DOI: 10.1111/cgf.14446

COMPUTER GRAPHICS *forum*
Volume 41 (2022), number 1 pp. 453–464

A Generative Framework for Image-based Editing of Material Appearance using Perceptual Attributes

J. Delanoy, M. Lagunas, J. Condor, D. Gutierrez and B. Masia

Universidad de Zaragoza, I3A, Zaragoza, Spain

Abstract

*Single-image appearance editing is a challenging task, traditionally requiring the estimation of additional scene properties such as geometry or illumination. Moreover, the exact interaction of light, shape and material reflectance that elicits a given perceptual impression is still not well understood. We present an image-based editing method that allows to modify the material appearance of an object by increasing or decreasing high-level perceptual attributes, using a single image as input. Our framework relies on a two-step generative network, where the first step drives the change in appearance and the second produces an image with high-frequency details. For training, we augment an existing material appearance dataset with perceptual judgements of high-level attributes, collected through crowd-sourced experiments, and build upon training strategies that circumvent the cumbersome need for original-edited image pairs. We demonstrate the editing capabilities of our framework on a variety of inputs, both synthetic and real, using two common perceptual attributes (*Glossy *and* Metallic*), and validate the perception of appearance in our edited images through a user study.*

Keywords: image processing, image and video processing

CCS Concepts: • Computing methodologies → Machine learning; Image processing; Neural networks; Perception

1. Introduction

Material appearance is one of the most important properties that determine how we perceive an object. The visual impression that it elicits, whether it appears metallic, glossy or matte, strongly impacts how we manipulate such objects and expect them to behave. This appearance does not only depend on the intrinsic properties of the material itself, but also on external factors such as the geometry or the illumination of the scene. Editing material appearance based on a single image is therefore a very challenging task. A common approach is to estimate illumination, geometry and reflectance properties (inverse rendering), and modify the latter. This approach faces two problems. First, inaccuracies in the estimation of any of those scene properties can strongly impact the final result. Second, even if they are correctly estimated, modifying the reflectance parameters to obtain a certain visual impression of the material is not a trivial task, since the exact interaction of light, shape and material reflectance that elicits a given perception of appearance is still not well understood.

We present an image-based method for appearance editing that does not rely on any physically based rendering of the image, but instead modifies directly the image cues that drive the perception of the material. It takes a single image of an object as input and modifies its appearance based on varying the intensity of high-level perceptual attributes (see Figure 1). However, since the image cues that drive the perception of such attributes can not be captured in a few image statistics [FS19, SAF21], we rely on generative neural networks to learn their relationship with appearance, and generate novel images with the edited material. Our networks additionally take as input a normal map that helps preserve the high-frequency details of the input geometry in the reconstructed images. Since normal maps are not available in photographs, we provide a normal map predictor that extends the applicability of our method to real input images.

A possible approach to training our framework would be to collect pairs of (original, edited) images, where the edited exemplars were manually produced given a target high-level attribute value. This is not only cumbersome, but could also lead to high variability that could hamper the learning process. Instead, and taking inspiration from existing works on face editing [LZU*17, CUYH20, LDX*19, KWKT15], we train our system using perceptual judgements of the attributes of a large set of training images, that we

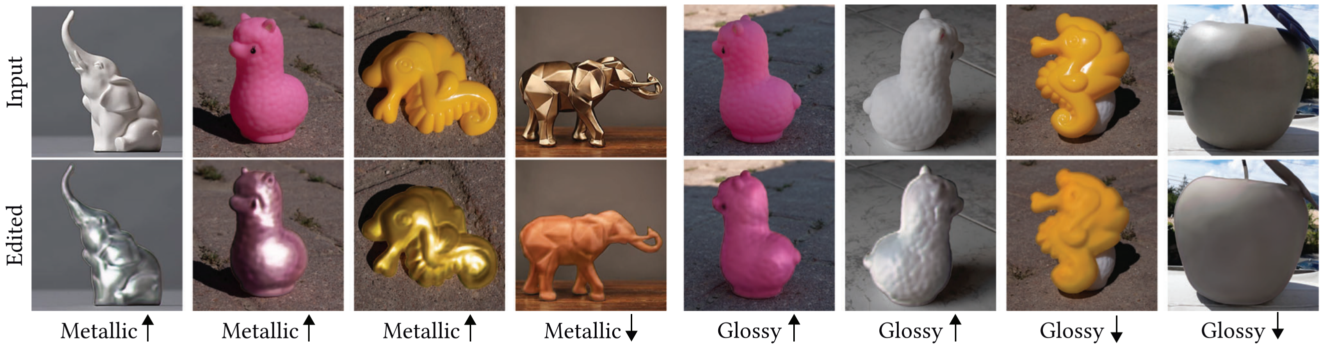

Figure 1: *Given a single image as input (top row), our framework allows to edit the appearance of objects using high-level perceptual attributes. It produces realistic edits (bottom row) for a variety of real images depicting objects with different material appearance, illumination and geometry. Note how illumination conditions are preserved in the edited results even though they were not explicitly modelled in the framework. Arrows indicate a high (pointing up) or low (pointing down) value of the target perceptual attribute.*

collect through crowd-sourced experiments. While these works benefit from a fixed camera location and exploit the fact that faces share similar geometry and features, we deal with a more unconstrained and varied set of potential input images. We thus devise a two-step framework, where the first step drives the change in appearance, while the second produces an image with high-frequency details.

To demonstrate the editing capabilities of our framework on a varied set of synthetic and real images, we focus on two attributes that are both common and easy to understand by participants: *Metallic* and *Glossy*. Without loss of generality, this allows us to collect robust human judgements of such attributes, while additionally assessing the perception of the appearance in our edited images through a user study. We validate our framework qualitatively, and by means of the aforementioned user study, as well as ablating each of its components. We will make our dataset of perceptual judgements publicly available to foster further research.

2. Related Work

2.1. Material perception

The exact way in which our visual system infers material properties from an image is yet to be understood [FDA03, FDA01]. It depends not only on the intrinsic properties of the material, but also on factors like motion [MLMG19, DFY*11], shape [VLD07, HFM16], illumination [HLM06, KFB10, VBF17] or the interactions between these [LSGM21].

A large body of work has been devoted to understand the visual cues that we use to infer isolated appearance properties such as glossiness [CK15, WAKB09, PFG00], translucency [GXZ*13, XZG*20, GWA*15] or softness [SFV20, CDD21], while others aimed at understanding the perceptual cues used by artists when depicting materials in realistic paintings [DCWP19, DSMG21]. Last, recent works have suggested that material perception might be driven by complex non-linear statistics, similar to the ones extracted by neural networks [FS19, SAF21, DLG*20]; our method is thus based on deep neural networks, which have the ability to model these complex visual cues and manipulate them in relation to perceptual data.

2.2. Editing of material appearance

Editing the appearance of materials is a complex task since there is a disconnect between their physical attributes and our perception [FWG13, TFCRS11, CK15]. We provide here a brief cross-section of different material editing approaches, and refer the interested reader to the more comprehensive review by Schmidt *et al.* [SPN*16].

Several perceptually based frameworks have been proposed to provide users with more intuitive controls over parametric appearance models [FPG01, PFG00, BPV18, KP10, DRCP14]. Non-parametric models such as measured BRDFs are harder to edit. Different approaches have been proposed, such as fitting the non-parametric BRDFs to parametric models [SJR18, BSH12, BP20], inverse shading trees [LBAD*06], polynomial bases [BAEDR08] or using deep-learning techniques [ZFWW20]. Closer to our work, other authors have proposed links between human perception and editing of non-parametric BRDFs through a set of intuitive perceptual traits [MGZ*17, SGM*16, MPBM03]. However, these methods only provide a new material definition that can later be used in a 3D scene, but do not allow to modify the material directly in an existing image.

Image-based material editing techniques allow the user to directly alter the pixels in an image without manipulating an underlying BRDF nor requiring to re-render a scene. The work of Khan *et al.* [KRFB06] exploits the fact that human vision is tolerant to many physical inaccuracies to propose a material editing framework requiring a single HDR image as input. Such approach was later extended to include global illumination [GSLM*08] or weathering effects [XWT*08]. Other methods are based on frequency-domain analyses [BBPA15], visual goals [NSRS13], or use a light field as input [BSM*18, JMB*14].

Since geometry and illumination also play a role in material appearance, several works focused on explicitly decomposing the image into material, illumination and geometry information [BM14, HFB*09, YS19, GMLMG12, OKT*19], allowing to manipulate each of these properties independently. Recently, neural networks have also been used for such decomposition [LCY*17, MLTFR19,

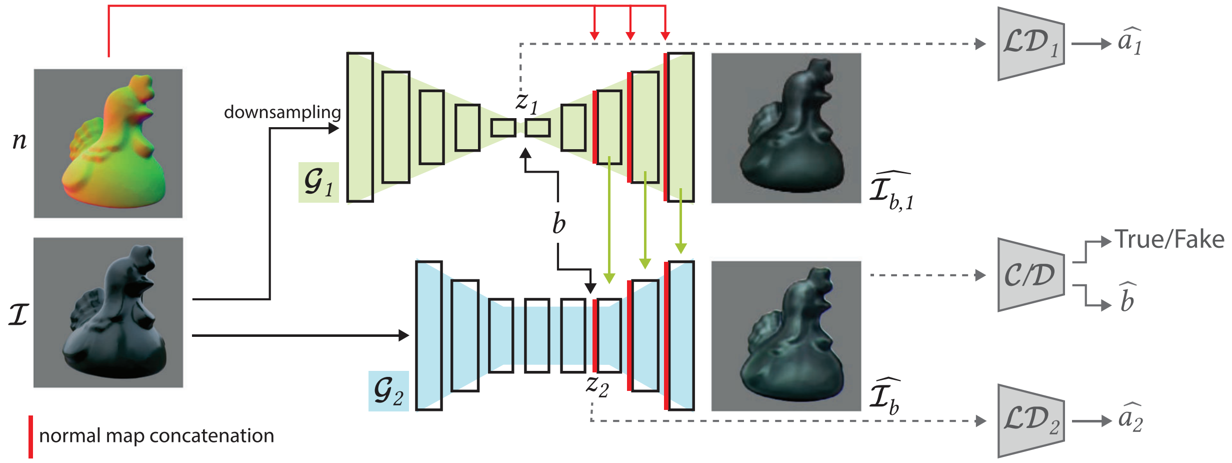

Figure 2: *Overview of the different components of our framework. The two networks $\mathcal{G}_1$ and $\mathcal{G}_2$ both take as input the image $\mathcal{I}$, its normal map n, and the target attribute value b. The image first goes through $\mathcal{G}_1$, whose decoder features are forwarded to the decoder of $\mathcal{G}_2$ (green arrows). $\mathcal{G}_2$ is in charge of producing the final output image $\widehat{\mathcal{I}_b}$. The three auxiliary networks ($\mathcal{LD}_1$, $\mathcal{LD}_2$ and $\mathcal{C}/\mathcal{D}$), shown in grey, are used at training time to guide the networks towards correctly interpreting the target attribute value b.*

RRF*16, LMF*19, GLD*19]. However, editing the material with such methods requires a robust estimation of all three layers. In contrast, we require only to estimate the geometry, providing plausible edits even under unknown lighting conditions. Our image-based framework is not based on visual goals; instead, it relies on appearance perception data, collected through crowd-sourced experiments, used to train a learning network.

3. Our Framework

3.1. Goal and overview

The goal of our method is to take as input an image $\mathcal{I}$ of an object, whose material appearance we wish to edit and without its background, and a target value $b_A \in [-1, 1]$ for a high-level perceptual attribute A (e.g. *Glossy* or *Metallic*), and from them produce a new image $\widehat{\mathcal{I}_b}$ that exhibits the same content as $\mathcal{I}$, but features a change in appearance according to the desired value of the perceptual attribute, b_A (hereafter, we drop the subindex A for clarity). Our method thus needs to extract or disambiguate the information of such attribute from the input image, and allow its subsequent manipulation to generate the final one. We leverage the success of generative neural networks on image-based editing tasks, and propose a framework based on them.

Producing a representation of $\mathcal{I}$ in which the information of the attribute has been disambiguated requires a *deep* model that can produce a compact latent code; however, such a model typically encompasses the loss of high-frequency details from the input image, hindering the reconstruction of the final image. We therefore propose a framework based on two generative networks, $\mathcal{G}_1$ and $\mathcal{G}_2$. $\mathcal{G}_1$ is a deeper network that aims at producing a compact latent code of $\mathcal{I}$ that is easy to control, and can be used to produce the final target appearance. Meanwhile, $\mathcal{G}_2$ is a shallower model that has the task of reconstructing the final image with high-frequency details, *guided by* the intermediate features of $\mathcal{G}_1$ that encode the relevant information on the final target appearance. An overview of our framework is shown in Figure 2, while the remainder of this section provides the details on the architecture, loss functions and training scheme used.

3.2. Model architecture

Both networks, $\mathcal{G}_1$ and $\mathcal{G}_2$, are based on an encoder–decoder architecture, in which the target attribute value b is concatenated at the bottleneck of each network (see Figure 2). Each encoder consists of a series of convolutional blocks that downscale the image by a factor of two, followed by a series of residual blocks. The output of these residual blocks is the latent code z_i ($i \in \{1, 2\}$), which we train to encode a representation of the input image $\mathcal{I}$ that does not contain information about the perceptual attribute. In particular, we have six convolutional blocks for $\mathcal{G}_1$, and three for $\mathcal{G}_2$. Each decoder consists of a series of convolutional blocks followed by bilinear upsampling that restore the original resolution of the image. The complete description of the architecture of each network can be found in the supplementary material.

One of the main drawbacks of encoder–decoder architectures such as ours is the loss of high-frequency information when reconstructing the image from the latent code z_i. A popular strategy to recover the missing information is to use *skip connections*, that forward feature maps between the encoder and the decoder, explicitly allowing to generate high frequencies. In our case, however, this strategy cannot be applied: our latent space is trained to be invariant to the attribute, so that the decoder can reconstruct the image with the target attribute value; adding skip connections would hamper this by forwarding information from the encoder to the decoder. We alleviate this problem by providing high-frequency information to the decoder through a normal map n of the object. This normal map is concatenated to the feature maps of the decoder at different scales (illustrated in red in Figure 2), allowing it to incorporate high-frequency information into the reconstruction of the target image. In the case of real images, where the normal map is not directly available, it can be obtained through a normal map predictor network (see Section 5).

Even with the use of normal map information, a single network such as $\mathcal{G}_1$ can succeed in obtaining an attribute-invariant latent code z_1, but struggles when generating a detailed reconstructed image: image $\widehat{\mathcal{I}_{b,1}}$ in Figure 2 has the desired appearance, but lacks fine detail. We therefore use $\mathcal{G}_1$ not to produce the final result, but as a means to generate a series of feature maps that encode a representation of the edited image with the target appearance. These feature maps will be used by the second network, $\mathcal{G}_2$, a shallow network capable of reconstructing high-frequency details. More precisely, we use the three last feature maps from $\mathcal{G}_1$, which include information at multiple scales, and concatenate them to the feature maps of $\mathcal{G}_2$ (as illustrated by the green vertical arrows in Figure 2). In this way, $\mathcal{G}_2$ is able to provide the output image $\widehat{\mathcal{I}_{b,2}} = \widehat{\mathcal{I}_b}$, which features the desired appearance specified by the target attribute value b while preserving the relevant high-frequency information of the input. As we will show in Section 5.1, the latent space of $\mathcal{G}_2$ alone has too much information from the input image $\mathcal{I}$ to allow for manipulation of the desired attribute.

As explained, we need to train the latent spaces from $\mathcal{G}_1$ and $\mathcal{G}_2$ to be invariant to the attribute of interest, while learning to generate a realistic target image $\widehat{\mathcal{I}_b}$. To do this, *during training*, we use three auxiliary networks. Two latent discriminators ($\mathcal{LD}_i$ in Figure 2) push the latent spaces z_i to not contain information on the attribute, while an attribute predictor and discriminator $\mathcal{C}/\mathcal{D}$, trained in an adversarial manner, guides the network towards generating a realistic image with the target attribute value b. The next subsection explains the training process and objectives.

3.3. Loss functions and training scheme

Image reconstruction loss. The first goal of each encoder–decoder network $\mathcal{G}_i$ (for clarity, we will use $\mathcal{G}$ instead of $\mathcal{G}_i$ hereafter) is to reconstruct the input image $\mathcal{I}$ when given the ground-truth perceptual attribute value a, and the normal map n. We use the L_1 loss between pixels as a measure of error, and define the reconstruction loss as:

$$\mathcal{L}_{rec}(\mathcal{G}) = \|\mathcal{I} - \mathcal{G}(\mathcal{I}, n, a)\|_1. \tag{1}$$

Attribute-invariant latent space loss. In order to force the decoder to exploit the target attribute b, we draw inspiration from FaderNet [LZU*17], and push the encoder to produce a latent space that does not contain information about the attribute. This is achieved with an adversarial training on the latent space, for which a latent discriminator $\mathcal{LD}$ is introduced. The goal of $\mathcal{LD}$ is to predict the ground truth attribute value a from the latent code z,

$$\mathcal{L}_{lat}(\mathcal{LD}) = \|a - \mathcal{LD}(z)\|_1, \tag{2}$$

while the goal of $\mathcal{G}$ is to prevent $\mathcal{LD}$ from being able to predict a from z:

$$\mathcal{L}_{lat}(\mathcal{G}) = -\|a - \mathcal{LD}(z)\|_1. \tag{3}$$

This adversarial training effectively pushes the encoder to generate an attribute-invariant latent space z, thus forcing the decoder to use the ground-truth attribute a to reach a good reconstruction.

Attribute predictor and discriminator losses. Until this point, the model has no feedback on its ability to edit images, since the target attribute is the ground-truth attribute value of the input image, $b = a$ (recall that the training data lacks original-edited image pairs). Therefore, in order to provide additional feedback to the model regarding the edited image, we introduce an attribute predictor $\mathcal{C}$. This predictor is trained to predict the attribute value of an image, using the following loss:

$$\mathcal{L}_{attr}(\mathcal{C}) = \|a - \mathcal{C}(\mathcal{I})\|_1. \tag{4}$$

Meanwhile, the network $\mathcal{G}$ is trained so that the attribute value of the edited image is correctly predicted by $\mathcal{C}$, using:

$$\mathcal{L}_{attr}(\mathcal{G}) = \|b - \mathcal{C}(\mathcal{G}(\mathcal{I}, n, b))\|_1. \tag{5}$$

However, trying to satisfy the attribute predictor can lead $\mathcal{G}$ to the generation of unrealistic artifacts in the reconstructed image. Thus, to additionally push the network to generate images that feature the same distribution as the original input data, we introduce a GAN loss together with an image discriminator $\mathcal{D}$. In particular, we use the losses from WGAN-GP [GAA*17] on both networks $\mathcal{G}$ and $\mathcal{D}$, $\mathcal{L}_{adv}(\mathcal{G})$ and $\mathcal{L}_{adv}(\mathcal{D})$ (the complete formulation can be found in the supplementary material).

Final loss functions. $\mathcal{G}_1$ is trained jointly with its latent discriminator $\mathcal{LD}_1$, by using the losses $\mathcal{L}_{lat}(\mathcal{G}_1)$ and $\mathcal{L}_{rec}(\mathcal{G}_1)$. We do not include the attribute predictor and discriminator module because $\mathcal{G}_1$ is intended to create a compact and editable latent space, rather than a high-quality output image. The resulting loss functions are:

$$\mathcal{L}(G_1) = \lambda^{\mathcal{G}}_{rec}\mathcal{L}_{rec}(\mathcal{G}_1) + \lambda^{\mathcal{G}}_{lat}\mathcal{L}_{lat}(\mathcal{G}_1), \tag{6}$$

$$\mathcal{L}(\mathcal{LD}_1) = \lambda^{\mathcal{LD}}_{lat}\mathcal{L}_{lat}(\mathcal{LD}_1). \tag{7}$$

$\mathcal{G}_2$ is trained jointly with its latent discriminator $\mathcal{LD}_2$, as well as the attribute predictor and discriminator module $\mathcal{C}/\mathcal{D}$. The resulting loss functions are:

$$\begin{aligned}\mathcal{L}(G_2) = {}& \lambda^{\mathcal{G}}_{rec}\mathcal{L}_{rec}(\mathcal{G}_2) + \lambda^{\mathcal{G}}_{lat}\mathcal{L}_{lat}(\mathcal{G}_2) \\ & + \lambda^{\mathcal{G}}_{adv}\mathcal{L}_{adv}(\mathcal{G}_2) + \lambda^{\mathcal{G}}_{attr}\mathcal{L}_{attr}(\mathcal{G}_2),\end{aligned} \tag{8}$$

$$\mathcal{L}(\mathcal{LD}_2) = \lambda^{\mathcal{LD}}_{lat}\mathcal{L}_{lat}(\mathcal{LD}_2), \tag{9}$$

$$\mathcal{L}(\mathcal{C}/\mathcal{D}) = \lambda^{\mathcal{D}}_{adv}\mathcal{L}_{adv}(\mathcal{D}) + \lambda^{\mathcal{C}}_{attr}\mathcal{L}_{attr}(\mathcal{C}). \tag{10}$$

In practice, $\mathcal{C}$ and $\mathcal{D}$ share the same convolutions and are trained as a unique network, thus the joint loss in Equation (10).

Training details. We optimize all losses using the Adam optimizer [KB14] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. To train the generators, we use a learning rate of 10^{-4}. $\mathcal{G}_1$ is trained with the following loss weights: $\lambda^{\mathcal{G}}_{rec} = 1$, $\lambda^{\mathcal{G}}_{lat} = 5$, while $\mathcal{G}_2$ is trained with $\lambda^{\mathcal{G}}_{rec} = 1$, $\lambda^{\mathcal{G}}_{lat} = 2.5$, $\lambda^{\mathcal{G}}_{adv} = 0.02$ and $\lambda^{\mathcal{G}}_{attr} = 2$. Both latent discriminators are optimized with a learning rate of $2.5 \cdot 10^{-5}$ for 12 iterations for every iteration on the generator. $\mathcal{C}/\mathcal{D}$ is optimized with

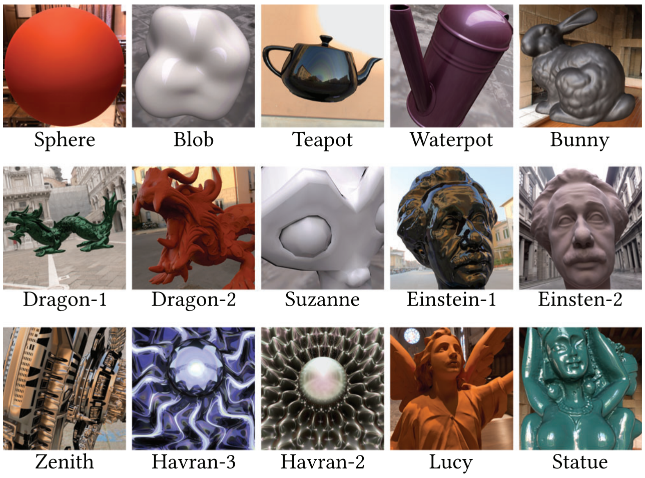

Figure 3: *Representative samples of the image dataset used for training. The images show each of the 15 scenes in the dataset (13 distinct geometries, two of them with two different viewpoints, for a total of 15 scenes), featuring different materials and illuminations.*

a learning rate of 10^{-4} for seven iterations for every iteration on the generator with loss weights $\lambda^{\mathcal{D}}_{adv} = 1$ and $\lambda^{\mathcal{C}}_{attr} = 3$. Our model is trained individually for each attribute. We first train $\mathcal{G}_1$ for 300 epochs, then train $\mathcal{G}_2$ for 50 epochs, freezing parameters for $\mathcal{G}_1$. We implemented our models using the Pytorch framework [PGM*19] and trained them using a Nvidia 2080Ti GPU. In total, training our framework took 2 days per attribute. Before feeding the images to the network, we fill the background with black colour and add the mask of the object as a fourth channel.

4. Training Dataset

Training our model to edit a certain attribute of material appearance requires images with realistic depictions of materials, on objects with different shapes and a variety of illuminations. For each of these images, we require the corresponding value for the attribute of interest. Since we are targeting high-level perceptual attributes of material appearance, this value needs to be obtained from subjective data gathered through subject responses. These image-attribute $(\mathcal{I}, a)$ pairs are used to train our network towards correctly interpreting such attributes.

Image data. We leverage the recent dataset by Lagunas *et al.* [LMS*19], designed specifically for learning tasks related to material appearance. It is composed of realistic renderings of 13 geometries of varied complexity (with two additional viewpoints, leading to 15 different scenes), illuminated with six captured environment maps [Deb]. The objects are rendered with 100 measured BRDFs from the MERL dataset [MPBM03], using the physically based renderer Mitsuba [Jak]. The dataset comprises a total of 9000 renderings, of which representative samples are shown in Figure 3.

Subjective attributes. The image dataset we use [LMS*19] includes associated subjective data, but in the form of similarity judgements between pairs of images, unsuitable for our goal. Other datasets include subjective measures of high-level perceptual attributes of material appearance for the materials in the MERL dataset, but for a single shape and illumination [SGM*16]. Since shape and illumination play an important role in the perception of material appearance [LSGM21, VLD07, NS98], we set out to gather our own subjective data of high-level perceptual attributes for the Lagunas *et al.* image dataset.

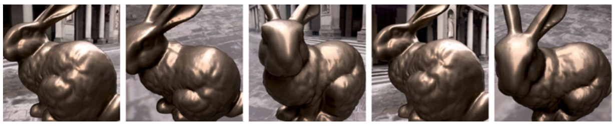

Figure 4: *Example of the five viewpoints used in the perceptual study on the* bunny *shape rendered with the* Uffizi *illumination and* alum–bronze *material.*

To do so, we follow the same methodology as Serrano *et al.* [SGM*16]: We carry out a perceptual experiment in which, for each image in the Lagunas *et al.* dataset, participants had to rate a number of high-level attributes on a Likert-type, 1-to-5 scale. To further increase the robustness of the obtained ratings, we augment Lagunas *et al.*'s dataset by creating, for each combination of material × shape × illumination, five different images with slight variations in the viewpoint (randomly sampled within a 45°cone around the original viewpoint). Examples of such images for the *bunny* shape are shown in Figure 4. Similar to previous large-scale studies, we relied on Amazon Mechanical Turk to collect the ratings. A total of 2600 paid subjects participated in the study, each of them seeing 15 different random images. Participants in the study had to go through a training session at the beginning of it, and control stimuli were used to reject invalid subjects. More details about the experiment can be found in the supplementary material.

Through our perceptual study we gather, for each attribute, 39,000 ratings (13 shapes × 6 illuminations × 100 materials × 5 viewpoints), leading to that number of image–attribute pairs. It is important to note that, due to the vast size of our dataset, we only gather one response per condition (per combination of material × shape × illumination × viewpoint), which can lead to variability in the data that may hinder the convergence of the training. In order to reduce it, we pool the perceptual ratings over viewpoint and shape by means of the median, more robust to outliers than the mean.

5. Results and Evaluation

In this section, we start by introducing our evaluation dataset, and showing results of our framework by applying it to two perceptual attributes: *Glossy* and *Metallic*. We then validate our design choices through a series of ablation studies (Section 5.1), and analyse the consistency of our editing across controlled geometry, illumination and material variations (Section 5.2). In addition, we perform user studies to assess whether our edits of the attributes do correlate with human perception (Section 5.3) and to evaluate the perceived quality of our images (Section 5.4). Finally, we compare our results with actual renderings of a modified scene (Section 5.5).

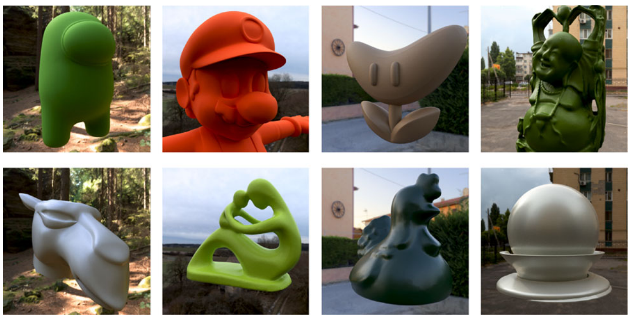

Figure 5: *Representative images of our synthetic evaluation dataset, showing the eight shapes and materials used in it. Each column is rendered with one of the four illuminations used.*

Evaluation data. Our evaluation data are composed of both synthetic images and real photographs. The *synthetic images* evaluation dataset is composed of images never seen during training by our framework. They are rendered using eight shapes collected from free online sources, four illuminations obtained from *HDRI-Haven* [HDR], and eight materials coming from Dupuy and Jakob's database [DJ18]. A representative subset is shown in Figure 5.

We collected *real images* for our evaluation dataset by browsing online catalogues of decorative items, as well as photographing objects ourselves in uncontrolled setups. Within each image, we masked the object of interest using an online API [Kal]. Since our framework requires a normal map, which is not directly available when using real photographs, we obtain the normal maps for these objects by using a *normal map predictor*. Inspired by image-to-image generative networks, we trained a new model to infer normal maps directly from the single-view RGB images. Our normal map predictor consists of a modified Pix2Pix network[IZZE17]. We carefully designed our architecture and losses to minimize convolution artifacts, high variance noise in the resulting normals, and maintain as much geometrical detail from the original images as possible, while reducing the influence of varying reflectance and illumination conditions. The model was trained on synthetic data coupled with ground-truth normal maps. Additional details about the architecture and losses used to train the normal predictor can be found in the supplementary material. Representative examples of our real evaluation dataset, together with their predicted normal maps, can be seen in Figure 6.

Results. Figure 1 shows editing results for a variety of real-world objects photographed in uncontrolled setups under different conditions, for our two attributes *Glossy* and *Metallic*. They include indoor and outdoor scenarios, varied shape complexity and different types of materials, yet our framework can handle them gracefully, producing compelling edits by just changing the high-level perceptual attribute. It is interesting to observe how, even though the illumination is not explicitly modelled during training, the edits seem to plausibly capture the lighting in the scene. Additionally, our framework is trained so that the attribute of interest can be sampled along its range, producing consistent results. This is shown in Figure 7 for two real images, where both attributes exhibit a coherent variation (see the supplementary material for additional results). Figures 1 and 7 also show that our normal map predictor is capable of yielding a normal map that allows for realistic editing of photographs.

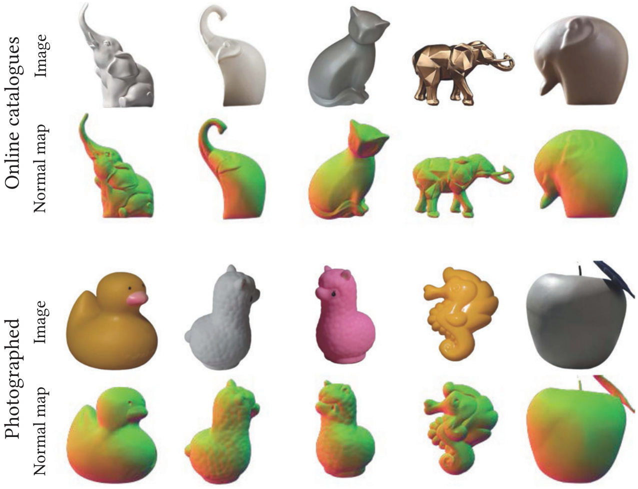

Figure 6: *Representative examples of our real images evaluation dataset, comprised of photos from online catalogues (top), and casually photographed objects (bottom). For each image, we also show its normal map, as obtained by our normal map predictor.*

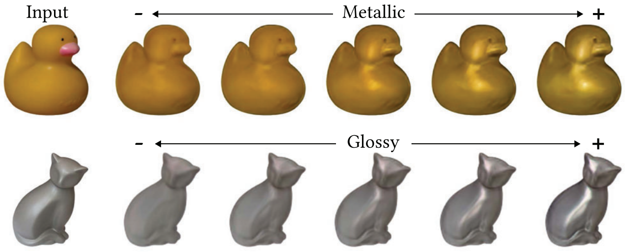

Figure 7: *Editing results by varying the perceptual attributes* Metallic *and* Glossy. *First column is the input image, following ones show the edited image when sampling the attribute as* $[-1,\ 0,\ 0.5,\ 0.75,\ 1]$ *for* Metallic *and* $[-1,\ -0.25,\ 0,\ 0.25,\ 1]$ *for* Glossy. *Our method produces a realistic editing of the input over the whole range.*

5.1. Ablation studies

We evaluate the utility of each of the components of our method through a series of ablation studies where the *Metallic* attribute is used. We generate five ablated versions of our framework, for which we show an illustrative result in Figure 8. First, the effect of the individual generative networks is shown in *Only* $\mathcal{G}_1$ and *Only* $\mathcal{G}_2$. When using only $\mathcal{G}_1$, the resulting image features the desired edit, but lacks high-frequency details. Meanwhile, $\mathcal{G}_2$ alone is able to reconstruct the fine detail of the input image, but cannot convincingly edit the appearance towards the target increased metallicity. We then investigate the effect of the auxiliary networks. When the latent discriminator $\mathcal{LD}_2$ is removed (*W/o* $\mathcal{LD}_2$), the generated image struggles to convey the appearance required by the

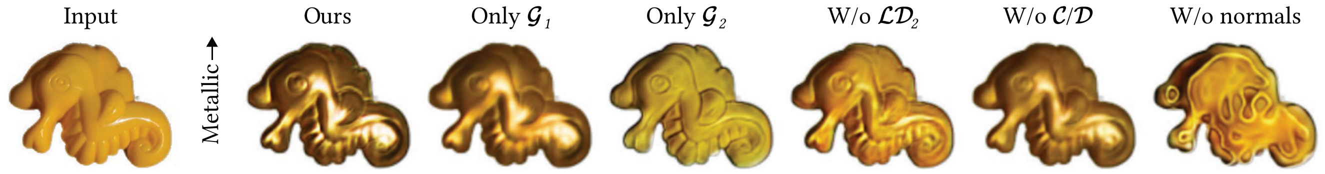

Figure 8: *Ablation studies where we trained and tested out each of the individual components of our framework. The leftmost image shows the input photograph, followed by the target attribute (*Metallic $+1$*). Then, from left to right: the resulting edited image using our method, only the $\mathcal{G}_1$ network, only the $\mathcal{G}_2$ network, training without the* latent discriminator $\mathcal{LD}_2$ *and its associated loss function, training without the* attribute predictor and discriminator $\mathcal{C}/\mathcal{D}$ *and its associated loss function, and training without using the normal map information of the input image. Our method qualitatively yields superior performance and allows for the creation of sharp highlights and realistic images.*

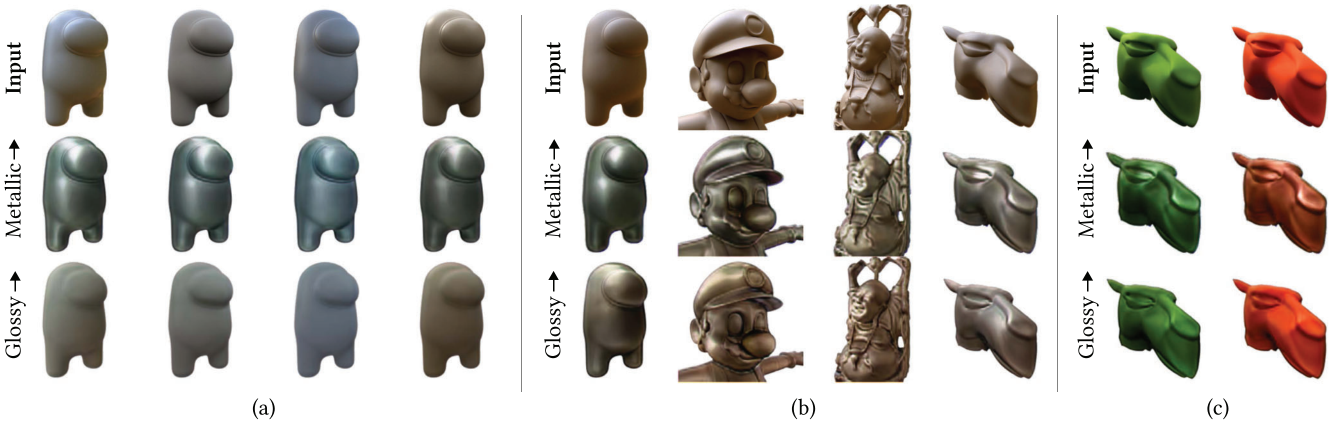

Figure 9: *Example results illustrating the consistency of our editing framework. (a) Same object and material, but different illuminations in the input image; (b) same illumination and material, but different geometry ; (c) same geometry and illumination, but different materials with similar reflectance properties. Our framework is capable of producing compelling and consistent edits in all cases. Arrows pointing up correspond to a target attribute value of $+1$, while arrows pointing down correspond to a value of -1.*

target edit. Additionally, without the attribute predictor and discriminator ($W/oC/D$), the framework is only slightly able to improve the edited result from the first network $\mathcal{G}_1$. Finally, we investigate the effect of the normal map information by removing them from the training (*W/o normals*). Without this information, the framework cannot reconstruct the geometry, leading to unrealistic results.

5.2. Consistency of the edits

We use our synthetic evaluation dataset to assess the consistency of our edits under different conditions. Figure 9(a) shows edits performed when both material and geometry are the same in the input image, and only the illumination changes. Our material edits are perceptually consistent, while illumination properties are preserved within the edits. Figure 9(b) shows results when only the geometry changes in the input images. Our edits yield consistent results across geometries, appearing to be all made of a similar material (within each row). Last, in Figure 9(c), we evaluate the consistency of our edits using two different materials with similar reflectance properties, namely *acrylic-felt-orange* and *acrylic-felt-green*. Again our framework yields consistent, plausible results for both attributes.

5.3. User study

We run an additional user study to assess the perception of the appearance in our edited images. In the study, participants were asked to rate the perceptual attribute in generated images in which such attribute had been edited with our framework. While we include here the main aspects, more details on the study, including the full set of stimuli used, can be found in the supplementary material.

Stimuli. We selected three images for each attribute (*Glossy* and *Metallic*), varied in shape, illumination and material, and edited them with our method by setting the target attribute value to -1, 0 and $+1$. This leads to two sets (one per attribute) of nine *edited images*. We also incorporated, for each attribute, nine other images from the training dataset, chosen such that they covered the whole range of attribute values; we will term them *training images*. Note that these images are unedited, and for each, we have the 'ground truth' attribute value gathered through our perceptual study that was used to train our framework (Section 4).

Procedure. The stimuli were shown to participants in two separate blocks, one per attribute. Each block thus consists of 18 images, for which the participants had to rate the attribute on a Likert-type 1-to-5 scale.

Table 1: *Pearson correlation coefficients (along with their* p*-value) between the expected attribute of the images shown in the user study, and the answers of the participants (collected attribute).*

	Metallic	Glossy
Edited images	0.90, $p < 0.001$	0.86, $p = 0.003$
Training images	0.92, $p < 0.001$	0.96, $p < 0.001$

Fifteen participants took part in the study, leading to 15 ratings for each image and attribute.

Results. For each image, we average the participants' ratings to obtain a perceived attribute value (to which we will refer here as *collected* value). Table 1 shows the results of the Pearson correlation between the collected and the expected attribute values. Note that the expected attribute value is the target attribute value for the edited images, and the 'ground-truth' attribute value for the training images.

For both, edited and training images, there is a strong (and significant) correlation between the collected and the expected attribute values. While for the *Metallic* attribute, the correlations for edited images are on par with the ones for training images (0.90 and 0.92, respectively), correlations for the *Glossy* attribute are lower for the edited images than for the training images (0.86 and 0.96, respectively). This can be due to the fact that our edited images do not cover the full range of glossiness, with the most glossy images (with a target attribute set to $+1$) being scored between 3 and 3.7 (on a scale of 1–5). However, the correlations for the edited images remain high, showing that our edited images are globally well perceived.

5.4. Perceived quality

During the development of this work, we noticed that users had difficulty telling a real image (photograph) from our edited results in many cases (see second and third images in Figure 1 for instance). We thus run an additional user study to evaluate the perceived subjective quality of our edited results, both starting from rendered images and real photographs. We took 20 edited images covering a wide variety of appearances. Since compositing an edited image over its original background can produce visual artifacts at the border of the object, we showed the images over a neutral background. Following common practice on subjective image quality evaluation (e.g. [MTM12]), we showed each image for 3 s, then asked the participants to rate its perceived quality according to the following scale: 'bad', 'poor', 'fair', 'good' and 'excellent'. We report the results numerically, associating to this scale scores between 1 and 5.

For comparison purposes, we run a similar study with eight *actual photographs* of real objects, also on a neutral background. The two studies were completed by 10 and 11 participants, respectively. We also queried the experience in computer graphics of our participants and collected the following answers: nine 'professional', seven 'intermediate', one 'beginner' and four 'no experience'. A representative subset of images is shown in Figure 10, along with their average rating and standard deviation.

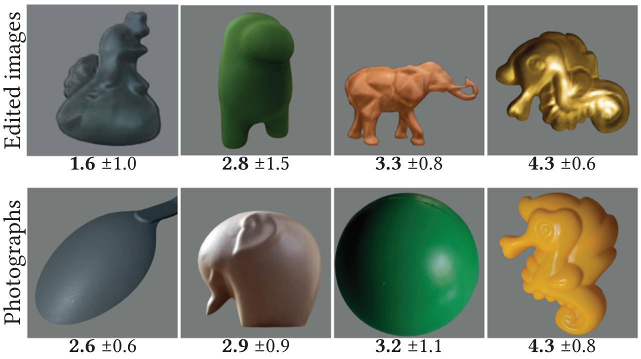

Figure 10: *Example stimuli shown in the user study to evaluate the perceived quality of the images. Top row: edited images with our method. Bottom row: photographs (not edited). For each one, we show its average rating and standard deviation.*

Our edited images were rated with an average score of 2.9 (standard deviation of 1.1) while the photographs were rated only slightly better, 3.4 in average with a standard deviation of 1.0 (see Figure 10). While this indicates that the perceived quality of our edited images is comparable to the photographs, we believe that the relative simplicity of the images (a single object on a neutral background) may have played a role in the lukewarm scores obtained in both cases (edited images and photographs). For instance, the grey spoon in Figure 10 was rated between 'poor' and 'fair' (average 2.6), despite being an actual photograph.

5.5. Comparison with BRDF editing

In addition, we compare our results on synthetic scenes to actual renderings editing a BRDF attribute. We use the method proposed by Serrano *et al*. [SGM*16] and edit the *Metallic* or *Glossy* attributes of several materials from the extended MERL dataset presented in the same work. Note that these materials were not present in our training dataset. In particular, we first render an object of material m into an image $\mathcal{I}_m$. We then manipulate a given material attribute in two different ways: rendering a new image $\mathcal{I}_{\widehat{m}}$ with the modified material attribute, and producing an edited image $\widehat{\mathcal{I}_m}$ from the original one, using our method. Figure 11 illustrates the results, making two objects less glossy and more metallic, respectively. In general, our edited results match the rendered images well when reducing glossiness (top row), while producing plausible but less accurate results the other way around (bottom row). This makes sense, since it is easier to remove existing information (highlights) in the first example than it is to deal with missing information in the second, such as the original lighting environment.

6. Discussion and Limitations

We have presented a framework to edit materials directly in images, under unknown illumination and geometric conditions, through the manipulation of high-level perceptual attributes. Our framework is based on two generative networks aiming at providing an editable latent space, and reconstructing high-frequency details, respectively. We have shown that our method produces plausible results, almost

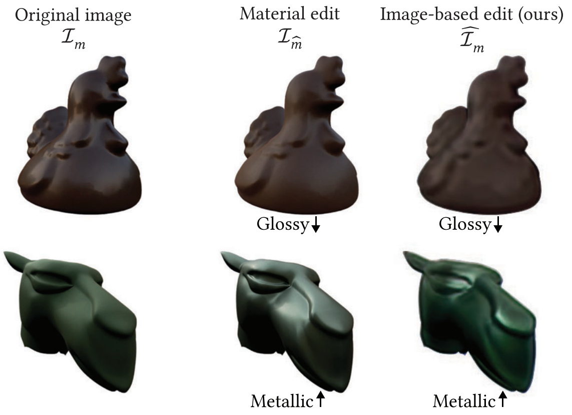

Figure 11: *Comparison with BRDF editing. Left: original rendered image. Middle: rendered result modifying a BRDF attribute (following Serrano* et al. *[SGM*16]). Right: our edited results modifying the same attribute. Our edited results match the rendered images well when reducing glossiness, while producing plausible images when increasing the presence of highlights.*

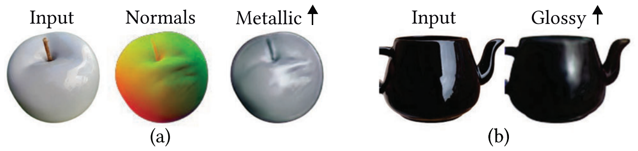

Figure 12: *Limitations of our framework: (a) noise in the prediction of the normals may lead to unpredicted editing results (left: input image, centre: inferred normal map, right: edited image with* Metallic *+1); (b) due to the lack of skip-connections, almost mirror-like reflections in the input image (left) are hard to model during editing when trying to reach high glossiness (right: edited image with* Glossy *+1).*

on par with real photographs, on an ample variety of input images (Figures 1 and 10). This was further validated through a user study.

Our framework is not free of limitations, which open up several possibilities for future work. Since no normal maps are provided for real pictures, we have introduced a normal map predictor; inaccuracies in its output may lead to distortions in the edited objects, especially visible around highlights, as shown in Figure 12(a); our framework would thus benefit from better models to infer normals. Besides, since our architecture does not allow for the use of skip-connections, high-frequency illumination details such as mirror-like reflections may also not be recovered properly when trying to reach high glossiness values, as shown in Figure 12(b). Similarly, our framework can only create fuzzy highlights when presented with an input image depicting a diffuse material that conveys only limited information about the illumination.

It would be interesting to combine our approach with recent neural rendering techniques, which can create such information about the illumination [TZN19, LSR*20].

Our framework was trained using the dataset by Lagunas *et al.* [LMS*19], which contains synthetic data using the isotropic BRDFs from MERL [MPBM03]. However, MERL materials are biased in terms of albedo and reflectance. To mitigate this, we have augmented our input data with changes in hue before feeding it to our framework (see the supplementary material for more information). Nevertheless, designing a dataset beyond isotropic BRDFs could allow the framework to edit a wider range of appearances. Moreover, since our dataset contains single-colour objects, we currently cannot edit spatially varying reflectance (such as the duck's beak in Figure 7).

We hope that our work inspires additional research and novel perceptually based applications. We will make our data and code available for further experimentation, in order to facilitate the exploration of these possibilities.

Acknowledgements

This work was funded by the European Union's Horizon 2020 research and innovation program, through the CHAMELEON project (European Research Council, Grant Agreement No. 682080), the DyViTo project (MSCA-ITN, Grant Agreement No. 765121) and the PRIME project (MSCA-ITN, Grant Agreement No. 956585).

References

[BAEDR08] Ben-Artzi A., Egan K., Durand F., Ramamoorthi R.: A precomputed polynomial representation for interactive BRDF editing with global illumination. *ACM Transactions on Graphics 27*, 2 (2008), 1–13.

[BBPA15] Boyadzhiev I., Bala K., Paris S., Adelson E.: Band-sifting decomposition for image-based material editing. *ACM Transactions on Graphics 34*, 5 (2015), 1–16.

[BM14] Barron J. T., Malik J.: Shape, illumination, and reflectance from shading. *IEEE Transactions on Pattern Analysis and Machine Intelligence 37*, 8 (2014), 1670–1687.

[BP20] Bieron J., Peers P.: An adaptive BRDF fitting metric. *Computer Graphics Forum 39*, (2020), 59–74.

[BPV18] Barla P., Pacanowski R., Vangorp P.: A composite BRDF model for hazy gloss. *Computer Graphics Forum 37*, (2018), 55–66.

[BSH12] Bagher M. M., Soler C., Holzschuch N.: Accurate fitting of measured reflectances using a shifted gamma microfacet distribution. *Computer Graphics Forum 31*, (2012), 1509–1518.

[BSM*18] Beigpour S., Shekhar S., Mansouryar M., Myszkowski K., Seidel H.-P.: Light-field appearance editing based on intrinsic decomposition. *Journal of Perceptual Imaging 1*, 1 (2018), 10502-1–10502-15.

[CDD21] Cavdan M., Drewing K., Doerschner K.: Materials in action: The look and feel of soft. *Journal of Vision 20*, 11 (2020), 514–514.

[CK15] Chadwick A., Kentridge R.: The perception of gloss: A review. *Vision Research 109* (2015), 221–235.

[CUYH20] Choi Y., Uh Y., Yoo J., Ha J.-W.: StarGAN v2: Diverse image synthesis for multiple domains. In *Proceedings of the Computer Vision and Pattern Recognition* (2020), pp. 8188–8197.

[DCWP19] Di Cicco F., Wijntjes M. W., Pont S. C.: Understanding gloss perception through the lens of art: Combining perception, image analysis, and painting recipes of 17th century painted grapes. *Journal of Vision 19*, 3 (2019), 7.

[Deb] Debevec P.: (2008), accessed on 10/09/2020 https://vgl.ict.usc.edu/Data/HighResProbes/.

[DFY*11] Doerschner K., Fleming R. W., Yilmaz O., Schrater P. R., Hartung B., Kersten D.: Visual motion and the perception of surface material. *Current Biology 21*, 23 (2011), 2010–2016.

[DJ18] Dupuy J., Jakob W.: An adaptive parameterization for efficient material acquisition and rendering. *ACM Transactions on Graphics 37*, 6 (2018), 1–14.

[DLG*20] Delanoy J., Lagunas M., Galve I., Gutierrez D., Serrano A., Fleming R., Masia B.: The role of objective and subjective measures in material similarity learning. In *Proceedings of the ACM SIGGRAPH Posters* (2020).

[DRCP14] Di Renzo F., Calabrese C., Pellacini F.: AppIm: Linear spaces for image-based appearance editing. *ACM Transactions on Graphics (TOG) 33*, 6 (2014), 1–9.

[DSMG21] Delanoy J., Serrano A., Masia B., Gutierrez D.: Perception of material appearance: A comparison between painted and rendered images. *Journal of Vision 21*, 5 (May 2021), 16.

[FDA01] Fleming R. W., Dror R. O. & Adelson E. H.: How do humans determine reflectance properties under unknown illumination? In *CVPR 2001 Workshop on Identifying Objects Across Variations in Lighting: Psychophysics and Computation* (2001) pp. 1–8.

[FDA03] Fleming R. W., Dror R. O., Adelson E. H.: Real-world illumination and the perception of surface reflectance properties. *Journal of Vision 3*, 5 (2003), 3.

[FPG01] Ferwerda J. A., Pellacini F., Greenberg D. P.: Psychophysically based model of surface gloss perception. In *Proceedings of the Human Vision and Electronic Imaging VI* (2001), vol. 4299, pp. 291–302.

[FS19] Fleming R. W., Storrs K. R.: Learning to see stuff. *Current Opinion in Behavioral Sciences 30* (2019), 100–108.

[FWG13] Fleming R. W., Wiebel C., Gegenfurtner K.: Perceptual qualities and material classes. *Journal of Vision 13*, 8 (2013), 9.

[GAA*17] Gulrajani I., Ahmed F., Arjovsky M., Dumoulin V. & Courville A. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems* (2017), vol. 30, Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/892c3b1c6dccd52936e27cbd0ff683d6-Paper.pdf.

[GLD*19] Gao D., Li X., Dong Y., Peers P., Xu K., Tong X.: Deep inverse rendering for high-resolution SVBRDF estimation from an arbitrary number of images. *ACM Transactions on Graphics (TOG) 38*, 4 (2019), 1–15.

[GMLMG12] Garces E., Munoz A., Lopez-Moreno J., Gutierrez D.: Intrinsic images by clustering. *Computer Graphics Forum 31*, (2012), 1415–1424.

[GSLM*08] Gutierrez D., Seron F. J., Lopez-Moreno J., Sanchez M. P., Fandos J., Reinhard E.: Depicting procedural caustics in single images. *ACM Transactions on Graphics 27*, 5 (2008), 1–9.

[GWA*15] Gkioulekas I., Walter B., Adelson E. H., Bala K., Zickler T.: On the appearance of translucent edges. In *Proceedings of the Computer Vision and Pattern Recognition* (2015), pp. 5528–5536.

[GXZ*13] Gkioulekas I., Xiao B., Zhao S., Adelson E. H., Zickler T., Bala K.: Understanding the role of phase function in translucent appearance. *ACM Transactions on Graphics 32*, 5 (2013), 147.

[HDR] HDRIhaven: https://www.hdrihaven.com/. Accessed on 01/02/2021.

[HFB*09] Haber T., Fuchs C., Bekaer P., Seidel H.-P., Goesele M., Lensch H. P.: Relighting objects from image collections. In *Proceedings of the Computer Vision and Pattern Recognition* (2009), pp. 627–634.

[HFM16] Havran V., Filip J., Myszkowski K.: Perceptually motivated BRDF comparison using single image. *Computer Graphics Forum 35*, (2016), 1–12.

[HLM06] Ho Y.-X., Landy M. S., Maloney L. T.: How direction of illumination affects visually perceived surface roughness. *Journal of Vision 6*, 5 (2006), 8.

[IZZE17] Isola P., Zhu J.-Y., Zhou T., Efros A. A.: Image-to-image translation with conditional adversarial networks. In *Proceedings of the Computer Vision and Pattern Recognition* (July 2017).

[Jak10] Jakob W.: Mitsuba renderer. (2010). http://www.mitsuba-renderer.org, accessed on 01/02/2021

[JMB*14] Jarabo A., Masia B., Bousseau A., Pellacini F., Gutierrez D.: How do people edit light fields. *ACM Transactions on Graphics 33*, 4 (2014), 4.

[Kal] Kalideo: (2018), Remove.bg. https://www.remove.bg, accessed on 1/03/2021

[KB14] Kingma D. P. & Ba J.: Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations*, Y. Bengio and Y. LeCun (Eds.). ICLR 2015 (2015), San Diego, CA, USA, Conference Track Proceedings. http://arxiv.org/abs/1412.6980

[KFB10] Křivánek J., Ferwerda J. A., Bala K.: Effects of global illumination approximations on material appearance. *ACM Transactions on Graphics (Proc. SIGGRAPH) 29*, 4 (2010), 112:1–112:10.

[KP10] Kerr W. B., Pellacini F.: Toward evaluating material design interface paradigms for novice users. *ACM Transactions on Graphics 29*, (2010), 35.

[KRFB06] Khan E. A., Reinhard E., Fleming R. W., Bülthoff H. H.: Image-based material editing. *ACM Transactions on Graphics 25*, 3 (2006), 654–663.

[KWKT15] Kulkarni T. D., Whitney W., Kohli P. & Tenenbaum J. B.: Deep convolutional inverse graphics network. In *Advances in Neural Information Processing Systems*. C. Cortes, N. Lawrence, D. Lee, M. Sugiyama and R. Garnett (Eds.). (2015), vol. 28, Curran Associates, Inc.

[LBAD*06] Lawrence J., Ben-Artzi A., DeCoro C., Matusik W., Pfister H., Ramamoorthi R., Rusinkiewicz S.: Inverse shade trees for non-parametric material representation and editing. *ACM Transactions on Graphics 25*, 3 (2006), 735–745.

[LCY*17] Liu G., Ceylan D., Yumer E., Yang J., Lien J.-M.: Material editing using a physically based rendering network. In *Proceedings of the International Conference on Computer Vision* (October 2017).

[LDX*19] Liu M., Ding Y., Xia M., Liu X., Ding E., Zuo W., Wen S.: STGAN: A unified selective transfer network for arbitrary image attribute editing. In *Proceedings of the Computer Vision and Pattern Recognition* (2019).

[LMF*19] LeGendre C., Ma W.-C., Fyffe G., Flynn J., Charbonnel L., Busch J., Debevec P.: Deeplight: Learning illumination for unconstrained mobile mixed reality. In *Proceedings of the Computer Vision and Pattern Recognition* (2019), pp. 5918–5928.

[LMS*19] Lagunas M., Malpica S., Serrano A., Garces E., Gutierrez D., Masia B.: A similarity measure for material appearance. *ACM Transactions on Graphics (Proc. SIGGRAPH) 38*, 4 (2019), 1–12.

[LSGM21] Lagunas M., Serrano A., Gutierrez D., Masia B.: The joint role of geometry and illumination on material recognition. *Journal of Vision 21* (2021), 2.

[LSR*20] Li Z., Shafiei M., Ramamoorthi R., Sunkavalli K., Chandraker M.: Inverse rendering for complex indoor scenes: Shape, spatially-varying lighting and SVBRDF from a single image. In *Proceedings of the Computer Vision and Pattern Recognition* (2020), pp. 2475–2484.

[LZU*17] Lample G., Zeghidour N., Usunier N., Bordes A., Denoyer L., Ranzato M. A.: Fader networks: Manipulating images by sliding attributes. In *Proceedings of the Advances in Neural Information Processing Systems* (2017), vol. 30.

[MGZ*17] Mylo M., Giesel M., Zaidi Q., Hullin M., Klein R.: Appearance bending: A perceptual editing paradigm for data-driven material models. In *Proceedings of the Vision, Modeling and Visualization* (2017), The Eurographics Association.

[MLMG19] Mao R., Lagunas M., Masia B., Gutierrez D.: The effect of motion on the perception of material appearance. In *Proceedings of the ACM Symposium on Applied Perception* (2019), pp. 1–9.

[MLTFR19] Maximov M., Leal-Taixé L., Fritz M., Ritschel T.: Deep appearance maps. In *Proceedings of the International Conference on Computer Vision* (2019), pp. 8729–8738.

[MPBM03] Matusik W., Pfister H., Brand M., McMillan L.: A data-driven reflectance model. *ACM Transactions on Graphics 22*, 3 (2003), 759–769.

[MTM12] Mantiuk R. K., Tomaszewska A., Mantiuk R.: Comparison of four subjective methods for image quality assessment. In *Computer Graphics Forum 31*, (2012), 2478–2491.

[NS98] Nishida S., Shinya M.: Use of image-based information in judgments of surface-reflectance properties. *Journal of the Optical Society of America A 15*, 12 (1998), 2951–2965.

[NSRS13] Nguyen C. H., Scherzer D., Ritschel T., Seidel H.-P.: Material editing in complex scenes by surface light field manipulation and reflectance optimization. *Computer Graphics Forum 32*, (2013), 185–194.

[OKT*19] Ono T., Kubo H., Tanaka K., Funatomi T., Mukaigawa Y.: Practical BRDF reconstruction using reliable geometric regions from multi-view stereo. *Computational Visual Media 5*, 4 (2019), 325–336.

[PFG00] Pellacini F., Ferwerda J. A., Greenberg D. P.: Toward a psychophysically-based light reflection model for image synthesis. In *Proceedings of the Conference on Computer Graphics and Interactive Techniques* (2000), pp. 55–64.

[PGM*19] Paszke A., Gross S., Massa F., Lerer A., Bradbury J., Chanan G., Killeen T., Lin Z., Gimelshein N. & Antiga L. et al.: Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*. H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox and R. Garnett (Eds.). (2019), vol. 32, Curran Associates, Inc. https://proceedings.neurips.cc/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf.

[RRF*16] Rematas K., Ritschel T., Fritz M., Gavves E., Tuytelaars T.: Deep reflectance maps. In *Proceedings of the Computer Vision and Pattern Recognition* (2016), pp. 4508–4516.

[SAF21] Storrs K. R., Anderson B. L., Fleming R. W.: Unsupervised learning predicts human perception and misperception of gloss. *Nature Human Behaviour 5*, (2021), 1–16.

[SFV20] Schmidt F., Fleming R. W., Valsecchi M.: Softness and weight from shape: Material properties inferred from local shape features. *Journal of Vision 20*, 6 (2020), 2.

[SGM*16] Serrano A., Gutierrez D., Myszkowski K., Seidel H.-P., Masia B.: An intuitive control space for material appearance. *ACM Transactions on Graphics 35*, 6 (November 2016), 186:1–186:12.

[SJR18] Sun T., Jensen H. W., Ramamoorthi R.: Connecting measured BRDFs to analytic BRDFs by data-driven diffuse-specular separation. *ACM Transactions on Graphics 37*, 6 (2018), 1–15.

[SPN*16] Schmidt T.-W., Pellacini F., Nowrouzezahrai D., Jarosz W., Dachsbacher C.: State of the art in artistic editing of appearance, lighting and material. *Computer Graphics Forum 35*, (2016), 216–233.

[TFCRS11] Thompson W., Fleming R., Creem-Regehr S., Stefanucci J. K.: *Visual Perception from a Computer Graphics Perspective* (1st edition). Natick, Massachusetts, United States: A. K. Peters, Ltd., 2011.

[TZN19] Thies J., Zollhöfer M., Nießner M.: Deferred neural rendering: Image synthesis using neural textures. *ACM Transactions on Graphics 38*, 4 (2019), 1–12.

[VBF17] Vangorp P., Barla P., Fleming R. W.: The perception of hazy gloss. *Journal of Vision 17*, 5 (2017), 19.

[VLD07] Vangorp P., Laurijssen J., Dutré P.: The influence of shape on the perception of material reflectance. *ACM Transactions on Graphics 26*, 3 (July 2007).77–es.

[WAKB09] Wills J., Agarwal S., Kriegman D., Belongie S.: Toward a perceptual space for gloss. *ACM Transactions on Graphics 28*, 4 (September 2009), 103:1–103:15.

[XWT*08] Xuey S., Wang J., Tong X., Dai Q., Guo B.: Image-based material weathering. *Computer Graphics Forum 27*, (2008), 617–626.

[XZG*20] Xiao B., Zhao S., Gkioulekas I., Bi W., Bala K.: Effect of geometric sharpness on translucent material perception. *Journal of Vision 20*, 7 (2020), 10.

[YS19] Yu Y., Smith W. A.: Inverserendernet: Learning single image inverse rendering. In *Proceedings of the Computer Vision and Pattern Recognition* (2019), pp. 3155–3164.

[ZFWW20] Zsolnai-Fehér K., Wonka P., Wimmer M.: Photorealistic material editing through direct image manipulation. *Computer Graphics Forum 39*, (2020), 107–120.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Figure 1: Screenshot of the perceptual study as seen by participants

Figure 2: Left: the two images used in our training session. Right: the four images used as controls

Figure 3: Input images and edited stimuli used in our user-study

Figure 4: Answers collected in our validation study for both attribute Metallic and Glossy and for the two sets of images

Figure 5: Average scores of quality collected in the user study over the 20 edited images (top) and the 8 real photographs (bottom)