Unsupervised Style-based Explicit 3D Face Reconstruction from Single Image

Heng Yu   Zoltán Á. Milacski   László A. Jeni

Robotics Institute, Carnegie Mellon University

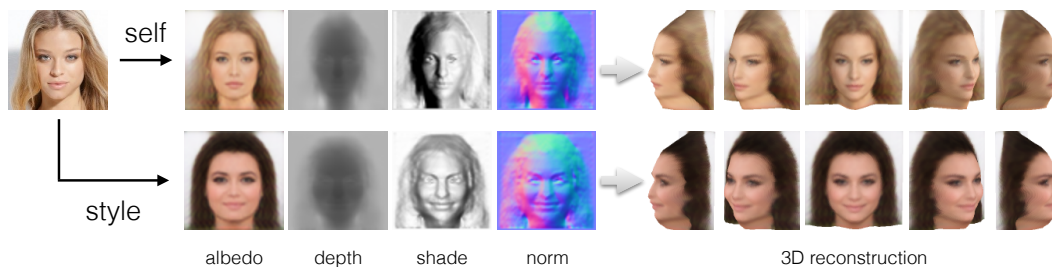{hengyu, zmilacsk}@andrew.cmu.edu   laszlojeni@cmu.edu

Figure 1. Given a single 2D image, we reconstruct the 3D object (top) by predicting the features of the rendering process: albedo, depth, shading and surface normals. Then the method can synthesize images from an arbitrary viewpoint with a new style, potentially including changes in both shape and appearance (bottom).

Abstract

Inferring 3D object structures from a single image is an ill-posed task due to depth ambiguity and occlusion. Typical resolutions in the literature include leveraging 2D or 3D ground truth for supervised learning, as well as imposing hand-crafted symmetry priors or using an implicit representation to hallucinate novel viewpoints for unsupervised methods. In this work, we propose a general adversarial learning framework for solving Unsupervised 2D to Explicit 3D Style Transfer (UE3DST). Specifically, we merge two architectures: the unsupervised explicit 3D reconstruction network of Wu et al. and the Generative Adversarial Network (GAN) named StarGAN-v2. We experiment across three facial datasets (Basel Face Model, 3DFAW and CelebA-HQ) and show that our solution is able to outperform well established solutions such as DepthNet in 3D reconstruction and Pix2NeRF in conditional style transfer, while we also justify the individual contributions of our model components via ablation. In contrast to the aforementioned baselines, our scheme produces features for explicit 3D rendering, which can be manipulated and utilized in downstream tasks.

1. Introduction

Reconstructing the underlying 3D structure of objects from few 2D images or even a single 2D image is a promising research area, since we live in the 3D world. Inferring accurate 3D models of human faces, bodies and other objects can greatly promote the development of several fields like animation, video games, virtual reality, augmented reality and the metaverse. This task is difficult, however, due to various forms of underdetermination, including depth ambiguity and occlusion. The straighforward resolution to these problems is applying supervised learning with 2D/3D ground truth information, keypoints [26, 42] or depth maps [25,64]. However, it is hard and sometimes even impossible to collect such supervisory signals, as this requires considerable human effort. Hence, recent techniques [36,59,69] exploit unsupervised learning to leverage human prior knowledge about the natural objects in the world. This alleviates the need for annotated data, but the prior has to be chosen wisely for each object class. E.g., human faces are roughly symmetric, and exploiting this as a prior yields considerable improvements [66].

Besides reconstructing the 3D structure, one may be also interested in altering its style to one that is different from that of the original input image. Consider, e.g., supporting — or even replacing — CG artists by automatically changing a 3D human model's gender, haircut, skin color or age. Interestingly, *Unsupervised 2D to Explicit 3D Style Transfer (UE3DST)* is a less discussed problem in the literature. By *explicit*, we mean that the method extracts a set of features, which allow recovering the 3D shape via using an explicit rendering function [66]. Unfortunately,

existing methods are either (i) implicit 3D, or (ii) supervised. The widely adopted solution for (i) is hallucinating novel 3D viewpoints using some form of a Generative Adversarial Network (GAN) [18] or a Neural Radiance Field (NeRF) [34]. An approach for (ii) is SofGAN [9], which recovers explicit 3D and can change styles such as pose, shape, and texture; however, it requires semantic segmentation maps as ground truth. While these techniques for (i)-(ii) are able to produce remarkable results on their own benchmarks, they are less applicable due to (i) lacking explicit 3D features and (ii) relying entirely on supervision.

In this paper, we propose a framework to tackle the UE3DST problem by unifying the solutions of (a) unsupervised 3D reconstruction and (b) style transfer. Specifically, we introduce an unsupervised end-to-end trainable framework (Fig. 1) that utilizes a GAN to generate an explicit 3D representation given a single input image and a style. While our framework supports arbitrary choices of architectures, throughout our quantitative experiments, we integrate the methods of Wu et al. [66] and StarGAN-v2 [12]. We test the hypothesis whether our combined method achieves better performance compared to existing methods, including DepthNet [35] for (a) and Pix2NeRF [7] for (b), as our solution can compete in their respective benchmarks. We perform our quantitative and qualitative experiments on three facial datasets: CelebA-HQ [27], 3DFAW [19, 24, 70, 71] and Basel Face Model (BFM) [44]. Lastly, we also perform an ablation study to test the efficiencies of different components of our model. The source code enabling the reproduction of our results will be made publicly available.

2. Related Works

In this section, we review 2D to 3D reconstruction and style transfer methods. Let us first introduce a categorizations of the former methods.

Explicit techniques extract features of the 3D shape, which facilitate the recovery of the shape through an explicit rendering function. These features can then be postprocessed or used for downstream tasks. In contrast, *implicit* approaches do not yield an explicit 3D representation of the object, but are still capable of producing an implicit one that cannot be manipulated and used easily.

2.1. 2D to 3D Reconstruction

Shallow Explicit 3D Reconstruction Traditional methods such as Structure from Motion (SfM) [61] use geometric relationships between matching 2D keypoints in multiple views of the same rigid scene to recover the 3D structure. Non-Rigid SfM [5, 52] extends this to deformable objects, but it still requires 2D key point annotations for training and testing. In addition, other information such as shading [69] and symmetry [36, 59] have been proven to be useful for 3D recovery.

Deep Explicit 3D Reconstruction With the recent development of deep learning technology, superior 3D reconstruction methods have emerged, facilitating the learning of a suitable shape prior from sufficient training data in order to alleviate the ill-posedness of 3D recovery [45, 65]. *Supervised* methods map images to their paired target variables, which include: 3D ground truth (e.g., meshes [63], voxels [51], points clouds [37]), 3D shape models, video [1, 41, 54, 62, 73], keypoints [10, 26, 55], 2D silhouettes, stereo image pairs [17] or even just object classes [26, 58]. In contrast, *unsupervised* schemes learn from raw images only, such as: 3D lifting for bottleneck codes of deformation field autoencoders [47], adversarial 3D mesh reconstruction [56], or using a network with disentangled rendering features with symmetry regularization for face depth and albedo [66].

Deep Implicit 3D Reconstruction As mentioned above, implicit schemes only output an implicit function representation of the object that is not useful on its own, requiring further processing, e.g., conversion to explicit form (see Marching Cubes [33]) or alternative ways of rendering (e.g., deep network based 3D viewpoint to 2D image hallucination, including differentiable ray casting). Here, we specifically focus on novel 3D viewpoint synthesis, which is inherently *unsupervised*.

Generative Adversarial Networks. A Generative Adversarial Network (GAN) [18] consists of two competing components: a generator and a discriminator. The generator takes a given latent (and optionally conditional) distribution as input, outputs 'generated' examples, and tries make them indistinguishable from the 'real' training set for the discriminator. The two networks are trained in alternating manner, thus both are improving the other by making the task of the counterpart more and more difficult during training. Note, however, that mode dropping should be avoided: the generator has to produce diverse examples instead of just fooling the discriminator by replicating few training samples. GANs for viewpoint generation include image conditional 3D-aware networks [2, 16, 31, 39, 46, 53] that accept viewpoint features as input and use them to transform the hidden representation in 3D. 3D aware images can be generated unconditionally from noise as well [13, 38, 50, 72].

Neural Radiance Fields. Neural Radiance Field (NeRF) [34] exploits differentiable volume rendering to learn density and color, and they must be trained on sufficiently large single-scene multi-view image sets, limiting applicability.

Hybrid Architectures. Recent works incorporate a NeRF into their GAN architectures [7, 8, 14, 20, 40, 48]. Typically, the generator is a NeRF and the discriminator classifies its hallucinated 2D images. However, these procedures are restricted to scene-specific or category-specific implicit representations.

Our work differs from these schemes, since in contrast

to recovering the original 3D structure, we also consider changing the 3D style.

2.2. Style Transfer

2D Style Transfer GANs achieve remarkable performance on 2D image generation [3, 6, 68] and style transfer (sometimes also referred to as image-to-image translation) [12, 23, 28, 74]. Pix2pix [23] solves the problem using conditional adversarial networks. Cycle-GAN [74] uses a special cyclic architecture and uses cycle consistency losses to achieve remarkable performance. Karras et al. [28] achieve unsupervised separation of high-level attributes and scale-specific control of the synthesis via Adaptive Instance Normalization (AdaIN) [22, 28] at each convolutional layer. Choi et al. [12] proposed StarGAN-v2 consisting of four modules, which significantly improve the diversity of generated images under multiple domains.

Explicit 3D Style Transfer SofGAN [9] achieves explicit 3D reconstruction and transferring styles such as pose, shape, and texture. However, it is supervised: it requires semantic segmentation maps as ground truth, limiting applicability.

Implicit 3D Style Transfer HoloGAN [38] learns to synthesize 3D-aware images from unlabeled 2D images and disentangles shape and appearance. π-GAN [8] leverages a SIREN-based NeRF to get multi-view consistency and high quality images. Pix2NeRF [7] is conditioned on a single input image based on π-GAN and disentangles pose and content.

The approaches listed here are not unsupervised and explicit at the same time, whereas our proposed method can tackle this scenario.

3. Methods

3.1. Neural Network Architecture

Inspired by StarGAN-v2 [12], our network architecture consists of three components: a generator G (i.e., an unsupervised explicit 3D reconstruction network), a discriminator D, and a style network S. These facilitate recovering explicit 3D structure via rendering, promoting the generation of a combination of styles, and generating a specific style from random noise, respectively. The overall framework is illustrated in Fig. 2. We introduce each component separately.

Generator The generator G takes image I (and optionally style s) as input and outputs the image in input (or optionally target) style along with all domain specific styles (e.g., male/female, cat/dog). It can be seen as the combination of the Generator and the Style encoder in StarGAN-v2 [12], i.e., the function $G(I, s)$ is generating the style transferred image, however, in contrast to their work, we also train the Style encoder $E_s(I)$ jointly with the rest of G,

which queries the domain specific styles. It is worth noting that all styles mentioned in the paper are domain specific, but this does not mean that each domain has only one style. In each domain, the style is related to an image and different images have different styles. We denote domain specific style as style for brevity in the rest of the paper.[1]

Our generator G is an unsupervised 3D reconstruction network [66]. We summarize its architecture as follows. The setup starts with an *autoencoder backbone* and five network *heads*, which together output our five rendering features: the global light direction l, the viewpoint w, the albedo image a, the depth map d and the confidence map σ); which are then used as input to our *lighting* and *reprojection* blocks to output our style transferred image. First, given input image I, we use our encoder E and two heads to infer the hidden representation h, the style code $s = E_s(I)$ and two rendering features (l and w). Optionally, at this point, the style code s can be altered for further control over the styles (e.g., using our style network $S(z)$). Next, given h and s, we apply our decoder D and three heads to compute the other three rendering features (a, d and σ). Afterwards, given a, d and l, our lighting block computes the shaded canonical image J as seen from the canonical viewpoint. Finally, given d, w and J, our reprojection block changes the viewpoint from canonical to w. We explain the details of each block in the following sections. For further information see the supplementary material.

Autoencoder Backbone and Heads. Unlike the structure proposed in [66], which uses separate instances of E and D with untied weights, our method uses weight sharing between these networks to generate all rendering features along with style codes for each domain. We claim that the shared parameters can help learn useful intrinsic features that contribute to all outputs. However, we also include untied head branches on top of the shared autoencoder for each output (resulting in an overall partially shared architecture). To keep enough network capacity, we use convolutional layers in the decoder and in the network heads. We apply Adaptive Instance Normalization (AdaIN) [22, 28] in the decoder and inject style code s into all AdaIN layers to provide scaling and shifting vectors.

Lighting. The shaded canonical image J can be generated from a, d and l. First, we compute the normal map n from the depth map d by computing the 3D surface normal for each pixel (u, v), i.e., $n \equiv t^u \times t^v$, where t^u and t^v are the surface tangent vectors along the u and v directions, respectively. Then, we can get the shaded canonical image J as follows:

$$J_{uv} = a_{uv} \cdot \left[k_s + k_d \max\left(0, \langle l, n_{uv} \rangle\right) \right], \qquad (1)$$

where k_s and k_d are scalars representing the weights of the

---

[1]For more details on the differences and connections between domain and style, we kindly forward the reader to [12].
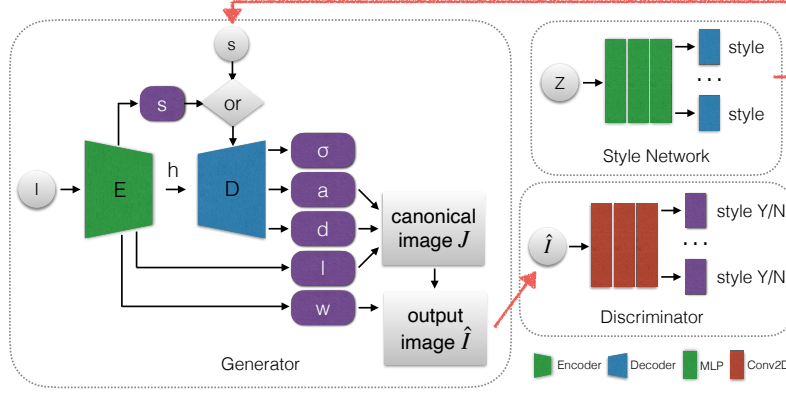
Figure 2. Schematic diagram of our unsupervised network architecture. The generator (left) is an autoencoder with heads and rendering blocks that takes as input a 2D image I (and optionally a style code s), infers five explicit 3D rendering features (confidence map σ, albedo a, depth d, light direction l and viewpoint w) and renders the 2D (optionally style transferred) image $\hat{I}$. The style network (right top) maps the latent noise z into style code s, promoting style diversity. The discriminator (right bottom) classifies whether the generated image is an instance of a style combination, and is fooled by the generator.

ambient and diffuse terms, respectively.[2]

Reprojection. We assume a perspective camera with relatively narrow field of view (FOV) of $\theta_{FOV} \approx 10°$, since the images are cropped accordingly. We also assume that the nominal distance between the object and the camera is $1\,\mathrm{m}$. Then mapping a 3D point $P \in \mathbb{R}^3$ from the reference frame of the camera to a pixel $p = (u, v, 1)$ can be expressed as:

$$p \equiv KP, \quad K = \begin{bmatrix} f & 0 & c_u \\ 0 & f & c_v \\ 0 & 0 & 1 \end{bmatrix}, \tag{2}$$

where $c_u = (W - 1)/2$, $c_v = (H - 1)/2$ and $f = (W - 1)/[2\tan(\theta_{FOV}/2)]$. Since the depth map d associates a depth value d_{uv} to each pixel at position (u, v) in the canonical view, we can get:

$$P = d_{uv} \cdot K^{-1} p. \tag{3}$$

The viewpoint $w \in \mathbb{R}^6$ is an Euclidean transformation. $w_{1:3}$ and $w_{4:6}$ represent the rotation angles and the translations along the x, y, z axes, respectively, which we denote by $(R, T) \in SE(3)$. Then, the warping transformation $A_{d,w}$ of pixels from the canonical view to the actual view is:

$$p' \equiv K(d_{uv}RK^{-1}p + T), \tag{4}$$

where $p' = (u', v', 1)$, and (u', v') is the pixel in the actual view mapped from pixel (u, v) in the canonical view. Thus, given a canonical image J, we output the image $\hat{I}$ as $\hat{I}_{u'v'} = J_{uv}$, where $(u, v) = A_{d,w}^{-1}(u', v')$.

Discriminator The discriminator D aims to determine whether the input image is a 'real' image of its styles or a

---

[2]Further details, e.g., the formulae for t^u, t^v and the light direction l can be found in [66].

'generated' image produced by G. It consists of N output branches, each of which is a binary classifier in each domain with output dimensions $K = 1$ for 'real'/'generated' classification. It is similar to the discriminator network in StarGAN-v2 [12].

Style Network The style network S is used to generate a style code s from the latent noise z as $s = S(z)$. S is a Multi-Layer Perceptron (MLP) with N output branches to provide style codes for all available domains, i.e., $s_y = S_y(z)$, where S_y outputs the style corresponding to the domain y. It is the same as the mapping network proposed in StarGAN-v2 [12]. By sampling the domain y and the latent noise z randomly, S can generate diverse style codes (as an alternative to obtaining the style code using the encoder as $s = E_s(I)$).

3.2. Training objectives

Reconstruction and Perceptual Losses Besides obtaining the output image $\hat{I}$ as above, we also compute a second output $\hat{I}'$ using the exact same procedure, except for using horizontally flipped depth $d' = flip(d)$ and horizontally flipped albedo $a' = flip(a)$. We then enforce $I \approx \hat{I}$ and $I \approx \hat{I}'$ using the following combination of losses:

$$\begin{aligned} \mathcal{L}_{rec} = &\big[\mathcal{L}(I, \hat{I}, \sigma) + \lambda_p \cdot \mathcal{L}_p(I, \hat{I}, \sigma)\big] \\ &+ \lambda_{rec} \cdot \big[\mathcal{L}(I, \hat{I}', \sigma') + \lambda_p \cdot \mathcal{L}_p(I, \hat{I}', \sigma')\big], \end{aligned} \tag{5}$$

where

$$\mathcal{L}(I, \hat{I}, \sigma) = -\frac{1}{|\Omega|} \sum_{uv \in \Omega} \ln\left[\frac{1}{\sqrt{2}\sigma_{uv}} \exp\left(-\frac{\sqrt{2}\ell_{1,uv}}{\sigma_{uv}}\right)\right], \tag{6}$$

$$\mathcal{L}_p(I, \hat{I}, \sigma) = -\frac{1}{|\Omega|} \sum_{uv \in \Omega} \ln\left[\frac{1}{\sqrt{2\pi}\sigma_{uv}} \exp\left(-\frac{\ell_{2,p,uv}^2}{2\sigma_{uv}^2}\right)\right] \tag{7}$$

are negative log-likelihood losses with factorized Laplacian and Gaussian priors, respectively, i.e., $\ell_{1,uv} = \|I_{uv} - \hat{I}_{uv}\|_1$ and $\ell_{2,p,uv} = \|VGG(I_{uv}) - VGG(\hat{I}_{uv})\|_2$, where VGG is the relu3_3 feature extractor of VGG16; λ_{rec}, λ_p are balancing factors; $\sigma, \sigma' \in \mathbb{R}^{W \times H}$ are the confidence maps estimating the level of asymmetry for each input pixel. This implicitly encourages $d \approx d'$, $a \approx a'$, i.e., symmetry in the canonical frame, as proposed in [66].

Adversarial Loss The generator G is trained to fool the discriminator D into thinking that the 'generated' image $\hat{I}$ is a 'real' instance of target style code $\tilde{s}$ via the following adversarial loss:

$$\mathcal{L}_{adv} = \mathbb{E}_I\left[\log D(I)\right] + \mathbb{E}_{I,z}\left[\log\left(1 - D\left[G(I,\tilde{s})\right]\right)\right]. \tag{8}$$

During training, the target style code $\tilde{s}$ is obtained either from the random noise (i.e., $\tilde{s} = S(z)$) or from the input image ($\tilde{s} = E_s(I)$).

Style Consistency We add a style consistency loss to enforce the generator G to be able to reconstruct the target style code $\tilde{s}$ from the generated the output image:

$$\mathcal{L}_{sty} = \mathbb{E}_{I,z}\left[\left\|\tilde{s} - E_s\left[G(I,\tilde{s})\right]\right\|_1\right]. \tag{9}$$

Intuitively, this will promote G to utilize $\tilde{s}$. Different from StarGAN-v2 [12], instead of using a separate encoder, we use the style encoder E_s of our generator G for obtaining the style code of the output image.

Source Image Consistency In line with recent works on adversarial learning [11, 12, 29, 75], we add another cycle consistency loss to enforce G to be able to recover the input image from its style transferred variant:

$$\mathcal{L}_{sou} = \mathbb{E}_{I,z}\left[\left\|I - G\left[G(I,\tilde{s}), E_s(I)\right]\right\|_1\right]. \tag{10}$$

Intuitively, this will promote preserving the original style-invariant characteristics of I. Note that the target style $\tilde{s}$ can be different from the inferred style of the input image $E_s(I)$.

Canonical Image Consistency In addition, we also enforce G to reconstruct the shaded canonical image $J = G_l(I)$ from itself using its canonical 3D reconstructor G_l:

$$\mathcal{L}_{can} = \mathbb{E}_I[\|G_l(I) - G_l(G_l(I))\|_1]. \tag{11}$$

Intuitively, this helps G to discover the correct canonical frontal view. During our experiments, we found that this also improves the stability of adversarial training.

Style Diversification Finally, we also add a style diversity loss to enforce G to make two style transferred versions of the same input image as different as possible:

$$\mathcal{L}_{sd} = -\mathbb{E}_{I,z_1,z_2}[\|G(I,\tilde{s}_1) - G(x,\tilde{s}_2)\|_1], \tag{12}$$

Table 1. Quantitative experimental results on the BFM dataset. We compare the methods in terms of the Scale-Invariant Depth Error (SIDE) [15] and Mean Angle Deviation (MAD) scores for 3D depth reconstruction, and their respective standard deviations. Winning numbers are highlighted in bold. †Numbers here for Wu et al. [66] are from our reproduction using CUDA 10.2, which is known to perform slightly worse than their published scores using CUDA 9.0.[3]

Method	Super-vised	SIDE ($\times 10^{-2}$) ↓	MAD (°) ↓
Wu et al. [66]†	yes	0.521±0.122	12.36±1.13
	no	0.892±0.175	16.98±1.07
Ours	yes	**0.481**±0.154	**10.64**±1.29
	no	**0.844**±0.117	**14.89**±1.33

where the target style codes $\tilde{s}_1$ and $\tilde{s}_2$ are either obtained from two random noise samples ($\tilde{s}_i = S(z_i), i \in \{1,2\}$) or from two images from the same domain ($\tilde{s}_i = E_s(I_i), i \in \{1,2\}$). Intuitively, minimizing this term forces G to explore and learn more diverse images even within the same style domain.

Full Objective The total loss function is as follows:

$$\min_{G,S} \max_D \ \mathcal{L}_{rec} + \mathcal{L}_{adv} + \lambda_{sty} \cdot \mathcal{L}_{sty} + \lambda_{sou} \cdot \mathcal{L}_{sou} \\ + \lambda_{can} \cdot \mathcal{L}_{can} + \lambda_{sd} \cdot \mathcal{L}_{sd}, \tag{13}$$

where λ_{sty}, λ_{sou}, λ_{can} and λ_{sd} are weighting factors.

4. Experiments

4.1. Datasets

We evaluated our method on three — two real and one synthetic — human face datasets: CelebA-HQ [27], 3DFAW [19, 24, 70, 71] and BFM [44]. CelebA-HQ is a high-quality version of CelebA [32], which is a large scale human face dataset and includes over 200k images of real human faces in the wild with bounding box annotations. CelebA-HQ consists of 30k images at 1024×1024 resolution. 3DFAW contains 23k images with 66 3D face keypoint annotations, which can be used for 3D face alignment and 3D facial landmark localization. Here, we used 3DFAW to evaluate our 3D prediction performance and compare with baselines at keypoint locations. Given that in-the-wild datasets lack ground-truth, we used the BFM model to generate a supervised dataset by sampling shapes, poses, textures, and illumination randomly according to the protocol of [49]. We also used images from SUN Database [67] as background and obtain ground truth depth maps for evaluation as proposed in [66].

---

[3]See https://github.com/elliottwu/unsup3d/issues/15 for details.

4.2. Metrics

We evaluated our method on two problems related to UE3DST, single view 2D to 3D reconstruction and 2D style transfer. We hypothesize that solving both tasks simultaneously regularizes each other.

For 3D reconstruction, we assessed 2D depth map recovery by applying Scale-Invariant Depth Error (SIDE) [15], Mean Angle Deviation (MAD) and Sum of Depth Correlations (SDC) [35]. SIDE can be written as:

$$E_{SIDE}(\bar{d}, d^*) = \sqrt{\frac{1}{WH} \sum_{uv} \left[\Delta_{uv}^2 - \left(\frac{\sum_{uv} \Delta_{uv}}{WH} \right)^2 \right]},$$
$$\Delta_{uv} = \log \bar{d}_{uv} - \log d^*_{uv}, \tag{14}$$

where d^* is the ground-truth depth map and $\bar{d}$ is the actual view depth map warped from predicted depth map d in the canonical view using the predicted viewpoint. MAD compares the angles between normals n^* and n computed from d^* and d, respectively, as in [66]. For keypoint depth evaluation, we computed the SDC score between 66 ground truth and predicted frontal facial keypoint locations (i.e., the score is between 0 and 66) as in [35]. For image synthesis evaluation, we used Fréchet Inception Distance (FID) [21] and Kernel Inception Distance (KID) [4].

4.3. Implementation Details

We split CelebA-HQ into two domains: male and female. For 3DFAW, we roughly cropped images around the head region and used the official training/validation/test splits. We resized images in all datasets to 64×64 in order to fit our method onto the GPUs available to us. All the experiments were conducted on a 11 GB NVIDIA GeForce RTX 2080Ti graphics card with CUDA 10.2 and PyTorch 1.4.0 [43]. We used Adam optimizer [30] with learning rate 1×10^{-4}, $\beta_1 = 0.0$, $\beta_2 = 0.99$ and weight decay 1×10^{-4}. We set $\lambda_{rec} = 0.5$, $\lambda_p = 1$, $\lambda_{sty} = 0.5$, $\lambda_{sou} = 0.5$, $\lambda_{can} = 0.3$, $\lambda_{sd} = 0.5$ for all experiments. We used batch size 16 and train 100k iterations in total. To stabilize the adversarial training process, first, we trained the generator G alone for the first 20k iterations, and then trained the whole network jointly for the remaining 80k iterations.

4.4. Results

We used the BFM dataset to compare the 2D depth map reconstruction quality obtained by our method with the baseline of Wu et al. [66], as well as their respective supervised rendering-free equivalents that regress the ground-truth depth maps directly using an ℓ_1 loss. We retrained

Table 2. Quantitative experimental results on the 3DFAW dataset. We compare the methods in terms of the Sum of Depth Correlations (SDC) [35] score for keypoint depth. Winning numbers are highlighted in bold for each category. [†]Numbers here for Wu et al. [66] are from our reproduction using CUDA 10.2, which is known to perform slightly worse than their published scores using CUDA 9.0.[4] Besides, the numbers may further differ due to using our custom implementation for evaluation.[5]

Method	Supervised	SDC ↑
Ground truth	N/A	66.00
AIGN [60]	yes [35]	50.81
DepthNetGAN [35]	yes [35]	**58.68**
MOFA [57]	yes, model-based [35]	15.97
DepthNet [35]	no, trained on [35]	26.32
DepthNet [35]	no	35.77
Wu et al. [66][†]	no	41.52
Ours	no	**44.86**

Table 3. Quantitative results on the CelebA-HQ 64×64 dataset. We compare the methods in terms of the Fréchet Inception Distance (FID) [21] and Kernel Inception Distance (KID) [4] scores for image synthesis. Winning numbers are highlighted in bold.

Method	3D	Img. Cond.	FID ↓	KID ($\times 100$) ↓
HoloGAN [38]	aware	no	N/A	2.87
π-GAN [8]	NeRF	no	**5.15**	**0.09**
Pix2NeRF [7]	NeRF	no	6.25	0.16
Pix2NeRF [7]	NeRF	yes	24.64	1.93
Ours	explic.	yes	**8.19**	**0.78**

Table 4. Quantitative ablation study results on the BFM dataset. We compare the methods in terms of the Scale-Invariant Depth Error (SIDE) [15] and Mean Angle Deviation (MAD) scores for 3D depth reconstruction, and their respective standard deviations. Winning numbers are highlighted in bold.

Method	SIDE ($\times 10^{-2}$) ↓	MAD (°) ↓
Ours, w/o albedo flip	3.148±0.322	41.21±2.01
Ours, w/o depth flip	1.463±0.241	29.87±2.10
Ours, w/o conf. map	0.885±0.131	15.16±1.23
Ours, full model	**0.844±0.117**	**14.89±1.33**

the baseline model from scratch using CUDA 10.2, since their published results were obtained with CUDA 9.0. The results are shown in Table 1. We can see that our method outperforms the baseline, which confirms that our shared encoder-decoder and adversarial losses together yield a better instance specific 3D representation.

Next, we compared the 3D keypoint depth reconstruc-

Figure 3. Reference-guided image synthesis results on CelebA-HQ. First row is source images and first column is reference images. We can see that our method can generate realistic images with the style of the reference image based on the source image.

Figure 4. Latent noise-guided image synthesis results on CelebA-HQ. First row is source images and we generate z randomly as input. We can see that our method can generate realistic images by controlling the style through the latent space.

tion quality with baselines Wu et al. [66], AIGN [60], MOFA [57] and DepthNet [35]. Note that AIGN, MOFA and DepthNet each have unfair advantage over our method, as the former two are supervised, whereas the latter takes 2D keypoint locations as input. For all algorithms except DepthNet, we the computed the full 2D depth map d and sampled it at keypoint locations, in order to compute the SDC score against the ground truth. We reimplemented the evaluation code based upon [66], since they did not release this part of their code, causing further potential mismatch between our replicated and their published performance, besides CUDA version differences. The results are collected in Table 2. We found that our model performs much better than other model-based or unsupervised methods, further confirming the advantages of using a tied weights autoencoder and style transfer regularization.

Lastly, for image synthesis quality evaluation, we computed FID and KID on the CelebA-HQ dataset, and compared with 3D aware and NeRF-integrated GAN baselines: HoloGAN, π-GAN and two variants of Pix2NeRF. Again,

note that some of these baselines are image unconditional, having an unfair advantage over our conditional procedure, since feeding image inputs limits the variability and diversity of output images, affecting both the FID and KID scores. The results are summarized in Table 3. We can see that our method outperformed the image conditional Pix2NeRF by a large margin; and is approaching the level of unconditional methods, while having explicit 3D reconstruction ability. We also provide qualitative examples to visualize our image synthesis results using style codes s both from reference images (Fig. 3) and from noise vectors z (Fig. 4). Fig. 5 depicts the recovered 3D structure of some style transferred images, which are spectacular for inferring from a single viewpoint only.

4.5. Ablation Study

We performed ablation experiments on the BFM dataset to confirm that the model exploits symmetry/asymmetry information and canonical consistency properly.

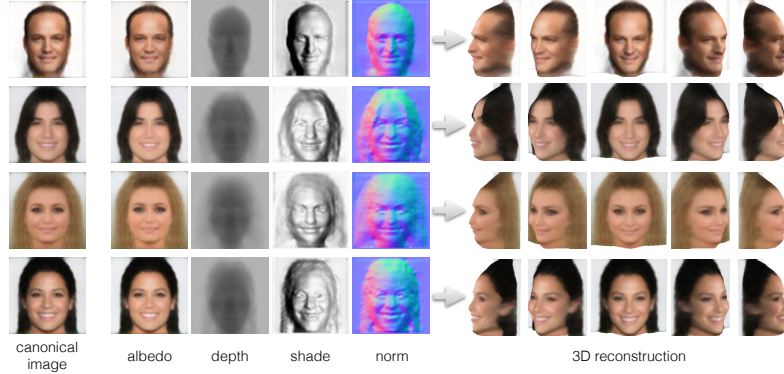First, we quantitatively evaluated the efficiency of the

Figure 5. 3D reconstruction from single view image visualization. Our method can effectively recover the 3D structure of style-transferred images.

Figure 6. Ablation study on canonical image consistency. We found that canonical image consistency plays a crucial role in the quality of the generated images.

flip operation and the confidence map, as in [66]. As shown in Table 4, we can see that performance degrades drastically without applying flipping on depth or albedo map. The lack of a confidence map affected the scores less, but was still worse than our full model.

Second, we show qualitative results for our method without canonical image consistency loss in Fig. 6. We can see that the model is not able learn good frontal view canonical images without the canonical image consistency loss, which confirms the rationality and necessity of this term.

5. Conclusions

In this paper, we proposed an end-to-end unsupervised network for 2D to explicit 3D style transfer. The method combines existing approaches for two related problems: the ill-posedness of 2D to 3D reconstruction is alleviated by utilizing albedo and depth symmetry, whereas adversarial training is stabilized by our cycle consistency losses. Our quantitative and qualitative experiments showed that our scheme achieves performance better or comparable to prior works on the two tasks, while solving both at the same time. Future work may consider using more powerful GAN architectures, richer shading models (e.g., shadows, specularity), more complex objects and priors beyond faces and symmetry (e.g., multiple canonical views) and integration with other explicit 3D representations (e.g., meshes).

References

[1] Pulkit Agrawal, Joao Carreira, and Jitendra Malik. Learning to see by moving. In *Proceedings of the IEEE international conference on computer vision*, pages 37–45, 2015. 2

[2] Hiroaki Aizawa, Hirokatsu Kataoka, Yutaka Satoh, and Kunihito Kato. agnostic image rendering. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3803–3812, 2021. 2

[3] David Berthelot, Thomas Schumm, and Luke Metz. Began: Boundary equilibrium generative adversarial networks. *arXiv preprint arXiv:1703.10717*, 2017. 3

[4] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018. 6

[5] Christoph Bregler, Aaron Hertzmann, and Henning Biermann. Recovering non-rigid 3d shape from image streams. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No. PR00662)*, volume 2, pages 690–696. IEEE, 2000. 2

[6] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 3

[7] Shengqu Cai, Anton Obukhov, Dengxin Dai, and Luc Van Gool. Pix2nerf: Unsupervised conditional pi-gan for single image to neural radiance fields translation. *arXiv preprint arXiv:2202.13162*, 2022. 2, 3, 6

[8] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5799–5809, 2021. 2, 3, 6

[9] Anpei Chen, Ruiyang Liu, Ling Xie, Zhang Chen, Hao Su, and Jingyi Yu. Sofgan: A portrait image generator with dynamic styling. *ACM Transactions on Graphics (TOG)*, 41(1):1–26, 2022. 2, 3

[10] Ching-Hang Chen, Ambrish Tyagi, Amit Agrawal, Dylan Drover, Stefan Stojanov, and James M Rehg. Unsupervised 3d pose estimation with geometric self-supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5714–5724, 2019. 2

[11] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797, 2018. 5

[12] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8188–8197, 2020. 2, 3, 4, 5

[13] Yu Deng, Jiaolong Yang, Dong Chen, Fang Wen, and Xin Tong. Disentangled and controllable face image generation via 3d imitative-contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5154–5163, 2020. 2

[14] Yu Deng, Jiaolong Yang, Jianfeng Xiang, and Xin Tong. Gram: Generative radiance manifolds for 3d-aware image generation. *arXiv preprint arXiv:2112.08867*, 2021. 2

[15] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*, 27, 2014. 5, 6

[16] Jiahao Geng, Tianjia Shao, Youyi Zheng, Yanlin Weng, and Kun Zhou. Warp-guided gans for single-photo facial animation. *ACM Transactions on Graphics (TOG)*, 37(6):1–12, 2018. 2

[17] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 270–279, 2017. 2

[18] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 2

[19] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-pie. *Image and vision computing*, 28(5):807–813, 2010. 2, 5

[20] Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. Stylenerf: A style-based 3d-aware generator for high-resolution image synthesis. *arXiv preprint arXiv:2110.08985*, 2021. 2

[21] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 6

[22] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017. 3

[23] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *CVPR*, 2017. 3

[24] László A Jeni, Jeffrey F Cohn, and Takeo Kanade. Dense 3d face alignment from 2d videos in real-time. In *2015 11th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, volume 1, pages 1–8. IEEE, 2015. 2, 5

[25] Yong Ju Jung, Aron Baik, Jiwon Kim, and Dusik Park. A novel 2d-to-3d conversion technique based on relative height-depth cue. In *Stereoscopic Displays and Applications XX*, volume 7237, pages 589–596. SPIE, 2009. 1

[26] Angjoo Kanazawa, Shubham Tulsiani, Alexei A Efros, and Jitendra Malik. Learning category-specific mesh reconstruction from image collections. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 371–386, 2018. 1, 2

[27] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 2, 5

[28] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks.

In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 3

[29] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. In *International conference on machine learning*, pages 1857–1865. PMLR, 2017. 5

[30] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[31] Jean Kossaifi, Linh Tran, Yannis Panagakis, and Maja Pantic. Gagan: Geometry-aware generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 878–887, 2018. 2

[32] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015. 5

[33] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. *ACM siggraph computer graphics*, 21(4):163–169, 1987. 2

[34] 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 *European conference on computer vision*, pages 405–421. Springer, 2020. 2

[35] Joel Ruben Antony Moniz, Christopher Beckham, Simon Rajotte, Sina Honari, and Chris Pal. Unsupervised depth estimation, 3d face rotation and replacement. *Advances in neural information processing systems*, 31, 2018. 2, 6, 7

[36] Dipti Prasad Mukherjee, Andrew Peter Zisserman, Michael Brady, and FT Smith. Shape from symmetry: Detecting and exploiting symmetry in affine images. *Philosophical Transactions of the Royal Society of London. Series A: Physical and Engineering Sciences*, 351(1695):77–106, 1995. 1, 2

[37] KL Navaneet, Ansu Mathew, Shashank Kashyap, Wei-Chih Hung, Varun Jampani, and R Venkatesh Babu. From image collections to point clouds with self-supervised shape and pose networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1132–1140, 2020. 2

[38] Thu Nguyen-Phuoc, Chuan Li, Lucas Theis, Christian Richardt, and Yong-Liang Yang. Hologan: Unsupervised learning of 3d representations from natural images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7588–7597, 2019. 2, 3, 6

[39] Thu H Nguyen-Phuoc, Chuan Li, Stephen Balaban, and Yongliang Yang. Rendernet: A deep convolutional network for differentiable rendering from 3d shapes. *Advances in Neural Information Processing Systems*, 31, 2018. 2

[40] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11453–11464, 2021. 2

[41] David Novotny, Diane Larlus, and Andrea Vedaldi. Learning 3d object categories by looking around them. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5218–5227, 2017. 2

[42] David Novotny, Ignacio Rocco, Samarth Sinha, Alexandre Carlier, Gael Kerchenbaum, Roman Shapovalov, Nikita Smetanin, Natalia Neverova, Benjamin Graham, and Andrea Vedaldi. Keytr: Keypoint transporter for 3d reconstruction of deformable objects in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5595–5604, 2022. 1

[43] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 6

[44] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3d face model for pose and illumination invariant face recognition. In *2009 sixth IEEE international conference on advanced video and signal based surveillance*, pages 296–301. Ieee, 2009. 2, 5

[45] Anurag Ranjan, Timo Bolkart, Soubhik Sanyal, and Michael J Black. Generating 3d faces using convolutional mesh autoencoders. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 704–720, 2018. 2

[46] Helge Rhodin, Mathieu Salzmann, and Pascal Fua. Unsupervised geometry-aware representation for 3d human pose estimation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 750–767, 2018. 2

[47] Mihir Sahasrabudhe, Zhixin Shu, Edward Bartrum, Riza Alp Guler, Dimitris Samaras, and Iasonas Kokkinos. Lifting autoencoders: Unsupervised learning of a fully-disentangled 3d morphable model using deep non-rigid structure from motion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 2

[48] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. Graf: Generative radiance fields for 3d-aware image synthesis. *Advances in Neural Information Processing Systems*, 33:20154–20166, 2020. 2

[49] Soumyadip Sengupta, Angjoo Kanazawa, Carlos D Castillo, and David W Jacobs. Sfsnet: Learning shape, reflectance and illuminance of facesin the wild'. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6296–6305, 2018. 5

[50] Yichun Shi, Divyansh Aggarwal, and Anil K Jain. Lifting 2d stylegan for 3d-aware face generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6258–6266, 2021. 2

[51] Daeyun Shin, Charless C Fowlkes, and Derek Hoiem. Pixels, voxels, and views: A study of shape representations for single view 3d object shape prediction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3061–3069, 2018. 2

[52] Vikramjit Sidhu, Edgar Tretschk, Vladislav Golyanik, Antonio Agudo, and Christian Theobalt. Neural dense non-rigid structure from motion with latent space constraints. In *European Conference on Computer Vision*, pages 204–222. Springer, 2020. 2

[53] Vincent Sitzmann, Justus Thies, Felix Heide, Matthias Nießner, Gordon Wetzstein, and Michael Zollhofer. Deepvoxels: Learning persistent 3d feature embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2437–2446, 2019. 2

[54] Jiaming Sun, Yiming Xie, Linghao Chen, Xiaowei Zhou, and Hujun Bao. Neuralrecon: Real-time coherent 3d reconstruction from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15598–15607, 2021. 2

[55] Supasorn Suwajanakorn, Noah Snavely, Jonathan J Tompson, and Mohammad Norouzi. Discovery of latent 3d keypoints via end-to-end geometric reasoning. *Advances in neural information processing systems*, 31, 2018. 2

[56] Attila Szabó, Givi Meishvili, and Paolo Favaro. Unsupervised generative 3d shape learning from natural images. *arXiv preprint arXiv:1910.00287*, 2019. 2

[57] 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 *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 1274–1283, 2017. 6, 7

[58] James Thewlis, Hakan Bilen, and Andrea Vedaldi. Modelling and unsupervised learning of symmetric deformable object categories. *Advances in Neural Information Processing Systems*, 31, 2018. 2

[59] Sebastian Thrun and Ben Wegbreit. Shape from symmetry. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 2, pages 1824–1831. IEEE, 2005. 1, 2

[60] Hsiao-Yu Fish Tung, Adam W Harley, William Seto, and Katerina Fragkiadaki. Adversarial inverse graphics networks: Learning 2d-to-3d lifting and image-to-image translation from unpaired supervision. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 4364–4372. IEEE, 2017. 6, 7

[61] Shimon Ullman. The interpretation of structure from motion. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 203(1153):405–426, 1979. 2

[62] Chaoyang Wang, José Miguel Buenaposada, Rui Zhu, and Simon Lucey. Learning depth from monocular videos using direct methods. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2022–2030, 2018. 2

[63] Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2mesh: Generating 3d mesh models from single rgb images. In *Proceedings of the European conference on computer vision (ECCV)*, pages 52–67, 2018. 2

[64] Yan Wang, Wei-Lun Chao, Divyansh Garg, Bharath Hariharan, Mark Campbell, and Kilian Q Weinberger. Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8445–8453, 2019. 1

[65] Jiajun Wu, Chengkai Zhang, Tianfan Xue, Bill Freeman, and Josh Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. *Advances in neural information processing systems*, 29, 2016. 2

[66] Shangzhe Wu, Christian Rupprecht, and Andrea Vedaldi. Unsupervised learning of probably symmetric deformable 3d objects from images in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1–10, 2020. 1, 2, 3, 4, 5, 6, 7, 8

[67] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE, 2010. 5

[68] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *International conference on machine learning*, pages 7354–7363. PMLR, 2019. 3

[69] Ruo Zhang, Ping-Sing Tsai, James Edwin Cryer, and Mubarak Shah. Shape-from-shading: a survey. *IEEE transactions on pattern analysis and machine intelligence*, 21(8):690–706, 1999. 1, 2

[70] Xing Zhang, Lijun Yin, Jeffrey F Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, and Peng Liu. A high-resolution spontaneous 3d dynamic facial expression database. In *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, pages 1–6. IEEE, 2013. 2, 5

[71] Xing Zhang, Lijun Yin, Jeffrey F Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, Peng Liu, and Jeffrey M Girard. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*, 32(10):692–706, 2014. 2, 5

[72] 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. *arXiv preprint arXiv:2010.09125*, 2020. 2

[73] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1851–1858, 2017. 2

[74] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017. 3

[75] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 5