GAN2X: Non-Lambertian Inverse Rendering of Image GANs

Xingang Pan[1]   Ayush Tewari[1,2]   Lingjie Liu[1]   Christian Theobalt[1]

[1]Max Planck Institute for Informatics   [2]MIT

{xpan,lliu,theobalt}@mpi-inf.mpg.de   ayusht@mit.edu

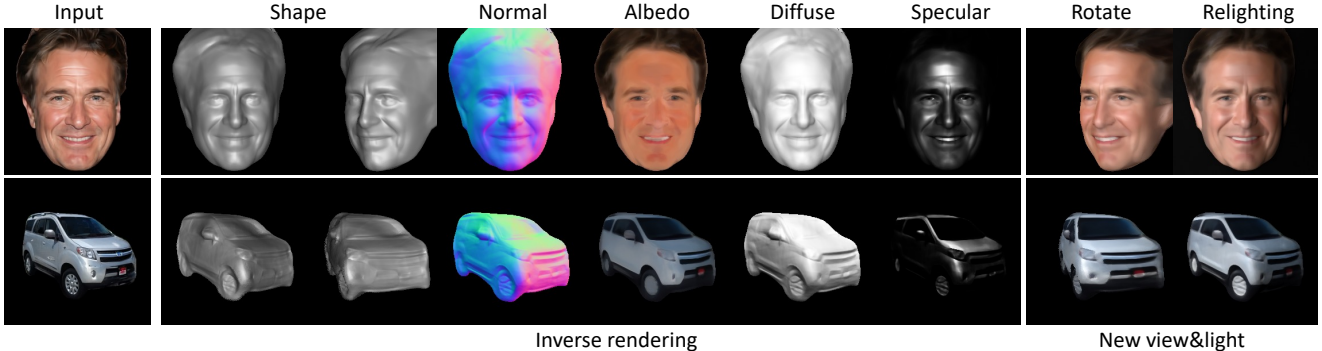Project page: https://vcai.mpi-inf.mpg.de/projects/GAN2X/

Figure 1: We study non-Lambertian inverse rendering of image GANs. For an image generated by a GAN pretrained on an unpaired image collection, we leverage its corresponding pseudo paired images to recover the underlying object shape, albedo, specular properties, and lighting. This allows to re-render the object under a novel viewpoint or lighting condition.

Abstract

2D images are observations of the 3D physical world depicted with the geometry, material, and illumination components. Recovering these underlying intrinsic components from 2D images, also known as inverse rendering, usually requires a supervised setting with paired images collected from multiple viewpoints and lighting conditions, which is resource-demanding. In this work, we present GAN2X, a new method for unsupervised inverse rendering that only uses unpaired images for training. Unlike previous Shape-from-GAN approaches that mainly focus on 3D shapes, we take the first attempt to also recover non-Lambertian material properties by exploiting the pseudo paired data generated by a GAN. To achieve precise inverse rendering, we devise a specularity-aware neural surface representation that continuously models the geometry and material properties. A shading-based refinement technique is adopted to further distill information in the target image and recover more fine details. Experiments demonstrate that GAN2X can accurately decompose 2D images to 3D shape, albedo, and specular properties for different object categories, and achieves state-of-the-art performance for unsupervised single-view 3D face reconstruction. We also show its applications in downstream tasks including real image editing and lifting 2D GANs to decomposed 3D GANs.

1. Introduction

Natural images are formed as a function of the objects' underlying physical attributes, such as geometry, material, and illumination. Estimating these object intrinsics from images is one of the core problems in computer vision, and is often referred to as inverse rendering, *i.e.*, the inverse process of rendering in computer graphics. It has wide applications in AR/VR and visual effects, such as relighting, material editing, and object pose editing.

A classic approach for inverse rendering is photometric stereo [1, 13], which requires multi-view and multi-lighting images of a scene to be captured with a light stage. These paired images provide sufficient information for estimating object intrinsics. However, the need of a sophisticated light stage setup makes it difficult to apply to diverse object categories, especially in-the-wild objects like cars. To get rid of this constraint, another line of work aims at performing inverse rendering in an unsupervised or weakly-supervised manner [18, 17, 12, 52], where only unpaired 2D image collections or weak annotations are given. Despite great progress, these approaches often miss fine-grained geometry details and cannot recover specular properties due to the inherent ambiguity.

While the lack of paired data is the main challenge of unsupervised inverse rendering, a continuous image distribution modelled by a GAN [14, 22] provides a possibility
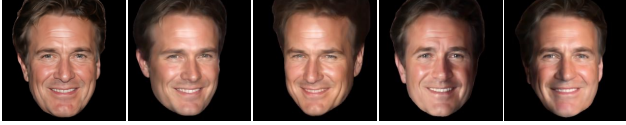
Figure 2: Images generated by a GAN [22] with different viewpoints and lighting conditions. The various specular highlight effects reveal the non-Lambertian material properties of the face.

to create pseudo light-stage data. GANs implicitly mimic the image formation process by learning from unpaired image collections. It is shown in the recent Shape-from-GAN works [32, 62] that the pseudo paired data generated by GAN can be used to reconstruct 3D shapes. However, these methods do not account for specular material properties and their mesh-based representation limits the precision of reconstructed 3D shapes. We believe that exploiting GANs for inverse rendering is still not fully explored. For instance, Fig. 2 provides GAN-generated images with variations on not only viewpoints but also the specular highlights, showing evidence that GANs implicitly capture the spatially-varying specular properties of the face.

Motivated by the above observation, in this work, we are interested to go one step further, exploring the problem of non-Lambertian inverse rendering of image GANs with finer details. In particular, we present a method that, for each image generated by a pretrained GAN, learns to predict its underlying 3D shape, material properties including albedo and specular components (*i.e.*, specular intensity and shininess), and lighting condition. Our method partially uses [32] as a tool, but we go beyond it to further explore if GANs can also be used to disentangle material properties. While the training is based on GAN-generated images, our method can also be applied to real images.

Concretely, to model high-quality 3D shape and non-Lambertian appearance, we devise an implicit neural surface representation to model the 3D surface, albedo, and specular properties. Rendering images from these intrinsics are conducted based on differentiable volume rendering [16] and Phong shading [35]. Inspired by [32], we adopt an exploration-and-exploitation algorithm to generate pseudo multi-view and multi-lighting images from a pretrained GAN. The neural surface representation can then be optimized with a reconstruction loss on the pseudo paired data along with regularization losses that facilitate training. However, the multi-view images generated by 2D GANs are not strictly 3D consistent, leading to imprecise results with noticeable mismatch to the target image. To address this issue, we adopt a shading-based refinement process that adapts our model to fully exploit information from the target image and effectively resolves the mismatch and recovers more fine details.

With the above designs, our approach, namely GAN2X,

can reconstruct high-quality 3D shape, albedo, and specular properties for different object categories, as shown in Fig. 1. The results show that image GANs do implicitly capture non-Lambertian object intrinsics, and thus provide a new way towards unsupervised inverse rendering. With such capability, GAN2X is also a powerful approach for unsupervised 3D shape learning from unpaired image collections. On the task of albedo and surface normal estimation and single-view 3D face reconstruction, GAN2X significantly outperforms existing unsupervised baselines. We also show promising results in downstream applications including image relighting and lifting 2D GANs to 3D with decomposed intrinsics.

Our contributions are summarized as follows: (1) We propose GAN2X that achieves high-quality non-Lambertian inverse rendering using GANs pretrained on unpaired images only. Our work reveals the large potential of GANs for learning spatially-varying non-Lambertian material properties. (2) GAN2X naturally serves as a strong unsupervised 3D shape learning approach. We show the state-of-the-art performance on unsupervised single-view 3D reconstruction of faces. (3) GAN2X enables a wide range of downstream applications including lifting 2D GANs to 3D with decomposed intrinsics and photorealistic visual effects like relighting and material editing.

2. Related Work

2.1. Generative Adversarial Networks

We have witnessed the great progress of Generative Adversarial Networks (GANs) in image synthesis starting from the first GAN model [14]. In particular, the StyleGAN family [21, 22, 20] has successfully produced photo-realistic and high-resolution imaginary. However, the image synthesis process of GANs is usually treated as a black box, which lacks physical interpretability (*e.g.*, 3D shape, material properties). There has been some works that extend GANs to enable 3D control [45, 11, 64, 46, 24, 38, 39], but such control is driven by the guidance of an external 3D morphable model [4] or a 3D mesh input. Different from these methods, our method does not require additional 3D geometry as input and can explicitly recover the physical attributes including 3D shape and material properties.

Another line of work is 3D-aware GANs, which adopt 3D representations (*e.g.*, voxel grids, and neural fields) or their integration with 2D generative models to enable 3D controllable image synthesis [40, 9, 31, 15, 8, 65, 53, 33, 54, 43]. However, most of these methods do not explicitly model the reflectance and the shading process, thus resulting in suboptimal 3D geometry and the lack of control on lighting. While [33] models illumination explicitly, it makes the assumption of Lambertian reflectance, which does not account for specular highlight. Besides,

the learned 3D shapes in [33] still lack fine-grained details due to limited resolution for training. In contrast, we make the first attempt to recover non-Lambertian material properties from pretrained GANs. Unlike previous works that adopt memory-consuming 3D representations to learn a 3D GAN, our work shows that the 3D geometry and material are also readily achievable from off-the-shelf 2D GANs that are more efficient to train.

2.2. Supervised Inverse Rendering

Inverse rendering has been well studied given paired images of an object from multiple viewpoints and light conditions. A typical way is to estimate object intrinsics by fitting photometric or geometric models to reconstruct the paired images. Conventional photometric stereo methods [1, 13, 28] perform this based on meshes while recently it is extended to implicit neural representations [5, 42, 61, 6, 60]. These methods are object-specific, *i.e.*, they do not generalize to unseen object instances. Some learning-based methods are also generalizable by learning from multi views, enabling test-time reconstruction with sparse views or a single view [59, 3, 7, 58, 2, 23]. However, collecting multi-view and multi-lighting images is resource-demanding and difficult to apply to in-the-wild objects like cars. Some methods use synthetic data for training [26, 29, 41, 25, 37], but training on such data requires us to solve the domain-gap problem for generalization to real in-the-wild images. To get rid of these limitations, in this work, we focus on the unsupervised setting where only unpaired image collections are available for training.

2.3. Unsupervised Inverse Rendering

In contrast to the supervised approaches, there has been a surge of interest to develop inverse rendering methods trained only on unpaired image collections. These methods predict the 3D geometry and appearance reconstruction from the input image and use image-space losses computed between the input and the rendered reconstruction. Since this is an ill-posed problem, early approaches focused on human faces and bodies using 3D priors [48, 17], with some methods also learning components of the prior from videos [44, 47]. Several methods learn to reconstruct the object shapes of general categories [12, 18, 55] using weak supervision like template shapes, masks, or hand-crafted priors (*e.g.*, object symmetry and smoothness). There are a few attempts for learning shape as well as material and illumination from unpaired image collections. Wu et al. [52] uses object symmetry and assumes Lambertian materials. The follow-up work [51] does not limit the object to be Lambertian, but requires objects to have rotational symmetry. Similar to us, [32] and [62] also exploit GANs to reconstruct 3D shapes. GAN2Shape [32] assumes Lambertian reflectance and does not recover high-quality albedo, while [62] does

not disentangle albedo and illumination. Concurrent to us, [50] relies on coarse 3D shape initialization to estimate the object intrinsics, and assumes globally shared specular intensity and shininess. In contrast, our approach does not rely on coarse shape, and recovers spatially-varying non-Lambertian material properties. In this work, we show that GANs implicitly capture object shape and material properties, and thus provide a new way for unsupervised inverse rendering with precise recovery of geometry and material properties.

3. Method

Our approach aims to recover the intrinsic components (3D shape, albedo, specular properties) at high quality for any image generated by an image GAN. In the following, we first provide some preliminaries on GANs. We then describe how we represent the intrinsic components and render images from them. Finally, we introduce our inverse rendering algorithm that distills the knowledge of the GAN to recover the intrinsic components.

Preliminaries on Generative Adversarial Networks (GANs). A GAN learns the data distribution via a min-max game between a generator G and a discriminator D [14]. After training from an image dataset, the generator can map a latent code z to an image. In this work, our study is based on StyleGAN2 [22], a classic GAN consisting of two parts. First, the latent code $z \in \mathcal{Z}$ is mapped to an intermediate latent code $w \in \mathcal{W}$ via a mapping network $F_w : \mathcal{Z} \mapsto \mathcal{W}$. Then w is used to produce the output image with a synthesis network. We refer to the synthesis network as G in the following sections.

3.1. Shape and Material Model

Scene representation. In order to model object shape and material with high fidelity, we adopt implicit neural fields to represent these factors. Specifically, for any image $\mathbf{I} \in \mathbb{R}^{3 \times H \times W}$ generated by the GAN, its 3D shape is represented by an MLP $F_s : (\mathbf{I}, \boldsymbol{x}) \mapsto s$, where $\boldsymbol{x} \in \mathbb{R}^3$ is the coordinate of any point and $s \in \mathbb{R}$ is its signed distance s to the object surface. Thus, the surface $\mathcal{S}$ of the object is $\mathcal{S} = \{\boldsymbol{x} \in \mathbb{R}^3 | F_s(\mathbf{I}, \boldsymbol{x}) = 0\}$. The material properties are represented by another MLP $F_m : (\mathbf{I}, \boldsymbol{x}) \mapsto (\boldsymbol{a}, k_s, p)$, where $\boldsymbol{a} \in \mathbb{R}^3$ is the diffuse albedo, $k_s \in [0, 1]$ is the specular intensity, and $p \in [p_{min}, p_{max}]$ is the shininess. Unlike NeRF [30] that models view-dependent color, in our formulation the material property is view-independent while view-dependent effects are explicitly modeled with shading as would be introduced later. The conditioning of F_s and F_m on $\mathbf{I}$ is implemented by concatenating the embedding of $\mathbf{I}$ with $\boldsymbol{x}$ as the input. The embedding can be obtained from 1) an additional encoder E_I and 2) the latent code w corresponding to $\mathbf{I}$ from the GAN. Their differences will be discussed latter.
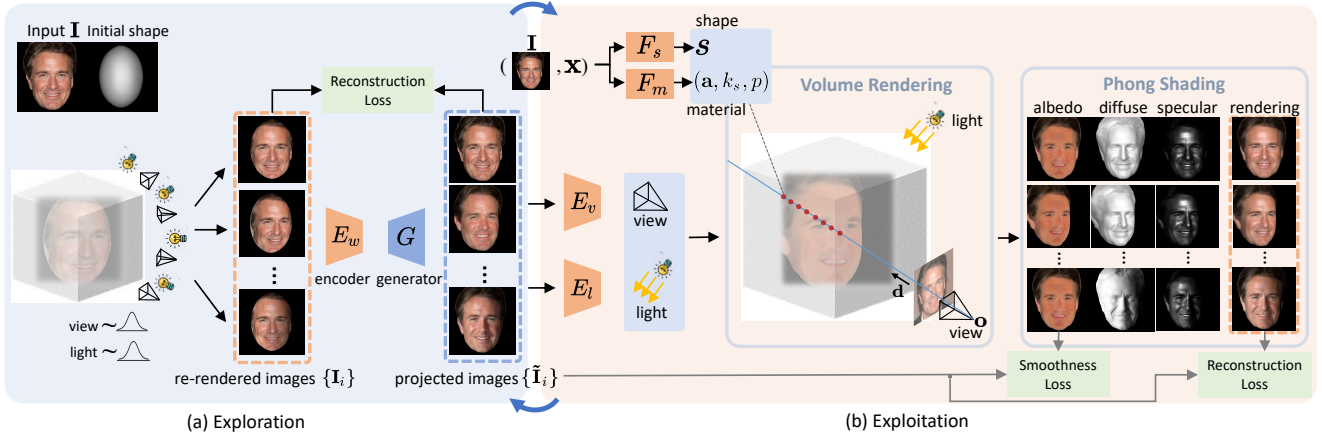
Figure 3: **Method Overview.** (a) In the exploration step, we use a convex shape prior to guide the GAN generator to produce projected images with various viewpoint and lighting conditions. (b) In the exploitation step, we leverage the projected images as pseudo paired data to optimize the underlying intrinsic components. The intrinsic components including shape, albedo, and specular properties are represented via implicit neural fields. Images can be rendered from this representation via volume rendering and Phong shading, which are naturally amenable to gradient-based optimization.

Apart from F_s and F_m, we also have a viewpoint encoder E_v that takes $\mathbf{I}$ as input and predicts the camera view $\boldsymbol{v} \in \mathbb{R}^6$ and a lighting encoder E_l that predicts the lighting condition $\boldsymbol{L} = (\boldsymbol{l}, k_a, k_d)$. Here $\boldsymbol{l} \in \mathbb{S}^2$ is the light direction, $k_a \in [0, 1]$ is the ambient coefficient, and $k_d \in [0, 1]$ is the diffuse coefficient. This scene representation allows to render an image $\hat{\mathbf{I}}$ from any viewpoint or lighting condition via a rendering process Φ as $\hat{\mathbf{I}} = \Phi(F_s(\mathbf{I}), F_m(\mathbf{I}), \boldsymbol{v}, \boldsymbol{L})$. Next, we introduce Φ in detail.

Rendering. Solving the inverse rendering problem requires our model to render images from the above intrinsics in a differentiable and easy-to-optimize manner. We use volume rendering [16] and Phong shading [35] to achieve this. To render the color $\boldsymbol{C}$ of a camera ray $\boldsymbol{r}(t) = \boldsymbol{o} + t\boldsymbol{d}$ with near and far bounds t_n and t_f, we first render the albedo $\boldsymbol{A}$, specular intensity K_s, shininess P, and surface normal $\boldsymbol{n} \in \mathbb{R}^3$ via:

$$\boldsymbol{A}(\boldsymbol{r}), K_s(\boldsymbol{r}), P(\boldsymbol{r}) = \int_{t_n}^{t_f} w(t) F_m(\boldsymbol{r}(t)) dt \quad (1)$$

$$\boldsymbol{n}(\boldsymbol{r}) = \hat{\boldsymbol{n}}(\boldsymbol{r}) / \|\hat{\boldsymbol{n}}(\boldsymbol{r})\|_2,$$

$$\text{where } \hat{\boldsymbol{n}}(\boldsymbol{r}) = -\int_{t_n}^{t_f} w(t) \nabla_{\boldsymbol{r}(t)} F_s(\boldsymbol{r}(t)) dt, \quad (2)$$

Here $w(t)$ is the weight function for volume rendering, for which we use the normalized S-density following NeuS [49]. We then perform Phong shading to get the final color $\boldsymbol{C}$:

$$\boldsymbol{C} = \tau\Big(k_a \boldsymbol{A} + k_d \big(\max(0, \boldsymbol{n} \cdot \boldsymbol{l}) \boldsymbol{A} + K_s \max(0, \boldsymbol{n} \cdot \boldsymbol{h})^P\big)\Big), \quad (3)$$

where $\boldsymbol{h}$ is the bisector of the angle between viewing direction $-\boldsymbol{d}$ and $\boldsymbol{l}$, and $\tau(\boldsymbol{x}) = \boldsymbol{x}^{1/\gamma}, \gamma = 2.2$ is a tone-

mappping function to ensure a more even brightness distribution. A limitation of this vanilla Phong shading is that it produces similar darkness at surface areas opposite to the light source. We observe that the surface with smaller $\boldsymbol{n} \cdot \boldsymbol{l}$ often tends to be darker due to complex reflection in the environment. Thus, we optionally replace $\max(0, \boldsymbol{n} \cdot \boldsymbol{l})$ with $\max(0, \boldsymbol{n} \cdot \boldsymbol{l}) + k_n \min(0, \boldsymbol{n} \cdot \boldsymbol{l}), k_n \in [0, 1]$, where the new negative term is to compensate for the darkness variation at opposite-light areas. We show that this revision leads to more realistic relighting in experiments.

Note that $w(t)$ in Eq. 1 and Eq. 2 has a learnable inverse standard deviation parameter σ (s in [49]) that controls the concentration of density on the surface. It is directly optimized during training in [49]. However, in our case, this would produce a sub-optimal solution as the 3D inconsistency of GAN generated images would impede the convergence of $1/\sigma$. Thus, we manually increase σ from σ_{min} to σ_{max} with an exponential schedule as $\sigma_i = \sigma_{max} + \exp(-i\beta)(\sigma_{min} - \sigma_{max})$, where i is the number of training iterations and β controls the decay.

3.2. Inverse Rendering

We perform inverse rendering based on the shape and material model introduced above. In order to generate a number of approximated paired images of various viewpoint and lighting conditions using the GAN, we adopt an exploration and exploitation algorithm following [32]. Different from [32], in this work the inverse rendering is based on the new non-Lambertian neural representation and rendering equation introduced before. Besides, we also devise a chromaticity-based smoothness loss to regularize the material and a viewpoint and lighting loss to stabilize training.

A shading-based refinement technique will also be introduced in the next subsection.

Exploration. We first initialize F_s to produce an ellipsoid as a convex shape prior following [32]. The camera viewpoint v and lighting L are initialized to be a canonical setting v_0 and L_0. We then optimize the material network F_m using the reconstruction loss $\mathcal{L}(\mathbf{I}, \Phi(F_s(\mathbf{I}), F_m(\mathbf{I}), v_0, L_0))$, where $\mathcal{L}$ is L1 loss. As shown in Fig. 3 (a), with this initial guess of the scene, we can re-render a number of new images $\{\mathbf{I}_i | i = 1, 2, ..., m\}$ from different viewpoints $\{v_i\}$ and lighting conditions $\{L_i\}$ using Φ, where $\{v_i\}$ and $\{L_i\}$ are randomly sampled from their prior distributions. These re-rendered images $\{\mathbf{I}_i\}$ roughly reveal the change of viewpoint and lighting, thus can serve as a guidance for the exploration in the GAN image manifold. Specifically, we reconstruct $\{\mathbf{I}_i\}$ using the GAN generator G by training an encoder E_w that predicts the latent codes. After training, this would produce the GAN-reconstructed images $\{\tilde{\mathbf{I}}_i = G(E_w(\mathbf{I}_i))\}$, namely projected images, which have viewpoints and lighting conditions that resemble the re-rendered images $\{\mathbf{I}_i\}$.

Exploitation. The projected images $\{\tilde{\mathbf{I}}_i\}$ obtained in the exploration process could be viewed as pseudo paired images for $\mathbf{I}$. Thus, we exploit their information to recover the intrinsic components. As shown in Fig. 3 (b), the viewpoint $\tilde{v}_i$ and lighting $\tilde{l}_i$ for each image $\tilde{\mathbf{I}}_i$ are predicted by the viewpoint encoder E_v and lighting encoder E_l respectively. They are then used to render images with the shared shape field and material field. The scene related networks F_s, F_m, E_v, E_l can be jointly optimized using the reconstruction loss:

$$\mathcal{L}_{recon} = \frac{1}{m} \sum_{i=0}^{m} \mathcal{L}'\left(\tilde{\mathbf{I}}_i, \Phi\left(F_s(\mathbf{I}), F_m(\mathbf{I}), E_v(\tilde{\mathbf{I}}_i), E_l(\tilde{\mathbf{I}}_i)\right)\right) \tag{4}$$

where $\mathcal{L}'$ is a combination of L1 loss and perceptual loss.

In addition, we also regularize the material using a chromaticity-based smoothness loss. For the material maps (A, K_s, P) rendered from $(F_s(\mathbf{I}), F_m(\mathbf{I}), E_v(\tilde{\mathbf{I}}_i), E_l(\tilde{\mathbf{I}}_i))$, we denote their concatenation as $\mathbf{M}_i$. The chromaticity-based smoothness loss is defined as:

$$\mathcal{L}_{reg} = \frac{1}{m} \sum_{i=0}^{m} \left(\|\psi(\delta_x \mathbf{\Gamma}_i)\delta_x \mathbf{M}_i\|_1 + \|\psi(\delta_y \mathbf{\Gamma}_i)\delta_y \mathbf{M}_i\|_1\right) \tag{5}$$

where $\mathbf{\Gamma}_i$ is calculated from $\tilde{\mathbf{I}}_i$ by taking its $a^* b^*$ channels in the CIELAB color space, δ_x and δ_y are used to signify the computation of image gradients, and $\psi(x) = \exp(-\|x\|^2/\gamma), \gamma = 10$ is a non-linear function. This regularization motivates pixels with close chromaticity to have similar materials.

As the projected images share similar viewpoints and lighting conditions as the re-rendered images, we can use $\{v_i\}$ and $\{L_i\}$ to guide the learning of E_v and E_l at the beginning of training:

$$\mathcal{L}_{vl} = \frac{1}{m} \sum_{i=0}^{m} \left(\|\{v_i\} - E_v(\tilde{\mathbf{I}}_i)\|_1 + \|\{l_i\} - E_l(\tilde{\mathbf{I}}_i)\|_1\right) \tag{6}$$

In summary, the final training objective is:

$$\theta_s, \theta_m, \theta_v, \theta_l = \underset{\theta_s, \theta_m, \theta_v, \theta_l}{\arg\min} \mathcal{L}_{recon} + \lambda_{reg}\mathcal{L}_{reg} + \lambda_{vl}\mathcal{L}_{vl} \tag{7}$$

where $\theta_s, \theta_m, \theta_v, \theta_l$ are the parameters of networks F_s, F_m, E_v, E_l respectively, and $\mathcal{L}_{vl}$ is only used in the first 1k training iterations.

With this training objective, the object shape, material, and lighting will be more correctly inferred. We then use these updated intrinsic components as the initialization and repeat the exploration and exploitation steps for a few times. This allows to further distill the knowledge of the GAN and refine the results.

Joint training. We have introduced how our method can be applied to an individual instance $\mathbf{I}$ generated by the GAN. In this case, the dependence of F_s and F_m on $\mathbf{I}$ is actually not necessary. Having this dependence allows us to further extend our method for joint training on multiple instances as done in [32], which improves generalization. For example, if the conditioning of F_s and F_m on $\mathbf{I}$ is achieved by training an additional image encoder E_I, then we can apply our model to real images after joint training. And if the conditioning is based on the latent code w in the $\mathcal{W}$ space of the GAN, then the StyleGAN mapping network F_w together with our F_s and F_m forms a 3D GAN with decomposed intrinsic components. We show the applications of these variants in experiments.

3.3. Shading-based Refinement

While the above approach already produces pleasing results, the reconstructed shape and appearance could possibly bear noticeable mismatch with the target image. This is because the pseudo paired data generated by a 2D GAN is not strictly 3D-consistent. To this end, we draw inspiration from shape-from-shading literature [66] and adopt a shading-based refinement (SBR) step that fully exploits the information in the target image $\mathbf{I}$. Let $\hat{\mathbf{I}}, \hat{\mathbf{D}}$, and $\hat{\mathbf{M}}$ denote the image, depth, and material map rendered from $(F_s(\mathbf{I}), F_m(\mathbf{I}), E_v(\mathbf{I}), E_l(\mathbf{I}))$. The shading-based refinement objective is to reconstruct the target image with several constraints:

$$\theta_s, \theta_m, \theta_v, \theta_l = \underset{\theta_s, \theta_m, \theta_v, \theta_l}{\arg\min} \mathcal{L}'(\mathbf{I}, \hat{\mathbf{I}}) + \lambda_d \|\mathbf{D}, \hat{\mathbf{D}}\|_1$$
$$+ \lambda_m \|\mathbf{M}, \hat{\mathbf{M}}\|_1 + \lambda_{reg}\mathcal{L}_{reg} \tag{8}$$

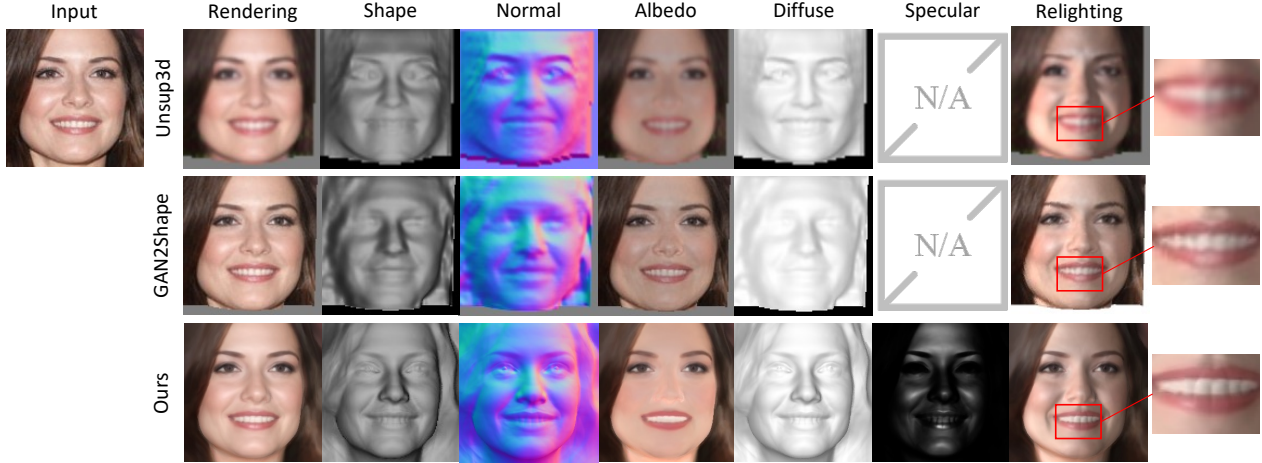| Input	Rendering	Shape	Normal	Albedo	Diffuse	Specular	Relighting

Figure 4: **Qualitative comparison.** Our approach achieves more accurate inverse rendering than baselines (zoom in to see details). Our relighting result successfully models the shift of specular highlight on the lip while baselines cannot.

where $\mathbf{D}$ and $\mathbf{M}$ are the initial values of $\hat{\mathbf{D}}$ and $\hat{\mathbf{M}}$ and are used to prevent the shape and material from deviating too much. Here $\mathcal{L}_{reg}$ is defined only for $\hat{\mathbf{M}}$. This optimization goal resolves the mismatch and recover more details from the target image while preserving a valid 3D shape and material, as we will show in experiments.

4. Experiments

We conduct extensive experiments to evaluate our approach GAN2X on unsupervised inverse rendering across different object categories including human faces, cat faces, and cars. We use datasets of unpaired image collections including CelebA [27], CelebAHQ [19], AFHQ cat [10], and LSUN car [57]. The GAN models we used are Style-GAN2 [22] pretrained on these datasets. We use off-the-shelf scene parsing models [56, 63] to remove the background. We recommend readers to refer to the supplementary material for more implementation details, qualitative results, and videos.

4.1. Inverse Rendering Results

Qualitative evaluation. Fig. 4 shows the qualitative comparison between our approach and two unsupervised inverse rendering baselines Unsup3d [52] and GAN2Shape [32] on the CelebA dataset. We visualize the rendering, 3D shape, surface normal, albedo, diffuse light map, and specular light map. Our approach reconstructs much more precise 3D shapes than the baselines, *e.g.*, we successfully recover the double-fold eyelids, teeth, and most wrinkles while the baselines miss fine details and produce blurry shapes. Besides, the baselines cannot decompose the specular light map. Consequently, the specular highlights are entangled into albedo maps. In contrast, our approach successfully disentangles the specular highlights from the diffuse albedo,

e.g., the highlights on the lips are correctly captured in the specular light map. This, in turn, results in a more accurate albedo map and produces more realistic relighting effects.

Our results on more datasets are shown in Fig. 5. GAN2X performs high-quality inverse rendering for human head, cat face, and cars, where the 3D shape, albedo, diffuse light map, and specular light map are accurately recovered. Besides, these reconstructed intrinsic components allow photorealistic re-rendering of the object under a novel viewpoint or lighting condition, showing large potential for image editing. We also visualize the learned specular components of different objects in Fig. 6. It is observed that our model successfully learns some meaningful object-varying and spatially-varying specular properties. For instance, the human lips have higher shininess than the human faces; the eyes of the cat have higher shininess than its fur; the car has higher specular intensity than other objects. These results verify our assumption that GANs implicitly capture object material properties. We note that learning spatially-varying specular intensity and shininess is very challenging and our method does not resolve all the ambiguities. For example, the tires of the car have high shininess, which is counter-intuitive. This is because its specular intensity is already very low and thus the shininess cannot get supervision. Addressing these ambiguities is left as a future work.

Quantitative evaluation. For quantitative evaluation, we first perform single-view 3D reconstruction on the H3DS dataset [36] to evaluate the 3D shape. The H3DS test set has ground truth 3D face scans for several identities. We take a single-view image as input and report the Chamfer Distance between our reconstructed shape and the ground truth. Apart from Unsup3d and GAN2Shape, we also compare with pi-GAN [9] and ShadeGAN [33] that can perform unsupervised 3D reconstruction via GAN inversion. All meth-
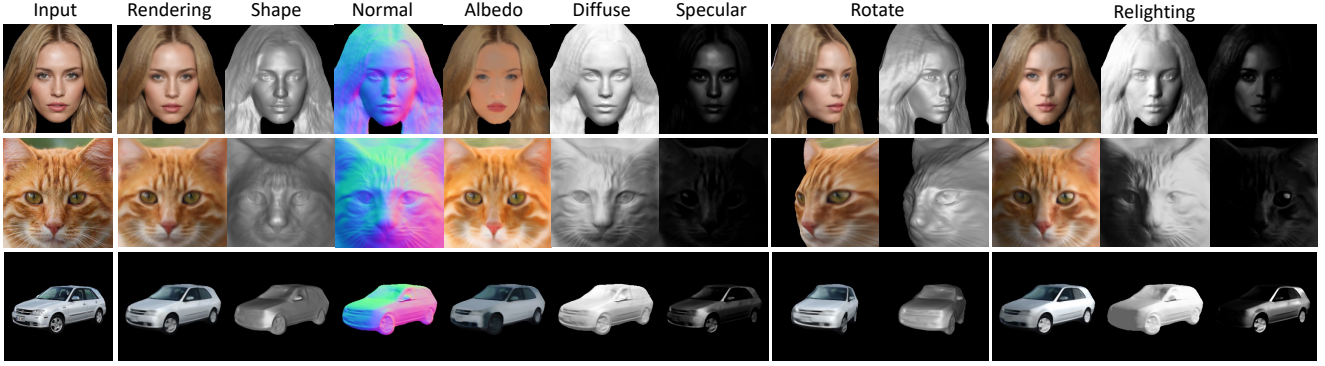
Figure 5: **Qualitative results.** We show inverse rendering and image editing results on the CelebaHQ, AFHQ Cat, and LSUN Car datasets.
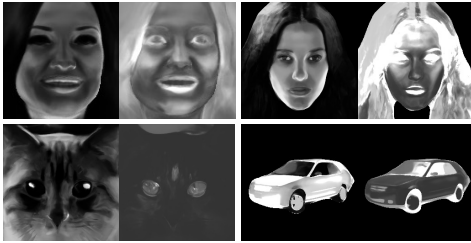

Figure 6: **Visualization of specular components** (*e.g.*, specular intensity and shininess) for objects in Fig. 4 and Fig. 5.
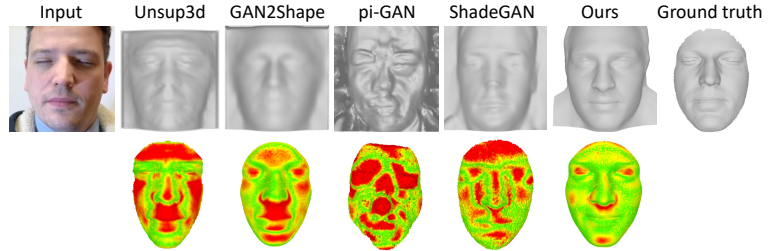

Figure 7: **Qualitative comparison** of single-view 3D reconstruction on H3DS dataset.

Method	Unsup3d	GAN2Shape	pi-GAN	ShadeGAN	Ours w/o SBR	Ours
CD↓	3.60	2.62	3.29	2.49	2.21	**2.08**

Table 1: **Single-view 3D reconstruction** on the H3DS dataset. We report chamfer distance (CD) between the predicted mesh and the ground truth mesh.

	Unsup3d	GAN2Shape	Ours
SIE ($\times 10^{-2}$)↓	3.21	3.05	**2.16**
MAD↓	18.66	21.75	**12.67**

Table 2: **Quantitative comparison** of albedo and surface normal on CelebA. We report scale-invariant error (SIE) for albedo and mean-angle deviation (MAD) for surface normal.

ods are trained on the CelebA dataset. Applying our model to these real images is achieved with the jointly trained image encoder E_I and F_s as discussed in Sec. 3.2. As shown in Fig. 7 and Tab. 1, our method reconstructs more accurate and more natural-looking 3D shapes than other baselines. Thus, our approach is also well suited for unsupervised 3D shape learning.

Evaluating albedo for our GAN-based inverse rendering setting is challenging as there lacks a suitable dataset with ground truth albedo. To this end, we leverage the state-of-the-art supervised approach Total Relighting [34] to produce pseudo ground truth, and report results on it as a reference. Total Relighting is trained on high-quality light-stage data, thus producing stable and reliable albedo and surface normal for human faces. We test on 500 images generated by the GAN trained on CelebA and report the scale-invariant error (SIE) for albedo and mean-angle deviation (MAD) for surface normal following [50]. As shown

in Tab. 2, our approach significantly outperforms baselines, showing better capability to recover face albedo and shape.

Ablation study. We further study the effects of several components in our pipeline, including the specular term in Eq. 3, the chromaticity-based smoothness loss in Eq. 5, the shading-based refinement (SBR), and the negative shading term introduced in Sec. 3.1. The results are shown in Tab. 3 and Fig. 8. It can be observed that including the specular term helps disentangle specular highlight with albedo. The reduced ambiguity in turn facilitates the learning of surface normal. Besides, the smoothness prior also leads to better disentanglement between albedo and shading, producing more natural albedo map. The example of Fig. 8 (a) is a challenging case where the rendering of our approach without SBR exhibits noticable mismatch with the input image. The SBR can effectively reduce the mismatch and recover more details in the target image. Quantitative results also
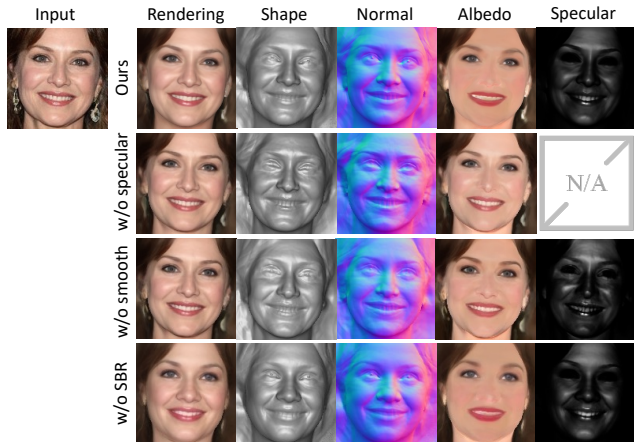
(a) Effects of specular component, smoothness prior, and SBR

(b) Effects of negative shading component on relighting

Figure 8: **Ablation study** of (a) specular component, smoothness prior, shading-based refinement (SBR) and (b) negative shading component.

	Ours	w/o specular	w/o smooth	w/o SBR	w/o neg
SIE ($\times 10^{-2}$) $\downarrow$	2.16	2.25	2.28	2.19	2.15
MAD $\downarrow$	12.67	12.88	12.63	12.74	12.66

Table 3: **Ablation study** of several design choices. The values have the same meaning as in Tab. 2.

confirm that SBR further improves the shape and albedo in our approach. Finally, the negative shading term has little effect on the quantitative results, but would improve qualitative relighting effects as it can compensate for darkness variation at surfaces opposite to the light source. As shown in Fig. 8 (b), the rendering without this negative shading term tends to produce constant darkness at the opposite-light areas while including this term reveals some geometry details and is more realistic.

4.2. Other Applications

Real image editing. We have shown that GAN2X can be used for single-view 3D reconstruction of real images, here we also demonstrate its applicability for other editing effects for real images. Fig. 9 provides an example of a real image, where our method successfully achieves satisfactory inverse rendering and image editing results. Here we also show material editing effects enabled by our method. By reducing the specular intensity and shininess, we can weaken the specular highlight on the face.
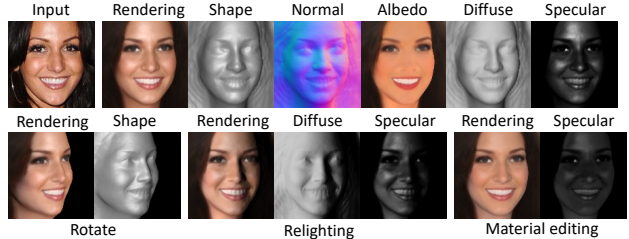

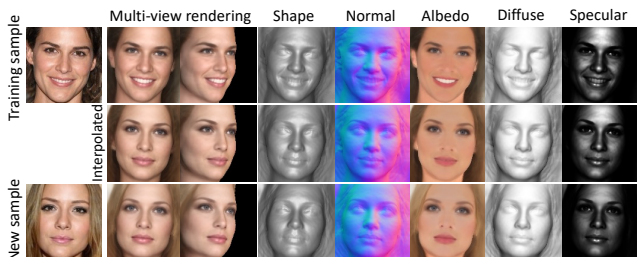Figure 9: Application on real image inverse rendering and editing.

Figure 10: Application on lifting 2D GANs to decomposed 3D GANs.

Lifting 2D GANs to decomposed 3D GANs. Finally, we observe that GAN2X jointly trained on multiple GAN samples is capable of inheriting the generative property of the 2D GAN. As shown in Fig. 10, our method not only works for samples used in training, but also generalizes to new samples of the 2D GAN. The different samples can also be naturally interpolated by interpolating the corresponding latent codes, which is an important property of GANs. The results demonstrate the possibility of lifting a 2D GAN to a decomposed 3D GAN with our approach.

5. Conclusion

We have presented a new approach for unsupervised inverse rendering. To overcome the shortage of paired data, we leverage the generative modeling capability of GANs to create pseudo paired data that reflect the underlying intrinsic components. With volume rendering and a Phong shading based implicit neural representation, our approach successfully recovers high-quality 3D shapes, albedo, and specular properties for different object categories, which opens up wide downstream applications. Notably, GAN2X is the first method that can learn spatially-varying specular properties from merely unpaired image collections. For future work, our approach could be further combined with recent advances in 3D GANs to further refine their geometry and decompose the material properties.

References

[1] Neil Alldrin, Todd Zickler, and David Kriegman. Photometric stereo with non-parametric and spatially-varying reflectance. In *CVPR*. 4321, 4323

[2] Jonathan T Barron and Jitendra Malik. Shape, illumination, and reflectance from shading. *TPAMI*, 37(8):1670–1687, 2014. 4323

[3] Sai Bi, Zexiang Xu, Kalyan Sunkavalli, David Kriegman, and Ravi Ramamoorthi. Deep 3d capture: Geometry and reflectance from sparse multi-view images. In *CVPR*, 2020. 4323

[4] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, 1999. 4322

[5] Mark Boss, Raphael Braun, Varun Jampani, Jonathan T Barron, Ce Liu, and Hendrik Lensch. Nerd: Neural reflectance decomposition from image collections. In *ICCV*, 2021. 4323

[6] Mark Boss, Varun Jampani, Raphael Braun, Ce Liu, Jonathan T. Barron, and Hendrik P.A. Lensch. Neural-pil: Neural pre-integrated lighting for reflectance decomposition. In *NeurIPS*, 2021. 4323

[7] Mark Boss, Varun Jampani, Kihwan Kim, Hendrik Lensch, and Jan Kautz. Two-shot spatially-varying brdf and shape estimation. In *CVPR*, 2020. 4323

[8] Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. In *CVPR*, 2022. 4322

[9] Eric R Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *CVPR*, 2021. 4322, 4326

[10] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *CVPR*, 2020. 4326

[11] Yu Deng, Jiaolong Yang, Dong Chen, Fang Wen, and Xin Tong. Disentangled and controllable face image generation via 3d imitative-contrastive learning. In *CVPR*, 2020. 4322

[12] Shubham Goel, Angjoo Kanazawa, and Jitendra Malik. Shape and viewpoint without keypoints. In *ECCV*, 2020. 4321, 4323

[13] Dan B Goldman, Brian Curless, Aaron Hertzmann, and Steven M Seitz. Shape and spatially-varying brdfs from photometric stereo. *TPAMI*, 32(6):1060–1071, 2009. 4321, 4323

[14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014. 4321, 4322, 4323

[15] Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. Stylenerf: A style-based 3d-aware generator for high-resolution image synthesis. In *ICLR*, 2022. 4322

[16] James T Kajiya and Brian P Von Herzen. Ray tracing volume densities. *ACM SIGGRAPH computer graphics*, 18(3):165–174, 1984. 4322, 4324

[17] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018. 4321, 4323

[18] Angjoo Kanazawa, Shubham Tulsiani, Alexei A Efros, and Jitendra Malik. Learning category-specific mesh reconstruction from image collections. In *ECCV*, 2018. 4321, 4323

[19] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *ICLR*, 2018. 4326

[20] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. In *NeurIPS*, 2021. 4322

[21] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, pages 4401–4410, 2019. 4322

[22] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, pages 8110–8119, 2020. 4321, 4322, 4323, 4326

[23] Tejas D Kulkarni, William F Whitney, Pushmeet Kohli, and Josh Tenenbaum. Deep convolutional inverse graphics network. In *NeurIPS*, 2015. 4323

[24] Thomas Leimkühler and George Drettakis. Freestylegan: Free-view editable portrait rendering with the camera manifold. 40(6), 2021. 4322

[25] Zhengqin Li, Mohammad Shafiei, Ravi Ramamoorthi, Kalyan Sunkavalli, and Manmohan Chandraker. Inverse rendering for complex indoor scenes: Shape, spatially-varying lighting and svbrdf from a single image. In *CVPR*, 2020. 4323

[26] Zhengqin Li, Zexiang Xu, Ravi Ramamoorthi, Kalyan Sunkavalli, and Manmohan Chandraker. Learning to reconstruct shape and spatially-varying reflectance from a single image. *ACM Transactions on Graphics (TOG)*, 37(6):1–11, 2018. 4323

[27] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, 2015. 4326

[28] Fujun Luan, Shuang Zhao, Kavita Bala, and Zhao Dong. Unified shape and svbrdf recovery using differentiable monte carlo rendering. In *Computer Graphics Forum*, volume 40, pages 101–113, 2021. 4323

[29] Abhimitra Meka, Maxim Maximov, Michael Zollhoefer, Avishek Chatterjee, Hans-Peter Seidel, Christian Richardt, and Christian Theobalt. Lime: Live intrinsic material estimation. In *CVPR*, 2018. 4323

[30] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 4323

[31] Michael Niemeyer and Andreas Geiger. Campari: Camera-aware decomposed generative neural radiance fields. In *3DV*, 2021. 4322

[32] Xingang Pan, Bo Dai, Ziwei Liu, Chen Change Loy, and Ping Luo. Do 2d gans know 3d shape? unsupervised 3d shape reconstruction from 2d image gans. In *ICLR*, 2021. 4322, 4323, 4324, 4325, 4326

[33] Xingang Pan, Xudong Xu, Chen Change Loy, Christian Theobalt, and Bo Dai. A shading-guided generative implicit model for shape-accurate 3d-aware image synthesis. In *NeurIPS*, 2021. 4322, 4323, 4326

[34] Rohit Pandey, Sergio Orts Escolano, Chloe Legendre, Christian Haene, Sofien Bouaziz, Christoph Rhemann, Paul Debevec, and Sean Fanello. Total relighting: learning to relight portraits for background replacement. *ACM Transactions on Graphics (TOG)*, 40(4):1–21, 2021. 4327

[35] Bui Tuong Phong. Illumination for computer generated pictures. *Communications of the ACM*, 18(6):311–317, 1975. 4322, 4324

[36] Eduard Ramon, Gil Triginer, Janna Escur, Albert Pumarola, Jaime Garcia, Xavier Giro-i Nieto, and Francesc Moreno-Noguer. H3d-net: Few-shot high-fidelity 3d head reconstruction. In *ICCV*, 2021. 4326

[37] Shen Sang and Manmohan Chandraker. Single-shot neural relighting and svbrdf estimation. In *ECCV*, 2020. 4323

[38] Kripasindhu Sarkar, Vladislav Golyanik, Lingjie Liu, and Christian Theobalt. Style and pose control for image synthesis of humans from a single monocular view, 2021. 4322

[39] Kripasindhu Sarkar, Lingjie Liu, Vladislav Golyanik, and Christian Theobalt. Humangan: A generative model of humans images. In *3DV*, 2021. 4322

[40] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. Graf: Generative radiance fields for 3d-aware image synthesis. *arXiv preprint arXiv:2007.02442*, 2020. 4322

[41] Soumyadip Sengupta, Angjoo Kanazawa, Carlos D Castillo, and David W Jacobs. Sfsnet: Learning shape, reflectance and illuminance of facesin the wild'. In *CVPR*, 2018. 4323

[42] Pratul P Srinivasan, Boyang Deng, Xiuming Zhang, Matthew Tancik, Ben Mildenhall, and Jonathan T Barron. Nerv: Neural reflectance and visibility fields for relighting and view synthesis. In *CVPR*, 2021. 4323

[43] Ayush Tewari, MalliKarjun B R, Xingang Pan, Ohad Fried, Maneesh Agrawala, and Christian Theobalt. Disentangled3d: Learning a 3d generative model with disentangled geometry and appearance from monocular images. In *CVPR*, 2022. 4322

[44] Ayush Tewari, Florian Bernard, Pablo Garrido, Gaurav Bharaj, Mohamed Elgharib, Hans-Peter Seidel, Patrick Pérez, Michael Zollhofer, and Christian Theobalt. Fml: Face model learning from videos. In *CVPR*, pages 10812–10822, 2019. 4323

[45] Ayush Tewari, Mohamed Elgharib, Gaurav Bharaj, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zollhofer, and Christian Theobalt. Stylerig: Rigging stylegan for 3d control over portrait images. In *CVPR*, 2020. 4322

[46] Ayush Tewari, Mohamed Elgharib, Mallikarjun BR, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zöllhofer, and Christian Theobalt. Pie: Portrait image embedding for semantic control. volume 39, December 2020. 4322

[47] Ayush Tewari, Hans-Peter Seidel, Mohamed Elgharib, Christian Theobalt, et al. Learning complete 3d morphable face models from images and videos. In *CVPR*, pages 3361–3371, 2021. 4323

[48] Ayush Tewari, Michael Zollhofer, Hyeongwoo Kim, Pablo Garrido, Florian Bernard, Patrick Perez, and Christian Theobalt. Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *ICCV*, 2017. 4323

[49] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. In *NeurIPS*, 2021. 4324

[50] Felix Wimbauer, Shangzhe Wu, and Christian Rupprecht. De-rendering 3d objects in the wild. *arXiv preprint arXiv:2201.02279*, 2022. 4323, 4327

[51] Shangzhe Wu, Ameesh Makadia, Jiajun Wu, Noah Snavely, Richard Tucker, and Angjoo Kanazawa. De-rendering the world's revolutionary artefacts. In *CVPR*, 2021. 4323

[52] Shangzhe Wu, Christian Rupprecht, and Andrea Vedaldi. Unsupervised learning of probably symmetric deformable 3d objects from images in the wild. In *CVPR*, 2020. 4321, 4323, 4326

[53] Xudong Xu, Xingang Pan, Dahua Lin, and Bo Dai. Generative occupancy fields for 3d surface-aware image synthesis. In *NeurIPS*, 2021. 4322

[54] Yinghao Xu, Sida Peng, Ceyuan Yang, Yujun Shen, and Bolei Zhou. 3d-aware image synthesis via learning structural and textural representations. In *CVPR*, 2022. 4322

[55] Yufei Ye, Shubham Tulsiani, and Abhinav Gupta. Shelf-supervised mesh prediction in the wild. In *CVPR*, pages 8843–8852, 2021. 4323

[56] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *ECCV*, 2018. 4326

[57] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015. 4326

[58] Ye Yu, Abhimitra Meka, Mohamed Elgharib, Hans-Peter Seidel, Christian Theobalt, and William AP Smith. Self-supervised outdoor scene relighting. In *ECCV*, 2020. 4323

[59] Ye Yu and William AP Smith. Inverserendernet: Learning single image inverse rendering. In *CVPR*, 2019. 4323

[60] Kai Zhang, Fujun Luan, Qianqian Wang, Kavita Bala, and Noah Snavely. Physg: Inverse rendering with spherical gaussians for physics-based material editing and relighting. In *CVPR*, 2021. 4323

[61] Xiuming Zhang, Pratul P Srinivasan, Boyang Deng, Paul Debevec, William T Freeman, and Jonathan T Barron. Nerfactor: Neural factorization of shape and reflectance under an unknown illumination. *ACM Transactions on Graphics (TOG)*, 40(6):1–18, 2021. 4323

[62] Yuxuan Zhang, Wenzheng Chen, Huan Ling, Jun Gao, Yinan Zhang, Antonio Torralba, and Sanja Fidler. Image gans meet differentiable rendering for inverse graphics and interpretable 3d neural rendering. In *ICLR*, 2021. 4322, 4323

[63] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017. 4326

[64] Hang Zhou, Jihao Liu, Ziwei Liu, Yu Liu, and Xiaogang Wang. Rotate-and-render: Unsupervised photorealistic face rotation from single-view images. In *CVPR*, 2020. 4322

[65] Peng Zhou, Lingxi Xie, Bingbing Ni, and Qi Tian. CIPS-3D: A 3D-Aware Generator of GANs Based on Conditionally-Independent Pixel Synthesis. 2021. 4322

[66] Michael Zollhöfer, Angela Dai, Matthias Innmann, Chenglei Wu, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Shading-based refinement on volumetric signed distance functions. *ACM Transactions on Graphics (TOG)*, 34(4):1–14, 2015. 4325