Eurographics Symposium on Rendering (2025)
B. Wang and A. Wilkie (Editors)

A Controllable Appearance Representation for Flexible Transfer and Editing

Santiago Jimenez-Navarro , Julia Guerrero-Viu & Belen Masia

Universidad de Zaragoza - I3A, Spain

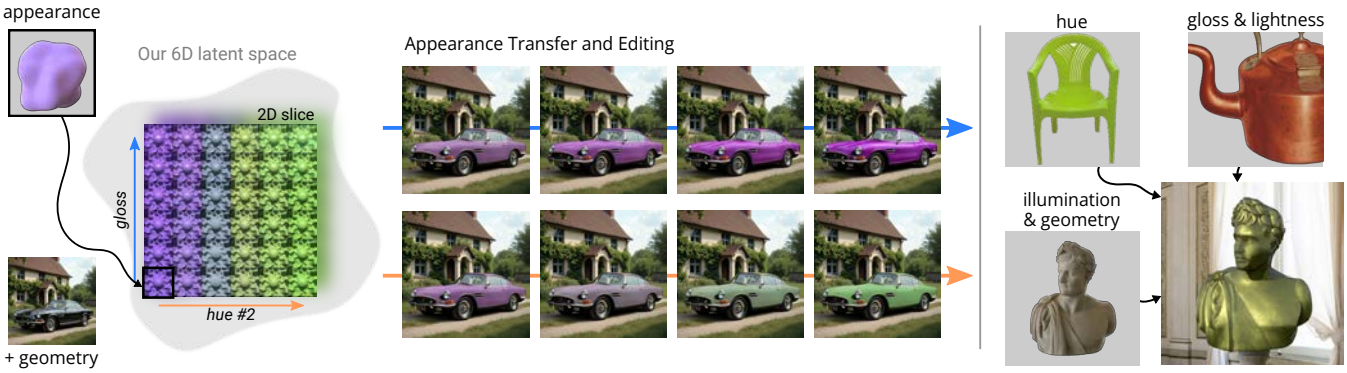

Figure 1: *Our model learns a disentangled and interpretable latent space of appearance in a self-supervised manner, without human-annotated data. Given an input image depicting a homogeneous object (left), it can be encoded into this space, which can then be traversed to generate meaningful variations of appearance (here, traversals along two dimensions encoding hue and gloss are shown). This encoded representation of appearance can be leveraged for appearance transfer and editing tasks: Given a target geometry (bottom left), we can transfer to it either the original appearance, or variations along any of the dimensions of the space (center). Since the space is disentangled and interpretable, it also enables selective appearance transfer (right): The resulting image (bottom right) is generated by selecting specific dimensions from each of the three inputs.*

Abstract

We present a method that computes an interpretable representation of material appearance within a highly compact, disentangled latent space. This representation is learned in a self-supervised fashion using a VAE-based model. We train our model with a carefully designed unlabeled dataset, avoiding possible biases induced by human-generated labels. Our model demonstrates strong disentanglement and interpretability by effectively encoding material appearance and illumination, despite the absence of explicit supervision. To showcase the capabilities of such a representation, we leverage it for two proof-of-concept applications: image-based appearance transfer and editing. Our representation is used to condition a diffusion pipeline that transfers the appearance of one or more images onto a target geometry, and allows the user to further edit the resulting appearance. This approach offers fine-grained control over the generated results: thanks to the well-structured compact latent space, users can intuitively manipulate attributes such as hue or glossiness in image space to achieve the desired final appearance.

Keywords: Latent representations; material appearance; self-supervised learning

CCS Concepts

• **Computing methodologies** → **Appearance and texture representations;**

1. Introduction

The visual appearance of a material in an image arises from the complex interplay of the material's reflectance properties, the lighting conditions, and the geometry of the object it is applied to. The combination of these gives rise to the proximal stimulus, which reaches our retinae and is interpreted by our visual system. Ade-

DOI: 10.2312/sr.20251187

quately characterizing material appearance is a fundamental goal of computer graphics. Traditionally, material properties are modeled through reflectance distribution functions, expressed either analytically, or in the form of tabulated data [MPBM03; DJ18]. More recent works rely on neural processes [ZZW*21] or SVBRDF maps, which allow characterization of complex textured materials from a single or few images [DAD*18; VMR*24]. These representations, while well suited for rendering, suffer from high-dimensionality, limited expressiveness, or a lack of interpretability, which hinders downstream tasks such as material compression or editing.

For these reasons, a number of works have been devoted to finding compact representations of material appearance, focusing on compression [HGC*20], interpolation [SSN18], editability [SWSR21], or the perceptual nature of the representation [SGM*16]. Depending on the application domain, different properties, such as interpretability or independence of dimensions, may be desirable in such representation. When aiming for interpretable spaces, existing approaches often rely on large quantities of human-annotated data. These are very costly to obtain, even more so across different attributes [SCW*21; DLC*22]. Besides, it is unclear a priori which attributes these should be [MPBM03; SGM*16; TGG*20]. Consequently, in this work we explore the use of self-supervised learning for identifying the underlying factors that determine material appearance, avoiding the need for labeled data in the creation of an interpretable and controllable latent space.

We leverage FactorVAE [KM18], a well-known method for disentangled representation learning, and build upon it to adapt it to our scenario. Specifically, we modify its original architecture to incorporate geometry information in the decoder, enforcing the bottleneck to learn the appearance separate from the geometry, and we also modify its loss function to avoid posterior collapse in our setup. Further, we carefully design a synthetic dataset that enables the FactorVAE to learn, in a self-supervised manner, explainable and independent dimensions encoding material appearance of homogeneous, opaque objects. Notably, and unlike many previous approaches, we work in image space, which has two advantages: first, we can encode the appearance of objects from images in the wild, and second, our model encodes visual appearance of the material, and not only material properties.

Our appearance encoder model thus enables encoding an input image, depicting an object, into a six-dimensional disentangled representation of the appearance of the object in the image. Despite the model being trained without labeled data, the six dimensions are interpretable, and encode hue (two dimensions), illumination (two dimensions), lightness and glossiness (note that both the illumination and the reflectance properties are included in this appearance representation). Since they are disentangled, the latent variable controlling gloss will change only with variation in object gloss, while the rest will remain constant. This greatly facilitates controllability, and therefore applications like editing, or selective attribute transfer. Fig. 1, left, shows the traversal of our learned latent space along the subspace spanned by two of its dimensions.

We demonstrate the potential of this disentangled, interpretable and controllable space for two applications: appearance transfer and editing (Fig. 1, center). We do this by using the latent appearance representation as a guidance to train a second model: a lightweight IP-Adapter [YZL*23] which effectively translates the information from the latent space to a diffusion pipeline, leveraging their generative ability. Specifically, we use a pre-trained latent diffusion model, based on Stable Diffusion XL [PEL*], and condition the generative process through two distinct branches: one that encodes the appearance via our latent space, and another one to integrate the target geometry. In this way, the appearance branch can either encode directly the appearance of a material in an input image (transfer) or be edited to generate variations of an existing material (editing). Interestingly, under this scheme, appearance transfer can be done from a single input image, or from multiple, performing selective attribute transfer from each; an example of this is shown in Fig. 1, right, where the appearance in the final image results from combining the hue of an exemplar, the gloss and lightness of another, and the geometry and illumination of a third one.

Performing appearance transfer in this way has advantages over existing approaches [CSM*24], since it offers a better disentanglement between appearance and geometry, and therefore more control over the transfer. This increased control is also an advantage of our method when compared to other methods for diffusion-based image editing [BHE23]: Diffusion models are typically trained on text-image pairs, relying on text prompts as the primary conditioning method. Despite its wide expressivity, the *text-only* control introduces ambiguities that can significantly limit its application for the particular case of material appearance, thus benefiting from image-based conditioning.

Overall, our latent representation offers increased controllability and interpretability, demonstrating its potential for appearance transfer and editing. We therefore make the following contributions:

- A self-supervised model that encodes an object in an input image into a disentangled and interpretable representation of its appearance.
- A large-scale dataset of almost 100,000 synthetic images carefully designed for self-supervised learning of homogeneous material appearance.
- Application of our representation to flexible appearance transfer and editing, through conditioning of a diffusion-based pipeline.

Our trained models, code, and dataset are available in `http://graphics.unizar.es/projects/mat_disentanglement_2025/`

2. Related Work

2.1. Low-Dimensional Material Appearance Representations

Material appearance is shaped by complex interactions between several factors, including surface properties, lighting, geometry, or viewing conditions [FDA03; LSGM21], often requiring high-dimensional data for accurate modeling. A number of works have attempted to find a compact representation of appearance, searching for low-dimensional BRDF or BTF embeddings [MPBM03; RJGW19; RGJW20; Kuz21], enabling applications such as compression or editing. However, they usually require costly human-annotated datasets [SGM*16; TGG*20; SWSR21], or produce

spaces that lack interpretability because their focus is on some other aspect, such as compression [HGC*20; SSN18; ZZW*21]. While the former work in material space, a series of approaches have focused on working directly in image space, to account for the interplay of confounding factors like geometry or illumination in the final perception of appearance [LMS*19; SCW*21]. Still, they rely on supervised learning-based methods and require large amounts of human annotations, which can be partially alleviated with weak supervision [GSS*24]. Some methods have focused directly on material editing in image space [DLC*22; SL23], leveraging GAN-based frameworks to allow controlled modification of specific attributes, like glossiness or metallicness, and also requiring ground-truth human labels for training. Unsupervised learning of appearance representations has been mostly limited to specific attributes, like gloss [SAF21] or translucency [LSX23], training on datasets with limited variability. In this work, we present a self-supervised model that learns a highly-compact latent space of material appearance. Remarkably, our model can disentangle interpretable factors like color or glossiness in images depicting homogeneous real-world materials, without any prior knowledge or annotated data.

2.2. Disentangled Representation Learning

Disentangled representation learning [WCWZ*24; LBL*19] aims to separate underlying factors of variation within data, improving interpretability in generative models. Variational Autoencoders (VAEs), including βVAE [HMP*17], achieve this by balancing reconstruction fidelity and latent space regularization, while FactorVAE [KM18] introduces a Total Correlation (TC) penalty via a discriminator network to enhance factor independence. Extensions like βTCVAE [CLGD18] propose different implementations of this TC term. A key challenge in these models is the posterior collapse, where latent variables lose informativeness by matching the prior too closely. Due to its relevance, this issue has been widely addressed [SRM*16; FLL*19; YWY*20; KOF*23]. Beyond VAEs, GAN-based [CDH*16] and diffusion models [YWLZ23] have also been explored for disentanglement. Nevertheless, the explicit modeling of a latent space in VAE-based models facilitates learning independent factors within a compact representation. Closer to our work is that of Benamira et al. [BSP22], that used βVAE to disentangle material appearance from measured BRDFs. In contrast, our model disentangles material appearance in *image space*, accounting for the influence of factors like illumination or geometry. We adapt FactorVAE to mitigate posterior collapse, and enable a self-supervised disentangled latent space to encode and modify material appearance (see Sec. 3).

2.3. Diffusion-Based Material Transfer and Editing

Since the seminal work from Ho et al. [HJA20], diffusion models have revolutionized image generation with exceptional quality and diversity, by progressively denoising from random noise [SCS*22; BGJ*23; RBL*22]. Conditioning mechanisms have recently played a central role to expand the functionality of diffusion models for *controlled* generation and editing. These include techniques like ControlNet [ZRA23] for multi-modal conditioning, LoRA [HSW*22] for task-specific fine-tuning, or IP-Adapter [YZL*23] and T2I-Adapter [MWX*24] for image-based conditioning, by injecting features from reference images alongside text prompts to influence the generation.

In the context of material appearance, diffusion models have been used for material generation [ZLX*24], physically-based synthesis [VBP*24], material capture [VMR*24], or texture editing [GHR*24]. In ColorPeel [BWVvdW24], they condition diffusion models on disentangled properties like color and texture, allowing for fine-grained editing but requiring labeled data to train the conditioning, which complicates generalization to novel properties. Cheng et al. [CSM*24] demonstrated zero-shot material transfer by injecting CLIP [RKH*21] embeddings of reference materials into a diffusion pipeline, achieving compelling results without any additional model training. However, their reliance on CLIP limits disentanglement between material properties and additional factors, such as lighting and geometry. Alchemist [SJL*24] showed impressive performance on material editing in image space, training diffusion models for specific material attributes with supervised ground-truth labels.

In our work, we train an IP-Adapter to condition a pre-trained diffusion model based on Stable Diffusion-XL [PEL*] on our compact self-supervised appearance representation, to showcase proof-of-concept applications of it (namely, appearance transfer and editing).

3. A Disentangled Latent Space for Appearance

In this section, we seek a material latent representation that disentangles the underlying factors responsible for its appearance in a given image. To do this, we train a variational autoencoder that reconstructs the appearance of an input image, while enforcing a latent space with independent dimensions encoding this appearance. The model (Sec. 3.1) takes as input an image of an object, made from a homogeneous material, and encodes it into a disentangled latent representation of the material's appearance in that image; this representation includes both the reflectance and the illumination. It can also decode such a representation, together with an input normal map, into an image of an object made of such material, whose geometry is determined by the input normals. This model is trained in a self-supervised manner, leveraging a dataset of almost 100,000 images specifically created for representation learning of material appearance (Sec. 3.2). We analyze the resulting latent space and the model's reconstruction ability in Sec. 4.

3.1. Method

We employ an encoder-decoder architecture to create a bottleneck of reduced dimensionality that encapsulates the internal representation of appearance (Fig. 2). We build our model on FactorVAE [KM18], a variational autoencoder for self-supervised learning of disentangled representations, for its effective balance between disentanglement and reconstruction quality. We introduce two key modifications to the original FactorVAE to make it suitable for our goal: (1) we modify the loss to avoid posterior collapse (Sec. 3.1.1), and (2) we input geometry information to the model in the form of normal maps, compelling the latent space to focus on appearance (Sec. 3.1.2).

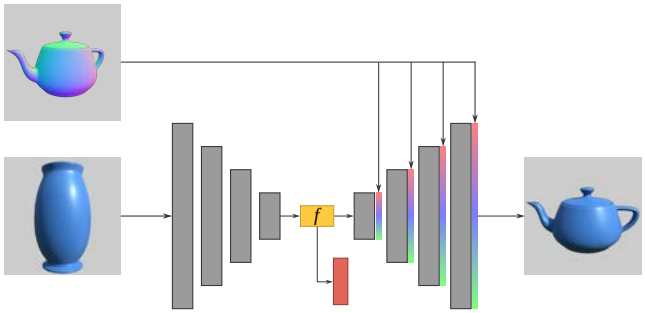

Figure 2: ***Diagram of our VAE-based architecture.*** *The encoder creates a low-dimensional, disentangled representation f of the appearance of the input image. The decoder learns to apply the incoming appearance to a reference geometry, specified with a normal map, which is concatenated in the decoder pipeline. The red box illustrates the discriminator used to compute the TC term (see text for details).*

3.1.1. Avoiding Posterior Collapse

The training loss proposed in the FactorVAE paper enforces the reconstruction of the input image, and the informativeness and independence of the dimensions of the latent space. To do so, it has three distinct terms: a term for reconstruction quality, a regularization term, and a term enforcing independence between factors (or Total Correlation term, TC). VAE-based models like this one, however, often suffer from a phenomenon called *posterior collapse*, in which the distribution learned by the encoder collapses to its prior, thus storing no relevant information about the generative factors of data [KM18; LTGN19]. Our proposed loss function $\mathcal{L}_{\theta,\phi}$ therefore is based on that of FactorVAE, with some modifications to avoid posterior collapse. It has the following formulation:

$$\begin{aligned} \mathcal{L}_{\theta,\phi}(x) &= \mathbb{E}_{z\sim q_\theta(z|x)}\left[\log p_\phi(x \mid z)\right] \\ -\beta D_{KL}(q_\theta(z \mid x), p(z), n) &- \gamma D_{KL}(q_\theta(z), \bar{q_\theta}(z), 1), \end{aligned} \tag{1}$$

where

$$\bar{q_\theta}(z) := \prod_{j=1}^{d} q_\theta\left(z_j\right), D_{KL}(Q,P,n) := \left\| \frac{1}{2}(\mu^2 + \sigma^2 - log(\sigma^2) - 1) \right\|^n$$

In Eq. 1, the probabilistic encoder $p(z \mid x)$ is approximated to the distribution $q_\theta(z \mid x)$, where θ corresponds to the weights learned by the network, and x describes a sample represented as z in the latent space. Analogously, $p_\phi(x \mid z)$ represents the probabilistic decoder, approximated with the weights ϕ.

For the reconstruction term we use a smooth L1 loss [Gir15] between the input image and the reconstructed one.

The second term serves as a regularizer of the latent space by minimizing the dimension-wise Kullback-Leibler (KL) divergence between the learned distribution and a prior $p(z)$. Being a variational model, the standard choice for the prior is a normal distribution. This term contains two key modifications with respect to the original formulation, aimed to address the posterior collapse issue. First, we apply a norm of order n (instead of the default summation) to the result of the KL operator. This encourages all dimensions to have a similar distance to the prior, thus storing a similar amount of information, effectively countering the posterior collapse. Second, we extend the FactorVAE loss by incorporating a weight β in this term, not present in the original definition. This weight allows, during training, to specify how much the learned distributions should resemble the prior. Additionally, we apply a linear annealing to the β term during training [SRM*16]. Gradually increasing the adherence to the prior distribution encourages the model to distribute information more evenly across the latent dimensions in the early training stages, further mitigating posterior collapse.

The TC term [Wat60] encourages the model to learn *independent* latent dimensions. In our application, this translates to learning different attributes in each dimension of the latent space. Following the original implementation of FactorVAE [KM18], we compute this TC term using an external discriminator.

3.1.2. Enforcing Appearance Encoding

In order for our representation to focus on the appearance of the surfaces, and not on their geometry, we input geometrical information to the model, in the form of normal maps, in the decoder [DLC*22], as shown in Fig. 2. This encourages the representation learnt by the encoder to focus on the reflectance and illumination. As a result, the geometry will not be identified as a factor of variation in the latent space. Moreover, this approach significantly improves the model's generalization ability, enabling it to more efficiently learn the desired explainable and independent factors during training.

As a result of the aforementioned adaptations (which we ablate in Sec. 4.4 and the supplementary material (S4)), our model learns a latent space with six dimensions, where each dimension represents an independent factor of material appearance. This represents a substantial reduction of dimensionality, from an RGB image (with a spatial resolution of 256 $\times$ 256 in our implementation) into a 6D feature vector f that successfully encodes the material's appearance, as shown in Sec. 4. Implementation details can also be found in the supplementary material (S1.1).

3.2. A Dataset for Material Appearance Disentanglement

Our model is trained in a self-supervised manner, learning to apply an input appearance onto a reference geometry, while building a well-structured latent space by optimizing the loss function (Eq. 1).

Existing datasets of material appearance are relatively abundant [SAF21; DLC*22; SCW*21]. However, they often include simple and unrealistic geometries with limited diversity, or are highly unbalanced with respect to appearance factors such as glossiness or hue. This is particularly problematic for our *self-supervised* training, as it could introduce unintended biases into the learned appearance representations. Moreover, we seek for a larger-scale dataset to improve generalization.

Therefore, we carefully designed a training dataset comprised of 98,550 synthetic images of 30 different objects rendered with 365 measured BRDFs [MPBM03; DJ18; SCW*21]. In order to facilitate the understanding of illumination by the network, we systematically vary the lighting for each object and material combination, leading to 9 lighting conditions (9 $\times$ 30 $\times$ 365 = 98,550).

Table 1: ***Quantitative metrics*** *obtained by our model and three baselines. All models are trained on our novel dataset (Sec. 3.2), and metrics are computed on our test dataset. Disentanglement and interpretability metrics are computed analyzing the structure of the latent space, using ground truth labels when necessary. Reported values are the mean and standard deviation of 5 independent trials, where each trial computes the metrics with a representative set of images. Arrows indicate desired performance, and the best value for each column is marked in bold.*

	Disentanglement		Interpretability		Reconstruction quality		
	GTC ↓	MIS ↓	Z-min ↑	MIR ↑	SSIM ↑	LPIPS ↓	PSNR ↑
βVAE	-	0.8825 ± 0.0197	0.5498 ± 0.0119	0.2224 ± 0.0046	0.5529 ± 0.0922	0.6642 ± 0.1638	29.129 ± 0.748
βTCVAE	-	0.8620 ± 0.0248	0.5717 ± 0.0061	0.2634 ± 0.0078	0.6177 ± 0.0916	0.6266 ± 0.1443	30.082 ± 1.490
FactorVAE	1.0633	0.8556 ± 0.0365	0.5521 ± 0.0041	0.2626 ± 0.0097	0.6556 ± 0.0596	0.5934 ± 0.0984	30.366 ± 1.290
Ours	**0.6686**	$\mathbf{0.8204 \pm 0.0514}$	$\mathbf{0.6955 \pm 0.0185}$	$\mathbf{0.2940 \pm 0.0089}$	$\mathbf{0.6775 \pm 0.0646}$	$\mathbf{0.4046 \pm 0.0648}$	$\mathbf{30.801 \pm 1.130}$

A representative set of images of the dataset, rendered with Mitsuba [Jak10], can be found in the supplementary material (S2).

4. Evaluation

We evaluate the ability of our model to find a disentangled latent space of material appearance, and to encode an input image into a suitable representation of the depicted material in this space. The evaluation is done both qualitatively and quantitatively, comparing its performance with various baselines. Specifically, we look at disentanglement and interpretability of the space, and at reconstruction quality of the model.

4.1. Quantitative Evaluation

Quantitatively, we compare our method to three baselines that are commonly used for disentangled representation learning (Sec. 2.2), all trained on our dataset (Sec. 3.2): βVAE [HMP*17], βTCVAE [CLGD18] and the vanilla FactorVAE [KM18]. The comparison evaluates disentanglement, interpretability and reconstruction quality. Our test set is comprised of images from the Serrano dataset [SCW*21]. From it, we select the ones featuring simple geometries (*blob* and *sphere*) to evaluate disentanglement and interpretability, to ensure that the lack of specialization of the baseline models in terms of geometric disentanglement is not overly penalized. To evaluate reconstruction quality we select images from a complex, unseen geometry (*statuette*). Note that, while some materials are both in our training and test sets, they are rendered with different illuminations and scene configurations, resulting in different appearance; besides, all baselines we compare to are trained and tested on the same sets.

Disentanglement is typically measured quantitatively with Gaussian Total Correlation (GTC) [Wat60] and the Mutual Information Score (MIS) [Sha48], which analyze the structure of the space and do not require labeled data. Measuring interpretability does require labeled data, and we measure it with the Z-min [KM18] and Mutual Information Ratio (MIR) [WDGB23] metrics, using ground truth labels available in the test dataset. We show all four metrics in Table 1. Our model clearly surpasses the rest, including the vanilla FactorVAE, showing the benefits of our modifications described in Sec. 3.1. Reconstruction quality is evaluated on the same test data with three widely-used metrics: SSIM [WBSS04], LPIPS [ZIE*18], and PSNR. Our model achieves better reconstructions, which can be attributed to the inclusion of geometry information in the decoder pipeline.

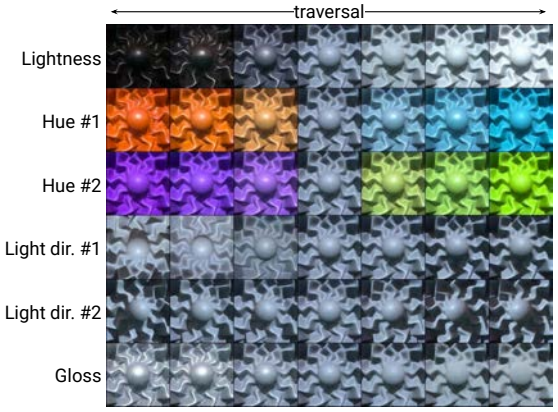

Figure 3: ***Prior traversals*** *sampling our 6D latent space. Each row samples a different dimension of the space, starting from a neutral, zero-valued feature vector (central column). The feature vector is fed to the decoder together with the normal map of the* Havran *geometry to generate the images shown. Our unsupervised model yields dimensions that are not only disentangled, but also interpretable, as indicated by the attributes we identify a posteriori.*

4.2. Qualitative Evaluation

Qualitatively, we can evaluate the disentanglement and interpretability of our latent space by visualizing the reconstructions generated by our decoder when sampling the latent space. To visualize the information encoded in the space, we take samples along each dimension (keeping the rest at a constant value of zero), generating feature vectors $f \in \mathbb{R}^6$ that are fed into the decoder together with a normal map. The results are the images shown in the *prior traversals* plot in Fig. 3. We chose the normal map of the *Havran* geometry for this visualization, as it is commonly used in perceptual studies, since it was specifically designed for broad light direction coverage, better showcasing a BRDF's appearance [HFM16]. In the figure, each row corresponds to the traversal of one latent dimension. We can see how the model identifies a slightly glossy gray material as the *neutral* one (zero-valued feature vector, central column). We can also clearly observe how, despite the lack of supervision, our model identifies interpretable factors that emerge from the data: Dimension 1 encodes lightness, dimensions 2 and 3 correspond to hue, dimensions 4 and 5 encode lighting direc-

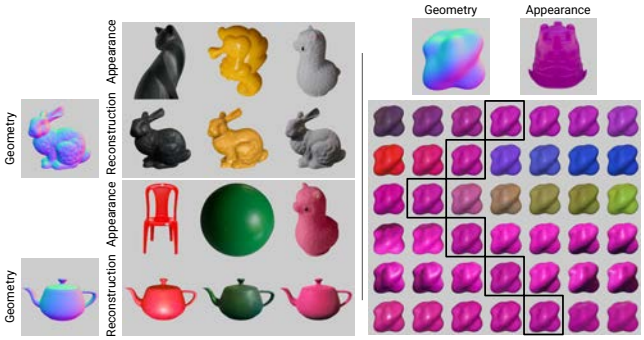

Figure 4: ***Left:*** *For six real-world, photographed objects (rows "Appearance", background has been masked), we encode their appearance with our model and decode it with a given normal map ("Geometry"). We see how the reconstructed objects (rows "Reconstruction", showing bunnies and teapots) exhibit the same appearance as their corresponding images, illustrating the ability of our model to encode appearance.* ***Right:*** *Posterior traversals of the latent space for the input image shown, reconstructed with the* blob *normal map (see text for details). Black boxes mark the appearance reconstructed from the input image.*

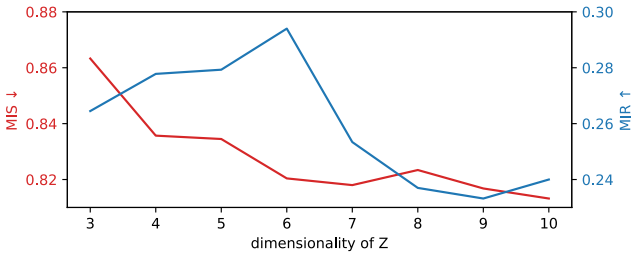

Figure 5: ***Latent space dimensionality analysis.*** *Evolution of metrics for interpretability (MIR, higher is better) and disentanglement (MIS, lower is better) for models whose latent space dimensionality ranges between 3 and 10. Our 6-dimensional space achieves the best balance between these two properties.*

Table 2: ***Ablation studies.*** *Our final model performs best both in terms of interpretability, as indicated by the MIR score, and reconstruction quality, measured with PSNR (see text for details).*

	MIR↑	PSNR↑
(a) Summation	0.2497	29.12
(b) Maximum $\beta = 1$	0.2306	30.08
(c) No β annealing	0.2130	30.36
Without Normals	0.2560	28.60
Ours	**0.2940**	**30.80**

tion (top to bottom and left to right, respectively), and dimension 6 represents gloss. Additional visualizations can be found in the supplemental material (S3).

Finally, we evaluate the ability of our model to encode the material appearance of a given image, *and* to modify such appearance along the dimensions of our latent space. Fig. 4, left, shows, for six real-world, photographed objects (the background has been masked), the result of encoding them with our model and decoding them using the normal map shown. The very similar appearance between the original and the reconstructed result shows the ability of our space to successfully encode the appearance of real objects. In Fig. 4, right, we also encode an input image of a real object into the latent space, obtaining its feature vector, and we then sample this space in each dimension, akin to what was done in the prior traversals plot in Fig. 3, but with the posterior, leading to *posterior traversals* plots. We can see how the material appearance of the input image is correctly captured (black boxes), and how we can modify the original appearance in a controlled manner by traversing the dimensions of the latent space. However, despite the great performance shown by the FactorVAE encoder in successfully distilling a disentangled and interpretable material appearance representation from the input image, the decoder pipeline proves insufficient when generating images of geometries very different from those seen during training (see supplementary material (S3.3) for more details). This highlights the need of a more powerful framework that translates the feature vector obtained into a final image, which is explored further in Sec. 5.

4.3. Dimensionality of the Latent Space

Our 6-dimensional space is highly compact, as compared to other alternatives in the literature (e.g., CLIP embeddings [RKH*21] can be 512D (ViT-B/32, ViT-B/16), 768D (ViT-L/14) or 1024D (ViT-H/14), and StyleGAN-based autoencoders [KLA19] are 512D and higher), facilitating interpretability. We explore the effect of modifying the dimensionality of our space, aiming to keep a balance between interpretability and disentanglement: larger latent spaces tend to dilute information in more dimensions, penalizing interpretability, while smaller ones need to embed the same information in a more constrained latent space, which often leads to worse disentanglement.

We analyze models from three to ten dimensions in the latent space, including quantitative metrics in Fig. 5 and qualitative results in Fig. 6, in which we show: (1) the factors captured in each dimension of the latent space, by computing the *prior traversals* plots, and (2) the informativeness of the space, by plotting the dimension-wise KL loss during training. Higher values in this plot represent that the learned distribution is different from the standard normal distribution $N(0,1)$, and thus are storing more information. Our model with six dimensions achieves the best balance between interpretability and disentanglement, as represented by MIR (higher is better) and MIS (lower is better) metrics, respectively (Fig. 5). This six-dimensional space is also more visually disentangled and interpretable, while not including uninformative dimensions (Fig. 6).

4.4. Ablation Studies

We evaluate the effectiveness of our design decisions by running ablation studies. We include here ablations on our proposed modifications of the loss function to tackle posterior collapse (Sec. 3.1.1),

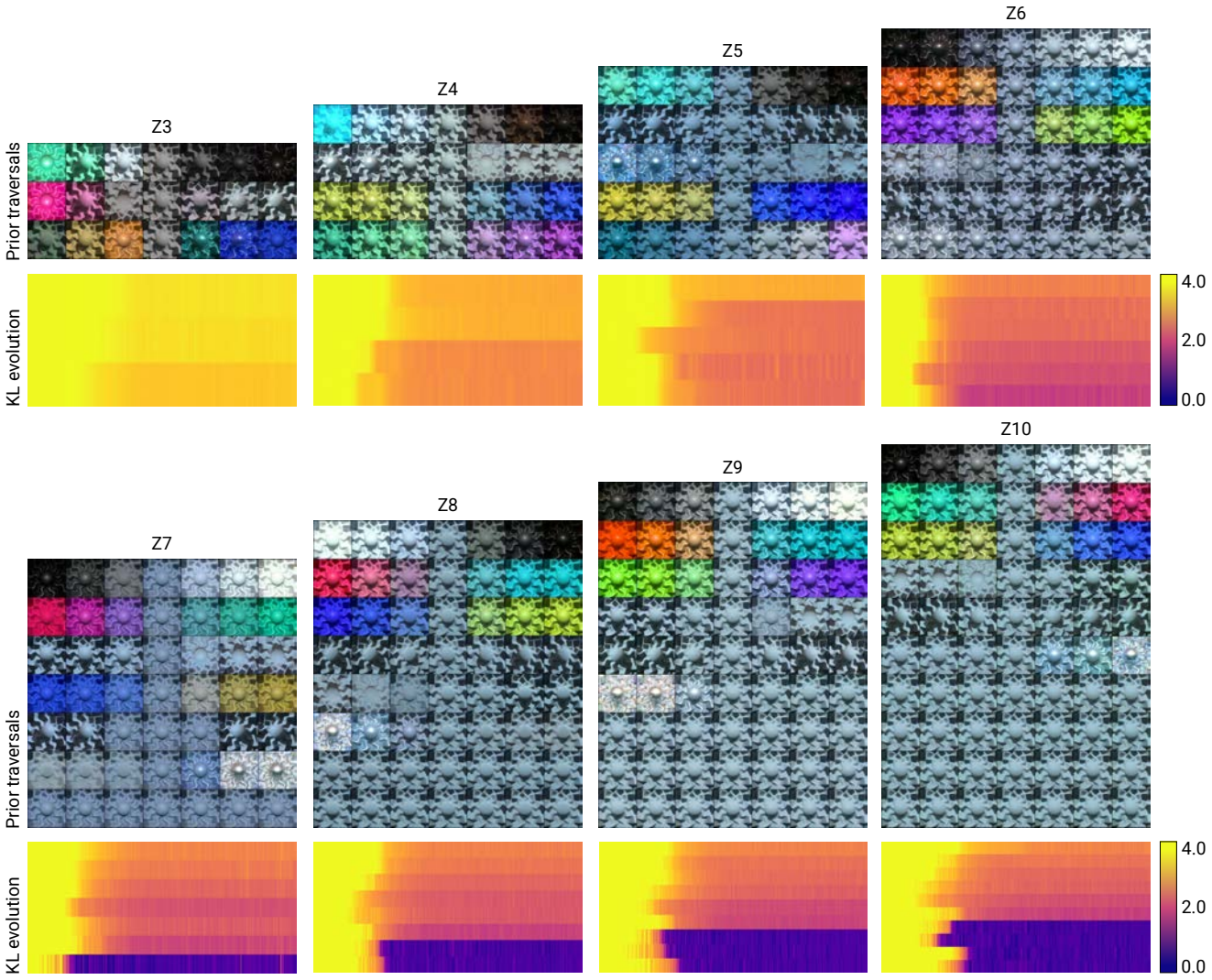

Figure 6: *Prior traversals and KL evolution plots of each model trained with latent space dimensionalities ranging from 3 to 10. KL evolution plots shows the dimension-wise evolution, during training, of the KL distance between the learned distributions and the standard normal distribution used to regularize (Sec. 3.1.1). The higher KL distance, the more information is stored in a given dimension.*

and on the use of normal maps to enforce the encoding of appearance (Sec. 3.1.2). For additional ablations on the reconstruction loss and normal map resampling, please refer to the supplementary material (S4).

4.4.1. Loss Function

We evaluate the effectiveness of the changes we propose with respect to the vanilla FactorVAE loss by training the following alternatives: (a) a model using the default summation instead of our proposed norm (i.e., $n = 1$ in the second term of Eq. 1), (b) a model with $\beta = 1$, and (c) a model without annealing on the β parameter with our default $\beta = 2$. Table 2 shows how systematically removing our modifications leads to models with reduced interpretability, as represented by lower MIR values.

4.4.2. Use of Normal Maps

In order to facilitate the disentanglement between appearance and geometry, we guide the reconstruction done by the model's decoder with normal maps. We ablate this modification in Table 2, training a model without this geometry guidance. As expected, we observe how leaving out this information diminishes reconstruction quality of the model, as measured by the PSNR metric. Additionally, the absence of normals leads to reduced interpretability (lower MIR) by requiring geometry to be encoded as an additional factor of variation.

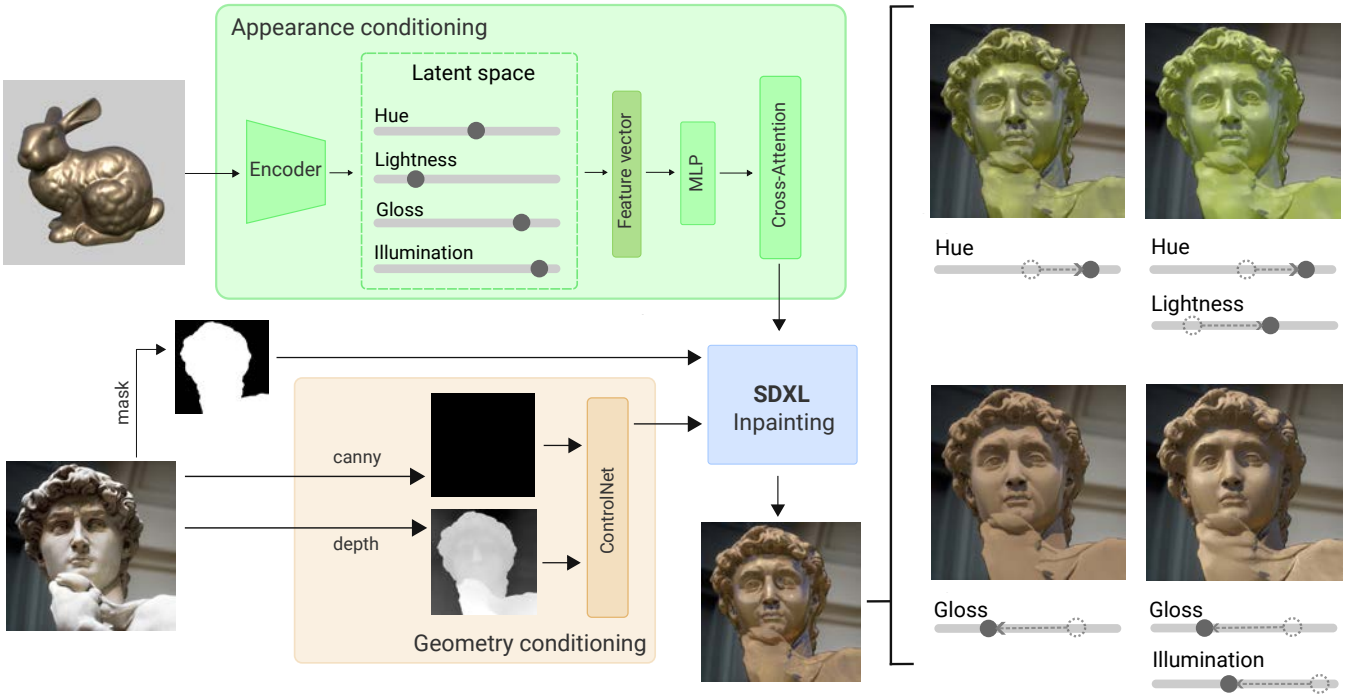

Figure 7: ***Diffusion-based pipeline for proof-of-concept applications of our space.*** *Our proposed pipeline uses two branches to condition the diffusion-based generative process with Stable Diffusion XL (SDXL). The appearance conditioning branch leverages our encoder to produce a 6D feature vector representing the desired appearance. This representation can be further edited along each of the six dimensions if desired, providing fine-grained control over the final appearance. The geometry branch leverages ControlNet to condition generation through Canny edges and depth information. We show here appearance transfer from an input image (*bunny*) to a target one (*David*), as well as editing along different dimensions of the latent space (right). Other uses, such as direct editing or selective transfer, are also possible.*

5. Applications: Appearance Transfer and Editing

We leverage our compact and controllable latent space (Sec. 3) for two applications: *appearance transfer*, which involves transferring the appearance of one or more reference exemplars to a target one, and *editing*, which modifies the visual appearance of an object in image space. We showcase these proof-of-concept applications by using our representation to condition a diffusion-based pipeline (Sec. 5.1), and evaluate its advantages and limitations (Sec. 5.2).

5.1. Diffusion-Based Pipeline

We design a diffusion-based pipeline that uses two sources of information as input: a *geometry reference* image, which defines the target geometry of the object, and an *appearance feature vector* in our latent space, which specifies the desired appearance to be applied to the target geometry. An overview of our pipeline is shown in Fig. 7. We use pre-trained latent diffusion model, RealisticVisionXL4.0, which is a fine-tune of the base model Stable Diffusion XL designed for photo-realism. We add our sources of information to condition the generative process through two distinct branches: an appearance conditioning branch and a geometry conditioning one.

The *appearance conditioning* branch encodes the material and illumination information from one or more reference images into an appearance feature vector by using our encoder (Sec. 3), effectively performing appearance transfer. Alternatively, one can directly sample the latent space to generate such vector. Further, the user can navigate the space, enabling controlled fine-grained adjustments to the generated images (appearance editing). Our appearance encoder, trained as explained in Sec. 3, is kept frozen during the training of the diffusion pipeline. We integrate the information of the appearance feature vector by training an IP-Adapter [YZL*23]. Following the IP-Adapter implementation, we plug an external network via a cross-attention mechanism, and train it by minimizing the original Stable Diffusion loss [RBL*22]. Only the weights of the external network are updated during training, in order to preserve the generative capabilities of the base model.

The *geometry conditioning* branch takes an image as input and incorporates the target geometry information into the diffusion pipeline via a combination of pre-trained ControlNets [ZRA23], designed to process Canny edges and depth maps. Depth information helps in transferring the general structure of the shape, while the Canny edges map allows to preserve the high-frequency details. For inference, we automatically obtain the depth map from the input image using a single-view depth predictor [KOH*24].

Finally, we follow previous work [CSM*24] and use an inpainting base model during inference, which restricts the generation to the area of the object, keeping the background intact. Please refer

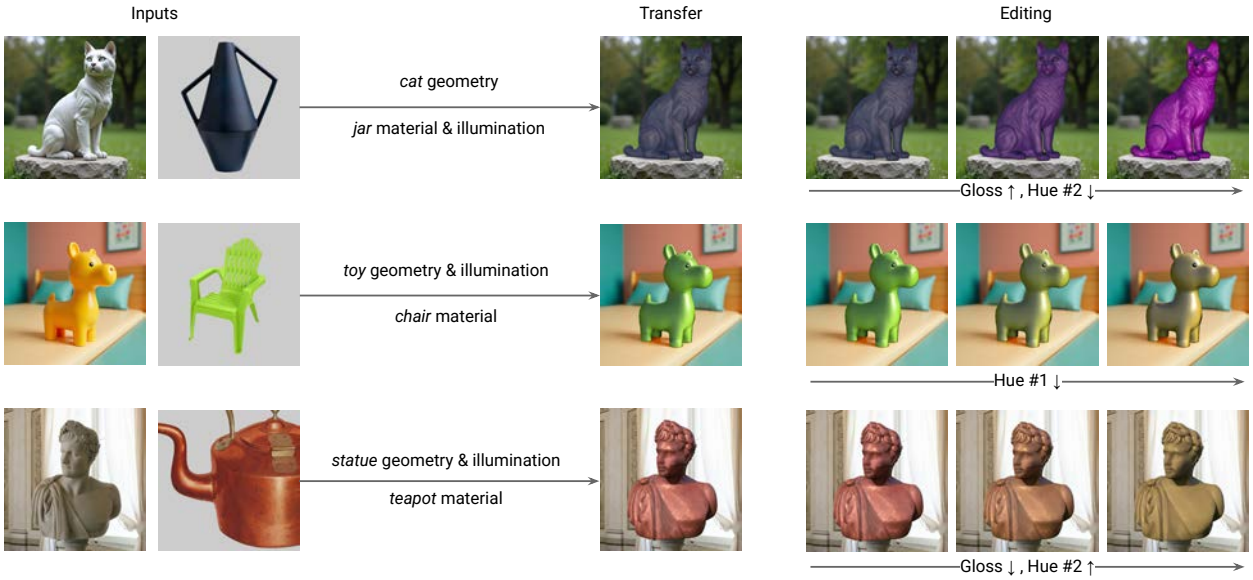

Figure 8: ***Appearance transfer and editing results*** *of our pipeline for real-world images leveraging our disentangled appearance representation. In each example, we use two input images, with the leftmost one as the target. We selectively transfer material and/or illumination properties from the other input image (background has been masked out) by encoding them into our latent space. We can further modify the appearance in this latent space, leading to fine-grained editing (right).*

to the supplementary material for full implementation details and ablations of our pipeline (S1.2).

Our pipeline is therefore versatile, and can be used for different applications depending on how the inputs are configured. Fig. 7 illustrates a canonical use case in which the appearance of the *bunny* is encoded with our model and transferred to the image of Michelangelo's *David* with our diffusion pipeline, performing appearance transfer. Once encoded, given the disentanglement and high controllability of our latent space, the user can easily modify the appearance vector and re-generate the image, thus performing fine-grained appearance editing, as shown. Our pipeline is however flexible to other use cases, such as introducing the same image for both appearance and geometry conditioning to perform direct editing (Fig. 11), selective transfer by extracting different factors of appearance from different input images (Fig. 8 and Fig. 1, right), manually defining the appearance of an object by sampling our latent space, or by interpolating in this space (Fig. 9).

5.2. Experiments

We evaluate our appearance-aware diffusion pipeline, showcasing applications of our disentangled appearance representation in different use cases. Additional results of our diffusion-based pipeline can be found in the supplementary material (S6).

5.2.1. Qualitative Evaluation

The fact that our space is disentangled enables integrating appearance information from two or more source images when performing appearance transfer. This is particularly useful for the illumination: one can perform material transfer from a source to a target image while keeping the illumination of the target (e.g., Fig. 8, second and third rows). Another example of selective transfer (i.e., transferring different dimensions from several images) beyond illumination dimensions is shown in Fig. 1 (right).

We include further results of our pipeline for both appearance transfer and editing tasks in Fig. 1 (center) and Fig. 8, using images not seen during training. In Fig. 8, for each row, we use two reference images as input (left) and perform appearance transfer between them (middle). Then, we modify this result by traversing the relevant dimension(s) of our latent space, performing disentangled editing (right). We show how our pipeline effectively captures the target appearance achieving realistic results, even when the reflectance properties are very different between the two input images (e.g., the copper glossy material on the statue, third row). The second and third rows further illustrate examples of integrating appearance information from two source images, as we extract the illumination from the feature vector of one of the images (two dimensions) and the material from the other (four dimensions). Despite being trained on synthetic data, our material and illumination appearance transfer generalizes well to real photographs. For editing, Fig. 8 shows how we can perform fine-grained modifications for different attributes while maintaining the identity of the image. We include an example modifying a single attribute (second row) to show the high disentanglement of our latent space. However, editing multiple attributes at once is also natively supported by our pipeline, as shown in the first and third rows.

Our pipeline can also be used to interpolate between two appearance feature vectors, as shown in Fig. 9. Despite significant differences in properties such as hue or gloss between both reference materials, the results exhibit realistic, smooth transitions.

Figure 9: ***Interpolation between two materials in our latent space.*** *The two ends of the progression show the result of transferring the material of two real-world objects, a blue glossy spoon and a pink rough llama, to a target geometry. All results follow the illumination of the target image. Progressively traversing our latent space results in an intuitive change of the appearance.*

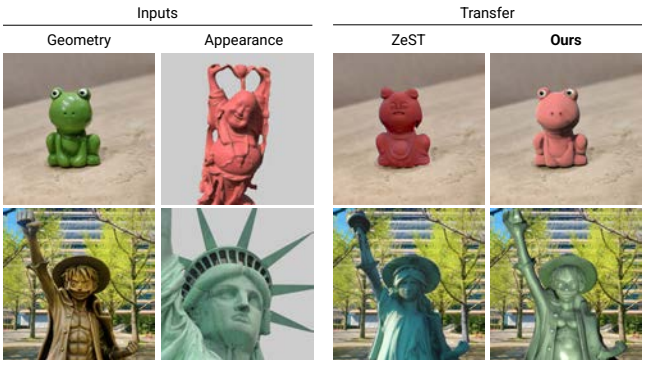

Figure 10: *Examples highlighting the disentanglement issue in ZeST [CSM*24] for appearance transfer. The appearance reference image of the first row is a rendering of a homogeneous buddha, while the second row uses a real photograph of the Statue of Liberty. In contrast to our method that properly disentangles appearance information from geometry, the ZeST results include geometric information being transferred to the output image.*

5.2.2. Benefits and Limitations for Transfer

The current state of the art in material transfer is ZeST [CSM*24]. An important difference with our approach lies in their use of a pre-trained *semantic* image encoder which projects the reference appearance image into the space of CLIP [RKH*21]. This encoder is more general than our appearance encoder, enabling ZeST to handle a wider variety of appearance.

However, unlike our method, CLIP has not been specifically trained to disentangle the appearance and geometry. As a result, some geometric information may be transferred through the appearance branch. Fig. 10 illustrates how this limitation of ZeST manifests in the generation of features such as the Buddha face or the Statue of Liberty's facial expression, which originate from the *appearance* image rather than the *geometry* image. As a result, details of the geometry reference image are modified or lost, and material appearance can be influenced by image semantics (e.g., Statue of Liberty). In contrast to CLIP, our appearance encoder explicitly encourages disentanglement between appearance and geometry, leading to a better performance than ZeST in this aspect (additional results in the supplementary material). Besides, our approach can be used to selectively transfer only certain attributes of a given set of inputs (Fig. 1, right).

5.2.3. Benefits and Limitations for Editing

We further evaluate our appearance editing task in Fig. 11, by comparing our results with two state-of-the-art editing methods in image space with publicly available code (Subias and Lagunas [SL23] and InstructPix2Pix [BHE23]), to highlight the benefits and limitations of our space. For every source image and method, we show appearance editing results by traversing two appearance factors, one after the other. The method by Subias and Lagunas is a GAN-based architecture trained supervisedly on specific attributes, so a direct comparison is only possible for the gloss attribute. InstructPix2Pix is a general image editing method conditioned on instructional text prompts, not solely targeted to appearance editing. While we lack such generality, compared to InstructPix2Pix, we can achieve finer-grained edits that only affect the desired material, as well as an increased control over the edit. This highlights the benefits of modeling an explicit disentangled latent space: although InstructPix2Pix has higher generative capacity, our approach is more suitable for *controllable* appearance editing. Finally, color shifts are a common challenge in generative models. While our method is not immune to this issue, its design allows for partial correction by manually adjusting the shifted dimensions to achieve the desired appearance, an ability that is often lacking in related works.

6. Discussion and Limitations

We show that we can, in a self-supervised manner and without the need for human-annotated data, learn a disentangled, interpretable and controllable space of appearance. We carefully evaluate the capabilities of such space, as well as alternative design decisions.

To illustrate potential uses of our space, we use it to condition a diffusion-based generative pipeline, enabling proof-of-concept applications: appearance transfer from one or more images, fine-grained editing, and interpolation within the space. We show these for real images, even though our models have been trained on synthetic data. We compare to existing dedicated material transfer or image editing methods for our sample downstream tasks. Despite the higher generative capacity of targeted methods, which we do not aim to surpass, our comparisons highlight the potential benefits of our space (i.e., fine-grained control and interpretability). Moreover, our representation of appearance could have alternative applications, such as generating a certain appearance from scratch by

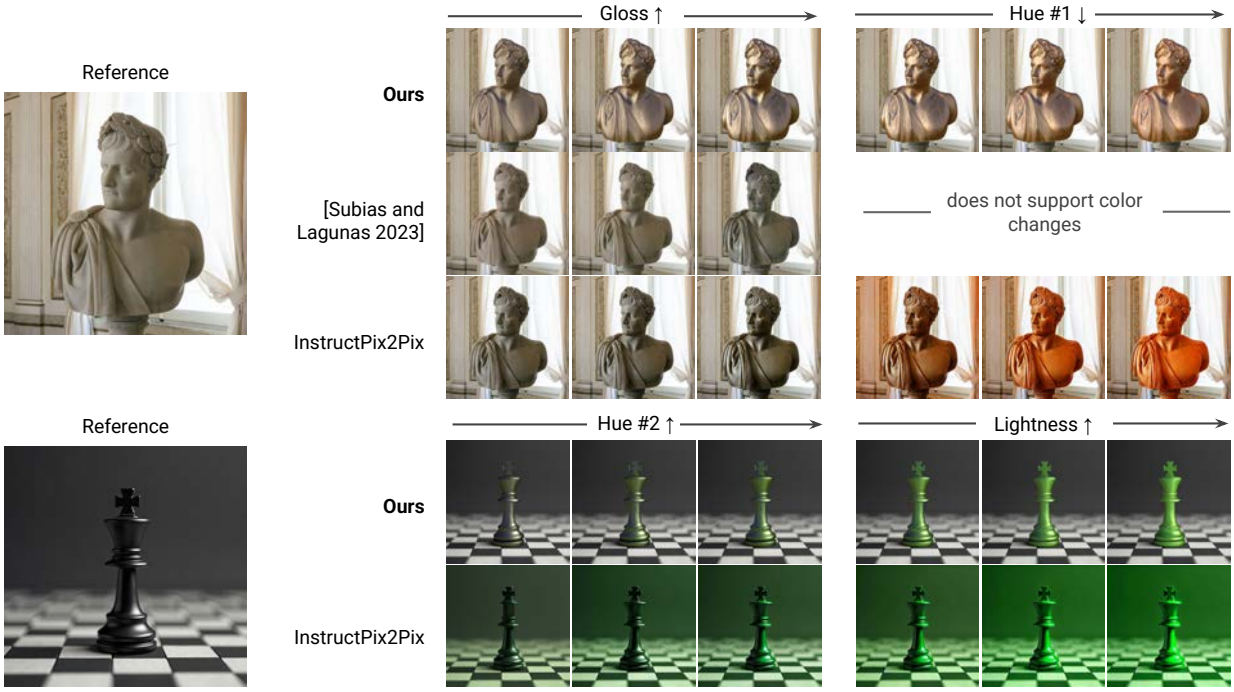

Figure 11: *Comparison of the appearance editing feature of our pipeline with two state-of-the-art solutions. Using a reference image (left), we show the progression of sequentially editing two attributes (right). The disentanglement of our latent space and the design of the diffusion pipeline allows for an intuitive and precise editing of the identified attributes.*

adjusting the different dimensions (exposed, e.g., as sliders), or to be used as a descriptor for attribute-based retrieval in large image databases.

Our appearance encoding model is limited to the range of appearances it was trained with, namely homogeneous, opaque materials and illuminations that do not exhibit very high frequency or strongly colored lighting. Fig. 12 shows reconstruction results when trying to encode samples with out-of-distribution, high-frequency illumination. Given the self-supervised nature of the approach, extension to a wider set of appearances increasing the training dataset remains as future work.

A promising direction for future research involves investigating the benefits of increasing the level of supervision in the training process. The current approach employs soft disentanglement constraints via the total correlation (TC) term in the loss function (Eq. 1). Still, we evaluate the disentanglement and interpretability of our space (Tab. 1), obtaining successful results. In future work, imposing hard constraints directly within the network architecture [KWKT15] could enable the enforcement of higher disentanglement in the learned latent dimensions. This could be implemented by fixing our extrinsic factors (geometry and illumination) and learning the intrinsic properties of the material.

Further, while the dimensions of our space are interpretable and controllable, they are not necessarily perceptually-linear, since no supervision enforces this. Although they are independent factors determining appearance, they may not be expressive enough and artists could prefer to control more, or alternative, dimensions.

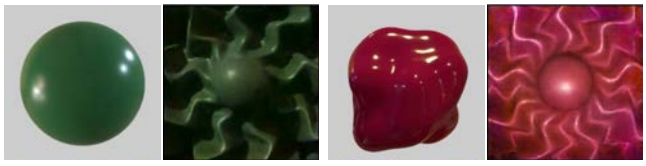

Figure 12: ***Limitations.*** *Examples of the behavior of our autoencoder when reconstructing out-of-distribution appearances. The* sphere *and* blob *samples are illuminated with high-frequency lighting. Reconstructing the* Havran *geometry using their respective embeddings, the model struggles and cannot recover appearance.*

Alternatively to our work, using supervision could allow to control specific factors of variation (e.g., varying parameters of an analytical BRDF model). However, it would necessarily require factors defining appearance to be defined *a priori*, which would introduce some bias, and could guide the space to learn the notion of, e.g., specular from the underlying model used. Instead, we have focused on self-supervised learning to investigate whether a meaningful appearance space can emerge from diverse realistic images, whose appearance factors of variation are a priori *unknown*. We hope our work inspires further exploration of self-supervised learning approaches to uncover the underlying factors that shape our perception of appearance.

Acknowledgments

This work has been partially supported by grant PID2022-141539NBI00, funded by MICIU/AEI/10.13039/501100011033 and by ERDF, EU, and by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 956585 (PRIME). Julia Guerrero-Viu was also partially supported by the FPU20/02340 predoctoral grant. We thank Daniel Martin for his help designing the final figures, Daniel Subias for proofreading the manuscript, and the members of the Graphics and Imaging Lab for insightful discussions. We also thank the people who participated in the user study, and the I3A (Aragon Institute of Engineering Research) for the use of its HPC cluster HERMES.

References

[BGJ*23] BETKER, JAMES, GOH, GABRIEL, JING, LI, et al. "Improving image generation with better captions". *Computer Science.* 2.3 (2023), 8 3.

[BHE23] BROOKS, TIM, HOLYNSKI, ALEKSANDER, and EFROS, ALEXEI A. "Instructpix2pix: Learning to follow image editing instructions". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, 18392–18402 2, 10.

[BSP22] BENAMIRA, ALEXIS, SHAH, SACHIN, and PATTANAIK, SUMANTA. "Interpretable Disentangled Parametrization of Measured BRDF with β-VAE". *arXiv preprint arXiv:2208.03914* (2022) 3.

[BWVvdW24] BUTT, MUHAMMAD ATIF, WANG, KAI, VAZQUEZ-CORRAL, JAVIER, and van de WEIJER, JOOST. "Colorpeel: Color prompt learning with diffusion models via color and shape disentanglement". *Proceedings of the European Conference on Computer Vision*. Springer. 2024, 456–472 3.

[CDH*16] CHEN, XI, DUAN, YAN, HOUTHOOFT, REIN, et al. "Infogan: Interpretable representation learning by information maximizing generative adversarial nets". *Advances in Neural Information Processing Systems* 29 (2016) 3.

[CLGD18] CHEN, RICKY TQ, LI, XUECHEN, GROSSE, ROGER B, and DUVENAUD, DAVID K. "Isolating sources of disentanglement in variational autoencoders". *Advances in Neural Information Processing Systems* 31 (2018) 3, 5.

[CSM*24] CHENG, TA-YING, SHARMA, PRAFULL, MARKHAM, ANDREW, et al. "Zest: Zero-shot material transfer from a single image". *Proceedings of the European Conference on Computer Vision*. Springer. 2024, 370–386 2, 3, 8, 10.

[DAD*18] DESCHAINTRE, VALENTIN, AITTALA, MIIKA, DURAND, FREDO, et al. "Single-image svbrdf capture with a rendering-aware deep network". *ACM Transactions on Graphics (TOG)* 37.4 (2018), 1–15 2.

[DJ18] DUPUY, JONATHAN and JAKOB, WENZEL. "An Adaptive Parameterization for Efficient Material Acquisition and Rendering". *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia)* 37.6 (Nov. 2018), 274:1–274:18 2, 4.

[DLC*22] DELANOY, JOHANNA, LAGUNAS, MANUEL, CONDOR, JORGE, et al. "A Generative Framework for Image-based Editing of Material Appearance using Perceptual Attributes". *Computer Graphics Forum*. Vol. 41. 1. Wiley Online Library. 2022, 453–464 2–4.

[FDA03] FLEMING, ROLAND W., DROR, RON O., and ADELSON, EDWARD H. "Real-world illumination and the perception of surface reflectance properties". *Journal of Visual Communication and Image Representation* 3.5 (2003). ISSN: 1534-7362 2.

[FLL*19] FU, HAO, LI, CHUNYUAN, LIU, XIAODONG, et al. "Cyclical Annealing Schedule: A Simple Approach to Mitigating KL Vanishing". *NAACL*. 2019 3.

[GHR*24] GUERRERO-VIU, JULIA, HASAN, MILOS, ROULLIER, ARTHUR, et al. "Texsliders: Diffusion-based texture editing in clip space". *ACM SIGGRAPH 2024 Conference Papers*. 2024, 1–11 3.

[Gir15] GIRSHICK, ROSS. "Fast R-CNN". *International Conference on Computer Vision (ICCV)*. 2015 4.

[GSS*24] GUERRERO-VIU, JULIA, SUBIAS, J DANIEL, SERRANO, ANA, et al. "Predicting Perceived Gloss: Do Weak Labels Suffice?": *Computer Graphics Forum*. Vol. 43. 2. Wiley Online Library. 2024, e15037 3.

[HFM16] HAVRAN, VLASTIMIL, FILIP, JIRI, and MYSZKOWSKI, KAROL. "Perceptually motivated BRDF comparison using single image". *Computer Graphics Forum*. Vol. 35. 4. Wiley Online Library. 2016, 1–12 5.

[HGC*20] HU, BINGYANG, GUO, JIE, CHEN, YANJUN, et al. "DeepBRDF: A Deep Representation for Manipulating Measured BRDF". *Computer Graphics Forum* 39.2 (2020), 157–166 2, 3.

[HJA20] HO, JONATHAN, JAIN, AJAY, and ABBEEL, PIETER. "Denoising diffusion probabilistic models". *Advances in Neural Information Processing Systems* 33 (2020), 6840–6851 3.

[HMP*17] HIGGINS, IRINA, MATTHEY, LOIC, PAL, ARKA, et al. "beta-vae: Learning basic visual concepts with a constrained variational framework." *ICLR (Poster)* 3 (2017) 3, 5.

[HSW*22] HU, EDWARD J, SHEN, YELONG, WALLIS, PHILLIP, et al. "LoRA: Low-Rank Adaptation of Large Language Models". *The Tenth International Conference on Learning Representations, ICLR*. 2022 3.

[Jak10] JAKOB, WENZEL. *Mitsuba renderer*. http://www.mitsuba-renderer.org. 2010 5.

[KLA19] KARRAS, TERO, LAINE, SAMULI, and AILA, TIMO. "A style-based generator architecture for generative adversarial networks". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, 4401–4410 6.

[KM18] KIM, HYUNJIK and MNIH, ANDRIY. "Disentangling by factorising". *International Conference on Machine Learning*. PMLR. 2018, 2649–2658 2–5.

[KOF*23] KINOSHITA, YURI, OONO, KENTA, FUKUMIZU, KENJI, et al. "Controlling posterior collapse by an inverse Lipschitz constraint on the decoder network". *International Conference on Machine Learning*. PMLR. 2023, 17041–17060 3.

[KOH*24] KE, BINGXIN, OBUKHOV, ANTON, HUANG, SHENGYU, et al. "Repurposing Diffusion-Based Image Generators for Monocular Depth Estimation". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024 8.

[Kuz21] KUZNETSOV, ALEXANDR. "Neumip: Multi-resolution neural materials". *ACM Transactions on Graphics (TOG)* 40.4 (2021) 2.

[KWKT15] KULKARNI, TEJAS D, WHITNEY, WILLIAM F, KOHLI, PUSHMEET, and TENENBAUM, JOSH. "Deep convolutional inverse graphics network". *Advances in Neural Information Processing Systems* 28 (2015) 11.

[LBL*19] LOCATELLO, FRANCESCO, BAUER, STEFAN, LUCIC, MARIO, et al. "Challenging common assumptions in the unsupervised learning of disentangled representations". *International Conference on Machine Learning*. PMLR. 2019, 4114–4124 3.

[LMS*19] LAGUNAS, MANUEL, MALPICA, SANDRA, SERRANO, ANA, et al. "A Similarity Measure for Material Appearance". *ACM Transactions on Graphics (TOG)* 38.4 (2019) 3.

[LSGM21] LAGUNAS, MANUEL, SERRANO, ANA, GUTIERREZ, DIEGO, and MASIA, BELEN. "The joint role of geometry and illumination on material recognition". *Journal of Vision* 21.2 (2021), 2–2 2.

[LSX23] LIAO, CHENXI, SAWAYAMA, MASATAKA, and XIAO, BEI. "Unsupervised learning reveals interpretable latent representations for translucency perception". *PLOS Computational Biology* 19.2 (2023), e1010878 3.

[LTGN19] LUCAS, JAMES, TUCKER, GEORGE, GROSSE, ROGER, and NOROUZI, MOHAMMAD. *Understanding Posterior Collapse in Generative Latent Variable Models*. 2019 4.

[MPBM03] MATUSIK, WOJCIECH, PFISTER, HANSPETER, BRAND, MATT, and MCMILLAN, LEONARD. "A Data-Driven Reflectance Model". *ACM Transactions on Graphics (TOG)* 22.3 (July 2003), 759–769 2, 4.

[MWX*24] MOU, CHONG, WANG, XINTAO, XIE, LIANGBIN, et al. "T2I-Adapter: learning adapters to dig out more controllable ability for text-to-image diffusion models". *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*. AAAI Press, 2024 3.

[PEL*] PODELL, DUSTIN, ENGLISH, ZION, LACEY, KYLE, et al. "SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis". *The Twelfth International Conference on Learning Representations, ICLR* 2, 3.

[RBL*22] ROMBACH, ROBIN, BLATTMANN, ANDREAS, LORENZ, DOMINIK, et al. "High-resolution image synthesis with latent diffusion models". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, 10684–10695 3, 8.

[RGJW20] RAINER, GILLES, GHOSH, ABHIJEET, JAKOB, WENZEL, and WEYRICH, TIM. "Unified neural encoding of BTFs". *Computer Graphics Forum*. Vol. 39. 2. Wiley Online Library. 2020, 167–178 2.

[RJGW19] RAINER, GILLES, JAKOB, WENZEL, GHOSH, ABHIJEET, and WEYRICH, TIM. "Neural BTF compression and interpolation". *Computer Graphics Forum*. Vol. 38. 2. Wiley Online Library. 2019, 235–244 2.

[RKH*21] RADFORD, ALEC, KIM, JONG WOOK, HALLACY, CHRIS, et al. "Learning transferable visual models from natural language supervision". *International Conference on Machine Learning* (2021), 8748–8763 3, 6, 10.

[SAF21] STORRS, KATHERINE R, ANDERSON, BARTON L, and FLEMING, ROLAND W. "Unsupervised learning predicts human perception and misperception of gloss". *Nature Human Behaviour* 5.10 (2021), 1402–1417 3, 4.

[SCS*22] SAHARIA, CHITWAN, CHAN, WILLIAM, SAXENA, SAURABH, et al. "Photorealistic text-to-image diffusion models with deep language understanding". *Advances in Neural Information Processing Systems* 35 (2022), 36479–36494 3.

[SCW*21] SERRANO, ANA, CHEN, BIN, WANG, CHAO, et al. "The effect of shape and illumination on material perception: model and applications". *ACM Transactions on Graphics (TOG)* 40.4 (2021) 2–5.

[SGM*16] SERRANO, ANA, GUTIERREZ, DIEGO, MYSZKOWSKI, KAROL, et al. "An intuitive control space for material appearance". *ACM Transactions on Graphics (TOG)* 35.6 (2016) 2.

[Sha48] SHANNON, CLAUDE ELWOOD. "A mathematical theory of communication". *The Bell System Technical Journal* 27.3 (1948), 379–423 5.

[SJL*24] SHARMA, PRAFULL, JAMPANI, VARUN, LI, YUANZHEN, et al. "Alchemist: Parametric control of material properties with diffusion models". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, 24130–24141 3.

[SL23] SUBIAS, J DANIEL and LAGUNAS, MANUEL. "In-the-wild Material Appearance Editing using Perceptual Attributes". *Computer Graphics Forum*. Vol. 42. 2. Wiley Online Library. 2023, 333–345 3, 10.

[SRM*16] SØNDERBY, CASPER KAAE, RAIKO, TAPANI, MAALØE, LARS, et al. "Ladder variational autoencoders". *Advances in Neural Information Processing Systems* 29 (2016) 3, 4.

[SSN18] SOLER, CYRIL, SUBR, KARTIC, and NOWROUZEZAHRAI, DEREK. "A versatile parameterization for measured material manifolds". *Computer Graphics Forum*. Vol. 37. 2. Wiley Online Library. 2018, 135–144 2, 3.

[SWSR21] SHI, WEIQI, WANG, ZEYU, SOLER, CYRIL, and RUSHMEIER, HOLLY. "A low-dimensional perceptual space for intuitive BRDF editing". *EGSR 2021-Eurographics Symposium on Rendering-DL-only Track*. 2021, 1–13 2.

[TGG*20] TOSCANI, MATTEO, GUARNERA, DAR'YA, GUARNERA, GIUSEPPE CLAUDIO, et al. "Three Perceptual Dimensions for Specular and Diffuse Reflection". *ACM Transactions on Applied Perception* 17.2 (May 2020). ISSN: 1544-3558 2.

[VBP*24] VAINER, SHIMON, BOSS, MARK, PARGER, MATHIAS, et al. "Collaborative control for geometry-conditioned PBR image generation". *Proceedings of the European Conference on Computer Vision*. Springer. 2024, 127–145 3.

[VMR*24] VECCHIO, GIUSEPPE, MARTIN, ROSALIE, ROULLIER, ARTHUR, et al. "ControlMat: A Controlled Generative Approach to Material Capture". *ACM Transactions on Graphics (TOG)* 43.5 (2024) 2, 3.

[Wat60] WATANABE, SATOSI. "Information Theoretical Analysis of Multivariate Correlation". *IBM Journal of Research and Development* 4.1 (1960), 66–82 4, 5.

[WBSS04] WANG, ZHOU, BOVIK, A.C., SHEIKH, H.R., and SIMONCELLI, E.P. "Image quality assessment: from error visibility to structural similarity". *IEEE Transactions on Image Processing* 13.4 (2004), 600–612 5.

[WCWZ*24] WANG, XIN, CHEN, HONG, WU, ZIHAO, ZHU, WENWU, et al. "Disentangled representation learning". *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024) 3.

[WDGB23] WHITTINGTON, JAMES C. R., DORRELL, WILL, GANGULI, SURYA, and BEHRENS, TIMOTHY. "Disentanglement with Biological Constraints: A Theory of Functional Cell Types". *The Eleventh International Conference on Learning Representations, ICLR*. 2023 5.

[YWLZ23] YANG, TAO, WANG, YUWANG, LU, YAN, and ZHENG, NANNING. "DisDiff: unsupervised disentanglement of diffusion probabilistic models". *Proceedings of the 37th International Conference on Neural Information Processing Systems*. 2023, 69130–69156 3.

[YWY*20] YAN, CHAOCHAO, WANG, SHENG, YANG, JINYU, et al. "Re-balancing variational autoencoder loss for molecule sequence generation". *Proceedings of the 11th ACM international conference on bioinformatics, computational biology and health informatics*. 2020, 1–7 3.

[YZL*23] YE, HU, ZHANG, JUN, LIU, SIBO, et al. "Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models". *arXiv preprint arXiv:2308.06721* (2023) 2, 3, 8.

[ZIE*18] ZHANG, RICHARD, ISOLA, PHILLIP, EFROS, ALEXEI A, et al. "The unreasonable effectiveness of deep features as a perceptual metric". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2018, 586–595 5.

[ZLX*24] ZHANG, YUQING, LIU, YUAN, XIE, ZHIYU, et al. "Dreammat: High-quality pbr material generation with geometry-and light-aware diffusion models". *ACM Transactions on Graphics (TOG)* 43.4 (2024), 1–18 3.

[ZRA23] ZHANG, LVMIN, RAO, ANYI, and AGRAWALA, MANEESH. "Adding conditional control to text-to-image diffusion models". *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, 3836–3847 3, 8.

[ZZW*21] ZHENG, CHUANKUN, ZHENG, RUZHANG, WANG, RUI, et al. "A Compact Representation of Measured BRDFs Using Neural Processes". *ACM Transactions on Graphics (TOG)* 41.2 (Nov. 2021). ISSN: 0730-0301 2, 3.