Revista Iberoamericana de Automática e Informática Industrial

Revista Iberoamericana de Automática e Informática Industrial 21 (2024) 374-382

www.revista-riai.org

Comprensión automática de escenas en imágenes de entornos submarinos

Cesar Borja*, Ana C. Murillo

Instituto Universitario en Ingeniería de Aragón (I3A), Universidad de Zaragoza, c/ María de Luna 1, 50018 Zaragoza, España.

To cite this article: Borja, C., Murillo, A.C. 2024. Visual scene understanding in underwater environments. Revista Iberoamericana de Automática e Informática Industrial 21, 374-382. https://doi.org/10.4995/riai.2024.21290

Resumen

La utilización de vehículos submarinos autónomos (AUV) representa un avance significativo en el campo de la monitorización del fondo marino. Sin embargo, el procesamiento de imágenes de datos adquiridos desde AUVs presenta un desafío único debido a las propiedades inherentes del entorno submarino, como la atenuación de la luz y la turbidez del agua. Este trabajo investiga técnicas para mejorar la comprensión automática del contenido de escenas submarinas a partir de imágenes monoculares. El sistema propuesto aprovecha modelos de aprendizaje profundo existentes junto con algoritmos simples de procesamiento de imágenes, eliminando la necesidad de entrenamiento supervisado adicional. El sistema estudia la combinación de un modelo de aprendizaje profundo pre-entrenado para la estimación de profundidad a partir de imágenes monoculares, con el algoritmo propuesto para distinguir regiones de agua del resto de elementos de la escena. El estudio presentado incluye una comparación detallada de la influencia en el resultado de varias alternativas y opciones de configuración del sistema. La validación experimental muestra cómo el sistema presentado obtiene resultados de segmentación más ricos en comparación con los algoritmos existentes utilizados como referencia. En particular, el sistema propuesto facilita la segmentación precisa de regiones de agua y facilita la detección de otros objetos de interés, incluyendo elementos suspendidos en el agua, que potencialmente pueden corresponder a peces u otros obstáculos móviles.

Palabras clave: Segmentación Semántica, Comprensión de escenas submarinas, Percepción y sensores.

Visual scene understanding in underwater environments

Abstract

The utilization of Autonomous Underwater Vehicles (AUVs) represents a significant advancement in the field of seabed monitoring. However, image processing of data acquired from AUVs presents a unique challenge due to the inherent properties of the underwater environment, such as light attenuation and water turbidity. This work investigates techniques to enhance underwater scene understanding from monocular images. The proposed system leverages existing deep learning methods in conjunction with simple image processing algorithms, eliminating the need for additional supervised training. The system studies the combinatio of a pre-trained deep learning model, for depth estimation from monocular images, with the proposed algorithm to distinguish water regions from the rest of the scene elements. The presented study includes comprehensive comparison of various system alternatives and configuration options. The experimental validation shows how the presented system obtains richer segmentation results compared to baseline algorithms. Notably, the proposed system facilitates the accurate segmentation of water regions and enables the detection of other objects of interest, including suspended elements in the water, which can potentially correspond to fish or other mobile obstacles.

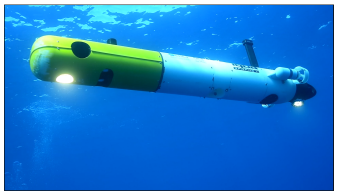
Keywords: Semantic Segmentation, Underwater scene understanding, Perception and Sensing.

*Autor para correspondencia: 800675@unizar.es

1. Introducción

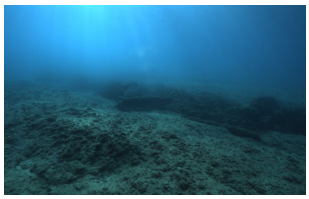
La comprensión de los ecosistemas marinos, los procesos geológicos y la biodiversidad marina son tareas esenciales en el estudio y conservación del medio ambiente, y la monitorización de mares y océanos es una práctica fundamental para facilitar estas tareas. El uso de robots autónomos en tareas de monitorización del fondo submarino está revolucionado la manera en la que se llevan a cabo estos estudios, permitiendo la obtención de datos en tiempo real de manera automatizada y en áreas de difícil acceso. La Figura 1a muestra un ejemplo de vehículo autónomo submarino (AUV, del inglés *Autonomous Underwater Vehicle*) y de imágenes de escenas submarinas.

Para conseguir que estos AUVs tengan mayor autonomía y capacidad para realizar diferentes tareas, resulta esencial equiparlos con sensores como por ejemplo cámaras o sensores de ultrasonidos, que es permitan percibir mejor que hay a su alrededor. En otras palabras, que les permitan *entender* su entorno. Esta mejora en las capacidades de los AUVs puede facilitar tareas como inspeccionar estructuras submarinas, explorar yacimientos arqueológicos o mapear el fondo marino. Resulta de gran interés automatizar parcial o completamente estas tareas, y los desarrollos recientes de visión por computador muestran grandes avances en, por ejemplo, sistemas de reconstrucción del color en imágenes submarinas Akkaynak and Treibitz (2019) o sistemas de *tracking* para AUVs Kumar et al. (2018).



(a)

(b)



(c)

Figura 1: Ejemplo de AUV ALICE Gutnik et al. (2022) (a) y de imágenes de escenas submarinas cedidas por el Laboratorio VISEAON [1] captadas desde un AUV (b), (c).

Sin embargo, características propias del medio subacuático, como la turbidez del agua o la atenuación de la luz, complican el estudio y procesado de estas imágenes. Los métodos existentes, el estado-del-arte de visión por computador, no funcionen con la misma fiabilidad en estos entornos. En la Figura 1b y 1c se muestran algunos ejemplos de imágenes captadas desde AUVs. En ellas se pueden observar estructuras y ruinas submarinas, y se aprecia cómo las condiciones subacuáticas dificultan la visibilidad en dichas escenas.

La contribución principal de este trabajo es un sistema para análisis automático del contenido de imágenes en entornos submarinos. El sistema presentado combina modelos de deep learning existentes con algoritmos sencillos de post-procesado para obtener un sistema mejorado de análisis de este tipo de imagen [2].

Una versión preliminar fue expuesta en las *XLIV Jornadas de Automática'23*, por Borja and Murillo (2023). El presente trabajo extiende dicha comunicación con una descripción más detallada del método e incluyendo modificaciones en el sistema propuesto que mejoran sustancialmente la calidad de los resultados con respecto a lo publicado en dicho trabajo. Estas mejoras afectan principalmente al sistema de post-procesado y al sistema de estimación de profundidad utilizado. Además, el presente artículo incluye experimentos mucho mas exhaustivos respecto al trabajo preliminar, ya que ahora se presentan resultados con el *benchmark* SUIM-dataset completo, en lugar de solo unas pocas imágenes, y un análisis mas exhaustivo de los resultados.

El sistema presentado ayuda a sobrevenir algunas de las dificultades específicas que se encuentran en el entorno submarino. Por un lado, se han estudiado modelos capaces de estimar la profundidad, respecto a la cámara, a la que están los distintos elementos de la imagen, tomada por una cámara monocular. En particular, se ha considerado el estado del arte para esta tarea, con métodos existentes basados en redes neuronales profundas, concretamente, *monodepth2*, de Godard et al. (2019), *monoUWNet*, de Amitai et al. (2023), y el reciente modelo *Depth Anything*, de Yang et al. (2024). Estos modelos reciben como entrada una imagen RGB y devuelven un valor de profundidad para cada píxel de la imagen. Con esta información se puede generar un modelo 3D sencillo (nube de puntos) de la escena. Estos modelos para estimación de profundidad en una imagen monocular se describen en más detalle en la siguiente sección.

Por otro lado, se han estudiado distintos métodos de segmentación de imagen para intentar separar las partes más o menos relevantes de la escena capturada en la imagen. El objetivo es conseguir un sistema que no necesite entrenar nuevos algoritmos supervisados específicos. Se han estudiado métodos más tradicionales de visión por computador, superpíxeles, como el trabajo presentado por Van den Bergh et al. (2012), y otros más recientes de segmentación basada en redes neuronales profundas, como el sistema de Kirillov et al. (2023).

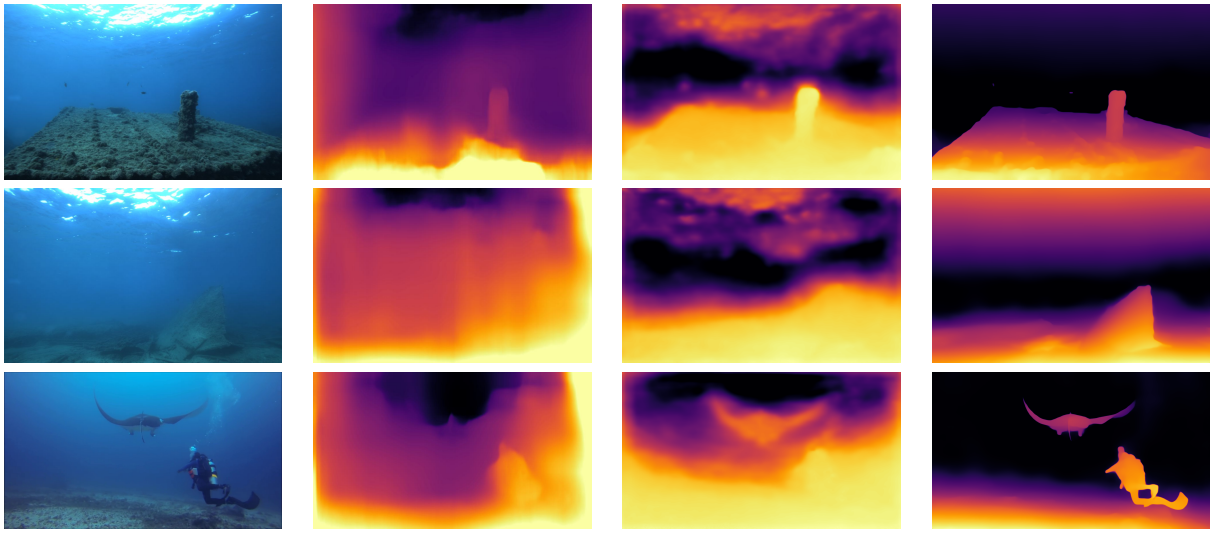
Para evaluar el sistema propuesto, se ha propuesto un conjunto heterogéneo de imágenes reales submarinas de distintas fuentes, para observar qué información útil sobre la escena se puede conseguir sin necesidad de entrenar nuevos modelos.

2. Trabajo relacionado

2.1. Modelos de estimación de profundidad

Estimar la profundidad de una escena a partir de una sola imagen es un problema de gran interés, por la gran cantidad de aplicaciones posibles. Encontramos numerosas propuestas en la literatura reciente, incluyendo HiMODE, de Junayed et al. (2022), la propuesta de Godard et al. (2017), o DINOv2,

[2] https://github.com/cborjamoreno/underwater_analysis.git



(a) Original (b) monodepth2 (c) monoUWNet (d) Depth Anything

Figura 2: Comparación cualitativa de los mapas de profundidad estimados con monodepth2, monoUWNet y Depth Anything para varias imágenes submarinas. Se observa como Depth Anything realiza una mejor estimación.

de Oquab et al. (2023). Este tipo de modelos de estimación de profundidad reciben una imagen RGB monocular como entrada y devuelven una matriz con las mismas dimensiones en la que cada elemento (x, y) de la matriz contiene el valor de profundidad estimado del píxel (x, y) de la imagen. Esta estimación es un valor a escala, es decir, no se mide en unidades de distancia.

El modelo Monodepth2, de Godard et al. (2019), es uno de los modelos pioneros para el uso del deep learning en esta tarea, y por lo tanto esencial en nuestras comparativas. Consigue resultados muy exitosos para estimar profundidad, a partir solo de imagen monocular (ver Figura 2b). Siguiendo algunas ideas de Monodepth2, pero adaptado especialmente para entornos submarinos, encontramos el modelo monoUWNet, de Amitai et al. (2023). La adaptación especifica al entorno submarino lo hacen un método esencial para incluir en nuestro estudio (ver Figura 2c). Por último, también resulta de especial interés, el modelo Depth Anything, de Yang et al. (2024), un modelo entrenado en una gran variedad de escenarios, tanto con datos etiquetados como no etiquetados, y que está presentando resultados muy prometedores en una multitud de escenarios diversos (ver Figura 2d). Nuestro trabajo analiza si también resulta un trabajo satisfactorio para algunas de nuestras tareas.

La estimación de distancias a partir de imágenes monoculares se realiza utilizando una red neuronal profunda entrenada a partir de un conjunto de pares de imágenes. Las dos imágenes de cada par se corresponden con la imagen sobre la que se quiere sacar la estimación y su *ground-truth* (GT) creado a partir de mediciones reales con sensores de profundidad. Sin embargo, dada la difícil adquisición de distancias GT (especialmente en entornos submarinos), tanto monodepth2 como monoUWNet proponen utilizar *frames* consecutivos para llevar a cabo un entrenamiento auto-supervisado. Utilizando estos *frames* se obtienen diferentes poses de la misma escena, permitiendo sacar una medida de profundidad teniendo en cuenta la estimación del movimiento entre *frames*. No obstante, pese a sus similitudes, y tal y como se muestra en el trabajo de monoUWNet, este incorpora una serie de mejoras y adaptaciones importantes enfocadas a mejorar los resultados en escenas submarinas que se demuestran en el trabajo original. En cuanto a Depth Anything, se trata de un modelo centrado en la generalización *zero-shot*, es decir, es capaz de producir estimaciones de profundidad en tipos de escenarios para los que no ha sido entrenado. Esta generalización ayuda a minimizar el problema de la adquisición de datos etiquetados como *ground-truth* en entornos submarinos.

En la Figura 2 se puede ver de manera cualitativa como Depth Anything ofrece estimaciones de profundidad más cercanas a la realidad que monodepth2 y monoUWNet. Realizando esta comparación, y basándonos en los resultados del trabajo original y de nuestros resultados preliminares con los otros modelos, se ha decidido utilizar Depth Anything como modelo base en el sistema final presentado en este trabajo.

2.2. *Segmentación de imagen*

Existen muchísimos trabajos en la literatura sobre segmentación de imagen. En nuestro trabajo resultan de especial interés y relevancia dos grandes grupos.

Segmentación en superpixels. Por un lado, existen numerosos métodos de segmentación no supervisada que agrupan zonas cercanas en la imagen, y similares en apariencia, para formar segmentos o superpíxeles. Alguno de los métodos más conocidos son SLIC, de Achanta et al. (2012), o SEEDS, de Van den Bergh et al. (2012). En nuestro sistema se va a trabajar con SEEDS (*Superpixels Extracted via Energy-Driven Sampling*), por presentar un buen compromiso entre precisión, facilidad de uso y rapidez. Este método comienza creando una malla cuadricular sobre toda la imagen, siendo cada uno de los cuadrados un superpíxel. Tras esto el algoritmo hace una optimización donde se favorece la homogeneidad de la distribución del color dentro de cada uno de los superpíxeles. Como resultado, cada superpíxel intercambia píxeles con sus vecinos cambiando la forma del borde de cada superpíxel hasta alcanzar la segmentación final.

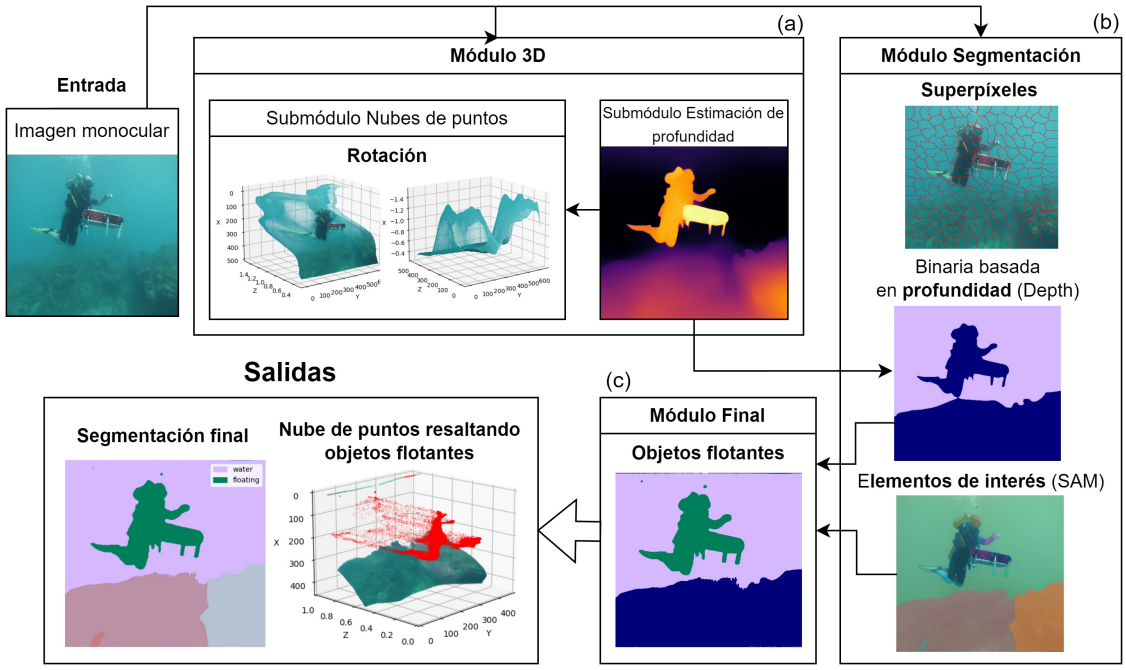


Figura 3: Diagrama del sistema desarrollado.

Segmentación con redes neuronales profundas. Por otro lado, resultan de gran interés los modelos más recientes de segmentación semántica basados en deep learning, ya que representan los resultados del estado del arte y presentan gran capacidad de generalización, como se viene demostrando en los últimos años con ejemplos como los trabajos de Feng et al. (2020), Yang and Yu (2021) o Girshick et al. (2014).

En particular, en este trabajo se hace uso del modelo *Segment Anything Model* (SAM), de Kirillov et al. (2023), ya que es capaz de obtener una segmentación genérica de los posibles objetos de la imagen, sin necesidad de conocer clases concretas de objetos. Este modelo funciona de manera muy eficaz sin ningún tipo de entrenamiento adicional, con lo cual resulta muy interesante para nuestro sistema. Sin embargo, este tipo de modelos conllevan un fuerte gasto computacional, que debe ser tenido en cuenta en los estudios para su uso en ciertas aplicaciones.

3. Sistema propuesto para análisis de la escena submarina

El sistema desarrollado en este trabajo consta de tres módulos principales, descritos a continuación. La entrada del sistema es una imagen monocular, mientras que la salida proporcionada tiene dos componentes: una imagen segmentada en la que se resaltan elementos de interés de la escena y un modelo 3D (en formato nube de puntos), que se post-procesa para incorporar anotaciones e información de interés. La Figura 3 muestra un diagrama donde se muestra el *pipeline* completo del sistema, a través de los tres módulos, con ejemplos de los pasos intermedios realizados en cada uno de ellos.

Módulo 3D. Este módulo obtiene la estimación de profundidad a partir de imágenes monoculares, utilizando el modelo pre-entrenado, descrito en la sección anterior, *Depth Anything*, y genera y maneja nubes de puntos 3D a partir de dicha información (ver Figura 3a).

Módulo de segmentación. Este módulo aplica diferentes métodos de segmentación, sin ningún tipo de entrenamiento adicional, para separar ciertas zonas de interés de la escena (ver Figura 3b). Principalmente, nos centramos en las zonas que corresponden con el agua, ya que generan ruido al resto de algoritmos de procesado, por ejemplo en las reconstrucciones 3D obtenidas en el modulo anterior, ya que son píxeles que no queremos recuperar ni incluir en el modelo 3D (al igual que en *point-clouds* en escenas convencionales no queremos recuperar los píxeles correspondientes al aire), y realmente corresponden con zonas sin interés para la detección de objetos y obstáculos.

Se propone cómo identificar la zona de la imagen que corresponde con el agua mediante los siguientes pasos:

1. **Binarización en imagen de profundidad**. Se estiman primero los valores de profundidad de la imagen y después se binariza el resultado con un *threshold*. Así, se clasifica cada píxel como "agua" si su valor de profundidad supera el valor del *threshold* o como "escena" en caso contrario. Este umbral no es un valor fijo, sino que es relativo a la profundidad máxima en cada escena. El criterio se estableció como el 90 % de la profundidad máxima ya que se observó que los valores de profundidad en las zonas de agua (las más profundas) eran muy significativos respecto al resto de la escena.
   Pese a que este método obtiene resultados prometedores por si mismo, como se observará en los resultados, esta segmentación binaria presenta algunos problemas de precisión, especialmente en imágenes con destellos sobre la superficie del agua, segmentando estas zonas como escena (ver Figura 6).

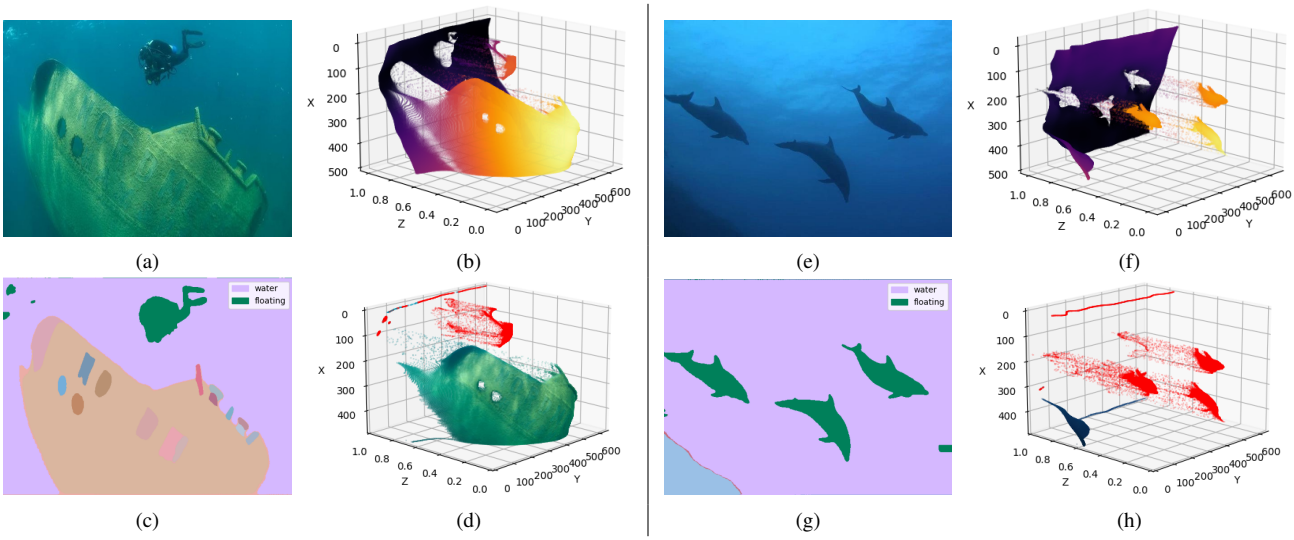


Figura 4: Ejemplos del modelo anotado final obtenido por el sistema propuesto dada una imagen original (a)(e). Puntos coloreados según el valor de profundidad estimados con monoUWNet (b)(f). Segmentación final 2D (c)(g). Se distinguen puntos pertenecientes a la masa de agua y objetos flotantes. Puntos con el color original en la imagen, salvo los objetos flotantes identificados, que se remarcan en rojo (d)(h).

2. **Segmentación combinada con modelo SAM.** Para conseguir una segmentación agua/escena más fiel a la realidad, se ha utilizado la segmentación del paso anterior en combinación con SAM. El objetivo en este paso es identificar qué segmento proporcionado por SAM corresponde con la masa de agua de la imagen. Para ello se calcula la *Intersection over Union* (IoU) de cada segmento de SAM con la zona de agua segmentada en el paso anterior (con la máscara binaria basada en profundidad) y se elige aquel que tenga mayor IoU. Si el segmento elegido supera un valor de IoU mínimo, se genera una nueva máscara binaria utilizando el segmento como zona de agua. En caso contrario se mantiene la segmentación basada únicamente en los valores de profundidad. Este valor de IoU mínimo se ha ajustado manualmente para requerir una intersección grande entre un segmento de SAM con la región de agua según la segmentación por profundidad, ya que el objetivo es que SAM refine los contornos de la zona identificada en el paso anterior. Para elegir un valor adecuado, se hicieron algunas pruebas considerando valores entre 0.6 y 0.9, y realizando una comparación cualitativa de los resultados en tres imágenes del VISEAON-dataset (sin *ground-truth* de segmentación semántica), y dos del SUIM-dataset (Sección 4.1). En la Figura 6 se muestran varios ejemplos de segmentación binaria utilizando las tres alternativas descritas en más detalle en la Sección 4.
3. **Segmentación de elementos suspendidos en el agua.** Además, se propone un algoritmo sencillo para identificar elementos que se encuentran flotando, suspendidos, en el agua, que resultan muy relevantes para el análisis de la escena. Este algoritmo se aplica sobre la segmentación agua/escena obtenida con los pasos anteriores, y consiste en la localización de componentes conexas en dicha imagen. La Figura 7 muestra varios ejemplos de los resultados de la segmentación de objetos flotantes.

Módulo Final. Este módulo combina los módulos anteriores para obtener la segmentación final (ver Figura 3c). Esta segmentación consiste en una imagen 2D y una nube de puntos 3D. La imagen 2D muestra la segmentación de la masa de agua y objetos suspendidos en el agua o "flotantes". El resto de elementos de la escena mantienen la segmentación realizada por SAM.

En cuanto a la nube de puntos, se post-procesan los puntos no relevantes, como por ejemplo los identificados como agua, que se eliminan del resultado final dado como modelo. Además, se anotan posibles elementos de interés, mediante los objetos flotantes, que representan posibles obstáculos o elementos interesantes para tareas de monitorización de, por ejemplo, animales u otros obstáculos móviles. La Figura 4 muestra dos ejemplos de la segmentación final devuelta por el sistema.

4. Experimentos

Esta sección describe los experimentos realizados para analizar y evaluar el sistema desarrollado.

4.1. *Configuración de los experimentos*

Datasets. Para conseguir un conjunto de evaluación heterogéneo, se han utilizado imágenes de varias fuentes.

VISEAON-dataset. Como datos muy relevantes, ya que son directamente capturados por un robot submarino, se han utilizado varias imágenes (12) capturadas desde un AUV por el laboratorio de investigación VISEAON [3].

SUIM-dataset. Para poder realizar un análisis cuantitativo más exhaustivo, y tener imágenes con *ground truth* para las localización de las zonas de agua de la imagen, se han incluido todas las imágenes de test (110 imágenes) del SUIM-dataset: Semantic Segmentation of Underwater Imagery, de Islam et al. (2020). Estos datos incluyen etiquetas para 8 categorías, entre

[3] https://www.viseaon.haifa.ac.il/

	SPX			Depth			Depth+SAM		
	Prec.	**Recall**	**t(s)**	**Prec.**	**Recall**	**t(s)**	**Prec.**	**Recall**	**t(s)**
Media	0.58	0.47	1.21	0.7	0.81	3.53	0.73	0.8	18.19
Mediana	0.78	0.46	1.06	0.81	0.99	3.47	0.96	0.96	17.75

Tabla 1: Comparación de precisión, *recall* (exhaustividad) y tiempo de ejecución en segundos de los tres métodos para segmentación del agua. Se muestra la media y mediana para las 110 imágenes utilizadas del SUIM-dataset.

ellas, la etiqueta de "agua", que es la que se utilizará para evaluar algunos de nuestros experimentos. Se ha elegido este dataset por su variedad de escenas subacuáticas, tanto en los propios objetos (buzos, gran variedad de fauna, embarcaciones hundidas, etc) como en la perspectiva de la cámara y el tono de color. Además, muchas de sus imágenes contienen elementos en suspensión con las que evaluar el algoritmo de segmentación de objetos flotantes.

Entorno de experimentación. Todos los experimentos se han ejecutado en un equipo con un microprocesador AMD Ryzen 5 5600X, 32GB de memoria RAM y una tarjeta gráfica NVIDIA GeForce RTX 3060.

4.2. Evaluación cuantitativa de segmentación binaria (agua)

Este experimento evalúa los resultados de los diferentes métodos implementados para la tarea de segmentación binaria de imágenes submarinas. En particular, comparamos los resultados obtenidos de usar las siguientes alternativas:

- **SPX**: Segmentación binaria utilizando superpíxeles obtenida de la siguiente manera. Se generan los superpíxeles en la imagen usando SEEDS, y se establece un rango de color en HSV que englobe los colores del agua. Para cada superpíxel se calcula su color medio, y se identifica como agua o no, dependiendo de si su color medio cae en el rango HSV establecido.
- **Depth**: Segmentación mediante *threshold* binario sobre los datos la estimación de profundidad, utilizando el modelo *Depth-Anything* (más detalles en el *módulo de segmentación* de la Sección 3).
- **Depth+SAM**: Combina la estimación de profundidad de la imagen con el resultado de la segmentación de objetos obtenida del modelo SAM. Este método corresponde con el método del *módulo final* descrito en la sección 3, coloreando todos los segmentos que no pertenecen al agua de la misma forma.

Para hacer una evaluación cuantitativa, se utilizarán las 110 imágenes del conjunto de test del SUIM-dataset, ya que el etiquetado semántico que ofrece permite calcular medidas de precisión (*Prec.*) y exhaustividad (*Recall*) de la segmentación de la masa de agua. La Tabla 1 muestra un resumen de los resultados obtenidos en este experimento. En general, podemos ver cómo los valores medios de precisión y *recall* son más bajos que los valores de sus medianas. Esto es debido a que para unas pocas imágenes, los valores de precisión y *recall* son muy bajos. En la Figura 5 vemos varios ejemplos de estos casos donde el sistema tiene mas dificultades.

El sistema siempre intenta segmentar una zona de agua, por tanto, aquellas imágenes en las que no se muestre la separación entre el fondo marino y la masa de agua (como en imágenes con la cámara orientada hacia el suelo) solo generarán falsos positivos.

Centrándonos en el análisis de los resultados de cada uno de las tres alternativas estudiadas, *Depth+SAM* obtiene los resultados más precisos. La mitad de imágenes tienen una precisión y *recall* mayor o igual que 0.96.

El método *SPX* tiene una media de exhaustividad más de dos décimas más baja que los otros dos métodos. Esto se debe a lo sensible que es el método a ligeros cambios en el tono de los colores de las imágenes.

Por último, *Depth* ofrece los resultados quizás más sorprendentes, obteniendo valores medios muy cercanos a *Depth+SAM*, y por lo tanto presentando el mejor compromiso entre calidad y coste de los resultados. En cuanto a las medianas, vemos como la precisión es notablemente más baja. Esta diferencia puede deberse a zonas de "escena" muy alejadas que el método clasifica incorrectamente como agua en algunos ejemplos. En la Figura 5 se muestra un ejemplo en la que el suelo del fondo marino, de arena, tras la formación rocosa, no es detectado correctamente por el modelo.

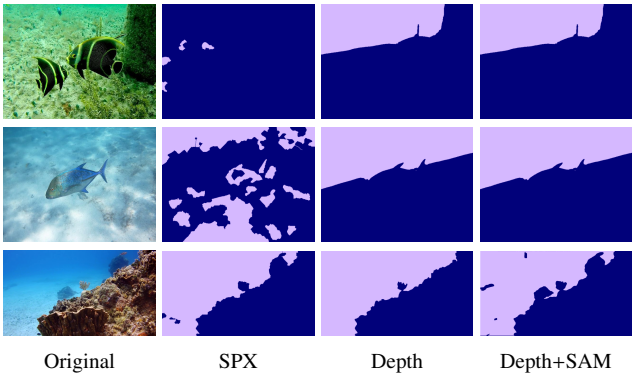


Figura 5: Ejemplos de imágenes para los cuales el sistema presenta dificultades en la segmentación binaria utilizando SPX, Depth y Depth+SAM. Las imágenes pertenecen al SUIM-dataset. Se puede apreciar como en las dos primeras la imagen original no contiene una zona clara de solo "agua" (porque la cámara apunta al suelo del fondo marino) y sin embargo Depth y Depth+SAM generan una zona clasificada como tal. El ejemplo de la tercera fila muestra una imagen de ejemplo de situaciones en que el modelo tiene problemas con los objetos que están a distancias muy largas.

Si tenemos en cuenta el tiempo de ejecución, SPX es el más rápido con una media de 1.21 segundos, seguido de Depth con 3.53 y la opción más costosa, Depth+SAM, con 18.19. Aunque ninguna de las implementaciones está especialmente optimizada, considerando los 3.53 segundos de media de tiempo de ejecución de la opción Depth (y teniendo en cuanta su buen rendimiento

en la segmentación), se puede considerar el más adecuado para aplicaciones con algún tipo de restricción computacional.

La Figura 6 muestra ejemplos representativos de los resultados obtenidos, para imágenes de ambos conjuntos de datos, con los tres métodos de segmentación binaria. Cualitativamente, se puede ver que Depth+SAM ofrece la segmentación más fiel a la realidad. Entre SPX y Depth, se puede observar como SPX, no obtiene resultados consistentes, segmentar bastante bien algunos casos, pero tiene problemas con la turbidez del agua y con algunos tipos de iluminación en los que el agua adopta colores más verdosos. En cambio Depth segmenta con más precisión la separación del agua y la escena, aunque tiende a segmentar como escena zonas de agua cercanas a la superficie.

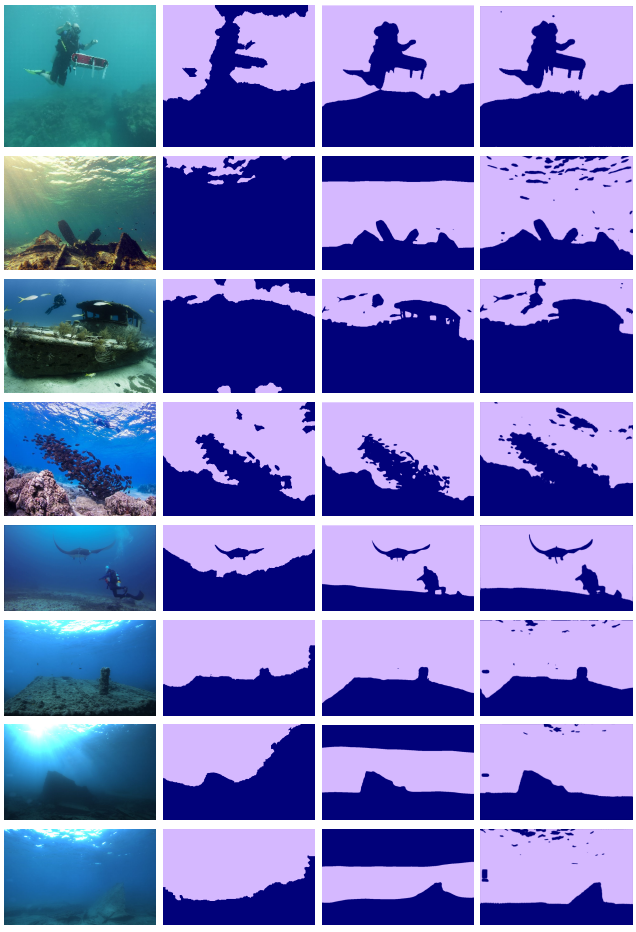


Figura 6: Comparación de los resultados de segmentación binaria utilizando SPX, Depth y Depth+SAM. Se puede apreciar como Depth+SAM consigue segmentar la escena y el agua mejor que SPX y Depth. Las 5 primeras imágenes pertenecen al SUIM-dataset mientras que las 3 últimas pertenecen al VISEAON-dataset.

4.3. *Evaluaciones cualitativas del sistema final*

En este apartado se muestran resultados cualitativos del método de segmentación de objetos flotantes y del sistema completo.

Segmentación de objetos flotantes. La Figura 7 muestra ejemplos de los resultados obtenidos con el método de segmentación de objetos flotantes para imágenes del SUIM-dataset. Esta alternativa sencilla para segmentar objetos en suspensión consigue resultados prometedores. Generalmente detecta correctamente los objetos flotando en el agua, aunque existen situaciones concretas en las que se suelen producir falsos positivos y falsos negativos. Los falsos positivos suelen darse por brillos en el agua o cambios de tonalidad en el color del agua. Los falsos negativos suelen deberse a que el objeto flotante se encuentra contiguo en la imagen al fondo marino (y por tanto no existe contorno para poder diferenciarlo). La perspectiva de la cámara es una limitación por tanto, ya que si la imagen es sacada apuntando hacia abajo, los objetos no se verán rodeados de agua.

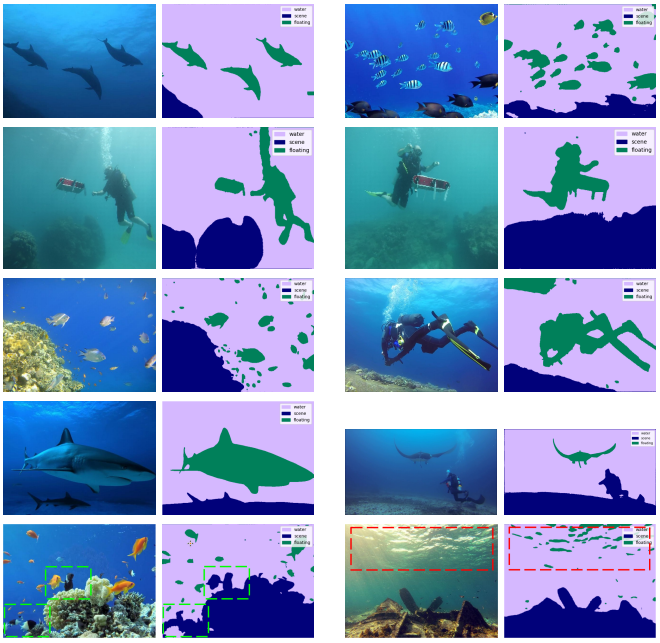


Figura 7: Ejemplos de segmentación de objetos flotantes en imágenes del SUIM-dataset. Para cada par de imágenes, se muestra la imagen original y la segmentación. En los dos ejemplos de la última fila, se ven marcados casos particulares de esta segmentación. Con un cuadrado verde, se resaltan falsos negativos en la segmentación. Esto se debe a que el contorno de los peces se junta con el contorno del coral, evitando por tanto que se forme un contorno cerrado. Por otro lado, un cuadrado rojo resalta como los brillos del agua producen falsos positivos.

Evaluación del sistema completo. Estos resultados ilustran las mejoras conseguidas respecto a la comprensión del contenido de la escena con los algoritmos básicos frente al sistema completo. En la Figura 8 se puede observar un mosaico con algunos resultados adicionales para diferentes imágenes. Primero se hace una estimación de profundidad, añadiendo una tercera dimensión a los datos. Sin embargo, estas nubes de puntos 3D contienen puntos de la masa de agua del mar que no son de interés y dificultan el análisis. Se hace una segmentación de la imagen separando el agua del resto de elementos de la escena, y se combina el resultado con la nube de puntos, eliminando los puntos de agua. Además, se puede aplicar la segmentación de objetos flotantes para resaltar objetos de interés en la nube de puntos, como posibles obstáculos o elementos móviles a monitorizar.

Esta evaluación muestra como los resultados del sistema favorecen considerablemente la comprensión automatizada de las escenas submarinas, detectando información relevante para monitorización o para sistemas autónomos de navegación o para realizar tareas de seguimiento en estos entornos.

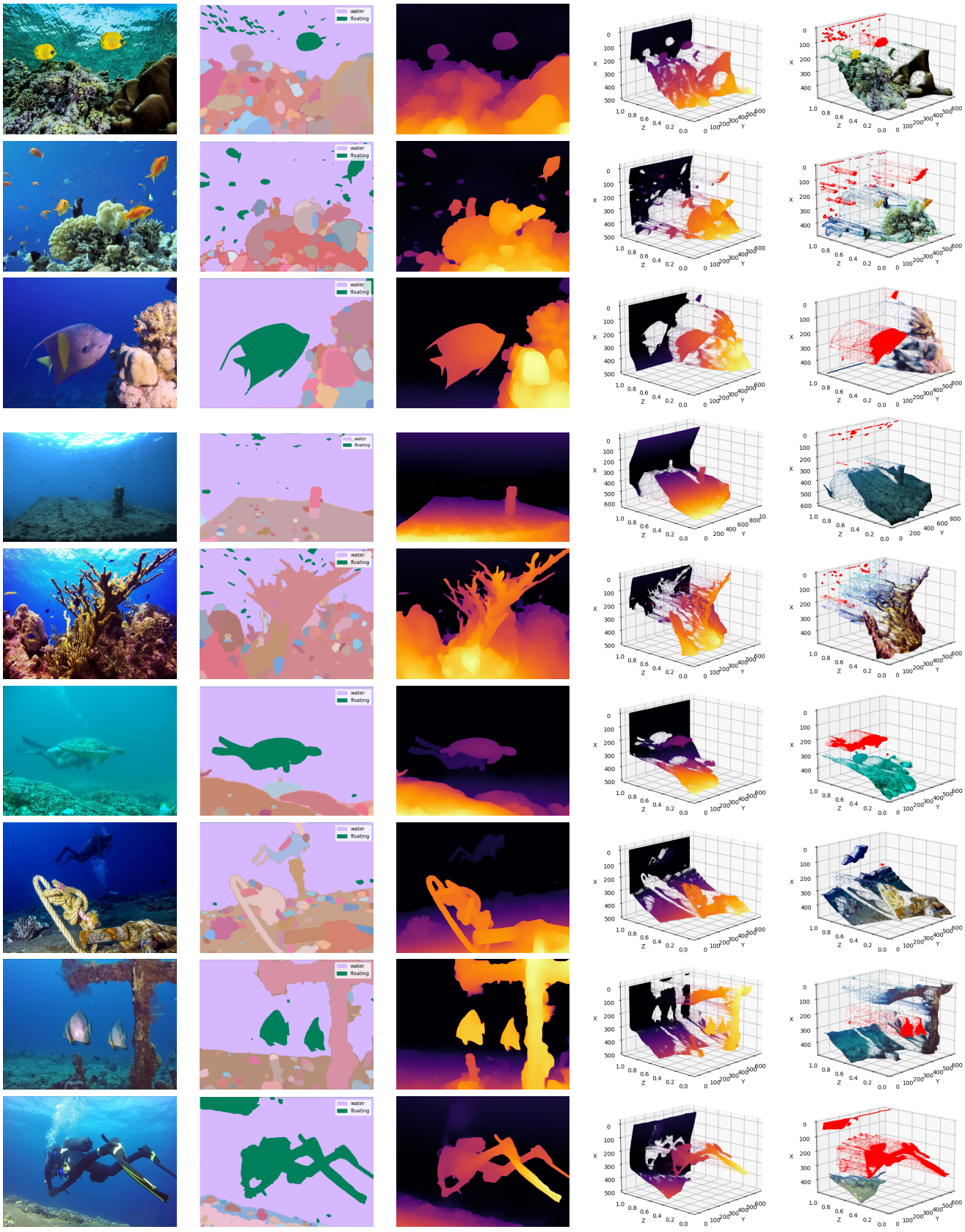


Figura 8: Resultados adicionales del sistema de análisis de escenas presentado. Se muestra la salida de varios pasos intermedios, similares a los mostrados en la Figura 4, obtenidos con imágenes de ambos conjuntos de datos utilizados. Cada columna (de izquierda a derecha) corresponde con: foto original, segmentación final (en verde objetos flotantes y morado el agua), estimación de profundidad, nube de puntos en 3D y nube de puntos limpia de agua destacando objetos flotantes (puntos en rojo) y con colores de la imagen original.

5. Conclusiones

Este trabajo presenta un sistema desarrollado para mejorar la comprensión automática de escenas submarinas, combinando información de profundidad con información semántica. Esta información se ha obtenido sin necesidad de re-entrenar ningún modelo adicional, con algoritmos sencillos de post procesado combinados con modelos del estado del arte basados en *Deep Learning*, mostrando el gran potencial de los modelos genéricos publicados recientemente, en particular los de estimación de profundidad (*Depth-Anything*) y segmentación de objetos (SAM).

El sistema propuesto es específico para imágenes subacuáticas, adaptando modelos generales, sin ningún re-entrenamiento, a problemáticas específicas de este entorno. Sin embargo, se podrían adaptar las ideas evaluadas para otros entornos.

Como pasos futuros, se plantean mejoras para segmentar objetos concretos que pudieran ser de interés, como corales o especies concretas de peces, combinando el trabajo realizado en sistemas de reconocimiento más específicos. Por otro lado, se podría integrar en un sistema robótico en tareas de navegación, aprovechando la información de profundidad de la nube de puntos y detectando posibles obstáculos como los elementos flotantes. Tal y como se ha comentado en la sección anterior, para esto último sería interesante utilizar el método de segmentación basado solo en el análisis de la imagen de profundidad, teniendo en cuenta su mejor rendimiento en tiempo de ejecución.

Agradecimientos

Este trabajo ha sido financiado parcialmente por FEDER/-Ministerio de Ciencia, Innovación y Universidades – Agencia Estatal de Investigación proyecto PID2021-125514NB-I00, y DGA T45_23R/FSE. Los autores agradecen el apoyo del Laboratorio VISEAON, de la Universidad de Haifa, Israel, a lo largo del trabajo. Además, el proyecto se ha desarrollado en el marco de la beca para *Prácticas de Estudiantes de Grado Universitario en el marco del TFG* del Instituto de Investigación en Ingeniería de Aragón (I3A) de la Universidad de Zaragoza.

Referencias

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Süsstrunk, S., 2012. Slic superpixels compared to state-of-the-art superpixel methods. IEEE transactions on pattern analysis and machine intelligence 34, 2274–2282.

Akkaynak, D., Treibitz, T., 2019. Sea-thru: A method for removing water from underwater images, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 1682–1691.

Amitai, S., Klein, I., Treibitz, T., 2023. Self-supervised monocular depth underwater, in: IEEE International Conference on Robotics and Automation (ICRA), IEEE. pp. 1098–1104.

Van den Bergh, M., Boix, X., Roig, G., de Capitani, B., Van Gool, L., 2012. Seeds: Superpixels extracted via energy-driven sampling. European Conference on Computer Vision, ECCV (7) 7578, 13–26.

Borja, C., Murillo, A.C., 2023. Análisis visual de escenas en entornos submarinos, in: XLIV Jornadas de Automática, Universidade da Coruña. Servizo de Publicacións. pp. 837–842.

Feng, D., Haase-Schütz, C., Rosenbaum, L., Hertlein, H., Glaeser, C., Timm, F., Wiesbeck, W., Dietmayer, K., 2020. Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. IEEE Transactions on Intelligent Transportation Systems 22, 1341–1360.

Girshick, R., Donahue, J., Darrell, T., Malik, J., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 580–587.

Godard, C., Mac Aodha, O., Brostow, G.J., 2017. Unsupervised monocular depth estimation with left-right consistency, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Godard, C., Mac Aodha, O., Firman, M., Brostow, G.J., 2019. Digging into self-supervised monocular depth estimation, in: Proceedings of the IEEE/CVF international conference on computer vision, pp. 3828–3838.

Gutnik, Y., Avni, A., Treibitz, T., Groper, M., 2022. On the adaptation of an auv into a dedicated platform for close range imaging survey missions. Journal of Marine Science and Engineering 10, 974.

Islam, M.J., Edge, C., Xiao, Y., Luo, P., Mehtaz, M., Morse, C., Enan, S.S., Sattar, J., 2020. Semantic segmentation of underwater imagery: Dataset and benchmark, in: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE. pp. 1769–1776.

Junayed, M.S., Sadeghzadeh, A., Islam, M.B., Wong, L.K., Aydın, T., 2022. Himode: A hybrid monocular omnidirectional depth estimation model, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5212–5221.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al., 2023. Segment anything. arXiv preprint arXiv:2304.02643 .

Kumar, G.S., Painumgal, U.V., Kumar, M.C., Rajesh, K., 2018. Autonomous underwater vehicle for vision based tracking. Procedia computer science 133, 169–180.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al., 2023. Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 .

Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H., 2024. Depth anything: Unleashing the power of large-scale unlabeled data. arXiv preprint arXiv:2401.10891 .

Yang, R., Yu, Y., 2021. Artificial convolutional neural network in object detection and semantic segmentation for medical imaging analysis. Frontiers in oncology 11, 638182.