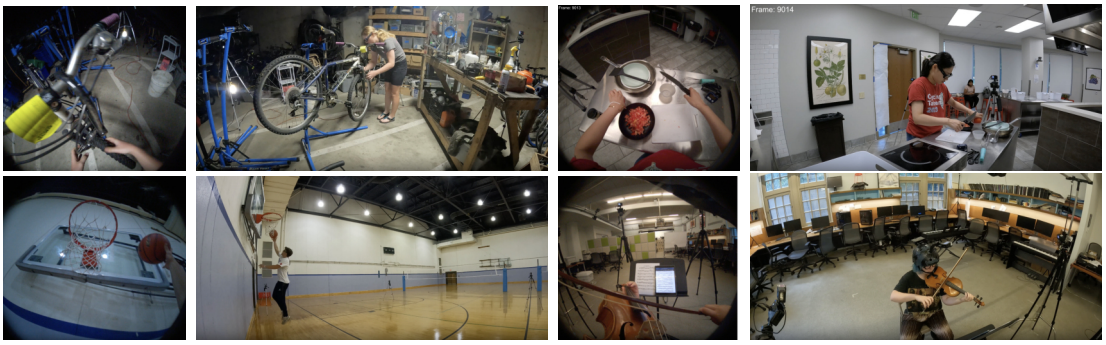


Figure 1.2: **Egocentric (first-person) vs. Exocentric (third-person) perspective in different scenarios.**

in visual data. More recently, the emergence of foundation models [16–18], pre-trained on large and diverse datasets, has enabled robust generalization across a wide range of downstream tasks.

However, in the context of egocentric vision, the tight coupling between the camera and the user's perspective introduces additional challenges while also requiring models to reason over fine-grained object manipulations. The first-person viewpoint is highly dynamic and subject to rapid and unpredictable camera motion due to natural head and body movements. This results in frequent motion blur, abrupt viewpoint changes, and partial occlusions. Unlike static third-person videos, the egocentric perspective captures unstructured and cluttered environments, with a strong emphasis on near-field interactions and a high proportion of non-informative frames. For instance, as shown in Figure 1.2, the first-person view highlights precise human–object interactions (e.g., the movement of a violin bow or a basketball shot), whereas the third-person view captures the full body pose together with the surrounding environment (e.g., the musician's posture or the basketball court). In robot learning, although several works [19–21] have demonstrated that third-person videos can effectively guide robot imitation and manipulation, the first-person perspective [22–24] is generally preferred, as it provides a more direct and natural view of human–object interactions.

Despite these challenges, progress in the field has been driven by the emergence of large-scale datasets such as EPIC-Kitchens [25], Ego4D [26], and Ego-Exo4D [27], which provide rich, realistic, and diverse first-person recordings of daily human activities. This large-scale data availability has enable progress in a wide variety of tasks such as action anticipation [28, 29], locomotion prediction [30, 31], next-active object detection [32, 33], action recognition [34, 35], object segmentation [36, 37], affordance segmentation [38, 39], episodic memory [40, 41], and object-state change [42–44].

Focused on improving egocentric perception, this thesis introduces methods for learning visual models of objects, affordances and environments, while also comple-

menting with other modalities such as language or the third-person perspective. Along the different chapters, *affordances* are continuously utilized to enhance core tasks such as detection, mapping and forecasting. To this end, I begin by introducing the term, reviewing the state of the art, and presenting the proposed methodologies.

1.1 Affordances

Definition and properties. Originally introduced by Gibson [45], affordances refer to the potential actions offered by the objects and environments to an agent, specifically in relation to the agent's motor and perception capabilities. Affordances encompass *all* possible actions enabled by the environment, making them well-suited to model complex and dynamic scenarios. For example, a chair may afford *sitting, standing on, hanging items, blocking a door* or serving as *climbing aid*. Gibson also emphasizes affordances as *relational properties*-that is, they emerge from the interaction between the agent, the object, and the surrounding environment-thus intrinsically linking perception and action. As Figure 1.3 shows, the affordances explain the different ways to interact with an object, and the potential actions that it offers. Due to this inherent connection between perceptual input and motor capabilities, the theory of affordances has inspired multiple works in robotics [46–51], enabling more natural interactions [49] and supporting more complex dexterous manipulations [50, 52]. In the context of computer vision, affordances provide a functional interpretation of objects [39, 53–64], enhancing generalization by abstracting beyond object category to action potential. They also support richer scene understanding [37, 65, 66], provide anticipatory cues for predicting future interactions [53, 67, 68] and serve as context-specific priors derived from expert demonstrations [69].

Detection. Classical approaches to affordance detection address the problem independently from the actual human-object interaction, resulting in "ungrounded" methods. These models are fully supervised using labeled masks, providing a precise localization very beneficial for applications such as robotic grasping [52, 70]. However, the requirement of manual masks limits the scalability of training datasets [54–57]. Given their similarity to semantic segmentation and object detection tasks, these methods typically adapt widely used architectures, such as encoder-decoder networks [71] or proposal-based detectors [58–60]. In Chapter 2, I begin my research following this ungrounded approach, proposing a Bayesian instance segmentation model that accurately segments objects parts according to their affordance labels. My main contribution is the introduction of Bayesian deep learning techniques, enabling the uncertainty quan-

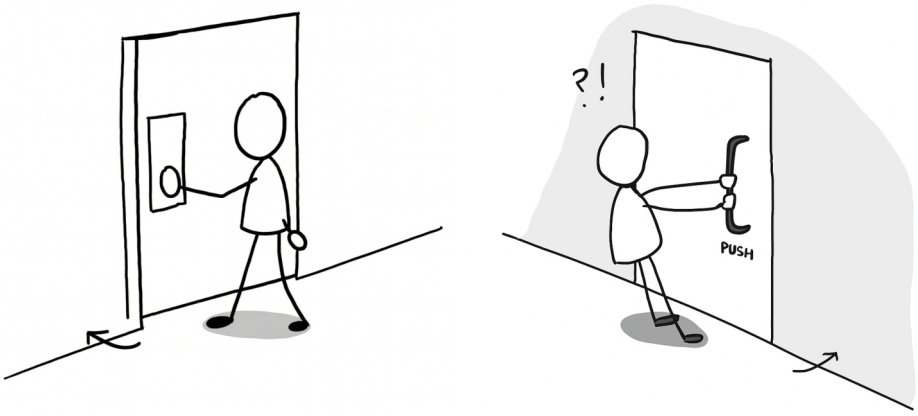


Figure 1.3: **Concept of affordances.** Left: a flat door only affords *pushing*. Right: a door with a bar handle affords both *pulling* or *pushing*.

tification, demonstrating improved calibration and superior performance compared to equivalent deterministic methods.

However, the close tight of the egocentric camera to the user provides a significant opportunity to capture fine-grained interactions directly by watching videos. This advantage is leveraged by "grounded" methods [39, 53, 57, 61–64], which learn affordance regions from human-object interactions using weak supervision based only on action labels. Demo2Vec [72] infers affordance keypoints from demonstration videos using action labels, Nagarajan et al. [53] identify interaction hotspots by extracting gradient-weighted attention maps from a video classifier, LOCATE [63] applies class activation mapping [73] on DINO-ViT [74] features to localize interaction regions and Yang. et al [64] transfer 2D image-based affordances to 3D point clouds using a multimodal paired dataset. Following, Luo et al. [57] proposed transferring affordance understanding from exocentric images to the egocentric perspective utilizing semantic labels as supervision, while Ego-Topo [39] constructed a topological graph of the scene to classify affordances from egocentric videos. However, existing grounded approaches have notable limitations in perception, as they are typically constrained to interaction hotspots [53, 63, 64, 72], image-level classification [39] or single-label segmentation [57]. To address these limitations, in Chapter 3, I introduce EPIC-Aff, the largest dataset of grounded affordances to date, which provides comprehensive, multi-label pixel-wise affordance annotations extracted from real-world interactions.

Mapping. When humans interact repeatedly within a familiar environment, we develop a form of physical cognition [75] that associates the spatial geometry with the objects and their attributions, effectively linking physical zones of interaction to the

activities they support. This process bridges the agent's capabilities with the surrounding environmental context. For example, a frying pan is associated with *cooking* when placed on a hob, whereas it affords *washing* when located near a dishwasher. Most existing works that relate semantics and geometry from egocentric video data primarily focus on semantic segmentation [76, 77] or tracking dynamic objects [40, 75, 78]. However, only a few studies investigate how action functionalities are grounded in specific 3D locations. Rhinehart et al. [65] learned 2D functional maps of actions, Liu et al. [66] recognized and localized activities within 3D voxel maps, and EPIC-Fields [37] registered dense camera poses in EPIC-Kitchens videos, demonstrating how object functionalities are tied to specific 3D regions in the environment. Motivated by these insights, in Chapter 3, I propose to integrate predicted 2D affordances with their corresponding 3D spatial location. By leveraging pre-extracted camera poses, this approach constructs a multi-label affordance map, enabling an environment representation in which affordances are expressed as relational properties between activity-centric regions. Rather than being a passive physical space, the environment becomes a set of activity-centric zones. Therefore, I leverage this multi-label affordance mapping to present a proof of concept for task-oriented navigation, in which the agent is guided from one location to another based on the final goal task. Then, in Chapter 5, I introduce a dynamic neural radiance field that distills video-language features from egocentric video. These features, pre-trained via contrastive learning by aligning narrative descriptions with videos, effectively localize the interaction hotspot corresponding to actions described in free-form language.

Forecasting. Studies such as [79] highlight that humans activities tend to exhibit consistent patterns across similar environments. Therefore, modeling the affordances distribution within a given environment provides informative cues about potential activities in novel, yet structurally similar, environments. Despite the promise of affordances for understanding human behavior, only a few works have explored their use in forecasting skills. Montesano et al. [51] predicted affordance effects for human-robot interaction, Nagarajan et al. [39] demonstrated that topological maps encoding scene affordances can improve long-term action anticipation and Koppula et al. [47] modeled object affordances to predict human behavior through the motion trajectories of both objects and agents. Other works [62, 67, 68, 80] track the past hand movements to predict the future hands location and infer the candidate regions with the next interaction. For instance, Liu et al. [67] highlighted how interaction hotspots predicted by forecasting hand motion can support action anticipation.

In Chapter 6, I propose leveraging environment affordances to enable grounded and

more accurate short-term object interaction anticipation. First, a robust representation of affordances in the training set is extracted, composed by activity-centric zones. Given a new video, the model associates it with similar environments via an attention mechanism, capturing the distribution of feasible interactions. While environment affordances indicate likely objects and potential actions, they do not specify where the interaction will occur. Following prior work [67, 68], my proposed method observes frames, hands, and objects to predict the interaction hotspot. This hotspot is then used to re-weight the confidence scores of short-term anticipation predictions, reducing the influence of false positives in regions unlikely to involve interaction.

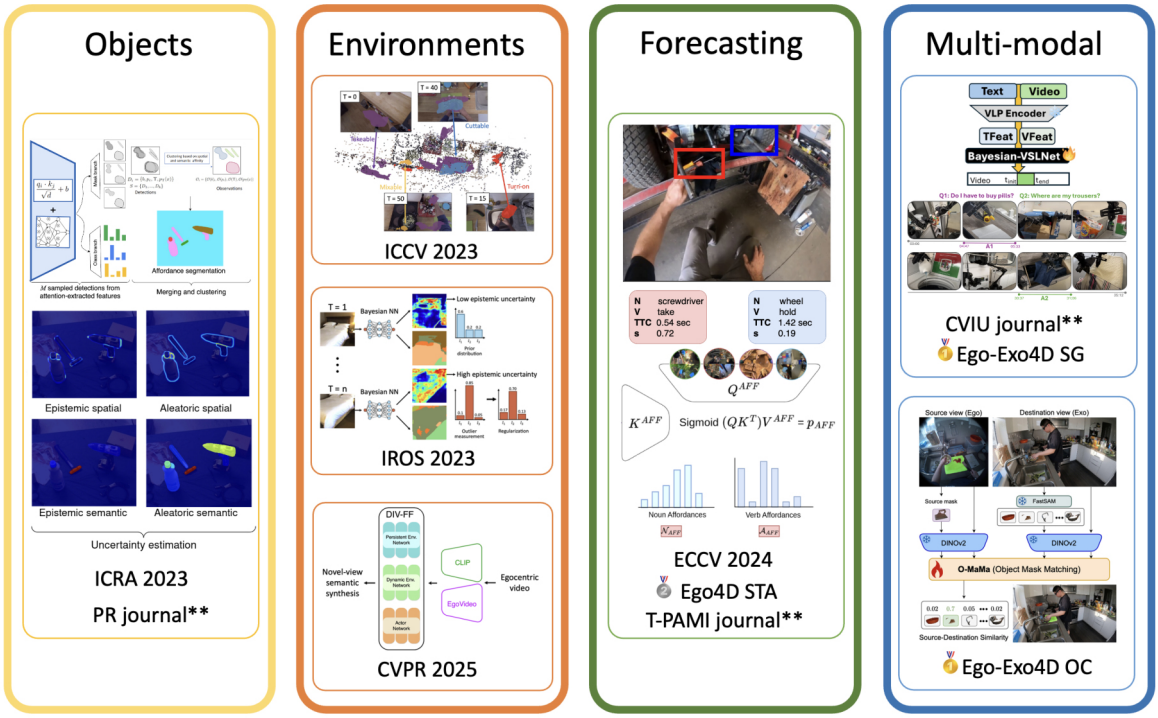


Figure 1.4: **My thesis research.** First, I propose to learn visual models of objects, affordances and environments for improving egocentric perception. Then, I apply these models in embodied tasks such as mapping or forecasting. In the last part, I explore how other modalities, such as language or third-person videos, complement egocentric perception. I denote as ** those works under review.

1.2 Outline

In this thesis, I advance egocentric perception by developing methods to learn visual models of objects, affordances, and environments, complemented by multi-modal information such as language and third-person viewpoints. Figure 1.4 shows a visual summary of my research. I consistently leverage affordances to improve key tasks such as detection, mapping, and forecasting. Specifically, my research is structured around the following problems:

– **Learning visual models of objects.** I first study in Chapter 2 how to segment object-part affordances and quantify the uncertainty using Bayesian deep learning techniques.

– **Learning visual models of environments.** Individual objects are not isolated in the real world; rather, they are part of a physical space the -environment- [81] that creates interconnections through their spatial configuration and functional relationships. To capture this structure, I propose a series of models for environment representation in Chapters 3 to 5.

– **Forecasting the short-term object interaction.** While the previous models focus on understanding the present or encoding the perceived environment, in Chapter 6 I introduce methods to anticipate the next object interaction.

– **Exploiting multi-modal cues for complementing egocentric vision.** I investigate how additional modalities—such as natural language (Chapter 7) or synchronized third-person videos (Chapter 8)—can complement egocentric visual understanding.

1.2.1 Learning visual models of objects.

I begin this thesis by learning visual models of object-part affordances-i.e., specific regions in the image that afford potential interactions. As a first step toward egocentric perception, a simple world model can be modeled as a set of functional objects distributed around the actor, where each object part affords a distinct interaction. This modeling is particularly relevant in visual prostheses for visually impaired people [9], where accurate affordance localization is essential to indicate crucial actionable regions—e.g., the handle of a knife affords *grasping*, while the blade affords *cutting*. Similarly, AR devices can provide expert-based guidance by highlighting relevant components—such as illuminating parts of a new engine for a novice mechanic—thereby facilitating intuitive procedural understanding and execution.

Particularly, in Chapter 2, I introduce a Bayesian instance segmentation model that accurately segments objects parts with the respective potential interaction label and quantifies the uncertainty of the predictions. As traditional deep learning models are overconfident sensors with low interpretability, they produce misclassified predictions with high confidence [82]. Therefore, incorporating Bayesian deep learning techniques is the basis for a more interpretable and robust egocentric perception [83]. My proposed uncertainty estimation is decomposed in semantic (associated with the affordance class) and spatial (related with the mask probability of encompassing an object). In each term, the contributions of the epistemic and aleatoric components are also computed, respectively. This part is presented in the following works:

– Lorenzo Mur-Labadia, Ruben Martinez-Cantin, and Jose J Guerrero. Bayesian deep learning for affordance segmentation in images. In *IEEE International Conference on Robotics and Automation (ICRA)* Core Ranking A*, pages 6981–6987, 2023

– Lorenzo Mur-Labadia, Ruben Martinez-Cantin, and Jose J Guerrero. Uncertainty Estimation in Instance Segmentation of Affordances via Bayesian Visual

Transformers. In *Pattern Recognition* Q1, (Under Review)

1.2.2 Learning visual models of environments

While learning visual models of the object affordances provides valuable information about their functionalities, it is insufficient for understanding the real-world. Individual objects are not isolated; rather, they are embedded within a broader physical space—the environment—which defines the context in which interactions occur [81]. The environment encompasses the spatial configuration of objects, their relative positions to the observer, and the functional relationships that emerge from this structure. For example, in human-robot collaboration, the environment is the common basis for grounding the different interactions. Similarly, assistive devices with such capabilities could track the 3D location of dynamic objects and improve their episodic memory capabilities, delivering spatial audio cues with the location of a specific lost object. In the following chapters, I introduce different approaches for capturing this structure, including multi-label affordance point-clouds, semantic voxel maps and dynamic image-video feature fusion fields.

In Chapter 3, I begin by constructing a multi-label affordance map that spatially links activity-centric zones. I first leverage this environment representation to develop a pipeline that automatically generates multi-label, pixel-wise, and grounded affordance annotations; representing a significant improvement over the ungrounded, single-label model introduced in Chapter 2. Next, I extend classical segmentation models with a multi-label prediction head to capture more diverse interaction regions, based on the assumption that a single object can afford multiple actions. Using camera pose estimation, I project the predicted affordance regions onto a 3D point cloud to generate a spatial map. Then, I show that this multi-label mapping can guide autonomous agents in task-oriented navigation. This work was published in ICCV 2023:

– Lorenzo Mur-Labadia, Jose J Guerrero, and Ruben Martinez-Cantin. Multi-label affordance mapping from egocentric vision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* Core Ranking A*, pages 5238–5249, 2023.

However, as this mapping approach assumes an static scene, it does not incorporate any temporal aggregation strategy; i.e., the predictions are not fused into a voxel grid or 3D surface, but they are simply projected (and accumulated) onto the 3D point cloud. Temporal fusion cannot be performed directly in the point cloud domain as the reconstructed 3D points lack stable correspondences across frames due to dynamic scene changes and depth estimation variability, which are required for an effective temporal

fusion. To address these limitations, in Chapter 4, I introduce a Bayesian semantic mapping that fuses predictions in a semantic voxel map. First, I propose a novel fusion strategy that regularizes the observations to mitigate the impact of prediction bias. Second, I incorporate epistemic uncertainty from the predictive distribution of the Bayesian neural network to reduce the influence of overconfident outlier predictions and to yield a more robust semantic voxel-map. This work, carried out in collaboration with David Morilla-Cabello, was published at IROS 2023:

– David Morilla-Cabello*, Lorenzo Mur-Labadia*, Ruben Martinez-Cantin, and Eduardo Montijano. Robust fusion for Bayesian semantic mapping. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS).* Core Ranking A, 2023

However, these two proposed solutions are based on *explicit* functions of the environment (point clouds and voxel maps, respectively). These *explicit* functions directly store the geometry and semantic information of the scene in a discrete form, where each voxel or point carries a semantic label. While such representations offer a straightforward approach to associate semantic information with the geometric structure, they are subject to inherent limitations. First, the scene size grows linearly with the number of frames, making scalability challenging. Second, the encoded information is typically limited to discrete semantic labels, where these representations tend to be sparse in less frequently observed areas. Finally, *explicit* functions are inherently rigid, which complicates the modeling and tracking of dynamic objects, widely present in egocentric videos. In contrast, humans develop an *implicit* model of the environment through repeated interactions, allowing them to infer unobserved structures, predict object behavior, and adapt to dynamic changes. *Implicit* functions, particularly those based on neural fields, provide a powerful alternative by capturing continuous and compact encodings of both geometry and semantics without relying on *explicit* discretization. Here, the scene is defined as the level set of a continuous function parameterized by a neural network, which maps spatial coordinates to scene properties such as occupancy or semantics.

Building on this insight, my latest contribution introduces an *implicit* neural representation capable of jointly modeling geometry, appearance, and semantic understanding from egocentric videos. Specifically, I propose Dynamic Image Video Feature Fields (DIV-FF), a language-embedded feature field that hierarchically decomposes the scene into three components: the actor, the dynamic objects, and the persistent environment. Although this provides a persistent long-term representation, it is dynamically updated as the user interacts, enabling a precise record of the location and state of

dynamic objects at every moment. This work was presented, and selected as Highlight, during the CVPR 2025 conference:

– Lorenzo Mur-Labadia, Jose J. Guerrero, and Ruben Martinez-Cantin. DIV-FF: Dynamic Image-Video Feature Fields for environment understanding in egocentric videos. Highlight in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* Core Ranking A*, 2025

1.2.3 Forecasting the short-term object interaction.

Anticipating the future is a fundamental embodied skill for autonomous agents, allowing them to act proactively in dynamic environments. For instance, a home assistant robot can plan tasks toward long-term goals, such as preparing ingredients in advance for a cooking recipe. In human-robot collaboration, anticipation enables temporal coordination and improves safety during complex joint manipulation. Similarly, a visual assistive device can help users avoid small obstacles, reducing the risk of collisions and falls. In Chapter 6, I address the Short-Term object interaction Anticipation (STA) task [26], which involves the simultaneous prediction of the action and object category, the object's future bounding box, and the time to contact.

My first contributions are STAformer and STAformer++, two novel end-to-end architectures specifically designed for the STA task. These models take as input a high-resolution image, complemented by a low-resolution video. The architectures incorporate attention-based components to facilitate image-video fusion, where the image provides fine-grained details of the scene complemented with the action dynamics from the video. The STAformer prediction head is based on Faster-RCNN [84], a proposal-based convolutional object detection architecture. In contrast, STAformer++ is a more advance version that leverages Detection Transformers (DETR) [85], a transformer-based approach that replaces traditional region proposals and post-processing with a direct set prediction mechanism, leading to improved detection accuracy in an end-to-end pipeline.

So far, I described how to detect affordances with uncertainty quantification (Chapter 2) or building environment representations (Chapters 3 to 5). While these models primarily enhance perceptual capabilities, the following section focuses on leveraging affordances as strong priors that encode future interactions. The core intuition is as follows: since affordances represent the set of possible actions in a given environment, an ideal affordance representation should inherently capture the user's next interaction. In this sense, affordances provide a natural bridge between perception and forecasting, by grounding predictions in past observed human behavior. The first approach, described

in Section 6.5.2, first builds an affordance database by extracting activity-centric zones from the training set videos. Then, an input observation is matched to a set of zones in the database. Only at inference time, the environment affordance priors—represented as verb and noun probability distributions—are used to refine the predicted verb and noun probabilities. A more advanced version, described in Section 6.5.3, also integrates environment affordances during training. Here, an attention mechanism links the input observation to relevant candidates in the affordance database. This learned prior guides the model toward more accurate and semantically grounded predictions. Lastly, I re-weight the STA confidence scores based on the probability values at the bounding box center coordinates, reducing confidence in false positive predictions that fall far from the interaction hotspot. A seminal part of this work was presented in ECCV 2024, and a follow-up version is currently under review. Besides, the STAformer architecture achieved the 2^{nd} position at the Ego4D STA Challenge during the EgoVIS Workshop at CVPR 2024.

- Lorenzo Mur-Labadia, Ruben Martinez-Cantin, Jose J Guerrero, Giovanni Maria Farinella, and Antonino Furnari. ZARRIO@ Ego4D Short Term Object Interaction Anticipation Challenge: Leveraging Affordances and Attention-based models for STA. 2^{nd} Position at Ego4D STA Challenge during EgoVIS Workshop CVPR 2024. *arXiv preprint arXiv:2407.04369*, 2024.

- Lorenzo Mur-Labadia, Ruben Martinez-Cantin, Jose J Guerrero, Giovanni Maria Farinella, and Antonino Furnari. Aff-ttention! affordances and attention models for short-term object interaction anticipation. In *European Conference on Computer Vision.* Core Ranking A*, pages 167–184. Springer, 2024

- Lorenzo Mur-Labadia, Ruben Martinez-Cantin, Jose J Guerrero, Giovanni Maria Farinella, and Antonino Furnari. Integrating Affordances and Attention models for Short-Term Object Interaction Anticipation. In *IEEE Transactions on Pattern Analysis and Machine Intelligence.* (Under Review)

1.2.4 Exploiting multi-modal cues for complementing egocentric vision.

In the previous chapters, I introduce different alternatives for learning objects, environments and affordances by watching exclusively egocentric videos. However, recent wearable devices [86] are equipped with extra sensors such as microphones, wireless connectivity, IMU and monocular cameras, complementing first-person video with other

modalities. In this final section, the proposed approaches combine egocentric video with language (Chapter 7) and the exocentric perspective (Chapter 8), respectively.

Aligning egocentric video with natural language is a crucial step to equip autonomous agents with a more natural communication. With the advent of Large Language Models (LLMs), this alignment has emerged as a promising avenue for bridging vision and language in real-world settings. For instance, a visual assistive device could retrieve the location of a specific object in a video based on user descriptions. Similarly, robots could exploit internet-scale instructional videos by parsing them into structure sequences of steps, facilitating a posterior imitation learning [87]. Specifically, I attempt the Step Grounding (SG) [3] task, which involves localizing the temporal boundaries of specific activities—described in free-form natural language—within long, untrimmed egocentric videos. Unlike traditional temporal action segmentation approaches [88–93], which are constrained to a closed set of action labels, the SG task benefits from the flexibility of open-vocabulary language descriptions. Moreover, it requires handling extremely long video sequences—often spanning several hours—with a long-tail distribution of step durations.

To address these challenges, I introduce Bayesian-VSLNet, a test-time refinement strategy that integrates temporal-order priors into the predictions. This simple yet effective refinement is guided by Bayes' theorem to align predictions according to the sequential order of steps, correcting common issues such as step repetition and cyclic actions. As a result, this method achieves state-of-the-art results on the Ego4D Goal-Step benchmark [3]. The work introduced in Chapter 7, done in collaboration with Carlos Plou, won the SG Ego4D challenge at the EgoVis Workshop during CVPR 2024. Next, a follow-up journal version is under review.

– Carlos Plou*, Lorenzo Mur-Labadia*, Ruben Martinez-Cantin, and Ana C Murillo. CarLor@ Ego4D Step Grounding Challenge: Bayesian temporal-order priors for test time refinement. Winner Solution at Ego4D Step Grounding challenge during EgoVis Workshop CVPR 2024 *arXiv preprint arXiv:2406.09575*, 2024.

– Carlos Plou*, Lorenzo Mur-Labadia*, Ruben Martinez-Cantin, and Ana C Murillo. Temporal Video Segmentation with Natural Language using Text-Video Cross Attention and Bayesian Order-priors. In *Recent Advances in Assistive Computer Vision and Robotics Special Issue of Computer Vision and Image Understanding Journal*, (Under Review)

As a final step in this thesis, I investigate how to integrate the egocentric perspective with the third-person view. Understanding the world from multiple viewpoints

is essential for intelligent systems operating in coordination, yet segmenting common objects across different views remains an open challenge. For instance, autonomous robots could replicate more effectively human demonstrations, or augmented reality systems could leverage external cameras to guide full-body pose estimation during supervised physical training. In this final part, I address the Ego-Exo4D Correspondence task, which aims to predict an object's mask in one view given a query mask from the other. Unlike traditional segmentation problems, this task presents additional challenges, including drastic viewpoint changes, scale variations, occlusions, and domain shifts caused by differences in camera optics and imaging conditions.

My solution, simple yet effective, redefines the complex cross-image segmentation task by reformulating it as object mask matching across egocentric and exocentric views, leveraging the powerful zero-shot segmentation capabilities of Segment Anything Models (SAM) [94, 95]. Object mask candidates are first extracted in the target view using FastSAM. Each candidate is then encoded with a Mask-Context Encoder that pools dense DINOv2 features, capturing both discriminative object details and surrounding contextual information. To enhance discrimination in cluttered scenes, a Hard Negative Adjacent Mining strategy selects neighboring objects as challenging negatives. Cross-view alignment is further improved through a novel Ego↔Exo Cross-Attention mechanism. The model is trained using a Mask Matching Contrastive Loss, which enables effective cross-view object correspondence by aligning both global scene context and fine-grained object features. This work, in collaboration with Maria Santos-Villafranca won the Ego-Exo4D Correspondence Challenge at the EgoVis Workshop during CVPR 2025, and a extended version was presented during ICCV 2025.

– Lorenzo Mur-Labadia*, Maria Santos-Villafranca*, Jesus Bermúdez Cameo, Alejandro Perez Yus, Ruben Martinez-Cantin and Jose J. Guerrero. O-MaMa @ Ego-Exo4D Correspondence Challenge: Object Mask Matching between Egocentric and Exocentric Views. Winner solution at EgoExo4D Correspondences challenge during EgoVis Workshop CVPR 2025.

– Lorenzo Mur-Labadia*, Maria Santos-Villafranca*, Jesus Bermúdez Cameo, Alejandro Perez Yus, Ruben Martinez-Cantin and Jose J. Guerrero. O-MaMa: Learning Object Mask Matching between Egocentric and Exocentric Views. In *International Conference on Computer Vision*. Core Ranking A*, 2025.

1.3 Other Merits

In addition to the aforementioned publications, I have carried out several other academic activities, including student supervision, teaching collaborations, and international mobility.

Student Supervision

1. **Miguel Marcos**. Master's Thesis: *GPianoroll: a Deep Learning System with Human Feedback for Music Generation*, MSc in Robotics, Graphics and Computer Vision. Final grade: 9.5/10, with honors. This work was later extended into a journal article currently under review: *Rotational Random Embedding Bayesian Optimization for Human-in-the-Loop Music Generation*, by Miguel Marcos, Lorenzo Mur-Labadia, and Rubén Martínez-Cantín, 2025.

2. **Alejandro de Nova Guerrero**. Master's Thesis: *Language-aware Neural Feature Fusion Fields for Egocentric Video*, MSc in Robotics, Graphics and Computer Vision. Final grade: 9.5/10.

3. **María Peribáñez**. Research Internship: *Revisiting Environments through Language-Aware Neural Feature Fields*, MSc in Robotics, Graphics and Computer Vision, 2025.

4. **Diego Lara**. Bachelor's Thesis: *Deep Learning Applications for Visual Assistive Devices*, BSc in Industrial Technologies Engineering, 2025.

Teaching Collaborations

1. 2024/2025 – *Machine Learning*, MSc in Robotics, Graphics and Computer Vision.

2. 2023/2024 – *Automatic Systems*, BSc in Mechanical Engineering.

3. 2022/2023 – *Industrial Robotics and Automation*, BSc in Industrial Technologies Engineering.

4. 2022/2023 – *Computer Vision and Robotics*, MSc in Industrial Technologies Engineering.

International Experience I conducted a research stay at the University of Catania (Italy) from September to December 2023. Additionally, I participated in the 2024 edition of the International Computer Vision Summer School (ICVSS). Most recently,

I will carry out a research internship at META Fundamental Artificial Intelligence Research (FAIR) in Paris, from July to December 2025, under the supervision of Adrien Bardes and Yann Le Cun working on the JEPA team.

Part I

Learning visual models of objects

Chapter 2

Uncertainty Estimation in Instance Segmentation of Affordances

I begin this thesis by learning visual models of object-part affordances—i.e., specific regions in the image that afford potential interactions. This modeling is particularly relevant in visual prostheses for visually impaired people [9], where accurate affordance localization is essential to indicate crucial actionable regions—e.g., the handle of a knife affords *grasping*, while the blade affords *cutting*. Particularly, I introduce a Bayesian instance segmentation model that accurately segments objects parts with the respective potential interaction label and quantifies the uncertainty of the predictions. The uncertainty estimation is decomposed in semantic (associated with the affordance class) and spatial (related with the mask probability of encompassing an object). In each term, the contributions of the epistemic and aleatoric components are also computed, respectively.

2.1 Introduction

Rooted in Gibson's foundational theory of perception, affordances are the potential actions that the environment offers to the agent based on its motor capabilities [45]. When transferred to autonomous robots [46], affordances would make them interact more naturally [47–49, 58] and obtain better grasping capabilities [50, 52]. In Augmented Reality (AR) systems, leveraging actions learned from expert demonstrations enhances user experience by providing context-specific guidance [69]. For instance, an AR assistive device could illuminate parts of a new engine model for a mechanic, indicating the appropriate procedural actions, thereby facilitating a more intuitive repair process. Similarly, the constrained resolution of visual assistive devices for the visually impaired [9] necessitates the isolation of specific regions indicative of potential user actions. Such applications require a detailed scene understanding that prioritizes lo-

calized affordance regions compared with diffuse saliency maps [53,72]. In this chapter, we segment the object boundaries and classify the affordable action associated with each object part, obtaining pixel-wise precision with a detailed estimation of the spatial and semantic uncertainty.

Recent advances in deep learning have enhanced visual perception: object detection [12,96,97], action anticipation [28,29,98,99], affordances detection [39,53,58] or hand-object interaction anticipation [32,62,100,101] are some examples of the improvements in this area. However, despite the significant advances in performance, deep neural networks are overconfident sensors with low interpretability, producing misclassified predictions with high confidence [82]. In this chapter, we overcome these deficiencies with Bayesian deep learning alternatives, which produce better-calibrated probabilities with uncertainty estimation, obtaining a more interpretable and robust perception [83,102].

Uncertainty estimation in the problem of instance segmentation of affordances aims to accurately segment object parts, classify them based on potential interactions, and quantify uncertainty for a more robust perception. A Bayesian detection model, inspired in Mask-RCNN [12] is extended with intermediate sampling layers for enhanced performance and uncertainty quantification. During inference, our proposed sampling-based approach executes N forward passes, generating multiple predictions—termed detections—from the same input. The detections are then clustered based on their spatial and semantic affinity in observations, thereby enhancing the model's reliability. Comparing the distribution of the grouped observations, we extract the epistemic and aleatoric variance both at spatial and semantic levels. The spatial uncertainty is associated with the binary probability of having a mask around an object, thus revealing pixel-level spatial uncertainty. Semantic uncertainty is derived from the variability within the class probability vectors, highlighting discrepancies in assigning an afforded action to the object part. In summary, our contributions in this chapter are as follows:

- We extend affordance segmentation to a probabilistic stage, extracting a per-pixel estimation of the aleatoric and epistemic variance at the spatial level.

- We compare different techniques for uncertainty estimation: Monte-Carlo Dropout (MC-Dropout) [103], Mask Ensembles (Mask-Ens) [104], Deep Ensembles (Deep-Ens) [105], and Snapshot Ensembles (Snap-Shot) [106].

- We design the novel Probability-based Mask Quality (PMQ) metric that evaluates the uncertainty estimation of the predictions.

- The results achieve a new state of the art at 90.6 % on the F_β^w score, which rep-

resents +7.4 percentage points of improvements compared with previous works in the IIT-Aff dataset. We report detailed ablation studies on the Bayesian techniques and the disposition of the sampling layers within the architecture, showing the benefits of Bayesian models over their respective deterministic versions.

The results presented in this chapter were initially published in ICRA 2023 and later extended in a follow-up version currently under review at Pattern Recognition Letters.

– Lorenzo Mur-Labadia, Ruben Martinez-Cantin, and Jose J Guerrero. Bayesian deep learning for affordance segmentation in images. In *IEEE International Conference on Robotics and Automation (ICRA)* Core Ranking A*, pages 6981–6987, 2023

– Lorenzo Mur-Labadia, Ruben Martinez-Cantin, and Jose J Guerrero. Uncertainty Estimation in Instance Segmentation of Affordances via Bayesian Visual Transformers. In *Pattern Recognition* Q1, (Under Review)

2.2 Related works

2.2.1 Visual perception of affordances

Gibson's psychological theory of affordances [107] has inspired multiple computer vision works due to its applications in robotics [47–50, 52, 58, 108], augmented reality systems [9], or scene understanding [38, 39, 69]. Koppula et al. [109] use support vector machines to model jointly human activities and object affordances for robot-assisted tasks. Following, Koppula et al. [47] anticipate future human activities through object affordances. The work by Montesano et al. [108] addresses the learning of affordances through robot-environment interactions using a Bayesian network model. Detailed affordance masks are exploited by Yang et al. [50] for learning a better robot manipulation by selecting accurately between *pickable* and *placeable* objects. Ego-Topo [53] constructed an affordance topological obtaining activity-centric zones on the nodes and used this representation for long-term activity recognition, showing the potential of affordances in video understanding. Mur-Labadia et al. [38] build a 3D multi-label mapping of affordances, which was exploited in task-oriented path planning.

In terms of perception affordance models, supervised approaches [54, 55, 58–60, 71] learn by direct supervision of manual annotated affordances masks, providing pixel-wise precision and more accurate location, very beneficial for robot grasping [52, 70]. The seminal UMD dataset [54] provides mask annotations for RGB-D images, but

the dataset is composed of pre-defined objects captured in isolated conditions. The IIT-Aff dataset [55] comprises real-world objects in multiple environments, enabling a better generalization in real conditions. Synthetic data reduces the cost of manual annotations, but the simplicity of the supervised synthetic images and the dataset gap limits its performance in real-world scenarios [110]. Nguyen et al. [71] introduced a convolutional encoder-decoder to learn affordances from a deep latent space, which was further refined using feature maps with Conditional Random Fields in a post-processing step in a subsequent work [55]. Affordance-Net [58] adapted an object detection for affordance perception with three significant contributions: a multi-task loss function, a resizing strategy, and a sequence of deconvolutional layers for producing high-resolution masks. Minh et al. [59] extended Do et al. [58] methodology by integrating ResNet101 for superior feature extraction and the Feature Pyramid Networks (FPN) to amalgamate multi-scale features, thereby elevating segmentation precision. Caselles et al. [60] demonstrated the effectiveness of reusing standard instance segmentation models for affordance perception. In our previous work [111], we shown that Bayesian models outperform their deterministic counterparts due to better mask refinement and generalization capabilities. Weakly supervised approaches learn from watching human-object interactions [39,72,112] but obtain diffuse interaction hotspots rather than precise affordance masks on the object part. The AGD20K [57] is a large-scale affordance dataset with egocentric and exocentric images, but it contains only the affordance label per image. Demo2Vec [72] inferred affordance key points on a static object by watching demonstration videos of the respective interaction using only the action label as supervision. Similarly, Nagarajan et al. [53] obtained interaction hotspots by extracting gradient-weighted attention maps from training an action classifier on videos. LOCATE [63] extracted interaction regions from an exocentric observation using the class activation mapping technique [73] from DINO-ViT features [74]. Zhai et al. [113] used collaboration learning for extracting common characteristics between objects with the same affordance. Yang et al. [64] presented a domain transfer from 2D interactions in images to the 3D object point cloud supported with a novel dataset with paired point cloud-image affordance data.

2.2.2 Neural Network Uncertainty Quantification

In the last years, the growing interest in uncertainty quantification has originated different approaches [114–119]. Sampling-based methods [103,104] incorporate intermediate sampling layers to train a single model, quantifying the uncertainty by comparing the divergence in the different predictions. Similarly, ensemble methods [105,106] require training multiple models to capture the diversity in the predictions. Feature-space

techniques [120–124] estimate the uncertainty with a single-pass by measuring the distance or density of the sample compared with the training data distribution in the new space.

Following the success of transformers in Natural Language Processing (NLP) tasks [14], attention-based architectures have been extended for uncertainty quantification. BayesFormer [125] applies a variational inference-based dropout framework within the transformer architecture, demonstrating its efficacy in NLP. On active learning tasks, Gleave et al. [126] propose a last-layer ensemble for uncertainty estimation. However, their estimated epistemic uncertainty is poorly calibrated since the ensemble models are very similar to each other. The proposed Uncertainty-Guided Transformer Reasoning by Yang et al. [127] uses the uncertainty estimation to guide the transformer and learn better the ambiguities present in camouflaged object detection. Following, [128] uses epistemic uncertainty estimation to guide both the training and inference processes. They approximate the scaled dot-product by sampling from a Gaussian distribution with an NLL loss. Uncertainty-based attention can identify more informative regions and improve the model robustness [129].

The uncertainty estimation increases the interpretability and robustness of deep neural networks, with applications in the robotics perception, mapping, and planning stages. In terms of perception, as Kendall et al. [83] shown on computer vision applications, the epistemic term appears in challenging pixels or occluded objects out of the distribution. Oppositely, aleatoric uncertainty appears in the object's contours or in far-away regions with higher camera noise [130, 131]. Other works show the benefits of uncertainty quantification in different tasks. Miller et al. [132] employ the epistemic uncertainty to identify false positive detections in object classification. Morilla et al. [102] combine Bayesian semantic predictions to mitigate the effect of overconfident outlier predictions and build more robust semantic maps. Feng et al. [133] reflect the environmental noise in Lidar sensors [134]. In active learning, the uncertainty is useful for selecting the most informative samples [135, 136] or for guiding the navigation in unseen areas [137].

2.3 Bayesian Instance Segmentation of Affordances

We start by detailing the theoretical foundations behind sampling-based methods for Bayesian deep learning. Then, we introduce our novel Bayesian instance segmentation model, capable of extracting uncertainty and segmenting object affordances. Our goal is to capture epistemic and aleatoric uncertainty [138] for both the semantic category (i.e., the affordance class) and the mask object contour.

2.3.1 Uncertainty estimation in Bayesian deep learning

Let $f(\cdot, \mathbf{w})$ be a deep neural network with model parameters $\mathbf{w}$ and $\mathcal{D} = \{\mathbf{X}, \mathbf{Y}\}$ be a training dataset with different inputs images $\mathbf{X} = \{\mathbf{X}_1, ..., \mathbf{X}_N\}$ and labels $\mathbf{Y} = \{\mathbf{y}_1, ..., \mathbf{y}_N\}$, where N is the dataset size. Placing a prior distribution $p(\mathbf{w})$ on the neural networks weights of the Bayesian neural network, the posterior distribution over the model parameters results in:

$$p(\mathbf{w}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathbf{w})p(\mathbf{w})}{p(\mathcal{D})} = \frac{\prod_{i=1}^{N} p(\mathbf{y}_i|\mathbf{X}_i, \mathbf{w})p(\mathbf{w})}{p(\mathcal{D})}. \tag{2.1}$$

The problem of Bayesian deep learning is that the true posterior $p(\mathbf{w}|\mathcal{D})$ in the general case is intractable, as the marginal probability $p(\mathcal{D})$ requires solving a high dimensional integral:

$$p(\mathcal{D}) = \int_{\mathcal{W}} p(\mathcal{D}|\mathbf{w})p(\mathbf{w})d\mathbf{w}. \tag{2.2}$$

Furthermore, in the case of uncertainty estimation, we need to propagate the model uncertainty to the predictive distribution, to see how it affects the network predictions. For that, we also need to solve another high dimensional integral for the predictive posterior distribution,

$$p(y^*|X^*, \mathcal{D}) = \int_{\mathcal{W}} p(y^*|X^*, \mathbf{w})p(\mathbf{w}|\mathcal{D})d\mathbf{w}, \tag{2.3}$$

where y^* corresponds to the prediction of a new image $\mathbf{X}^*$. As commented before, we are going to study a set of sample-based approaches to approximate the true posterior distribution $p(\mathbf{w}|\mathcal{D})$. This results in computing a sampling distribution of network weights $\{\mathbf{w}_m\}_{m=1}^{M} \sim q(\mathbf{w})$ where $q(\mathbf{w})$ is the approximate distribution of the true posterior. Thus, this results in the following approximation of equation (2.3):

$$\hat{p}(y^*|X^*, \mathcal{D}) \approx \frac{1}{M} \sum_{m=1}^{M} f(\mathbf{X}^*, \mathbf{w}_m). \tag{2.4}$$

For simplicity, in the following sections, we consider that for every detection, the network outputs a class probability vector $\mathbf{p}_m = f(\mathbf{X}^*, \mathbf{w}_m)$ associated to a the m-th sample. Therefore, we rewrite equation (2.4) as $\bar{\mathbf{p}} = \frac{1}{M} \sum_{m=1}^{M} \mathbf{p}_m$.

2.3.2 Sampling-based and ensemble methods

We compare several sampling-based and ensemble methods for quantifying uncertainty, including Monte Carlo dropout (MC-Dropout) [103], Mask Ensembles (Mask-Ens) [104], Deep Ensembles (Deep-Ens) [105], and Snapshot Ensembles (Snap-Ens) [106]. A common characteristic of these methods is that, at inference time, they perform M multiple forward passes, resulting in a set of M samples of the prediction, $\{\mathbf{p}_m\}_{m=1}^{M}$, that

approximate the posterior predictive distribution $\hat{p}(y^*|X^*, \mathcal{D})$, as shown in Equation (2.4).

Monte Carlo [103, 115] approaches generate a sample distribution of the network weights $\{\mathbf{w}_m\}_{m=1}^M$ from the posterior distribution. MC-Dropout, introduced by Gal et al. [103], obtains M weight samples by first training the network and obtaining an optimal set of parameters from the maximum posterior distribution $\mathbf{w}^* = \mathbf{w}_{MAP}$. The sampling distribution is then generated by applying dropout at inference, effectively multiplying the optimal weights by a Bernoulli-distributed mask, $\mathbf{Ber}(d)$, which determines whether a neuron is dropped. This results in the sampled weights $\{\mathbf{w}_m\}_{m=1}^M$, where $\mathbf{w}_m = \mathbf{w}^* \cdot z_m$ with $z_m \sim \mathbf{Ber}(d)$. In practice, this is equivalent to keeping dropout layers active at inference. While MC-Dropout effectively minimizes the Kullback-Leibler divergence between the true posterior $p(\mathbf{w}|\mathcal{D})$ and a tractable variational approximation $q_\theta(\mathbf{w})$ [103], the correlation between $\mathbf{w}_m$ samples may lead to an underestimation of the uncertainty.

Mask-Ens [104] also trains a single model with intermediate sampling layers, similar to MC-Dropout. However, unlike the stochastic nature of random dropout, Mask-Ens pre-computes a set of binary masks that, when applied, create M sampling layers with mask-based dropout. Mask-Ens offer two key advantages over MC-Dropout. First, the sampling masks exhibit lower correlation, leading to improved calibration performance [104]. Second, because the masks can be predefined, the layer size can be adjusted to compensate for dropped neurons, ensuring a constant number of active neurons equivalent to the original network without dropout. The mask generation process incorporates a controllable degree of overlap, regulated by a scale parameter s. At inference, the individual model weights $\{\mathbf{w}_m\}_{m=1}^M$ are obtained as $\mathbf{w}_m = \mathbf{w}^* \cdot \text{Mask}(s, m)$.

Deep-Ens [105] trains an ensemble of M models with randomly initialized network parameters, dataset shuffling, and random data augmentation. This method ensures that the weights $\mathbf{w}_m$ converge to different local optima. Notably, it effectively captures the multi-modality of the posterior distribution $p(\mathbf{w}|\mathcal{D})$, even with a relatively small number of ensemble models [105, 139].

Snap-Ens [106] mitigates the computational overhead of Deep-Ens, which requires training M models from scratch, by instead training all M models sequentially within a single training loop. This is achieved by cyclically increasing the learning rate once a model has converged, leading to M distinct solutions that capture multiple local minima throughout training. Snap-Ens does not require architectural modifications, as $\mathbf{w}_m$ are extracted at different training stages, making it easily applicable to pre-trained models.

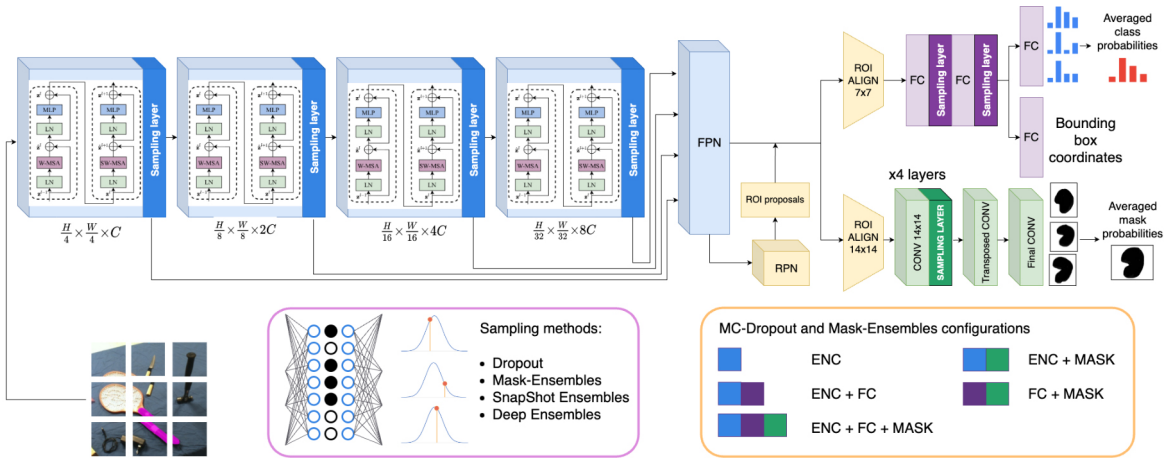


Figure 2.1: **Our model architecture is composed of an attention-based backbone extended with sampling layers.** We compare the performance and calibration of different sample-based (MC-Dropout, Mask-Ens) and ensembles (Deep-Ens, Snap-Ens) approaches for the Bayesian inference.

2.3.3 Bayesian Instance Segmentation

Prior works [54,55,58–60,71] consider affordances segmentation as a deterministic problem and do not evaluate uncertainty quantification. Our seminal work [111] introduced the Bayesian segmentation of affordances, pioneering the extraction of spatial uncertainty related to the predicted object boundary masks. In this chapter, we extend the analysis of semantic uncertainty associated with the affordance class for each detection. Our goal is to segment the boundaries of each object part, classifying the instance with the corresponding afforded action, and obtaining a well-calibrated uncertainty estimation. Additionally, we replace the commonly used convolutional encoders [12, 58, 111] with an attention-based backbone [140] adapted with internal sampling layers.

Training. Following prior works [60], we adopt Mask-RCNN as the baseline model. Figure 2.1 illustrates the general architecture of our approach. First, an encoder extracts hierarchical multi-scale feature maps from the image. We adopted the Swin Transformer [140], which processes images through a hierarchical self-attention mechanism [14], gradually combining small-size image patches that reduce the resolution while increasing the representation dimension. This structure mirrors the pyramid-like design of convolutional-based encoders, enabling efficient processing of high-resolution images and the extraction of multi-scale feature maps. The encoder progresses through multiple stages, each consisting of several Swin Transformer blocks that process the input image at different scales. Each block includes a Window-based Multi-head Self-Attention (W-MSA) module, a Multi-Layer Perceptron (MLP), a Shifted-Window Multi-head Self-Attention (SW-MSA) module, and a second MLP. W-MSA computes

self-attention within non overlapping windows, capturing local information, while SW-MSA shifts these windows to share patches with the previous layer, facilitating effective global information integration while maintaining linear complexity with the image size.

To generate a probabilistic representation in the feature space, we incorporate a sampling layer at the end of each encoder stage. The output from the four encoder stages consists of hierarchical feature maps across multiple scales, each capturing distinct semantic characteristics due to varying depths. Additionally, we integrate a Feature Pyramid Network (FPN) [141] with lateral connections and top-down propagation, which produces high-resolution feature maps at all levels, enriched with high-level semantic features.

Based on the FPN features, the Region Proposal Network (RPN) generates Region Proposals (RPs) that highlight areas likely to contain objects. The Region of Interest (RoI) Align block standardizes the size of the proposed RPs, ensuring that the extracted features correspond to the input regions. Subsequently, we obtain the semantic class and the bounding box coordinates of each RPs using respective Multi-Layer Perceptrons (MLPs), containing intermediate sampling layers. Then, a softmax converts the affordance class logit predictions into a probability vector $\mathbf{p}^c$. In parallel, we obtain the mask prediction after applying a 14×14 RoI Align block, four convolutional layers with intermediate sampling layers, and a transposed convolution to upsample the mask resolution. The binary probability masks $\mathbf{p}^h$ are obtained after a final convolution layer and a sigmoid.

We train the model end-to-end following [12] with a multi-task loss applied on each RoI as $\mathcal{L} = \mathcal{L}_{cls} + \mathcal{L}_{box} + \mathcal{L}_{mask} + \mathcal{L}_{RPNbox} + \mathcal{L}_{RPNcls}$. The classification loss $\mathcal{L}_{cls}$ computes the log loss to classify each anchor as an object or background accurately. The bounding box regression $\mathcal{L}_{box}$ refines the coordinates of the RP bounding boxes based on the ground truth offsets with a smooth L1 loss. The mask loss $\mathcal{L}_{mask}$ computes only for the positive instances the binary cross-entropy loss at the pixel level for the mask prediction. The RPN classification loss $\mathcal{L}_{RPNcls}$ computes a cross-entropy loss to train the RPN to distinguish between foreground and background anchors, while the RPN bounding box loss $\mathcal{L}_{RPNbox}$ applies a smooth L1 regression to refine the anchor box predictions generated by the RPN.

Inference. We extend the deterministic object affordances detection to a probabilistic sample-based approach in three steps: sampling, clustering, and merging.

1. **Sampling.** We conduct M forward passes to generate multiple detections per inference pass. Each detection $\mathbf{d} = \{\mathbf{b}, \mathbf{p}^c, \mathbf{p}^h\}$ consists of a bounding box $\mathbf{b} = (x, y, w, h)$, a class probability vector $\mathbf{p}^c$ and a *heatmap* representing the binary

probability of each pixel belonging to the detection or background $\mathbf{p}^h$. Depending on the uncertainty extraction method (Deep-Ens, Snap-Ens, MC-Dropout, Mask-Ens), the modification of the network weights $\mathbf{w}_m$ differs in each of the M forward passes.

2. **Clustering.** As defined by Miller et al. [142], an observation is a collection of detections with high spatial and semantic affinity $\mathcal{O}_d = \{\mathbf{d}_1, ..., \mathbf{d}_k\}$. Note that the number of times an instance is detected, k, is always less than or equal to the number of forward passes, M, i.e., $k \leq M$. Following [142], we cluster detections into observations using the Basic Sequential Algorithm Scheme (BSAS) with the Intersection over Union (IoU) of the detected masks $\mathbf{p}^h$ as the similarity metric. We only group detections with the same label class to avoid the merging of overlapped objects whose masks are spatially close. Compared to previous approaches limited to object detection with bounding boxes [142, 143], we compare pixel-wise masks of instances, enabling improved merging of clustered or irregularly shaped objects.

3. **Averaging.** Following the Bayesian approach in Equation (2.4), we average the clustered detections to obtain the final predictions as:

$$\mathcal{O} = \Big\{ \frac{1}{M} \sum_{m=1}^{M} \mathbf{b}_m, \frac{1}{M} \sum_{m=1}^{M} \mathbf{p}_m^c, \frac{1}{M} \sum_{m=1}^{M} \mathbf{p}_m^h \Big\}. \tag{2.5}$$

Consequently, an observation consists of a set of detections from different sampling steps, representing an approximation of the true posterior distribution of the observation.

Uncertainty estimation. Our novel Bayesian instance segmentation approach extends deterministic models by incorporating per-pixel uncertainty estimation, σ, which is the sum of the epistemic and aleatoric contributions,

$$\sigma = \sigma_e + \sigma_a. \tag{2.6}$$

The epistemic uncertainty σ_e represents the variability of the model parameters $\mathbf{w}$, which can be used to identify when a sample is out of the training distribution. Consequently, it can be reduced with a larger and more diverse dataset, as the model incorporates more knowledge. Following [83,130], we compute the epistemic covariance matrix as the average difference between the squared matrix formed by multiplying $\mathbf{p}_m - \bar{\mathbf{p}}$ by its transpose, as shown in Equation (2.7). It quantifies the deviation of each

prediction $\mathbf{p}_m$ from the mean probability $\bar{\mathbf{p}}$, which is computed as the average of the sampled probability vectors $\bar{\mathbf{p}} = \frac{1}{M}\sum_{m=1}^{M}\mathbf{p}_m$, and it is computed as follows:

$$\sigma_e = \frac{1}{M}\sum_{m=1}^{M}(\mathbf{p}_m - \bar{\mathbf{p}}) \cdot (\mathbf{p}_m - \bar{\mathbf{p}})^T . \tag{2.7}$$

The aleatoric uncertainty σ_a is inherent to data noise and cannot be reduced with more data, reflecting observations noise like the camera motion or object boundaries. Since it is intrinsic to the data distribution, collecting more data does not reduce it. The aleatoric covariance matrix is the average difference between a diagonal matrix formed with the components of the m-th prediction vector $\mathbf{p}_m$ and the outer product of the prediction vector $\mathbf{p}_m$ with itself, computed as:

$$\sigma_a = \frac{1}{M}\sum_{m=1}^{M}\mathbf{diag}(\mathbf{p}_m) - \mathbf{p}_m \cdot \mathbf{p}_m^T. \tag{2.8}$$

Previous works [111, 142–144] either ignore the differences between the two sources of uncertainty or solely evaluate the variance associated with mask segmentation. We introduce a novel formulation for extracting both spatial uncertainty σ^{sp} and semantic uncertainty σ^{sem}, as well as their respective epistemic and aleatoric variance contributions.

Spatial uncertainty is derived from the binary probability mask vector $\mathbf{p}^h$ and captures pixel-level differences in the segmentation. Analytically, we compute $\sigma^{sp} = \sigma_a^{sp} + \sigma_e^{sp}$, where $\mathbf{p}_m = \mathbf{p}_m^h$ represents the output after the sigmoid of the final mask convolution in the m-th forward pass and therefore, the spatial uncertainty σ^{sp} is extracted per-pixel and distributed in a 2D map along the image.

Semantic uncertainty reflects the ambiguities associated with the semantic affordance class and it is constant along all the detected object pixels. Similarly, we obtain $\sigma^{sem} = \sigma_a^{sem} + \sigma_e^{sem}$ where $\mathbf{p}_m = \mathbf{p}_m^c$ represents the probability of the affordance class, computed as the softmax output associated with the prediction of the semantic class of the m-th sampling. Therefore, we assume that the semantic uncertainty σ^{sem} is constant for all the pixels within the predicted affordance mask. We collapse the respective covariance by summing the trace of the matrix.

2.4 Experiments

Following previous works [58–60], we adopt the F_β^w score [145] to evaluate the affordance segmentation performance. This score assesses foreground maps and adjusts pixel-level errors based on three key assumptions: dependency, interpolation, and equal importance. Next, we assess the uncertainty of probabilistic affordance detections

using our novel Probabilistic Mask Quality (PMQ) metric. Additionally, we evaluate the calibration of the predictions through Expected Calibration Error (ECE) and Area Under the Sparsification Curve (AUSE). We provide a detailed calibration analysis for both spatial uncertainty (associated with mask contours) and semantic uncertainty (associated with the affordance class).

2.4.1 Probabilistic Mask Quality metric

We design our novel PMQ inspired by the Probabilistic-based Detection Quality (PDQ) score [146]. While the PDQ analyzes the spatial and semantic uncertainties of probabilistic object detections based on bounding box predictions, our novel PMQ score compares the averaged predicted masks $\bar{\mathbf{p}}^h$, providing a more fine-grained evaluation and preserving boundary detections. Compared to metrics based on the Average Precision score, PMQ does not rely on arbitrary IoU thresholds that filter the number of detections, nor does it evaluate foreground/background quality separately.

We define each i ground-truth object annotation as $\mathcal{G}_i = \{\hat{\mathbf{b}}_i, \hat{\mathbf{c}}_i, \hat{\mathcal{M}}_i\}$, where $\hat{\mathbf{b}}_i$ is the ground-truth bounding-box coordinates, $\hat{c}_i$ the discrete label class and $\hat{\mathcal{M}}_i$ the mask around the object. The output of our probabilistic model is a set of j averaged detections, where $\mathcal{O}_j = \{\bar{\mathbf{b}}_j, \bar{\mathbf{p}}^c_j, \bar{\mathbf{p}}^h_j\}$. We start evaluating the semantic estimation with the label quality Q_l. It is the probability score of the respective ground-truth class and not the highest predicted class score, formally defined as:

$$Q_L(\mathcal{G}_i, \mathcal{O}_j) = [\mathbb{1}_{c=\hat{c}_i}]^T \cdot \bar{\mathbf{p}}^c_j, \tag{2.9}$$

where $\mathbb{1}_{c=\hat{c}}$ is the indicator function, being 1 for $c = \hat{c}$ and 0 otherwise. The spatial quality Q_s measures how the probabilistic mask captures the contour of the ground-truth affordance mask. It is the exponential of the negative sum between the foreground loss L_{FG} and the background loss L_{BG} at the mask level

$$Q_S(\mathcal{G}_i, \mathcal{O}_j) = \exp(-(L_{FG}(\mathcal{G}_i, \mathcal{O}_j) + L_{BG}(\mathcal{G}_i, \mathcal{O}_j))). \tag{2.10}$$

While the computation of the foreground loss L_{FG} and the background loss L_{BG} is similar, they evaluate opposite effects. The L_{FG} penalizes predicted pixels inside the ground truth mask with low probability, while the L_{BG} penalizes pixels outside the ground truth mask with high probability. The L_{FG} is the average negative logarithm of the mask probabilities $\bar{\mathbf{p}}^h_j$ inside the ground-truth segments. The L_{BG} is defined as the sum of the negative log-pixel probabilities evaluated in the set $\mathcal{V}_{ij} = \{\bar{\mathbf{p}}^h_j - \hat{\mathcal{M}}_i\}$, corresponding to detected pixels $\bar{\mathbf{p}}^h_j$ but outside the ground truth masks $\hat{\mathcal{M}}_i$. We define formally these losses as:

$$L_{FG}(\mathcal{G}_i, \mathcal{O}_j) = -\frac{1}{|\hat{\mathcal{M}}_i|} \sum_{x \in \hat{\mathcal{M}}_i} \log(\bar{\mathbf{p}}_j^h(\mathrm{x}))), \tag{2.11}$$

$$L_{BG}(\mathcal{G}_i, \mathcal{O}_j) = -\frac{1}{|\hat{\mathcal{M}}_i|} \sum_{x \in \mathcal{V}_{ij}} \log(1 - \bar{\mathbf{p}}_j^h(\mathrm{x})). \tag{2.12}$$

Then, we compute the pairwise PMQ (pPMQ) for each observation $\mathcal{O}_j$ and object label $\mathcal{G}_i$ as the geometric mean of the respective semantic Q_L and spatial Q_S qualities computed as follows:

$$pPMQ(\mathcal{G}_i, \mathcal{O}_j) = \sqrt{Q_S(\mathcal{G}_i, \mathcal{O}_j) \cdot Q_L(\mathcal{G}_i, \mathcal{O}_j)}. \tag{2.13}$$

At this point, the pairwise PMQ contains a PMQ value between each label G_i and observation $\mathcal{O}_j$ pair. We apply the Hungarian algorithm [147] to obtain the optimal assignment of the N_{TP} true positives, denoted as $\mathbf{q} = [q_1, ..., q_{TP}]$, between the observations $\mathcal{O}_j$ and the label $\mathcal{G}_i$. An optimal assignment is zero when a ground-truth object is not detected (false negative) or a observation does not represent any affordable object-part (false positive). The final PMQ, compute along all the frames N_F in the evaluation set is computed by averaging the $pPMQ$ along the total number of false positives N_{FP} (the observation does not match with the ground-truth), the total false negatives N_{FN} (no exits ground-truth) and all the true positives N_{TP}, as follows:

$$PMQ(\mathcal{G}, \mathbf{d}) = \frac{1}{\sum_{f=1}^{N_F} \left(N_{TP}^f + N_{FN}^f + N_{FP}^f \right)} \sum_{f=1}^{N_F} \sum_{i=1}^{N_{TP}} \mathbf{q}^f(i). \tag{2.14}$$

2.4.2 Calibration metrics

We evaluate model calibration with the Expected Calibration Error (ECE) and the Area Under the Sparsification Error (AUSE). The ECE [82,148] measures the absolute calibration error of a probabilistic classification. A well-calibrated classification model produces predictions where the predicted confidence accurately reflects the likelihood of being correct. To compute ECE, predictions are first grouped into N bins based on their confidence scores. The calibration error is then measured as the difference between the accuracy and confidence of each bin. The bin accuracy, $\mathrm{acc}(B_n)$, is the fraction of correctly classified instances within the bin. The bin confidence, $\mathrm{conf}(B_n)$, is the average predicted probability of the instances in the bin:

$$\mathrm{ECE} = \sum_{n=1}^{N} \frac{|B_n|}{N_s} |\mathrm{acc}(B_n) - \mathrm{conf}(B_n)|, \tag{2.15}$$

where N_s is the total number of samples, and B_n is the number of samples in the respective bin.

The AUSE metric provides a relative measure of uncertainty, computed as the difference between the estimated uncertainty and the true error of the model, measured in terms of the Brier Score (BS) [149]. The BS is the mean squared difference between the predicted probabilities p and the corresponding one-hot ground-truth vector $\mathbb{1}_{c=\hat{c}}$, being 1 for $c = \hat{c}$ and 0 otherwise:

$$\mathrm{BS} = \frac{1}{N_p} \sum_{n=1}^{N_p} (\mathrm{p}_n - \mathbb{1}_{c=\hat{c}})^2. \tag{2.16}$$

In Equation (2.16), for the semantic AUSE, N_p is the total number of detections in the entire test dataset, while for the spatial AUSE, N_p is the total number of pixels with an estimated variance (thus, background regions between bounding boxes are ignored). In order to compute the AUSE metric, we gradually remove pixels based on their true error (Oracle) or their respective estimated variance (Model). Lower values of the AUSE metric indicate that the estimated variance represents better the true error.

2.4.3 Implementation details

We train the models with ResNet-50 [11] and ResNeXt-101 [150] using the Adam optimizer [151] with learning rates of 10^{-3} and 10^{-5}, respectively. The models with Swin-T [140] as the encoder were trained with AdamW [152], a learning rate of 10^{-5}, a weight decay of 0.05, and $\beta = (0.9, 0.99)$. We initialize the weights with the respective pre-trained versions on COCO [153] to ensure convergence and improve performance, starting with a linear warm-up of the learning rate during the first 1,000 iterations. We run our experiments in an NVIDIA GeForce 4090 GPU. We conducted the experiments on the IIT-AFF dataset [55], which consists of 8,835 real-world images depicting seven different affordance categories (*contain, cut, display, engine, grasp, hit, pound, support,* and*w-grasp*). The dataset is specifically designed for robotics-based scenarios and represents common manipulation capabilities for autonomous agents. We also report results in the UMD dataset [54], which contains 28,843 images of 17 object categories and 7 affordance actions, captured on a rotating table in clutter-free conditions.

Method	Backbone				F_β^w
	VGG	R50	R101	Swin-T	
ED-RGB [71]	✓	-	-	-	57.64
R-FCN [55]	-	-	✓	-	69.62
AffordanceNet [58]	✓	-	-	-	79.90
RelaNet [154]	-	✓	-	-	78.92
BPN [155]	-	✓	-	-	79.64
GSE [156]	-	-	✓	-	82.33
Mask-RCNN [60]	-	✓	-	-	84.40
Ours, Bayesian	-	✓	-	-	86.90
Ours, Deterministic	-	-	-	✓	88.32
Ours, Bayesian	-	-	-	✓	**90.60**

Table 2.1: **Affordance segmentation comparative with the state-of-the-art in the IIT-Aff dataset.**

Method	Backbone				F_β^w
	VGG	R50	R101	Swin-T	
AffordanceNet [58]	✓	-	-	-	79.90
EfficientNet [157]	✓	-	-	-	82.30
RelaNet [154]	-	✓	-	-	86.13
BPN [155]	-	✓	-	-	86.21
GSE [156]	-	-	✓	-	85.50
Mask-RCNN [60]	-	✓	-	-	84.21
Ours, Deterministic	-	-	-	✓	86.03
Ours, Bayesian	-	-	-	✓	**87.50**

Table 2.2: **Affordance segmentation comparative with the state-of-the-art in the UMD dataset.**

2.5 Results

2.5.1 Comparative with the state-of-the-art

We first compare our model with the state-of-the-art for the IIT-Aff and UMD datasets in Tables 2.1 and 2.2, respectively. Our approach demonstrates significant improvements on IIT-Aff, where our best model version (Mask-Ens) achieves 90.65 F_β^w, outperforming the previous Mask-RCNN [60] (84.40 F_β^w) by +6.1 percentage points (p.p). This notable gain highlights the superior generalization capabilities of our model in robotics and real-world scenarios. On the UMD dataset, the Bayesian model achieves a competitive performance of 87.50 F_β^w as shows Table 2.2. As UMD comprises isolated objects, model performance tends to saturate, resulting in smaller performance gains. Nevertheless, the Bayesian approach consistently outperforms the deterministic vari-

					Ours						
	[71]	[55]	[58]	[60]	Rx101 Determ.	Rx101 MC-Drop	Swin-T Determ.	Swin-T SnapShot Ens	Swin-T Deep Ens	Swin-T Mask Ens	Swin-T MC-Drop
contain	66.4	75.6	79.6	83.6	85.8	85.6	87.2	89.4	89.2	**89.5**	88.8
cut	60.7	69.9	75.7	84.7	86.4	84.6	86.7	88.8	**89.0**	87.7	88.9
display	55.4	72.0	77.8	86.3	89.9	90.6	92.8	**93.9**	93.8	93.3	93.4
engine	56.3	72.8	77.5	88.9	90.5	90.9	91.9	92.5	**92.6**	92.1	92.4
grasp	59.0	63.7	68.5	71.1	77.5	78.2	85.6	84.5	85.2	84.6	**88.9**
hit	60.8	66.6	70.8	92.3	94.3	95.4	94.5	**95.5**	95.4	95.4	95.4
pound	54.3	64.1	69.6	81.8	80.5	80.7	85.1	87.2	**87.9**	86.9	87.2
support	55.4	65.0	69.8	86.7	85.4	88.7	88.6	92.3	**92.4**	**92.4**	91.2
w-grasp	50.7	67.3	71.0	83.7	84.7	86.6	87.0	90.2	89.9	**90.8**	89.5
Average	57.6	68.6	73.4	84.4	86.1	86.9	88.3	90.5	**90.6**	90.3	90.2
Nº train parameters (M)	-	-	-	43.7	107.1	107.1	47.4	1258.2	1138.5	52.4	47.4
Inf. Time (ms)	-	-	-	45	47	1207	42	1015	1015	1015	1015

Table 2.3: **Per-class $F_\beta^w(\uparrow)$ affordance segmentation scores on the IIT-Aff test split dataset.**

ant (86.03 F_β^w) in the UMD dataset, highlighting the benefits of aggregating multiple detections.

Following, Table 2.3 presents a detailed per-class comparative on the IIT-Aff dataset against previous works [55, 58, 60, 71] across different versions of our approach. Our Swin-T Deep-Ens achieves a 90.6 % F_β^w score, with particularly strong results on the *cut* 89.0 % F_β^w, *engine* 92.6 % F_β^w, *pound* 87.9 % F_β^w and *support* 92.4 % F_β^w classes. Comparing the model size and the inference speed, the key advantage of MC-Drop and Mask-Ens is that they require training only a single model, resulting in a similar number of parameters as [60], while offering improved segmentation performance and uncertainty quantification. However, the computational cost of Bayesian methods is reflected in the inference time, which increases linearly with the number of samples.

2.5.2 Ablation study

Bayesian vs. Deterministic methods: Tables 2.3 and 2.4 highlight the differences between Bayesian and deterministic methods. Bayesian models obtain similar performance on the affordance segmentation (90.5 Snap-Ens, 90.6 Deep-Ens, 90.3 Mask-Ens, 90.2 MC-Drop % F_β^w score), but they show a strong improvement compared with the deterministic Swin-T (88.3 % F_β^w score) and lower calibration errors. The global consensus of multiple networks makes Bayesian models outperform the respective deterministic model due to a better generalization and mask refinement. Probabilistic models produce better calibrated probabilities, both at the semantic vectors or the binary masks, which means that the model is less overconfident and that the confidence of the probability is closer to the accuracy of the predictions.

		Calibration metrics				Performance metrics				
Backbone	Method	Semantic AUSE (↓)	Semantic ECE (↓)	Spatial AUSE (↓)	Spatial ECE (↓)	F_β^w score (↑)	PMQ (↑)	pPMQ (↑)	Q_L (↑)	Q_S (↑)
R50	Deterministic	0.308	0.0214	0.252	0.00483	84.4	27.7	55.5	89.9	53.0
	MC-Dropout	0.280	0.0112	0.132	0.00246	85.9	41.9	71.9	84.7	63.2
	Mask-Ens	0.225	0.0125	0.186	0.00263	84.9	36.3	68.7	78.9	61.9
	Deep-Ens	0.241	0.0134	0.158	0.00257	87.1	50.5	71.1	85.2	62.5
	Snap-Ens	0.176	0.0128	0.177	0.00253	86.3	40.7	70.5	82.0	64.1
Rx101	Deterministic	0.223	0.0187	0.146	0.00382	86.1	28.4	62.4	85.1	62.4
	MC-Dropout	0.207	0.0106	0.134	0.00244	86.9	44.1	75.6	89.8	65.8
	Mask-Ens	0.225	0.0112	0.153	0.00274	87.4	48.3	73.9	87.2	66.2
	Deep-Ens	0.208	0.0111	0.108	0.00202	88.6	51.3	72.9	87.4	63.7
	Snap-Ens	0.168	0.0139	0.106	0.00222	87.4	58.7	76.4	93.2	63.4
Swin-T	Deterministic	0.213	0.0146	0.155	0.00209	89.0	29.9	63.6	94.2	61.4
	MC-Dropout	0.196	**0.0084**	0.129	0.00215	89.6	37.8	73.2	82.2	66.5
	Mask-Ens	0.166	0.0089	0.102	0.00174	89.7	46.1	75.6	85.9	67.9
	Deep-Ens	**0.136**	0.0105	0.082	0.00167	**90.6**	57.5	77.3	**94.6**	68.1
	Snap-Ens	0.145	0.0112	**0.080**	**0.00159**	90.5	**60.3**	**77.5**	92.0	**68.2**

Table 2.4: **Affordance segmentation performance and calibration metrics** We compare different encoders (ResNet-50, ResNeXt-101, Swin-T) and uncertainty estimation methods (MC-Dropout, Mask-Ens, Deep-Ens, Snap-Ens).

Backbone configurations: Results in Table 2.4 show that the backbone affects significantly to the performance and the uncertainty estimation: attention-based encoders obtain best affordance segmentation and show better calibration metrics. For example, comparing the analogous Mask-Ens Resnet-50 and Swin-T configurations the differences are 84.9 vs. 89.7 % F_β^w, 0.225 vs. 0.166 Sem AUSE, 0.0125 vs. 0.0089 Sem ECE, 0.186 vs. 0.102 Sp AUSE and 0.00263 vs. 0.00174 Sp. ECE. We also report in Table 2.4 our novel PMQ metric, which evaluates the uncertainty quality of the predicted masks. The differences with the previous conference version [111] are due to the computation of the spatial probability as the mean of the per-pixel masks probabilities and not as the number of times that a pixel is detected ($\mathbf{p}^{mask} > 0.5$). The improvement of Swin-T against Resnet-50 is also reflected on this metric (46.1 vs. 36.3 PMQ, 75.6 vs. 68.7, 85.9 vs. 78.9 Q_L, 67.9 vs. 61.9 Q_S for the Mask-Ens Resnet-50 and Swin-T, respectively).

Comparative of Bayesian methods: Table 2.4 also compares different Bayesian methods for uncertainty estimation. Ensemble methods [105, 106] achieve superior segmentation results and exhibit lower calibration errors compared to MC-Drop or Mask-Ens. The Swin-T Deep-Ens and Swin-T Snap-Ens configurations achieve the highest performance with lower calibration errors, which we attribute to their capacity to capture diverse local optima and effectively represent the multi-modal nature of the underlying distribution. However, this improvement comes at the cost of a linear increase in training parameters, as Table 2.4 shows. Notably, while Deep-Ens requires

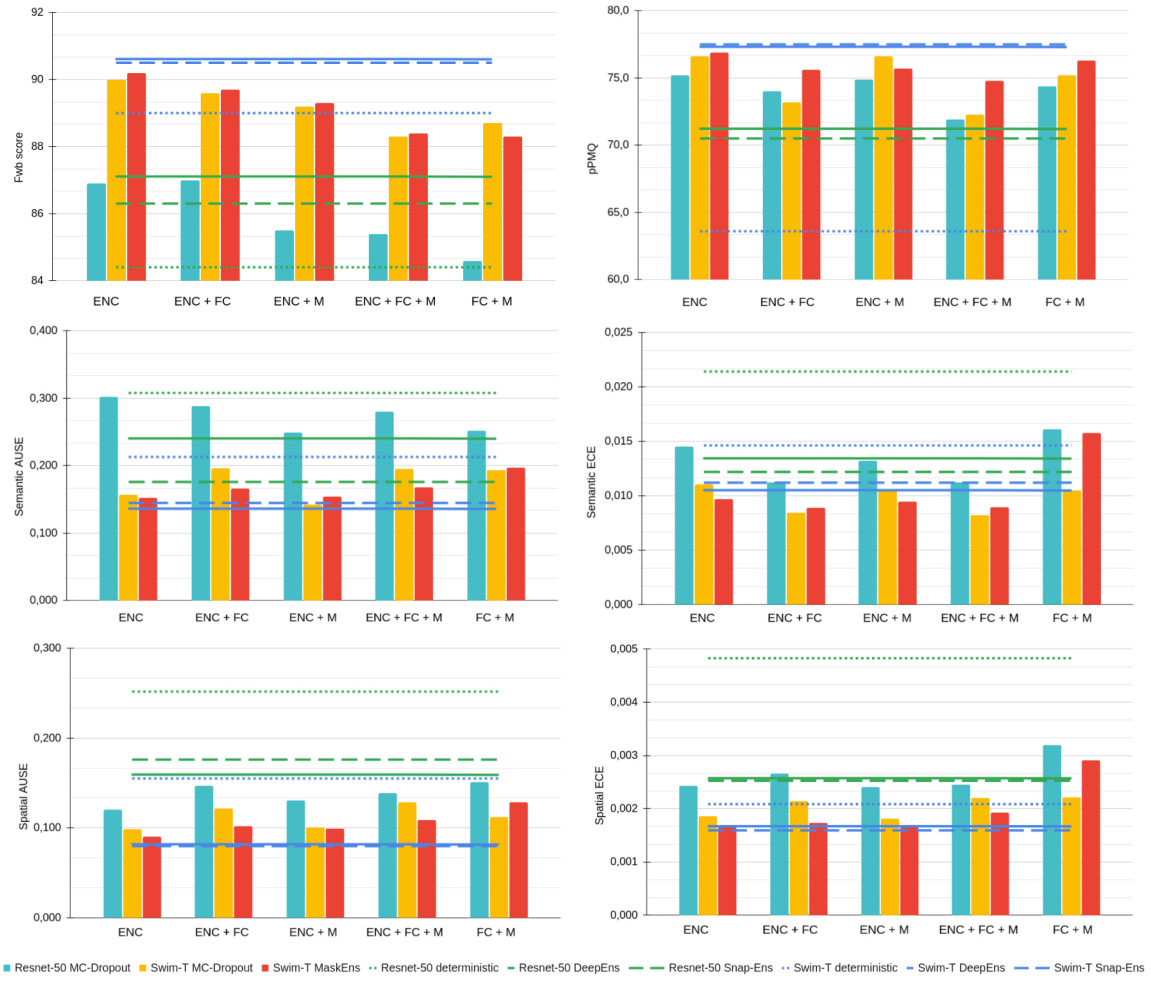


Figure 2.2: **Ablation study showing the performance and calibration.** We compare different configurations of the sampling layers on R50 MC-Dropout (blue bars), Swin-T MC-Dropout (yellow bars) and Swin-T Mask-Ens (red bars). We also report the behavior of their corresponding deterministic version with a straight line.

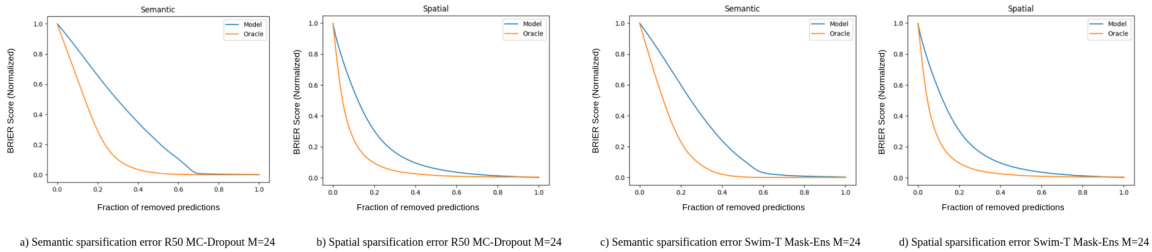


Figure 2.3: **Sparsification error curves for the semantic and spatial probabilities for Swin-T.**

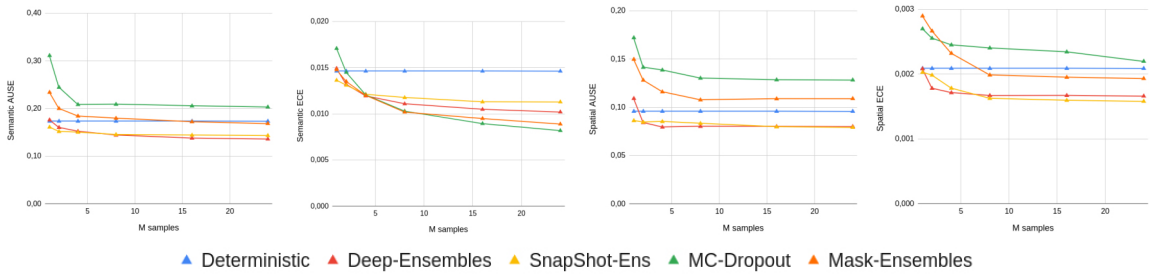


Figure 2.4: **Evolution of the calibration metrics with the number of samples M.** We compare different uncertainty extraction versions on the Swin-T encoder.

training M independent models, Snap-Ens mitigates computational overhead by extending the training of a single model through cyclical learning rate adjustments. In comparison, Mask-Ens outperforms MC-Drop in calibration, benefiting from reduced correlation among the sampling masks.

Disposition of sampling layers: We conduct an ablation study on the placement of MC-Dropout sampling layers, as illustrated in Figure 2.2. For reference, we also report the performance of ensembles and deterministic models, which do not require architectural modifications. When sampling layers are positioned in the encoder (ENC), the model learns a probabilistic latent space that propagates through the architecture, leading to better-calibrated probability estimates. In contrast, when sampling layers are applied only in the final layers (FC + M), enforcing probabilistic outputs at this stage results in less consistent and less well-calibrated predictions. Additionally, we observe that ENC + M provides better calibration for binary masks, whereas ENC + FC improves the calibration of semantic probabilities due to the proximity of the dropout layers. Comparing Swin-T MC-Dropout with Mask-Ens, the results indicate that Mask-Ens slightly outperforms MC-Dropout in both performance and calibration. This improvement is attributed to the lower correlation among its dropout layers, allowing Mask-Ens to generalize more effectively.

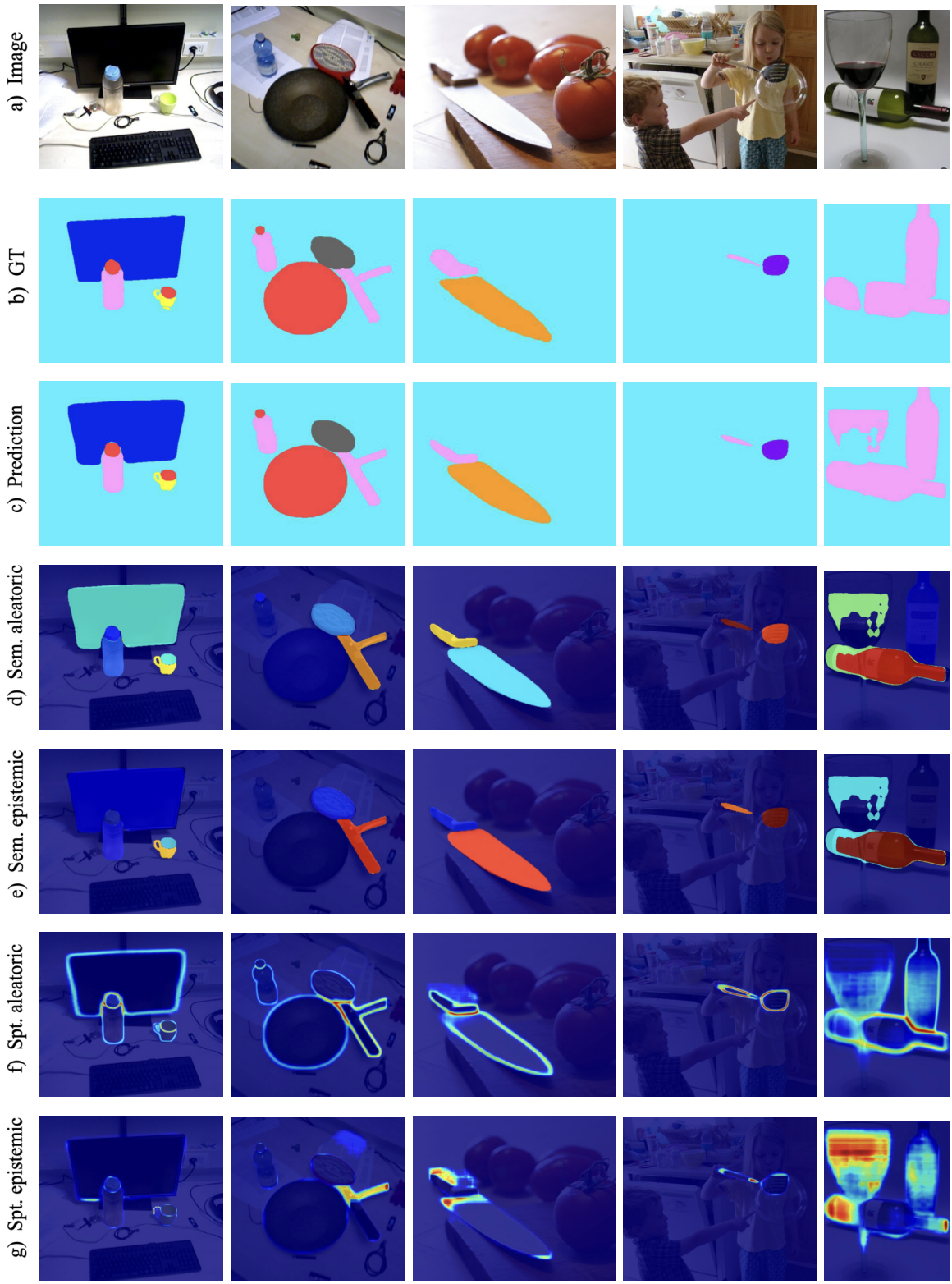


Figure 2.5: **Qualitative results and uncertainty prediction obtained by the Swin-T Mask-Ens Bayesian Instance segmentation model.**

Number of samples: We also present the evolution of calibration metrics as a function of the number of forward passes in Figure 2.4. The curves indicate rapid convergence within the first $M = 8$ samples, followed by a plateau phase. Additionally, the sparsification error curves in Figure 2.3 demonstrate that spatial uncertainty more accurately approximates the true error than semantic uncertainty, as evidenced by the smaller area difference between the corresponding curves in the spatial case. The reduction in calibration errors further confirms that the estimated probabilities more closely align with the true probabilities.

2.5.3 Qualitative results

We show an extensive example of the qualitative results in Figure 6. The predicted affordance maps are very close to the ground-truth labels with smooth borders as a consequence of the multiple mask averaging. While the semantic uncertainty is constant for all the predicted pixels of the object class, the spatial uncertainty is pixel-wise distributed. The spatial aleatoric uncertainty, due to the data noise, is mostly present in the object's contours. We encounter the highest values of this uncertainty at the intersection of different masks. The spatial epistemic variance reflects the model's ignorance about the detection and adds comprehensiveness when the model fails the segmentation. By [83], epistemic uncertainty appears in challenging pixels out-of-the-distribution. For example, in the third example, the model fails when predicting a part of the knife handle, due to the high reflections that make these pixels challenging to segment.

2.6 Conclusions

In this chapter, we introduced the instance segmentation of affordances with uncertainty estimation. We extend an attention-based backbone with different techniques for uncertainty quantification. We conducted a comparative analysis of ensemble methods (Deep-Ens [105] and Snap-Ens [106]) against sampling-based techniques (MC-Dropout [103] and Mask-Ens [104]). A detailed ablation study further examines the effects of MC-Dropout sampling layer configurations and the influence of the quantity of training parameters on model performance. We obtain a new state-of-the art at 90.6 $\%F_\beta^w$ score on the IIT-Aff dataset, which represents a significant improvement over [60]. We further extend the uncertainty quantification with the estimation of fine-grained spatial and semantic variance maps, both with the epistemic and aleatoric contributions and evaluated quantitatively with our novel PMQ metric.

Part II

Learning visual models of environments

Chapter 3

Multi-label affordance mapping from egocentric vision

Previously, I proposed learning affordances at the object level by training an instance segmentation model using manual annotations, which were ungrounded in real interaction and restricted to a single category per object part. However, while learning visual object models provides valuable information about their functionalities, it is insufficient for understanding the real world, where individual objects are not isolated; rather, they are embedded within a broader physical space—the environment—which defines the context in which interactions occur. In this chapter, I begin by constructing a multi-label affordance map that spatially links activity-centric zones. I first leverage this environment representation to develop a pipeline that automatically generates multi-label, pixel-wise, and grounded affordance annotations; representing a significant improvement over the ungrounded, single-label model introduced in the previous chapter. Next, I extend classical segmentation models with a multi-label prediction head to capture more diverse interaction regions, based on the assumption that a single object can afford multiple actions. Using camera pose estimation, I project the predicted affordance regions onto a 3D point cloud to generate a spatial map. Finally, I show that the multi-label mapping can guide autonomous agents in task-oriented navigation.

3.1 Introduction

When humans repeatedly interact in a close environment, we associate a set of affordable actions with a certain distribution of objects. For example, we associate a pan on a stove with cooking, but the same pan on the sink with washing. A joined spatial-semantic understanding contains powerful insights to understand human be-

havior. This requires a close combination of perception, mapping and navigation algorithms; with potential applications in augmented reality systems [158, 159], but also guiding a robot [47, 58] or assistive devices [9].

In the last years, the ability of deep learning models to extract high-level representations has improved the perception of autonomous agents, while egocentric vision offers a powerful viewpoint for modeling human-object interaction understanding. Recent advances include anticipating future actions [28, 29, 98], model the hands-object manipulation [32, 62, 100, 101], detect the change in an object state [42], identify interaction hotspots [53, 72] or create topological maps [39]. Despite the fast movements of a headset camera, egocentric perception has also contributed to the mapping and planning phases: localizing the agent in a known 3D map [66], performing visual navigation [160, 161] or building allocentric maps [76, 162].

Gibson's perception theory presents affordances as the potential actions that the environment offers to the agent based on its motor capabilities [45]. For example, the person can afford *taking* a glass, but the affordances of a soup in a pan can be *mixing, emptying, scooping* and *pouring* simultaneously. This multiplicity models better complex dynamic environments and opens the door to multi-agent collaboration with task synchronization. Although some authors have focused on more complex affordance models [51, 163], affordance perception is typically defined as a classification problem. Some authors have focused instead on learning affordances grounded on natural human-object interactions [72], which provide a more flexible setup and are truly associated with motor capabilities, showing improvements in action anticipation [57]. However, most learning approaches in affordance perception consider the problem ungrounded to the agent interaction with the object, requiring previous annotations of each affordance occurrence [54, 55, 58–60, 71]. While ungrounded methods have the advantage of providing pixel-wise precision, which we denominate metric understanding of the scene, many grounded approaches rely on full image classification losing any metric meaning.

In this chapter, we propose a grounded multi-label approach with pixel-wise precision, which enables a detailed perception while maintaining the flexibility of grounded methods. Close to our approach is the work of Nagarajan et al. [53], which presented a grounded approach for extracting interaction hotspots by directly observing videos. Similar to other previous works, the hotspots are modeled as a single available affordance. Instead, we propose to consider the multiplicity of affordances for a single object or spatial zone through multi-label pixel-wise predictions. Therefore, we present the following contributions:

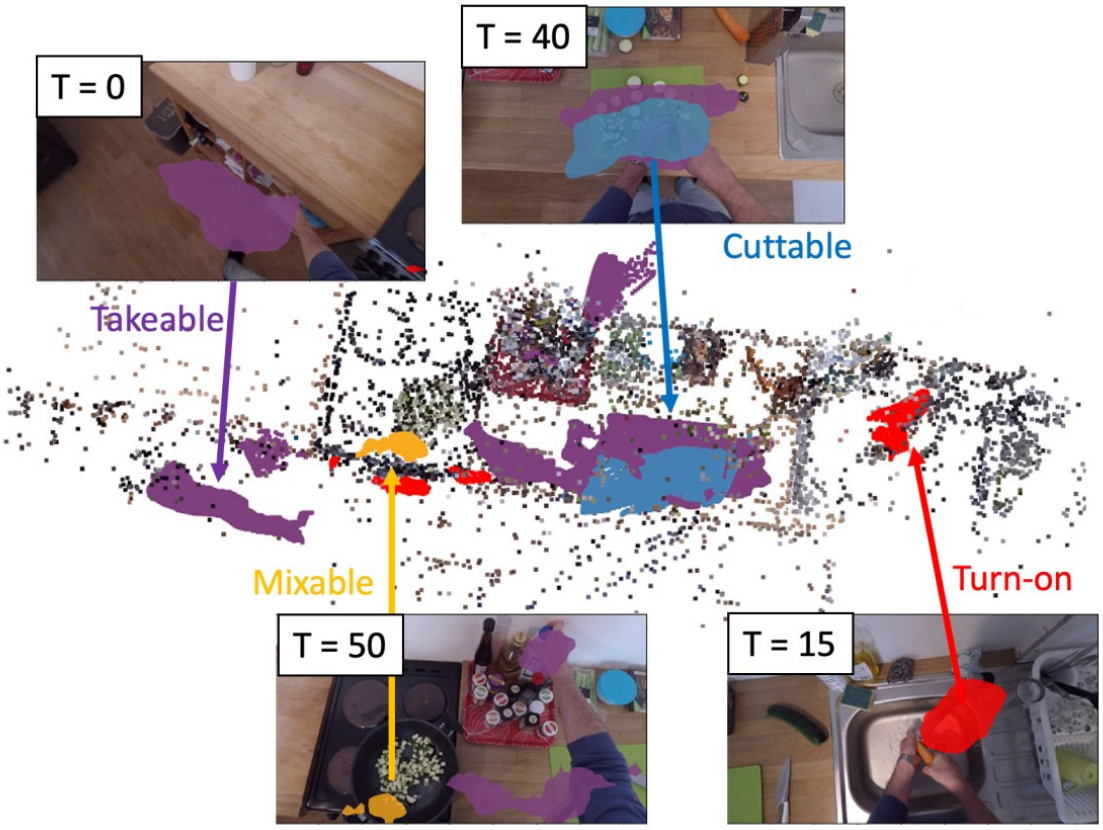


Figure 3.1: **Multi-label affordance mapping.** From a sequence of egocentric observations, the proposed approach creates a spatial-metric multi-label representation of the affordances, enabling a task-oriented navigation.

– We propose a pipeline to automatically collect multi-label, pixel-wise annotations from real-world interactions by leveraging a spatial and semantic representation of the environment. Applied to the EPIC-Kitchens dataset [25], this method enables the construction of EPIC-Aff, the largest dataset for grounded affordance segmentation to date.

– We extend several segmentation architectures to the multi-label setting, enabling richer scene understanding under the assumption that individual objects may support multiple affordances simultaneously.

– We conduct an extensive quantitative evaluation of the adapted architectures, comparing different heuristics to select multiple affordance labels from the predicted probability vectors.

– We introduce a mapping approach that translates the multi-label affordance segmentation into a spatial map linking activity-centric zones, as illustrated in Figure 3.1. This map provides a metric representation of environmental affordances, facilitating goal-directed navigation for embodied agents.

This work was published in ICCV 2023:

– Lorenzo Mur-Labadia, Jose J Guerrero, and Ruben Martinez-Cantin. Multi-label affordance mapping from egocentric vision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* Core Ranking A*, pages 5238–5249, 2023.

3.2 Related works

3.2.1 Learning Visual Affordances

Ungrounded approaches [54, 55, 58–60, 71] for affordance perception are fully supervised by manually labeled masks. Due to their similarity with semantic segmentation or object detection tasks, these works use common architectures such as an encoder-decoder [71], proposal-based detectors [55, 58–60] or Bayesian instance segmentation [111]. While these models [54, 55, 57, 58, 111] assume a single-label affordance per object and lose a valuable amount of information, our approach predicts multiple affordance categories. Concerning grounded works, Fang et al. [72] extracted a latent representation from demonstration videos or Luo et al. [57] transferred the learning from exocentric images to the egocentric perspective using only the semantic label as supervision. Nagarajan et al. [53] identified interaction hotspots by computing gradient-weighted attention maps during the training of an action classifier on videos. Then Ego-Topo [39] built a topological graph of the scene to perform affordance classification from egocentric videos, grouping each node visually and temporally coherent frames with similar object and action distributions, discovering activity-centric zones based on their visual content.

3.2.2 Multi-label perception

In the multi-label segmentation task, we assign two or more categories to a single pixel. A particular case is an amodal segmentation where the relevance of occluded parts depends on the depth order [164, 165]. The most common applications of multi-label segmentation are biomedical works, where there are multiple overlapped non-exclusive levels of tissues. Existing architectures extend a U-Net with minor changes such as a dynamic segmentation head [166], shuffle-attention mechanisms in the skip-connections [167], a combination of appearance and pose features [168] and split-attention modules [169]. The closest approach to multi-label segmentation is multi-label image recognition, where the label imbalance between positives and negatives in each binary classifier and the extraction of features from multiple objects make this

task more complex [170]. Class distribution aware losses [171] such as the asymmetric loss [172], the focal loss [173] or the Multi-Label Softmax loss [174] correct the over-suppression of negative samples. On the other hand, Graph Neural Networks such as [175] deal with the feature extraction from multiple objects by creating a dynamic graph for each image that leverages the content-aware category representations. Finally, the transformer architecture extracts multiple attention maps in the different regions of interest [176, 177], guiding the multi-label classification [170] or ranking the class of the pixels considering only the categories selected by the classifier [178].

3.3 Grounded Affordance Labeling

We extract automatic, interaction-grounded, multi-label pixel-wise and spatial affordance annotations from a sequence of real-world images in complex and cluttered environments, as shows Figure 3.2. Our multi-label segmentation approach learns *all* the potential options and does not reduce the perception to a single action. For example, a potato on a chopping board offers *cutting, putting, peeling, removing* and *taking* simultaneously. Current affordance segmentation works [54, 55, 57, 58, 111] assume a single-label affordance per object and lose a valuable amount of information. Although other affordance models allow for multiple predictions, these works ignore the segmentation of the interaction hotspot in the image and lose the pixel-level accuracy of the segmentation models. For example, topological maps extract multiple affordances from an image [39], or action anticipation models predict a probability distribution of the different possibilities [28, 29]. Our methodology gets the best of two worlds producing multi-label metric masks, resulting in a full distribution of affordances. It enables a deeper understanding of the manipulation task such as the grasping points of the tool [179] or the evolution of the manipulation process over time [67]. Similar to previous unsupervised or weakly supervised methods [51], we extract affordance labels from weak VISOR and EPIC Kitchens annotations grounded on actual interactions.

We join the affordances with their 3D spatial location by extracting the camera poses. The spatial approach to affordance perception is not new for the community. Rhinehart et al. [65] associate the functionality of regions with specific spatial locations, showing that that defining an affordance based solely on semantics is insufficient due to the significant influence of the physical context. For example, a frying pan is only *cookable* when it is on the hob or a plate is *washable* when the agent is next to the dishwasher. However, their method results in smooth 2D maps, which can be problematic for the fine-grained affordances in 3D space in our EPIC-Aff dataset. Instead, our method is able to scale up to large environments while maintaining the detail by using

neural networks. Other previous works [180, 181] also use SLAM for action prediction but with addressing different problems. In those works, the action is set on the human, while the image provides context; while in our case the action/affordance is set on the environment and the user provides context. In our work, we use COLMAP [182] to extract the relative pose between sparse frames with a filter of the dynamic objects, registering up to 93 % of the frames compared with the 44 % of the frames registered by ORB-SLAM [183] on EPIC Kitchens [39]. Recently, EPIC Fields [184] registered the camera pose of the dense videos in EPIC Kitchens using neural rendering techniques.

3.3.1 Affordance datasets

Following our motivation, we conduct a study along the visual affordance datasets shown in Table 3.1. The ungrounded datasets are subjected to the annotator's consideration and required to draw pixel-wise semantic mask to each object part [54–57, 113] or additional sensors [185, 186], decreasing the object variability and limiting the scalability due to the annotation costs. The UMD dataset contains a semantic affordance map for objects in isolated conditions and with low variability, which prevents generalization [54]. The IIT-Aff dataset [55] provides the most comprehensive annotations designed for use in robotics, including multiple objects in a single image. The ADE-Aff dataset [56], built on top of ADE20K scenes, examines the social acceptability of actions about context but is limited to only three affordance classes. The AGD20k [57] dataset includes the largest number of categories and actions by transferring from an exocentric to an egocentric viewpoint perspective. On the other hand, grounded works learn from observing interactions on the EPIC-Kitchens sequences [25], internet demonstration videos [72] or with gaze point with eye-tracking devices [187]. The annotations provided are only used for evaluation since they do not require strong supervision. However, these approaches ignore the pixel-wise precision [39] or the multi-label modality of our approach [53].

Based on the mentioned limitations, our novel dataset EPIC-Aff provides multi-label pixel-wise affordance annotations with the camera pose. It contributes to a diverse and comprehensive affordance database with the largest number of images up to date. This better captures the complexity, dynamics, multiplicity and variability of real-world environments, such as preparing a recipe in a kitchen. Finally, as our labels are automatically extracted, we enable the application of our method to other egocentric datasets.

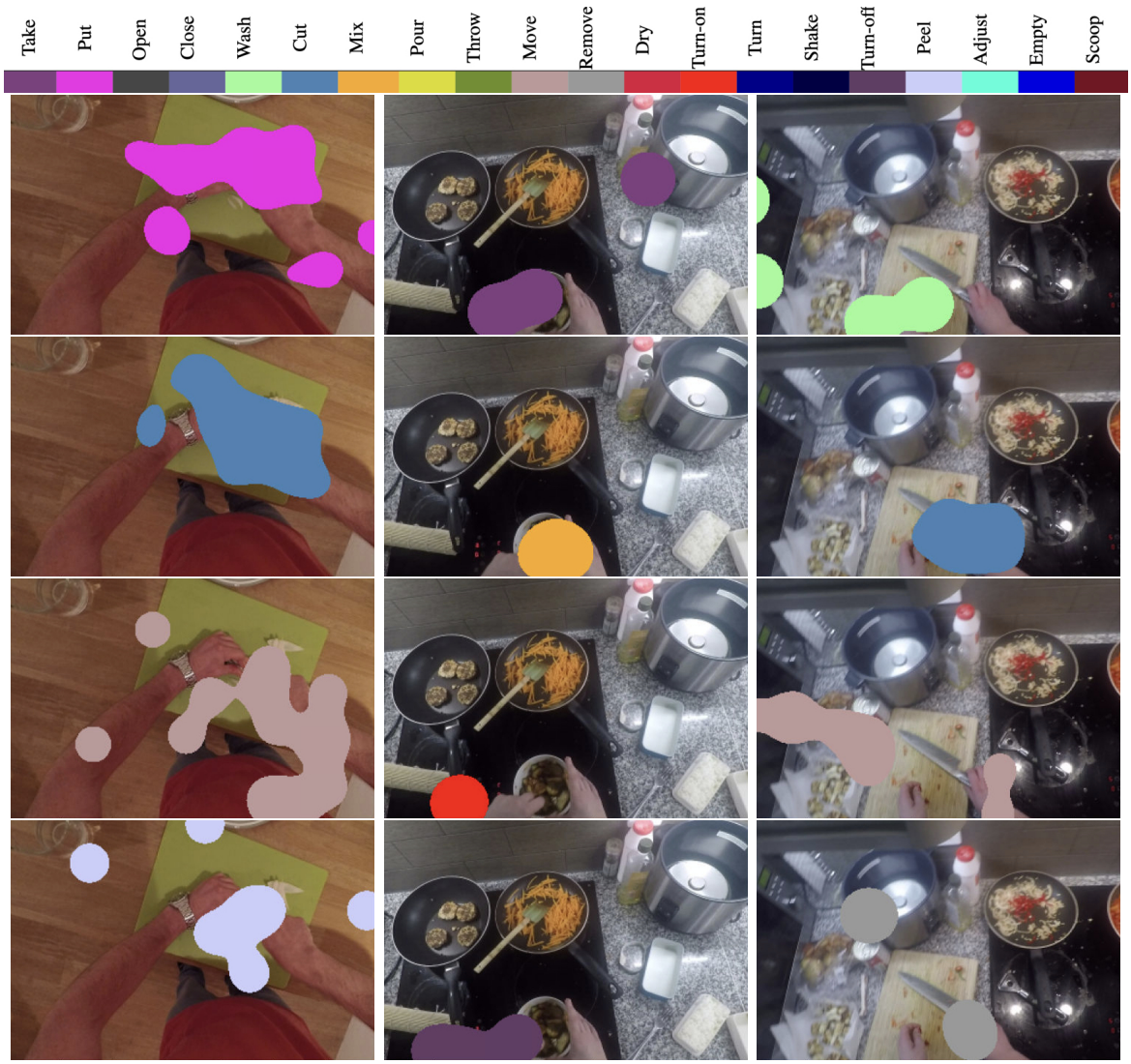


Figure 3.2: **EPIC-AFF Ground truth examples.** For visualization purposes, we show a single label of the affordable action on its location, although these are overlapped for the same sample. The food in the bowl affords *taking* or *mixing*, while the cutting board on the left affords *putting*, *cutting*, *moving* and *peeling*.

Dataset	Year	IG	Pix	ML	CP	#Obj.	#Aff.	#Imgs.
UMD [54]	2017	X	✓	X	X	17	7	30,000
IIT-Aff [55]	2017	X	✓	✓	X	10	9	8,835
ADE-Aff [56]	2018	X	✓	✓	X	150	3	10,000
OPRA [72]	2018	✓	✓	X	X	-	7	20,774
Grounded I.H [53]	2018	✓	✓	X	X	31	20	1,800*
Ego-Topo [39]	2020	✓	X	✓	X	304	75-120	1,020-1,115*
PAD v2 [113]	2021	X	✓	X	X	72	31	30,000
AGD20k [57]	2022	X	✓	X	X	50	36	23,816
EPIC-Aff	2023	✓	✓	✓	✓	304	20-43	38,876

Table 3.1: **Visual affordance datasets statistics**. I.G: Interaction Grounded. Pix: pixel-wise annotations. ML: multi-label. CP: camera poses #Obj: Number of objects. #Aff: Number of affordances. #Imgs: total number of images. * The affordance labels are only for evaluation, the model is trained supervised only by action labels.

Figure 3.3: **Extracting the center of the interaction.** Using the masks provided by VISOR Kitchens, we define the intersection between the object and the hand bounding boxes as the center of the interaction. We show in yellow the bounding box of the non-interacting objects, in green the bounding box of the hands and in blue the bounding box of the interacting object.

3.3.2 EPIC-Aff dataset

We detail the procedure shown in Figure 3.4 for our grounded affordance labeling. EPIC-Aff is composed of 38,876 images with up to 43 different affordable actions $\mathcal{K}$. We choose the EPIC-Kitchens as the base dataset because of its sequential and repetitive nature, which allows us to extract the 3D geometry, and because the kitchen is a scenario with multi-step and structured activities very rich in semantics. We cover all the object categories present in the EPIC-100 annotations, which constitute a wide, large and diverse knowledge base.

From a sequence of video, we join the EPIC-100 narrations [25] and the VISOR Kitchen annotations [188] to obtain a sparse sequence of frames $\mathcal{S}_{\mathcal{M}} = (f_1, ..., f_N)$ with the localization of the interactions on the image, as shows Figure 3.3. The EPIC-100 labels [25] contains narrations formed by an action verb $\mathcal{V}$ with an associated object $\mathcal{O}$,

i.e: "add steak", for more than 100 hours of video. VISOR Kitchens [188] interpolates from sparse annotations to generate semantic masks $\mathcal{M}$ and bounding boxes $\mathcal{B}$ on the active objects. We set the center of the interaction $x_i = \{u_i, v_i\}$ in the middle of the intersection between the hand $\mathcal{B}_h$ and the interacted object $\mathcal{B}_O$ bounding-boxes given by the narration $\mathcal{V} + \mathcal{O}$.

Then, we apply COLMAP, a Structure-from-Motion (SfM) algorithm [189] to obtain the camera poses T_w^c and a point cloud of the environment $\{X_p\}$. In the EPIC-Kitchens [25], each kitchen is composed of multiple videos, thus, we join all the sparse frames with interactions $\mathcal{S}_K = \{\mathcal{S}_1 \dots \mathcal{S}_M\}$ to relate frames from different videos in a common 3D reference. Furthermore, we use the VISOR semantic masks to avoid including dynamic objects in the point cloud.

Next, we employ a robust depth estimator based on a neural network $d_{NN} = f_d(\cdot)$ [190] to predict the depth of interaction points $d_{NN}(x_i)$. Because the neural network computes the depth up-to-scale, we compute a scale correction factor per image to fit the network scale to the SfM scale: $scale = median(d^{SfM}(X_p)/median(d^{NN}(X_p))$ [191], where $d^{SfM}(X_p)$ is the depth of all the points $\{X_p\}$ visible from the current image and $d^{NN}(X_p)$ is the depth of the same points given by the network estimator. Note that we employ an off-the-shelf depth estimator [190], since ground-truth depth annotations are unavailable in EPIC-Kitchens videos.

Using the predicted depth and the scale, we project the interaction point x_i in the 3D global coordinates X_i using the respective camera pose T_w^c as $X_i^w = T_w^c \cdot X_i^c$. Therefore, as it is shown in Figure 3.5, we obtain the history of all the interactions that occured in the kitchen in a common reference $\mathcal{I}_k = \{X_1^w, X_2^w, \dots, X_k^w\}$, cross-generalizing for the different sequences. This constitutes our knowledge base that follows our hypothesis that the distribution of affordances is spatially linked to pre-determined physical spaces (*i.e* you only wash in the dishwasher), not only to the semantic context of a topological graph [39].

Then, once we store all the past interactions $\mathcal{I}_k$ with their $\mathcal{V}_k$ and $\mathcal{O}_k$ labels, we reproject them back to the new camera reference system $X_i^c = T_c^w \cdot X_i^w$. Instead of considering all the object semantic masks as the affordance region, we center a Gaussian distribution over each affordance re-projected point X_i^c and build an additive heatmap. Then, the affordance masks $\mathcal{M}_i^{aff}$ are defined as the regions where the heatmap is greater than 0.25. This is grounded in how humans interact with objects [53] and allows us to consider the different affordability of the object parts (a knife is only *graspable* with the handle). In order to generate the affordance labels $\mathcal{A} = \{(\mathcal{V}_1, \mathcal{O}_1, \mathcal{M}_1^{aff}), ..., (\mathcal{V}_j, \mathcal{O}_j)\}$, we select only those verbs whose associated object $\mathcal{O}_i$ was present in the VISOR annotations $\{\mathcal{M}, \mathcal{B}\}$. With this procedure, we are

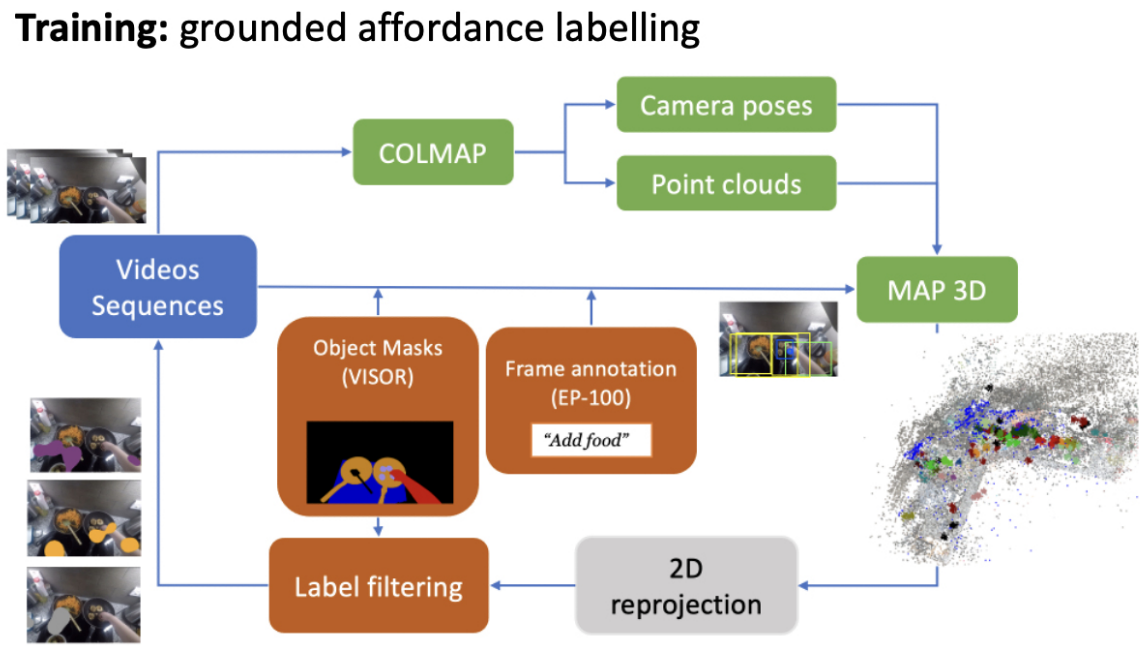


Figure 3.4: **Automatic extraction of EPIC-Aff labels.** We combine the EPIC-100 narration with the VISOR masks annotations to extract the interaction point. Then, using the camera pose extracted from COLMAP, we project all the interactions in a common 3D global reference. Finally, we reproject all the past interactions to each frame, and filter the affordance annotation by the objects present at the image.

grounding our dataset in the past interactions in that environment and associating multiple affordances to a single object. We show qualitative samples of the EPIC-Aff in Figure 3.2.

We provide two different versions of the dataset: the easy EPIC-Aff and the complex-EPIC Aff, with 20 and 43 affordance classes respectively. There is a challenging class imbalance, as shown the Figure 3.6 with a significant frequency gap between the most common class (*open*, with a 16.0 %) and the less represented (*dry*, with a 0.3 %). In Fig. 3.7 we show the pixel ratio, which reflects that semantically similar or opposite actions are associated with the same space (*i.e, turn-on, turn-off, adjust* or *cut* and *peel*) and the importance of the multi-level approach. This shows that activity-centric zones are physical spaces where there occur multiple common activities, both synonyms or antonyms.

3.4 Multi-label segmentation and mapping

In this section, we explain our inference procedure. First, we describe the modifications needed to obtain a multi-label segmentation model. We then show how our approach can be applied to mapping and planning tasks.

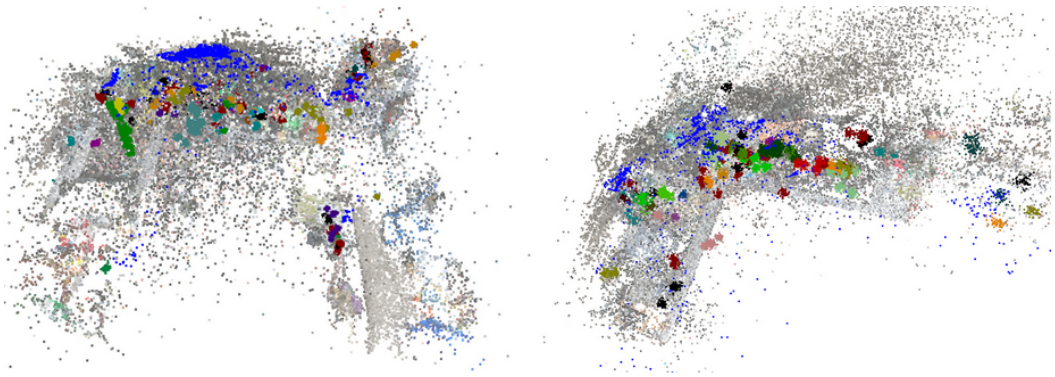


Figure 3.5: **Historical with all past interaction hotspots within the environment.** The blue dots represent the camera poses of the sparse frames from all the sequences.

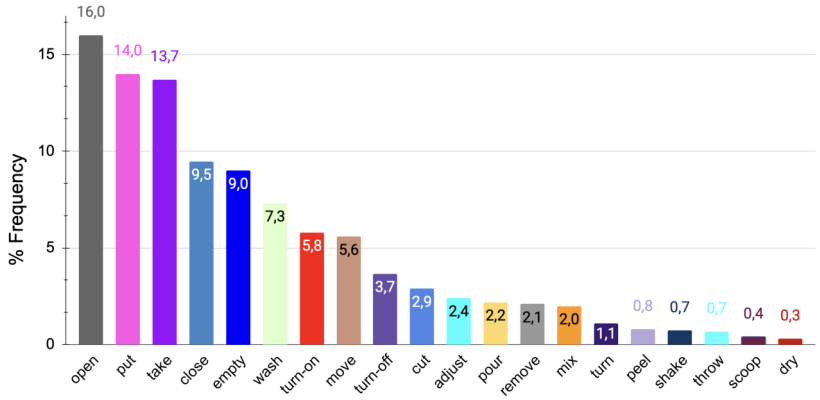


Figure 3.6: **Distribution of the 20 classes in the easy-EPIC Aff dataset.** The long-tail distribution shows a significant class imbalance.

3.4.1 Multi-label segmentation

In this section we describe how we transform classical semantic segmentation models to a multi-label version. While there exists lots of single-label segmentation [12,13,178, 192–194] and multi-label image classification works [170, 171, 174, 195], the multi-label segmentation is a more unexplored task restricted to small domains like biomedical images [167, 168].

Given an input image $\mathbf{X}$, the multi-label segmentation goal is to predict a group of categories for each pixel. Therefore, we assume that each pixel could represent multiple affordances (*takeable*, *cuttable*, *washable*, ..., etc.) or not belong to any category. For a total number of $\mathcal{K}$ classes we define the label y for each pixel of the image as $y = [y_1, ..., y_k]$, where $y_k = 1$ if the pixel contains the $\mathcal{K}$-category label, otherwise $y_k = 0$. In order to predict multi-label segmentation masks, we have evaluated two different approaches. First, we use a standard multiclass segmentation networks and evaluate three different heuristics to select multiple labels per pixel. Then, we modify

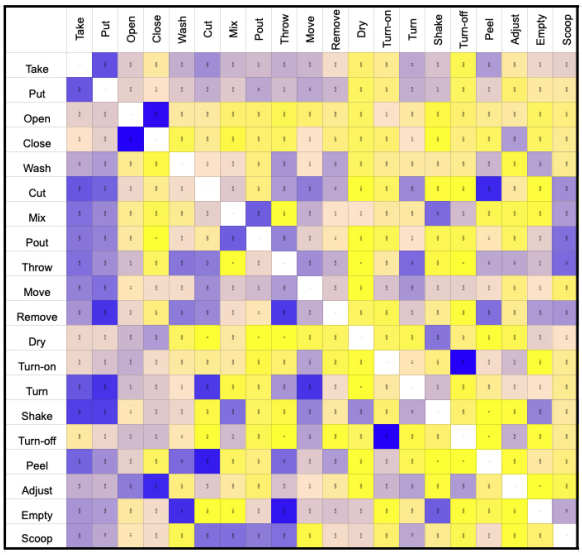


Figure 3.7: **Pixel ratio of the 20 classes in the easy-EPIC Aff dataset.** The *blue* color represents high correlation and *yellow* low concurrence in the same pixel.

the segmentation networks to output multiple binary classifiers which enable multiple labels to be active.

For the multiclass scenario, we assume that the network output is a categorical distribution for all the classes and use the standard supervision loss, the cross-entropy. Then, we transform the probability vector $p = [p_1, ..., p_k], \sum_{k=1}^{k} p_k = 1$ with three heuristics to choose the multiple winning-classes, as shows Figure 3.9. On the first method, we select the top-k classes with the largest probability value p_k. Note that we do not considers predictions with a $s_k < 1/k$, as it occurs to k_1, k_5 on Figure 3.9. The second alternative is max-θ, which consists in selecting all the possible classes whose p_k is greater than a threshold θ. Finally, the last heuristic is a dynamic θ_d threshold. We select the classes whose difference with the next class is larger than a θ_d.

On the multi-label scenario, the model outputs $\mathcal{K}$ independent Bernouilli distributions, generating binary probabilities $p = [p_1, ..., p_k]$, where we assume a detection if $p_k > 0.5$. Then, we substitute the cost function by a class-weighting Binary Cross Entropy (BCE) loss, obtaining $\mathcal{K}$ binary classifiers. One disadvantage of having independent binary classifiers is that the performance is more sensitive to the class imbalance in the dataset. To alleviate that, we use the Asymmetric (Asym) loss $\mathcal{L}_{asym}$ [172] shown in Equation 3.1. It combines the focal loss [196] with the margin loss [197] to reduce the contribution of easy negative samples, which rejects mislabeled samples with a continuous gradient and it is computed as follows:

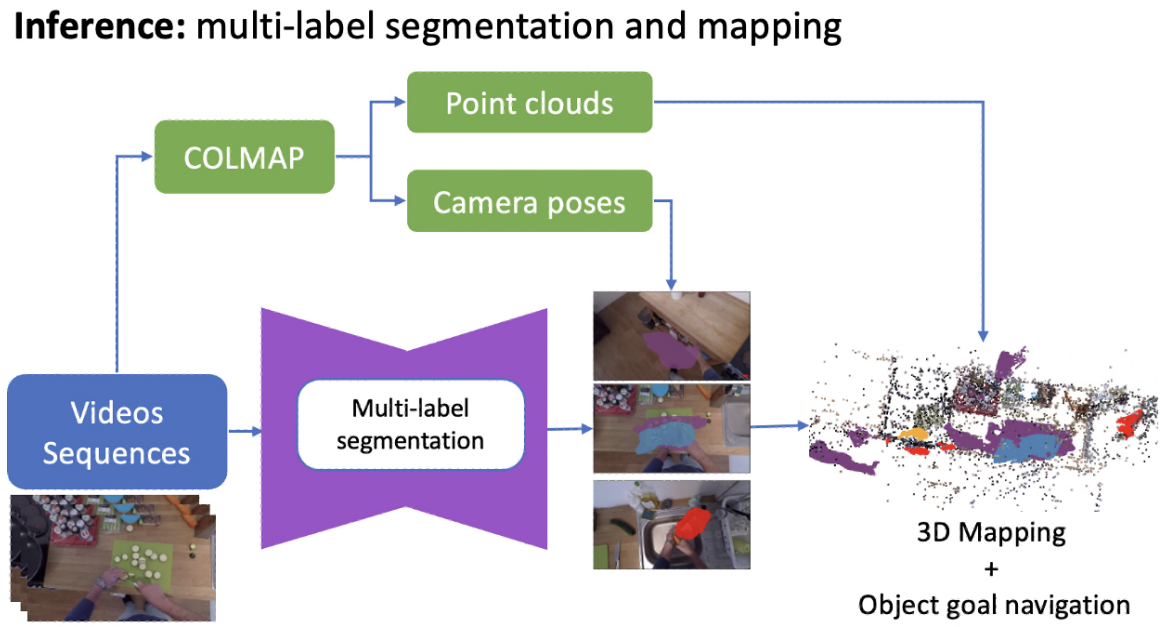


Figure 3.8: **Inference:** the multi-label masks predictions from our model are leveraged to a 3D map.

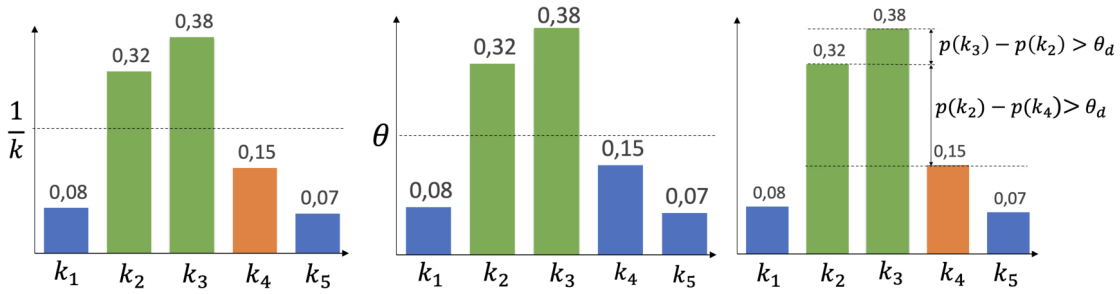


Figure 3.9: **Heuristics to select multiple labels from a probability vector.**

$$
\begin{aligned}
ASL_k &= \begin{cases} \log(p_k)(1-p_k)^{\gamma+}, & y_k = 1 \\ \log(1-p_k)(p_k)^{\gamma-}, & y_k = 0 \end{cases} \\
\mathcal{L}_{asym} &= \frac{1}{N} \sum^{N} \frac{w_k}{\mathcal{K}} \sum_{k=1}^{\mathcal{K}} ASL_k.
\end{aligned}
\tag{3.1}
$$

For each training image, $\mathbf{X}$ with N total pixels, $\mathcal{L}_{asym}$ computes a different term depending on if the y_k binary label indicating that the class k is present or not in the pixel. We apply a weighting average w_k depending on the ratio between positive and negative samples for class k to avoid the class imbalance. Following the original paper [172], we set $\gamma^+ = 4$ and $\gamma^- = 1$.

3.4.2 Embodied skill applications

Given that our system provides metric information of the affordance location, and it has information of the camera poses and has multi-label affordance detection, we can apply it to common spatial tasks such as mapping and navigation.

Mapping of activity-centric zones. We take the video sequence and sample unseen frames $\{f_1, ..., f_t\}$ during training. Following the same procedure as in the extraction of labels, we reproject the inferred semantic masks on the pixels i, j to its respective 3D location x, y, z using the camera intrinsic K_{int}, COLMAP pose R_w^c, t_w^c and the scaled depth $d_{i,j}$ as:

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = d_{i,j}(R_w^c)^{-1} K_{int}^{-1} \begin{bmatrix} i \\ j \\ 1 \end{bmatrix} - t_w^c. \tag{3.2}$$

We accumulate in a global map the COLMAP key-points to represent the geometry and the segmented affordances regions. We do not perform any fusion on voxels or octrees, since our multi-label approach assumes that a zone can represent several predicted classes. Note that the map representation is common for all the sequences of the same environment, with the potential of linking zones across multiple episodes and learning from past interactions.

Task-oriented navigation. Finally, we introduce a task-oriented navigation experiment to show the relevance of the map representation. We use the COLMAP key-points to build an occupancy grid with the available free space. Then, the agent is initialized in a random localization and asked to navigate to perform a certain action. Once it selects the location from the semantic-metric representation, the agent decides the path planning using a A* search with the Euclidean distance on the free space. We use the point cloud from COLMAP to create an occupancy grid.

3.5 Experiments

3.5.1 Models and metrics

In our experiments, we modify three popular semantic segmentation architectures [13, 198, 199] and compare them with an instance segmentation model [12] plus an interaction hotspots model [53].

– **Grounded Interaction Hotspots (GIH)** [53]: this work provides a trained version on EPIC-Kitchens scenes, which re-use to extract predictions from our

	KLD ↓	SIM ↑	AUC-J ↑	mIoU ↑	F1-Score ↑	mAP ↑	AP50 ↑
GIH [53]	2.381	0.116	0.511	17.5	29.4	14.2	15.5
Mask-RCNN [12]	1.365	0.150	0.841	40.1	56.5	**59.3**	**62.6**
U-Net [198] top-$\mathcal{K}$	2.532	0.341	0.830	9.5	17.4	22.0	30.5
U-Net [198] max-θ	2.532	0.341	0.830	13.2	23.6	22.0	30.5
U-Net [198] dyn-θ	2.532	0.341	0.830	13.2	23.7	22.0	30.5
Ours, U-Net + BCE	2.718	0.304	0.949	20.9	34.2	48.2	44.7
Ours, U-Net + Asym	0.783	0.665	0.857	17.7	29.9	15.6	32.3
FPN [199] top-$\mathcal{K}$	2.229	0.362	0.812	8.9	15.6	18.9	24.7
FPN [199] max-θ	2.229	0.362	0.812	12.4	21.8	18.9	24.7
FPN [199] dyn-θ	2.229	0.362	0.812	12.9	23.6	18.9	24.7
Ours, FPN + BCE	1.613	0.365	0.955	22.2	35.7	48.7	44.5
Ours, FPN + Asym	0.789	0.546	0.956	39.8	56.8	44.1	59.3
DeepLab-v3 [13] top-$\mathcal{K}$	4.947	0.192	0.911	18.9	31.9	35.0	40.9
DeepLab-v3 [13] max-θ	4.947	0.192	0.911	19.2	32.3	35.0	40.9
DeepLab-v3 [13] dyn-θ	4.947	0.192	0.911	19.5	32.7	35.0	40.9
Ours, DeepLab-v3 + BCE	1.276	0.179	0.964	31.3	47.2	58.6	56.2
Ours, DeepLab-v3 + Asym	**0.603**	**0.668**	**0.965**	**42.3**	**60.1**	43.6	58.5

Table 3.2: **Affordance multi-label segmentation on easy-EPIC Aff test set (20 classes).** Note that except the mIoU and the F1-Score, the rest of the metrics are common for the three versions of the multi-class segmentation models.

	KLD ↓	SIM ↑	AUC-J ↑	mIoU ↑	F1-Score ↑	mAP ↑	AP50 ↑
Mask-RCNN	2.287	0.211	0.756	17.1	27.3	**40.1**	**46.7**
Ours, U-Net + Asym	1.104	0.320	0.657	12.9	24.8	11.2	17.9
Ours, FPN + Asym	0.530	**0.673**	0.921	28.1	42.9	24.8	43.4
Ours, DeepLab-v3 + Asym	**0.520**	0.670	**0.931**	**31.1**	**46.5**	27.4	43.9

Table 3.3: **Affordance multi-label segmentation on complex-EPIC Aff test set (43 classes).**

images. To reduce the gap, we crop our scenes to represent a single object and compare for the same number of affordable actions $\mathcal{K}$ in the easy-EPIC Aff dataset.

– **Mask-RCNN** [12]: we assume an overlapping in the bounding boxes between two different instances. We do not consider the amodal Mask-RCNN versions [165, 200] which treat differently visible and occlusion masks, since our affordance classes $\mathcal{K}$ are not ranked in order.

– **Semantic segmentation architectures.** We compare the performance of three popular semantic segmentation models: U-Net [198], Feature Pyramid Networks (FPN) [141, 199] and DeepLab v-3 [13].

We train the segmentation models with an input resolution of 232×348 for 100 k iterations usign Adam as optimizer with weight decay of 10^{-4}, batch size of 8 and a initial learning rate of 10^{-4}, using a polynomial decay up to 10^{-6}. We apply random crop, color jitter, resize and flipping as data augmentation. In the same way, we train

	Take	Put	Open	Close	Wash	Cut	Mix	Pour	Throw	Move
GIH [53]	22.1	21.9	13.8	10.8	16.3	25.8	21.2	23.7	14.0	16.9
Mask R-CNN [12]	**37.7**	**36.9**	**47.1**	**43.9**	**51.5**	41.4	**46.4**	38.1	43.6	**42.9**
U-Net dyn-θ [198]	0.1	0.7	5.4	11.9	22.4	17.1	22.2	17.3	11.3	15.9
Ours, U-Net BCE	22.3	22.5	30.9	24.0	30.2	23.7	21.1	17.7	17.1	23.4
Ours, U-Net Asym	14.3	13.7	13.8	14.7	21.3	17.9	18.3	18.7	32.5	15.7
FPN dyn-θ [199]	2.4	2.4	5.6	10.2	21.7	13.2	17.7	17.0	11.5	13.6
Ours, FPN BCE	25.7	25.9	33.3	26.7	33.2	22.4	21.8	15.4	18.5	23.9
Ours, FPN Asym	36.3	34.7	46.1	42.0	46.8	42.7	42.2	37.5	43.3	41.7
Deep-Lab v3 dyn-θ [13]	10.1	11.0	15.4	17.3	19.1	19.4	25.2	19.1	14.7	17.9
Ours, Deep-Lab v3 BCE	33.3	34.2	44.1	37.6	43.1	32.0	30.9	26.2	28.9	33.7
Ours, Deep-Lab v3 Asym	31.6	32.9	37.3	37.8	44.5	**43.9**	45.0	**41.8**	**53.4**	42.3

	Remove	Dry	Turn-on	Turn	Shake	Turn-off	Peel	Adjust	Empty	Scoop
GIH [53]	17.2	12.8	10.3	20.5	16.6	10.8	26.1	9.5	13.9	25.8
Mask R-CNN [12]	38.4	13.1	**52.5**	43.7	30.8	**50.9**	35.3	47.1	33.6	26.0
U-Net dyn-θ [198]	21.0	4.5	14.8	21.1	16.3	18.4	12.9	20.6	0.5	9.6
Ours, U-Net BCE	18.5	8.7	27.3	22.4	13.2	23.8	16.2	22.2	19.0	13.1
Ours, U-Net Asym	15.6	16.6	15.2	18.9	22.2	19.5	19.5	24.3	5.7	15.7
FPN dyn-θ [199]	20.0	4.6	13.8	22.6	12.5	15.5	14.4	17.3	0.8	9.7
Ours, FPN BCE	21.4	7.4	3.8	20.8	13.3	28.4	13.6	24.5	23.5	11.6
Ours, FPN Asym	**39.6**	21.3	47.4	**43.7**	34.3	45.0	33.8	46.3	**38.0**	33.2
Deep-Lab v3 dyn-θ [13]	17.1	9.2	20.4	31.9	25.3	26.5	24.3	31.7	18.0	18.1
Ours, Deep-Lab v3 BCE	27.6	13.2	41.6	27.6	22.2	39.1	22.3	35.1	32.3	20.7
Ours, Deep-Lab v3 Asym	39.4	**33.1**	45.5	**52.2**	**44.0**	46.7	**43.5**	**51.1**	32.3	**46.6**

Table 3.4: **Class-wise IoU scores on the easy-EPIC Aff test set**. All scores are in [%].

Mask-RCNN SGD and 10^{-2} as initial learning rate. We use a Resnet-50 backbone pre-trained on Imagenet for all the models in order to perform a fair comparative.

Following the evaluation of Nagarajan et al, [53], we report the Kullback-Leibler Divergence (KLD) [72], the Similarity metric (SIM) and the Area Under the Curve (AUC-J) [201, 202] which provide different metrics for the mismatch of the distribution of heatmaps or affordance regions considering the predictive probability. We also report metrics from segmentation literature, such as the mean Intersection over the Union (mIoU) and the F1-Score to measure the performance of the semantic segmentation, and the Average Precision (AP) AP-50 and mAP to report the performance of the detection metrics.

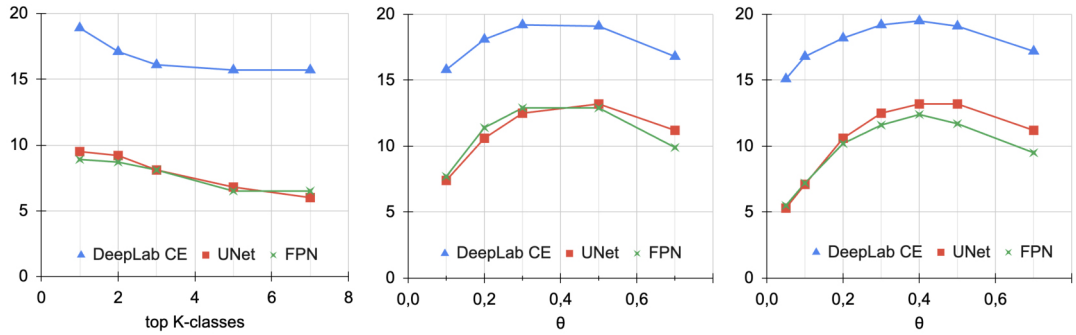


Figure 3.10: **Evolution of the mIoU for different heuristics to select multiple winning classes from a multi-class probability vector.** Left: top-$\mathcal{K}$. Center: max-θ. Right: dyn-θ.

3.5.2 Quantitative results

We compare the performance of different popular architectures on the multi-label affordance segmentation task in Table 3.2 on the easy-EPIC Aff dataset. DeepLab-v3 trained with the Asymmetric loss obtains the best performance on the segmentation and saliency metrics (42.3 % mIoU 60.1 % F-1 score, 0.603 KLD, 0.668 SIM, 0.965 AUC-J). Since the backbone of the three semantic segmentation models is the same (Resnet-50), the different results are due to the configuration of the different decoders. In the dataset, since the labels represent interaction hotspots, they are not aligned with the borders of the objects and represent a more high-level zone. Thus, the atrous convolution of DeepLab enables to enlarge of the filter's field of view and better captures these regions. We show the per-class segmentation performance in terms of the IoU in Table 3.4. Comparing with the apparition frequency of the classes in the dataset shown in Figure 3.6, Mask R-CNN fails at the low-represented classes since is not trained with the Asymmetric loss. However, it is the best architecture on the detection metrics (59.3 % mAP, 62.6 % AP50). Compared with previous works, the pre-trained version of [53] achieves intermediate results on the segmentation metrics (17.5 % mIoU, 29.4 % F-1 score) but low on the AP scores.

The results in Figure 3.10 show the impact of the hyper-parameters when adapting the multi-class models. The top-$\mathcal{K} = 1$ represents the classical multi-label case. The results show how its performance is far from the multi-label versions, supporting the need for specific architectural changes for this scenario. In Figure 3.10 left, when we increase the number of winning classes, the performance decreases by introducing too many false positives. The other two heuristics achieve better performance since they better reject these outliers. For example, the dyn-θ adapts dynamically to the probability distribution shape, obtaining higher mIoU and F-1 in the three cases (see Table 3.2). Finally, we appreciate similar results on the complex-EPIC Aff, shown in

Table 3.3. In this case, the overall performance decreases due to the higher number of classes and its imbalance.

3.5.3 Mapping: metric distribution of affordances

We show on Figure 3.11 qualitative results of the multi-label interaction hotspots from affordances predicted by the DeepLab-v3 Asym model. This is consistent along different time-steps. For example, the microwave of the left-map in Figure 3.11 is detected as *turn-on* both at $t = 0, 25, 52$. The qualitative results clearly motivate our multi-label approach: the milkshake on the right map affords *mixing*, *pouring* and *taking*, or the sink in the center map affords *drying* and *washing*. Our pixel-wise conception is also required, since it reflects the interactions hotspots rather than highlight the complete object (for example *grasping* a pan only with the handle).

3.5.4 Task-oriented navigation

Finally, we use the spatial localization of the affordances to show a proof-of-concept "task-oriented" navigation. As we illustrate on Figure 3.12, we guide the agent according to the action possibilities that the environment offers to him. Therefore, we can ask our system to perform certain action, meaning to *go to where the object and affordance are available*. The A* indicates to the agent the shortest path from its current location to the position where it took the action in the past. For example, this could guide a visually impaired person with an assistant device [203].

3.6 Limitations

Our current approach presents several limitations. At the dataset extraction, we assume that the interaction occurs in the intersection between the object and the hand bounding-boxes, thus it depends on the bounding-box aligned to the actual object. This could be mitigated with a detection model for grasping points, but we wanted a simpler version for our prototype as a more convoluted approach might introduce further biases, difficult to detect. Also, the camera poses from COLMAP can be distorted by noisy-frames or dynamic objects non-suppressed by the mask. Furthermore, a real-time mapping system would require a SLAM system such as ORB-SLAM [183] which might reduce the accuracy of COLMAP. Similarly, our dataset is fully based on Kitchen sequences and it does not incorporate another environments introducing important dataset bias in the trained models. However, our automatic labeling pipeline could be easily used to extend the dataset in other scenarios.

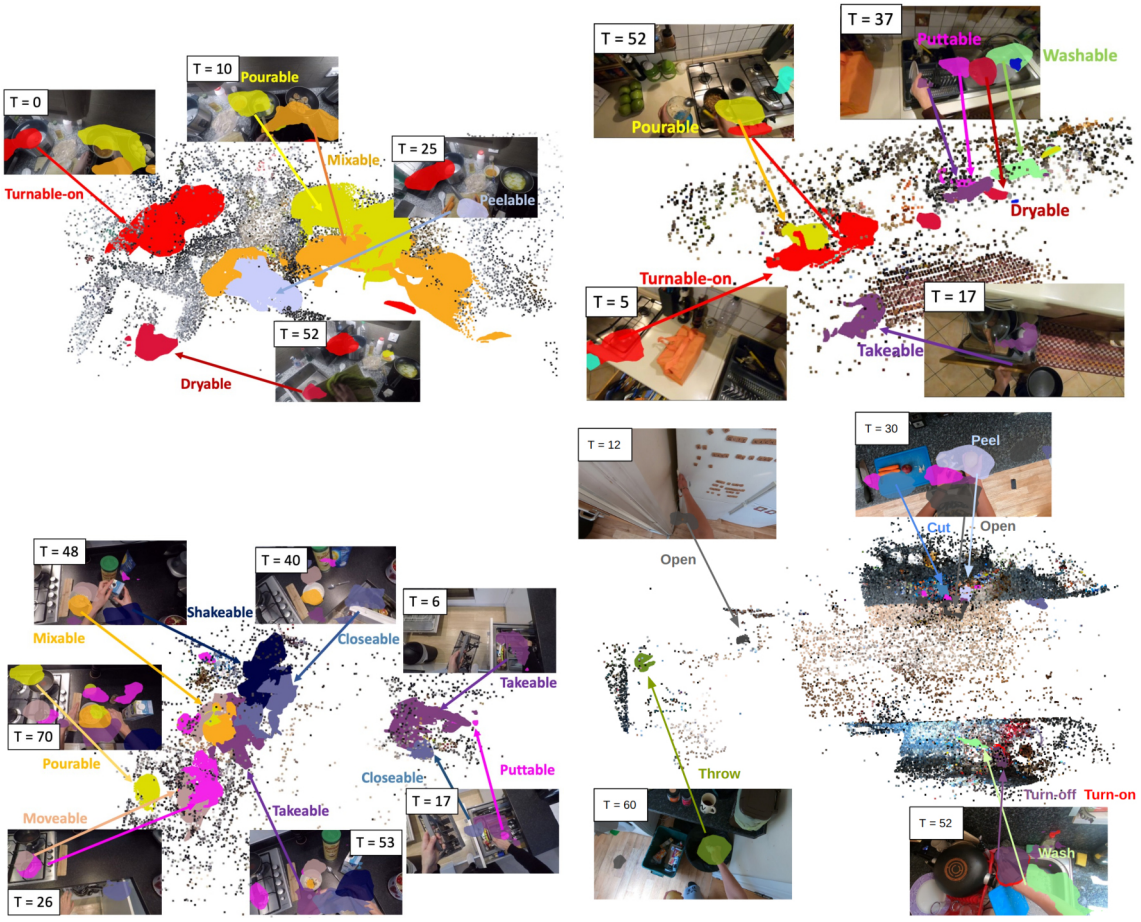


Figure 3.11: **Multi-label affordance mapping.** We show examples on four different scenarios.

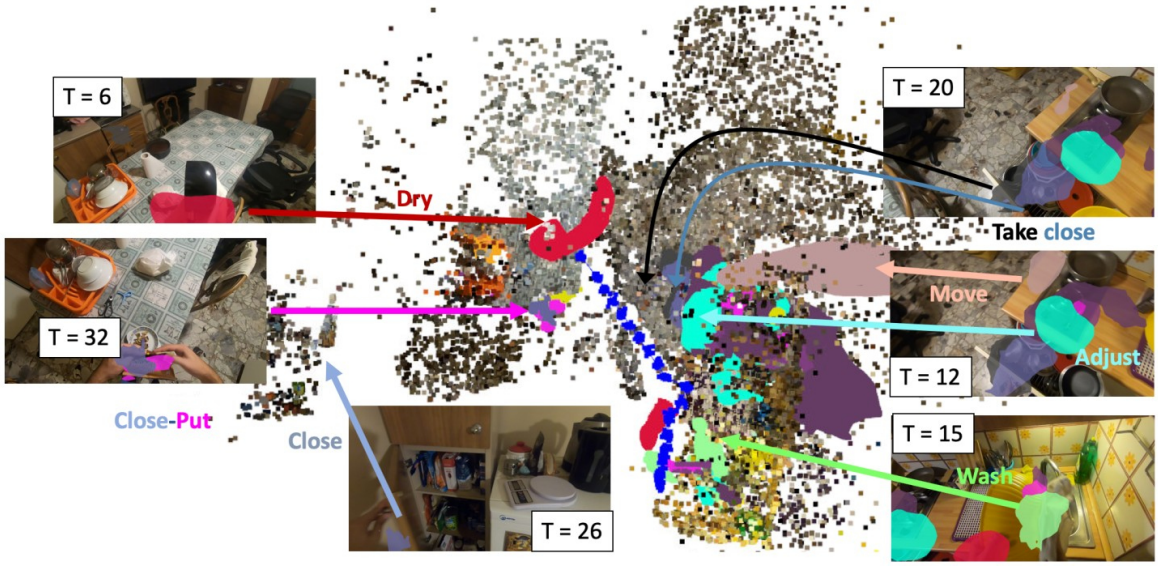


Figure 3.12: **Goal oriented path-planning.** In the example, at t=36 we indicate the user the trajectory from the sink to the place where it used to dry the crockery. The blue points represents the steps of the path planning

3.7 Conclusions

We introduced a novel multi-label, metric and spatial-oriented perception of affordances. First, we present a method for extracting grounded affordances labels based on egocentric interaction videos through a common metric representation of all the past interactions in a common reference. We use this pipeline to build the most complete affordance dataset based on the classic EPIC-Kitchen dataset. This constitutes EPIC-Aff, the largest semantic segmentation dataset of affordances grounded on the human interactions. We also motivate a method for grounded affordance detection with pixel precision using multi-label predictors, which enhances the perception and the representation of the environment. Furthermore, we show that the metric representation obtained can be used to build detailed affordance maps and to guide the user to perform task-oriented navigation tasks.

Chapter 4

Robust Fusion for Bayesian Semantic Mapping

In the previous chapter, the proposed multi-label mapping lacked any temporal aggregation strategy—that is, the predictions were not fused into a voxel grid or 3D surface, but were instead projected (and accumulated) onto a 3D point cloud. To address this limitation, I introduce in this chapter a Bayesian semantic mapping that fuses predictions in a semantic voxel map. First, I present a novel fusion strategy to regularize observations and mitigate the impact of prediction bias. Second, my approach utilizes the Dirichlet distribution from the output of a Bayesian neural network to incorporate epistemic uncertainty, thereby reducing the influence of overconfident outlier predictions and resulting in a more robust semantic voxel representation.

4.1 Introduction

Robots rely on understanding their surroundings to work autonomously and make informed decisions. Semantic information is critical for enabling reasoning on a higher abstraction level in complex tasks such as recognizing different objects [204], driving safely [205] or helping in our homes [206]. Neural networks have played an important role in making this possible, but their application in robotic perception pipelines still presents multiple challenges [207]. In this chapter, we address the problem of robustness when combining multiple, potentially biased, neural network predictions in a semantic mapping pipeline. Compared to traditional sensors, which are well understood and calibrated, the behavior of neural network models still lacks interpretability. Firstly, neural network predictions assume that they are always within the trained distribution and are agnostic to the real distribution of the input data. Secondly, even if the output

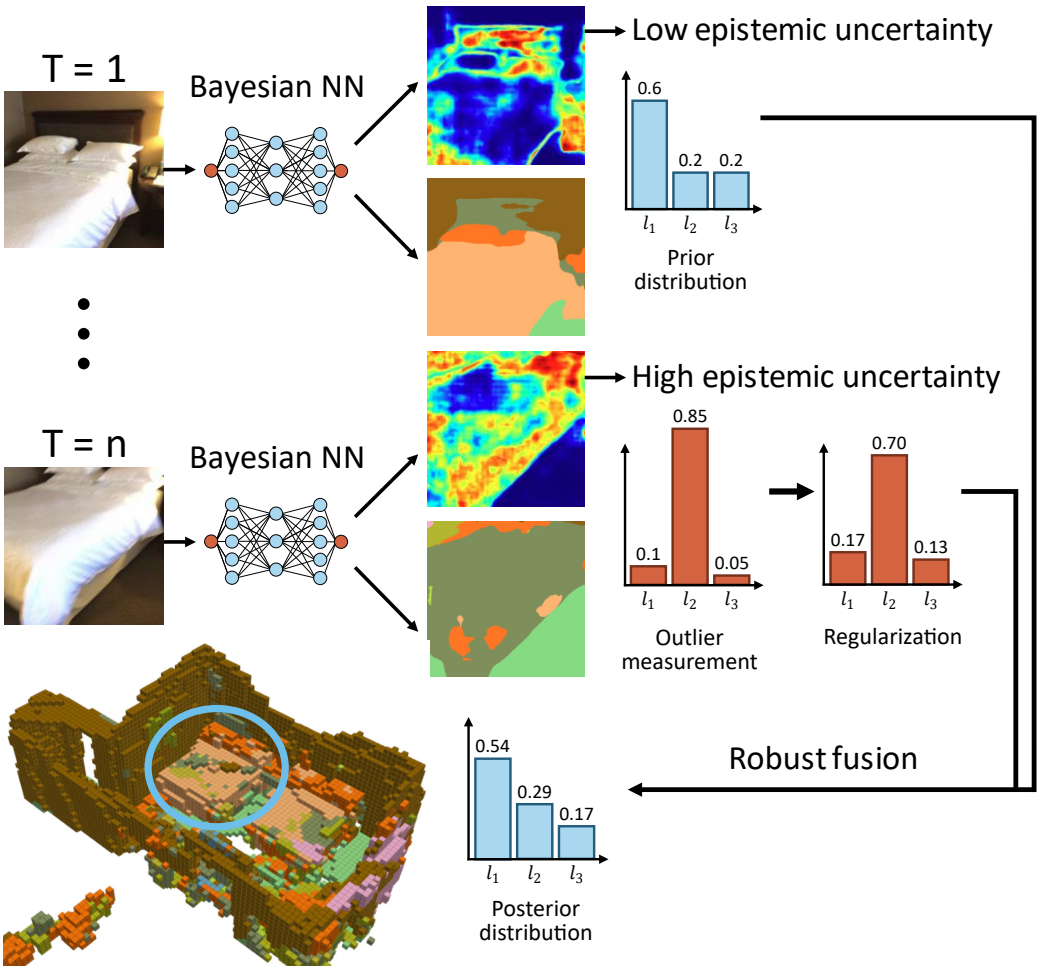


Figure 4.1: **Robust Fusion for Bayesian semantic mapping.** Since neural networks are overconfident sensors, they output misclassified predictions with high confidence. Using Bayesian neural networks, we regularize their output and weight observations according to their epistemic uncertainty, obtaining a fusion method robust to outlier detections.

of traditional neural networks can be understood as a probability (i.e., classification confidence among a set of classes), these are often overconfident and produce misclassified predictions with a high confidence level [82]. Despite this, existing fusion methods [208–211] in semantic mapping fully trust the neural network predictions, which makes them vulnerable to the aforementioned limitations.

To prevent these problems, our main contribution in this chapter is a new fusion method that exploits the information given by a Bayesian neural network to increase the robustness of the resulting semantic map. As we show in Chapter 2, Bayesian deep learning techniques offer more interpretability by extracting uncertainty from the predictions. Similarly, we adopt the previously discussed sample-based Bayesian deep learning techniques, specifically we extend a common segmentation neural network with intermediate dropout layers for performing MC-Dropout.

Although uncertainties have been successfully utilized in embodied active learning [136, 212, 213] and domain adaptation [214], their potential in calibrating neural networks as sensors and enhancing the robustness of semantic information acquisition at test time has scarcely been investigated. Therefore, our goal is to leverage the uncertainties provided by Bayesian neural networks to obtain a better semantic mapping. In particular, the contributions of this chapter are the following:

- We apply a regularization in the observations to reduce the influence of overconfident outlier measurements on the posterior distribution.
- We obtain a Dirichlet distribution from the output of a Bayesian neural network to incorporate the lack of knowledge from the network, which allows to fuse the data considering the epistemic uncertainty of each measurement.
- We validate the method in simulated and real environments, showing improvements with respect to non-Bayesian approaches.

The work described in this chapter was presented in IROS 2023:

- David Morilla-Cabello*, Lorenzo Mur-Labadia*, Ruben Martinez-Cantin, and Eduardo Montijano. Robust fusion for Bayesian semantic mapping. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS).* Core Ranking A, 2023

4.2 Related Works

4.2.1 Semantic Mapping

There are several ways to use neural networks for semantic mapping, including end-to-end, averaging, and fully Bayesian methods. For the first type, the data association and update of the semantic probabilities are learned in an end-to-end pipeline including recurrent structures to account for the temporal dimension [215,216]. Even though these are interesting approaches, end-to-end pipelines are complex to analyze and require specific training.

Averaging methods aggregate directly the predictions obtained from the neural network. In some cases, this is done considering only the winning class in each observation and performing a voting scheme [217, 218]. In other cases, the probabilities given by the network are averaged for the total number of observations [209,219]. The former is motivated by the excessive confidence typically shown by semantic segmentation neural networks. While these approaches are very efficient, Bayesian methods provide a better framework to handle uncertainties and probability distributions.

Bayesian methods offer an interesting framework to combine the prediction of the network in a probabilistic fashion. Among these types of works, Semantic Fusion [208] updates the probability distribution of the surfels generated by their SLAM system using a recursive Bayesian update. This is also used by current semantic mapping approaches [210]. Similarly, the work by Asgharivaskasi et al. [211] sums the log probabilities. Panoptic mapping [220] separates meaningful objects from the background and combines the probability of the object classes by weighting based on the class confidences. All these methods propose a Bayesian update based on the probability distributions coming from the neural network. Thus, they assume that the network's predictions are correct and that the confidences are well-calibrated. In contrast, we leverage Bayesian neural networks to model the uncertainty in sensor observations and perform robust semantic fusion

4.2.2 Bayesian Deep Learning

Bayesian neural networks (BNNs) infer a distribution over the network weights to model the uncertainty of the training process. Since the analytical computation of the posterior is intractable, several methods propose computing approximate distributions. Monte-Carlo (MC) dropout [221] and deep ensembles [105, 222] are two popular sampling techniques that approximate the true posterior using discrete samples. On the other hand, feature-space techniques [120, 123, 223] estimate the uncertainty with a single-pass by measuring the distance or density of the samples in feature space compared with the training data.

BNN applications in robotic systems show multiple uses of uncertainty estimation in perception and planning for learning, but not in guiding semantic mapping. At the perception stage, the uncertainty in the model, or epistemic uncertainty, appears in regions where the model parameters are less confident or poorly estimated, leading to ambiguous or challenging pixels [83, 111] and identifies false positive detections [132]. The aleatoric uncertainty, or entropy of the data distribution, stems from inherent randomness or variability in the data distribution. It is present in the contour of objects and noisy regions [131]. Other works employ the uncertainty at the planning stage to perform active learning [135], which reduces the labeling cost by selecting the most informative samples [136] and allows for domain adaptation [214]. Finally, epistemic uncertainty has been used in object goal navigation to make informed decisions about exploring uncertain areas or reaching known target objects [137]. To the best of our knowledge, few works consider epistemic uncertainty as a weight for semantic fusion in object detection tasks [224] and planning [225]. In our case, we apply the epistemic uncertainty of BNNs to semantic fusion over all the pixels in the image.

4.3 Approach

Our objective is to perform robust semantic mapping using a given uncertainty-aware NN on a mobile sensor. This is accomplished by considering the model's confidence and regularizing the map fusion in order to account for re-observations and reject outliers coming from overconfident predictions. First, we describe the semantic map structure and introduce an overview of Bayesian neural networks to estimate the observations' uncertainties. Then, we introduce the novel Bayesian fusion method based on confidence weighting and regularization.

4.3.1 Map Description

We consider a generic representation for the map using a voxel set, $\mathcal{M}$, for all the detected surfaces. Each voxel, $m \in \mathcal{M}$ has a semantic label, $l_m \in \mathcal{K} = \{1, \dots, K\}$, where $\mathcal{K}$ is the set of pre-defined possible classes. The objective of the semantic mapping algorithm is to infer the semantic class for all the voxels. In order to do that, the categorical probability distribution of a voxel m is defined by the probability of the voxel belonging to each of the defined classes,

$$p(l_m = l_i|\mathbf{p}) = \prod_{j=1}^{k} p_j^{[j=i]}, \tag{4.1}$$

where $\mathbf{p} = (p_1, \dots, p_k)$ such that p_j represents the probability of the voxel belonging to class j, the exponent, $[j = i]$, is the indicator function, and $\sum_{j=1}^{k} p_j = 1$. Without loss of generality, we describe the whole semantic fusion process for a single voxel m, denoting $p(l_i) \equiv p(l_m = l_i|\mathbf{p})$, understanding that the same procedure is applied to all the voxels concurrently.

Semantic inference is carried out by aggregating all the measurements that are taken of the voxel in a Bayesian fashion. Denoting $\mathbf{X}_t$ as the measurement acquired at time t, the objective is to obtain the posterior distribution, $p(l_i|\mathbf{X}_{1:t})$, where the initial prior is a uniform distribution, to model the fact that there is no initial knowledge of the class. In this chapter, $\mathbf{X}_t$ is the sensor input, which can be an image or a LIDAR scan, for example. Then, we let $p(\mathbf{Y}_t|\mathbf{X}_t)$ be the class probability distribution associated with each data $\mathbf{Y}_t = f_\omega(\mathbf{X}_t)$ (pixel, LIDAR point, etc.), which is obtained from a semantic segmentation neural network $f_\omega(\cdot)$. Under this observation model, standard semantic Bayesian fusion [208] approximates the posterior of a voxel by

$$p(l_i|\mathbf{X}_{1:t}) \propto p(l_i|\mathbf{X}_{1:t-1}) \prod_j p(y_j = l_i|\mathbf{X}_t), \tag{4.2}$$

where the product, $\prod_j$, includes, for the time t, the NN outputs $y_j \in \mathbf{Y}_t$ (labeled pixels, points, etc.) from $\mathbf{X}_t$ whose projection falls in the corresponding voxel. We assume the computation of this projection is given by existing geometric mapping algorithms, e.g., SLAM, noting that both Eq. (4.2) and our proposed fusion method can be applied independently of how the metric part is computed.

The inclination of NNs to return over-confident observations is an important problem for (4.2) because a single outlier with high probability has a strong negative impact in the posterior distribution. A simple numeric example of this problem is illustrated in Figure 4.2, where it can be observed how one measurement shifts the posterior distribution to wrong values independently of the number of measurements in $\mathbf{X}_{1:t-1}$.

4.3.2 BNN for Semantic Observation

As a first step towards obtaining a more robust semantic fusion mechanism, we consider the use of a Bayesian neural network, instead of a standard neural network.

Considering a posterior distribution $p(\omega|\mathcal{D})$ on the neural network weights after training on dataset $\mathcal{D}$, the semantic observation of the network can be defined by the predictive posterior distribution:

$$p(y_j|\mathbf{X}_t, \mathcal{D}) = \int_\omega p(y_j|\mathbf{X}_t, \omega)p(\omega|\mathcal{D})d\omega. \tag{4.3}$$

Without loss of generality, we use Monte-Carlo dropout to compute the posterior distribution $p(\omega|\mathcal{D})$, but other methods such as deep ensembles, or feature density could be used. We approximate the predictive posterior distribution using directly the Monte-Carlo samples as follows:

$$p(y_j|\mathbf{X}_t, \mathcal{D}) \approx \frac{1}{M}\sum_{i=1}^{M} p(y_j|\mathbf{X}_t, \omega^{(i)}, \tag{4.4}$$

where $p(y_j|\mathbf{X}_t, \omega^{(i)})$ is the output of the network for sample $\omega^{(i)}$. The Monte-Carlo samples are generated at test time, sampling a Bernoulli distribution that multiplies the values of each weight in the network. In practice, this is implemented by dropout layers that remain active at test time.

Importantly, this definition of the observation enables the computation of the uncertainty associated with the output of the network $\Sigma(\mathbf{X}_t)$, which can be divided into $\Sigma = \Sigma_a + \Sigma_e$. In this case, $\Sigma_a(y_j)$ is called the aleatoric uncertainty, it is related to the data noise and it is already encoded in the entropy of the output probabilities. Notice that, higher $\Sigma_a(y_j)$ will have a low influence in Eq. 4.2 with the uniform distribution having no effect. The epistemic uncertainty $\Sigma_e(y_j)$ represents the uncertainty of the

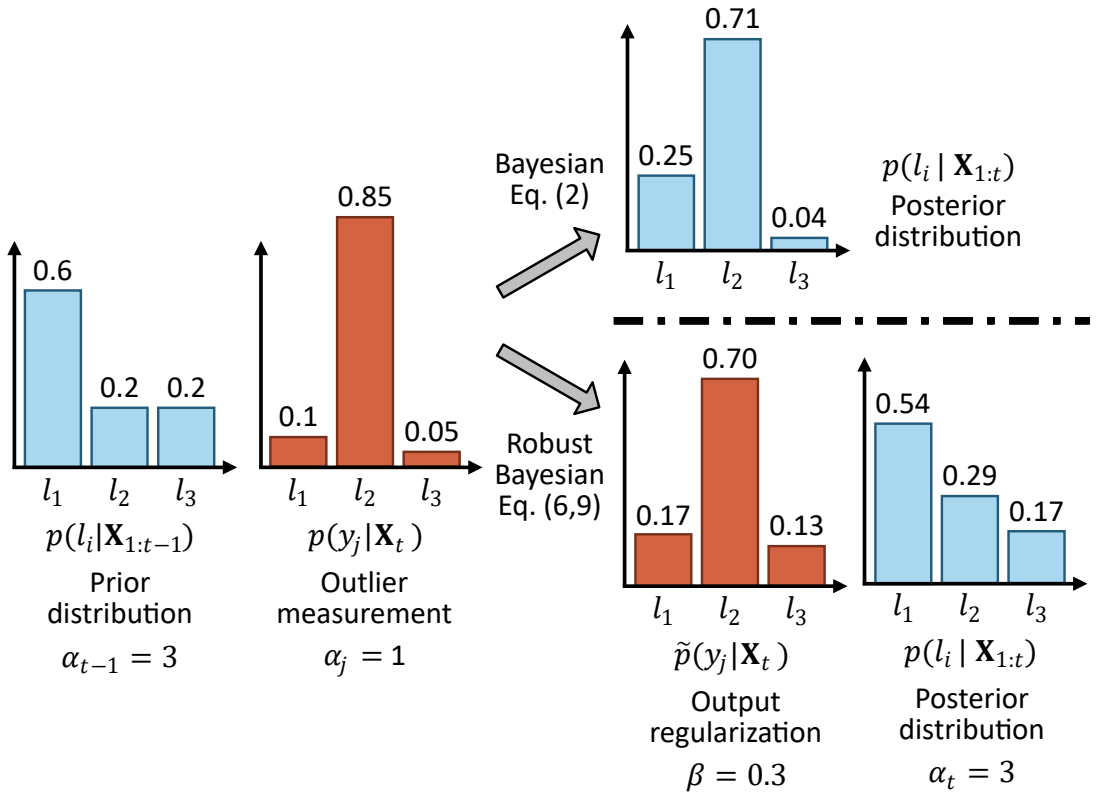


Figure 4.2: **Influence of an outlier in traditional approaches vs. our robust semantic Bayesian fusion.** One wrong observation (middle distribution) can shift drastically the prior distribution (left), to values where the highest probability belongs to the wrong class (right upper). Our method (right lower) first regularizes the measurements to avoid overconfidence and considers the epistemic uncertainty of the model in the Bayesian fusion with the α term.

model or the lack of knowledge with respect to a new input $\mathbf{X}_t$,

$$\Sigma_e(y_j) = \frac{1}{M}\sum_{i=1}^{M}\left[\left(p(y_j|\mathbf{X}_t,\omega^{(i)}) - p(y_j|\mathbf{X}_t,\mathcal{D})\right)\cdot\left(p(y_j|\mathbf{X}_t,\omega^{(i)}) - p(y_j|\mathbf{X}_t,\mathcal{D})\right)^T\right]. \tag{4.5}$$

The lower this quantity, the more we can trust the prediction of the network about the observation.

4.3.3 Robust Fusion Algorithm

In order to reduce the sensitivity of the fusion to outliers with low aleatoric uncertainty, i.e., wrong observations with high confidence in the winning class, we first include a

constant regularization term, β, in the output of the network,

$$\tilde{p}(y_j|\mathbf{X}_t, \mathcal{D}) \approx (1-\beta)\frac{1}{M}\sum_{i=1}^{M} p(y_j|\mathbf{X}_t, \omega^{(i)}) + \beta\mathcal{U}, \tag{4.6}$$

where $\mathcal{U} = (\frac{1}{k}, \ldots, \frac{1}{k})$ is a uniform distribution over the semantic class set $\mathcal{K}$. This equation can be understood as the marginalization of the probability distribution given by the neural network, conditioned to the probability it has to make a mistake, defined by β.

The positive consequence of the regularization is that overconfident outlier measurements will have less influence on the posterior distribution. On the other hand, the mapping algorithm will require more measurements of each voxel before committing to a specific class.

Second, we modify the categorical distributions from Eq. (4.2) by Dirichlet distributions $\mathcal{D}ir(p, \alpha)$, where the concentration parameters $\alpha = \{\alpha_i\}_{i=1}^{K}$ are inversely related to the epistemic uncertainty component, that is,

$$\alpha_{t,j,i} = -\log(\Sigma_e(y_j, l_i)), \tag{4.7}$$

where $\Sigma_e(y_j, l_i))$ represents the marginal variance associated with label l_i. Therefore, our Bayesian fusion equation is

$$p(l_i|\mathbf{X}_{1:t}) \propto p(l_i|\mathbf{X}_{1:t-1})^{\frac{\bar{\alpha}_{t-1,i}}{\bar{\alpha}_{t,i}}} \cdot \prod_j \tilde{p}(y_j = l_i|\mathbf{X}_t, \mathcal{D})^{\frac{\alpha_{t,j,i}}{\bar{\alpha}_{t,i}}}, \tag{4.8}$$

where $\bar{\alpha}_{t,i}$ is a normalizing constant equal to the the maximum value up to time t for label l_i, that is,

$$\bar{\alpha}_{t,i} = \max_{\tau<t,j}\{\alpha_{\tau,j,i}\}. \tag{4.9}$$

We can think that the Dirichlet distribution in this case represents the confidence on the underlying distribution. For example, if we have an observation with a higher concentration, it is equivalent to fusing multiple observations with that value, therefore having a higher multiplicity in Eq. (4.2). Figure 4.2 illustrates the effect of the regularization in the measurement and the effect of the fusion of Dirichlet distributions in the computation of the posterior.

4.4 Experiments

We evaluate the impact of our robust fusion method on the semantic mapping task and investigate the advantages it offers compared to existing fusion methods. To achieve this, we use an RGB-D camera to capture images of different environments along a fixed

Figure 4.3: **Environment visualizations.** Overview of the *Office* and *House* virtual scenes and two environments from the real dataset, ScanNet (sequences 6 and 40).

trajectory with known poses, so that uncertainties and errors related to geometry do not affect the analysis. These images are then input to a Bayesian semantic segmentation network, and the resulting predictions are projected onto a map using the camera's depth information and fused considering different methods.

4.4.1 Environments

Ideally, we would like to abstract the problem of semantic fusion from other uncertainties arising from sensor pose and depth estimation. Furthermore, having accurate semantic ground truth (GT) is also advantageous. Therefore, we initially test our semantic mapping in a photo-realistic simulation environment using Unreal Engine 4[1] and the Airsim simulator to model a virtual camera in the scene. The simulator provides the sensor poses, RGB images, and GT depth and semantic labels. We model two different environments called *Office* and *House*. Figure 4.3 offers an overview of the environments. We conduct mapping over three different trajectories within each environment and evaluate the aggregation of the results.

We further evaluate our approach on three sequences from ScanNet (6, 9 and 40). This is a real dataset with annotated 3D reconstructions of indoor scenes that provide GT camera poses and depth measurements [226]. This dataset validates the application of our method to real-world scenarios and how it performs under noisy depth measurements coming from a real device.

[1] `https://www.unrealengine.com/en-US`

For the mapping framework, we use a simple voxel hash-map. By relaxing the need to deal with occupancy probability, we only store surface (i.e., occupied) voxels and avoid storing the *free* space, reducing the storage and increasing the efficiency, resulting in a simple scheme to study the semantic fusion problem. We always use a voxel size of $0.1m$.

4.4.2 Sensor configuration

We use a BNN based on the semantic segmentation model DeepLab-v3 [227]. We place a dropout layer after each ResNet block with a dropout rate $d = 0.3$, following [111]. At inference, we keep the dropout layers active and obtain 32 samples from the approximate posterior using MC dropout. For the baseline models, we use the same network with inactive dropout layers, as in traditional deterministic models.

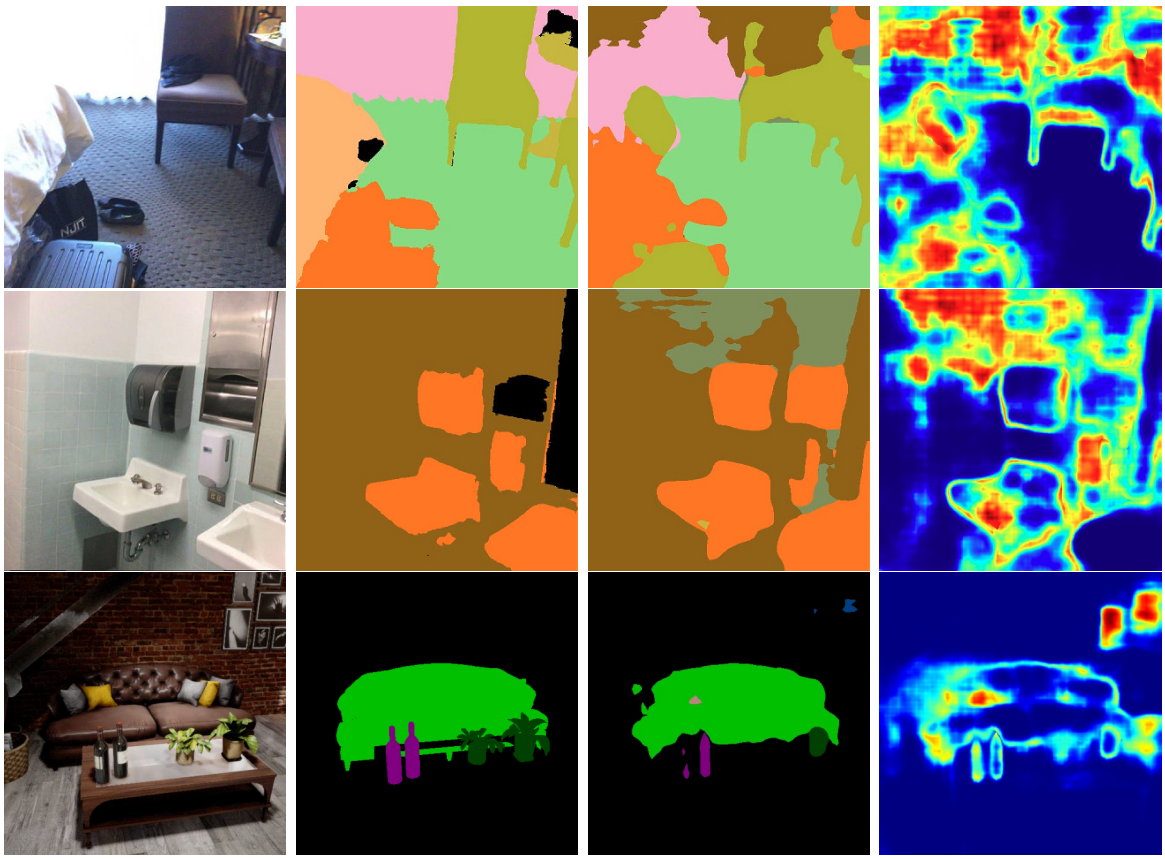
Real applications require robust methods that can generalize to new situations. We introduce this challenge in our experiments by training on different environments than that of the testing. For the simulated environment, we train on real data using the VOC12 Dataset [228]. We select the samples that contain any of the six classes of interest (see Table 4.1), corresponding to indoor objects present in the simulation setup. For the real environment, we train on the NYU-v2 dataset [229] with 12 labels and evaluate the semantic mapping in sequences from the ScanNet dataset [226].

4.4.3 Metrics and Baselines

In order to assess the quality of the mapping, we measure the Intersection over Union (IoU) over each of the classes and compute the mean IoU (mIoU). We also compute the Accuracy for all the voxels in the map. To evaluate each part of the robust fusion, we perform an ablation study using the regularization (R), the Dirichlet model (D), and both at the same time (D+R). The parameter β used in the regularization is set to 0.3 for all the experiments. Adapting this value depending on the environment and the class might improve the results and will be explored in the future. We also compare our method with three current semantic fusion methods using a deterministic version of our neural network: summing the predicted probabilities [219], counting the predicted classes as labels [218], and the Bayesian fusion [208] described in (4.2).

4.4.4 Results

Figure 4.4 (a) shows different examples of images in the real and simulated sequences. There, it is possible to see the difference between the GT semantic labels (b) and the output of the BNN (c), which motivates the need for better fusion models. Finally, we



(a) RGB Image (b) Ground Truth (c) BNN Prediction (d) Epistemic Unc.

Figure 4.4: **Qualitative results of the Bayesian semantic segmentation in ScanNet scenes (top) and in our simulated environment (down).** We find higher values of the epistemic uncertainty (in red) in the regions where the BNN fails in the prediction, showing its degree of confidence.

show the epistemic uncertainty associated with the predictions in (d). Notably, the red areas in (d), which correspond to pixels with high uncertainty, are associated in many cases with errors in the classification.

Simulated environments In Table 4.1, we present the quantitative results for the virtual environment. All baselines achieve comparable mIoU and accuracy, which is expected since the neural network accuracy will be similar. However, the fusion approach has a strong influence on the final map. The ablation rows show that, when applied individually, each term in our fusion (D and R) improves the quality of the semantic fusion, making them a valuable addition to the current state-of-the-art methods. Finally, the combination of D and R obtains the best results for the majority of classes.

Real environments The quantitative results on the real scenario of the ScanNet dataset are shown in Table 4.2. These results show a similar trend as those obtained

Method		Backgr.	Bottle	Chair	Table	Plant	Sofa	TV	mIoU	Acc.
Sum. Probs. [219]		89.2	25.4	30.7	12.5	18.0	43.7	6.4	32.3	89.1
Sum. Labels. [218]		89.9	27.5	32.3	15.1	19.9	47.6	7.2	34.2	89.5
Bayesian [208]		90.5	20.6	44.6	21.9	18.1	51.5	5.3	36.1	90.8
Robust Fusion (Ours)	R	91.1	26.9	**46.5**	**23.2**	20.1	58.2	9.9	39.4	91.4
	D	91.4	24.1	42.7	21.3	19.8	68.9	3.4	38.8	91.8
	D + R	**91.9**	**30.7**	45.5	22.4	**21.5**	**71.2**	**10.6**	**42.0**	**92.2**

Table 4.1: **Quantitative results on the virtual environment.** We report the IoU per class, the mIoU, and the mAcc to evaluate the quality of the semantic mapping. We aggregate the measurements from three trajectories in each of the two environments.

Method		Bed	Ceiling	Chair	Floor	Furniture	Objects	Picture	Sofa	Table	TV	Door	Window	mIoU	Acc.
Sum. Probs [219]		22.9	16.8	21.4	48.6	32.7	20.4	5.2	24.2	15.4	14.1	59.2	10.2	20.8	53.1
Sum. Labels [218]		27.5	19.6	22.3	51.1	32.0	20.2	5.8	23.6	15.3	**33.2**	59.7	11.2	22.9	53.8
Bayesian [208]		29.3	21.7	26.9	59.7	35.4	23.5	7.4	33.5	15.7	31.9	63.6	11.7	25.8	57.9
Robust Fusion (Ours)	R	30.9	22.0	27.5	60.1	35.9	23.6	7.4	34.7	15.8	28.9	63.6	12.4	25.9	58.0
	D	41.6	**28.2**	**33.3**	72.4	37.5	28.6	**23.9**	**44.5**	18.5	14.1	65.7	**32.2**	**30.1**	63.7
	R+D	**42.1**	**28.2**	32.8	**72.5**	**38.2**	**28.7**	**23.9**	43.8	**19.0**	12.6	**66.0**	31.4	30.0	**63.9**

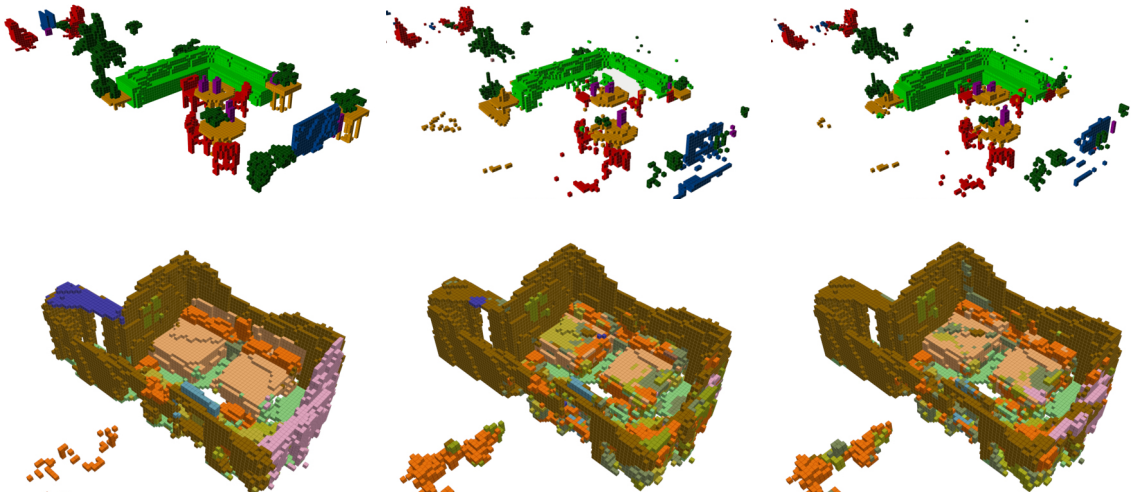
Table 4.2: **Quantitative results on the ScanNet dataset.** We report the IoU per class, the mIoU and the mAcc to evaluate the quality of the semantic mapping. We aggregate the measurements from three different scenes.

for the simulated environment, showing that our approach generalizes well to real scenarios. Notably, for the case of these scenes, the contribution of the epistemic term (D) is more significant than in the simulated case. We argue that this might be produced by a larger fraction of out-of-distribution samples than in the simulated environments.

Finally, we show qualitative results of the generated maps in Figure 4.5 for both configurations. Compared to the baseline, our mapping is less sensitive to outliers. For example, the sofa on the top example and the beds on the bottom are better segmented with less spurious classes. Overall, our approach shows promise for practical applications in real-world scenarios, demonstrating its potential to enhance the performance of semantic mapping in new environments, dealing with unknown data and outliers in a reliable manner.

4.5 Conclusion

In this chapter, we proposed a robust fusion method for semantic mapping that leverages Bayesian neural networks to consider the uncertainty of the network in the semantic fusion process. We achieved this by combining a regularization term, to miti-



(a) Ground-truth (b) Standard fusion (c) Robust fusion (ours)

Figure 4.5: **Qualitative examples of the experiments performed in virtual (top) and real (bottom) scenarios.** Voxels corresponding to the background class were removed for clarity. Our robust Bayesian fusion method is able to improve the mapping by reducing the influence of wrong measurements on the map.

gate the overconfidence in the predictions, together with a Dirichlet representation of the observations, using the epistemic uncertainty as a concentration parameter. Our method showed advantages over currently used methods when evaluated in both virtual photo-realistic and real environments, suggesting the importance of considering model uncertainties in tasks that require semantic understanding of the scene.

Chapter 5

DIV-FF: Dynamic Image-Video Feature Fields For Environment Understanding in Egocentric Videos

In the two previous chapters, I introduced environment representations based on explicit functions such as voxel maps and point clouds. However, these representations scale with scene size, capture sparse and limited information (only RGB color and the semantic class), and assume a static scene. In this chapter, I introduce Dynamic Image Video Feature Fields (DIV-FF), a language-embedded feature field that decomposes the scene into three components: the actor, the dynamic objects, and the persistent environment. This approach constitutes an implicit neural representation capable of jointly modeling geometry, appearance, and semantic understanding from egocentric video. While it provides a persistent, long-term representation that captures fine-grained affordance descriptions, it is also dynamically updated to maintain an accurate record of the location and state of dynamic objects over time.

5.1 Introduction

Egocentric videos offer a unique way to understand human activities from a first-person perspective, benefiting applications like mobile robotics, augmented reality and assistive devices. However, in egocentric videos an actor continuously moves to interact sporadically with multiple dynamic objects in a static scene, breaking the usual rigid scene assumption made in previous chapters. This tight integration between objects, actions and the dynamic scene introduces both opportunities and challenges for environmental understanding from egocentric videos.

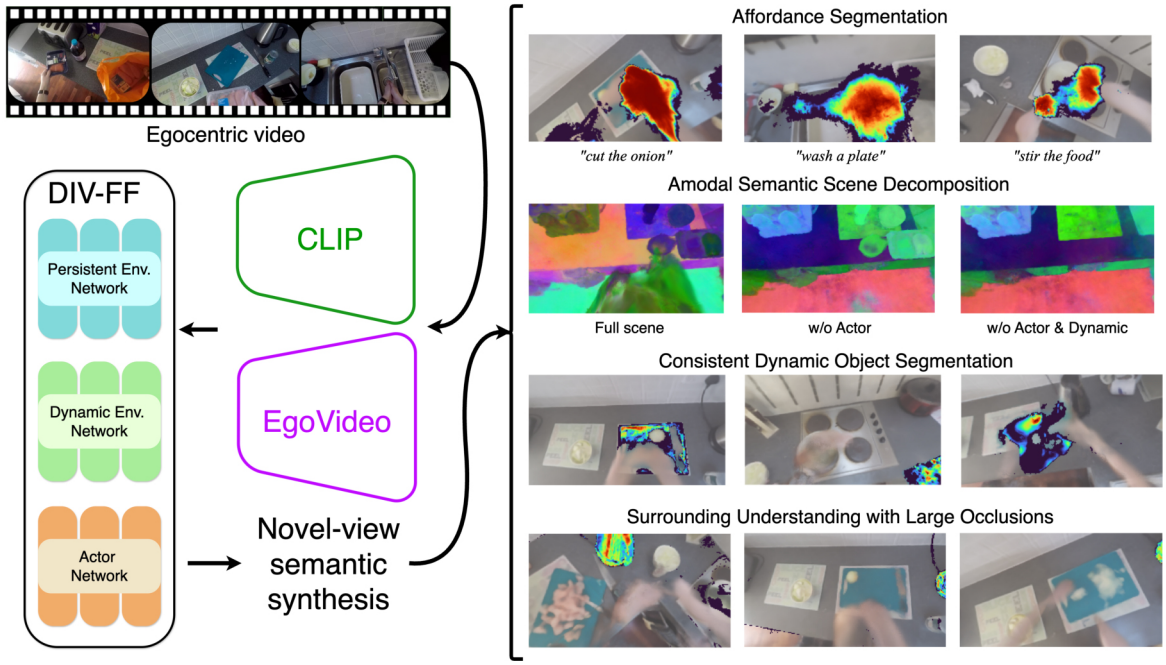


Figure 5.1: **Dynamic Image Video Feature Fields (DIV-FF) for egocentric videos.** DIV-FF distills image and video language features in a triple stream feature field tailored to egocentric videos with numerous interactions and camera wearer movements. This approach achieves a deep understanding of the environment, supporting precise affordance segmentation, semantic scene decomposition and consistent segmentation of dynamic objects. With its implicit 3D representation, DIV-FF comprehends not just novel views but also surrounding areas.

Most existing methods in egocentric environment understanding either consider a short video clip isolated from the physical space [28,29,230–232] or they provide a strong spatial representation but with low semantic understanding [65,66,181,233]. However, when humans interact repeatedly in a fixed environment, we develop a physical and semantic model that integrates the spatial distributions of the elements around us, both persistent and dynamic. The semantics capture detailed information about objects and their attributes through natural language descriptions. Additionally, we encode the available action (i.e.: affordance) locations in the environment, linking physical zones of interaction to the likely activities they support. Besides, this implicit semantic model is dynamically updated as the actor interacts, recording the location and state of dynamic objects at every moment. In that sense, some approaches propose intermediate representations between a pure semantic understanding of the video without explicit representation and a pure geometrical representation. They adopt semantic topological maps [39], local environment state representations [234] or explicit representations [38,75] for improving the environment understanding of the scene. In this chapter, we build an implicit (neural network) model that is able to jointly capture the geometry, appearance and semantic understanding encoded in the video, and enable predictions

in novel-view points using Neural Radiance Fields (NeRFs).

NeRFs provide a compact implicit representation of the geometry and visual appearance of a scene [235]. The implicit representation of NeRFs can also be used for semantic encoding, supporting multiple applications like robot manipulation [236,237], navigation [238], or scene editing [239,240]. For example, Neural Feature Fusion Fields (N3F) [241] extends the NeRFs predictive capabilities in a teacher-student fashion, where a teacher model that predicts semantic features in image space is used to train a NeRF-like student to predict semantic features in 3D space. These semantic capabilities are further extended in Language Embedded Radiance Fields (LERF) [242], enabling natural language query in 3D locations by volume rendering CLIP embeddings. However, LERF assumes a rigid scene which limits its applicability to egocentric videos where the actor is interacting with the environment. Furthermore, semantic distillation is based on single-image semantic features (e.g., CLIP features) which do not capture the dynamic nature of actions or changing elements.

In this chapter, we propose DIV-FF (Dynamic Image-Video Feature Fields[1]), the first language embedded feature field capable of decomposing both the geometry and the semantics of the scene for the actor, and also for the persistent and dynamic elements via three different streams. While previous works focus on image-language embeddings such as CLIP [243], DIV-FF also introduces video-language embeddings (based on EgoVideo [7]) to understand fine-grained action descriptions. This encodes the environment affordances, possible actions available in the environment for the actor, linking specific activities to physical zones where interactions are likely to occur. A parallel feature field, based on image-language features from CLIP, captures detailed information about objects and their attributes, categorizing them through natural language descriptions rather than fixed semantic tags, even from novel viewpoints. Its implicit representation, similar to NeRFs, ensures that even areas not visible from the egocentric camera remain strongly connected in the environment model. Although this environment model provides a persistent long-term representation, it is dynamically updated as the user interacts, enabling a precise record of the location and state of dynamic objects at every moment. The main contributions of this chapter are as follows:

- We introduce an approach to adapt language embedded feature fields to dynamic egocentric videos by dividing the radiance and feature fields depending on whether they are from the actor, dynamic, or persistent elements.

- We propose to distill video-language embeddings (from EgoVideo) to understand

[1]We use *embedding radiance fields* and *feature fields* interchangeably.

temporally dependent semantics, such as affordances (available actions), which single-image models like CLIP cannot capture.

– We present a robust image-language feature field enhanced by leveraging SAM masks, which also includes the temporal dependency and achieves a consistent segmentation of the dynamic objects over time.

– Our results demonstrate significant improvements in dynamic object (+40.5%) and affordance segmentation (+69.7 %) by using text query relevancy maps. Furthermore, this model effectively connects the egocentric view with the semantics of the surroundings and decomposes the scene into different levels.

This work was presented, and selected as Highlight, during the CVPR 2025 conference:

– Lorenzo Mur-Labadia, Jose J. Guerrero, and Ruben Martinez-Cantin. DIV-FF: Dynamic Image-Video Feature Fields for environment understanding in egocentric videos. Highlight in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* Core Ranking A*, 2025

5.2 Related works

Egocentric environment understanding using geometric representations. Some works that consider the physical layout build semantic explicit representations from videos of indoor scenes using visual SLAM systems. Rhinehart et al. [65] learn 2D maps with the functionality of different actions. Semantic MapNet [77] propose a birds-eye-view spatial memory for mapping, which is updated with recurrent neural networks to remember places visited in the past. Cartillier et al. [76] encode the egocentric frame, project its features, and then decode the semantic labels in a 2D map. Liu et al. [66] recognize and localize activities in an existing 3D voxel map from an egocentric video. The limitations in extracting the camera pose from egocentric video [244] due to the quick camera movements and motion blur have hampered the unification of 3D geometry and video understanding. Recently, the arrival of egocentric 3D datasets with camera poses [27,37,38,245] and the improvement of 3D sensors like project ARIA [246] has unlocked the arrival of novel works. Plizzari et al. [75] track active objects through their appearance and spatial consistency in the 3D scene, even when they are out of view. Mur-Labadia et al. [38] extract 2D affordance segmentation maps to build later a point cloud of the environment encoding those labels. Tschernezki et al. [78] proposed

a 3D-aware instance object tracking by keeping a long-term consistency. EgoLoc [40] extend episodic memory to 3D by estimating the relative 3D object pose to the user.

Egocentric environment understanding without geometric representations. Most egocentric video understanding works just consider a short time window of the video. Although these works obtain a remarkable semantic understanding in multiple tasks like action recognition [230, 231], object segmentation [36, 37], action forecasting [28,232] or capturing activity threads [247], these approaches ignore the underlying physical space of the scene. Some approaches [39, 234, 248] extract environment-aware features via alternative representations that avoid the geometric reconstruction problems from SLAM in egocentric videos [244]. Ego-Topo [39] builds a topological map, where the nodes represent environment zones with a coherent set of interactions linked by their spatial proximity. EgoEnv [234] encodes the relative directions of the objects to the camera wearer in a local state vector, learning an environment-aware video representation. Ramakrishnan et al. [248] capture the inherent statistics of indoor environments to learn an environment predictive coding, which applies later for navigation.

Dynamic Radiance Fields. Neural Radiance Fields (NeRFs) [240] allow capturing and rendering complex 3D scenes from a set of multi-view posed images. Using an implicit function and via differentiable volume rendering, NeRFs map spatial coordinates and viewing directions to colors and densities. Early methods for rendering dynamic scenes [249,250] use pre-trained motion segmentation methods to mask moving objects, guiding separate NeRFs networks to disentangle motion-based components. Liang et. al [251] leverage DINO features to identify salient foreground regions along spacetime, while Wu et al. [252] decouple moving objects from the static background in a self-supervised manner with two neural radiance fields. NeuralDiff [241] separates the static background, dynamic objects and the actor's body via inductive biases, obtaining a different implicit representation for each part of the scene. Recently, Zhang et al. [253] optimize 3D Gaussians to reconstruct the scene and track the 3D object motions from an egocentric video, but requiring pre-extracted hand-object interaction masks.

Feature Distillation in NeRFs. Several works extend radiance fields to integrate 2D semantic labels into the 3D space during the optimization [254–257]. In contrast, the objective of 3D feature distillation methods [239, 241, 258, 259] is to transfer 2D image features from a teacher model (i.e, a self-supervised model like DINO [74]) into a 3D student neural renderer. Expanding on this, 3D language feature fields distill image-text CLIP features [243], enabling querying the 3D student with open-vocabulary text descriptions to obtain relevancy maps. LERF [242] fuses multi-scale patch-level CLIP features conditioned on the scale. N2F2 [260] addresses the need for evaluating the rendering at the different scales by learning a unified feature field, where the different

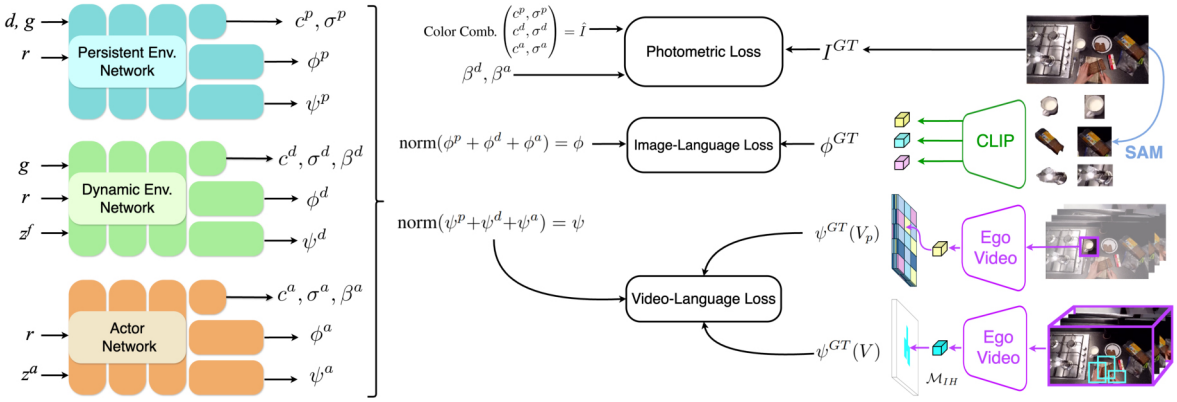


Figure 5.2: **Overview of DIV-FF.** Our three-stream architecture field predicts the color c, the density σ, the material aleatoric uncertainty β, the image-language features ϕ and the video-language features ψ along a ray r with direction d given the camera viewpoint g and a frame specific code z. We first extract SAM masks and bounding boxes from the image, that we leverage to obtain a unique CLIP descriptor ϕ_{GT} in all the pixels within the respective mask. We supervise the video-language feature field with local patch features $\psi^{GT}(V_p)$ and a global video embedding $\psi^{GT}(V)$ assigned only to pixels in the interaction hotspot $\mathcal{M}_{IH}$, computed with a pre-trained hand-object detector.

semantic granularities are encoded in a high-dimensional feature space. LangSplat [261] adopts 3D Gaussians [262] and combines CLIP features with multi-scale SAM masks, improving the segmentation quality. EgoLifter [263] augments 3D Gaussian Splatting with instance features from egocentric videos, but it only reconstructs the static part of the scene by filtering out the actor and the dynamic objects.

5.3 Methods

Our approach is to build a language embedding feature field that decomposes the 3D representation in three components (persistent environment, dynamic environment and actor) for accounting the inherent dynamics present in egocentric videos. In addition, we incorporate a second modality stream of embeddings based on video-language models which can capture the action semantics only present in the video modality. Besides, we introduce a time-dependent module on the dynamic and actor stream, capturing the temporal evolution of the feature fields.

5.3.1 Dynamic Neural Radiance Fields

The geometry model [264] captures the dynamic scene by integrating three different radiance fields, illustrated in Figure 8.2. The persistent environment network predicts the color c_k^p and density σ_k^p at each point along a ray r_k, given a viewing position g_t and unit-norm viewing direction d_t. Formally, it is defined as $(c_k^p, \sigma_k^p) = \text{MLP}^p(g_t r_k, d_t)$.

To model the dynamic objects in the scene, a second dynamic environment network $(c_k^d, \sigma_k^d, \beta_k^d) = \text{MLP}^d(g_t r_k, z_t^d)$ estimates the density σ_k^d and the color as a Gaussian distribution $\mathcal{N}(c_k^d, \beta_k^d)$, where β_k^d represents the heteroscedastic aleatoric uncertainty associated with the color. It also includes as input a frame-specific code z_t^d that accounts for temporal variations of the dynamic objects, which exhibit sporadic motion relative to the global reference frame. The actor network is similar to the dynamic environment network, but since the actor moves continuously linked to the camera, it removes the projection of the ray to the world coordinate system $(c_k^a, \sigma_k^a, \beta_k^a) = \text{MLP}^a(r_k, z_t^a)$. Here, z_t^a is a frame-specific parameter designed to capture the continuous motion of the actor. The predicted material uncertainty terms β_k^d, β_k^a indicate the confidence levels associated with each ray r_k for representing the dynamic objects and the actor, respectively. By employing improved color mixing techniques and inductive biases during training, the model accurately reconstructs scene dynamic geometry as a composite of the three radiance fields. For more details on the geometric model, please refer to [264].

5.3.2 Image-Language Feature Field

We extend the three-stream geometry model to distill image-language semantic features from CLIP [243]. Since the CLIP image encoder is a global image descriptor, it lacks pixel-aligned embeddings. To address this, LERF [242] extracts multi-scale patch-level features, which often fail to encompass the target object or add excessive contextual information. It results in blurred object boundaries and noise, requiring DINO [74] for regularization.

As shown in Figure 8.2, our model incorporates pixel-aligned CLIP features by leveraging accurate object masks generated by Segment Anything Model (SAM) [17] inspired by recent works [260, 261]. Specifically, we extract CLIP features per each segmented mask region $\phi_{\mathcal{M}}^{GT}$ and its respective bounding box $\phi_{\mathcal{B}}^{GT}$. We assign the same weighted descriptor $\phi^{GT} = 0.75 \cdot \phi_{\mathcal{M}}^{GT} + 0.25 \cdot \phi_{\mathcal{B}}^{GT}$ to all the pixels within the mask. This balanced approach achieves pixel-level alignment while preserving semantic context. Furthermore, the use of precise semantic masks with sharp object boundaries eliminates the need for DINO regularization used in previous works [242], leading to the following loss formulation:

$$\mathcal{L}_I = \left\| \hat{\phi} - \phi^{GT}(I_p) \right\|^2. \tag{5.1}$$

5.3.3 Video-Language Feature Field

While the CLIP image-language features contain fine-grained and accurate details of the objects, they ignore interaction semantics present in egocentric videos as they require temporal information. Therefore, we incorporate in parallel a video-language feature field to capture dynamic semantics, such as *affordances* and potential interactions. We leverage Video-Language Pre-trained (VLP) models [6, 7], which offer richer and action-oriented context by pairing narrative descriptions with video using contrastive learning. We select Ego-Video [7], the state-of-the-art in multiple Ego4D [26] challenges, for this task. Similar to CLIP, the video encoder of Ego-Video outputs a single descriptor from a video patch, not pixel-aligned features. In this case we cannot use object masks as in Section 5.3.2, because our goal is to identify *interaction hotspot* regions, including both the hands and the object parts (e.g. "knife edge", "spatula handle"), not just the entire object. While SAM's small masks could localize these areas, their limited size loses essential action context.

Therefore, we distill the video-language feature field with patch and global-level embeddings. We first pre-compute video descriptors $\psi^{GT}(V_p)$ from medium-sized video patches V_p, balancing fine-grained details with action context. Second, we derive a global descriptor $\psi^{GT}(V)$ for the entire video, assigned solely to the pixels within the interaction hotspot area $\mathcal{M}_{IH}$ [53]:

$$\mathcal{L}_V = \left\| \hat{\psi} - \psi^{GT}(V_p) \right\|^2 + \mathcal{M}_{IH} \left\| \hat{\psi} - \psi^{GT}(V) \right\|^2 . \tag{5.2}$$

This improves the feature field's capability to capture relevant interaction regions. We obtain the interaction hotspot mask $\mathcal{M}_{IH}$ as the union of the hands and active objects bounding box, pre-extracted with an existing hands-object detector [36]. Additionally, the training of this feature field is regularized with pixel-aligned DINO [74] features thanks to its object decomposition properties [242].

5.4 Experimental Settings

Implementation details. We extend the three stream architecture of NeuralDiff [264] by incorporating 4-layer, 256-width MLPs for the image ϕ and video language ψ feature fields, respectively. Both the coarse and fine models use 64 samples, while we select the best 32 samples for the feature distillation. The representations are summed and normalized post-rendering. We use an Adam optimizer with a learning rate of 5 $\times 10^{-4}$ and a cosine annealing scheduler. We train the geometry for 10 epochs with a batch size of 1024, then distill semantic features in two phases: training only the semantic heads for 5,000 iterations, followed by the full model for 3 epochs on an NVIDIA 4090.

Feature extractors. To extract the CLIP image embeddings, we utilize the OpenCLIP ViT-B/16 model [265] trained on the LAION-2D dataset following [242] for fair comparison. We prompt SAM [17] with a 32×32 grid, filtering redundant masks by 0.7 IoU, 0.85 stability score, and 0.7 overlap rate. We reduce the dimensionality of CLIP descriptors by training a scene-wise language auto-encoder [261] to reduce the memory cost. We sample 4 video frames at 60 fps with a temporal stride of 15. Local-patch video features are extracted using a patch size of 33% the image size and a stride factor of 0.5. For masking the global video descriptor, we employ a hand-object detector [36], specifically finetunned for egocentric sequences.

Dataset. We conduct our experiments on the EPIC-Diff subset [264] of the EPIC-Kitchens [25] dataset. On average, each sequence comprises of 900 calibrated frames, spanning 14 minutes of egocentric video, featuring multiple viewpoints and a large number of manipulated objects. Our evaluation encompasses both our method and the baselines on the test set, which includes frames not utilized during model training. This set facilitates assessments of new-view synthesis and segmentation capabilities.

Baselines. We compare against the following baselines:

- **LERF** [242] assumes a static scene, using a single stream for geometry and semantics. The image-language field is distilled from multi-scale patch-level CLIP features.

- **OWL** [266]. We apply this open-vocabulary object detector on the novel-view rendered images produced by Neural-Diff [264].

- **OWL+SAM** [17, 266] obtains the object's masks from the bounding box coordinates provided by the OWL baseline.

Ablations. We compare different versions of DIV-FF.

- **DIV-FF (CLIP in patches)** keeps the CLIP patch features from LERF $\phi^{GT} = \phi^{GT}_{\mathcal{P}}$, but it introduces the dynamic geometry model from [264].

- **DIV-FF (CLIP in SAM masks)** substitutes CLIP patch features by embeddings from SAM masks $\phi^{GT} = \phi^{GT}_{\mathcal{M}}$.

- **DIV-FF (full model, image inference)** incorporates the bounding box to obtain the CLIP descriptor $\phi^{GT} = 0.75 \cdot \phi^{GT}_{\mathcal{M}} + 0.25 \cdot \phi^{GT}_{\mathcal{B}}$, where $\phi^{GT}_{\mathcal{B}}$.

- **DIV-FF (full model, video inference).** In the full model, we render from the parallel video-language feature field.

5.5 Results

Once trained, our DIV-FF model predicts the color c, density σ, CLIP ϕ and EgoVideo ψ semantic features of a novel-view in an specific time-step and separates the actor and the dynamic elements from the persistent environment. We evaluate this comprehensive spatio-temporal semantic understanding in different downstream tasks.

5.5.1 Dynamic Object Segmentation

In each scene, we identify a subset of objects that move throughout the video and evaluate in the novel-views the relevancy maps originated by the text queries in the ϕ_{img} image-language feature field. Following the method proposed by LERF [242], we compute the relevancy score as: $\min_i \frac{\exp(\phi_{img} \cdot \phi_{quer})}{\exp(\phi_{img} \cdot \phi_{quer}) + \exp(\phi_{img} \cdot \phi^i_{can})}$. This formula evaluates how closely the rendered embedding ϕ_{img} matches the query embedding ϕ_{quer} compared to a set of predefined canonical phrases ϕ^i_{can} (*"object", "thing", "stuff", "texture", "hands"*). Segmentation masks are generated for relevance scores above a specified threshold. For the evaluation, we leverage existing annotations from [241] and report the mean intersection over union (mIoU). We visualize the text query relevancy maps by normalizing from 50 % to the maximun relevancy.

Table 5.1 presents quantitative results on EPIC-Diff scenes. The full version of DIV-FF achieves the best performance (30.5 mIoU), surpassing the OWL+SAM detector (21.7 mIoU) by +40.5%, illustrating that distilling semantic features outperforms traditional open vocabulary object detection from novel views, since the OWL model fails due to artifacts and the blurry hand effects in the novel view rendered. The CLIP patch-level version of DIV-FF (19.8 mIoU) significantly improves upon LERF by explicitly considering the dynamics parts in the semantic and geometric fields with the triple stream architecture of DIV-FF. This leads to sharper reconstructions, particularly for moving objects as shown in Figure 5.4. Subsequently, leveraging SAM to extract object-level CLIP features further improves performance (26.2 mIoU), and generates more accurate and consistent semantic renderings compared to CLIP patch-level embeddings. Finally, the introduction of contextual information from the object bounding boxes ultimately yields to the best performance (30.5 mIoU).

Figure 5.3 showcases novel-view renderings for various text queries in two scenes, effectively capturing fine-grained details like the *"countertop"* borders. The uniform assignation of the same CLIP descriptor across all object pixels allows DIV-FF to segment objects of any size, such as *"sink"* or *"banana"*. As Figure 5.5 shows, our approach also segments consistently dynamic objects across multiple novel views in different time-steps of the sequence due to the combined impact of object-level CLIP

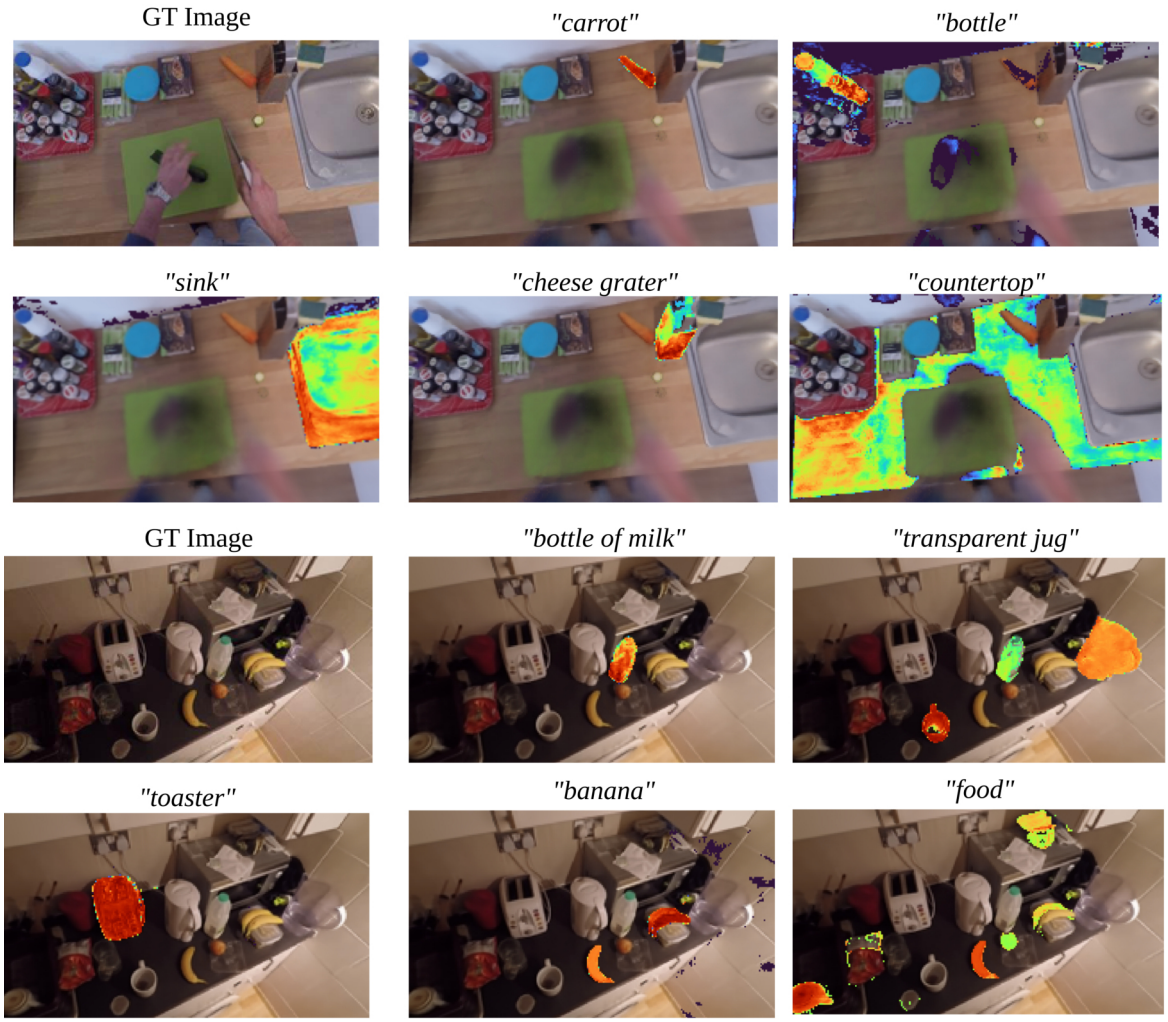


Figure 5.3: **DIV-FF Image-Language relevancy maps in novel-views.** We can see the performance of various text queries for dynamic object segmentation. We can see how the object contours are well defined as we used masks during training.

features and the temporal encoding in the frame-specific codes. Unlike egocentric methods limited to short time windows, our environment understanding extends beyond the current view to the surrounding regions. Figure 5.7 illustrates this capability, showing how our 3D semantic implicit model segments the *"pot"* and *"sink"*, despite being almost occluded in the edge of the image.

5.5.2 Affordance Segmentation

We identify affordable actions in each scene and generate Ego-Video [7] text queries ψ_{quer} describing the interaction, which are more complex than simple object labels as they capture nuanced action dynamics. We compute the relevancy score from the video-text feature field ψ against a different set of canonical phrases ψ^i_{can} (*"general task", "indistinct movement", "unclear action", "background"*). We manually annotate

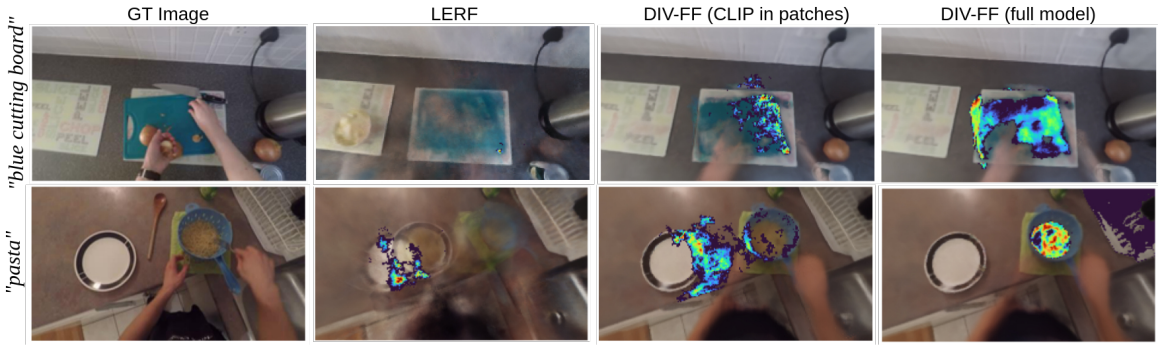


Figure 5.4: **Ablations on the image-language feature field.** Treating the egocentric video as a dynamic scene enhances geometric reconstruction, while utilizing SAM masks further improves object segmentation accuracy.

Method	S01	S02	S03	S04	S05	S06	S07	S08	S09	S10	Average mIoU
LERF	22.1	10.1	11.7	9.7	13.2	18.6	12.5	6.2	19.0	5.5	12.8
NeuralDiff + OWL-ViT	9.4	10.2	13.2	15.4	9.4	13.3	14.5	23.2	23.7	28.9	16.1
NeuralDiff + OWL-ViT + SAM	8.7	12.6	23.2	23.9	13.8	15.9	17.8	**28.0**	**32.9**	**41.1**	<u>21.7</u>
DIV-FF (CLIP in patches)	26.9	21.7	18.3	16.8	18.1	24.9	17.3	12.3	17.9	23.6	19.8
DIV-FF (CLIP in SAM masks)	30.7	19.3	29.6	24.9	**31.3**	26.1	28.8	14.8	23.8	35.1	26.2
DIV-FF video infer.	16.1	15.4	9.3	9.5	21.8	20.7	10.7	18.5	17.9	20.6	16.6
DIV-FF image infer.	**40.3**	**30.4**	**37.4**	**29.8**	29.5	**32.6**	**30.6**	15.1	25.1	33.6	**30.5** (+40.5%)

Table 5.1: **Dynamic Object Segmentation by CLIP image-language feature field.** Compared with LERF, DIV-FF considers a dynamic scene in the geometric reconstruction. Our full model assigns the same descriptor to all the pixels within a SAM mask. This descriptor is a weighted average between the CLIP of the mask and the bounding box.

affordance segmentation masks for five affordable actions per sequence, resulting ≈ 700 masks. We report mIoU.

Table 5.2 demonstrates the effectiveness of video-language features in capturing actions. Previous methods that rely on single-image CLIP features miss the dynamic action context in egocentric videos. Consequently, the video-language feature field of DIV-FF excels in the affordance segmentation, achieving 20.7 mIoU (+69.7 %), benefiting from training on video narrations, unlike CLIP models that use static image captions. We visualize these differences in Figure 5.6, showing the relevancy scores for text queries detailing specific actions. The image-language model performs well when actions are explicitly linked to objects, such as *"cut the onion"* or *"add ingredients to the mixture"*.

However, it struggles with verbs or semantic contexts that imply a location, like *"wash a kitchen utensil"*—which suggests the sink—or *"toast the bread"*, associated

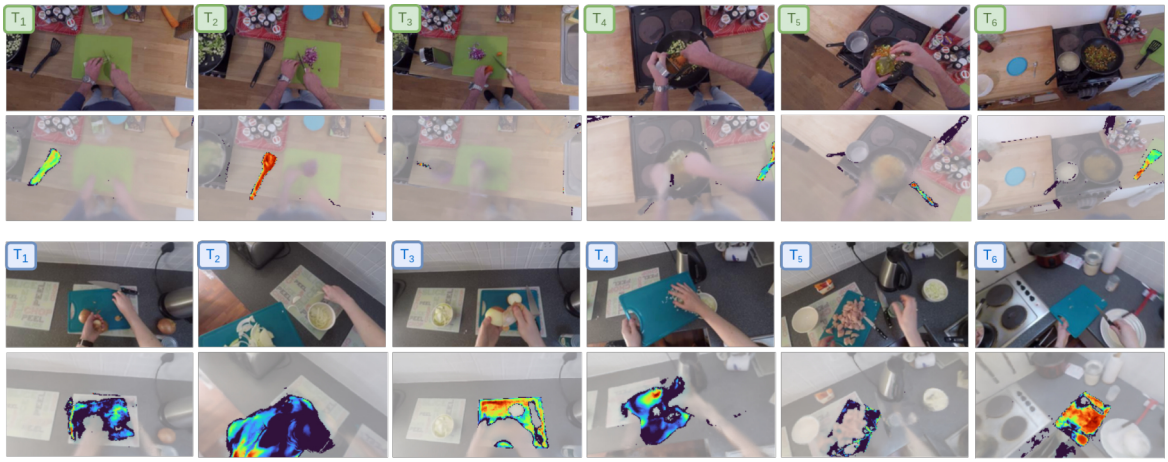


Figure 5.5: **Consistent Dynamic Object Segmentation along different time-steps in novel views:** The dynamic and actor streams contain respective frame-specific codes z_t^f andz_t^a. This time encoding is also propagated to the semantic feature field, obtaining consistent segmentations despite the continuous movement of the *"spatula"* and *"blue cutting board"*.

Method	S01	S02	S03	S04	S05	S06	S07	S08	S09	S10	Average mIoU
OWL-ViT	4.8	4.2	1.4	5.6	4.8	2.3	13.2	4.0	5.4	4.4	5.0
OWL-ViT + SAM	4.6	5.4	1.9	4.6	5.8	1.1	8.6	4.5	7.7	8.3	5.3
LERF	18.2	17.4	6.8	11.5	11.9	18.4	11.7	7.5	15.2	4.2	12.2
DIV-FF (CLIP in patches)	17.1	15.6	7.1	9.4	12.9	19.7	12.4	11.3	15.3	12.6	13.3
DIV-FF (full m., image infer.)	17.3	13.7	6.2	13.7	19.1	8.1	18.5	7.1	11.1	3.6	11.8
DIV-FF (full m., video infer.)	**20.6**	**19.9**	**14.4**	**22.4**	**30.1**	**22.3**	**20.1**	**16.8**	**17.1**	**23.1**	**20.7** (+69.7%)

Table 5.2: **Affordance Segmentation.** We compare the segmentation masks of a set of affordable actions in the scene. The full version of DIV-FF is composed by two parallel semantic feature fields, image (CLIP + SAM + boxes descriptors) and video (Ego-Video) respectively.

with the toaster. In these instances, the video-language model distinctly outperforms, accurately identifying the action's interaction hotspot. We also highlight that the localization of these fine-grained areas is due to the additional global supervision to the local medium-size Ego-Video patch features, as Figure 5.9 shows. The joined effect of the two losses improves the relevancy maps by explicitly guiding the optimization toward the interaction hotspot regions. Table 5.2 also shows that for single-image models, patch-based methods (LERF, CLIP in patches) outperform the full model using object masks, as we suggested in Section 5.3.3.

5.5.3 Amodal Scene Understanding.

Our DIV-FF model comprises three distinct levels of geometry and semantics, representing different scene levels as illustrated in Figure 5.8. This introduces significant versatility in the environment understanding. For example, we can remove the actor's hands to reveal the dynamic objects without occlusions. Additionally, eliminating both

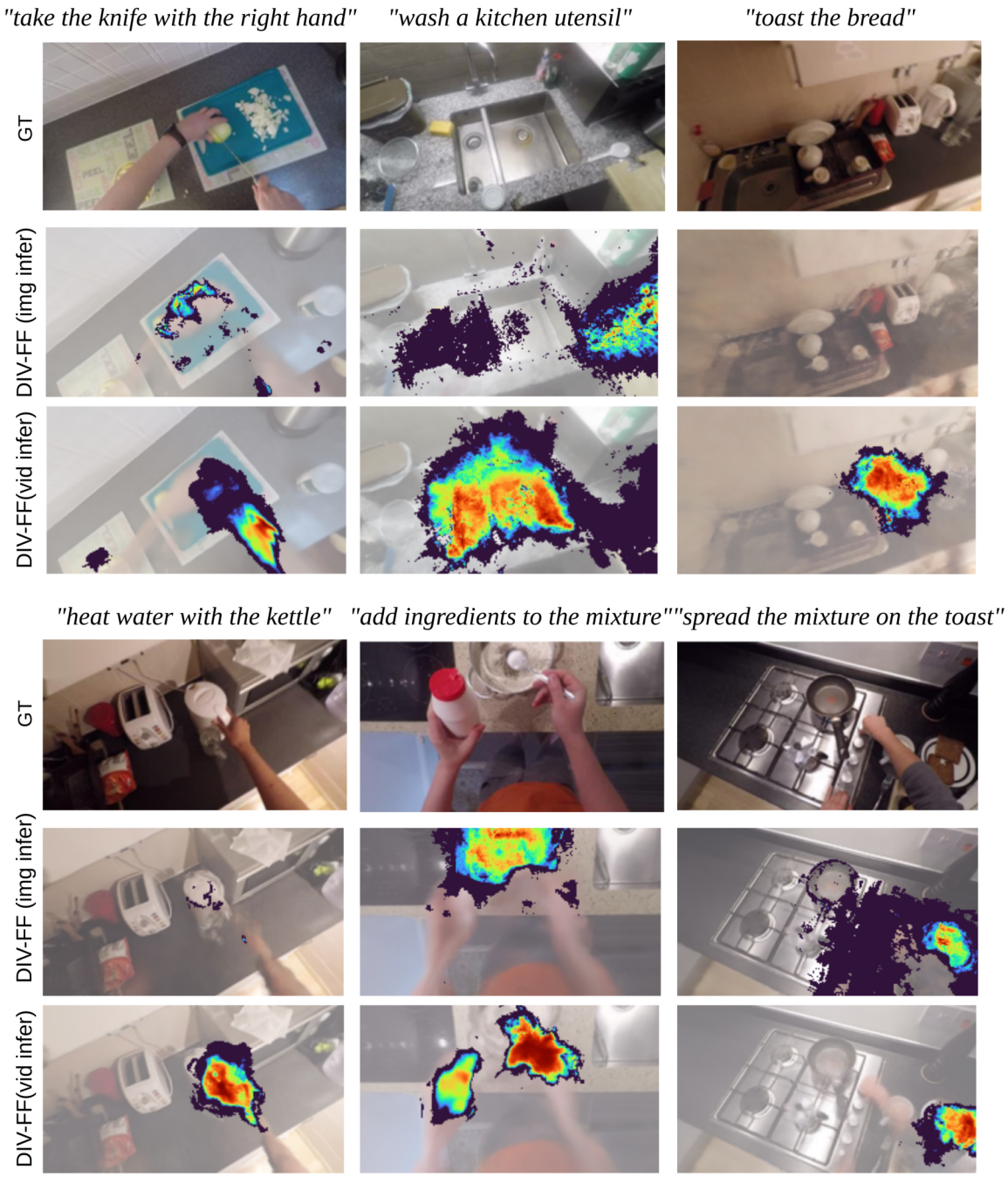


Figure 5.6: **Affordance Segmentation qualitative examples.** We compare the relevancy maps produced by the image-language field against those from the video-language field of DIV-FF, based on a detailed action description text query.

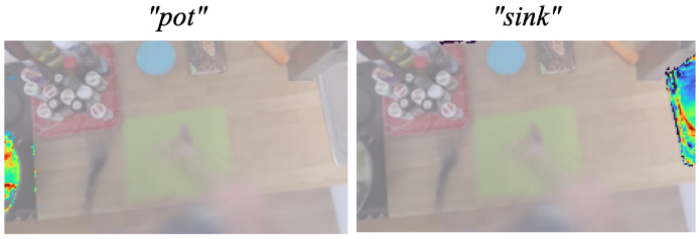


Figure 5.7: **Surrounding Understanding.** DIV-FF understands the novel view and the surrounding environment, enabling segmentation of objects at the image's edges with limited observability.

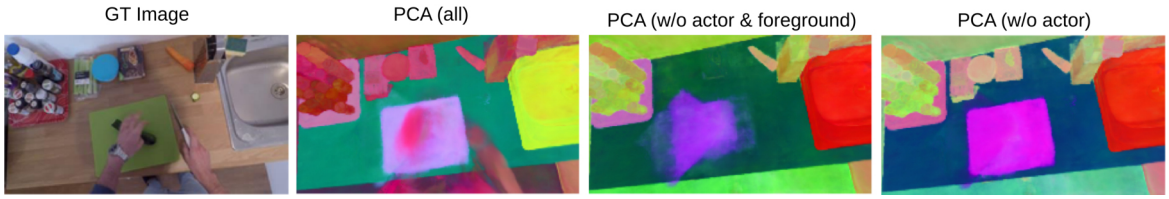


Figure 5.8: **Amodal Scene Understanding.** We visualize the PCA components obtained from the different composition of the image-text feature fields, showing accurate decomposition of the objects contours due to the SAM masks regularizing effect.

the actor and dynamic elements exposes only the persistent parts of the scene. Our intuition is that this static spatial-semantic representation contains strong priors that can be exploited when the user revisits the scene at another time.

5.6 Limitations

The image-language field of DIV-FF inherits several limitations from SAM, notably in the excessive segmentation of objects that omits some of its parts. This is evident in cases such as the *"cup"* in P04-01, *"plate"* in P13-03 and *"sink"* in P21-01 examples of Figure 5.10. The segmentation produced by SAM either omits some parts of the objects or introduces artifacts such as holes. Another limitation of DIV-FF is the degradation associated to the geometry, specially when rendering the actor's hands (scenes P01-01 and P13-03 in Figure 5.10). The actor's hands continuous movement and the biased top-view (egocentric) perspective of all the images pose a significant challenge in accurately rendering the hands in novel views, which is later reflected in the relevancy maps for the *"hands"* text query. Despite the inclusion of persistent, dynamic and actor streams in DIV-FF to enhance the capture of egocentric video, the rendering quality of objects in contact with the actor, such as the *"green cutting board"* in P01-01 or *"pasta"* in P13-03 in Figure 5.10, is compromised. This degradation is primarily due to frequent occlusions by the actor's hands, disrupting the continuity of views

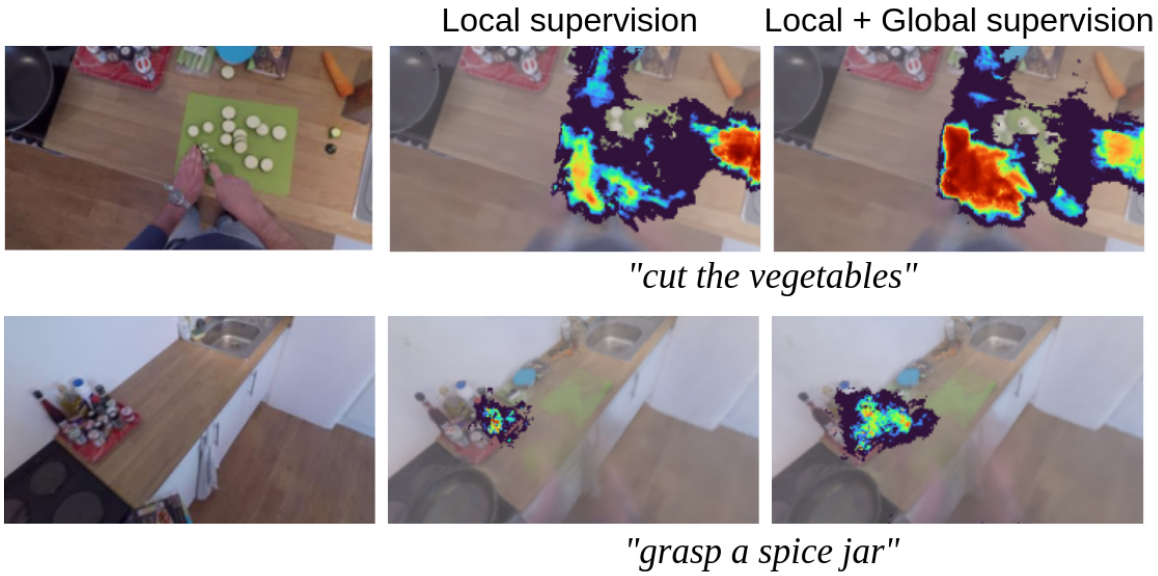


Figure 5.9: **Video-Language Loss ablation.** Including the global supervision term in the interaction hotspot mask produces sharper relevancy maps compared to just having the patch-level (local) term of the loss.

The main limitation in the video-language feature field of DIV-FF is the rendering of diffuse relevancy maps, which introduce excessive noise (*"pour soap on the sponge"* in P01-01 A or *"control the stove"* in P09-02 A, both in Figure 5.11) in some cases.

5.7 Conclusions

We proposed Dynamic Image-Video Feature Fields (DIV-FF) to address the limitations of existing egocentric video understanding methods. By decoupling the scene into persistent, dynamic, and actor streams and integrating video-based semantics, our approach achieves robust and consistent semantic segmentation over time. The model's ability to perceive and reason about both persistent and evolving scene elements marks a significant improvement in affordance and dynamic object understanding. Experimental results highlight DIV-FF's effectiveness in representing the rich and dynamic nature of egocentric environments, setting a promising direction for future work in spatial-temporal scene modeling and interaction-aware perception.

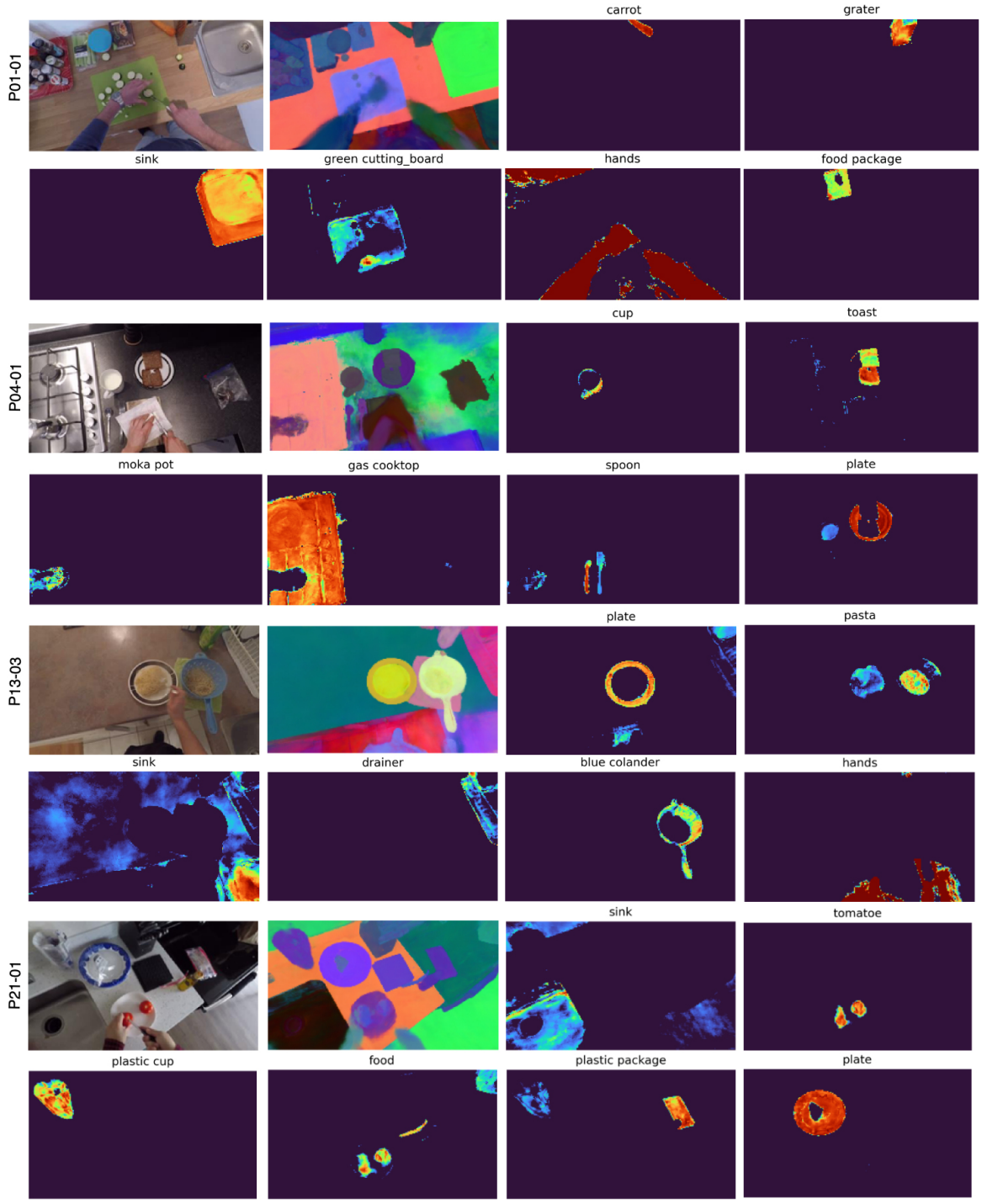


Figure 5.10: **Additional results of the DIV-FF Image Language relevancy map in novel views.** We visualize the ground-truth image, the PCA of the image-language features and different relevancy maps for different text queries.

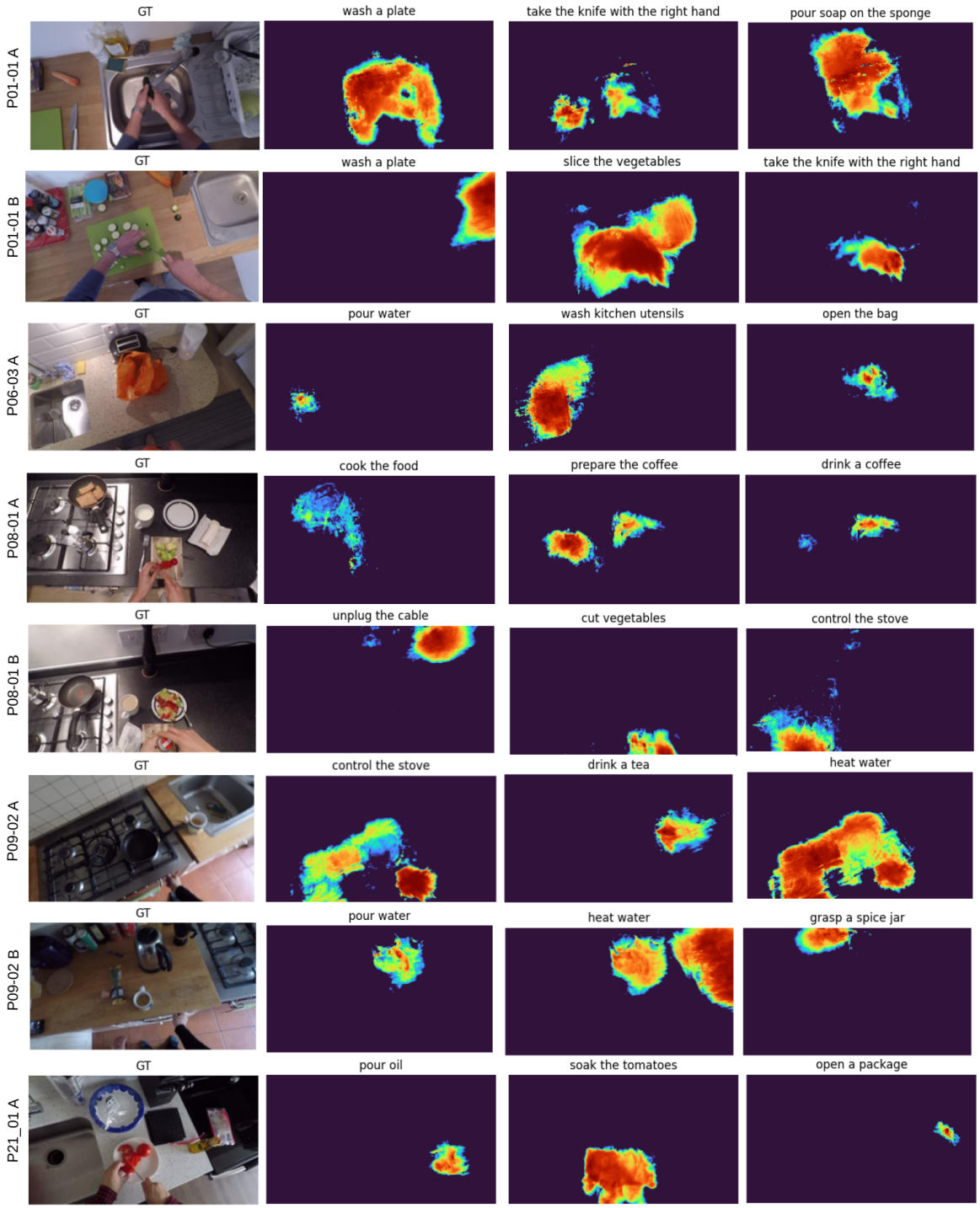


Figure 5.11: **Additional results of the DIV-FF Image Language relevancy map in novel views.** We visualize the ground-truth image and three different relevancy maps of the video-language feature field corresponding of affordable interactions.

Part III

Forecasting the short-term object interaction

Chapter 6

Integrating Affordances and Attention models for Short-Term Object Interaction Anticipation

While in the previous chapters I focused on understanding the present or encoding the perceived environment, in this chapter I introduce methods to anticipate the next object interaction. My first contribution is STAformer and STAformer++, two novel end-to-end architectures specifically designed for the short-term anticipation task. The STAformer prediction head is based on Faster R-CNN [84], a proposal-based convolutional object detection architecture. In contrast, STAformer++ is a more advanced version that leverages Detection Transformers (DETR) [85] to enhance detection performance in an end-to-end manner. While previous affordance models primarily improve perception, in this chapter I investigate how to leverage affordances as strong priors that encode future interactions. Specifically, I propose refining the verb and noun probabilities from environment affordances, which are extracted by matching the input observation with a set of activity-centric zones. Additionally, I re-weight the predicted confidence scores according to their location within the interaction hotspot.

6.1 Introduction

Anticipating the future is a fundamental ability for assistive egocentric devices and to support human-robot interaction. For example, a smart wearable device could alert an electrical operator before they short-circuit a switchboard, or a home robot can support the user by turning on appliances or moving objects according to their forecasted long-term goal. Predicting the future state of the scene from egocentric visual observations is a growing research area [267, 268], with works tackling action

anticipation [29,269–274], locomotion prediction [30,31,233,275,276], hands trajectory forecasting [67, 68, 277], and next-active object detection [32, 33, 80, 278]. Recently, Grauman et al. [26] defined the Short-Term object interaction Anticipation (STA) task as the simultaneous prediction of the action and object category, the object's bounding box, and the time to contact, and introduced an international challenge within the forecasting benchmark of the Ego4D dataset. Inspired by this challenge, the community proposed different approaches [279–285].

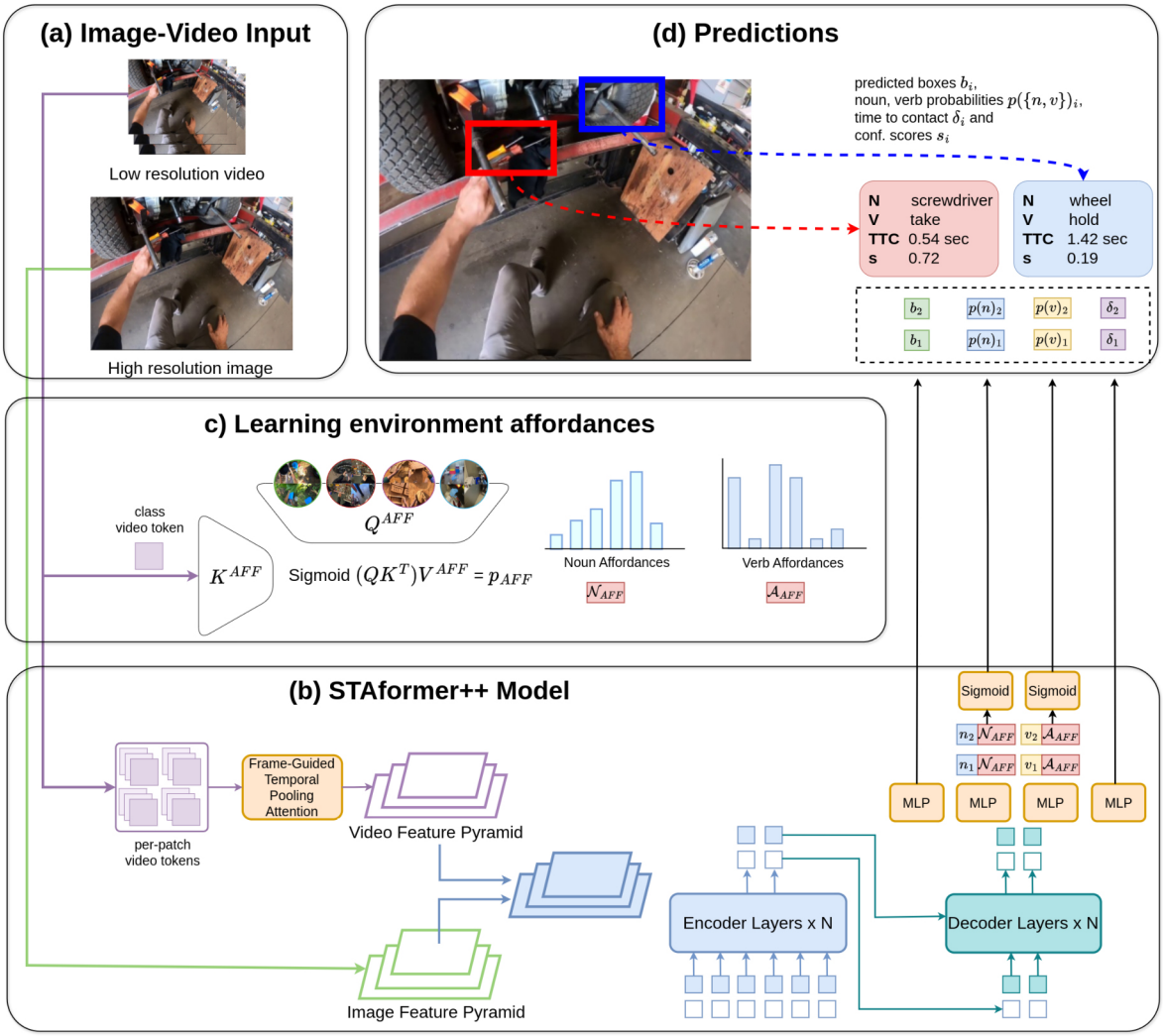


Figure 6.1: **Short-Term Object Interaction Anticipation.**(a) Our approach takes as input an image-video pair. (b) The input is processed by the novel STAformer++, an end-to-end STA model based on transformers which predicts object bounding boxes, the associated verb/noun probabilities, time-to-contact estimates and confidence scores. (c) In a dynamically and flexible way during training, the model grounds the predictions on the environment affordances. (e) The final predictions.

The aim of this chapter is to advance research in STA in two main directions. First, we propose a new architectural design based on transformers to provide a principled and modern end-to-end architecture for STA which can be easily extended. Specifically, we

introduce STAformer, which combines a transformer backbone with a Faster R-CNN detection head, and STAformer++, which improves upon STAformer by including a novel transformer-based detection head adapted from DETR [85] for the STA task. Differently from previous approaches [26, 281, 284], these two architectures operate on an image-video input pairs, introducing novel attention-based components for image-video fusion, such as a per-scale frame-guided temporal pooling and dual-cross attention fusion. Besides, these methods leverage the modeling capacity of state-of-the-art feature extractors such as DINOv2 [16], Swin-T [140], EgoVideo [7] and TimeSformer [286].

Second, to tackle the challenges associated with relating past visual observations to future events from video, we propose to ground predictions into human behavior by modeling environment affordances. In this chapter, we refer to environment affordances as the possible interactions that the agent can perform in a given environment. As highlighted in recent studies [79], human activities exhibit consistency in similar environments. Our intuition is that linking a novel video across similar environments captures a description of the feasible interactions, grounding predictions into previously observed human behavior. We hence propose to leverage a precomputed distribution of environment affordances. By matching the input observation with the affordance database, we obtain the noun and verb affordance probabilities. During inference, these affordance distributions are used to refine the predicted verb and noun probabilities. In a more advanced version, we integrate affordance information during training. An attention mechanism links a new video to all relevant candidates in the affordance database, enabling a more flexible approach that avoids selecting a fixed number of database members to construct the distribution. Finally, we predict interaction hotspots [67] to re-weigh confidence scores of STA predictions depending on the object's locations, linking predictions to spatial priors for interactions in the current frame. In sum, our contributions in this chapter are:

- We introduce STAformer and STAformer++, two principled architectures for STA based on transformers. The two architectures have been designed to process an input image-video pair and their implementations are open-sourced to support further research.

- We propose different approaches to ground STA predictions in human behavior by modeling environments affordances. This is achieved with two different methods: late-fusing pre-learned affordances in a static way, and learning to integrate the predicted affordances during the training of the end-to-end STA model.

- We further investigate the integration of interaction hotspots to refine the bounding box scores, benefiting predictions close to the interaction hotspots.

– We contribute a novel set of STA annotations, curated from public EPIC-Kitchens labels. This effectively provides the research community with a second large-scale and challenging benchmark for the STA task, besides the popular Ego4D.

– Experiments on Ego4D [26] and EPIC-Kitchens [25] highlight the effectiveness of the proposed approach, obtaining state-of-the-art results in the validation splits of Ego4D [26] and in a novel set of curated STA annotations on the EPIC-Kitchens dataset [25].

– We report extensive ablation studies, an analysis of the influence of the amount of video seen and a comparative of our affordance against naive priors, highlighting the challenges of this task.

A seminal part of this work was presented in ECCV 2024, and a follow-up version is currently under review at T-PAMI journal. Besides, the STAformer architecture achieved the 2^{nd} position at the Ego4D STA Challenge during the EgoVIS Workshop at CVPR 2024.

– Lorenzo Mur-Labadia, Ruben Martinez-Cantin, Jose J Guerrero, Giovanni Maria Farinella, and Antonino Furnari. ZARRIO@ Ego4D Short Term Object Interaction Anticipation Challenge: Leveraging Affordances and Attention-based models for STA. 2^{nd} Position at Ego4D STA Challenge during EgoVIS Workshop CVPR 2024. *arXiv preprint arXiv:2407.04369*, 2024.

– Lorenzo Mur-Labadia, Ruben Martinez-Cantin, Jose J Guerrero, Giovanni Maria Farinella, and Antonino Furnari. Aff-ttention! affordances and attention models for short-term object interaction anticipation. In *European Conference on Computer Vision.* Core Ranking A*, pages 167–184. Springer, 2024

– Lorenzo Mur-Labadia, Ruben Martinez-Cantin, Jose J Guerrero, Giovanni Maria Farinella, and Antonino Furnari. Integrating Affordances and Attention models for Short-Term Object Interaction Anticipation. In *IEEE Transactions on Pattern Analysis and Machine Intelligence.* (Under Review)

6.2 Related works

In this section we review the key advancements in short-term object interaction anticipation, placing it within the broader context of video forecasting. We also discuss the role of affordances for anticipation, and explore the evolution of object detection architectures from convolutional to transformer-based models.

6.2.1 Short-term Object Interaction Anticipation

Furnari et al. [32] initially introduced the concept of Next-Active Objects (NAO), proposing to detect future interacted objects by analyzing their trajectories as observed from the first-person point of view. Differently from action anticipation [25], the NAO detection task is designed to provide grounded predictions in the form of bounding boxes, which can be particularly informative for wearable AI assistants or embodied robotic agents. Unlike traditional object detection [287], NAO prediction requires the ability to model the dynamics of the scene and anticipate the user's intention. Jiang et al. [80] developed a method to predict the next-active object location in the form of a Gaussian heatmap from a single RGB image, combining visual attention with probabilistic maps of hand locations. Ego-OMG [278] segments the NAO and predicts the interaction time using a contact anticipation map that captures scene dynamics. While previous works considered different task formulations and evaluation approaches, Grauman et al. [26] formalized NAO prediction by introducing the STA task and an associated challenge on the EGO4D dataset [26]. The initial baseline is composed of a Faster R-CNN branch to detect objects [287] and a SlowFast 3D CNN [288] for video processing. Subsequent research introduced architectural enhancements and alternative approaches. Chen et al. [279] employed pre-computed object detections using a DETR model and substituted SlowFast with a VideoMAE pre-trained ViT [280]. Pasca et al. [281] proposed TransFusion, which employs a language encoder for action context summary, performing multi-modal fusion with visual features. While previous works leveraged pre-extracted object detections for 2D image understanding, Ragusa et al. [282] introduced StillFast, an end-to-end framework unifying the processing of 2D images and video in a combined backbone. Thakur et al. [283] proposed GANO, an end-to-end model based on a transformer architecture including a novel guided attention mechanism. Guided attention was integrated within a StillFast architecture in [284], achieving state-of-the-art results. Thakur et al. [285] introduced NAOGAT, a multi-modal transformer that attends detected objects and includes a motion decoder to track object trajectories. Recently, a video-language foundation model denominated EgoVideo [7] achieved state-of-the-art performance in the STA task. The authors selected 7M ego video-text clips from multiple datasets and trained the model with standard video-text contrastive learning. The video encoder was then finetunned to the STA task using the StillFast [282] prediction head. Compared with previous works, we propose a novel architecture that fuses the image-video pair with attention-based components and that integrates affordances for refining the predictions.

6.2.2 Affordances for Anticipation

The computational perception of affordances has been investigated in different forms. A line of works predicts affordance labels of object parts, requiring strong supervision in the form of manually annotated masks [54,55,58,289]. However, these methods are not "grounded" in human behavior as the annotator declares interaction regions outside of any interaction context [53]. Other works considered the problem of grounding affordance regions in images by leveraging videos depicting human-object interactions in a weakly supervised way, where only the action label is used as supervision without spatial annotations [53,61–63]. Nagarajan et al. [53] introduced the concept of "interaction hotspots" as the potential spatial regions where the action can occur. Mur-Labadia et al. [38] create a 3D multi-label mapping of affordances extracted from egocentric video. Another line of work infers interaction hotspots from video by forecasting future hand movements to select candidate regions for future interactions [62,67,68,80]. Few works studied scene affordances to predict a list of likely actions that can be performed in a given scene [39, 65]. In particular, Nagarajan et al. [39] proposed Ego-Topo, a procedure to decompose a set of egocentric videos into a topological map encoding scene affordances. Despite the interest in affordances, only a few works investigated how to exploit them for future predictions. Montesano et al. [51] predicted affordance effects for human-robot interaction. Koppula et al. [47] used object affordances to anticipate human behavior in the form of motion trajectories of objects and humans. Nagarajan et al. [39] showed how scene affordances learned from egocentric video can improve long-term action anticipation. Liu et al. [68] tackled action anticipation by jointly predicting egocentric hand motion, interaction hotspots, and future actions. Liu et al. [67] highlighted how interaction hotspots predicted by forecasting hand motion can support action anticipation. In this chapter we integrate affordances in an unified architecture for the short therm anticipation task by the first time. In accordance to literature, we show that affordances are beneficial for performance due to its generalization capabilities. Moreover, we study how to use then during training time.

6.2.3 Object Detection Architectures

Object detectors based on convolutional networks are categorized as either two-stage or one-stage models, relying on hand-crafted anchors or reference points for object localization, respectively. Two-stage detectors [12, 290, 291] involve a Region Proposal Network (RPN) that generates boxes candidates that are subsequently refined. Faster-RCNN [290] applies a Region of Interest (RoI) alignment and a set of linear layers for accurate prediction of bounding boxes and semantic class for object detection. One-

stage approaches, such as YOLO [292] directly predict offset from predefined anchors without the proposal stage, notably reducing the inference time. However, convolutional models still require manual components like Non-Maximun Suppression (NMS) to eliminate redundant boxes and rely heavily on anchor generation methods, affecting overall performance.

These limitations were solved by the arrival of the DEtection TRansformer (DETR) [85], an end-to-end transformer-based architecture for object detection. DETR introduces the concept of "object queries", a fixed number of learned embeddings decoded to predict objects in an image, eliminating the need for hand-crafted components. During training, these queries interact with the image encoded features through cross-attention in the transformer decoder. Since each object query ultimately corresponds to a potential detected object, DETR applies simple linear layers to predict the bounding boxes and the class labels for each object. However, the instability in the the Hungarian algorithm for matching the targets with the object queries, the lack of inductive biases like anchor boxes and the global attention mechanism make very slow the convergence of DETR. Deformable DETR [293] focuses on selecting a set of sampling points and applying a deformable attention that attends to a small set of points around the sampled point, improving both the efficiency and accuracy of the model. DAB-DETR [294] formulates the positional part of the decoder queries as dynamic 4D anchor box coordinates (x, y, h, w), which provides a reference query point (x, y) and a reference anchor size (w, h) that simplifies the refinement process. DN-DETR [295] introduces a DeNoising (DN) training strategy that accelerates the DETR convergence by solving the instability of the bipartite matching. It feeds noisy ground truth samples into the decoder and trains to recover the original, uncorrupted data with an additional denoising loss. DINO-DETR [296] combines DAB-DETR and DN-DETR with the deformable attention for its computational efficiency, including a contrastive denoising training, a mixed query selection and a novel "look forward twice" scheme, achieving significant improvements both in accuracy and convergence. In this chapter, we benchmark both convolutional and transformer based heads with DINO-v2 [16] and Swin Transformer (Swin-T) [140] features. We also highlight the importance of the video encoder for modeling the action dynamics, and the importance of the intermediate components for fusing both modalities in order to obtain the better predictions.

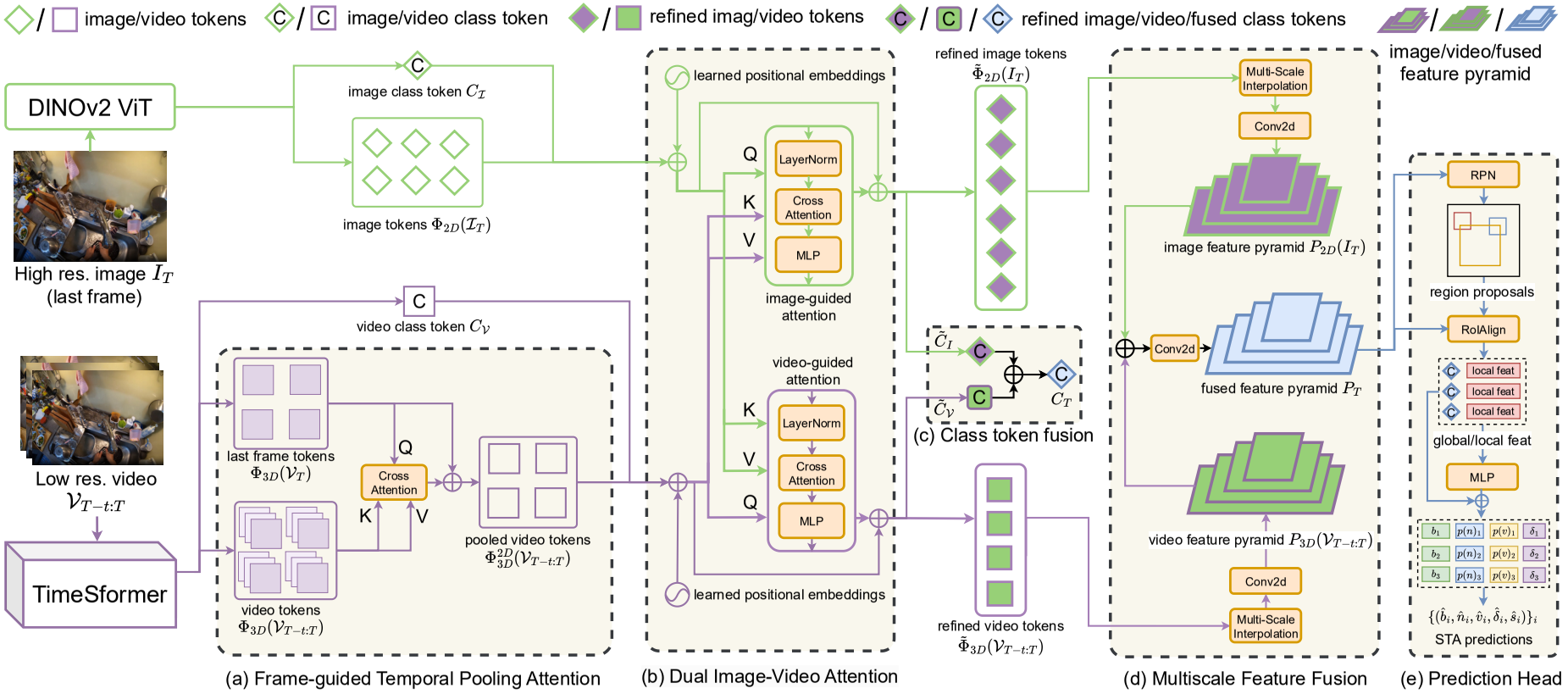


Figure 6.2: **STAformer architecture.** DINO-v2 and TimeSformer extract 2D and 3D features form the image-video input. (a) Frame-guided temporal pooling attention spatially aligns video to image features. (b) Dual image-video attention enriches 2D features with temporal dynamics and 3D features with fine-grained image details. Image and video representations are joined to obtain a global class token (c) and a feature pyramid (d), from which we obtain the STA predictions (e).

6.3 STAformer: a Transformer-based Architecture for Short-Term Anticipation

As defined in [26], the goal of Short-Term object interaction Anticipation (STA) is to detect the next-active object from the observation of the image frame at time T, $I_T \in \mathbb{R}^{h_s \times w_s \times c}$, and sequence of frames $\mathcal{V}_{T-t:T} \in \mathbb{R}^{t \times h_f \times w_f \times c}$ taken t time-steps before T. The model's predictions are a set of detections, defined as a tuple $(b_i, n_i, v_i, \delta_i, s_i)$, denoting future interacted objects in the last observed frame I_T. Each bounding box b_i is associated with an object category label n_i (noun), a verb label indicating the interaction mode v_i, a time-to-contact δ_i indicating that the interaction will take place at time $T + \delta_i$, and a confidence score s_i.

STAformer is a novel architecture that leverages pre-trained transformer models for image and video feature extraction [6, 16] and introduces novel attention-based components for image-video representation fusion. We now describe each of the components of STAformer in detail, as Figure 6.2 shows.

I. **Feature Extraction** Following previous work [26, 282], we process a high resolution image $I_T \in \mathbb{R}^{h_s \times w_s \times 3}$ sampled from the input video $\mathcal{V}_{:T}$ at time T and a sequence of low-resolution frames $\mathcal{V}_{T-t:T} \in \mathbb{R}^{t \times h_f \times w_f \times 3}$ taken t time-steps before T. First, we extract high-resolution 2D features from I_T with a DINOv2 model [16], obtaining a set of 2D image tokens $\Phi_{2D}(I_T)$ and a class token C_I offering a global representation of the image. The high-level semantics and dense localization of DINOv2 features makes them very suitable for the object detection task. We also extract spatio-temporal 3D features from $\mathcal{V}_{T-t:T}$ using a TimeSformer model [286], obtaining a set of video tokens $\Phi_{3D}(\mathcal{V}_T)$ and a class token $C_\mathcal{V}$ that captures a global representation of the input clip.

II. **Frame-guided Temporal Pooling Attention (Figure 6.2(a)).** While the overall video tokens provide a spatio-temporal representation of the input video, STA predictions need to be aligned to the spatial location of the last video frame. The frame-guided temporal pooling attention maps video tokens to the spatial reference system of the last video frame, compressing the 3D representation obtained by the TimeSformer to a 2D one. The 3D video tokens $\Phi_{3D}(\mathcal{V}_{T-t:T})$ are mapped to 2D pooled video tokens denoted as $\Phi^{2D}_{3D}(\mathcal{V}_{T-t:T})$ adopting a residual cross-attention mechanism. Specifically, we compute query vectors from last-frame video tokens $\Phi_{3D}(\mathcal{V}_T)$ with a linear projection W_Q, while key and value vectors are computed from the overall video tokens $\Phi_{3D}(\mathcal{V}_{T-t:T})$ using the W_K and W_V linear projection layers. We obtain pooled video tokens with a residual

multi-head attention (A) layer as follows:

$$\begin{gathered}\Phi_{3D}^{2D}(\mathcal{V}_{T-t:T}) = \Phi_{3D}(\mathcal{V}_T) + A(Q_{TP}, K_{TP}, V_{TP}) \\ Q_{TP} = \Phi_{3D}(\mathcal{V}_T) \cdot W_Q, \quad K_{TP} = \Phi_{3D}(\mathcal{V}_{T-t:T}) \cdot W_K, \quad V_{TP} = \Phi_{3D}(\mathcal{V}_{T-t:T}) \cdot W_V.\end{gathered} \tag{6.1}$$

Used as queries, last-frame tokens guide an adaptive temporal pooling that summarizes the spatio-temporal video feature map computed and maps it to the 2D reference space of the last observed frame. The residual connection facilitates learning and lets the attention mechanism focus on enriching last-frame tokens with video tokens.

III. **Dual Image-Video Attention fusion (Figure 6.2(b)).** Image tokens $\Phi_{2D}(I_T)$ and pooled video tokens $\Phi_{3D}^{2D}(\mathcal{V}_{T-t:T})$ are spatially aligned, but carry different information, with image tokens encoding fine-grained visual features and video tokens encoding scene dynamics. This module adopts a residual dual cross-attention that aims to enrich image tokens with scene dynamics information coming from video tokens through image-guided cross-attention and, vice versa, video tokens with fine-grained visual information coming from image tokens through video-guided cross-attention. Prior to forwarding image and video tokens to the multi-head cross-attention modules, these are summed with learnable positional embeddings to capture insightful spatial relationships and normalized through a Layer Norm. The residual image-guided cross-attention is as follows:

$$\begin{aligned}[\tilde{\Phi}_{2D}(I_T), \tilde{C}_I] &= [\Phi_{2D}(I_T), C_I] + A(Q_{CA}, W_{CA}, V_{CA}) \\ Q_{CA} &= [\Phi_{2D}(I_T), C_I] \cdot W_Q, \\ K_{CA} &= [\Phi_{3D}^{2D}(\mathcal{V}_{T-t:T}), C_\mathcal{V}] \cdot W_K, \\ V_{CA} &= [\Phi_{2D}^{3D}(\mathcal{V}_{T-t:T}), C_\mathcal{V}] \cdot W_V,\end{aligned} \tag{6.2}$$

where $[\cdot, \cdot]$ denotes concatenation along batch dimension, and W_Q, W_K, and W_V are linear projection layers. After the multi-head attention layer, the refined image representation $[\tilde{\Phi}_{2D}(I_T), \tilde{C}_I]$ is passed through a residual MLP. The video-guided cross-attention works in a similar way to compute refined video tokens $\tilde{\Phi}_{3D}(\mathcal{V}_{T-t:T})$ and video class tokens $\tilde{C}_\mathcal{V}$, but queries are computed from video tokens while keys and values are computed from image tokens.

IV. **Feature Fusion and Fast-RCNN based STA prediction head (Figure 6.2(c)-(e)).** Refined image and video class tokens are summed to obtain the

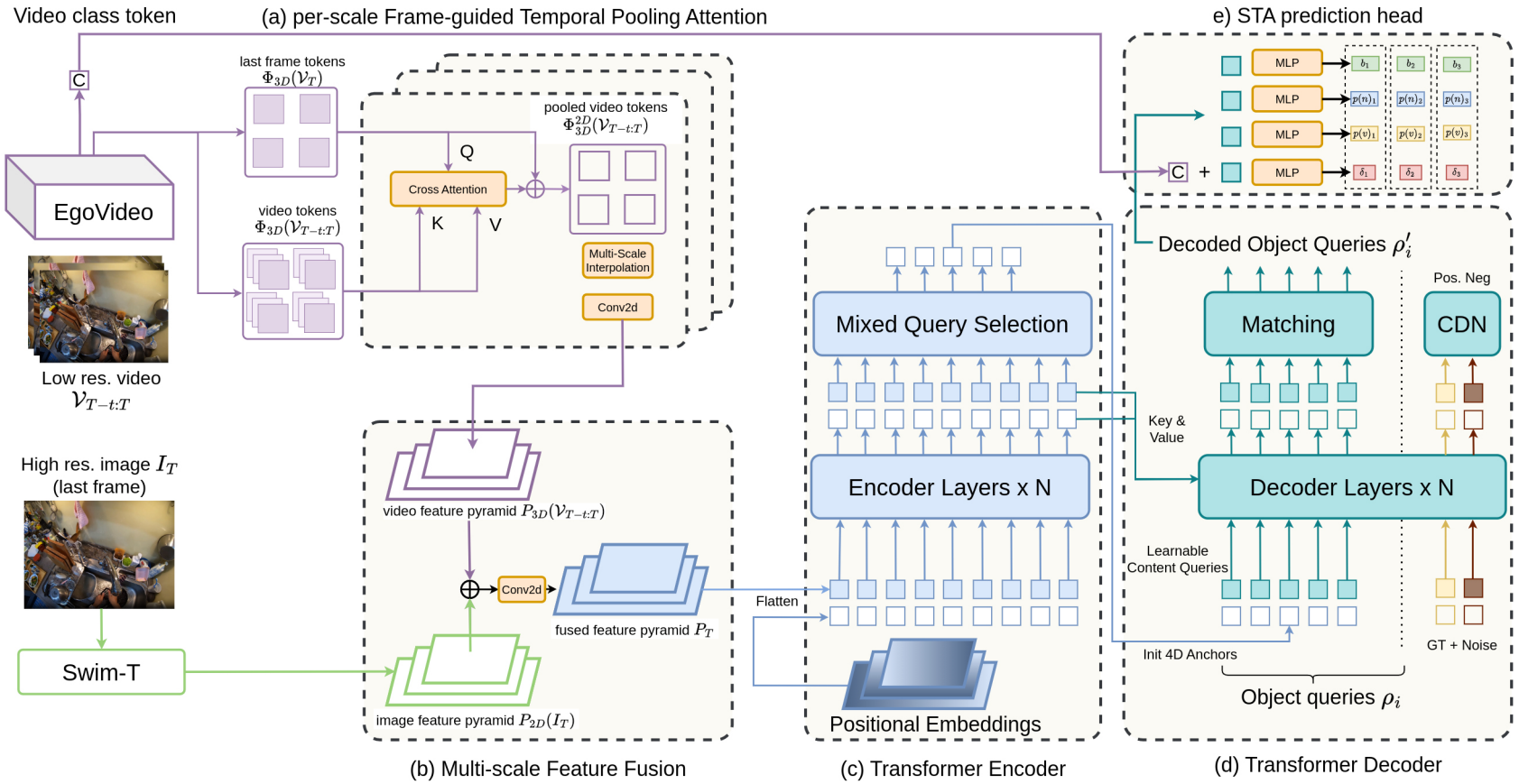


Figure 6.3: **STAformer++ architecture.** The Swin-T backbone extracts hierarchical multi-scale 2D feature maps from the high-resolution image, while the EgoVideo backbone extracts spatio-temporal 3D features. a) We compute per-scale Frame-guided temporal pooling, and then resize the pooled video tokens to the respective image map. b) The two feature maps are summed to obtain the fused feature pyramid P_T. c) The DETR Encoder enhances the features and applies the Mixed Query Selection to initialize the positional part of the object queries ρ_i, while the content parts are kept as learnable parameters. d) The DETR Decoder incorporates the refined image-video features to the object queries. We accelerate the convergence using a Contrastive DeNoising (CDN) part with positive and negative samples. e) The STA prediction head applies independent MLP layers to obtain the final predictions $(\hat{b}_i, \hat{n}_i, \hat{v}_i, \hat{\delta}_i, \hat{s}_i)$.

overall class token $C_T = \tilde{C}_I + \tilde{C}_\mathcal{V}$, a global representation of the input image-video pair (Figure 6.2(c)). Refined image tokens $\tilde{\Phi}_{2D}(I_T)$ are mapped to a multi-scale feature pyramid [141] $P_{2D}(I_T)$ by rescaling $\tilde{\Phi}_{2D}(I_T)$ to multiple resolutions using bilinear interpolation, followed by a 3×3 convolution to compensate for interpolation artifacts. Refined video tokens $\tilde{\Phi}_{3D}(\mathcal{V}_{T-t:T})$ are mapped to a feature pyramid $P_{3D}(\mathcal{V}_{T-t:T})$ in the same way. The two feature pyramids are summed and passed through a 2D 3×3 convolution to obtain the fused feature pyramid P_T (Figure 6.2(d)). We adopt the prediction head proposed in Stillfast [282] to obtain the final predictions $(\hat{b}_i, \hat{n}_i, \hat{v}_i, \hat{\delta}_i, \hat{s}_i)$, which modifies the Faster-RCNN [287] head integrating components specialized for STA prediction. In short, P_T is passed to a Region Proposal Network (RPN), which computes object proposals. Such proposals are then used to extract local features from P_T with RoI Align [12], mapping bounding boxes to appropriate layers of the pyramid following [141]. Each extracted local feature vector is concatenated with the fused class token C_T and passed through an MLP with a residual connection. Linear layers are used to compute noun probabilities $p(n)_i$, verb probabilities $p(v)_i$ and time-to-contact (ttc) predictions. Note that while [282] uses global average pooling to obtain a global representation of the scene, we naturally use the class token C_T learned from the input image-video pair.

6.4 STAformer++: End-to-End Short-Term Anticipation with Transformers

While STAformer delivers state of the art performance, it still makes use of components based on convolutional object detectors, notably in the detection head, which may limit its performance. We investigate whether the inclusion of a transformer-based detection head can further improve performance and propose STAformer++, a redesign of the original STAformer architecture. Specifically, we substitute the Fast-RCNN STA head by a prediction head based on DETR. We also replace the DINOv2 [16] image features by Swin-T [140], a multi-scale image transformer. We subsequently compute per-scale the frame-guided temporal pooling to pool more robust temporal features to the object size. The TimeSformer [6] video feature extraction is substituted by EgoVideo [7], the state-of-the-art in multiple Ego4D challenges. Our STAformer++ pipeline is the following, as Figure 6.3 shows.

I. **Feature extraction.** We process the high resolution image I_T with Swin-T [140] to extract hierarchical multi-scale feature maps $P_{2D}(I_T)$. Swin-T alternates window multi-head self-attention with shifted window partitioning attention, which

introduces cross-window connections. Since its computational cost grows linearly, it is a great candidate for dense vision tasks with high input image resolution. We use EgoVideo [7] for extracting the video tokens $\Phi_{3D}(\mathcal{V}_{T-t:T})$ and the video class token $C_{\mathcal{V}}$ from the low-resolution video $\mathcal{V}_{T-t:T}$.

II. **Per-Scale Frame-guided Temporal Pooling Attention and Feature Fusion (Figure 6.3(a-b)).** The per-scale frame-guided temporal pooling attention maps video tokens to the spatial reference system of the last video frame by implicitly considering the scale of the feature map. Specifically, we repeat the frame-guided temporal pooling attention for each scale of the image features. As detailed in Section 3.2, we adopt a residual cross-attention mechanism between the video tokens of the last frame $\Phi_{3D}(\mathcal{V}_T)$ and the full stack of 3D video features $\Phi_{3D}(\mathcal{V}_{T-t:T})$. The pooled video tokens are followed by a bilinear interpolation and a 2D Convolution to obtain the video feature pyramid $P_{3D}(\mathcal{V}_{T-t:T})$. Then, we sum the two multi-scale feature maps to obtains the fused feature pyramid P_T.

III. **Detection Transformer (Figure 6.3(c-d)).** We flatten the fused feature map P_T to obtain a sequence of tokens which are then forwarded to the DETR Encoder. We sum a fixed positional encoding to incorporate the spatial relationships of the patches. The DETR Encoder consists of standard multi-head self-attention layers followed by feed-forward networks. The self-attention mechanism of the encoder aggregates context from the entire image.

   Then, the DETR Decoder processes the object queries ρ. We follow the mixed query selection strategy proposed by Liu et al. [294] to dynamically initialize the positional part of the object queries, while its content part remains static to accelerate the convergence. Specifically, the positional part are 4D anchor boxes composed by the reference query points (x, y) and anchor sizes (w, h), obtained after a query selection of the top-K encoder features. We apply the deformable attention [293] to layer-by-layer integrate the comprehensive context from the image-video into the object queries. We further accelerate the convergence by feeding noise-altered ground-truth labels and boxes into the DETR Decoder [294], which trains the model for accurate ground-truth reconstruction. Moreover, as proposed by Zhang et al. [296], we adopt the Contrastive DeNoising (CDN) to discard irrelevant anchors and the "look forward twice" for more efficient training.

IV. **DETR based STA prediction head (Figure 6.3(e)).** From the processed object queries ρ', we obtain the final predictions $(\hat{b}_i, \hat{n}_i, \hat{v}_i, \hat{\delta}_i, \hat{s}_i)$. Bounding-box

coordinates are computed with a 3-layer Multi-Layer Perceptron (MLP), predicting the normalized center, height, and width relative to the input image. The noun $p(n)_i$ and verb probabilities $p(v)_i$ are predicted with two independent 3-layer MLP followed by a Sigmoid function, and considering an additional special class label $\emptyset$, which indicates that no object is detected. The score s_i of the joined prediction is the multiplication of the respective noun and verb probabilities. Finally, we concatenate the class token of the video model $C_{\mathcal{V}}$ with the decoded object queries ρ' for explicitly incorporating the action dynamics. We regress the time-to-contact ttc_i with a final 3-layer MLP.

6.5 Environment affordances for human behavior grounding

While end-to-end STA architectures predict the next interaction directly from input video, in this section we show that it is beneficial to ground the predictions on past observed human behavior. Environment affordances [39] refer to all potential interactions that can be performed in a given physical zone. By learning a map of the environment affordances, using egocentric videos of human activities, we are able to guide the next interaction prediction using the similarities and correlations among human activities and scenarios. We describe how to build an affordance semantic map (Figure 6.4-a)) by grouping and connecting similar training videos. Then, we present two methods for grounding predictions on environment affordances. Our first solution (Figure 6.4-b)) pre-computes a fixed affordance distribution of the current scene based on the affordance distributions of similar scenes or videos, where we use the cosine similarity. This affordance distribution is used during inference to refine noun and verb probabilities. Our second proposed strategy (Figure 6.4-c)) learns the affordance distribution during training using a flexible attention mechanism.

6.5.1 Building a persistent memory of affordances

We start extracting activity-centric zones from the training set following [39] in order to build an affordance map as Figure 6.4 a) shows. Each affordance zone is a group of image-video pairs with high visual similarity in a certain environment. We create positive and negative frame pairs labels by counting homography estimation inliers, evaluating temporal coherence, and computing visual similarity. We consider two frames similar if they are less than 15 frames apart or if they share 10 inlier homography key-points. We extract SuperPoint keypoint descriptors [297] and use RANSAC

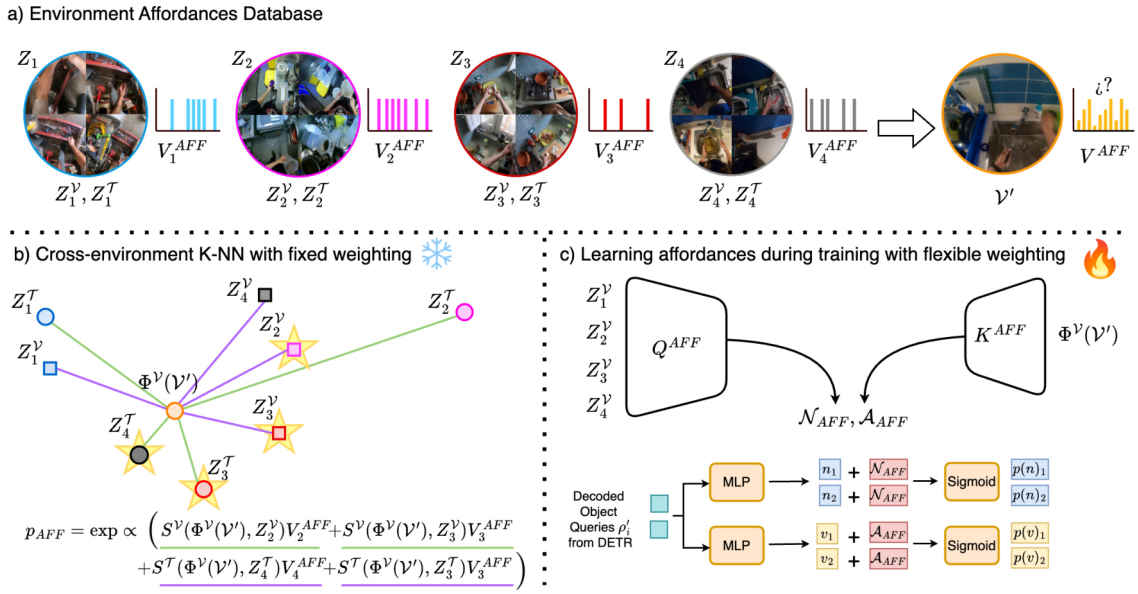


Figure 6.4: **Environment affordances in forecasting.** a) We build an affordance database by linking training videos according to their visual similarity, obtaining activity-centric zones with affordances values V_i^{AFF} and respective video $Z_i^{\mathcal{V}}$, text $Z_i^{\mathcal{T}}$ descriptors. b) Our first approach matches the input encoded video $\Phi^{\mathcal{V}}(\mathcal{V}')$ to the affordance database by selecting the K nearest neighbors in terms of the cosine similarity with the visual $Z^{\mathcal{V}}$ and text $Z^{\mathcal{T}}$ zone descriptors. The affordance probability p_{AFF} is obtained by weighting the counts of nouns present in the top-2K nearest zones ($\bigstar$) according to the respective similarity $\mathcal{S}$. This will be late-fused with the predictions made by the end-to-end model. Example for K=2. c) In our second methodology, an attention mechanism (Q^{AFF}, K^{AFF}) learns to associate a novel video $\mathcal{V}'$ with all the potential zone candidates Z_i in the affordance database. This dynamically obtains the noun $\mathcal{N}_{AFF}$ and verb $\mathcal{A}_{AFF}$ affordance distributions, which are summed to the DETR predicted nouns n_i and verb v_i logits during model training. The final binary class probabilities $p(n)_i$, $p(v)_i$ are obtained after a Sigmoid layer.

for the homography estimation. We measure the visual similarity from pre-trained ResNet-152 features [11] to select dissimilar frames. Based on the positive and negative pairs, we train a Siamese network $\mathbb{L}$, composed by a Resnet-18 [11] followed by a 5 layer multi-layer perceptron (MLP), on these pairs and used to predict the probability $\mathbb{L}(I, I')$ that two frames I and I' belong to the same zone. We then process all frames in a video sequence with $\mathbb{L}$ to group video frames according to their visual similarity in different zones.

Each zone Z represents an activity-centric region composed of the group of visually similar images I_i^Z, their corresponding videos $\mathcal{V}_i^Z$, the associated narrations $\mathcal{T}_i^Z$, sets of nouns $\mathcal{N}_i^Z$ and action verbs $\mathcal{A}_i^Z$ appearing at least once in the STA annotations of all images I_i^Z. We define the affordance distribution in each zone as a unnormalized distribution $V_i^{\mathcal{N}_{AFF}} = \mathbb{1}_{n \in Z_i}$, $V_i^{\mathcal{A}_{AFF}} = \mathbb{1}_{a \in Z_i}$ that considers if the noun $\mathcal{N}$, verb $\mathcal{A}$ appears in the zone i. Since each zone captures the different interactions that the user

performed in that specific environment, this database represents a sort of *persistent memory* on how humans behave in each space. We obtain a visual descriptor $Z^{\mathcal{V}}$ and a text descriptor $Z^{\mathcal{T}}$ for each zone Z using video language pre-trained models [6, 7] to extract the zone video $\Psi^{\mathcal{V}}(\mathcal{V}_i^Z)$ and text $\Psi^{\mathcal{T}}(\mathcal{T}_i^Z)$ descriptors as follows:

$$Z^{\mathcal{V}} = \sum_{i=1}^{|Z|} \Psi_{\mathcal{V}}(\mathcal{V}_i^Z)/|Z|, \quad Z^{\mathcal{T}} = \sum_{i=1}^{|Z|} \Psi_{\mathcal{T}}(\mathcal{T}_i^Z)/|Z|. \tag{6.3}$$

6.5.2 Fixed pre-inferred environment affordances

At inference time, we predict the nouns and verbs affordance distribution by matching a novel video $\mathcal{V}'$ to zones related to functionally similar environments in the affordance database. Since we can only extract a visual descriptor from the novel video, $\Psi^{\mathcal{V}}(\mathcal{V}')$, we compute the visual cosine similarity $\mathcal{S}^{\mathcal{V}}(\Psi^{\mathcal{V}}(\mathcal{V}'), Z^{\mathcal{V}})$ and the video-text cross cosine similarity $\mathcal{S}^{\mathcal{T}}(\Psi^{\mathcal{V}}(\mathcal{V}'), Z^{\mathcal{T}})$ between the clip and each zone Z in the database. Beyond retrieving visually similar zones, the video-text cross cosine similarity relates different locations with similar functionality that affords the same interaction (i.e, painting a wall in India or painting a canvas with watercolor in Spain both afford to dip the brush in the paint). As illustrated in Figure 6.4 a), we employ the K-Nearest Neighbor algorithm to identify the most similar zones to the given input $\mathcal{V}'$. We define the top-K visual zones $\mathcal{K}^{\mathcal{V}}$, where $S_k^{\mathcal{V}}$ is a shorthand notation for $S_k^{\mathcal{V}}(\Psi(\mathcal{V}'), Z_k^{\mathcal{V}})$, and the top-K narrative zones $\mathcal{K}^{\mathcal{T}}$, defined as:

$$\mathcal{K}^{\mathcal{V}} = \{(Z_1^{\mathcal{V}}, S_1^{\mathcal{V}}), ..., (Z_K^{\mathcal{V}}, S_K^{\mathcal{V}})\}, \quad \mathcal{K}^{\mathcal{T}} = \{(Z_1^{\mathcal{T}}, S_1^{\mathcal{T}}), ..., (Z_K^{\mathcal{T}}, S_K^{\mathcal{T}})\}. \tag{6.4}$$

Combining both sets, $\mathcal{K} = \mathcal{K}^{\mathcal{V}} \cup \mathcal{K}^{\mathcal{T}} = \{(Z_i, S_i)\}_{i=1}^{2K}$ yields a total of $2K$ zones and their respective similarity scores, which we assume to share affordances with $\mathcal{V}'$. We then define the probability of each noun $p_{\text{aff}}(n|\mathcal{V}')$ as an exponential distribution by weighting the noun and verb appearance in each neighboring zone according to the respective similarity S_i:

$$p_{\text{aff}}(n|\mathcal{V}') \propto \exp(\sum_{(Z_i,S_i)\in\mathcal{K}} S_i \cdot V_i^{\mathcal{N}_{AFF}}), \tag{6.5}$$

$$p_{\text{aff}}(v|\mathcal{V}') \propto \exp(\sum_{(Z_i,S_i)\in\mathcal{K}} S_i \cdot V_i^{\mathcal{A}_{AFF}}). \tag{6.6}$$

Based on the environment affordances, we can predict probability distributions over *possible* nouns $p_{\text{aff}}(n|\mathcal{V})$ or verbs $p_{\text{aff}}(v|\mathcal{V}')$ given past interactions in functionally similar zones. Differently, the STA model will predict probability distributions of given

nouns and verbs being the next interactions $p_{\text{sta}}(n|\mathcal{V}', I')$ and $p_{\text{sta}}(v|\mathcal{V}', I')$ directly from the input image-video pair, without explicitly considering the set of possible actions. We assume independence between the two predictions[1] and perform data fusion by computing the unnormalized joint likelihoods:

$$\begin{aligned} p_{\text{fus}}(n|I', \mathcal{V}') &\propto p_{\text{aff}}(n|\mathcal{V}') \cdot p_{\text{sta}}(n|\mathcal{V}', I'), \\ p_{\text{fus}}(v|I', \mathcal{V}') &\propto p_{\text{aff}}(v|\mathcal{V}') \cdot p_{\text{sta}}(v|\mathcal{V}', I'). \end{aligned} \tag{6.7}$$

6.5.3 Learning of environment affordances

The approach described in Section 6.5.2 uses affordances to refine predictions at inference time. To further improve the exploitation of affordances, here we propose an alternative approach that learns the affordances directly during the training of the end-to-end STA prediction model. Specifically, we adopt an attention mechanism between the input video descriptor $\Phi(\mathcal{V}')$ and the descriptors $\Phi(\mathcal{Z}_i^{\mathcal{V}})$ of the affordances zones. We interpret the attention mechanism [14] as a learnable way of querying the most similar situations from the agent's past memory. In order to obtain the affordance keys K^{AFF}, we project with a linear layer W_K the zone video embeddings $\mathcal{Z}^{\mathcal{V}} = \sum_{i=1}^{|Z|} \Psi_{\mathcal{V}}(\mathcal{V}_i^Z)/|Z|$, which represent descriptors of that environment in the memory. In this case, the per-zone video embedding $\Psi^{\mathcal{V}}(\mathcal{V}_i^Z)$ is the mean EgoVideo [7] class token $\overline{C}_{\mathcal{V}}$ of the videos inside each affordance zone $\mathcal{Z}$. We represent the affordance query Q^{AFF} by processing the EgoVideo class token of the novel video $C_{\mathcal{V}'}$ with a learnable linear layer W_Q, while keys are computed from the affordance database with a linear layer W_K . The W_K, W_Q layers learn to compute the similarity of a novel video with respect to all the past environment observations. In contrast with a rigid similarity, here we learn how to best associate the input video to the learned affordance zones as:

$$Q^{AFF} = C_{\mathcal{V}'} \cdot W_Q, \quad K^{AFF} = \mathcal{Z}_i^{\mathcal{V}} \cdot W_K. \tag{6.8}$$

As values, we use $V^{\mathcal{N}_{AFF}}$ and $V^{\mathcal{A}_{AFF}}$, the non-normalized affordance distributions within the zone Z_i, defined as $\mathbb{1}$ if the noun $\mathcal{N}$ or action verb $\mathcal{A}$ is present, or zero otherwise

$$V^{\mathcal{N}_{AFF}} = \mathbb{1}_{n \in \mathcal{N}^{Z_i}}, \quad V^{\mathcal{A}_{AFF}} = \mathbb{1}_{v \in \mathcal{A}^{Z_i}}. \tag{6.9}$$

Then, we obtain the affordances distribution from the attention scores and the unnormalized values. Rather than taking the Softmax and compute the weighted sum

[1] In practice, we build the two models with different architectures and training objectives to make the dependence weak.

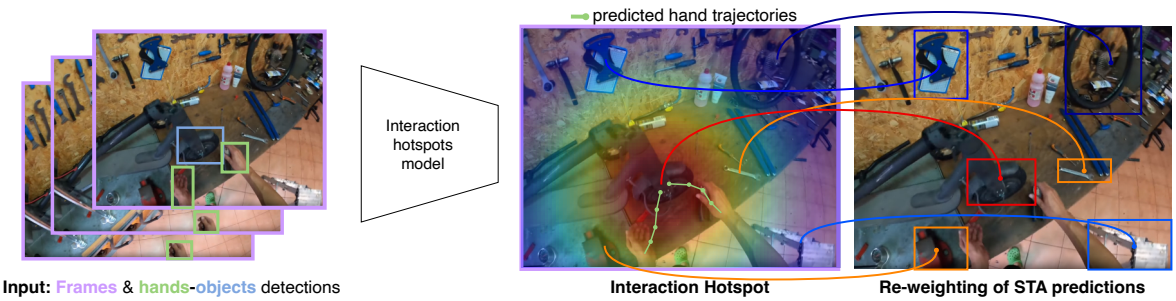


Figure 6.5: **Refinement of confidence scores based on the interaction hotspots.** The interaction hotspot model observes frames, hands, and objects and forecasts a map encoding the probability of the interaction in each pixel. STA confidence scores are re-weighted based on the probability values at the bounding box coordinate centers, reducing confidence in false positive predictions falling far from the interaction hotspot.

across all the different environment zones, we apply a Sigmoid on the attention score to obtain binary probabilities indicating whether a given environment is relevant to our query video. We hence compute the maximum probability value for each verb and noun across all zones. The choice of binary probabilities is due to the fact that DETR models predict a binary probability for each class. The max operation makes the distribution less sensitive to the long-tail distribution of verbs and nouns. We show this computation in the following formula:

$$\begin{aligned} p_{\text{aff}}(n|\mathcal{V}') &= \max[\text{Sigmoid}(Q^{AFF} \cdot (K^{AFF})^T)] \cdot V_{\mathcal{N}}^{AFF}, \\ p_{\text{aff}}(v|\mathcal{V}') &= \max[\text{Sigmoid}(Q^{AFF} \cdot (K^{AFF})^T)] \cdot V_{\mathcal{A}}^{AFF}. \end{aligned} \tag{6.10}$$

Our base DETR [296] prediction head uses the Binary Cross Entropy loss for predicting the object class. Therefore, we transform the nouns/verbs affordance probability to the logits space following $\log(p_{AFF} + \epsilon) - \log(1 - p_{AFF} + \epsilon)$ and fuse there to use the numerically stable loss function requiring logits. In this way, we ground the learning of the model on the past human behavior, contributing to the full model training. In this more flexible approach, the model adapts dynamically to each novel video $\mathcal{V}'$, as we do not rely on a fixed distance or number of zones and learns to attend to a memory available at test time.

6.6 Leveraging interaction hotspots

While our affordance database gives us information on which objects (nouns) and interaction modes (verbs) are likely to appear in the current scene, it does not give us any information on where the interaction will take place in the observed images. As noted in previous works [67, 68], observing how hands move in egocentric videos can allow us to predict the interaction hotspot [39, 67], a distribution over image regions

indicating possible future interactions locations. We exploit this concept and include a module to predict an interaction hotspot by observing frames, hands, and objects. As Figure 6.5 illustrates, we hence re-weigh the confidence scores s_i of STA predictions according to the location of the respective bounding box centers in the predicted interaction hotspot, to reduce the influence of false positive detections falling in areas of unlikely interaction.

6.6.1 Inferring interaction hotspots

We base our interaction hotspot module on the work presented in [67] with some improvements. First, we fine-tune the hand object detector presented in [36] on EGO4D-SCOD [26] annotations, rather than using it out-of-the-box. Second, we extract stronger egocentric-aware frame features with the video part of the dual-encoder version of EgoVLP [6] pre-trained on Ego4D [6], instead of using a ConvNet as in [67].[2] The model takes as inputs the features of the observed frames, besides the coordinates and features of both hands and pre-detected objects, and is trained to forecast the hand trajectory, from which it predicts a distribution over plausible future contact points. Given the observed image-video pair $(I_T, \mathcal{V}_{T-t:T})$, the output of the model is a probability distribution over the spatial locations of I_T indicating the probability of interaction of each pixel denoted as $p_{ih}(x, y|I_T, \mathcal{V}_{T-t:T})$.

6.6.2 Fusing STA predictions with interaction hotspots:

We exploit the interaction hotspots to refine the predictions of the STA model, assuming that regions close to the predicted interaction hotspots are more likely to contain the next active objects. Given a predicted box $\hat{b}_i$, we re-weigh its related confidence score $\hat{s}_i$ according to the location of the bounding box center $(\hat{c}_i^x, \hat{c}_i^y)$ in the interaction hotspot as following: $\hat{s}_i \cdot p_{ih}(\hat{c}_i^x, \hat{c}_i^y|I_T, \mathcal{V}_{T-t:T})$.

6.7 Experimental setup

Following the official benchmark [26], we adopt standard Noun (N), Noun+Verb (N+V), Noun+time-to-contact (N+δ) and Noun+Verb+time-to-contact (All) Top-5 mean Average Precision (mAP). We also provide detailed comparative on the Top-5 Average Precision metric (AP) as defined in [284].

[2] See supp. for more information on the interaction hotspot prediction module.

6.7.1 Datasets

We validate our method on Ego4D [26] and EPIC-Kitchens [25], two large-scale datasets of egocentric videos with high diversity and long-tail distributions classes.

Ego4D. We consider both the first and second versions of Ego4D STA annotations. Version 1 (v1) of the Ego4D STA split is composed of 27,801 training 17,217 validation and 19,870 test instances with 87 noun and 74 verb categories. Version 2 (v2) extends v1 with additional videos and annotations, for a total of 98,276 training, 47,385 validation and 19,870 test videos, with a taxonomy of 128 nouns and 81 verb classes. Ego4D STA contains a single test split which is compatible with v1 and v2, hence models trained on either versions can be compared on the same test split.

EPIC-Kitchens. Since Ego4D is the only dataset containing STA annotations to date, we extend EPIC-Kitchens dataset [25] by post-processing its active object and action segment annotations. We first merge active object bounding boxes into tracks by grouping neighboring annotations of the same object class and removing tracks with multiple bounding boxes for the same class. Then, we match each object track to one of the annotated action segments. Specifically, if an action segment including the same noun as the object track is found, this is matched to the object track. Next, we truncate object tracks to exclude frames depicting the action, enabling anticipation of future actions. Finally, we attach the following data to a given bounding box: the noun associated to the track as the object category, the associated action segment as the interaction verb, the distance from the time-step of the current frame to the beginning of the associated action segment as the time to contact. The final set of annotations contains 33,804 training and 7,055 validation instances with 104 noun and 51 verb categories, which we release to the community.

6.7.2 Implementation details

We train STAformer with Adam as an optimizer, an initial learning rate of 10^{-4} with linear warm-up, and a weight decay of 10^{-6}, on 4 Tesla V100 GPUs. Similarly, we train STAformer++ following the official procedure of DINO-DETR [296] with AdamW as optimizer and an initial learning rate of $2 \cdot 10^{-5}$. For the image encoders, we first adopt the DINOv2-B [16] visual transformer, composed by 12 blocks. Alternatively, we extract multi-scale image features with the Swin-T [140] Large version pre-trained on COCO dataset formed by 24 blocks grouped in hierarchical depths. We fine-tune the last 3 blocks in both cases. We utilize the video encoder of EgoVIDEO [7] composed of 38 blocks. For TimeSformer, we sample 16 frames at 30 FPS, while the available version of EgoVideo only allows processing 4 frames, which we sample at 7.5 FPS to

Model	N	N + V	N + δ	All
FRCNN+SF [26]	17.55	5.19	5.37	2.07
FRCNN+Feat. [298]	22.01	5.52	5.54	1.78
StillFast [282]	16.21	7.47	4.94	2.48
Transfusion [281]	20.19	7.55	6.17	2.60
STAformer	21.71	10.75	7.24	3.53
STAformer & AFF (fixed)	24.36	12.00	7.66	3.77
STAformer++	32.07	15.00	8.53	4.31
STAformer++ & AFF (learned)	**33.21**	**15.94**	**8.98**	**4.66**
Gain (rel %)	+36.3	+24.5	+17.2	+23.6

Table 6.1: **Results in mAP on the validation split of Ego4D-STA v1.**

	B	B+N	B+V	B+δ	B+N+V	B+N+δ	B+V+δ	All
Slowfast	40.50	24.50	0.34	8.16	0.34	5.00	0.06	0.06
Slowfast (w/Transformer)	40.50	24.50	8.20	7.50	8.20	4.50	1.30	0.73
AVT	40.50	24.50	8.45	7.12	8.45	4.39	1.15	0.71
ANACTO	40.50	24.50	8.90	7.47	8.90	4.55	1.54	0.91
MeMVIT	40.50	24.50	10.05	9.27	10.04	4.95	2.11	1.34
GANO (w/o guided attn.)	40.50	24.50	7.10	9.01	7.10	4.20	1.22	0.75
GANO (w/ guided attn.)	45.30	25.80	10.56	10.1	10.56	5.90	2.77	1.70
StillFast	27.78	17.75	10.21	7.33	7.01	4.61	2.68	1.77
STAformer	38.38	28.36	16.66	12.27	12.66	8.89	5.47	4.06
STAformer++	**49.68**	38.83	18.02	**12.92**	14.00	10.58	**5.75**	4.63
STAformer++ & AFF(learned)	48.60	**39.12**	**19.45**	12.29	**15.77**	**10.28**	5.71	**4.77**
Gain	+9.7	+37.9	+16.7	+5.3	+24.5	+15.6	+4.4	+17.5

Table 6.2: **Results in AP on the validation split of Ego4D-STA v1.**

cover the same video segment. In both cases, we fine-tune the last 4 blocks of the video model. The DETR model weights are initialized using the 12-epoch version of [296].

6.8 Results

We compare our model against several STA baselines which either provide open source implementations [26, 282] or report results in their papers [26, 279, 281, 282, 284, 298]. We also report multiple ablation studies showing the contribution of each individual component of our approach.

6.8.1 Comparison with the state-of-the-art

Ego4D v1 validation split (Tables 6.1-6.2). Our initial version of STAformer achieves 21.71 N, 10.75 N+V, 7,24 N+δ and 3.53 All mAP, while our the most advanced version of STAformer plus the incorporation of learned affordances, scores a 33.21 N, 15.94 N+V, 8.89 N+δ and 4.88 All mAP in the v1 split, as Table 6.1 shows. We

Model	N	N + V	N +δ	All
FRCNN+SF [26]	21.00	7.45	7.07	2.98
InternVideo [279]	19.45	8.00	6.97	3.25
StillFast [282]	20.26	10.37	7.26	3.96
GANO v2 [284]	20.52	10.42	7.28	3.99
STAformer	27.51	14.68	9.63	5.50
STAformer & AFF (fixed)	29.39	15.38	9.94	5.67
STAformer++	36.78	17.26	11.03	5.87
STAformer++ & AFF (learned)	**37.41**	**18.51**	**11.14**	**6.26**
Gain (rel %)	+27.8	+20.3	+12.7	+10.4

Table 6.3: **Results in mAP on the validation split of Ego4D-STA v2.**

	B	B+N	B+V	B+δ	B+N+V	B+N+δ	B+V+δ	All
STAformer	43.24	33.53	20.88	14.84	16.52	11.23	7.70	5.89
STAformer++	55.95	47.02	24.40	**16.91**	20.24	**14.14**	8.40	**6.85**
STAformer++ & AFF(learned)	**55.98**	**47.63**	**24.82**	16.40	**20.77**	13.86	**8.42**	6.77
Gain	+29.3	+42.2	+18.9	+13.9	+25.7	+25.9	+9.3	+14.9

Table 6.4: **Results in AP on the validation split of Ego4D-STA v2.**

obtain a relative gain[3] up to +23.6 % in the mAP All metric compared with our previous conference version [299]. Tables 6.2 and 6.4 compare our method with previous approaches reporting results using the AP metric. In this case, the initial version of STAformer, based on a Faster-RCNN architecture, shows a lower detection performance (38.38 B AP) compared with the most advanced version of GANO [285] (45.30 Box AP). However, the novel DETR-based architecture of STAformer++ notably improves the quality of the predicted bounding boxes up to 49.68 B AP (+9.7 % relative gain), which is reflected in the overall metric where it achieves a 4.77 (+ 17.5 % relative gain). The results also show the benefits of leveraging affordances in the short-term anticipation task, which are specially relevant in the semantic metrics (+ 37.9 % B+N AP, +16.7 % B+V AP, +24.5 % B+N+V AP, + 36.3 % N mAP, + 24.5 % N+V mAP), since this prior just refines the noun and verb probabilities.

Ego4D v2 validation split (Tables 6.3-6.4). The overall improvement is also significant in the v2 split, a larger version of Ego4D containing also v1 annotations. Our most advanced version, STAformer++ with learned affordances, scores 37.41 N mAP, 18.51 N+V mAP, 11.14 N+δ and 6.26 All mAP, showing a relative gain of +10.4 % in the overall metric and demonstrating significant improvements both in semantic, detection, and temporal performance.

Ego4D test split (Tables 6.5-6.6). Since the test set of Ego4D is private, we are only able to compare approaches showing test results in their papers. For fair

[3]We compute the relative gain% of x relative to y as $100 \cdot (\frac{x-y}{y})$.

Model	N	N + V	N + δ	All
FRCNN+SF. [26]	20.45	6.78	6.17	2.45
FRCNN+Feat. [298]	20.45	4.81	4.40	1.31
InternVideo [279]	24.60	9.18	7.64	3.40
Transfusion [281]	24.69	9.97	7.33	3.44
StillFast [282]	19.51	9.95	6.45	3.49
STAformer	24.39	12.49	7.54	4.03
STAformer & AFF (fixed)	26.52	13.15	7.78	4.06
STAformer++	33.78	14.28	**10.14**	4.97
STAformer++&AFF(learned)	**34.06**	**15.94**	10.10	**5.24**
Gain (rel %)	+28.4	+21.2	+29.8	+29.1

Table 6.5: **Results in mAP on the test split of Ego4D-STA of models trained on the v1 training split.**

Model	N	N + V	N + δ	All
StillFast [282]	25.06	13.29	9.14	5.12
GANO v2 [284]	25.67	13.60	9.02	5.16
Language NAO	30.43	13.45	10.38	5.18
EgoVideo	31.08	16.18	12.41	**7.21**
STAformer	30.61	16.67	10.06	5.62
STAformer & AFF (fixed)	32.39	17.38	10.26	5.70
STAformer++	41.96	19.16	**13.05**	6.92
STAformer++&AFF(learned)	**42.07**	**19.51**	12.73	6.26
Gain (rel %)	+29.9	+12.3	+5.2	-4.1

Table 6.6: **Results in mAP on the test split of Ego4D-STA of models trained on the v2 training split.**

comparisons, we report two settings with methods trained on v1 or v2. Our method achieves significant gains with respect to trained methods on v1, for instance, obtaining a +28.4% N mAP, +21.2% N+V mAP, + 29.8% N+δ mAP and +29.1% in mAP All. We observe similar improvements when training on v2, with +29.9% N mAP, +12.3% N+V mAP and +5.2% N+δ mAP. However, our model does not outperform EgoVideo in the overall metric, scoring 6.92 vs. 7.21 All mAP. Note that the participating version of EgoVideo is fully fine-tuned to the STA task on Ego4D and covers 16 frames, while, for computational constraints, our STAformer++ model just trains the last 4 blocks of a simpler general model which only processes 4 frames. It is worth noting that our approach also benefits from training on larger datasets, improving from 5.24 mAP All when it is trained on v1 (Table 6.5) to 6.92 mAP All when training on v2 Table 6.6.

EPIC-Kitchens STA (Table 6.7). Since this benchmark is new, we train the official implementation of StillFast [282] on EPIC-Kitchens as a baseline, obtaining 21.24 N mAP, 12.41 N+V mAP, 6.22 N + δ mAP and 3.28 All mAP. The introduction of more powerful backbones (DINO and TimeSformer) and the dual cross-attention mechanism

Model	N	N + V	N + δ	All
StillFast [282]	21.24	12.41	6.22	3.28
STAformer	25.25	17.17	9.10	6.13
STAformer & AFF (fixed)	28.37	18.95	9.29	6.60
STAformer++	44.96	24.67	14.01	7.87
STAformer++&AFF(learned)	**45.34**	**25.82**	**14.06**	**8.67**
Gain (rel %)	+59.5	+36.20	+51.6	+31.5

Table 6.7: **Results in mAP on the validation split of EPIC-Kitchens.**

in STAformer achieve a 28.37 N mAP, 18.95 N+V mAP, 9.29 N + δ mAP and 6.60 All mAP. The performance gains are particularly notable with STAformer++, achieving 45.34 N mAP, 25.82 N+V mAP, 14.06 N+δ mAP and 8.67 All mAP, representing a +31.5 % increase in All mAP. This highlights the generality of our framework in different training regimes and datasets.

6.8.2 Ablation Study on STAformer and STAformer++ components

Table 6.8 ablates the performance effect of the proposed components of the STA models: the image encoder, the video encoder, the temporal pooling, the 2D-3D fusion module and the prediction head (Faster-RCNN or DETR based).

Image encoder and STA head (Table 6.8, Exp A). We first encode the image with DINOv2 (Exp A.1) and discard the video, obtaining small gains with respect to the baseline [282]. While [282] fully trains both image-video encoders, the A1 version trains solely the Faster-RCNN STA prediction head and reflects the modeling capacity of DINOv2. Then, we replace the Faster-RCNN head by the DETR [296] in Experiments A2-A4. When the entire Swin-T is frozen and uses existing weights pre-trained on COCO(Exp. A2), it losses the generalization capabilities of DINOv2 features, obtaining a drop to 1.91 mAP in the All metric. Then, when we refine the last blocks of both image encoders the performance increases, achieving a 2.55 mAP All for the Swin-T version and a 2.88 mAP All for the DINOv2 model. The main benefit of using a DETR-based model is its superior detection capabilities, as the Noun mAP shows an improvement from 17.48 (DINOv2 with Faster-RCNN) to 29.33 (DINOv2 with DETR).

Video encoder (Table 6.8, Exp B). Using per-frame DINOv2 features with mean temporal pooling (Exp. B1 vs. A1) reduces performance, highlighting DINOv2's limitations in capturing video dynamics. However, incorporating an specific video encoder like the X3D 3D CNN [288] (Exp. B2 vs. A1) achieves better results, indicating the advantage of appropriately encoding video dynamics. Experiments B3, B4 and B5 show

Exp.	Image Encoder	Video Encoder	Temporal Pooling	2D-3D Fusion	Detection Head	N	N + V	N + δ	All
[282]	R50	X3D	Mean	Sum	Fast-RCNN	16.21	7.52	4.94	2.48
A1	DINOv2	-	-	-	Fast-RCNN	17.48	8.64	5.20	2.52
A2	Swin-T	-	-	-	DETR	27.69	9.57	5.43	2.71
A3	Swin-T	-	-	-	DETR	28.77	11.04	6.12	2.85
A4	DINOv2	-	-	-	DETR	29.33	11.65	6.46	2.98
B1	DINOv2	DINOv2	Mean	Sum	Fast-RCNN	15.82	7.65	4.11	2.19
B2	DINOv2	X3D	Mean	Sum	Fast-RCNN	18.84	8.84	5.56	2.57
B3	DINOv2	TimeSformer	Mean	Sum	Fast-RCNN	16.67	8.38	5.16	2.63
B4	DINOv2	TimeSformer	Mean	Sum	DETR	26.11	10.65	6.90	3.02
B5	Swin-T	TimeSformer	Mean	Sum	DETR	24.55	9.49	6.01	2.89
B5	Swin-T	EgoVideo	Mean	Sum	DETR	31.82	13.15	6.81	3.24
B6	Swin-T	EgoVideo	Mean	Sum	DETR	32.50	14.72	7.73	3.65
C1	DINOv2	TimeSformer	Conv	Sum	Fast-RCNN	17.36	8.75	6.05	2.94
C2	DINOv2	TimeSformer	SH.Attn	Sum	Fast-RCNN	19.78	10.04	6.35	3.39
C3	Swin-T	EgoVideo	MH.Attn	Sum	DETR	31.31	14.22	8.05	4.07
C4	Swin-T	EgoVideo	per-Scale MH.Attb	Sum	DETR	**32.57**	**15.10**	**8.53**	4.31
C5	DINOv2	EgoVideo	MH.Attn	Sum	DETR	31.15	13.72	7.71	3.84
C6	DINOv2	EgoVideo	per-Scale MH.Attb	Sum	DETR	31.73	14.10	8.21	4.26
D1	DINOv2	TimeSformer	SH.Attn	Dual $I \leftrightarrow \mathcal{V}$ Attn	Fast-RCNN	20.08	10.21	6.51	3.47
D2	DINOv2	TimeSformer	SH.Attn	Dual $I \leftrightarrow \mathcal{V}$ Attn	Fast-RCNN	21.71	10.75	7.24	3.53
D3	DINOv2	TimeSformer	SH.Attn	Single $I \rightarrow \mathcal{V}$ Attn	Fast-RCNN	20.01	10.04	5.80	3.01
D4	DINOv2	TimeSformer	SH.Attn	Single $\mathcal{V} \rightarrow I$ Attn	Fast-RCNN	20.12	10.31	6.30	3.35
D5	DINOv2	TimeSformer	MH.Attn	MH.Dual $I \leftrightarrow \mathcal{V}$ Attn	Fast-RCNN	23.02	11.57	7.86	3.85
D6	Swin-T	EgoVideo	per-Scale MH.Attb	MH.Dual $I \leftrightarrow \mathcal{V}$ Attn	DETR	31.80	15.06	7.95	**4.54**
D7	DINOv2	EgoVideo	per-Scale MH.Attb	MH.Dual $I \leftrightarrow \mathcal{V}$ Attn	DETR	31.82	13.92	7.98	4.21

Table 6.8: **Ablation study of the architectural components of STA-former on the validation split of Ego4D-v1.** Encoder frozen Encoder fine-tuned. For fair comparison, we fine-tune 3 blocks in the image encoders and 4 blocks in the video encoders and the video comprises 0.5 sec. We refer to STAformer when the detection head is based in Faster-RCNN, while STAformer++ makes reference to DETR-based models.

that adopting TimeSformer as video model only leads to marginal improvements with respect to A4, B2 and A3, respectively. Incorporating EgoVideo improves the semantic and temporal reasoning. Indeed, finetunning the last blocks of EgoVideo (Exp. B6) achieves a 7.73 N+δ mAP, due to the direct connection of the EgoVideo class token $C_{\mathcal{V}}$ with the temporal MLP.

Temporal Pooling (Table 6.8, Exp C). We start showing the effects of temporal pooling in Experiments B3 (mean temporal pooling), C1 (temporal convolution) and C2 (frame-guided temporal pooling). The temporal convolution helps capturing the video dynamics and obtaining more accurate time to contact estimates, improving N+δ mAP up to 6.05. However, our frame-guided attention mechanism (Exp C2) enhances spatio-temporal understanding of the video, achieving significant improvements from 8.75 to 10.04 N+V mAP and from 2.94 to 3.39 All mAP. Unlike convolutional pooling,

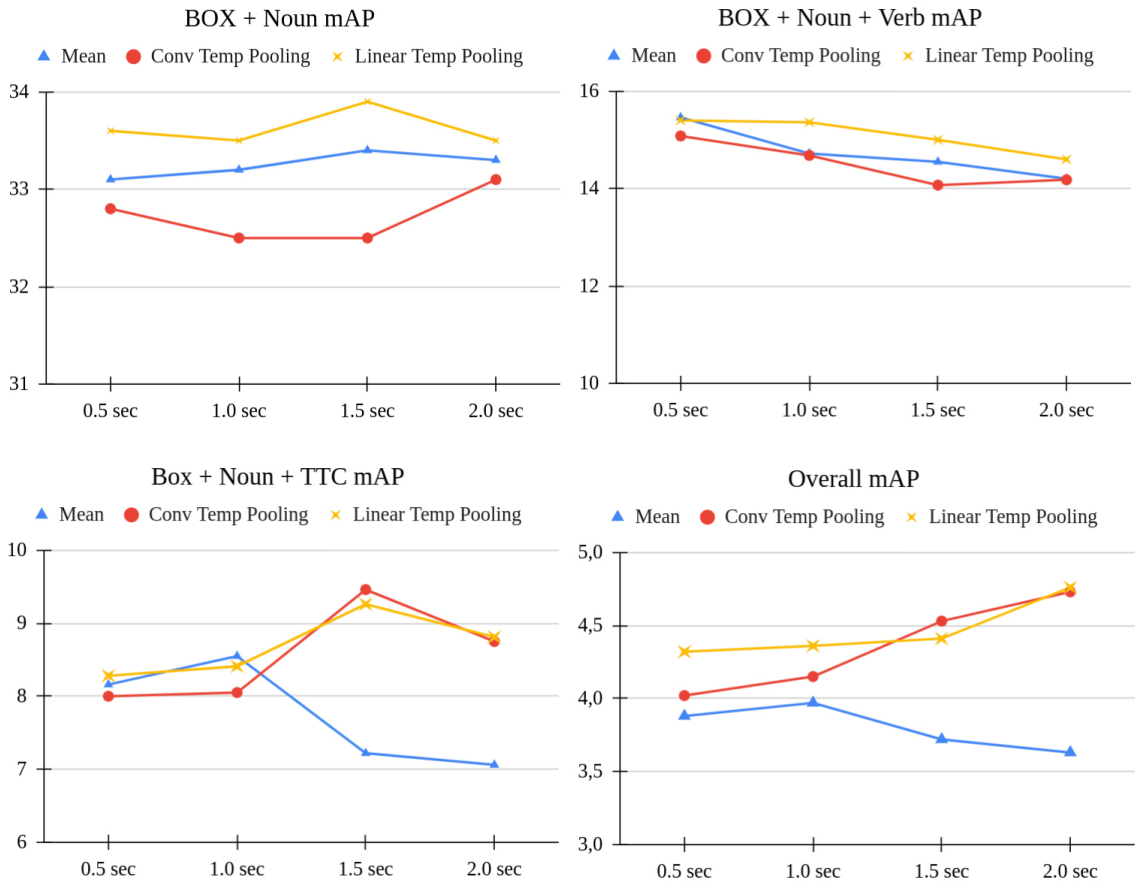


Figure 6.6: **Performance evolution according to the amount of video seen.** We report the mAP N, mAP N+V, mAP N+δ, mAP Overall on the validation split of Ego4D-STA v1.

which focuses solely on temporal dynamics, our attention mechanism joins a spatio-temporal understanding of the video by mapping to the 2D reference space of the last observed frame the pooled video features. This advantage extends to DETR head versions, with significant improvements on the time to contact score up to 8.53 N + δ mAP in Exp C4 (multi-head attention). Finally, performing the temporal pooling per-scale further enhances performance, achieving up to 32.58 N mAP, 15.00 N+V mAP, 8.53 N+δ mAP, and 4.71 All mAP, demonstrating more robust spatio-temporal feature learning adaptable to object sizes.

Feature Fusion (Table 6.8, Exp D). Experiments D1-D7 of Table 6.8 compare the contribution of the proposed Dual Image-Video Attention module for 2D-3D feature fusion. Comparing experiments D1 vs. C2 shows small but consistent gains when dual image-video attention is used for fusion in STAformer, as compared to simple sum fusion (20.08 vs. 19.78 N, 10.21 vs. 10.04 N + V, 6.51 vs. 6.35 N + δ and 3.47 vs. 3.39 All mAP). However, using cross-attention only with image tokens ($I \rightarrow \mathcal{V}$-Exp.D3) or video tokens ($\mathcal{V} \rightarrow I$ -Exp.D4) as queries, performs worse than the proposed dual

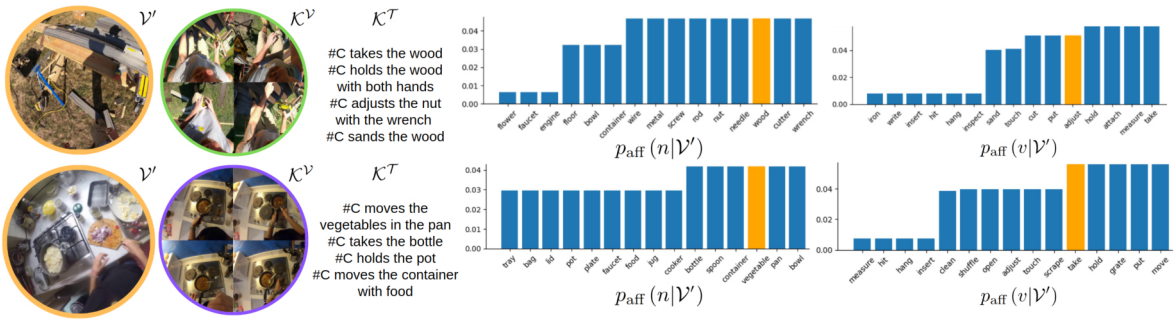


Figure 6.7: **Fixed affordances distribution extracted for refining predictions only at inference.** We visualize the closest environments in terms of the visual $\mathcal{K}^{\mathcal{V}}$ and narrative $\mathcal{K}^{\mathcal{T}}$ cosine similarity. We show in orange the STA ground-truth label.

image-video attention (Exp. D2), suggesting the need to incorporate the refinement of both modalities. Incorporating multi-head attention on the temporal pooling and on the 2D-3D fusion (Exp.D5) produces a consistent improvement in all the metrics due to its ability to capture diverse patterns simultaneously. However, we do not see any systematic improvement when we apply the MH.Dual Cross Attention on the STAformer++ model. Since we are operating at multi-scale levels, this introduces a long sequence of tokens that makes this mechanism very computational consuming, leading a trade-off between the number of fine-tuned video blocks and the dual cross attention.

Dependence on video length (Figure 6.6). We analyze the performance with different time windows for the video model, as Figure 6.6 shows. We compare a temporal pooling with two versions of our frame-guided temporal pooling: the Conv. Temp.Pooling uses convolutional weights on the Q_{TP}, K_{TP}, V_{TP} projection layers, while the Linear Temp.Pooling adopts linear layers plus a positional encoding. Averaging features of longer videos reduces the spatial alignment due to the camera movement. However, computing our frame-guided temporal attention pooling projects the video features into the last frame via the attention mechanism, capturing better aligned spatio-temporal features. This effect is visualized in the N + δ and Overall mAP plots: while a larger time-window degrades the performance when computing the temporal mean, it benefits the temporal reasoning of the model, obtaining better results when the video covers 1.5 secs rather than 0.5 sec.

6.8.3 Ablation Study on Affordances

Tables 6.9, 6.10, 6.11 detail the influence of Environment Affordances (E.AFF) and Interaction Hotspots (I.H), when integrated, separately and jointly, showing in all cases consistent improvements.

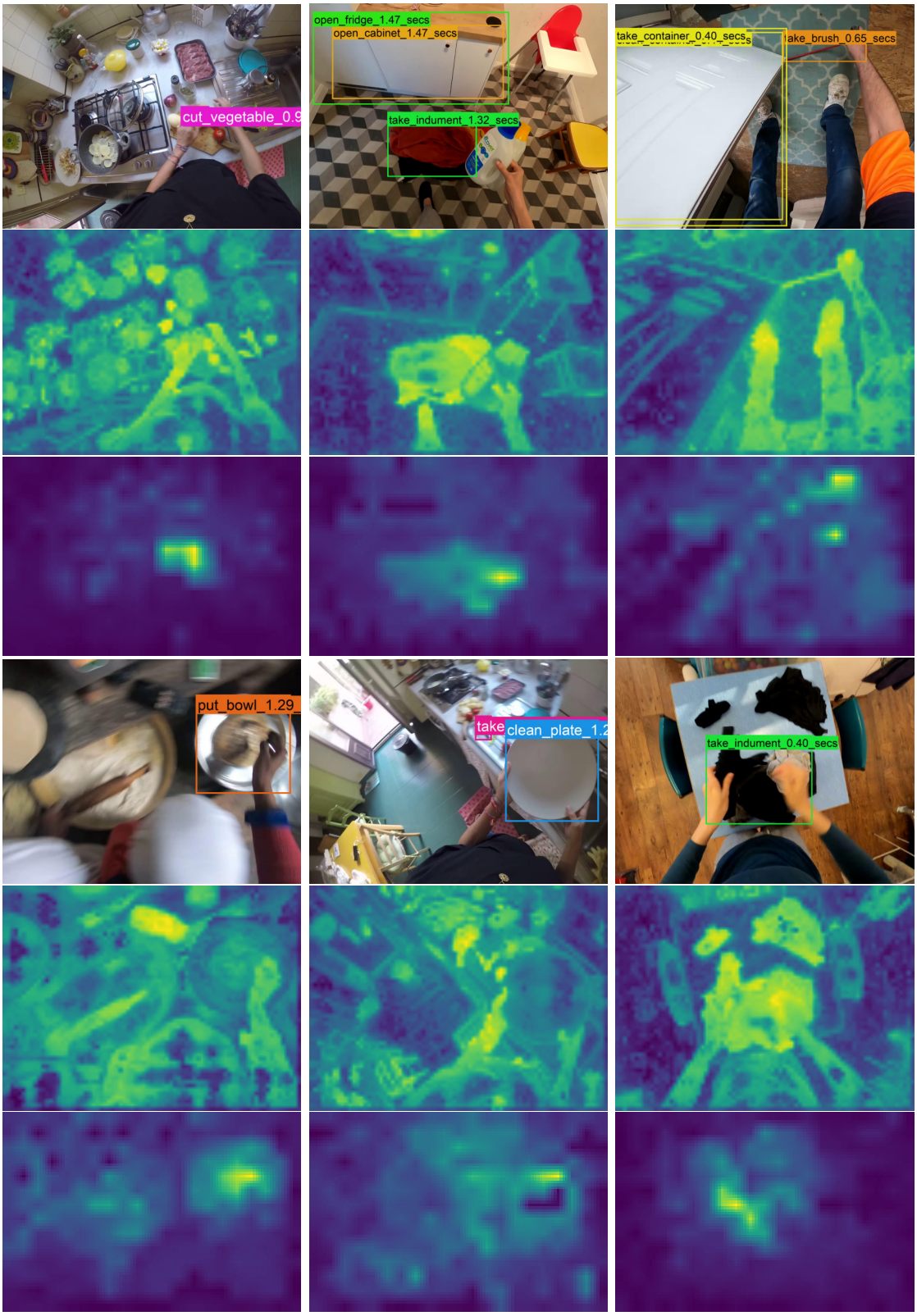


Figure 6.8: **Dual image-video attention maps, qualitative results.** Top to bottom: final predictions, attention map of pooled video tokens (queries) on image tokens (keys and values) and attention of image tokens (queries) on pooled video tokens (keys and values). Video tokens attend fine-grained object information from the high-resolution image; image features focus on objects which are important for future interactions.

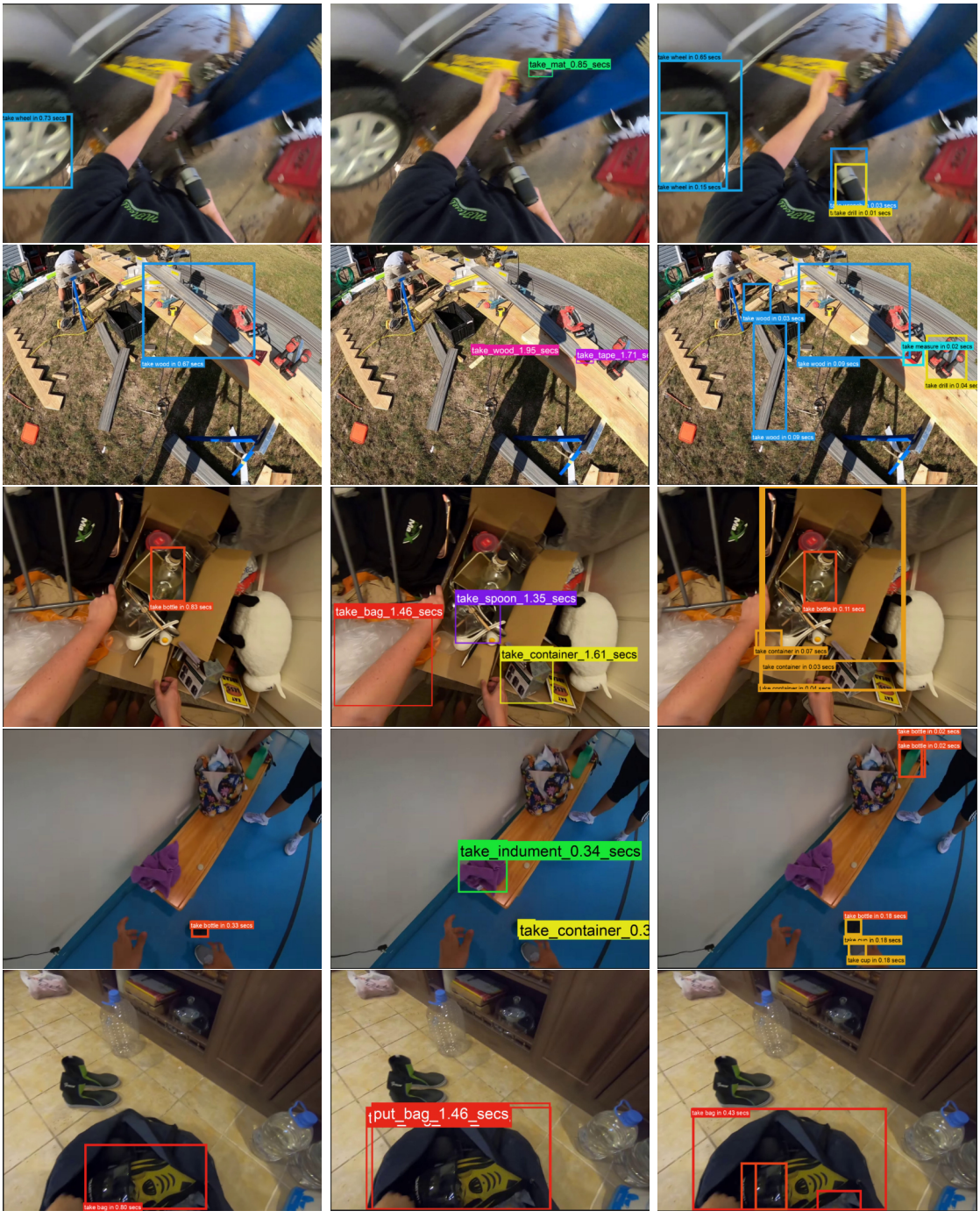


Figure 6.9: **Ego4D Qualitative results** Left to right: ground truth, STAformer predictions and STAformer++ predictions in Ego4D v2 validation split. We visualize the top-5 detections by the model. It is appreciated how the STAformer++ detections capture better the contour of the object, and that the whole model achieves a better understanding of the potential interactions in the video.

		N	N+V	N + TTC	All
Base model		16.20	7.47	4.94	2.48
Env. Aff.	Count-based priors	16.44	7.84	4.50	2.39
	Ego-Topo [39]	14.92	6.45	4.01	2.14
	Fixed Weighted (Ours)	18.44	8.46	5.47	2.85
Int. Hot.	Center Prior	14.44	6.86	3.90	2.05
	Hands Proximity	13.86	6.15	3.71	1.86
	Ours	17.82	7.62	5.05	2.53
Both (Ours)		**19.34**	**8.58**	**5.55**	**2.95**
Gain		+19.3	+14.9	+12.4	+18.9

Table 6.9: **Comparative of the affordances priors effect on Stillfast.** Results in mAP on the validation split of Ego4D v1.

		N	N+V	N + TTC	All
Base model		21.71	10.75	7.24	3.53
Env. Aff.	Count-based Prior	21.96	10.98	6.80	3.56
	Ego-Topo [39]	17.21	8.45	5.32	2.64
	Fixed Weighted (Ours)	23.55	11.75	7.55	3.74
Int. Hot.	Center Prior	17.70	8.82	5.22	2.62
	Hands Proximity	16.35	7.91	4.49	2.30
	Ours	23.63	11.38	7.51	3.66
Both (Ours)		**24.36**	**12.00**	**7.66**	**3.77**
Gain		+12.2	+11.6	+5.8	+6.8

Table 6.10: **Comparative of the affordances priors effect on STAformer.** Results in mAP on the validation split of Ego4D v1.

Environment Affordances on Stillfast (Table 6.9) and STAformer (Table 6.10). We first evaluate a naive Count-based Prior, re-weighting nouns and verbs probabilities by their frequency in the training dataset. While it slightly improves some metrics, it highlights the need to relate test samples to the specific scene's affordances. Training a NN classifier as in [39] does not produce a useful distribution of the affordances for fusion with STA probabilities. Our intuition is that the NN overfits to the interactions in the scene which are more obvious, losing the generalist quality of our predictions across environments. Our Fixed Env.Aff approach significantly refines nouns and verbs probabilities, obtaining consistent gains in $N + V$ Top-5 mAP (8.45 vs. 7.47 in Stillfast and 11.75 vs. 10.75 in STAformer). Figure 6.7 visualizes the fixed noun and verb affordance distribution for two query videos, showing the closest zones in appearance and narration. Although the ground truth STA class is not top-ranked, it appears in both predicted verb and noun affordances, supporting the observation that similar scenes afford similar interactions.

Interaction Hotspots on Stillfast (Table 6.9) and STAformer (Table 6.10).

Env.Aff	N	N+V	N+ttc	All
Base model	32.07	15.00	8.53	4.31
Fixed Uniform	29.95	14.15	8.56	3.70
Fixed Weighted	31.00	14.17	8.69	3.84
Learned	**33.21**	**15.94**	**8.98**	**4.66**
Gain	+3.5	+6.3	+5.3	+8.1

Table 6.11: **Comparative of the environment affordances effect on the STAformer++ model.** Results in mAP on the validation split of Ego4D v1.

We start evaluating the interaction hotspots with simple spatial priors. A center prior, that benefits bounding box predictions in the center of the scene, is detrimental to performance due to the complexity of egocentric video in which the objects appearing in the peripheral areas can be interacted with in the future. Similarly, re-weighting based on the current hands location with respect to the object proves ineffective, highlighting the importance of explicitly modeling future hand motion to predict the next interacted objects. Re-weighing confidence scores based on the spatial prior provided by the interaction hotspots produces a general improvement in all the metrics (e.g., N mAP of 17.82 vs 16.20 in StillFast and 23.63 vs 21.71 in STAformer - mAP All of 2.53 vs 2.48 in StillFast and 3.66 vs 2.53 in STAformer) by accounting for future interaction locations. Combining environment affordances and hotspots brings significant improvements in both StillFast and STAformer. For instance, STAformer improves N mAP from 21.72 to 24.36 and All mAP from 3.53 to 3.77.

Learned vs. Fixed Environment Affordances in STAformer++. We compare our approaches for grounding environment affordances in the STA task in Table 6.11. Learning affordances during training shows consistent gains in all the metrics, from 32.07 to 33.21 N mAP, 15.00 to 15.94 N+V mAP, 8.53 to 8.98 N+δ mAP and 4.31 to 4.66 All mAP. At higher performance levels, the use of a fixed distribution results in degradation, demonstrating the significance of a flexible and adaptive affordance representation for refining the probabilities.

6.8.4 Qualitative results

Figure 6.8 reports attention maps produced within the dual image-video attention module and final predictions (top). Video tokens attend fine-grained object information in the high-resolution image (middle), while image tokens attend scene dynamics in video features, which correspond to regions important for future interactions, such as moving hands or objects (bottom). We illustrate in Figure 6.9 a qualitative comparative on the Ego4D dataset between our two proposed models: STAformer and STAFormer++. The results show qualitatively the improvements achieved our novel architecture ver-

sion. First, the detected bounding boxes delimit significantly better the objects contour (i.e, the "wood" in the second column or the "bag" in the final example). Next, they predict more correctly the semantic class of the detected objects (i.e, "tape" in the second, or "container" in the fourth column). Finally, STAformer++ captures better the action dynamics and offers more plausible next-interactions according to the scene context, as the two first examples show.

6.9 Conclusions

In this chapter, we addressed the problem of Short-Term object-interaction Anticipation (STA). We proposed two novel architectures for STA. First, STAformer leveraged transformer models for feature extraction, and introduces novel components for image-video fusion, as the frame-guided temporal pooling or the dual cross-attention. Next, with STAformer++ we further improve performance by adopting a DETR prediction head. Our work also explores the contributions of environment affordances and interaction hotspots for refining the probabilities of STA models. We first propose a fixed representation which we exploit at inference with late fusion. A second approach enables STAformer++ to learn to extract the similarity of the current video with a memory of the past interactions in order to extrapolate the affordances distribution. Our results showcase the improvements given by the proposed architecture and affordance modules, which scores first on all splits of the challenging Ego4D and EPIC-Kitchens benchmarks. We also detailed the contribution of each individual component through ablations and showed that the integration of affordances is beneficial also to other STA architecture besides the proposed one.

Part IV

Exploiting multi-modal cues

Chapter 7

Temporal Video Segmentation with Natural Language using Text-Video Cross Attention and Bayesian Order-priors

So far, I have proposed visual models of objects, environments, and affordances that learn by exclusively observing videos or images from a first-person view. However, aligning egocentric video with natural language is a crucial step toward enabling autonomous agents to communicate more naturally. This chapter focuses on the step grounding task, which involves localizing the temporal boundaries of specific activities—described in free-form natural language—within long, untrimmed egocentric videos. While this task benefits from the flexibility of open-vocabulary language descriptions, it also requires handling extremely long video sequences with a long-tail distribution of step durations. To address these challenges, I introduce Bayesian-VSLNet, a test-time refinement strategy that incorporates temporal-order priors into the predictions. This simple yet effective refinement leverages Bayes' theorem to align predictions with the sequential order of steps, correcting common issues such as step repetition and cyclic actions.

7.1 Introduction

The proliferation of wearable and mobile devices and the growing availability of assistive robotic platforms present plenty of opportunities to develop and integrate assistive technologies into users' daily lives. Many of these innovative applications are enabled by video perception systems, making understanding video content a crucial task for

these domains. For example, wearable video devices are opening up new possibilities for health and safety among other fields, including applications such as assistance to impaired people with supermarket shopping [300], or detection of mistakes in procedural egocentric videos [301]. These kinds of applications require detailed video understanding. The temporal grounding of events of interest, often described by open language descriptions, is a challenging task towards this goal.

Understanding video demonstrations and aligning video content with natural language descriptions is also an essential perception task for robot learning [302]. In particular, imitation learning and reinforcement learning using video demonstrations [303–306] show great potential in robotic applications, as videos can be easily reproduced multiple times with minimal cost and even transferred between robots. However, most related work is focused on learning using short videos and simple tasks, splitting complex behaviors into simple steps or single actions that can be trained individually [307]. This prior splitting limits the rich information that can be extracted from videos, and hinders the performance for complex plans. Autonomous robots, operating in real-world environments, gather vast amounts of visual data throughout the day [1, 308] and numerous video demonstrations can be easily found on Internet [303].

In this context of vast amounts of unlabeled video data, the Step Grounding (SG) [3] task is crucial. The objective of this task is to localize the temporal boundaries of activities, described in free-form natural language, within long and untrimmed videos. Assistive devices require strong episodic memory capabilities [26], in order to identify the location of certain objects in a full video, discarding multiple irrelevant frames and focusing on certain actions. In robotics, it is usually needed to decompose complex tasks -such as object manipulation [309] or household chores [310]- into manageable steps, facilitating decision-making [311] and execution, or easing a posterior imitation learning [87]. Figure 7.1 illustrates an example of the proposed approach applied to video captured by a robot. To identify the specific video segments corresponding to the task of *washing clothes*, the activity is described through simple natural language instructions, such as *"Prepare the soap"* or *"Put the washed clothes in the dryer"*.

Step grounding addresses two key challenges that are critical for real-world applications (Figure 7.1). First, unlike traditional temporal action segmentation methods that rely on a fixed set of action labels [88–93], SG introduces flexibility by using natural language descriptions to identify actions. Second, the SG task handles long, untrimmed videos, enabling assistive vision devices and autonomous robots to manage large-scale visual data and prolonged sequences. In contrast, previous works typically focus on short clips of only a few minutes [34, 93, 312, 313]. The extended duration of videos intensifies the "needle in a haystack" problem, where irrelevant frames interfere

with the precise alignment between the video content and the textual query, leading to a loss of contextual detail.

In this chapter, we propose Bayesian-VSLNet, a method that addresses the challenges presented in SG. This approach first extracts a video-text feature representation for each processed video and the text query using Video Language Pre-trained (VLP) models, which are specially trained via contrastive learning to align both modalities in a common feature space. Bayesian-VSLNet extends VSLNet [314], a video span localizing network, with a novel training strategy that groups all the identical text queries of a video, and a head that predicts the binary probability of each video segment representing the text query. Our key contribution of this chapter is a test-time refinement strategy that integrates temporal-order priors into the predictions. This simple yet effective refinement is guided by Bayes' theorem, aligning predictions according to the sequential order of steps. A enhanced model, Bayesian-VSLNet++, processes video sequences spanning a few hours, allowing robots to interpret longer and more intricate action sequences. This approach is evaluated on the Ego4D Goal Step dataset [3], which comprises videos of procedural activities. This dataset presents additional challenges, such as cyclic actions, a long-tail distribution of step durations, and very long videos lasting up to 5 hours.

All things considered, our contributions in this chapter are:

- We design a novel Bayesian temporal-order method that adjusts for cyclic and repetitive actions, refining predictions during inference. This approach achieves state-of-the-art results on the step grounding task at the Ego4D 2024 challenge. We also include an extensive experimental analysis of the approach and its design choices.

- We propose two novel temporal grounding metrics that measure different aspects related to the step grounding task.

- We show qualitative results in household assistive scenarios demonstrating its applicability to real-world robotic and assistive use cases.

The work introduced in this Chapter won the Step Grounding Ego4D challenge at the EgoVis Workshop during CVPR 2024. Next, a follow-up journal version is currently under review.

- Carlos Plou*, Lorenzo Mur-Labadia*, Ruben Martinez-Cantin, and Ana C Murillo. CarLor@ Ego4D Step Grounding Challenge: Bayesian temporal-order priors for test time refinement. Winner Solution at Ego4D Step Grounding challenge during EgoVis Workshop CVPR 2024 *arXiv preprint arXiv:2406.09575*, 2024.

Figure 7.1: **Step Grounding task:** localize the segment in a long untrimmed video that represents the free-form natural language description of the step. The example represents the SG task (step 7 and step 12) along a video captured by an autonomous robot performing household chores.

– Carlos Plou*, Lorenzo Mur-Labadia*, Ruben Martinez-Cantin, and Ana C Murillo. Temporal Video Segmentation with Natural Language using Text-Video Cross Attention and Bayesian Order-priors. In *Recent Advances in Assistive Computer Vision and Robotics Special Issue of Computer Vision and Image Understanding Journal*, (Under Review)

7.2 Related works

7.2.1 Egocentric Video Understanding.

Egocentric (first-person) vision offers a unique perspective for understanding human behavior, as it closely captures fine-grained hand–object interaction details. The arrival of large-scale datasets such as Ego4D [26] and Epic-Kitchens [25] has driven progress in action recognition [34, 35], action anticipation [232, 270], affordance segmentation [38, 39], and episodic memory [40, 41]. These advancements in perception capabilities have led to various applications in assistive technologies. For example, [300] present an augmented reality system to guide individuals with impairments during supermarket shopping; [315] assist industrial operators in retrieving information on tools, equipment, and safety procedures; [301] detect mistakes in procedural egocentric videos; [316] combine GPS with visual information to guide users in urban environments; and [317] provide step-by-step assistance in instructional videos through visual affordances.

The unique perspective of egocentric vision for capturing interactions has also converted it into promising way to scale up learning in robotics [21, 318]. For exam-

ple, [62] learn interaction regions and afforded grasps from attending the hands movements, [319] leverage egocentric YouTube videos to learn navigation policies, while [320] fine-tune video captioning models on egocentric data to enable high-level reasoning over long-horizon tasks. However, the egocentric video modality introduces additional challenges, such as the sensor gap and the need for models to process extremely long recordings with a low density of informative frames. In this chapter, we introduce a test-time refinement strategy that incorporates the expected step order, improving the model's ability to handle long videos involving multiple action steps.

7.2.2 Video in robotics

Automatically understanding video content is a fundamental perception task in robotics, enabling a wide range of applications such as aerial drone action recognition [321,322], forecasting in autonomous driving [323,324], and real-world surveillance [325]. By leveraging video data, robotic systems can track objects with higher spatial and temporal resolution, facilitating more accurate navigation in complex environments [326–330]. Moreover, video provides fine-grained features and temporally consistent representations, helping to compensate for the lack of visual detail in texture-less modalities such as point clouds [331, 332]. For instance, [331] employed a video-language model to address the texture deficiency in 4D point clouds, effectively aligning both modalities to improve action recognition. Moreover, video is frequently employed in reinforcement learning for robotic manipulation [303, 304, 307], as it provides rich cues about interactions and the sequential structure of complex tasks. [307] combined data collected through robot interactions with observed videos of the same tasks to learn more effective control policies.

7.2.3 Bayesian statistics for video understanding

Bayesian statistics has also been applied to video understanding. [333] incorporated a Bayesian dropout-based variational layer into an audio-visual activity classifier to capture epistemic uncertainty in predictions and [334] introduced a teacher-student Bayesian evidential deep learning model for uncertainty quantification in online action detection. [335] incorporated Laplace ensembles into the final layer of a ViT to reduce overconfidence in action recognition. Rather than explicitly computing uncertainty, other approaches leverage Bayes' theorem to encode prior knowledge of step sequences, thereby enhancing robustness. [336] introduced a hierarchical Bayesian model that captures the realization of repeated procedural actions. Similarly, [337, 338] exploited the temporal ordering of action steps to learn more robust video representations. In

this chapter, we propose a Bayesian temporal order prior that adjusts for cyclic and repetitive actions, widely present in long, untrimmed videos.

7.2.4 Video Temporal Segmentation

From all the tasks related to video understanding, our work is focused on temporal segmentation, which aims to predict the start and end of actions. Initially, the Temporal Action Localization (TAL) task assumed a fixed and closed set of action labels [88–93]. ActionFormer [88] predicts actions boundaries using a multi-scale feature representation processed by a self-attention based transformer. [90] introduced a weakly supervised technique that achieved comparable SOTA results using less than 1% of the fully supervised labels. The precise temporal segmentation of the actions has facilitated human-robot interaction in various scenarios, such as surgical cooperation with robots [339–341], assistance in daily living tasks [342] and cooking [343]. [344] leverage the sequential nature of activities to guide a clustering algorithm in an unsupervised approach. However, TAL approaches are constrained by the set of action labels used during training, which limits its development in real-world scenarios with a higher variety of actions and object semantics.

The arrival of Video-Language Pre-training (VLP) methods [6, 7, 345–347] allows the opportunity of more complicated challenges, like Natural Language Queries (NLQ) task [26]. It consists of identifying the moment in a video that answers a text query. VLP methods [345–347] learn transferable representations from a large-scale training on pairs of video and the respective text narration, using a contrastive learning objective that aligns both modalities. EgoVLPv2 [6] incorporates a cross-modal fusion mechanism inside the video and text backbones, learning stronger video-text representation. EgoVideo [7] collects a massive egocentric and exocentric video-text dataset, obtaining the SOTA in a wide variety of tasks. NLQ approaches [234, 311, 348–350] leverage VLP models to propose multiple solutions. GroundNLQ [351] incorporated a text-aware temporal pyramid to capture temporal intervals of varying lengths. [348] proposed a sliding window technique to pre-filter candidate windows, preserving temporal resolution. EgoEnv [234] contextualizes videos within their 3D physical environment, rather than relying on naive temporal feature aggregation. [350] transforms the common video-text narrations into training data for NLQ, substantially boosting the performance across top models. Lastly, [352] introduces a novel clip selection approach to search the core clip iteratively. Specifically, we build upon VSLNet [353], the gold-standard neural network for open-vocabulary video localization which proposes a video span localization network that employs a shared feature encoder followed by context-query attention to learn cross-modal features.

7.3 Background

In this chapter, we focus on the step grounding task [3], which aims to identify the temporal clip (start_time, end_time) corresponding to a given free-form language description, representing a step within a procedural task. Our proposed method, Bayesian-VSLNet, extends the VSLNet architecture [314,353], the gold-standard neural network for natural language video localization, designed to localize a single query within a video. We adopt the VSLNet baseline due to its suitability for our target applications (robotics and assistive vision devices), offering a lightweight architecture compatible with embedded devices and enabling real-time inference with an average processing time of 125 ms per text query.

7.3.1 Problem definition

Given a set of long videos of procedural tasks, let $\mathcal{V}$ represent one of these videos. Assume that $\mathcal{V}$ has a duration of D seconds and contains a process of n steps. Each video $\mathcal{V}$ is linked with a text description of the process in the form of a set of natural language descriptions $\mathcal{T} = \{t_j\}_{j=1}^{n}$, where each t_j describes a specific step of the procedural task. Thus, our goal is, given a video $\mathcal{V}$ and a text query -i.e. natural language description of a step- t_j, to predict its starting and ending times (s_j, e_j) inside the video. We will refer to this time interval as a step clip.

7.3.2 VSLNet

Given a video $\mathcal{V}$ and a text query t_j, VSLNet method starts pre-extracting the video and text representations through a video encoder (Omnivore-L [4]) and a text encoder (BERT [5]), respectively. Afterwards, for long videos, it adopts a sparse sampling technique for compressing the video features. This technique involves dividing the video into K uniform segments and averaging the video features within each segment. This transformation makes the problem more efficient by converting each step clips (s_j, e_j) into a discrete representation $(k_j^s, k_j^e) \in [0, K]$,

$$k_j^s = \lfloor \frac{s_j \cdot K}{D} \rfloor, \qquad k_j^e = \lfloor \frac{e_j \cdot K}{D} \rfloor. \tag{7.1}$$

The pre-extracted visual and text representations are processed with a shared feature encoder formed by four convolutional layers followed by multi-head attention. Then, a context-query attention extracts cross-modal interaction features between the refined visual and text representations. This obtains a feature per segment with the content of the video refined with its significance in the query. The original VSLNet's

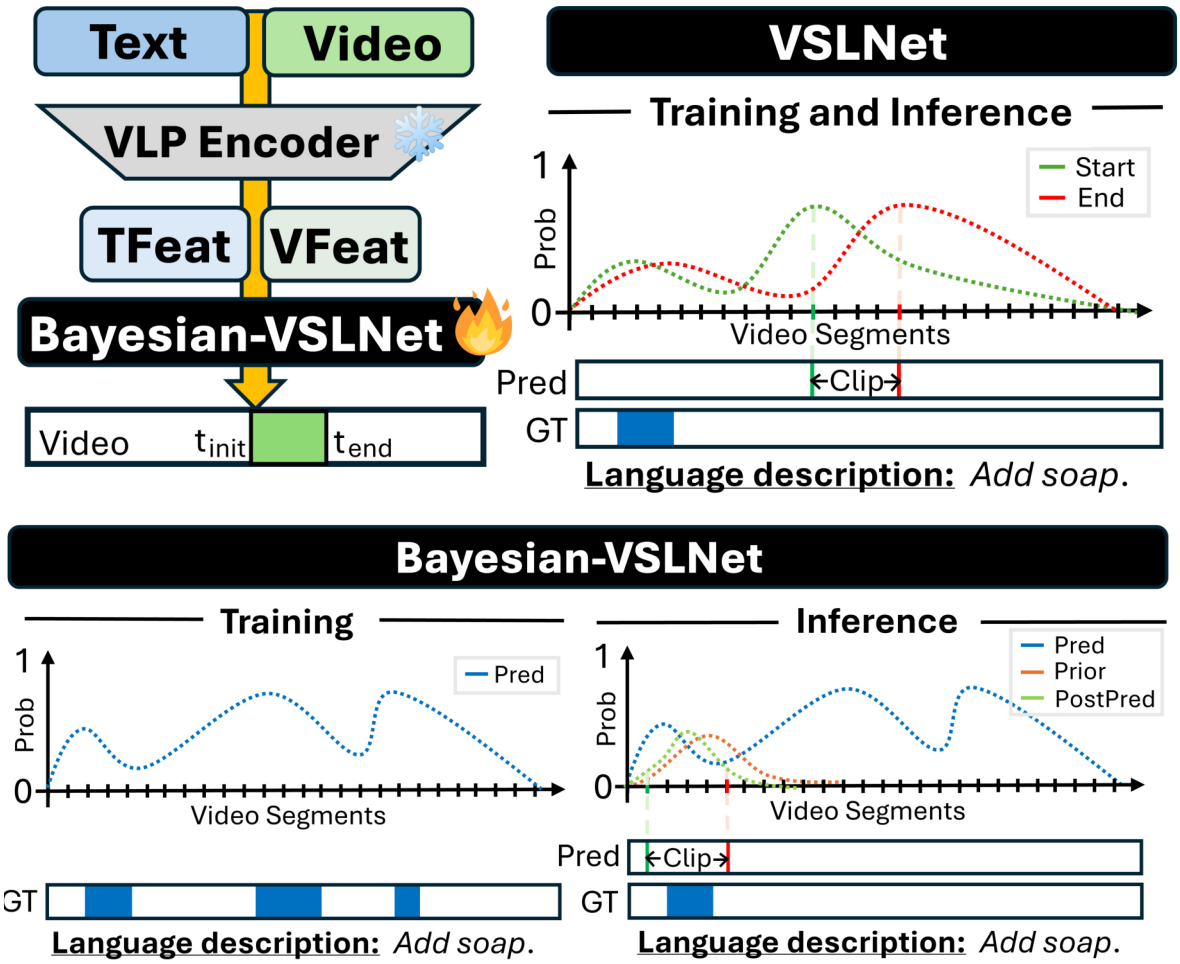


Figure 7.2: **Bayesian-VSLNet.** (Left) Our architecture is an extension of VSLNet with two novel components: a novel head predicts the probability of the text query in each video segment and a Bayesian temporar-order prior refines the predictions during the inference stage. (Center) VSLNet predicts each step independently, producing a prediction probability for each segment of the video, resulting in inconsistent results for long videos or descriptions with multiple steps. (Right) Bayesian-VSL use a step prior based on the order of the sequence of steps which improves the accuracy for long videos and guarantees consistency in the process description. During training, a step description might be repeated multiple times (see training GT) which confuses VSLNet. However, at inference time we want to segment the video to the exact occurrence of that step where the ordering prior plays a fundamental role.

head outputs two probability vectors $\hat{\mathbf{p}}_\mathbf{j}^\mathbf{s}, \hat{\mathbf{p}}_\mathbf{j}^\mathbf{e} \in [0,1]^K$ that estimate the likelihood of starting and ending the step described by the text query t_j in each of the K segments, as Figure 7.2 shows. After crossing both vectors $\hat{\mathbf{p}}_\mathbf{j}^\mathbf{s}$ and $\hat{\mathbf{p}}_\mathbf{j}^\mathbf{e}$, VSLNet predicts the starting and the ending segment $(\hat{k}_j^s, \hat{k}_j^e)$, which will result in the final predicted step clip $(\hat{s}_j, \hat{e}_j)$.

However, VSLNet [314] handles each step of the video independently, making it challenging to model scenarios where different steps share the same text description,

such as repeated actions within a procedural task. This is particularly problematic during training: when identical text queries appear, the model treats each instance separately, potentially leading to contradictions between iterations. Besides, during inference, all step clips within a video that share the same natural language description are assigned identical predictions, forgetting about the temporal order of the steps.

7.4 Method

Our method, Bayesian-VSLNet, extends the VSLNet architecture which fails to address the two key challenges of the step grounding task: cyclic actions and long videos.

1. The procedural nature of Ego4D videos involves repeated and cyclic actions, which complicate training under VSLNet's original formulation [314] as we mentioned in Section 7.3.2. As illustrated in Figure 7.2, VSLNet partitions the video into segments and estimates the probability of each segment marking the start and end of an action, which fails when there are repeated steps with the same text description. In contrast, our Bayesian-VSLNet directly predicts, for each video segment, the probability of being associated with the text query, mitigating the adverse effects of repeated actions.

2. The extended duration of the videos dilutes fine-grained correspondences and significantly increases the number of potential moment candidates. To address this, we introduce a novel test-time refinement strategy that leverages Bayes' rule, incorporating temporal-order priors into our predictions. These priors refine the output probabilities in order to accurately predict every moment when the step appears in the video and focus the prediction according to the step order in the sequence.

Therefore, while Bayesian-VSLNet preserves the efficiency of VSLNet architecture [314], it introduces some modifications in the training and inference procedures that allow it to tackle the main challenges of the task.

7.4.1 Text and Video Representations

Following previous work [348], we enhance our video representations by aggregating features from various video encoders [4, 6, 7]. Omnivore-L [4] is a multi-modal vision model trained on images, videos, depth maps and 3D data using supervised learning, enabling it to generalize across different visual input modalities and producing modality-invariant embeddings. Additionally, we employ EgoVideo [7] and the dual-encoder version of EgoVLP-v2 [6], which are egocentric VLP models. Both models are

trained via contrastive learning to align video and text embeddings, obtaining notable performance on episodic memory or action recognition tasks. Similar to VSLNet, we divide the video into K uniform segments and average the video features within each segment. In the same way as Eq. 7.1, we transform the ground truth step clips from time reference (s_j, e_j) into segments reference $(k_j^s, k_j^e) \in [0, K]$.

7.4.2 Bayesian-VSLNet

We detail the training and inference procedures of our novel Bayesian-VSLNet, which are depicted in Figure 7.2.

Training. Given a video $\mathcal{V}$ and a text query t_j, we start grouping all the step clips of the video whose natural language description is identical to t_j. Then, we build a ground truth "event" vector $\mathbf{p}_j = \{p_j^k\}_{k=1}^K \in \{0,1\}^K$ that contains all the segments of these grouped step clips. In mathematical terms,

$$p_j^k = \begin{cases} 1, \text{if } \exists t_q \in \mathcal{T} \text{ s.t. } \begin{array}{l} 1\ t_q = t_j \\ 2\ k_q^s \leq k \leq k_q^e \end{array} \\ 0, \text{otherwise} \end{cases} \tag{7.2}$$

We replace the VSLNet head with a single LSTM layer followed by a sigmoid layer, producing $\hat{\mathbf{p}}_j \in [0,1]^K$. The model is trained using Binary Cross-Entropy loss.

Inference. Given a video $\mathcal{V}$ and a text query t_j, we predict the starting and ending segment $(\hat{k}_j^s, \hat{k}_j^e)$ from the event probability distribution $\hat{\mathbf{p}}_j$. To achieve this goal, we search the most likely segment, that is, $k_j^* = \arg\max_k \hat{p}_j^k$ and, from it, we extend the clip forward $(k_j^* \rightarrow k)$ and backward $(k \leftarrow k_j^*)$ until $\hat{p}_j^k$ is under the α-percentile $\hat{p}_j^\alpha$, where α controls the segment amplitude. In other terms,

$$\begin{aligned} \hat{k}_j^s &: k \leq k_j^* \text{ s.t.} \qquad \hat{p}_j^{k-1} < \hat{p}_j^\alpha \leq \hat{p}_j^k, \\ \hat{k}_j^e &: k \geq k_j^* \text{ s.t.} \qquad \hat{p}_j^k \geq \hat{p}_j^\alpha > \hat{p}_j^{k+1}. \end{aligned} \tag{7.3}$$

However, let us assume that a given text description is repeated m times in a video. Each text query yields the same prediction $\hat{\mathbf{p}}_j$, resulting in a single clip $(\hat{k}_j^s, \hat{k}_j^e)$, and loosing the fact that the step occurs m times. To mitigate this drawback, we exploit Bayes rule, adding a temporal-order prior $\mathbf{q}_j = \{q_j^k\}_{k=1}^K$, defined as a Gaussian distribution:

$$q_j^k := \mathcal{N}\left(k; \frac{j \cdot K}{n}, K \cdot \beta\right), \tag{7.4}$$

where β controls the prior shape distribution and therefore its influence on the final posterior. Hence, we redefine our final predictions as

$$\hat{p}_j^k := \frac{\hat{p}_j^k \cdot q_j^k}{\max(\mathbf{q}_j)}, \tag{7.5}$$

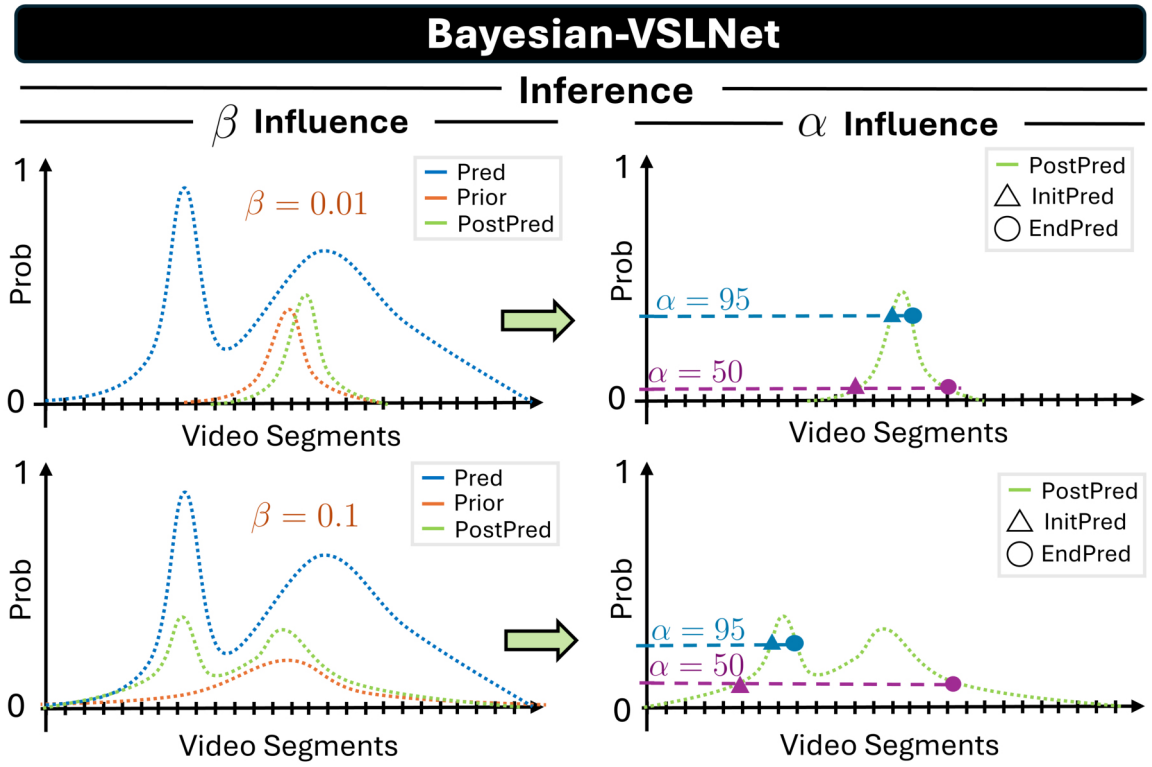


Figure 7.3: **Influence of the α and β hyper-parameters at the inference stage.** β determines the variance of the prior that controls the smoothness of the posterior. It can be seen as the weight that we give to the step ordering. Once we have the posterior, α sets the threshold (α-percentile of the posterior probability value p_j^k) that controls the length of the predicted clip and can be used to control the ratio of true positives and false positives.

which, as shown in Figure 7.2, results in a more accurate predicted step clip $(\hat{k}_j^s, \hat{k}_j^e)$. Figure 7.3 illustrates the roles of hyper-parameters α and β.

7.5 Experimental Setup

7.5.1 Dataset

We conduct our experiments on the Ego4D Goal-Step dataset [3], which comprises egocentric videos of procedural activities, allowing us to evaluate our model in real-world scenarios that require processing long video sequences [1]. The dataset consists of 368 hours of egocentric video footage, spanning 851 videos with durations ranging from 15 seconds to 5 hours, with an average length of 26 minutes. These videos exhibit a long-tail distribution of segment lengths, capturing procedural activities that vary from brief atomic actions lasting a few seconds to extended activities spanning several minutes. As shown in Figure 7.4, this variability necessitates both fine-grained and global video understanding. In total, the dataset contains 48K step annotations, densely labeled across the videos, averaging 56 annotations per video. Each annotation includes a time interval and a free-form natural language description of the ongoing action. Following previous work [3], we extract video features with a stride of 16 frames—equivalent to 1.875 features per second—using a sliding window of 32 frames.

7.5.2 Metrics

Temporal grounding tasks commonly use the Intersection over Union (IoU) to measure the temporal similarity between the predicted step clip $\hat{g}_j = (\hat{s}_j, \hat{e}_j)$ and the ground truth step clip $g_j = (s_j, e_j)$ for each text query t_j. We will refer to this metric as the IoU-IndOrder metric.

However, IoU-IndOrder does not account for cases where some step clips from the procedural video $\mathcal{V}$ share identical natural language descriptions. As a result, the predicted step clip $\hat{g}_j$ for a text query t_j could perfectly overlap with another step clip g_k that shares the same description ($t_k = t_j$), yet the IoU-IndOrder value would be zero. To address this limitation and provide more flexibility, we propose two alternative temporal grounding metrics: IoU-IndAll and IoU-Grouped. Next, we will provide a detailed explanation of these metrics.

• IoU-IndOrder: it computes the IoU between a predicted segment $\hat{g}_j$ and the ground truth segment g_j located in the same order position j.

$$\text{IoU} - \text{IndOrder}\left(\hat{g}_j, g_j\right) = \text{IoU}\left(\hat{g}_j, g_j\right). \tag{7.6}$$

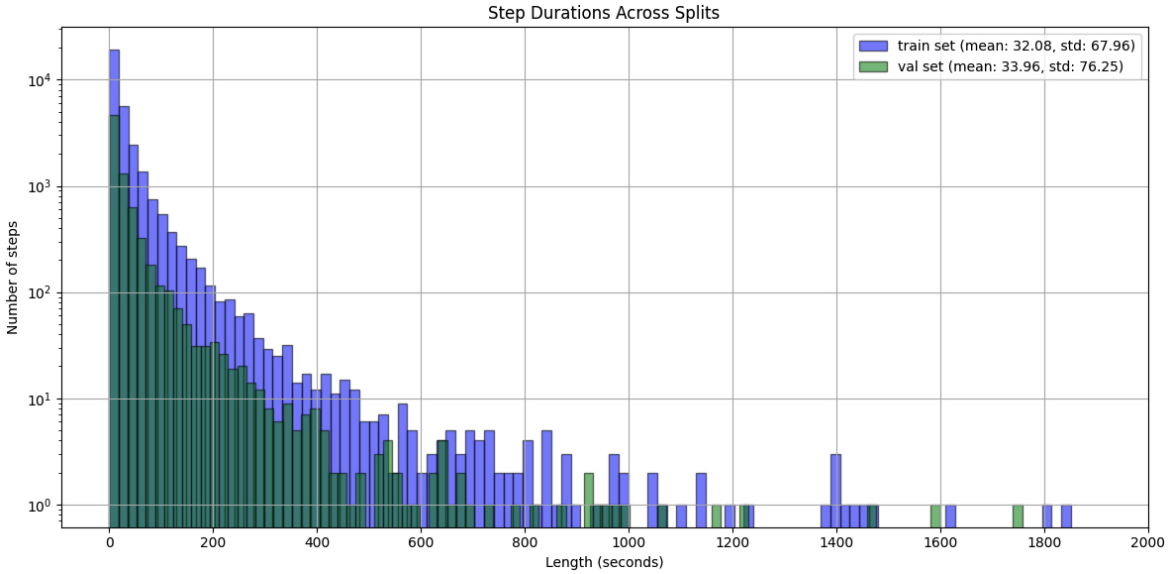


Figure 7.4: **Distribution of the step durations in the Ego4D Goal-Step dataset**.

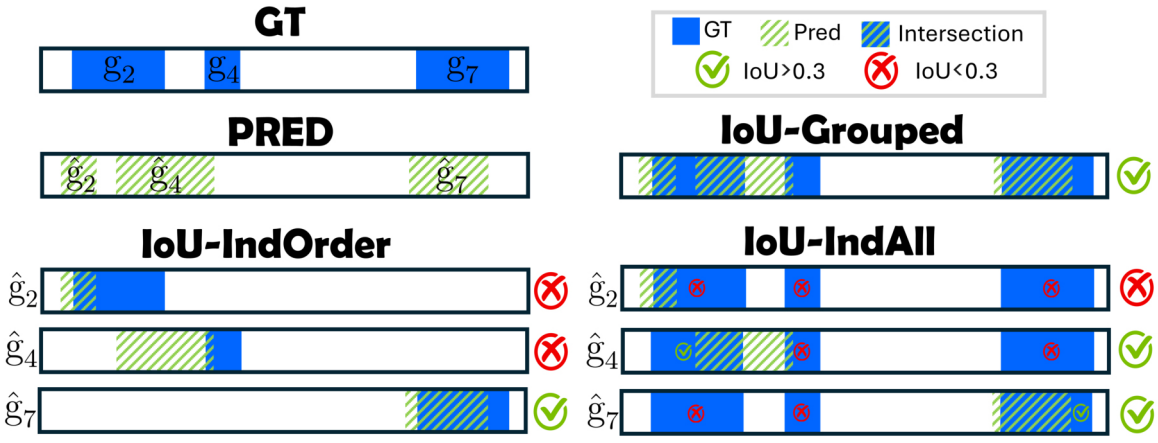


Figure 7.5: **Visualization of the metrics** (IoU-IndOrder, IoU-IndAll, IoU-Grouped) for a video and three step clips g_2, g_4, g_7 that share the same natural language description. Here R@1mIoU=0.3 values would be 33.3, 66.6, and 100, respectively.

• IoU-IndAll: For every predicted clip $\hat{g}_j$, this metric computes the IoU with each ground truth clip of the video $\mathcal{V}$ that has an identical text query, retaining the maximum IoU value obtained. In mathematical terms,

$$\text{IoU-IndAll}\left(\hat{g}_j, g_j\right) = \max_{q|t_q=t_j} \text{IoU}\left(\hat{g}_j, g_q\right). \tag{7.7}$$

• IoU-Grouped: This score calculates the IoU between the set of predicted step clips $\hat{G}_j = \{\hat{g}_q | t_q = t_j\}$ and the set of ground truth step clips $G_j = \{g_q | t_q = t_j\}$ whose text queries share the same natural language description.

$$\text{IoU-Grouped}\left(\hat{g}_j, g_j\right) = \text{IoU}\left(\hat{G}_j, G_j\right). \tag{7.8}$$

Specifically, Ego4D Goal-Step dataset leverages the IoU-IndOrder to report the Recall-at-one (R@1) for mIoU=0.3 and mIoU=0.5[1] was the metric used to rank the Ego4D 2024 challenge [3, 26], which measure the percentage of predicted clips that get an mean IoU value greater than 0.3 and 0.5, respectively. We show in Figure 7.5 the main differences among the three temporal grounding metrics when leveraging them to report the R@1,mIoU=0.3.

7.6 Results

In this section, we analyze the impact of the improvements introduced by Bayesian-VSLNet for the step grounding task. Specifically, we evaluate two configurations of Bayesian-VSLNet, both of which share the same architecture, training, and inference procedures described in Section 7.4.2. The key difference between these configurations lies in their video/text feature representations and hyperparameter settings.

The first configuration of our model, Bayesian-VSLNet-v0, utilizes video features from Omnivore-L [4] and EgoVLPv2 [6], with test-time refinement hyperparameters set to $\beta = 0.1$ and $\alpha = 90$. As Table 7.2 shows, this model won the Ego4D Step Grounding challenge at the CVPR 2024 EgoVis workshop. Next, we developed an improved version, called Bayesian-VSLNet++, which employs a more robust video representation composed of features from Omnivore-L [4], EgoVLPv2 [6], and EgoVideo [7], along with refined test-time hyperparameters. However, since the evaluation server was promptly closed, we report results for the Bayesian-VSLNet++ only on the validation set, as Table 7.1 shows. All experiments were conducted using two NVIDIA GeForce RTX 4090 GPUs. We employed the AdamW optimizer with a linear learning rate scheduler for training.

[1] R@1,mIoU=0.3

	Validation-R@1 mIoU					
	Grouped		IndAll		IndOrder	
Model Name	0.3	0.5	0.3	0.5	0.3	0.5
VSLNet (Baseline)	19.17	12.98	19.62	12.27	12.15	7.67
Bayesian-VSLNet-v0	24.28	12.70	24.83	12.01	18.15	8.97
Bayesian-VSLNet++	**30.43**	**18.71**	**32.48**	**19.62**	**23.75**	**14.41**

Table 7.1: **Results in the Ego4D Goal-Step validation set [3].** We measure R@1, mIoU=0.3 and R@1, mIoU=0.5 for each temporal grounding metric.

	Test-R@1mIoU- IndOrder	
Model Name	0.3	0.5
VSLNet (Baseline)	19.04	12.04
FlyFishing*	29.69	18.99
iLearn*	33.00	26.37
EgoVideo [7]	34.06	**26.97**
Bayesian-VSLNet-v0 (Ours)	**35.18**	20.48

Table 7.2: **Results for the different methods evaluated in the test set of the Ego4D Goal-Step dataset by the evaluator server as part of the Step Grounding challenge.** It measures R@1,mIoU=0.3 (primary metric) and R@1,mIoU=0.5 for the IoU-IndOrder metric.

7.6.1 Quantitative results

Table 7.1 shows that our proposed training and inference strategy (Section 7.4.2) enables Bayesian-VSLNet to significantly surpass the baseline VSLNet across all evaluated metrics. Notably, in the primary metric R@1 mIoU-IndOrder=0.3, Bayesian-VSLNet-v0 achieves a 49.5% improvement over the baseline (18.15 vs. 12.15 R@1 mIoU-IndOrder=0.3), while Bayesian-VSLNet++ exhibits an even more substantial gain, with a 95.5 % increase (23.75 vs. 12.15 R@1mIoU-IndOrder=0.3). These results highlight the effectiveness of incorporating Bayesian priors in tasks where specifying the exact step instance is crucial, particularly when steps are duplicated in the description. Additionally, the improvements remain consistent across other metrics (mIoU-Grouped and mIoU-IndAll), indicating that our approach achieves more robust video understanding while effectively leveraging the temporal priors of Bayesian-VSLNet.

Next, we report the results of the Ego4D Step Grounding CVPR 2024 EgoVis workshop in Table 7.2. In order to ensure fair competition, it featured an evaluation server with a withheld ground truth test set. Our best configuration up to that point, Bayesian-VSLNet-v0, achieved the state of the art by surpassing all other teams in the primary metric (35.18 R@1, mIoU- IndOrd=0.3) and winning the challenge. As the

Hyper.		Validation-R@1mIoU					
		Grouped		IndAll		IndOrd	
α	β	0.3	0.5	0.3	0.5	0.3	0.5
90	0.1	29.17	16.99	29.81	16.71	22.45	12.82
85	0.1	26.87	15.17	26.57	14.15	20.13	10.91
95	0.1	**30.43**	18.71	**32.48**	19.62	**23.75**	**14.41**
98	0.1	27.79	16.97	29.63	18.49	21.86	13.50
95	0.05	28.39	16.51	29.89	16.93	22.82	13.03
95	0.15	29.89	**18.81**	32.41	**20.11**	23.12	14.37

Table 7.3: **Ablation study of α and β hyper-parameters for the Bayesian-VSLNet model in the Ego4D Goal-Step validation set.** We measure R@1,mIoU=0.3 and R@1,mIoU=0.5 for each of the temporal grounding metrics. Gray row shows the α and β values used in the Bayesian-VSLNet++ configuration.

primary metric (R@1 mIoU-IndOrd=0.3) considers relevant the step order in repeated actions, the impact of our temporal priors results crucial, as it discards false positives outside the relevant video region.

7.6.2 Ablation studies

First, we provide an ablation study of both α and β hyper-parameters from Bayesian-VSLNet++ configuration in Table 7.3. The best results were obtained with a percentile threshold of $\alpha = 95$ and $\beta = 0.1$. The high α value ensures the selection of segments that are not excessively long, even after the smoothing effect introduced by an intermediate β covariance value in the prior, which reduces probability differences between consecutive segments.

Next, we evaluated the impact of encoding video and the step description across different models in Table 7.4. The combination of Omnivore-L with BERT features yields the lowest performance (15.20 R@1mIoU-IndOrder=0.3 and 7.32 mIoU=0.5) since both encoders are trained with data of their respective modality and consequently, they lack the fine-grained alignment necessary for effective step grounding. In contrast, using specialized video-language pre-trained models significantly improves performance. For instance, EgoVideo [7] achieves 19.90 R@1mIoU-IndOrder=0.3 and 11.29 mIoU=0.5, as it is trained via contrastive learning to align video and text embeddings, producing more representative features that enhance the subsequent temporal segmentation by our Bayesian-VSLNet head. Next, we combined video features from different models to obtain a more comprehensive video representation, leading to the best performance for both VSLNet (18.19 R@1mIoU-IndOrder=0.3 and 13.55 mIoU=0.5) and Bayesian-VSLNet (21.05 R@1mIoU-IndOrder=0.3 and 13.50 mIoU=0.5). These results further

Model		Video Features			Text Features			K	mIoU-IndOrd	
Name	Config	[4]	[6]	[7]	[5]	[6]	[7]		0.3	0.5
VSLNet		✓	-	-	✓	-	-	128	11.77	7.77
VSLNet		✓	✓	-	-	✓	-	128	13.13	8.75
VSLNet		✓	✓	-	-	✓	-	512	16.26	11.81
VSLNet		✓	✓	-	-	✓	-	1024	15.28	11.18
VSLNet		-	-	✓	-	-	✓	512	17.74	12.32
VSLNet		✓	✓	✓	-	-	✓	512	18.19	13.55
Bay-VSLNet		✓	-	-	✓	-	-	128	15.20	7.32
Bay.VSLNet	v0	✓	✓	-	-	✓	-	512	18.15	8.97
Bay.VSLNet		-	-	✓	-	-	✓	512	19.90	11.29
Bay.VSLNet		✓	✓	✓	-	-	✓	512	21.05	11.53
Bay.VSLNet	++	✓	✓	✓	-	-	✓	1024	**23.75**	**14.41**
Bay.VSLNet		✓	✓	✓	-	-	✓	2048	22.27	13.70

Table 7.4: **Ablation study about number of segments K and feature extractors in the Ego4D Goal-Step validation set for both VSLNet and Bayesian-VSLNet architectures.** We leverage Omnivore-L [4], BERT [5], Ego-VLP [6] and EgoVIDEO [7] features. We take R@1,mIoU-IndOrd at 0.3 and 0.5 as reference metric.

highlight the advantages of our Bayesian-VSLNet approach, which consistently outperforms its VSLNet counterpart, demonstrating its effectiveness in improving step grounding accuracy.

Further, we also present a comparison of the number of sampled segments in Table 7.4. As the results show, the optimal configuration, Bayesian-VSLNet++, is achieved with 1024 segments. This configuration reports 23.75 mIoU=0.3 and 14.41 mIoU=0.5. This is a good balance between the model complexity and the information loss due to the diffusion of individual video features in the segment.

7.6.3 Qualitative results

Figure 7.6 presents qualitative results from our best-performing configuration, Bayesian-VSLNet++, across five egocentric videos covering different activities, illustrating the impact of our novel temporal priors. The figure demonstrates how Bayesian-VSLNet predicts the probability of each step description corresponding to a given video segment. Notably, similar descriptions (e.g., *"Adds minced cocoa into milk?"* vs. *"Mix minced cocoa and milk together?"* in the first example) yield nearly identical predictions, highlighting the challenge of achieving precise segmentations. As shown in the fourth example, when a video contains multiple repeated procedural actions, an effective temporal prior becomes crucial for accurately segmenting steps. Our temporal ordering mechanism refines the network's probability predictions, aligning them with

the correct temporal sequence to enhance segmentation accuracy. However, the sparse-sampling technique poses a limitation, as it causes the model to struggle with short atomic actions due to the loss of fine-grained temporal information during sampling.

7.6.4 Qualitative results on assistive robotics data

We present qualitative results in a real-world assistive robotics scenario to demonstrate the potential of our approach in enhancing human-robot interaction in practical applications. Specifically, we used two near-egocentric videos from the Mobile Aloha project [1], each about 2 minutes long, featuring a robot performing household chores with approximately 30 and 15 steps, respectively. We queried the model with the free-form description of three different steps, without any prior fine-tuning on this data. Our approach achieves efficient execution, with an inference time of 125 ms per sample, ensuring real-time processing capabilities and supporting embedded applications as Figure 7.7 shows.

7.7 Conclusions

Video understanding is a key perception task for assistive applications, where robotic platforms or wearable devices frequently rely on video data to function. In this paper, we presented Bayesian-VSLNet to improve existing video understanding capabilities. Our novel approach for step grounding in long, untrimmed videos exploits video text cross attention and Bayesian temporal-order priors to significantly improve temporal segmentation of activities from natural language descriptions. Our approach addresses challenges in real-world applications, like repetitive and cyclic actions. Our main limitation comes in very long videos, where a sparse-sampling technique is not enough. We achieve state-of-the-art results on the Ego4D Goal-Step dataset, where this approach won the step grounding challenge at the CVPR 2024 EgoVis workshop, and demonstrate qualitative results on real-world robotics data, showing its potential for more natural interactions and robot learning in assistive applications.

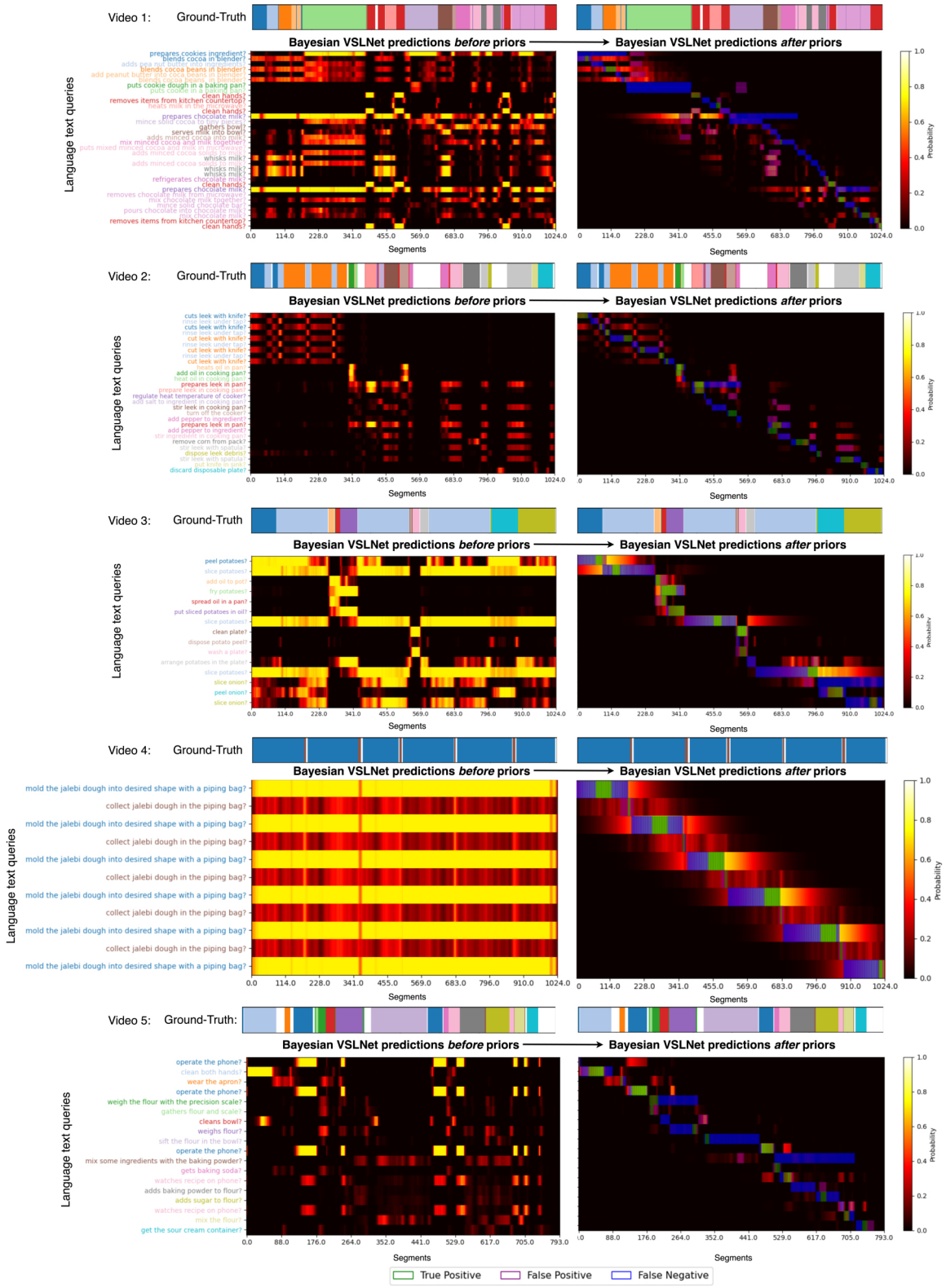


Figure 7.6: **Bayesian-VSLNet++ qualitative results on the Ego4D Goal-Step dataset.** The plots in the left column show the predicted probabilities by our Bayesian-VSLNet per each step description, while the plots in the right column display the refined probabilities after applying our temporal-order prior. The final predicted step clip is extracted from the refined probabilities. We report the true positive segments in green, the false positives in purple and the false negatives in blue.

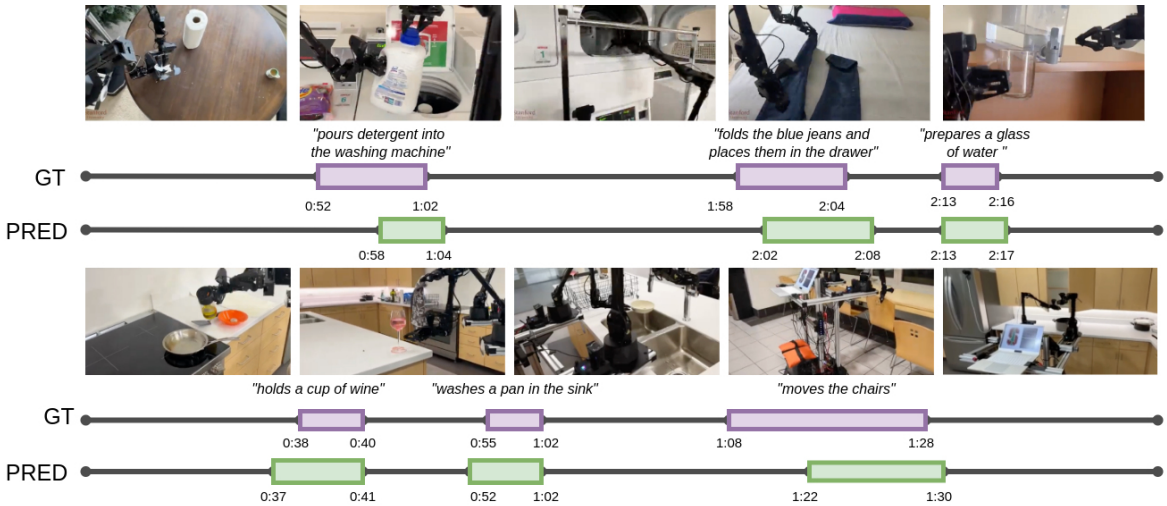


Figure 7.7: **Qualitative examples in real-world robotics scenarios.** Bayesian-VSLNet predicts with high precision the moment associated to the provided step description without fine-tuning on this type of data. Video sourced from from the Mobile Aloha project [1].

Chapter 8

O-MaMa: Learning Object Mask Matching between egocentric and exocentric views

Although the egocentric perspective offers multiple cues for understanding human behavior, relying exclusively on this viewpoint limits the overall scope of comprehension. Therefore, in this final chapter, I investigate how to complement the first-person view with the third-person (exocentric) perspective. Specifically, I address the Ego-Exo4D Correspondence task, which aims to predict an object's mask in one view given a query mask from the other. My proposed solution redefines the complex cross-image segmentation task by reformulating it as object mask matching across egocentric and exocentric views, leveraging the powerful zero-shot segmentation capabilities of Segment Anything Models (SAM). Object mask candidates are first extracted in the target view using FastSAM. Each candidate is then encoded with a Mask-Context Encoder that pools dense DINOv2 features, capturing both discriminative object details and surrounding contextual information. The model is trained using a Mask Matching Contrastive Loss, which enables effective cross-view object correspondence by aligning both global scene context and fine-grained object features.

8.1 Introduction

Nowadays, intelligent agents need to collaborate while performing cooperative tasks. This includes applications like multi-robot manipulation [354, 355], augmented reality assistants [356, 357], and human-robot collaboration [358, 359]. In robot learning, a robot could infer human dexterous manipulation skills by analyzing both third-person demonstrations and first-person execution. Similarly, AI tutors in online education

could enhance learning by aligning first-person demonstrations with third-person instructional videos, such as guiding a piano student's hand placement from multiple perspectives. Perception plays a crucial role in this scenario, but each agent typically has access only to its own sensors or cameras, and each one perceives the environment from a different perspective. Consequently, understanding object correspondences between egocentric (first-person) and exocentric (third-person) views is essential to align multi-agent perception and establish a shared basis for interaction. Despite the advances in segmentation [360–363] and object detection [12, 364] from single images, cross-view segmentation between egocentric and exocentric perspectives remains an open challenge.

This chapter addresses the Ego-Exo4D Correspondences task, which aims to predict an object's mask in one perspective given a query mask from the other. Unlike traditional segmentation problems, this task introduces additional challenges, including drastic viewpoint transformations, scale variations, occlusions, and domain shifts due to differences in camera optics and imaging conditions. While the exocentric view captures both the full environment and the person's body, it contains objects at multiple scales. Conversely, the egocentric view offers rich details on hand-object interactions, but it is highly dynamic, suffering from motion blur and frequent occlusions due to the ongoing interactions. These challenges make establishing precise object-level correspondences across views particularly difficult, requiring fine-grained segmentation and strong cross-view semantic reasoning.

We propose simplifying the complex cross-image segmentation task by reformulating it as Object Mask Matching (O-MaMa) across ego and exo views, leveraging the excellent zero-shot segmentation capabilities of Segment Anything Models (SAM) [17]. An overview of our proposed method is shown in Figure 8.1. First, we extract a set of object mask candidates in the destination view using FastSAM [95]. To obtain mask descriptors, each object mask is encoded with a Mask-Context Encoder, which pools dense DINOv2 [16] semantic features from both object masks and their extended bounding boxes, combining discriminative object features with contextual information. In cluttered scenes, nearby objects often share similar context while containing distinct object descriptors. Therefore, our proposed Hard Negative Adjacent Mining strategy selects neighboring object candidates to encourage the model to better differentiate between nearby objects. Next, cross-view global features are incorporated using a novel Ego↔Exo Cross Attention mechanism. Finally, our model is trained with a Mask Matching Contrastive Loss, which selects the best mask candidate in the destination view by learning a cross-view feature alignment that captures both global scene context and fine-grained object features.

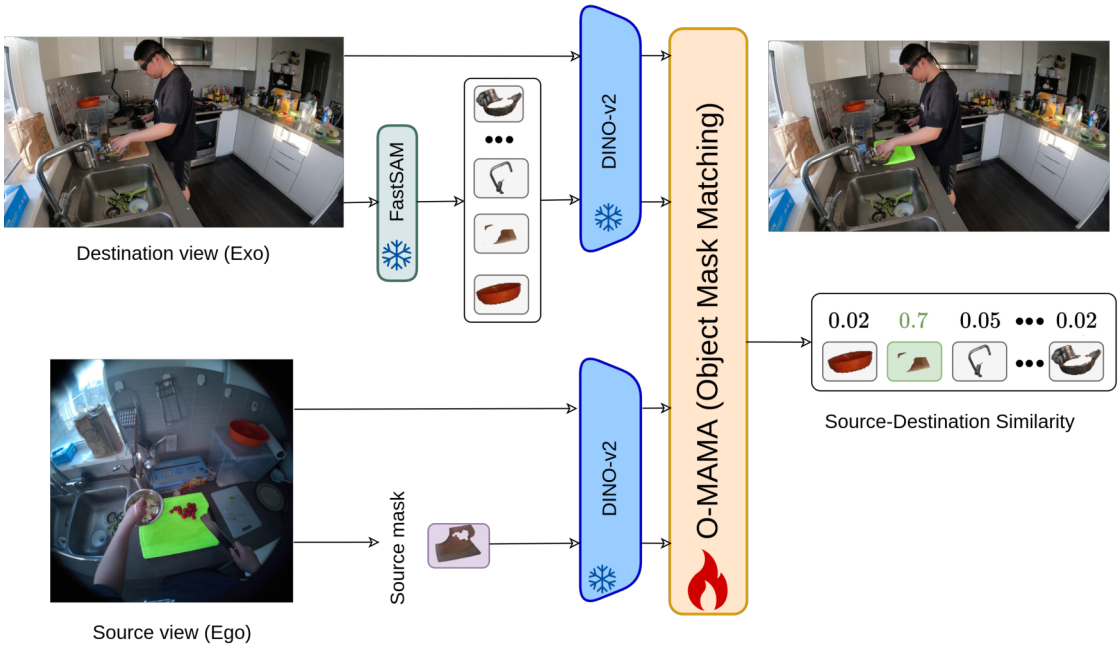


Figure 8.1: **Overview of the proposed Object Mask Matching (O-MaMa).** Instead of attempting the complex cross-view segmentation task, first FastSAM extracts a set of mask candidates in the destination view. Through contrastive learning, the mask candidate that best matches the source mask is selected.

Our proposed method is both simple and effective, achieving state of the art in the Ego-Exo4D Correspondences benchmark. O-MaMa obtains 45.8 (Ego2Exo) and 48.6 (Exo2Ego) IoU in the test v2 split, which represents a relative gain of +31.2% and +94.4% against the official challenge baselines [27], respectively. Moreover, in the validation v1 set, this method scores 50.1 and 54.2 IoU in the Ego2Exo and Exo2Ego scenarios, showing a +13.1 % and +6.5 % against the previous SOTA model [365] using only 1% of training parameters.

In summary, our contributions in this chapter are as follows:

1. We design the O-MaMa (Object Mask Matching) approach, which reformulates the complex cross-view segmentation Ego-Exo Correspondences task as a cross-view object mask matching.

2. We propose an Ego↔Exo Cross Attention mechanism that introduces cross-view global features in the object embedding and a Hard Negative Adjacent Mining that disambiguate between nearby but distinct objects with similar context. Our results show that our proposed integration of local and global information results in an object mask selection more sensitive to the cross-view relationship.

3. O-MaMa reports strong improvements compared with previous works, achieving the state of the art in the Ego-Exo Correspondences task, while requiring only 1

% of the trainable parameters compared to [365].

The results introduced in this Chapter won the Ego-Exo4D Correspondence Challenge at the EgoVis Workshop during CVPR 2025, and a extended version was presented during ICCV 2025.

– Lorenzo Mur-Labadia*, Maria Santos-Villafranca*, Jesus Bermúdez Cameo, Alejandro Perez Yus, Ruben Martinez-Cantin and Jose J. Guerrero. O-MaMa @ Ego-Exo4D Correspondence Challenge: Object Mask Matching between Egocentric and Exocentric Views. Winner solution at EgoExo4D Correspondences challenge during EgoVis Workshop CVPR 2025.

– Lorenzo Mur-Labadia*, Maria Santos-Villafranca*, Jesus Bermúdez Cameo, Alejandro Perez Yus, Ruben Martinez-Cantin and Jose J. Guerrero. O-MaMa: Learning Object Mask Matching between Egocentric and Exocentric Views. In *International Conference on Computer Vision*. Core Ranking A*, 2025.

8.2 Related Works

Ego-Exo understanding. Exocentric (third-person) vision has been extensively studied in action recognition [366,367], segmentation [368–370] and tracking [371,372]. Alternatively, egocentric vision offers a unique viewpoint for capturing human behavior. The arrival of large-scale datasets [25, 26] has driven progress in action recognition [34,35], action anticipation [232,270], affordance segmentation [38,39], and episodic memory [40, 41]. Since egocentric and exocentric views provide complementary visual cues of the scene, combining both perspectives has emerged as a promising direction for learning generalizable view-invariant representations. Some works [373, 374] improve model training in one perspective by leveraging data properties from the other view. Other works explore the benefits of learning cross-view invariant features for action recognition [375–377], affordance segmentation [63], activity progression [378], and temporal action segmentation [379]. In contrast, we propose learning view invariant features at object level to match masks across synchronized views, which requires fine-grained pixel-level predictions.

Learning correspondences. Traditionally, detector-based local feature matching methods first detect key-points in an image and then extract descriptors to establish correspondences following hand-crafted [380–382] or learning [297,383] approaches, using nearest neighbor search or an attentional graph neural network [384] to find matches between the extracted interest points. More recently, detector-free methods [2,385,386]

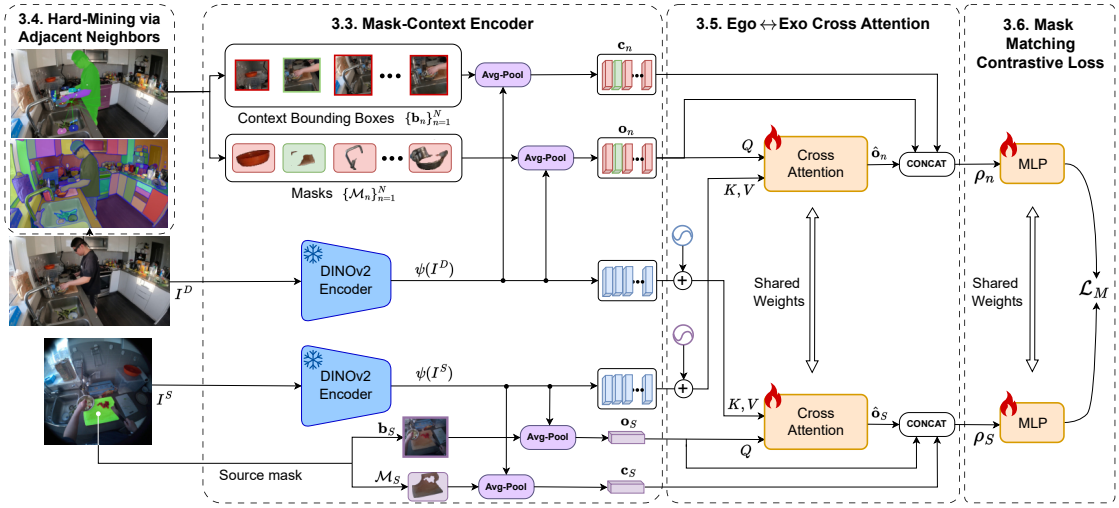


Figure 8.2: **O-MaMa architecture.** In the destination view, we generate a set of mask candidates with FastSAM. We extract descriptors on both source and destination masks by pooling dense DINOv2 features, and we aggregate global cross-view features with respective cross-attention mechanisms. We learn view-invariant features in a latent space via contrastive learning, and we select the most similar mask embedding to obtain the corresponding mask.

have gained attention for their ability to directly obtain dense feature matches without an explicit keypoint detection. For instance, RoMa [2] achieves robust feature matching under extreme changes in scale, viewpoint and illumination. While these methods establish correspondences between images of the same scene or object instances, semantic correspondence approaches [387–389] find dense matches between semantically similar images. For example, Zhang [389] fuses DINOv2, that provides sparse but accurate matches, with stable diffusion features, which contain high-quality spatial information. Here, we propose leveraging powerful semantic features from DINOv2 as well, but in a different context. Instead of keypoints or semantic parts, we establish correspondences between entire object masks across different perspectives.

Segmentation Models. Traditional approaches are categorized into semantic segmentation [360, 361], instance segmentation [12, 289, 364], panoptic segmentation [199, 390–392], and video object segmentation [393–395]. SAM [17] introduced the promptable segmentation task [95, 396–398], where fine-grained segmentation masks are generated from a spatial prompt (i.e, a point, or a bounding box). Recently, several works [399–401] have leveraged the input token flexibility of Large Language Models (LLMs) to support diverse input conditions and output formats.

Despite the significant advances in single-image segmentation, few works address object segmentation across multiple images. For instance, semi-supervised video object segmentation [393] methods segment an object along the video given its mask in

the initial frame [402–406]. Similarly, image co-segmentation models aim to identify common objects across different images [407–409]. Adapting LLMs to the cross-view segmentation, Object-Relator [365] fine-tuned PSALM [400] with a dedicated module that enforces view-invariant embeddings through a self-supervised alignment. In this chapter, we propose to reformulate the cross-view segmentation as a mask matching task, incorporating FastSAM [95] in our pipeline to leverage its fine-grained, zero-shot segmentation capabilities; showing that we can achieve state-of-the-art results while significantly reducing the number of trainable parameters.

8.3 Methods

8.3.1 Task Formulation

Given a pair of images from two different views (egocentric and exocentric) and a query object mask $\mathcal{M}_S$ in the source view I^S, the objective of the Ego-Exo Correspondences task is to predict the corresponding object mask in the destination view I^D. In the Ego2Exo task, the source view corresponds to the egocentric image, and the destination view is the exocentric image, and the opposite happens in the Exo2Ego setting. The task is restricted to use only visual information as input for the model, excluding semantic labels, object names, or camera pose information.

Additionally, the task involves significant viewpoint variation and challenges due to the inherent characteristics of each viewpoint: egocentric views suffer camera motion and occlusions due to the ongoing interactions, while the exocentric perspective contains objects at multiple scales distributed along all the scene. See Figure 8.3 for some examples of these limitations in the dataset. All these factors require combining both a very fine-grained segmentation capability and a global cross-view understanding to effectively locate the corresponding objects across views.

8.3.2 Method overview

We reformulate the challenging cross-view segmentation task as a cross-view Object Mask Matching (O-MaMa) problem. This novel approach exploits the high-quality zero-shot segmentation capabilities exhibited by SAM [17, 95] to simplify the segmentation problem. Figure 8.2 shows the architecture of our method. The goal of O-MaMa is to select the object mask candidate in the destination view that matches the source mask best. We first extract a set of object mask proposals in the destination view using FastSAM [95], from which we compute object-level and contextual descriptors with the Mask-Context Encoder (Section 8.3.3). We also include cross-view global

features through a novel Ego↔Exo Cross Attention mechanism (Section 8.3.5). We adopt a novel Mask Matching Contrastive Loss (Section 8.3.6) to learn view-invariant features by minimizing the distance between paired samples and maximizing the distance between negative (unpaired) samples in a shared latent space. During training, we enforce the model to learn more robust and discriminative object descriptors with a Hard-Mining via Adjacent Neighbors (Section 8.3.4).

8.3.3 Mask-Context Encoder

We first generate a set of dense mask proposals in the destination view using FastSAM [95], which segments intuitive regions of the scene, such as entire objects, object parts or surfaces, with comparable quality to SAM [17], while being 50× faster. Specifically, from the destination image, I^D, we generate N mask candidates $\{\mathcal{M}_n\}_{n=1}^{N}$.

Then, we compute a descriptor of each mask segment. We leverage DINOv2 [16], a self-supervised learning model, due to its high-level semantics, object decomposition capabilities and dense feature localization properties. We extract a local object descriptor $\mathbf{o}_n = \text{Avg-Pool}(\mathcal{M}_n, \psi(I^D))$ by pooling the corresponding object mask from the DINOv2 feature map of the destination image, denoted as $\psi(I^D)$. Some studies suggest [410–412] that humans leverage visual contextual associations among objects to represent scenes. Inspired by this, we extract a context descriptor $\mathbf{c}_n = \text{Avg-Pool}(\mathbf{b}_n, \psi(I^D))$ by pooling features from an extended bounding box $\mathbf{B}$ around each mask. In both cases, we upsample ×4 the DINOv2 feature map size in order to retain feature's regions fine granularity [413]. Similarly, we extract object $\mathbf{o}_S$ and context embeddings $\mathbf{c}_S$ of the source mask $\mathcal{M}_S$ in the other view.

8.3.4 Hard-Mining via Adjacent Neighbors

While the object embedding $\mathbf{o}_n$ contains very discriminative object features, the context embedding $\mathbf{c}_n$ incorporates surrounding information to help localizing the object in the other view, but this surrounding context also introduces ambiguity in cluttered environments, where nearby objects share a similar context. To address this, we introduce a hard-negative mining strategy based on adjacent neighbors, encouraging the model to disambiguate between nearby but distinct objects with similar context. In the destination view, we construct a graph of mask segments based on the pixel centers of each mask using the Delaunay Triangulation, as fig:hard negs shows. This results in a binary adjacency matrix $\mathcal{A} \in \{0,1\}^{N\times N}$, defining the connectivity between segments. We define $\mathcal{N}(\mathbf{o}_n)$ to the set of neighbors of object $\mathbf{o}_n$. Then, we take the second order neighbor set $\mathcal{N}^2(\mathbf{o}_n) = \{\mathcal{N}(\mathbf{o}_j) \;\; \forall \mathbf{o}_j \in \mathcal{N}(\mathbf{o}_n)\}$. Finally, we consider the joint set of first

Figure 8.3: **Examples of complex scenarios.** The target object may appear on the edges of the image, be partially occluded or be extremely small.

and second order neighbors as the hard negative candidates $\mathcal{O}_n^- = \{\mathcal{N}(\mathbf{o}_n) \cup \mathcal{N}^2(\mathbf{o}_n)\}$.

8.3.5 Ego↔Exo Cross Attention

Although the mask context embedding incorporates surrounding contextual information, it lacks a global representation across views. Therefore, we introduce a Ego↔Exo Cross Attention mechanism, which enhances the object embedding by extracting its corresponding semantic features in the other view. Specifically, we compute a cross attention operation [14] between the candidate object masks $\mathbf{o}_n$ and the source image feature map $\psi(I^S)$ as follows:

$$\hat{\mathbf{o}}_n = \text{Softmax}\left(\frac{\mathbf{o}_n W_Q \cdot \left(\psi\left(I^S\right) W_K\right)^\top}{\sqrt{d}}\right) \cdot \psi\left(I^S\right) W_V.$$

We compute query vectors from the candidate object descriptors $\mathbf{o}_n$ with a linear projection W_Q, while key and value vectors represent the overall source image features $\psi(I^S)$ using the W_K and W_V linear layers. Before the cross-attention operation, we incorporate a learnable positional embedding to encode the spatial location of the patch tokens and a standard Layer Norm operation. Intuitively, the cross-view embedding of the object-mask candidates $\hat{\mathbf{o}}_n$ captures how each potential mask candidate is represented in the source view. Similarly, we compute the cross-view embedding of the source mask $\hat{\mathbf{o}}_S$ using the source mask descriptor $\mathbf{o}_S$ for the queries and the overall destination image features $\psi(I^D)$ for the keys and values.

8.3.6 Mask Matching Contrastive Loss

The final descriptor ρ_n is obtained from the n-th candidate mask by concatenating its refined cross-view embedding $\hat{\mathbf{o}}_n$, the context embedding $\mathbf{c}_n$ and the object embedding $\mathbf{o}_n$; while the final descriptor ρ_S of the source mask is the result of concatenating $\hat{\mathbf{o}}_S$, $\mathbf{c}_S$ and $\mathbf{o}_S$, respectively. Then, a shadow multi-layer perceptron $f_\theta(x) \in \mathbb{R}^{d_f}$ maps the

Figure 8.4: **Hard Negatives mining examples** We visualize 2^{nd} order adjacent neighbors both in ego (left) and exo (right) scenarios.

cross-view embeddings to a latent feature representation, where d_f is the dimension of the common feature space. This mapping ensures that both egocentric and exocentric masks are embedded within a shared latent space, enabling a cross-view comparison.

Our contrastive loss is based on InfoNCE [414]. We select a batch $\mathcal{B}$ of $|\mathcal{B}|$ elements, one positive and $|\mathcal{B}|-1$ negatives from the list of closest neighbors around the target object in the other view $\mathcal{O}_n^-$ as defined in Section 8.3.4. If the number of neighbors is greater than the negative batch size, $|\mathcal{O}_n^-| > |\mathcal{B}|$ we randomly select a subset of $|\mathcal{B}|$ elements from the neighbor set $\mathcal{O}_n^-$. If $|\mathcal{O}_n^-| < |\mathcal{B}|$ we also include masks from random objects from the rest of the image. If the number of segmented objects is less than the neighbor set $\mathcal{O}_n^-$, the objects are duplicated in order to fill the negative batch size $|\mathcal{B}|$. Finally, we apply the pairwise cosine similarity $\text{sim}(\cdot,\cdot)$ between the source mask embedding $f_\theta(\rho_S)$ and the batch of $|\mathcal{B}|$ mask candidates embeddings $\{f_\theta(\rho_n)\}_{n=1}^{|\mathcal{B}|}$ for computing the training loss:

$$\mathcal{L}_M(\rho^+, \rho_S) = -\log \frac{\exp(\text{sim}(f_\theta(\rho^+), f_\theta(\rho_S))/\tau)}{\sum_{n=1}^{|\mathcal{B}|} \exp(\text{sim}(f_\theta(\rho_n), f_\theta(\rho_S))/\tau)},$$

where $f_\theta(\rho^+)$ is the positive element in the batch. This *mask matching contrastive loss* aligns the corresponding cross-view object embeddings while it separates the remaining object candidates in the shared feature space.

8.3.7 Inference

At inference, we choose the object candidate whose embedding is closest to the source object in the latent space,

$$\mathcal{M}_{n^*}, \text{ where } n^* = \arg\max(\text{sim}\,[f_\theta(\rho_n), f_\theta(\rho_S)]).$$

We use the source-destination similarity score to decide if the object is visible. Intuitively, if the object does not appear in the destination view, even the similarity of the

closest object candidate should be low.

8.4 Experiments

8.4.1 Experimental Setup

Evaluation. Following the official Ego-Exo4D Correspondences benchmark [27], we adopt the Intersection over Union (IoU) as the primary evaluation metric. We also report the Visibility Accuracy (Vis.A) [415] to assess whether the model predicts when an object is visible or occluded in the target view, the Contour Accuracy (Cont.A) [393] to measure the similarity between predicted and ground-truth mask contours after translation and the Location Error (Loc.E), which quantifies the normalized distance between the predicted and ground-truth mask centroids.

Implementation details. We employ FastSAM [95] with default hyper-parameters (0.9 IoU, 0.4 confidence score) to generate dense candidate masks in the destination view. FastSAM achieves performance comparable to SAM [17] while being 50× faster and utilizing just 68M parameters. For feature extraction, we use a DINOv2 [16] ViT-B/14 model, which consists of 86M parameters. Our model is trained with the AdamW optimizer [152] and an initial learning rate of $8{\cdot}10^{-5}$ with cosine annealing scheduling. We use a batch size of 24 image pairs, sampling 32 masks candidates in each destination image during training. We conduct our experiments on two NVIDIA GeForce RTX 4090.

Training dataset. We use the novel Ego-Exo4D [27] dataset for our experiments. Ego-Exo4D is a massive-scale multi-modal video dataset containing synchronized egocentric and exocentric recordings of human activities. Specifically, we consider the Ego-Exo4D Correspondences set, which includes 1.8M synchronized object masks annotated at 1 FPS, covering 5.6K objects across 1335 unique videos and six different activities (*cooking, bike repair, health, music, basketball,* and *soccer*).

8.4.2 Baselines Models

We compare our approach against the following baselines:

- **XSegTx and XView-XMem** are the official baselines [27]. XSegTx adapts an image co-segmentation model [409], extended with a cross-view temporal memory [416].
- **CMX** [417] is a transformer-based segmentation model that fuses two modalities. We concatenate the query mask with the source image, and we adapt the decoder to predict both visibility and the mask.

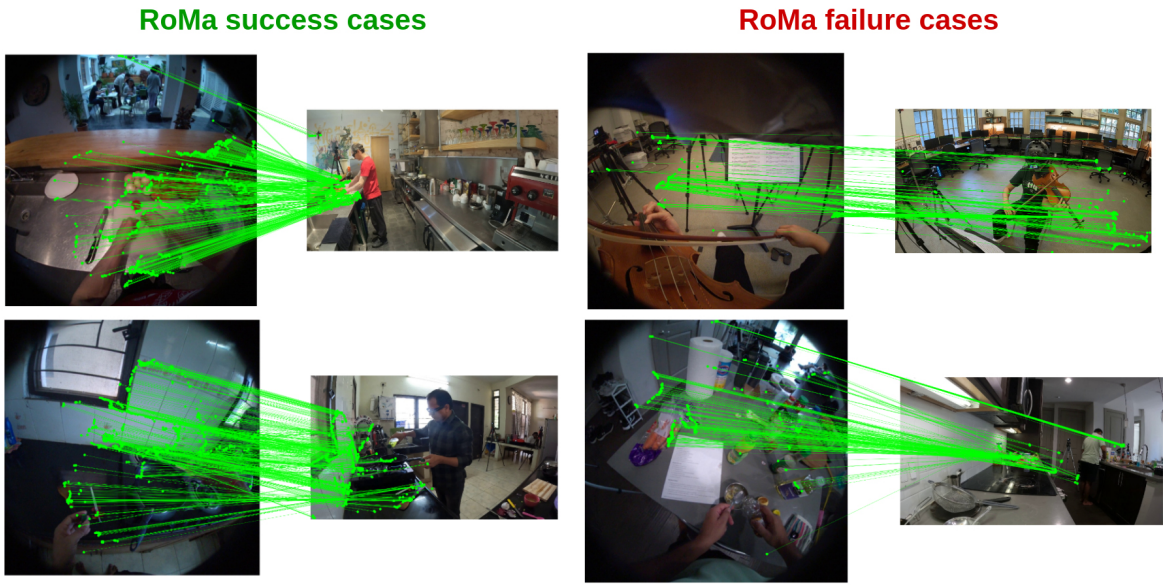


Figure 8.5: **RoMa success and failure cases.** The extreme view variance makes that, even SOTA methods in geometry matching like RoMa [2], fail in extracting matches.

- **PSALM** [400] combines a LLM with Mask2Former [392] to perform zero-shot segmentation. **ObjectRelator** [365] trains PSALM with specialized cross-view modules.

- **k-Nearest Neighbors (k-NN)**. This is a naïve version of our approach. We extract descriptors of the generated mask candidates in the destination view, and we select the most similar to the query mask in the source view.

- **Geometry Methods** [2]. We restrict the k-NN search to only the masks that satisfy the epipolar line restriction, in order to evaluate the geometrical constraints of traditional geometrical matching methods. We tried LightGlue [418] using either SuperGlue [384] (43.8% success rate), SIFT [380] (42% success rate), DISK [419] (31.2% success rate) or ALIKED [420] (40.2% success rate), and compared them using RoMa [2] (67.60% success rate). Due to its more view variance robustness, we select this last method to obtain the fundamental matrix and transfer the mask centroid in the source view to its epipolar line in the destination view, and we discard those candidate masks further than a certain threshold to the epipolar line.

8.4.3 Comparison with the State of the Art

Table 8.1 presents results on the EgoExo4D v2 Correspondences test set, demonstrating the our approach's effectiveness. Even our simplest version, the k-NN baseline, already surpasses the official XMem+XSegTx, achieving 35.3 IoU in Ego2Exo and 34.8 IoU in

	Ego2Exo				Num. Param. (M)	
Method	IoU↑	Vis.A↑	Loc.E↓	Cont.A↑	Total	Train
PSALM (Zero-shot)	7.3	-	0.270	0.121	1587.1	0
CMX	6.8	92.8	0.110	0.137	138.0	17.3
XSegTx	18.9	66.3	0.070	0.386	12.1	3.6
XMem	19.3	64.4	0.151	0.262	62.2	62.2
XMem + XSegTx	34.9	66.8	**0.038**	0.559	75.6	67.1
Ours (k-NN baseline)	35.3	84.7	0.188	0.445	154.0	0
Ours (O-MaMa)	**45.8**	**99.7**	0.077	**0.598**	165.6	11.6

	Exo2Ego				Num. Param. (M)	
Method	IoU↑	Vis.A↑	Loc.E↓	Cont.A↑	Total	Train
PSALM (Zero-shot)	2.1	-	0.290	0.058	1587.1	0
CMX	12.0	90.5	0.166	0.177	138.0	17.3
XSegTx	27.1	82.0	0.104	0.358	12.1	3.6
XMem	16.6	60.3	0.160	0.240	62.2	62.2
XMem + XSegTx	25.0	59.7	0.117	0.237	75.6	67.1
Ours (k-NN baseline)	34.8	87.6	0.182	0.400	154.0	0
Ours (O-MaMa)	**48.6**	**91.7**	**0.103**	**0.563**	165.6	11.6

Table 8.1: **Results on the Ego-Exo4D Correspondences v2 test split.**

Exo2Ego tasks. Our full method, O-MaMa, further improves performance, reaching 45.8 Ego2Exo and 48.6 Exo2Ego IoU, representing considerable relative gains[1] of up to +31.2% and +94.4% over XMem+XSegTx. The improvement is consistent in the other metrics, where O-MaMa obtains 99.7 Vis.A, 0.077 Loc.E, 0.598 Cont.A in the Ego2Exo task and 91.7 Vis.A, 0.103 Loc.E and 0.563 Cont.A in the Exo2Ego task.

As ObjectRelator [365] reports only results on the outdated EgoExo4D v1 Correspondences validation split, we also show results in this split in Table 8.2. O-MaMa also achieves the best performance, obtaining a 50.1 Ego2Exo IoU and 54.2 Exo2Ego IoU, while requiring only 1% of the trainable parameters compared to [365]. The potential of our approach is further demonstrated by comparing our k-NN baseline with PSALM [400] in zero-shot inference. While PSALM achieves only 7.9 and 9.6 IoU in the Ego2Exo and Exo2Ego tasks, respectively, our k-NN baseline scores 40.5 and 40.6 IoU while using approximately 10% of the total number of parameters. This highlights the underlying challenges of the cross-view segmentation task and showcases that reformulating the problem as an object mask matching task significantly boosts zero-shot performance.

[1] We compute the relative gain% of x relative to y as $100 \cdot (\frac{x-y}{y})$.

Method	Ego2Exo IoU	Exo2Exo IoU	Total Param.(M)	Train Param.(M)
XSegTx	6.2	30.2	12.1	3.6
XMem	17.2	20.7	62.2	62.2
XMem + XSegTx	36.9	36.1	75.6	67.1
PSALM (zero-shot)	7.9	9.6	1587.1	0
PSALM (fine-tuned)	41.3	44.1	1587.1	1587.1
ObjectRelator	44.3	50.9	1587.3	1587.3
Ours (k-NN baseline)	40.5	40.6	154.0	0
Ours (O-MaMa)	**50.1**	**54.2**	165.6	11.6

Table 8.2: **Ego-Exo4D Correspondences v1 val split results.**

						Ego2Exo			
Exp.	$\mathcal{L}_M$	Context	Adj. Neg	C.Attn	Global Union	IoU↑	Vis.A↑	Loc.E↓	Cont.A ↑
Base.	-	-	-	-	-	35.2	90.6	0.191	0.455
A	✓	-	-	-	-	42.2	58.6	0.074	0.571
B	✓	✓	-	-	Concat	42.7	78.8	0.069	0.577
C	✓	✓	✓	-	Concat	46.9	88.3	0.079	0.599
D	✓	✓	✓	✓	Weighted Sum.	47.3	83.4	0.064	0.611
E	✓	✓	✓	✓	Concat	**48.3**	**98.1**	**0.062**	**0.621**
Relative Gain % of x with respect to y $\frac{(x-y)}{y}$						+37.2%	+8.3%	+67.5%	+36.5%
						Exo2Ego			
Exp.	$\mathcal{L}_M$	Context	Adj. Neg	C.Attn	Global Union	IoU↑	Vis.A ↑	Loc.E↓	Cont.A↑
Base.	-	-	-	-	-	34.9	94.3	0.163	0.423
A	✓	-	-	-	-	44.7	80.8	0.112	0.546
B	✓	✓	-	-	Concat	44.4	75.3	0.116	0.543
C	✓	✓	✓	-	Concat	45.6	81.8	0.107	0.548
D	✓	✓	✓	✓	Weighted Sum.	46.8	89.3	0.112	0.543
E	✓	✓	✓	✓	Concat	**49.6**	**98.8**	**0.101**	**0.576**
Relative Gain % of x with respect to y $\frac{(x-y)}{y}$						+42.1%	+5.1%	+38.0%	+36.2%

Table 8.3: **Ablation study on the O-MaMa proposed modules on the validation set.**

8.4.4 Ablation study.

O-MaMa architecture. Table 8.3 details the contribution of each O-MaMa component. Experiment A highlights the benefits of training a simple MLP with our novel Mask Matching Contrastive Loss $\mathcal{L}_M$, which aligns cross-view embeddings in a common latent space and improves IoU from 35.2 to 42.2 (Ego2Exo) and 34.9 to 44.7 (Exo2Ego). Second, Experiments A, B and C show that incorporating regional context is only beneficial when we sample adjacent negatives during training, as the hard-mining strategy forces the model to learn more fine-grained discriminative embeddings in nearby candidates with similar context but different mask descriptor. Next, our Ego↔Exo Cross Attention mechanism introduces cross-image content and global information into the object embedding, significantly improving Vis.A (98.1 Ego2Exo, 98.8 Exo2Ego). As

Figure 8.6 shows, this module incorporates the object features from the other perspective, smoothing the cross-view alignment and improving the final performance. Finally, the joined effect of all our proposed modules specially improves the Loc.E, with relative improvements of +67.5% (Ego2Exo) and +38.0% (Exo2Ego), which yields a final gain of +37.2% Ego2Exo and +42.1% Exo2Ego IoU. This demonstrates that, while the k-NN baseline is agnostic to the candidate mask location (it just selects the most similar match), our proposed integration of local and global information results in an object mask selection more sensitive to the cross-view relationship.

Mask Descriptors. Table 8.4 compares different pooling strategies for obtaining a mask descriptor. The k-NN baseline, which relies solely on object semantic similarity, shows that averaging DINOv2 upsampled features over mask pixels provides the best results (35.2 and 34.9 IoU), as it retains the fine-grained object representation from the dense DINOv2 feature map. This strategy outperforms CLIP-based descriptors (24.5 Ego2Exo and 23.9 Exo2Ego IoU), DINOv2 average pooling over the bounding box (21.8 Ego2Exo and 21.2 Exo2Ego IoU) or DINOv2 mask centroid (25.6 Ego2Exo and 24.1 Exo2Ego IoU). Table 8.4 also reports that applying geometric constraints yields a minor performance gain (35.2 vs. 35.4 Ego2Exo IoU, and 34.9 vs. 36.6 Exo2Ego IoU when pooling DINOv2 mask features) due to the low success rate of camera pose estimation methods. As Figure 8.5 shows, even RoMa [2] struggles with the high viewpoint variance between ego and exo perspectives. This improvement is even less significant when compared to learning view-invariant features with $\mathcal{L}_M$ (35.2 vs. 42.2 Ego2Exo IoU and 34.9 vs. 44.7 Exo2Ego IoU), highlighting the need to extract stronger visual cues.

Detailed performance per task and mask size. Figure 8.8 shows the IoU across different scenarios in the Ego2Exo task, where O-MaMa outperforms XMem + XSeg-Tx in most cases, including the challenging *cooking* and *bike repair* activities, which involve cluttered environments and objects of multiple sizes. Figure 8.7 analyzes the segmentation performance across the target mask sizes, showing that O-MaMa excels in medium and large-size objects. However, it still struggles with very small objects, as extracting a meaningful mask descriptor remains challenging.

8.4.5 Qualitative results

We show qualitative examples for the Ego2Exo (Figure 8.10) and Exo2Ego (Figure 8.9) tasks. The results show that top mask candidates are closely aligned due to their similar context, but our method correctly matches the top-1 mask candidate with the target object. FastSAM's fine-grained zero-shot capabilities yield high-quality segmentation masks (e.g, the *tire* in Figure 8.10 and the *knife* or *bottle* in Figure 8.9). However, as a

			Ego2Exo			
Method	Geometry	$\mathcal{L}_M$	IoU↑	Vis.A↑	Loc.E↓	Cont.A↑
Max-Pool(**b**) DINOv2 [16]	✗	✗	5.8	**97.6**	0.306	0.119
	✓	✗	6.9	97.4	0.302	0.132
Centroid($\mathcal{M}$) DINOv2 [16]	✗	✗	25.6	73.6	0.202	0.357
	✓	✗	26.5	73.4	0.190	0.378
Avg-Pool(**b**) DINOv2 [16]	✗	✗	21.8	94.8	0.245	0.324
	✓	✗	23.2	94.2	0.238	0.345
	✗	✓	27.8	62.5	0.092	0.426
Avg-Pool(**b**) CLIP [243]	✗	✗	24.5	95.8	0.257	0.325
	✓	✗	26.2	95.3	0.220	0.359
	✗	✓	27.5	90.2	0.170	0.379
Avg-Pool($\mathcal{M}$) DINOv2 [16]	✗	✗	35.2	90.6	0.191	0.455
	✓	✗	35.4	90.2	0.184	0.467
	✗	✓	**42.2**	58.6	**0.074**	**0.571**
			Exo2Ego			
Method	Geometry	$\mathcal{L}_M$	IoU↑	Vis.A↑	Loc.E↓	Cont.A↑
Max-Pool(**b**) DINOv2 [16]	✗	✗	17.8	**97.2**	0.216	0.253
	✓	✗	20.2	96.9	0.210	0.278
Centroid($\mathcal{M}$) DINOv2 [16]	✗	✗	24.1	83.4	0.178	0.326
	✓	✗	26.0	83.2	0.172	0.346
Avg-Pool(**b**) DINOv2 [16]	✗	✗	21.2	95.6	0.201	0.291
	✓	✗	23.7	95.2	0.195	0.314
	✗	✓	44.1	64.9	**0.111**	0.537
Avg-Pool(**b**) CLIP [243]	✗	✗	23.9	94.6	0.234	0.301
	✓	✗	26.7	94.0	0.209	0.335
	✗	✓	40.4	40.7	0.155	0.477
Avg-Pool($\mathcal{M}$) DINOv2 [16]	✗	✗	34.9	94.3	0.163	0.423
	✓	✗	36.6	94.2	0.156	0.440
	✗	✓	**44.7**	80.8	0.112	**0.546**

Table 8.4: **Ablation study on the mask descriptors and the influence of learning and geometry constraints.** We compare the effects of leveraging inferred camera pose constraints or training a simple MLP with our $\mathcal{L}_M$, configuration that corresponds to Exp.A in Table 8.3.

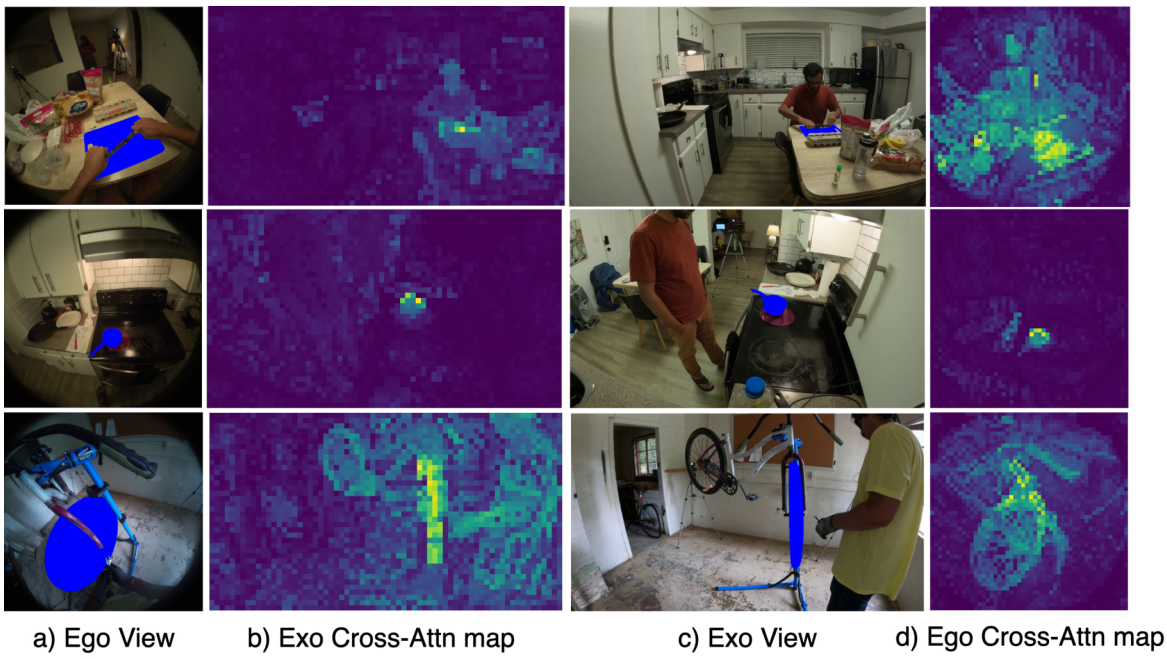


Figure 8.6: **Ego↔Exo Cross-Attention maps**

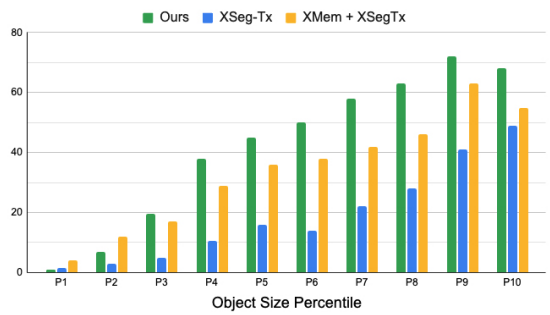


Figure 8.7: **Ego2Exo IoU performance across different object sizes in the destination view.**

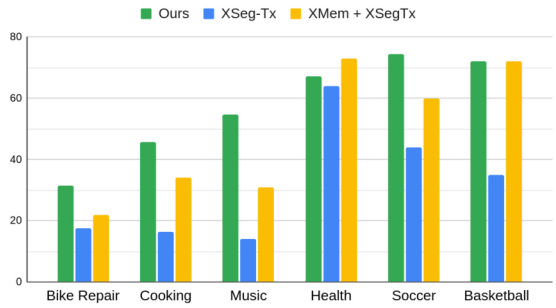


Figure 8.8: **Per-task Ego2Exo IoU performance.**

limitation of our approach, they may produce partial segmentations when they capture only a part of the object (e.g, the *saucepan* in Figure 8.10). Finally, the total inference time of our approach is 250ms on average, of which 70ms correspond to the FastSAM mask extraction.

8.5 Conclusions

In this work, we address the problem of ego-exo object correspondences, a key step for multi-agent perception. We demonstrate that reformulating cross-view segmentation as an object mask matching problem simplifies the task while improving accuracy under zero-shot conditions. Our Mask Matching Contrastive Loss effectively aligns cross-view embeddings, while DINOv2 pooled mask features preserve fine-grained details. The

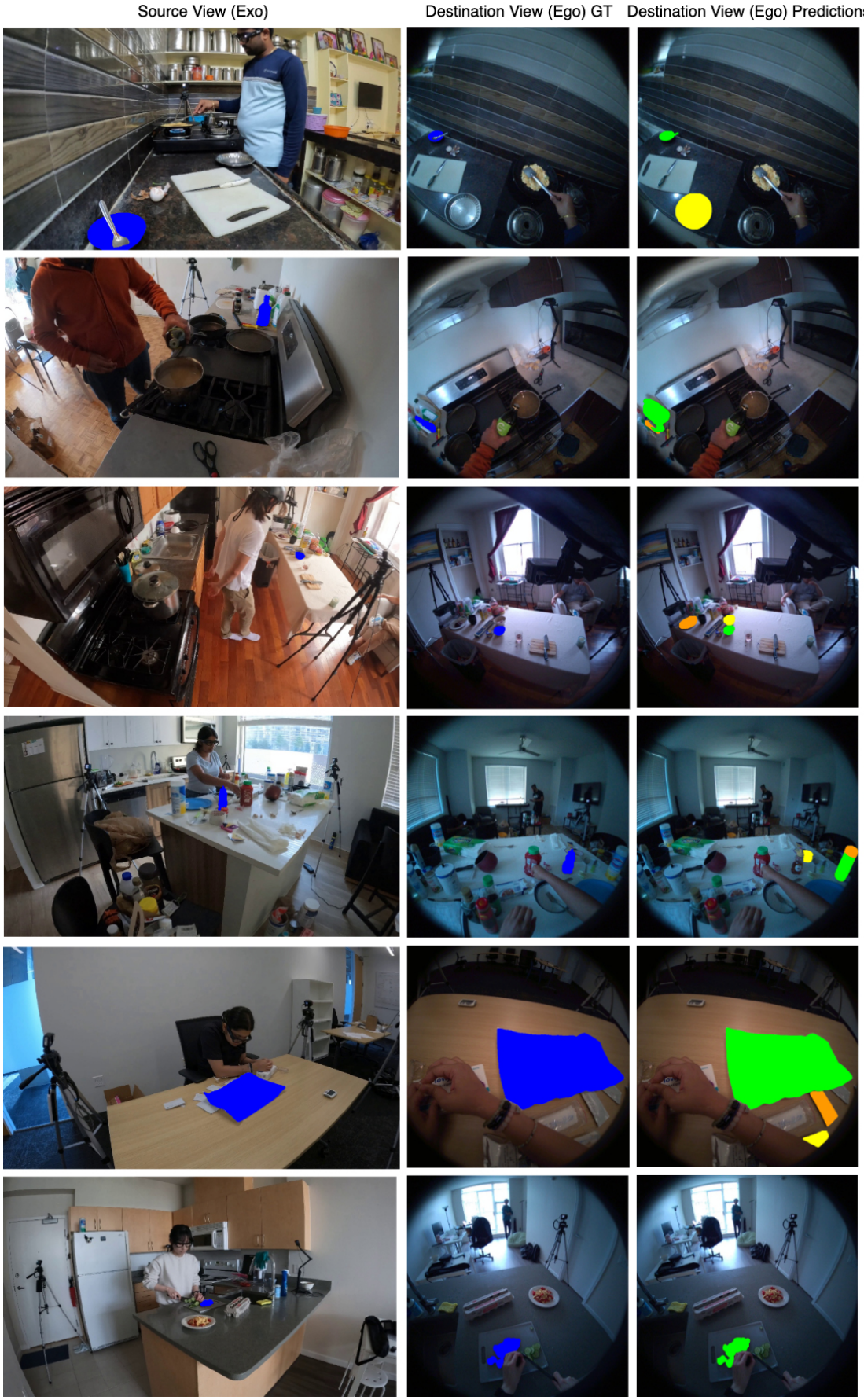


Figure 8.9: **Exo2Ego Qualitative Results.** We show the source mask in blue and the top 3 target masks in green, yellow and orange.

Figure 8.10: **Ego2Exo Qualitative Results.** For visualization purposes, we show the top 3 masks in green, yellow and orange.

proposed Hard Negative Adjacent Mining strategy enhances object differentiation, and Ego↔Exo Cross Attention integrates global cross-view context. As a result, O-MaMa achieves state of the art performance on the EgoExo4D Correspondences task while using considerably fewer parameters, obtaining a unified fine-grained segmentation and strong cross-view understanding.

Chapter 9

Conclusions and Future Work

In this thesis, we have advanced egocentric perception by learning visual models of objects, environments, and affordances, complemented by multi-modal representations. Regarding visual models of objects, we first extended Mask-RCNN [12] with intermediate sampling layers to support instance segmentation with uncertainty estimation. We then adapted two popular object detection architectures—Faster-RCNN [290] and Detection Transformers (DETR) [85]—to the short term anticipation task. Specifically, we introduced attention-based components that enable effective image-video fusion, improving the spatial, temporal, and semantic understanding of the next interaction.

In learning visual models of environments, we first introduced a multi-label affordance mapping, and demonstrated its potential to support embodied skills such as task-oriented navigation. Then, we improved the robustness of environment functions by exploiting the predictive distribution of a Bayesian neural network. Our approach first regularizes observations by reducing the influence of overconfident outlier measurements and incorporates epistemic uncertainty by obtaining a Dirichlet distribution from the output of the Bayesian neural network. Next, we proposed learning environment representations using implicit functions, specifically through a decomposed neural radiance field consisting of three components—persistent, dynamic, and actor—to better model the inherent dynamics of egocentric videos. This representation was further enriched with image-language and video-language feature fields, effectively capturing affordances and enabling temporally consistent segmentation of dynamic objects.

Following, throughout all the thesis, we investigated the role of affordances for improving detection, mapping, and forecasting tasks. We first improved detection by learning both ungrounded and grounded affordance segmentation models, enabling uncertainty quantification and multi-label segmentation, respectively. Then, we integrated affordances into an environment representation, such as a point cloud, joining the 3D geometry with activity-centric zones and showing its potential for task-oriented navigation. Next, we presented different alternatives for grounding the predictions

of a forecasting model in past observed human behavior by leveraging environment affordances and interaction hotspots.

In the final section, we enhanced egocentric perception through the integration of multi-modal representations. First, we aligned language with long, untrimmed egocentric videos by temporally localizing activities described in textual queries. To this end, we introduced Bayesian-VSLNet, a test-time refinement strategy that incorporates temporal-order priors based on the sequential structure of the task steps. Second, we reformulated the cross-view object segmentation task as an object mask matching problem. The proposed model learns to segment an object from one viewpoint using a query mask from another, effectively bridging first- and third-person perspectives at the object level.

Returning to our motivational Figure 1.1, this thesis has presented a series of computer vision algorithms that conceptually address the different situations depicted. For instance, DIV-FF enables the creation of a spatial memory capable of recalling the location of specific objects; STA-former++ anticipates user actions to prevent potential errors while following a cooking recipe; and O-MaMa allows monitoring of body pose by integrating egocentric and exocentric views.

The key contributions of this thesis are summarized as follows:

I. In Chapter 2, we extended affordance segmentation to a probabilistic framework by extracting per-pixel estimation of both aleatoric and epistemic variance at spatial and semantic levels, achieving state-of-the-art results in the IIT-Aff dataset.

II. We introduced a new Bayesian fusion method that exploits the uncertainty quantification to increase the robustness of voxel-based semantic maps in Chapter 4.

III. To address the limitations of ungrounded affordance methods, we presented in Chapter 3 a pipeline for automatically collecting multi-label, pixel-wise annotations from real-world interactions, resulting in a novel dataset, EPIC-Aff. Then, we adapted popular segmentation architectures to the multi-label setting and proposed various heuristics for selecting multiple labels from the predicted probability vector.

IV. We presented multiple approaches for capturing environment affordances: a multi-label point cloud representation in Chapter 3, a distilled video-language feature field in Chapter 5, and a video-to-activity zone matching strategy in Chapter 6.

V. In Chapter 5, we adapted language-embedded feature fields to dynamic egocentric videos by decomposing the radiance and feature field representations into

actor-centric, dynamic, and persistent components. This was complemented with a robust image-language feature field that leveraged Segment Anything Model (SAM)-derived object masks to achieve temporally consistent segmentation of dynamic objects.

VI. We designed two principled architectures fully based on transformers for the Short-Term object interaction Anticipation (STA) task in Chapter 6, achieving state-of-the-art performance on the Ego4D validation splits [26] and on a newly curated set of STA annotations for the EPIC-Kitchens dataset.

VII. In Chapter 6, we proposed three approaches to ground forecasting predictions in human behavior: (1) late fusion of pre-computed environment affordances at inference time, (2) integration of affordance information during training via an attention mechanism, and (3) confidence re-weighting based on the location of the object bounding box within interaction hotspot

VIII. We formulated a test-time refinement strategy for the Step Grounding task in Chapter 7, which incorporates temporal-order priors into the predictions. This effective refinement, guided by Bayes' theorem, aligns predictions according to the sequential order of steps, achieving state-of-the-art performance on the Step Grounding Ego4D benchmark.

IX. We reformulated the cross-view segmentation problem as a cross-view object mask matching task in Chapter 8. The proposed method integrates key components—such as the Mask-Context Encoder, Hard Negative Adjacent Mining, Ego↔Exo Cross Attention, and a Mask Matching Contrastive Loss—yielding significant improvements over prior work and setting a new state of the art on the Ego-Exo4D Object Correspondences benchmark, while requiring only 1 % of the trainable parameters compared to the best-competing method.

Overall, the proposed methods achieved state-of-the-art performance on three official benchmarks in egocentric perception: Short-Term Anticipation on Ego4D [26], Step Grounding on Ego4D-Goal Step [3], and Object Correspondences on Ego-Exo4D [27]. In addition, we contributed with two novel datasets to the field of egocentric vision: EPIC-Aff and a new set of curated labels on Epic-Kitchens for the short-term anticipation task. These efforts were further complemented by the design of four principled architectures: STA-former, STA-former++, DIV-FF, and O-MaMa. These contributions have been presented at top-tier computer vision and robotics conferences, including ICRA 2023 [111], IROS 2023 [102], ICCV 2023 [38], ECCV 2024 [232], CVPR

2025 [421] and ICCV 2025, and have been further extended into three journal submissions currently under review.

9.1 Future Work

While these contributions represent significant progress, they also open up several promising and exciting directions for future research:

Active Visual Perception. The proposed Bayesian semantic mapping in Chapter 4 constitutes an intermediate step toward end-to-end active visual perception. In the context of exploring an unseen environment, future navigation algorithms could leverage the epistemic uncertainty encoded in the map to guide the agent toward ambiguous or poorly modeled regions.

Exploiting environment representations. The environment representations presented from Chapters 3 to 5 exhibit certain limitations that suggest opportunities for future research.

- **Long-term memory.** Currently, DIV-FF processes 1,000 sparsely sampled frames from a 30-minute video, which is insufficient for modeling extended time periods such as an entire day or week of continuous activity. Investigating scalable and memory-efficient mechanisms to retain long-term representations is a promising direction of study.
- **Revisiting environments.** From a single egocentric video, DIV-FF captures a function of the environment, the dynamic objects and the actor. However, real-world environments are continuously revisited, offering new challenges and opportunities. For example, could the persistent environment be leveraged to construct a prior that guides the user when performing new tasks? How could the environment function be updated with novel experiences while avoiding catastrophic forgetting?
- **Multi-modal distillation.** DIV-FF distills video-language (Ego-Video) and image-language (CLIP) features, enabling the identification of actions and objects described through free-form language. However, humans also rely on auditory cues to localize events in 3D space—for instance, associating the sound of a ringing phone with its location. As modern wearable devices increasingly incorporate microphones, future environment functions will likely integrate audio signals into implicit feature fields.

Video foundational models. Foundational models represent a promising direction for learning robust and generalizable strong features across multiple tasks. In the image domain, models such as DINO-v2 [16] have demonstrated strong performance and emergent capabilities—such as object part recognition and implicit scene geometry—by leveraging large-scale self-supervised training on 142 million images. In the video domain, the recently proposed V-JEPA (Joint Embedding Prediction Architecture) [18], trained in "only" 2 M videos, already outperforms previous approaches and it is particularly effective in motion understanding, but its performance is still limited in capturing fine-grained action details. Scaling the pretraining video distribution and incorporating multi-modal inputs represent promising directions toward a unified architecture, which will enhance video understanding across a wide range of tasks.

Cross-environment object correspondences and cross-level action correspondences. In Chapter 8, O-MaMa shows how to capture fine-grained object instance correspondences across synchronized views. The natural next step is to extend this capability to *non-synchronized* videos recorded in different environments. For example, the violin bow position of a professional violinist from an Internet video could be paired from the egocentric viewpoint of a novice player. Following, another interesting future research is the discovery of fine-grained cross-level *action* correspondences, even when the execution quality differs significantly. Continuing with the violinist example, an expert may perform smooth, precise bowing movements, while a beginner may produce uneven or scratchy sounds. Identifying these action correspondences despite skill-level disparities will facilitate effective knowledge transfer from expert demonstrations to novice users, opening avenues for skill assessment using wearable devices.

Cooperative perception. The arrival of Ego-Exo4D [27] unlocked the multi-view paradigm in egocentric perception, capturing a scene from a dynamic firs-person view combined with multiple static third-person cameras. The next frontier is the development of multi-agent egocentric perception—modeling scenarios in which multiple individuals perform coordinated tasks with a common goal, such as in team sports, professional kitchens, or the staging of a musical performance. Learning cooperative egocentric models will require reasoning not only about each individual perspective, but also about inter-agent dependencies and synchronized behavior, constituting a very promising and exciting research direction.

Conclusiones

En esta tesis, se ha avanzado la percepción egocéntrica mediante el aprendizaje de modelos visuales de objetos, entornos y affordances, complementados con representaciones multimodales.

En lo que respecta a los modelos visuales de objetos, en primer lugar se amplió Mask-RCNN [12] incorporando capas intermedias de muestreo, lo que permitió realizar segmentación de instancias con estimación de incertidumbre. Posteriormente, se adaptaron dos arquitecturas populares de detección de objetos —Faster-RCNN [290] y DETR [85]— para la tarea de anticipación a corto plazo. En concreto, se introdujeron componentes basados en atención que facilitaron la fusión eficaz de imágenes y vídeos, mejorando así la comprensión espacial, temporal y semántica de las interacciones futuras.

En el aprendizaje de modelos visuales de entornos, e primer lugar se presentó un mapeado de affordances multi-etiqueta, demostrando su potencial para respaldar habilidades como la navegación orientada a tareas. A continuación, se mejoró la robustez de las funciones de representación del entorno mediante el aprovechamiento de la distribución predictiva de una red neuronal bayesiana. Esta enfoque regulaba las observaciones, reduciendo la influencia de mediciones atípicas con excesiva confianza e incorporó incertidumbre epistémica a través de la obtención de una distribución de Dirichlet generada a partir de la salida de la red neuronal bayesiana. Posteriormente, se propuso el aprendizaje de representaciones del entorno mediante funciones implícitas, en particular a través de un campo neuronal de radiancia descompuesto en tres componentes —persistente, dinámica y actor—, con el fin de modelar de forma más precisa las dinámicas inherentes a los vídeos egocéntricos. Dicha representación se enriqueció además con campos de características imagen-lenguaje y vídeo-lenguaje, lo que permitió capturar affordances de manera efectiva y lograr segmentaciones de objetos dinámicos temporalmente coherentes.

A lo largo de la tesis, se ha investigado el papel de las affordances en la mejora de las tareas de detección, mapeo y predicción. Primero, se mejoró la detección mediante el aprendizaje de modelos de segmentación de affordances, lo que permitió la

cuantificación de la incertidumbre y la segmentación multi-etiqueta, respectivamente. Posteriormente, se integraron las affordances en la representación del entorno usando nubes de puntos que combinaban la geometría 3D con zonas de actividad, demostrando su utilidad para la navegación orientada a tareas. Asimismo, se presentaron diversas estrategias para fundamentar las predicciones de los modelos de anticipación en el comportamiento humano previamente observado, aprovechando las affordances del entorno y las zonas de interacción preferente (interaction hotspots).

En la sección final, se amplió la percepción egocéntrica mediante la integración de representaciones multimodales. En primer lugar, se alineó el lenguaje con vídeos egocéntricos extensos y no segmentados, localizando temporalmente las actividades descritas en consultas textuales. Para ello, se introdujo Bayesian-VSLNet, una estrategia de refinamiento que incorporaba información a priori sobre el orden temporal en función de la estructura secuencial de los pasos de la tarea. En segundo lugar, se reformuló la tarea de segmentación de objetos entre diferentes vistas (cross-view object segmentation) como un problema de correspondencia de máscaras de objetos. El modelo propuesto aprendió a segmentar un objeto desde una vista utilizando como referencia una máscara procedente de otra vista, lo que permitió una integración eficaz entre las perspectivas en primera y tercera persona a nivel de objeto.

Chapter 10

Bibliography

[1] Zipeng Fu, Tony Z. Zhao, and Chelsea Finn. Mobile aloha: Learning bimanual mobile manipulation with low-cost whole-body teleoperation. In *arXiv*, 2024.

[2] Johan Edstedt, Qiyu Sun, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. Roma: Robust dense feature matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19790–19800, 2024.

[3] Yale Song, Eugene Byrne, Tushar Nagarajan, Huiyu Wang, Miguel Martin, and Lorenzo Torresani. Ego4d goal-step: Toward hierarchical understanding of procedural activities. *Advances in Neural Information Processing Systems*, 36, 2024.

[4] Rohit Girdhar, Mannat Singh, Nikhila Ravi, Laurens Van Der Maaten, Armand Joulin, and Ishan Misra. Omnivore: A single model for many visual modalities. In *CVPR*, pages 16102–16112, 2022.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[6] Shraman Pramanick, Yale Song, Sayan Nag, Kevin Qinghong Lin, Hardik Shah, Mike Zheng Shou, Rama Chellappa, and Pengchuan Zhang. Egovlpv2: Egocentric video-language pre-training with fusion in the backbone. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5285–5297, 2023.

[7] Baoqi Pei, Guo Chen, Jilan Xu, Yuping He, Yicheng Liu, Kanghua Pan, Yifei Huang, Yali Wang, Tong Lu, Limin Wang, et al. Egovideo: Exploring egocentric foundation model and downstream adaptation. *arXiv preprint arXiv:2406.18070*, 2024.

[8] Eshed OhnBar, Kris Kitani, and Chieko Asakawa. Personalized dynamics models for adaptive assistive navigation systems. In *Conference on Robot Learning*, pages 16–39. PMLR, 2018.

[9] Melani Sanchez-Garcia, Ruben Martinez-Cantin, and Jose J Guerrero. Semantic and structural image segmentation for prosthetic vision. *Plos one*, 15(1):e0227677, 2020.

[10] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[12] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2961–2969, 2017.

[13] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations*, 2021.

[16] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[17] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al.

Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023.

[18] Adrien Bardes, Quentin Garrido, Jean Ponce, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *arXiv:2404.08471*, 2024.

[19] Pratyusha Sharma, Deepak Pathak, and Abhinav Gupta. Third-person visual imitation learning via decoupled hierarchical controller. *Advances in Neural Information Processing Systems*, 32, 2019.

[20] Vidhi Jain, Maria Attarian, Nikhil J Joshi, Ayzaan Wahid, Danny Driess, Quan Vuong, Pannag R Sanketi, Pierre Sermanet, Stefan Welker, Christine Chan, et al. Vid2robot: End-to-end video-conditioned policy learning with cross-attention transformers. *arXiv preprint arXiv:2403.12943*, 2024.

[21] Shikhar Bahl, Abhinav Gupta, and Deepak Pathak. Human-to-robot imitation in the wild. *arXiv preprint arXiv:2207.09450*, 2022.

[22] Simar Kareer, Dhruv Patel, Ryan Punamiya, Pranay Mathur, Shuo Cheng, Chen Wang, Judy Hoffman, and Danfei Xu. Egomimic: Scaling imitation learning via egocentric video. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 13226–13233. IEEE, 2025.

[23] Ashish Kumar, Saurabh Gupta, and Jitendra Malik. Learning navigation subroutines from egocentric videos. In *Conference on Robot Learning*, pages 617–626. PMLR, 2020.

[24] Marion Lepert, Jiaying Fang, and Jeannette Bohg. Masquerade: Learning from in-the-wild human videos using data-editing. *arXiv preprint arXiv:2508.09976*, 2025.

[25] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 720–736, 2018.

[26] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu,

et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022.

[27] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19383–19400, 2024.

[28] Antonino Furnari and Giovanni Maria Farinella. What would you expect? anticipating egocentric actions with rolling-unrolling lstms and modality attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6252–6261, 2019.

[29] Rohit Girdhar and Kristen Grauman. Anticipative video transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13505–13515, 2021.

[30] Francesco Marchetti, Federico Becattini, Lorenzo Seidenari, and Alberto Del Bimbo. Multiple trajectory prediction of moving agents with memory augmented networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[31] Kris M Kitani, Brian D Ziebart, James Andrew Bagnell, and Martial Hebert. Activity forecasting. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part IV 12*, pages 201–214. Springer, 2012.

[32] Antonino Furnari, Sebastiano Battiato, Kristen Grauman, and Giovanni Maria Farinella. Next-active-object prediction from egocentric videos. *Journal of Visual Communication and Image Representation*, 49:401–411, 2017.

[33] Francesco Ragusa, Antonino Furnari, Salvatore Livatino, and Giovanni Maria Farinella. The meccano dataset: Understanding human-object interactions from egocentric videos in an industrial-like domain. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1569–1578, 2021.

[34] Siddhant Bansal, Chetan Arora, and CV Jawahar. My view is the best view: Procedure learning from egocentric videos. In *European Conference on Computer Vision*, pages 657–675. Springer, 2022.

[35] Gorjan Radevski, Dusan Grujicic, Matthew Blaschko, Marie-Francine Moens, and Tinne Tuytelaars. Multimodal distillation for egocentric action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5213–5224, 2023.

[36] Dandan Shan, Jiaqi Geng, Michelle Shu, and David F Fouhey. Understanding human hands in contact at internet scale. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9869–9878, 2020.

[37] Vadim Tschernezki, Ahmad Darkhalil, Zhifan Zhu, David Fouhey, Iro Laina, Diane Larlus, Dima Damen, and Andrea Vedaldi. Epic fields: Marrying 3d geometry and video understanding. *Advances in Neural Information Processing Systems*, 36, 2024.

[38] Lorenzo Mur-Labadia, Jose J Guerrero, and Ruben Martinez-Cantin. Multi-label affordance mapping from egocentric vision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5238–5249, 2023.

[39] Tushar Nagarajan, Yanghao Li, Christoph Feichtenhofer, and Kristen Grauman. Ego-topo: Environment affordances from egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 163–172, 2020.

[40] Jinjie Mai, Abdullah Hamdi, Silvio Giancola, Chen Zhao, and Bernard Ghanem. Egoloc: Revisiting 3d object localization from egocentric videos with visual queries. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 45–57, 2023.

[41] Leonard Bärmann and Alex Waibel. Where did i leave my keys?-episodic-memory-based question answering on egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1560–1568, 2022.

[42] Jean-Baptiste Alayrac, Ivan Laptev, Josef Sivic, and Simon Lacoste-Julien. Joint discovery of object states and manipulation actions. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2127–2136, 2017.

[43] Zihui Xue, Kumar Ashutosh, and Kristen Grauman. Learning object state changes in videos: An open-world perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18493–18503, 2024.

[44] Jiangwei Yu, Xiang Li, Xinran Zhao, Hongming Zhang, and Yu-Xiong Wang. Video state-changing object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20439–20448, 2023.

[45] James J Gibson. The theory of affordances. *Hilldale, USA*, 1(2):67–82, 1977.

[46] Robin R Murphy. Case studies of applying gibson's ecological approach to mobile robots. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 29(1):105–111, 1999.

[47] Hema S Koppula and Ashutosh Saxena. Anticipating human activities using object affordances for reactive robotic response. *IEEE transactions on pattern analysis and machine intelligence*, 38(1):14–29, 2015.

[48] Aris Alissandrakis, Chrystopher L Nehaniv, and Kerstin Dautenhahn. Imitation with alice: Learning to imitate corresponding actions across dissimilar embodiments. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 32(4):482–496, 2002.

[49] Manuel Lopes and José Santos-Victor. Visual learning by imitation with motor representations. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 35(3):438–449, 2005.

[50] Shuo Yang, Wei Zhang, Ran Song, Jiyu Cheng, Hesheng Wang, and Yibin Li. Watch and act: Learning robotic manipulation from visual demonstration. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2023.

[51] Luis Montesano, Manuel Lopes, Alexandre Bernardino, and José Santos-Victor. Learning object affordances: from sensory–motor coordination to imitation. *IEEE Transactions on Robotics*, 24(1):15–26, 2008.

[52] Chunfang Liu, Bin Fang, Fuchun Sun, Xiaoli Li, and Wenbing Huang. Learning to grasp familiar objects based on experience and objects' shape affordance. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 49(12):2710–2723, 2019.

[53] Tushar Nagarajan, Christoph Feichtenhofer, and Kristen Grauman. Grounded human-object interaction hotspots from video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8688–8697, 2019.

[54] Austin Myers, Ching L Teo, Cornelia Fermüller, and Yiannis Aloimonos. Affordance detection of tool parts from geometric features. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1374–1381. IEEE, 2015.

[55] Anh Nguyen, Dimitrios Kanoulas, Darwin G Caldwell, and Nikos Tsagarakis. Object-based affordances detection with convolutional neural networks and dense conditional random fields. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5908–5915. IEEE, 2017.

[56] Ching-Yao Chuang, Jiaman Li, Antonio Torralba, and Sanja Fidler. Learning to act properly: Predicting and explaining affordances from images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 975–983, 2018.

[57] Hongchen Luo, Wei Zhai, Jing Zhang, Yang Cao, and Dacheng Tao. Learning affordance grounding from exocentric images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2252–2261, 2022.

[58] Thanh-Toan Do, Anh Nguyen, and Ian Reid. Affordancenet: An end-to-end deep learning approach for object affordance detection. In *2018 International Conference on Robotics and Automation (ICRA)*, pages 5882–5889. IEEE, 2018.

[59] Chau Nguyen Duc Minh, Syed Zulqarnain Gilani, Syed Mohammed Shamsul Islam, and David Suter. Learning affordance segmentation: An investigative study. In *2020 Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–8. IEEE, 2020.

[60] Hugo Caselles-Dupré, Michael Garcia-Ortiz, and David Filliat. Are standard object segmentation models sufficient for learning affordance segmentation? *arXiv preprint arXiv:2107.02095*, 2021.

[61] Hongchen Luo, Wei Zhai, Jing Zhang, Yang Cao, and Dacheng Tao. Learning visual affordance grounding from demonstration videos. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[62] Mohit Goyal, Sahil Modi, Rishabh Goyal, and Saurabh Gupta. Human hands as probes for interactive object understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3293–3303, 2022.

[63] Gen Li, Varun Jampani, Deqing Sun, and Laura Sevilla-Lara. Locate: Localize and transfer object parts for weakly supervised affordance grounding. In *Proceed-*

ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10922–10931, 2023.

[64] Yuhang Yang, Wei Zhai, Hongchen Luo, Yang Cao, Jiebo Luo, and Zheng-Jun Zha. Grounding 3d object affordance from 2d interactions in images. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[65] Nicholas Rhinehart and Kris M Kitani. Learning action maps of large environments via first-person vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 580–588, 2016.

[66] Miao Liu, Lingni Ma, Kiran Somasundaram, Yin Li, Kristen Grauman, James M Rehg, and Chao Li. Egocentric activity recognition and localization on a 3d map. In *European Conference on Computer Vision*, pages 621–638. Springer, 2022.

[67] Shaowei Liu, Subarna Tripathi, Somdeb Majumdar, and Xiaolong Wang. Joint hand motion and interaction hotspots prediction from egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3282–3292, 2022.

[68] Miao Liu, Siyu Tang, Yin Li, and James M Rehg. Forecasting human-object interaction: joint prediction of motor attention and actions in first person video. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 704–721. Springer, 2020.

[69] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[70] Luis Montesano and Manuel Lopes. Learning grasping affordances from local visual descriptors. In *2009 IEEE 8th International Conference on Development and Learning*, pages 1–6. IEEE, 2009.

[71] Anh Nguyen, Dimitrios Kanoulas, Darwin G Caldwell, and Nikos G Tsagarakis. Detecting object affordances with convolutional neural networks. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2765–2770. IEEE, 2016.

[72] Kuan Fang, Te-Lin Wu, Daniel Yang, Silvio Savarese, and Joseph J Lim. Demo2vec: Reasoning object affordances from online videos. In *Proceedings of*

the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2139–2147, 2018.

[73] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2921–2929, 2016.

[74] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9650–9660, 2021.

[75] Chiara Plizzari, Shubham Goel, Toby Perrett, Jacob Chalk, Angjoo Kanazawa, and Dima Damen. Spatial cognition from egocentric video: Out of sight, not out of mind. *arXiv preprint arXiv:2404.05072*, 2024.

[76] Vincent Cartillier, Zhile Ren, Neha Jain, Stefan Lee, Irfan Essa, and Dhruv Batra. Semantic mapnet: Building allocentric semantic maps and representations from egocentric views. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

[77] Joao F Henriques and Andrea Vedaldi. Mapnet: An allocentric spatial memory for mapping environments. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8476–8484, 2018.

[78] Yash Bhalgat, Vadim Tschernezki, Iro Laina, João F Henriques, Andrea Vedaldi, and Andrew Zisserman. 3d-aware instance segmentation and tracking in egocentric videos. *arXiv preprint arXiv:2408.09860*, 2024.

[79] Chiara Plizzari, Toby Perrett, Barbara Caputo, and Dima Damen. What can a cook in italy teach a mechanic in india? action recognition generalisation over scenarios and locations. In *ICCV2023*, 2023.

[80] Jingjing Jiang, Zhixiong Nan, Hui Chen, Shitao Chen, and Nanning Zheng. Predicting short-term next-active-object through visual attention and hand position. *Neurocomputing*, 433:212–222, 2021.

[81] Tushar Nagarajan. *Learning affordance, environment and interaction representations by watching people in video.* PhD thesis, University of Austin, Texas, 2023.

[82] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR, 2017.

[83] Alex Kendall and Yarin Gal. What uncertainties do we need in Bayesian deep learning for computer vision? *31st Conference on Neural Information Processing Systems (NIPS)*, 2017.

[84] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[85] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[86] Zhaoyang Lv, Nickolas Charron, Pierre Moulon, Alexander Gamino, Cheng Peng, Chris Sweeney, Edward Miller, Huixuan Tang, Jeff Meissner, Jing Dong, et al. Aria everyday activities dataset. *arXiv preprint arXiv:2402.13349*, 2024.

[87] Haresh Karnan, Garrett Warnell, Xuesu Xiao, and Peter Stone. Voila: Visual-observation-only imitation learning for autonomous navigation. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2497–2503. IEEE, 2022.

[88] Chen-Lin Zhang, Jianxin Wu, and Yin Li. Actionformer: Localizing moments of actions with transformers. In *European Conference on Computer Vision*, pages 492–510. Springer, 2022.

[89] Dingfeng Shi, Yujie Zhong, Qiong Cao, Lin Ma, Jia Li, and Dacheng Tao. Tridet: Temporal action detection with relative boundary modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18857–18866, 2023.

[90] Fan Yang, Shigeyuki Odashima, Shoichi Masui, and Shan Jiang. Is weakly-supervised action segmentation ready for human-robot interaction? no, let's improve it with action-union learning. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 9800–9807. IEEE, 2023.

[91] Elahe Vahdani and Yingli Tian. Deep learning-based action detection in untrimmed videos: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4302–4320, 2022.

[92] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970, 2015.

[93] Hilde Kuehne, Ali Arslan, and Thomas Serre. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 780–787, 2014.

[94] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

[95] Xu Zhao, Wenchao Ding, Yongqi An, Yinglong Du, Tao Yu, Min Li, Ming Tang, and Jinqiao Wang. Fast segment anything. *arXiv preprint arXiv:2306.12156*, 2023.

[96] Jianxin Li, Guannan Si, Xinyu Liang, Zhaoliang An, Pengxin Tian, Fengyu Zhou, and Xiaoliang Wang. Multimodal fusion via voting network for 3d object detection in indoors. *Pattern Recognition*, page 111501, 2025.

[97] Heng Hu, Sibao Chen, Zhihui You, and Jin Tang. Fsenet: Feature suppression and enhancement network for tiny object detection. *Pattern Recognition*, page 111425, 2025.

[98] Yazan Abu Farha, Alexander Richard, and Juergen Gall. When will you do what?-anticipating temporal occurrences of activities. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5343–5352, 2018.

[99] Young Hwi Kim, Seonghyeon Nam, and Seon Joo Kim. Temporally smooth online action detection using cycle-consistent future anticipation. *Pattern Recognition*, 116:107954, 2021.

[100] Lingzhi Zhang, Shenghao Zhou, Simon Stent, and Jianbo Shi. Fine-grained egocentric hand-object segmentation: Dataset, model, and applications. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIX*, pages 127–145. Springer, 2022.

[101] Dima Damen, Teesid Leelasawassuk, and Walterio Mayol-Cuevas. You-do, i-learn: Egocentric unsupervised discovery of objects and their modes of interaction towards video-based guidance. *Computer Vision and Image Understanding*, 149:98–112, 2016.

[102] David Morilla-Cabello, Lorenzo Mur-Labadia, Ruben Martinez-Cantin, and Eduardo Montijano. Robust fusion for Bayesian semantic mapping. *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2023.

[103] Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning (ICML)*, pages 1050–1059. PMLR, 2016.

[104] Nikita Durasov, Timur Bagautdinov, Pierre Baque, and Pascal Fua. Masksembles for uncertainty estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13539–13548, 2021.

[105] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

[106] Gao Huang, Yixuan Li, Geoff Pleiss, Zhuang Liu, John E Hopcroft, and Kilian Q Weinberger. Snapshot ensembles: Train 1, get m for free. *International Conference on Learning Representations (ICLR)*, 2017.

[107] James J Gibson. *The ecological approach to visual perception: classic edition*. Psychology Press, 2014.

[108] Luis Montesano, Manuel Lopes, Alexandre Bernardino, and José Santos-Victor. Learning object affordances: from sensory–motor coordination to imitation. *IEEE Transactions on Robotics*, 24(1):15–26, 2008.

[109] Hema Swetha Koppula, Rudhir Gupta, and Ashutosh Saxena. Learning human activities and object affordances from rgb-d videos. *The International journal of robotics research*, 32(8):951–970, 2013.

[110] Fu-Jen Chu, Ruinian Xu, and Patricio A Vela. Learning affordance segmentation for real-world robotic manipulation via synthetic images. *IEEE Robotics and Automation Letters*, 4(2):1140–1147, 2019.

[111] Lorenzo Mur-Labadia, Ruben Martinez-Cantin, and Jose.J Guerrero. Bayesian deep learning for affordance segmentation in images. In *2023 International Conference on Robotics and Automation (ICRA)*. IEEE, 2023.

[112] Joya Chen, Difei Gao, Kevin Qinghong Lin, and Mike Zheng Shou. Affordance grounding from demonstration video to target image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6799–6808, 2023.

[113] Wei Zhai, Hongchen Luo, Jing Zhang, Yang Cao, and Dacheng Tao. One-shot object affordance detection in the wild. *International Journal of Computer Vision*, 130(10):2472–2500, 2022.

[114] Soopil Kim, Philip Chikontwe, Sion An, and Sang Hyun Park. Uncertainty-aware semi-supervised few shot segmentation. *Pattern Recognition*, 137:109292, 2023.

[115] Sai Harsha Yelleni, Deepshikha Kumari, et al. Monte carlo dropblock for modeling uncertainty in object detection. *Pattern Recognition*, 146:110003, 2024.

[116] Zihan Li, Dihan Li, Cangbai Xu, Weice Wang, Qingqi Hong, Qingde Li, and Jie Tian. Tfcns: A cnn-transformer hybrid network for medical image segmentation. In *International conference on artificial neural networks*, pages 781–792. Springer, 2022.

[117] Chen Wang, Xiang Wang, Jiawei Zhang, Liang Zhang, Xiao Bai, Xin Ning, Jun Zhou, and Edwin Hancock. Uncertainty estimation for stereo matching based on evidential deep learning. *pattern recognition*, 124:108498, 2022.

[118] Qingsen Yan, Haishen Wang, Yifan Ma, Yuhang Liu, Wei Dong, Marcin Woźniak, and Yanning Zhang. Uncertainty estimation in hdr imaging with bayesian neural networks. *Pattern Recognition*, 156:110802, 2024.

[119] Pham Thanh Huu, Nguyen Thai An, and Nguyen Ngoc Trung. Contextual and uncertainty-aware approach for multi-person pose estimation. *Pattern Recognition*, page 111454, 2025.

[120] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018.

[121] Jishnu Mukhoti, Andreas Kirsch, Joost van Amersfoort, Philip H. S. Torr, and Yarin Gal. Deterministic neural networks with appropriate inductive biases capture epistemic and aleatoric uncertainty. *International Conference on Machine Learning (ICML) Workshop*, 2021.

[122] Jishnu Mukhoti, Joost van Amersfoort, Philip H. S. Torr, and Yarin Gal. Deep deterministic uncertainty: A new simple baseline. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[123] Janis Postels, Hermann Blum, Cesar Cadena, Roland Siegwart, Luc Van Gool, and Federico Tombari. Quantifying aleatoric and epistemic uncertainty using density estimation in latent space. *Proceedings of the 39th Conference on Uncertainty in Artificial Intelligence*, 2023.

[124] Joost Van Amersfoort, Lewis Smith, Yee Whye Teh, and Yarin Gal. Uncertainty estimation using a single deep deterministic neural network. In *International conference on machine learning*, pages 9690–9700. PMLR, 2020.

[125] Karthik Abinav Sankararaman, Sinong Wang, and Han Fang. Bayesformer: Transformer with uncertainty estimation. *arXiv preprint arXiv:2206.00826*, 2022.

[126] Adam Gleave and Geoffrey Irving. Uncertainty estimation for language reward models. *arXiv preprint arXiv:2203.07472*, 2022.

[127] Fan Yang, Qiang Zhai, Xin Li, Rui Huang, Ao Luo, Hong Cheng, and Deng-Ping Fan. Uncertainty-guided transformer reasoning for camouflaged object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4146–4155, 2021.

[128] Hongji Guo, Hanjing Wang, and Qiang Ji. Uncertainty-guided probabilistic transformer for complex action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20052–20061, 2022.

[129] Hongji Guo, Zhou Ren, Yi Wu, Gang Hua, and Qiang Ji. Uncertainty-based spatial-temporal attention for online action detection. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IV*, pages 69–86. Springer, 2022.

[130] Yongchan Kwon, Joong-Ho Won, Beom Joon Kim, and Myunghee Cho Paik. Uncertainty quantification using Bayesian neural networks in classification: Ap-

plication to biomedical image segmentation. *Computational Statistics & Data Analysis*, 142:106816, 2020.

[131] Kumar Shridhar, Felix Laumann, and Marcus Liwicki. Uncertainty estimations by softplus normalization in Bayesian convolutional neural networks with variational inference. *arXiv preprint arXiv:1806.05978*, 2018.

[132] Dimity Miller, Niko Sünderhauf, Michael Milford, and Feras Dayoub. Uncertainty for identifying open-set errors in visual object detection. *IEEE Robotics and Automation Letters*, 7(1):215–222, 2021.

[133] Di Feng, Ali Harakeh, Steven L Waslander, and Klaus Dietmayer. A review and comparative study on probabilistic object detection in autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, 2021.

[134] Di Feng, Zining Wang, Yiyang Zhou, Lars Rosenbaum, Fabian Timm, Klaus Dietmayer, Masayoshi Tomizuka, and Wei Zhan. Labels are not perfect: Inferring spatial uncertainty in object detection. *IEEE Transactions on Intelligent Transportation Systems*, 2021.

[135] Gaston Lenczner, Adrien Chan-Hon-Tong, Bertrand Le Saux, Nicola Luminari, and Guy Le Besnerais. Dial: Deep interactive and active learning for semantic segmentation in remote sensing. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15:3376–3389, 2022.

[136] Julius Rückin, Liren Jin, Federico Magistri, Cyrill Stachniss, and Marija Popović. Informative path planning for active learning in aerial semantic mapping. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 11932–11939. IEEE, 2022.

[137] Georgios Georgakis, Bernadette Bucher, Karl Schmeckpeper, Siddharth Singh, and Kostas Daniilidis. Learning to map for active semantic goal navigation. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022.

[138] Alex Kendall and Yarin Gal. What uncertainties do we need in Bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.

[139] Lars Kai Hansen and Peter Salamon. Neural network ensembles. *IEEE transactions on pattern analysis and machine intelligence*, 12(10):993–1001, 1990.

[140] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10012–10022, 2021.

[141] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2117–2125, 2017.

[142] Dimity Miller, Feras Dayoub, Michael Milford, and Niko Sünderhauf. Evaluating merging strategies for sampling-based uncertainty techniques in object detection. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 2348–2354. IEEE, 2019.

[143] Dimity Miller, Lachlan Nicholson, Feras Dayoub, and Niko Sünderhauf. Dropout sampling for robust object detection in open-set conditions. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3243–3249. IEEE, 2018.

[144] D Morrison, A Milan, and E Antonakos. Uncertainty-aware instance segmentation using dropout sampling. In *Proceedings of the Robotic Vision Probabilistic Object Detection Challenge (CVPR 2019 Workshop), Long Beach, CA, USA*, pages 16–20, 2019.

[145] Ran Margolin, Lihi Zelnik-Manor, and Ayellet Tal. How to evaluate foreground maps? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2014.

[146] David Hall, Feras Dayoub, John Skinner, Haoyang Zhang, Dimity Miller, Peter Corke, Gustavo Carneiro, Anelia Angelova, and Niko Sünderhauf. Probabilistic object detection: Definition and evaluation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1031–1040, 2020.

[147] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

[148] Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using Bayesian binning. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[149] Glenn W Brier et al. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3, 1950.

[150] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1492–1500, 2017.

[151] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[152] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations, ICLR*, 2019.

[153] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proceedings of the IEEE/CVF European Conference on Computer Vision*, pages 740–755. Springer, 2014.

[154] Xue Zhao, Yang Cao, and Yu Kang. Object affordance detection with relationship-aware network. *Neural Computing and Applications*, 32(18):14321–14333, 2020.

[155] Congcong Yin and Qiuju Zhang. Object affordance detection with boundary-preserving network for robotic manipulation tasks. *Neural Computing and Applications*, 34(20):17963–17980, 2022.

[156] Yang Zhang, Huiyong Li, Tao Ren, Yuanbo Dou, and Qingfeng Li. Multi-scale fusion and global semantic encoding for affordance detection. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2022.

[157] Qipeng Gu, Jianhua Su, and Lei Yuan. Visual affordance detection using an efficient attention convolutional neural network. *Neurocomputing*, 440:36–44, 2021.

[158] Rodrigo Quesada and Yiannis Demiris. Holo-spok: Affordance-aware augmented reality control of legged manipulators. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 856–862. IEEE, 2022.

[159] Rodrigo Chacón Quesada and Yiannis Demiris. Proactive robot assistance: Affordance-aware augmented reality user interfaces. *IEEE Robotics & Automation Magazine*, 29(1):22–34, 2022.

[160] Tushar Nagarajan, Santhosh Kumar Ramakrishnan, Ruta Desai, James Hillis, and Kristen Grauman. Egocentric scene context for human-centric environment understanding from video. *arXiv preprint arXiv:2207.11365*, 2022.

[161] Santhosh Kumar Ramakrishnan, Tushar Nagarajan, Ziad Al-Halah, and Kristen Grauman. Environment predictive coding for visual navigation. In *International Conference on Learning Representations*, 2021.

[162] Daniel Maturana, Po-Wei Chou, Masashi Uenoyama, and Sebastian Scherer. Real-time semantic mapping for autonomous off-road navigation. In *Field and Service Robotics: Results of the 11th International Conference*, pages 335–350. Springer, 2018.

[163] Zecheng Yu, Yifei Huang, Ryosuke Furuta, Takuma Yagi, Yusuke Goutsu, and Yoichi Sato. Fine-grained affordance annotation for egocentric hand-object interaction videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2155–2163, 2023.

[164] Yan Zhu, Yuandong Tian, Dimitris Metaxas, and Piotr Dollár. Semantic amodal segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1464–1472, 2017.

[165] Rohit Mohan and Abhinav Valada. Amodal panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21023–21032, 2022.

[166] Ruining Deng, Quan Liu, Can Cui, Zuhayr Asad, Yuankai Huo, et al. Single dynamic network for multi-label renal pathology image segmentation. In *International Conference on Medical Imaging with Deep Learning*, pages 304–314. PMLR, 2022.

[167] Michael Lempart, Martin P Nilsson, Jonas Scherman, Christian Jamtheim Gustafsson, Mikael Nilsson, Sara Alkner, Jens Engleson, Gabriel Adrian, Per Munck af Rosenschöld, and Lars E Olsson. Pelvic u-net: multi-label semantic segmentation of pelvic organs at risk for radiation therapy anal cancer patients using a deeply supervised shuffle attention convolutional neural network. *Radiation Oncology*, 17(1):1–15, 2022.

[168] Savinien Bonheur, Darko Štern, Christian Payer, Michael Pienn, Horst Olschewski, and Martin Urschler. Matwo-capsnet: a multi-label semantic seg-

mentation capsules network. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 664–672. Springer, 2019.

[169] Minho Lee, JeeYoung Kim, Regina EY Kim, Hyun Gi Kim, Se Won Oh, Min Kyoung Lee, Sheng-Min Wang, Nak-Young Kim, Dong Woo Kang, ZunHyan Rieu, et al. Split-attention u-net: a fully convolutional network for robust multi-label segmentation from brain mri. *Brain Sciences*, 10(12):974, 2020.

[170] Shilong Liu, Lei Zhang, Xiao Yang, Hang Su, and Jun Zhu. Query2label: A simple transformer way to multi-label classification. *arXiv preprint arXiv:2107.10834*, 2021.

[171] Tong Wu, Qingqiu Huang, Ziwei Liu, Yu Wang, and Dahua Lin. Distribution-balanced loss for multi-label classification in long-tailed datasets. In *European Conference on Computer Vision*, pages 162–178. Springer, 2020.

[172] Tal Ridnik, Emanuel Ben Baruch, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. Asymmetric loss for multi-label classification. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 82–91. IEEE, 2021.

[173] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[174] Zongyuan Ge, Dwarikanath Mahapatra, Suman Sedai, Rahil Garnavi, and Rajib Chakravorty. Chest x-rays classification: A multi-label and fine-grained problem. *arXiv preprint arXiv:1807.07247*, 2018.

[175] Jin Ye, Junjun He, Xiaojiang Peng, Wenhao Wu, and Yu Qiao. Attention-driven dynamic graph convolutional network for multi-label image recognition. In *European Conference on Computer Vision*, pages 649–665. Springer, 2020.

[176] Xing Cheng, Hezheng Lin, Xiangyu Wu, Dong Shen, Fan Yang, Honglin Liu, and Nian Shi. MLTR: multi-label classification with transformer. In *IEEE International Conference on Multimedia and Expo, ICME 2022, Taipei, Taiwan, July 18-22, 2022*, pages 1–6. IEEE, 2022.

[177] Jack Lanchantin, Tianlu Wang, Vicente Ordonez, and Yanjun Qi. General multi-label image classification with transformers. In *Proceedings of the IEEE/CVF*

Conference on Computer Vision and Pattern Recognition, pages 16478–16488, 2021.

[178] Haodi He, Yuhui Yuan, Xiangyu Yue, and Han Hu. Rankseg: Adaptive pixel classification with image category ranking for segmentation. In *European Conference on Computer Vision*, pages 682–700. Springer, 2022.

[179] Paola Ardón, Èric Pairet, Ronald PA Petrick, Subramanian Ramamoorthy, and Katrin S Lohan. Learning grasp affordance reasoning through semantic relations. *IEEE Robotics and Automation Letters*, 4(4):4571–4578, 2019.

[180] Jiaqi Guan, Ye Yuan, Kris M Kitani, and Nicholas Rhinehart. Generative hybrid representations for activity forecasting with no-regret learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 173–182, 2020.

[181] Nicholas Rhinehart and Kris M Kitani. First-person activity forecasting with online inverse reinforcement learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3696–3705, 2017.

[182] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[183] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. Orb-slam: a versatile and accurate monocular slam system. *IEEE transactions on robotics*, 31(5):1147–1163, 2015.

[184] Vadim Tschernezki, Ahmad Darkhalil, Zhifan Zhu, David Fouhey, Iro Laina, Diane Larlus, Dima Damen, and Andrea Vedaldi. Epic fields: Marrying 3d geometry and video understanding. *arXiv preprint arXiv:2306.08731*, 2023.

[185] Hema S Koppula and Ashutosh Saxena. Physically grounded spatio-temporal object affordances. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part III 13*, pages 831–847. Springer, 2014.

[186] Claudio Castellini, Tatiana Tommasi, Nicoletta Noceti, Francesca Odone, and Barbara Caputo. Using object affordances to improve object recognition. *IEEE transactions on autonomous mental development*, 3(3):207–215, 2011.

[187] Alireza Fathi, Yin Li, James M Rehg, et al. Learning to recognize daily actions using gaze. *ECCV (1)*, 7572:314–327, 2012.

[188] Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma, Amlan Kar, Richard Higgins, Sanja Fidler, David Fouhey, and Dima Damen. Epic-kitchens visor benchmark: Video segmentations and object relations. *Advances in Neural Information Processing Systems*, 35:13745–13758, 2022.

[189] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise View Selection for Unstructured Multi-View Stereo. In *European Conference on Computer Vision (ECCV)*, 2016.

[190] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3), 2022.

[191] Maria Klodt and Andrea Vedaldi. Supervising the new with the old: learning sfm from sfm. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 698–713, 2018.

[192] Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision transformer adapter for dense predictions. *arXiv preprint arXiv:2205.08534*, 2022.

[193] Guosheng Lin, Chunhua Shen, Anton Van Den Hengel, and Ian Reid. Efficient piecewise training of deep structured models for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3194–3203, 2016.

[194] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017.

[195] Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo. Multi-label image recognition with graph convolutional networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5177–5186, 2019.

[196] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9627–9636, 2019.

[197] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019.

[198] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[199] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9404–9413, 2019.

[200] Lu Qi, Li Jiang, Shu Liu, Xiaoyong Shen, and Jiaya Jia. Amodal instance segmentation with kins dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3014–3023, 2019.

[201] Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand. What do different evaluation metrics tell us about saliency models? *IEEE transactions on pattern analysis and machine intelligence*, 41(3):740–757, 2018.

[202] Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba. Learning to predict where humans look. In *2009 IEEE 12th international conference on computer vision*, pages 2106–2113. IEEE, 2009.

[203] João Guerreiro, Daisuke Sato, Dragan Ahmetovic, Eshed Ohn-Bar, Kris M Kitani, and Chieko Asakawa. Virtual navigation for blind people: Transferring route knowledge to the real-world. *International Journal of Human-Computer Studies*, 135:102369, 2020.

[204] Shengqi Duan, Guohui Tian, Zhongli Wang, Shaopeng Liu, and Chenrui Feng. A semantic robotic grasping framework based on multi-task learning in stacking scenes. *Engineering Applications of Artificial Intelligence*, 121:106059, 2023.

[205] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 3354–3361, 2012.

[206] Zhongli Wang, Guohui Tian, and Xuyang Shao. Home service robot task planning using semantic knowledge and probabilistic inference. *Knowledge-Based Systems*, 204:106174, 2020.

[207] Niko Sünderhauf, Oliver Brock, Walter Scheirer, Raia Hadsell, Dieter Fox, Jürgen Leitner, Ben Upcroft, Pieter Abbeel, Wolfram Burgard, Michael Milford, et al. The limits and potentials of deep learning for robotics. *The International journal of robotics research*, 37(4-5):405–420, 2018.

[208] John McCormac, Ankur Handa, Andrew Davison, and Stefan Leutenegger. SemanticFusion: Dense 3D semantic mapping with convolutional neural networks. In *IEEE Int. Conf. on Robotics and Automation*, pages 4628–4635, 2017.

[209] Niko Sunderhauf, Trung T. Pham, Yasir Latif, Michael Milford, and Ian Reid. Meaningful maps with object-oriented semantic mapping. In *IEEE Int. Conf. on Intelligent Robots and Systems*, pages 5079–5085, 2017.

[210] Antoni Rosinol, Marcus Abate, Yun Chang, and Luca Carlone. Kimera: An Open-Source Library for Real-Time Metric-Semantic Localization and Mapping. In *IEEE Int. Conf. on Robotics and Automation*, pages 1689–1696, may 2020.

[211] Arash Asgharivaskasi and Nikolay Atanasov. Active Bayesian Multi-class Mapping from Range and Semantic Segmentation Observations. In *IEEE Int. Conf. on Robotics and Automation*, pages 1–7, 2021.

[212] Devendra Singh Chaplot, Murtaza Dalal, Saurabh Gupta, Jitendra Malik, and Russ R Salakhutdinov. Seal: Self-supervised embodied active learning using exploration and 3d consistency. *Adv. in neural information proc. systems*, 34:13086–13098, 2021.

[213] David Nilsson, Aleksis Pirinen, Erik Gärtner, and Cristian Sminchisescu. Embodied visual active learning for semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2373–2383, 2021.

[214] René Zurbrügg, Hermann Blum, Cesar Cadena, Roland Siegwart, and Lukas Schmid. Embodied active domain adaptation for semantic segmentation via informative path planning. *IEEE Robotics and Automation Letters*, 7(4):8691–8698, 2022.

[215] Yu Xiang and Dieter Fox. DA-RNN: Semantic Mapping with Data Associated Recurrent Neural Networks. In *Robotics: Science and Systems*, volume 13, 2017.

[216] Zhiliu Yang and Chen Liu. TUPPer-Map: Temporal and Unified Panoptic Perception for 3D Metric-Semantic Mapping. In *IEEE Int. Conf. on Intelligent Robots and Systems*, pages 1094–1101, 2021.

[217] Margarita Grinvald, Fadri Furrer, Tonci Novkovic, Jen Jen Chung, Cesar Cadena, Roland Siegwart, and Juan Nieto. Volumetric instance-aware semantic mapping and 3d object discovery. *IEEE Robotics and Automation Letters*, 4:3037–3044, 2019.

[218] Lukas Schmid, Jeffrey Delmerico, Johannes L. Schonberger, Juan Nieto, Marc Pollefeys, Roland Siegwart, and Cesar Cadena. Panoptic Multi-TSDFs: a Flexible Representation for Online Multi-resolution Volumetric Mapping and Long-term Dynamic Scene Consistency. In *IEEE ICRA*, pages 8018–8024, 2022.

[219] John McCormac, Ronald Clark, Michael Bloesch, Andrew Davison, and Stefan Leutenegger. Fusion++: Volumetric Object-Level SLAM. In *Int. Conf. on 3D Vision*, pages 32–41, 2018.

[220] Gaku Narita, Takashi Seno, Tomoya Ishikawa, and Yohsuke Kaji. PanopticFusion: Online Volumetric Semantic Mapping at the Level of Stuff and Things. In *IEEE International Conference on Intelligent Robots and Systems*, pages 4205–4212. Institute of Electrical and Electronics Engineers Inc., nov 2019.

[221] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Int. conf. on machine learning*, pages 1050–1059, 2016.

[222] Fredrik K Gustafsson, Martin Danelljan, and Thomas B Schon. Evaluating scalable Bayesian deep learning methods for robust computer vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 318–319, 2020.

[223] Jishnu Mukhoti, Andreas Kirsch, Joost van Amersfoort, Philip H.S. Torr, and Yarin Gal. Deep deterministic uncertainty: A new simple baseline. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, pages 24384–24394, June 2023.

[224] Yuri Feldman and Vadim Indelman. Bayesian viewpoint-dependent robust classification under model and localization uncertainty. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3221–3228, 2018.

[225] Vladimir Tchuiev and Vadim Indelman. Epistemic uncertainty aware semantic localization and mapping for inference and belief space planning. *Artificial Intelligence*, 319:103903, 2023.

[226] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Comp. Vision and Pattern Recog. (CVPR), IEEE*, 2017.

[227] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *CoRR*, abs/1706.05587, 2017.

[228] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012).

[229] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012.

[230] Minghuang Ma, Haoqi Fan, and Kris M Kitani. Going deeper into first-person activity recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1894–1903, 2016.

[231] Swathikiran Sudhakaran, Sergio Escalera, and Oswald Lanz. Lsta: Long short-term attention for egocentric action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9954–9963, 2019.

[232] Lorenzo Mur-Labadia, Ruben Martinez-Cantin, Jose J Guerrero, Giovanni Maria Farinella, and Antonino Furnari. Aff-ttention! affordances and attention models for short-term object interaction anticipation. In *European Conference on Computer Vision*, pages 167–184. Springer, 2024.

[233] Hyun Soo Park, Jyh-Jing Hwang, Yedong Niu, and Jianbo Shi. Egocentric future localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4697–4705, 2016.

[234] Tushar Nagarajan, Santhosh Kumar Ramakrishnan, Ruta Desai, James Hillis, and Kristen Grauman. Egoenv: Human-centric environment representations from egocentric video. *Advances in Neural Information Processing Systems*, 36, 2024.

[235] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

[236] Justin Kerr, Chung Min Kim, Mingxuan Wu, Brent Yi, Qianqian Wang, Ken Goldberg, and Angjoo Kanazawa. Robot see robot do: Imitating articulated object manipulation with monocular 4d reconstruction. *arXiv preprint arXiv:2409.18121*, 2024.

[237] Haozhe Lou, Yurong Liu, Yike Pan, Yiran Geng, Jianteng Chen, Wenlong Ma, Chenglong Li, Lin Wang, Hengzhen Feng, Lu Shi, et al. Robo-gs: A physics consistent spatial-temporal model for robotic arm with hybrid representation. *arXiv preprint arXiv:2408.14873*, 2024.

[238] Abdelrhman Werby, Chenguang Huang, Martin Büchner, Abhinav Valada, and Wolfram Burgard. Hierarchical open-vocabulary 3d scene graphs for language-grounded robot navigation. In *First Workshop on Vision-Language Models for Navigation and Manipulation at ICRA 2024*, 2024.

[239] Sosuke Kobayashi, Eiichi Matsumoto, and Vincent Sitzmann. Decomposing nerf for editing via feature field distillation. *Advances in Neural Information Processing Systems*, 35:23311–23330, 2022.

[240] Weining Ren, Zihan Zhu, Boyang Sun, Jiaqi Chen, Marc Pollefeys, and Songyou Peng. Nerf on-the-go: Exploiting uncertainty for distractor-free nerfs in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8931–8940, 2024.

[241] Vadim Tschernezki, Iro Laina, Diane Larlus, and Andrea Vedaldi. Neural feature fusion fields: 3d distillation of self-supervised 2d image representations. In *2022 International Conference on 3D Vision (3DV)*, pages 443–453. IEEE, 2022.

[242] Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. Lerf: Language embedded radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19729–19739, 2023.

[243] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[244] Suvam Patra, Kartikeya Gupta, Faran Ahmad, Chetan Arora, and Subhashis Banerjee. Ego-slam: A robust monocular slam for egocentric videos. In *2019*

IEEE Winter Conference on Applications of Computer Vision (WACV), pages 31–40. IEEE, 2019.

[245] Fei Xia, Amir R Zamir, Zhiyang He, Alexander Sax, Jitendra Malik, and Silvio Savarese. Gibson env: Real-world perception for embodied agents. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9068–9079, 2018.

[246] Jakob Engel, Kiran Somasundaram, Michael Goesele, Albert Sun, Alexander Gamino, Andrew Turner, Arjang Talattof, Arnie Yuan, Bilal Souti, Brighid Meredith, et al. Project aria: A new tool for egocentric multi-modal ai research. *arXiv preprint arXiv:2308.13561*, 2023.

[247] Will Price, Carl Vondrick, and Dima Damen. Unweavenet: Unweaving activity stories. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13770–13779, 2022.

[248] Santhosh Kumar Ramakrishnan and Tushar Nagarajan. Environment predictive coding for visual navigation. *ICLR 2022*, 2022.

[249] Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. Neural scene flow fields for space-time view synthesis of dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6498–6508, 2021.

[250] Chen Gao, Ayush Saraf, Johannes Kopf, and Jia-Bin Huang. Dynamic view synthesis from dynamic monocular video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5712–5721, 2021.

[251] Yiqing Liang, Eliot Laidlaw, Alexander Meyerowitz, Srinath Sridhar, and James Tompkin. Semantic attention flow fields for monocular dynamic scene decomposition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21797–21806, 2023.

[252] Tianhao Wu, Fangcheng Zhong, Andrea Tagliasacchi, Forrester Cole, and Cengiz Oztireli. D^ 2nerf: Self-supervised decoupling of dynamic and static objects from a monocular video. *Advances in neural information processing systems*, 35:32653–32666, 2022.

[253] Daiwei Zhang, Gengyan Li, Jiajie Li, Mickaël Bressieux, Otmar Hilliges, Marc Pollefeys, Luc Van Gool, and Xi Wang. Egogaussian: Dynamic scene un-

derstanding from egocentric video with 3d gaussian splatting. *arXiv preprint arXiv:2406.19811*, 2024.

[254] Shuaifeng Zhi, Tristan Laidlow, Stefan Leutenegger, and Andrew J Davison. In-place scene labelling and understanding with implicit scene representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15838–15847, 2021.

[255] Suhani Vora, Noha Radwan, Klaus Greff, Henning Meyer, Kyle Genova, Mehdi SM Sajjadi, Etienne Pot, Andrea Tagliasacchi, and Daniel Duckworth. Nesf: Neural semantic fields for generalizable semantic segmentation of 3d scenes. *arXiv preprint arXiv:2111.13260*, 2021.

[256] Yawar Siddiqui, Lorenzo Porzi, Samuel Rota Buló, Norman Müller, Matthias Nießner, Angela Dai, and Peter Kontschieder. Panoptic lifting for 3d scene understanding with neural fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9043–9052, 2023.

[257] Xiao Fu, Shangzhan Zhang, Tianrun Chen, Yichong Lu, Lanyun Zhu, Xiaowei Zhou, Andreas Geiger, and Yiyi Liao. Panoptic nerf: 3d-to-2d label transfer for panoptic urban scene segmentation. In *2022 International Conference on 3D Vision (3DV)*, pages 1–11. IEEE, 2022.

[258] Rahul Goel, Dhawal Sirikonda, Saurabh Saini, and PJ Narayanan. Interactive segmentation of radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4201–4211, 2023.

[259] Yuanwen Yue, Anurag Das, Francis Engelmann, Siyu Tang, and Jan Eric Lenssen. Improving 2d feature representations by 3d-aware fine-tuning. In *European Conference on Computer Vision*, pages 57–74. Springer, 2025.

[260] Yash Bhalgat, Iro Laina, João F Henriques, Andrew Zisserman, and Andrea Vedaldi. N2f2: Hierarchical scene understanding with nested neural feature fields. *arXiv preprint arXiv:2403.10997*, 2024.

[261] Minghan Qin, Wanhua Li, Jiawei Zhou, Haoqian Wang, and Hanspeter Pfister. Langsplat: 3d language gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20051–20060, 2024.

[262] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023.

[263] Qiao Gu, Zhaoyang Lv, Duncan Frost, Simon Green, Julian Straub, and Chris Sweeney. Egolifter: Open-world 3d segmentation for egocentric perception. In *European Conference on Computer Vision*, pages 382–400. Springer, 2025.

[264] Vadim Tschernezki, Diane Larlus, and Andrea Vedaldi. Neuraldiff: Segmenting 3d objects that move in egocentric videos. In *2021 International Conference on 3D Vision (3DV)*, pages 910–919. IEEE, 2021.

[265] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829, 2023.

[266] Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. Scaling open-vocabulary object detection. *Advances in Neural Information Processing Systems*, 36, 2024.

[267] Chiara Plizzari, Gabriele Goletto, Antonino Furnari, Siddhant Bansal, Francesco Ragusa, Giovanni Maria Farinella, Dima Damen, and Tatiana Tommasi. An outlook into the future of egocentric vision. *arXiv preprint arXiv:2308.07123*, 2023.

[268] Ivan Rodin, Antonino Furnari, Dimitrios Mavroeidis, and Giovanni Maria Farinella. Predicting the future from first person (egocentric) vision: A survey. *Computer Vision and Image Understanding*, 211:103252, 2021.

[269] Debaditya Roy, Ramanathan Rajendiran, and Basura Fernando. Interaction region visual transformer for egocentric action anticipation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6740–6750, 2024.

[270] Antonino Furnari and Giovanni Maria Farinella. Rolling-unrolling lstms for action anticipation from first-person video. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):4021–4036, 2020.

[271] Hyung-gun Chi, Kwonjoon Lee, Nakul Agarwal, Yi Xu, Karthik Ramani, and Chiho Choi. Adamsformer for spatial action localization in the future. In *Proceed-*

ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 17885–17895, 2023.

[272] Megha Nawhal, Akash Abdu Jyothi, and Greg Mori. Rethinking learning approaches for long-term action anticipation. In *European Conference on Computer Vision*, pages 558–576. Springer, 2022.

[273] Zeyun Zhong, David Schneider, Michael Voit, Rainer Stiefelhagen, and Jürgen Beyerer. Anticipative feature fusion transformer for multi-modal action anticipation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6068–6077, 2023.

[274] Olga Zatsarynna, Yazan Abu Farha, and Juergen Gall. Multi-modal temporal convolutional network for anticipating actions in egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2249–2258, 2021.

[275] Yong Jae Lee, Joydeep Ghosh, and Kristen Grauman. Discovering important people and objects for egocentric video summarization. In *2012 IEEE conference on computer vision and pattern recognition*, pages 1346–1353. IEEE, 2012.

[276] Huikun Bi, Ruisi Zhang, Tianlu Mao, Zhigang Deng, and Zhaoqi Wang. How can i see my future? fvtraj: Using first-person view for pedestrian trajectory prediction. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*, pages 576–593. Springer, 2020.

[277] Wentao Bao, Lele Chen, Libing Zeng, Zhong Li, Yi Xu, Junsong Yuan, and Yu Kong. Uncertainty-aware state space transformer for egocentric 3d hand trajectory forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13702–13711, 2023.

[278] Eadom Dessalene, Chinmaya Devaraj, Michael Maynord, Cornelia Fermuller, and Yiannis Aloimonos. Forecasting action through contact representations from first person video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[279] Guo Chen, Sen Xing, Zhe Chen, Yi Wang, Kunchang Li, Yizhuo Li, Yi Liu, Jiahao Wang, Yin-Dong Zheng, Bingkun Huang, et al. Internvideo-ego4d: A pack of champion solutions to ego4d challenges. *arXiv preprint arXiv:2211.09529*, 2022.

[280] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022.

[281] Razvan-George Pasca, Alexey Gavryushin, Yen-Ling Kuo, Otmar Hilliges, and Xi Wang. Summarize the past to predict the future: Natural language descriptions of context boost multimodal object interaction. *arXiv preprint arXiv:2301.09209*, 2023.

[282] Francesco Ragusa, Giovanni Maria Farinella, and Antonino Furnari. Stillfast: An end-to-end approach for short-term object interaction anticipation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3635–3644, 2023.

[283] Sanket Thakur, Cigdem Beyan, Pietro Morerio, Vittorio Murino, and Alessio Del Bue. Enhancing next active object-based egocentric action anticipation with guided attention. *In International Conference on Image Processing*, 2023.

[284] Sanket Thakur, Cigdem Beyan, Pietro Morerio, Vittorio Murino, and Alessio Del Bue. Guided attention for next active object@ ego4d sta challenge. *CVPR23 EGO4D Workshop STA Challenge*, 2023.

[285] Sanket Thakur, Cigdem Beyan, Pietro Morerio, Vittorio Murino, and Alessio Del Bue. Leveraging next-active objects for context-aware anticipation in egocentric videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 8657–8666, 2024.

[286] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, volume 2, page 4, 2021.

[287] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[288] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019.

[289] Lorenzo Mur-Labadia, Ruben Martinez-Cantin, and Jose J Guerrero. Bayesian deep learning for affordance segmentation in images. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 6981–6987, 2023.

[290] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016.

[291] Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, et al. Hybrid task cascade for instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4974–4983, 2019.

[292] J Redmon. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[293] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.

[294] Shilong Liu, Feng Li, Hao Zhang, Xiao Yang, Xianbiao Qi, Hang Su, Jun Zhu, and Lei Zhang. Dab-detr: Dynamic anchor boxes are better queries for detr. *arXiv preprint arXiv:2201.12329*, 2022.

[295] Feng Li, Hao Zhang, Shilong Liu, Jian Guo, Lionel M Ni, and Lei Zhang. Dn-detr: Accelerate detr training by introducing query denoising. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13619–13627, 2022.

[296] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022.

[297] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperPoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 224–236, 2018.

[298] Ego4D Team. Short-Term object-interaction Anticipation quickstart. `https://colab.research.google.com/drive/1Ok_6F1O6K8kX1S4sEnU62HoOBw_CPngR?usp=sharing`, 2023. [Online; accessed 03-March-2024].

[299] Lorenzo Mur-Labadia, Ruben Martinez-Cantin, Jose J Guerrero, Giovanni Maria Farinella, and Antonino Furnari. Aff-ttention! affordances and attention models for short-term object interaction anticipation. In *European Conference on Computer Vision*, pages 167–184. Springer, 2025.

[300] Michele Mazzamuto, Francesco Ragusa, Antonino Furnari, Irene D'Ambra, Antonia Guarriera, Armando Sorbello, and Giovanni Maria Farinella. A mixed reality application to help impaired people rehabilitate outside clinical environments. In *2024 IEEE International Conference on Metrology for eXtended Reality, Artificial Intelligence and Neural Engineering (MetroXRAINE)*, pages 42–47. IEEE, 2024.

[301] Alessandro Flaborea, Guido Maria D'Amely Di Melendugno, Leonardo Plini, Luca Scofano, Edoardo De Matteis, Antonino Furnari, Giovanni Maria Farinella, and Fabio Galasso. Prego: online mistake detection in procedural egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18483–18492, 2024.

[302] Oier Mees, Lukas Hermann, Erick Rosete-Beas, and Wolfram Burgard. Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. *IEEE Robotics and Automation Letters (RA-L)*, 7(3):7327–7334, 2022.

[303] Yusuf Aytar, Tobias Pfaff, David Budden, Thomas Paine, Ziyu Wang, and Nando De Freitas. Playing hard exploration games by watching youtube. *Advances in neural information processing systems*, 31, 2018.

[304] Bowen Baker, Ilge Akkaya, Peter Zhokov, Joost Huizinga, Jie Tang, Adrien Ecoffet, Brandon Houghton, Raul Sampedro, and Jeff Clune. Video pretraining (vpt): Learning to act by watching unlabeled online videos. *Advances in Neural Information Processing Systems*, 35:24639–24654, 2022.

[305] Yilun Du, Sherry Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Josh Tenenbaum, Dale Schuurmans, and Pieter Abbeel. Learning universal policies via text-guided video generation. *Advances in Neural Information Processing Systems*, 36, 2024.

[306] Alejandro Escontrela, Ademi Adeniji, Wilson Yan, Ajay Jain, Xue Bin Peng, Ken Goldberg, Youngwoon Lee, Danijar Hafner, and Pieter Abbeel. Video prediction models as rewards for reinforcement learning. *Advances in Neural Information Processing Systems*, 36, 2024.

[307] Karl Schmeckpeper, Oleh Rybkin, Kostas Daniilidis, Sergey Levine, and Chelsea Finn. Reinforcement learning with videos: Combining offline observations with interaction. In *Conference on Robot Learning*, pages 339–354. PMLR, 2021.

[308] Yiyi Liao, Jun Xie, and Andreas Geiger. Kitti-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3292–3310, 2022.

[309] Jensen Gao, Bidipta Sarkar, Fei Xia, Ted Xiao, Jiajun Wu, Brian Ichter, Anirudha Majumdar, and Dorsa Sadigh. Physically grounded vision-language models for robotic manipulation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 12462–12469. IEEE, 2024.

[310] Zhihao Cao, Zidong Wang, Siwen Xie, Anji Liu, and Lifeng Fan. Smart help: Strategic opponent modeling for proactive and adaptive robot assistance in households. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18091–18101, 2024.

[311] Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, et al. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*, 2022.

[312] Z. et al. Yi. Unified fully and timestamp supervised temporal action segmentation via sequence-to-sequence learning. In *European Conference on Computer Vision (ECCV)*, 2022.

[313] Zijia Lu and Ehsan Elhamifar. Fact: Frame-action cross-attention temporal modeling for efficient action segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18175–18185, June 2024.

[314] Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. Span-based localizing network for natural language video localization. In *Proc. of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6543–6554. Association for Computational Linguistics, July 2020.

[315] Claudia Bonanno, Francesco Ragusa, Antonino Furnari, and Giovanni Maria Farinella. Hero: A multi-modal approach on mobile devices for visual-aware conversational assistance in industrial domains. In *International Conference on Image Analysis and Processing*, pages 424–436. Springer, 2023.

[316] Genci Capi, Mitsuki Kitani, and K Ueki. Guide robot intelligent navigation in urban environments. *Advanced Robotics*, 28(15):1043–1053, 2014.

[317] Benita Wong, Joya Chen, You Wu, Stan Weixian Lei, Dongxing Mao, Difei Gao, and Mike Zheng Shou. Assistq: Affordance-centric question-driven task completion for egocentric assistant. In *European Conference on Computer Vision*, pages 485–501. Springer, 2022.

[318] Andrew N Meltzoff. The role of imitation in understanding persons and developing a theory of mind. *Understanding other minds: perspectives from autism.*, pages 335–366, 1993.

[319] Matthew Chang, Arjun Gupta, and Saurabh Gupta. Semantic visual navigation by watching youtube videos. *Advances in Neural Information Processing Systems*, 33:4283–4294, 2020.

[320] Pierre Sermanet, Tianli Ding, Jeffrey Zhao, Fei Xia, Debidatta Dwibedi, Keerthana Gopalakrishnan, Christine Chan, Gabriel Dulac-Arnold, Sharath Maddineni, Nikhil J Joshi, Pete Florence, Wei Han, Robert Baruch, Yao Lu, Suvir Mirchandani, Peng Xu, Pannag Sanketi, Karol Hausman, Izhak Shafran, Brian Ichter, and Yuan Cao. RoboVQA: Multimodal long-horizon reasoning for robotics. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 645–652, 2024.

[321] Divya Kothandaraman, Ming Lin, and Dinesh Manocha. Diffar: Differentiable frequency-based disentanglement for aerial video action recognition. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8254–8261. IEEE, 2023.

[322] Xijun Wang, Ruiqi Xian, Tianrui Guan, Celso M de Melo, Stephen M Nogar, Aniket Bera, and Dinesh Manocha. Aztr: Aerial video action recognition with auto zoom and temporal reasoning. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1312–1318. IEEE, 2023.

[323] Jiachen Li, Xinwei Shi, Feiyu Chen, Jonathan Stroud, Zhishuai Zhang, Tian Lan, Junhua Mao, Jeonhyung Kang, Khaled S Refaat, Weilong Yang, et al. Pedestrian crossing action recognition and trajectory prediction with 3d human keypoints. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1463–1470. IEEE, 2023.

[324] Yung-Chi Kung, Arthur Zhang, Junmin Wang, and Joydeep Biswas. Looking inside out: Anticipating driver intent from videos. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5608–5614. IEEE, 2024.

[325] Minseok Seo, Donghyeon Cho, Sangwoo Lee, Jongchan Park, Daehan Kim, Jaemin Lee, Jingi Ju, Hyeoncheol Noh, and Dong-Geol Choi. A self-supervised sampler for efficient action recognition: Real-world applications in surveillance systems. *IEEE Robotics and Automation Letters*, 7(2):1752–1759, 2021.

[326] Jinrong Yang, En Yu, Zeming Li, Xiaoping Li, and Wenbing Tao. Quality matters: Embracing quality clues for robust 3d multi-object tracking. *arXiv preprint arXiv:2208.10976*, 2022.

[327] Riccardo Pieroni, Simone Specchia, Matteo Corno, and Sergio Matteo Savaresi. Multi-object tracking with camera-lidar fusion for autonomous driving. *arXiv preprint arXiv:2403.04112*, 2024.

[328] Liangliang Yao, Changhong Fu, Sihang Li, Guangze Zheng, and Junjie Ye. Sgdvit: saliency-guided dynamic vision transformer for uav tracking. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3353–3359. IEEE, 2023.

[329] Alberto Dionigi, Alessandro Devo, Leonardo Guiducci, and Gabriele Costante. E-vat: An asymmetric end-to-end approach to visual active exploration and tracking. *IEEE Robotics and Automation Letters*, 7(2):4259–4266, 2022.

[330] Irene Ballester, Alejandro Fontán, Javier Civera, Klaus H Strobl, and Rudolph Triebel. Dot: Dynamic object tracking for visual slam. In *2021 IEEE international conference on robotics and automation (ICRA)*, pages 11705–11711. IEEE, 2021.

[331] Zhichao Deng, Xiangtai Li, Xia Li, Yunhai Tong, Shen Zhao, and Mengyuan Liu. Vg4d: Vision-language model goes 4d video recognition. *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024.

[332] Yunze Liu, Changxi Chen, Zifan Wang, and Li Yi. Crossvideo: Self-supervised cross-modal contrastive learning for point cloud video understanding. *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024.

[333] Mahesh Subedar, Ranganath Krishnan, Paulo Lopez Meyer, Omesh Tickoo, and Jonathan Huang. Uncertainty-aware audiovisual activity recognition using deep bayesian variational inference. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6301–6310, 2019.

[334] Hongji Guo, Hanjing Wang, and Qiang Ji. Bayesian evidential deep learning for online action detection. In *European Conference on Computer Vision*, pages 283–301. Springer, 2024.

[335] Carlos Plou, Nerea Gallego, Alberto Sabater, Eduardo Montijano, Pablo Urcola, Luis Montesano, Ruben Martinez-Cantin, and Ana C. Murillo. Eventsleep: Sleep activity recognition with event cameras, 2024.

[336] Karan Goel and Emma Brunskill. Learning procedural abstractions and evaluating discrete latent temporal structure. In *International Conference on Learning Representations*, 2019.

[337] Basura Fernando, Efstratios Gavves, Jose M Oramas, Amir Ghodrati, and Tinne Tuytelaars. Modeling video evolution for action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5378–5387, 2015.

[338] Hsin-Ying Lee, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Unsupervised representation learning by sorting sequences. In *Proceedings of the IEEE international conference on computer vision*, pages 667–676, 2017.

[339] Mahtab Jahanbani Fard, Sattar Ameri, Ratna Babu Chinnam, and R Darin Ellis. Soft boundary approach for unsupervised gesture segmentation in robotic-assisted surgery. *IEEE Robotics and Automation Letters*, 2(1):171–178, 2016.

[340] Giacomo De Rossi, Marco Minelli, Serena Roin, Fabio Falezza, Alessio Sozzi, Federica Ferraguti, Francesco Setti, Marcello Bonfè, Cristian Secchi, and Riccardo Muradore. A first evaluation of a multi-modal learning system to control surgical assistant robots via action segmentation. *IEEE Transactions on Medical Robotics and Bionics*, 3(3):714–724, 2021.

[341] Daniele Meli and Paolo Fiorini. Unsupervised identification of surgical robotic actions from small non-homogeneous datasets. *IEEE Robotics and Automation Letters*, 6(4):8205–8212, 2021.

[342] Chun Zhu and Weihua Sheng. Wearable sensor-based hand gesture and daily activity recognition for robot-assisted living. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 41(3):569–573, 2011.

[343] Tsukasa Fukuda, Yasushi Nakauchi, Katsunori Noguchi, and Takashi Matsubara. Sequential human behavior recognition for cooking-support robots. *Journal of Robotics and Mechatronics*, 17(6):717, 2005.

[344] Anna Kukleva, Hilde Kuehne, Fadime Sener, and Jurgen Gall. Unsupervised learning of action classes with continuous temporal embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12066–12074, 2019.

[345] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1728–1738, 2021.

[346] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2630–2640, 2019.

[347] Kevin Qinghong Lin, Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Z XU, Difei Gao, Rong-Cheng Tu, Wenzhe Zhao, Weijie Kong, et al. Egocentric video-language pretraining. *Advances in Neural Information Processing Systems*, 35:7575–7586, 2022.

[348] Zhijian Hou, Wanjun Zhong, Lei Ji, Difei Gao, Kun Yan, Wing-Kwong Chan, Chong-Wah Ngo, Zheng Shou, and Nan Duan. An efficient coarse-to-fine alignment framework@ ego4d natural language queries challenge 2022. *arXiv preprint arXiv:2211.08776*, 2022.

[349] Jie Lei, Tamara L Berg, and Mohit Bansal. Detecting moments and highlights in videos via natural language queries. *NeurIPS*, 34:11846–11858, 2021.

[350] Santhosh Kumar Ramakrishnan, Ziad Al-Halah, and Kristen Grauman. Naq: Leveraging narrations as queries to supervise episodic memory. In *CVPR*, pages 6694–6703, 2023.

[351] Zhijian Hou, Lei Ji, Difei Gao, Wanjun Zhong, Kun Yan, Chao Li, Wing-Kwong Chan, Chong-Wah Ngo, Nan Duan, and Mike Zheng Shou. Groundnlq@ ego4d natural language queries challenge 2023. *arXiv preprint arXiv:2306.15255*, 2023.

[352] Xiang Fang, Daizong Liu, Wanlong Fang, Pan Zhou, Zichuan Xu, Wenzheng Xu, Junyang Chen, and Renfu Li. Fewer steps, better performance: Efficient cross-modal clip trimming for video moment retrieval using language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 1735–1743, 2024.

[353] H. Zhang, A. Sun, W. Jing, L. Zhen, J. T. Zhou, and R. S. M. Goh. Natural language video localization: A revisit in span-based question answering framework. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[354] Patrik Schmuck and Margarita Chli. Ccm-slam: Robust and efficient centralized collaborative monocular simultaneous localization and mapping for robotic teams. *Journal of Field Robotics*, 36(4):763–781, 2019.

[355] Marco Karrer, Mohit Agarwal, Mina Kamel, Roland Siegwart, and Margarita Chli. Collaborative 6dof relative pose estimation for two uavs with overlapping fields of view. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 6688–6693, 2018.

[356] Daiwat Amit Vyas and Dvijesh Bhatt. Augmented reality (ar) applications: A survey on current trends, challenges, & future scope. *International Journal of Advanced Research in Computer Science*, 8(5), 2017.

[357] Tram Thi Minh Tran, Shane Brown, Oliver Weidlich, Mark Billinghurst, and Callum Parker. Wearable augmented reality: Research trends and future directions from three major venues. *IEEE Transactions on Visualization and Computer Graphics*, 29(11):4782–4793, 2023.

[358] Mai Lee Chang, Greg Trafton, J Malcolm McCurry, and Andrea Lockerd Thomaz. Unfair! perceptions of fairness in human-robot teams. In *30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*, pages 905–912, 2021.

[359] Houston Claure, Mai Lee Chang, Seyun Kim, Daniel Omeiza, Martim Brandao, Min Kyung Lee, and Malte Jung. Fairness and transparency in human-robot interaction. In *17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 1244–1246, 2022.

[360] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[361] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090, 2021.

[362] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7061–7070, 2023.

[363] Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. Scaling open-vocabulary object detection. *Advances in Neural Information Processing Systems*, 36:72983–73007, 2023.

[364] Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. Yolact: Real-time instance segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9157–9166, 2019.

[365] Yuqian Fu, Runze Wang, Yanwei Fu, Danda Pani Paudel, Xuanjing Huang, and Luc Van Gool. Objectrelator: Enabling cross-view object relation understanding in ego-centric and exo-centric videos. *arXiv preprint arXiv:2411.19083*, 2024.

[366] Matthew S Hutchinson and Vijay N Gadepally. Video action understanding. *IEEE Access*, 9:134611–134637, 2021.

[367] Kunchang Li, Xinhao Li, Yi Wang, Yinan He, Yali Wang, Limin Wang, and Yu Qiao. Videomamba: State space model for efficient video understanding. In *European Conference on Computer Vision*, pages 237–255. Springer, 2024.

[368] Yi-Hsuan Tsai, Ming-Hsuan Yang, and Michael J Black. Video segmentation via object flow. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3899–3908, 2016.

[369] Dahun Kim, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Video panoptic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9859–9868, 2020.

[370] Sanghyun Woo, Dahun Kim, Joon-Young Lee, and In So Kweon. Learning to associate every segment for video panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2705–2714, 2021.

[371] Mengmeng Wang, Teli Ma, Shuo Xin, Xiaojun Hou, Jiazheng Xing, Guang Dai, Jingdong Wang, and Yong Liu. Visual object tracking across diverse data modalities: A review. *arXiv preprint arXiv:2412.09991*, 2024.

[372] Gioele Ciaparrone, Francisco Luque Sánchez, Siham Tabik, Luigi Troiano, Roberto Tagliaferri, and Francisco Herrera. Deep learning in video multi-object tracking: A survey. *Neurocomputing*, 381:61–88, 2020.

[373] Yanghao Li, Tushar Nagarajan, Bo Xiong, and Kristen Grauman. Ego-exo: Transferring visual representations from third-person to first-person videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6943–6953, 2021.

[374] Zi-Yi Dou, Xitong Yang, Tushar Nagarajan, Huiyu Wang, Jing Huang, Nanyun Peng, Kris Kitani, and Fu-Jen Chu. Unlocking exocentric video-language data for egocentric video representation learning. *arXiv preprint arXiv:2408.03567*, 2024.

[375] H. Yu, M. Cai, Y. Liu, and F. Lu. Joint attention learning for first and third person video co-analysis. *ACM MM*, 2019.

[376] Zihui Sherry Xue and Kristen Grauman. Learning fine-grained view-invariant representations from unpaired ego-exo videos via temporal alignment. *Advances in Neural Information Processing Systems*, 36:53688–53710, 2023.

[377] H. Rahmani and A. Mian. Learning a non-linear knowledge transfer model for cross-view action recognition. In *CVPR*, 2016.

[378] Gerard Donahue and Ehsan Elhamifar. Learning to predict activity progress by self-supervised video alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18667–18677, 2024.

[379] Camillo Quattrocchi, Antonino Furnari, Daniele Di Mauro, Mario Valerio Giuffrida, and Giovanni Maria Farinella. Synchronization is all you need: Exocentric-to-egocentric transfer for temporal action segmentation with unlabeled synchronized video pairs. In *European Conference on Computer Vision*, pages 253–270. Springer, 2024.

[380] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60:91–110, 2004.

[381] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. ORB: An efficient alternative to SIFT or SURF. In *2011 International conference on computer vision*, pages 2564–2571. Ieee, 2011.

[382] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-up robust features (surf). *Computer vision and image understanding*, 110(3):346–359, 2008.

[383] Kwang Moo Yi, Eduard Trulls, Vincent Lepetit, and Pascal Fua. Lift: Learned invariant feature transform. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14*, pages 467–483. Springer, 2016.

[384] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4938–4947, 2020.

[385] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8922–8931, 2021.

[386] Wei Jiang, Eduard Trulls, Jan Hosang, Andrea Tagliasacchi, and Kwang Moo Yi. Cotr: Correspondence transformer for matching across images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6207–6217, 2021.

[387] Yanbin Liu, Linchao Zhu, Makoto Yamada, and Yi Yang. Semantic correspondence as an optimal transport problem. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4463–4472, 2020.

[388] Seokju Cho, Sunghwan Hong, Sangryul Jeon, Yunsung Lee, Kwanghoon Sohn, and Seungryong Kim. Cats: Cost aggregation transformers for visual correspondence. *Advances in Neural Information Processing Systems*, 34:9011–9023, 2021.

[389] Junyi Zhang, Charles Herrmann, Junhwa Hur, Luisa Polania Cabrera, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence. *Advances in Neural Information Processing Systems*, 36:45533–45547, 2023.

[390] Bowen Cheng, Maxwell D Collins, Yukun Zhu, Ting Liu, Thomas S Huang, Hartwig Adam, and Liang-Chieh Chen. Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation. In *Proceedings of the*

IEEE/CVF conference on computer vision and pattern recognition, pages 12475–12485, 2020.

[391] Rohit Mohan and Abhinav Valada. Efficientps: Efficient panoptic segmentation. *International Journal of Computer Vision*, 129(5):1551–1579, 2021.

[392] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022.

[393] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 724–732, 2016.

[394] Zhao Yang, Qiang Wang, Luca Bertinetto, Weiming Hu, Song Bai, and Philip HS Torr. Anchor diffusion for unsupervised video object segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 931–940, 2019.

[395] Pavel Tokmakov, Karteek Alahari, and Cordelia Schmid. Learning video object segmentation with visual memory. In *Proceedings of the IEEE international conference on computer vision*, pages 4481–4490, 2017.

[396] Xiangtai Li, Haobo Yuan, Wei Li, Henghui Ding, Size Wu, Wenwei Zhang, Yining Li, Kai Chen, and Chen Change Loy. Omg-seg: Is one model good enough for all segmentation? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 27948–27959, 2024.

[397] Lei Ke, Mingqiao Ye, Martin Danelljan, Yu-Wing Tai, Chi-Keung Tang, Fisher Yu, et al. Segment anything in high quality. *Advances in Neural Information Processing Systems*, 36:29914–29934, 2023.

[398] Yunyang Xiong, Bala Varadarajan, Lemeng Wu, Xiaoyu Xiang, Fanyi Xiao, Chenchen Zhu, Xiaoliang Dai, Dilin Wang, Fei Sun, Forrest Iandola, et al. Efficientsam: Leveraged masked image pretraining for efficient segment anything. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16111–16121, 2024.

[399] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 9579–9589, 2024.

[400] Zheng Zhang, Yeyao Ma, Enming Zhang, and Xiang Bai. Psalm: Pixelwise segmentation with large multi-modal model. In *European Conference on Computer Vision*, pages 74–91. Springer, 2024.

[401] Xueyan Zou, Zi-Yi Dou, Jianwei Yang, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Harkirat Behl, Jianfeng Wang, Lu Yuan, et al. Generalized decoding for pixel, image, and language. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15116–15127, 2023.

[402] Raghav Goyal, Wan-Cyuan Fan, Mennatullah Siam, and Leonid Sigal. Tamvt: Transformation-aware multi-scale video transformer for segmentation and tracking. *arXiv preprint arXiv:2312.08514*, 2023.

[403] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[404] Liulei Li, Wenguan Wang, Tianfei Zhou, Jianwu Li, and Yi Yang. Unified mask embedding and correspondence learning for self-supervised video segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18706–18716, 2023.

[405] Sergi Caelles, Jordi Pont-Tuset, Federico Perazzi, Alberto Montes, Kevis-Kokitsi Maninis, and Luc Van Gool. The 2019 davis challenge on vos: Unsupervised multi-object segmentation. *arXiv preprint arXiv:1905.00737*, 2019.

[406] Ning Xu, Linjie Yang, Yuchen Fan, Jianchao Yang, Dingcheng Yue, Yuchen Liang, Brian Price, Scott Cohen, and Thomas Huang. Youtube-vos: Sequence-to-sequence video object segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 585–601, 2018.

[407] Michael Rubinstein, Armand Joulin, Johannes Kopf, and Ce Liu. Unsupervised joint object discovery and segmentation in internet images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1939–1946, 2013.

[408] Tatsunori Taniai, Sudipta N Sinha, and Yoichi Sato. Joint recovery of dense correspondence and cosegmentation in two images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4246–4255, 2016.

[409] Xi Shen, Alexei A Efros, Armand Joulin, and Mathieu Aubry. Learning co-segmentation by segment swapping for retrieval and discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5082–5092, 2022.

[410] Kartik Garg, Sai Shubodh Puligilla, Shishir Kolathaya, Madhava Krishna, and Sourav Garg. Revisit anything: Visual place recognition via image segment retrieval. In *European Conference on Computer Vision*, pages 326–343. Springer, 2024.

[411] Moshe Bar. Visual objects in context. *Nature Reviews Neuroscience*, 5(8):617–629, 2004.

[412] Helene Intraub. The representation of visual scenes. *Trends in cognitive sciences*, 1(6):217–222, 1997.

[413] Michal Shlapentokh-Rothman, Ansel Blume, Yao Xiao, Yuqun Wu, Sethuraman TV, Heyi Tao, Jae Yong Lee, Wilfredo Torres, Yu-Xiong Wang, and Derek Hoiem. Region-based representations revisited. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17107–17116, 2024.

[414] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[415] Kay Henning Brodersen, Cheng Soon Ong, Klaas Enno Stephan, and Joachim M Buhmann. The balanced accuracy and its posterior distribution. In *20th International Conference on Pattern Recognition*, pages 3121–3124. IEEE, 2010.

[416] Ho Kei Cheng and Alexander G Schwing. XMem: Long-term video object segmentation with an atkinson-shiffrin memory model. In *European Conference on Computer Vision*, pages 640–658. Springer, 2022.

[417] Jiaming Zhang, Huayao Liu, Kailun Yang, Xinxin Hu, Ruiping Liu, and Rainer Stiefelhagen. Cmx: Cross-modal fusion for rgb-x semantic segmentation with transformers. *IEEE Transactions on intelligent transportation systems*, 24(12):14679–14694, 2023.

[418] Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. Lightglue: Local feature matching at light speed. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17627–17638, 2023.

[419] Michał Tyszkiewicz, Pascal Fua, and Eduard Trulls. Disk: Learning local features with policy gradient. *Advances in Neural Information Processing Systems*, 33:14254–14265, 2020.

[420] Xiaoming Zhao, Xingming Wu, Weihai Chen, Peter CY Chen, Qingsong Xu, and Zhengguo Li. Aliked: A lighter keypoint and descriptor extraction network via deformable transformation. *IEEE Transactions on Instrumentation and Measurement*, 72:1–16, 2023.

[421] Lorenzo Mur-Labadia, Josechu Guerrero, and Ruben Martinez-Cantin. Div-ff: Dynamic image-video feature fields for environment understanding in egocentric videos. *arXiv preprint arXiv:2503.08344*, 2025.