3D-Aware Generative Model for Improved Side-View Image Synthesis

Kyungmin Jo[1,*] Wonjoon Jin[2,*] Jaegul Choo[1] Hyunjoon Lee[3] Sunghyun Cho[2]

[1]KAIST
Daejeon, Korea
{bttkm, jchoo}@kaist.ac.kr

[2]POSTECH
Pohang, Gyeongbuk, Korea
{jinwj1996, s.cho}@postech.ac.kr

[3]Kakao Brain
Seongnam-si, Gyeonggi-do, Korea
malfo.lee@kakaobrain.com

Figure 1: Our method robustly produces high-quality images of human faces, regardless of the camera pose while the baselines (π-GAN [2] and EG3D [1]) generate blurry images at the steep pose. The images are rendered with horizontal rotation from the frontal view to the side view.

Abstract

While recent 3D-aware generative models have shown photo-realistic image synthesis with multi-view consistency, the synthesized image quality degrades depending on the camera pose (e.g., a face with a blurry and noisy boundary at a side viewpoint). Such degradation is mainly caused by the difficulty of learning both pose consistency and photo-realism simultaneously from a dataset with heavily imbalanced poses. In this paper, we propose SideGAN, a novel 3D GAN training method to generate photo-realistic images irrespective of the camera pose, especially for faces of side-view angles. To ease the challenging problem of learning photo-realistic and pose-consistent image synthesis, we split the problem into two subproblems, each of which can be solved more easily. Specifically, we formulate the problem as a combination of two simple discrimination problems, one of which learns to discriminate whether a synthesized image looks real or not, and the other learns to discriminate whether a synthesized image agrees with the camera pose. Based on this, we propose a dual-branched discriminator with two discrimination branches. We also propose a pose-matching loss to learn the pose consistency of 3D GANs. In addition, we present a pose sampling strat-

*Both authors contributed equally to this research. Also, this work was done during an internship at Kakao Brain.

egy to increase learning opportunities for steep angles in a pose-imbalanced dataset. With extensive validation, we demonstrate that our approach enables 3D GANs to generate high-quality geometries and photo-realistic images irrespective of the camera pose.

1. Introduction

Generative Adversarial Networks (GANs) [9] have shown remarkable success in photo-realistic image generation [13, 14] by learning the distributions of high-resolution image datasets. Recent studies have taken this success one step further by extending GANs to pose-controllable image generation based on the guidance of a 3DMM prior [25, 5] or a differentiable renderer [28]. However, they produce inconsistent results across different poses and also suffer from limited pose controllability as they learn to generate 2D images for different poses independently without considering the 3D face structure.

Therefore, 3D-aware GANs have emerged to achieve multi-view consistent image generation. Recent studies [19, 2, 10, 27, 1, 23, 17] have tackled this problem by modeling the 3D structure of a face using neural radiance fields [16], enabling explicit view control. Combining volumetric feature projection with convolutional neural networks (CNNs) enables 3D GANs to generate photo-realistic face images in high resolution [10, 18, 1]. Albeit their ability to synthesize photo-realistic images with explicit view control, their results do not have a stable quality depending on the camera pose (Fig. 1). To be specific, side-view facial images generated by such methods show degraded qualities compared to photo-realistic images of frontal viewpoints (*e.g.,* a blurry and a noisy facial boundary).

This unstable image quality is caused by the challenge for 3D-aware GANs to simultaneously learn to generate *pose-consistent* and *photo-realistic* images from a pose-imbalanced dataset (Fig. 2) such as the FFHQ dataset [13] where most images are frontal-view images. Specifically, EG3D [1], the state-of-the-art 3D GAN approach, formulates the problem as a learning problem of a pose-conditional distribution of real images. Unfortunately, learning the distribution of real images for each pose can be extremely challenging, especially for poses with only a small number of real images. GRAM [6] casts the problem as a combination of the learning of real/fake image discrimination and pose estimation. Nevertheless, pose estimation from degraded side-view images is not trivial to learn either. As a result, images generated by the existing 3D GANs are blurry or have noisy boundaries in the face region at steep pose angles (Fig. 1).

To tackle this problem, we propose SideGAN, a novel 3D GAN training method to generate photo-realistic images irrespective of the viewing angle. Our key idea is as follows. To ease the challenging problem of learning photo-realistic and multi-view consistent image synthesis, we split the problem into two subproblems, each of which can be solved more easily. Specifically, we formulate the problem as a combination of two simple discrimination problems, one of which learns to discriminate whether a synthesized image looks real or not, and the other learns to discriminate whether a synthesized image agrees with the camera pose. Unlike the formulations of the previous methods, which try to learn the real image distribution for each pose, or to learn pose estimation, our subproblems are much easier as each of them is analogous to a basic binary classification problem.

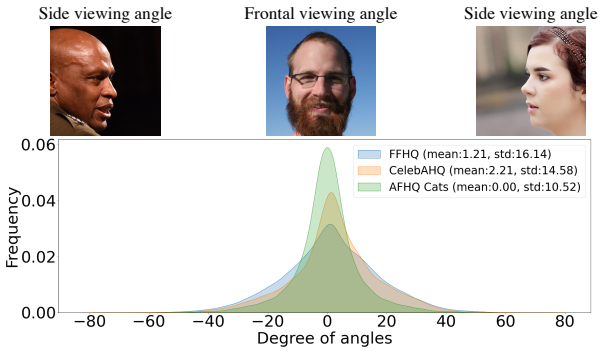

Figure 2: Real-world face datasets generally have an imbalanced pose distribution, which is mainly concentrated on the frontal viewpoint.

Based on this key idea, we propose a *dual-branched discriminator*, which has two branches for learning photo-realism and pose consistency, respectively. As these branches are supervised explicitly for their respective purposes, high-quality images with pose consistency can be produced at each viewing angle, and consequently, the generator creates high-quality images and shapes. In addition, we propose a *pose-matching loss* to give supervision to the discriminator for the pose consistency, by considering a positive pose (*i.e.*, rendering pose or ground truth pose) and a negative pose (*i.e.*, irrelevant pose) for a given image. For example, the frontal viewpoint is one of the irrelevant poses for a side-view image. As reported in the experiments, this loss helps improve image and shape quality. Compared to the previous pose estimation strategy [6], our pose-matching loss provides a more effective way to learn pose-consistent image generation, as the pose-matching loss casts the learning of pose-consistent image generation as the learning of simple binary classification that is much easier than the learning of accurate pose regression.

Additionally, we suggest a simple but effective training strategy to alleviate the degradation caused by insufficient semantic knowledge at steep poses in a pose-imbalanced dataset. As shown in Fig. 2, most in-the-wild face datasets [13, 12, 3] usually have pose distributions concentrated on the frontal angle, causing the degradation of generated images at steep poses. While we may con-

struct a pose-balanced dataset in a controlled environment, it requires a significant amount of effort, and is also hard to guarantee the diversity like in the in-the-wild datasets. Instead, we present an additional uniform pose sampling (AUPS) strategy that draws camera poses from both a uniform distribution and the actual camera pose distribution to enhance learning opportunities for steep angles during training. Our experiments show that this simple pose sampling strategy substantially improves the generation quality for side-view images.

Our contributions are summarized as follows:

- We split the problem of learning of 3D GANs into two easier subproblems: real/fake image discrimination and pose-consistency discrimination.
- We propose a dual-branched discriminator and a pose-matching loss to effectively learn the pose consistency by considering both positive and negative poses of a given image.
- We also present a simple but effective pose sampling strategy to compensate for the insufficient amount of side-view images in pose-imbalanced in-the-wild datasets.
- With extensive evaluations, SideGAN shows the state-of-the-art image and shape quality irrespective of the camera pose, especially at steep view angles.

2. Related work

Extending 2D GANs to have pose controllability. GANs [9] have achieved significant success in photo-realistic 2D image generation [13, 14]. Extending 2D GANs to provide pose controllability has been addressed by disentangling 3D information from GAN's latent space. Finding meaningful directions for editing pose in the latent space can be done with supervision from pre-trained classifiers [20] or in an unsupervised manner [21]. Editing the camera pose can be implemented by disentangling the pose factor from the latent space with guidance from a 3DMM prior [25, 5]. Zhang et al. [28] utilize inverse graphics with a differentiable renderer for pose-controllable image generation by fine-tuning StyleGAN to have disentangled pose attributes. Shi et al. [7] exploit a depth prior to disentangle the latent codes of geometry and appearance for RGBD generation with pose controllability. Unfortunately, these studies based on 2D GANs fundamentally lack multi-view consistency or accurate pose controllability since they do not consider the 3D structure of faces.

3D-aware GANs. Recent work incorporating neural 3D representations into GANs enables multi-view consistent image generation with explicit camera control. GRAF [19] and π-GAN [2] adopt fully implicit volumetric fields with differentiable volumetric rendering for 3D scene generation. However, these methods suffer from a large memory burden due to fully implicit networks, restricting image resolution and expressiveness. To enable high-resolution image synthesis, GRAM [6] restricts point sampling to regions near the learned implicit surface. StyleNeRF [10], StyleSDF [18] and GIRAFFE [17] combine CNN-based upsamplers with volumetric feature projection in their multi-view consistent image generation. EG3D [1], which is the most recent and related to our work, achieves photo-realistic image synthesis based on their tri-plane representation and StyleGAN feature generator. While previous 3D GAN studies have made significant progress in 3D-aware image synthesis, they have a limitation that the image quality degrades as the viewpoint shifts from frontal angles to steeper angles. To the best of our knowledge, our work is the first one to tackle the ineffectiveness of training 3D GANs from a pose-imbalanced dataset for photo-realistic multi-view consistent image generation irrespective of the camera pose.

3. SideGAN Framework

Our framework generates photo-realistic images irrespective of the camera pose even though most images in the training dataset are frontal-view images. As shown in Fig. 3, the main architecture is composed of two components. The first component is a generator G_θ for generating images from latent vectors $\mathbf{z}_{fg}$ and $\mathbf{z}_{bg}$ for the foreground and background regions, respectively, and a rendering camera parameter $\boldsymbol{\xi}^+$. The second component is a dual-branched discriminator D_ϕ for discriminating a generated image $\hat{\mathbf{I}}$ from a real image $\mathbf{I}$ and for discriminating whether a generated image agrees with a camera pose $\boldsymbol{\xi}$. In the following, we describe each component. More details on our framework are provided in Sec. B.

3.1. Generator

Existing 3D GAN models [1, 2] mainly render the background and foreground together by a single network. This causes the 3D structures in the background region to mingle with the 3D structures in the foreground region and makes it difficult to create photo-realistic side-view images (Fig. 4). To address this issue, we design our generator G_θ to separately produce the foreground and background regions to avoid mingled foreground and background structures. Specifically, our generator G_θ is composed of two components: an image generator and a background network, inspired by EpiGRAF [23]. The image generator has two roles: it produces features for the foreground region (*i.e.,* the facial region), and produces a final high-resolution image using both foreground and background features. Meanwhile, the background network produces features for the background region, which are used by the image generator.

For the image generator, we adopt the generator of a state-of-the-art 3D GAN model [1]. The image generator forms tri-plane features from the latent code $\mathbf{z}_{fg}$ and the

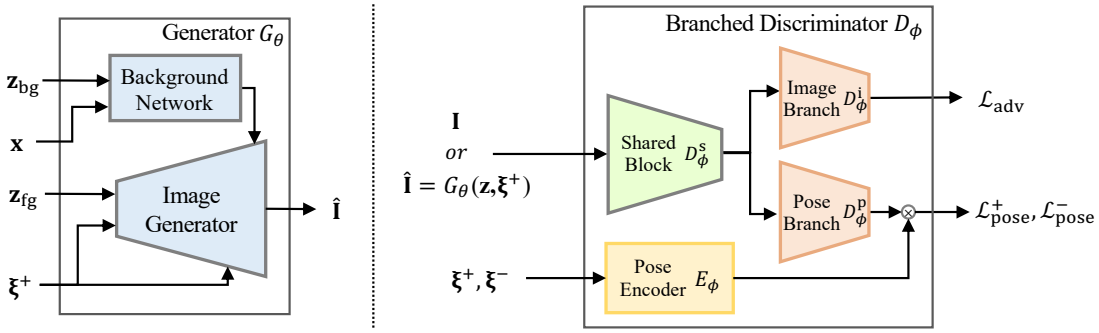

Figure 3: Illustration of main architecture. The generator takes latent codes $\mathbf{z}_{fg}$ and $\mathbf{z}_{bg}$, camera parameters $\boldsymbol{\xi}^+$, and 3D position $\mathbf{x}$ as inputs and synthesizes an image $\hat{\mathbf{I}}$. The dual-branched discriminator takes either a real image $\mathbf{I}$ or a generated image $\hat{\mathbf{I}}$ and camera parameters $\boldsymbol{\xi}$ and outputs separably logits for image distribution and image-pose consistency.

camera parameter $\boldsymbol{\xi}^+ \in \mathbb{R}^{25}$. Then, the generator samples 3D positions according to $\boldsymbol{\xi}^+$, then obtains features for the sampled positions from the tri-plane. After that, the foreground feature maps are obtained through a decoder and volume rendering. Then, the feature maps are integrated with the background feature maps according to the transmittance of the foreground to generate a low-resolution feature map. Finally, a high-resolution image is obtained from the low-resolution features through a super-resolution module in the image generator.

For the background network, we adopt the background network of EpiGRAF [23]. The background network is a multi-layer perceptron (MLP) that takes a latent code $\mathbf{z}_{bg}$ and a 3D position $\mathbf{x}$ as inputs and outputs a feature vector. To generate background features, we first sample 3D positions according to the camera pose $\boldsymbol{\xi}^+$, and feed them to the background network to obtain feature vectors for the sampled 3D positions. After aggregating all the background features, we feed them to the image generator.

3.2. Dual-Branched Discriminator

As shown in Fig. 3, the dual-branched discriminator D_ϕ takes an image and a camera pose as inputs. The input pose can be either positive ($\boldsymbol{\xi}^+$) or negative ($\boldsymbol{\xi}^-$), where a positive pose means that the pose agrees with the input image, while a negative pose means it does not. From the inputs, the discriminator predicts whether the input image is real or fake, and whether the input image agrees with the input camera pose using two output branches.

The discriminator D_ϕ comprises four components: a shared block D^{s}_ϕ, a pose encoder E_ϕ, an image branch D^{i}_ϕ, and a pose branch D^{p}_ϕ. The shared block extracts features from an input image, which will be used by the image and pose branches, while the pose encoder E_ϕ projects an input camera parameter $\boldsymbol{\xi}$ to an embedding space. The image branch D^{i}_ϕ predicts whether the input image is real or fake using the output of the shared block D^{s}_ϕ. The pose branch D^{p}_ϕ extracts pose features of the input image from the output of shared block D^{s}_ϕ, which are then combined with the features from the pose encoder to discriminate whether the input image agrees with the input camera pose.

4. Training for a Wider Range of Angles

In this section, we describe our training strategy including the pose-matching loss and AUPS.

4.1. Pose-Matching Loss

To promote pose consistency between the input pose to the generator and its corresponding synthesized image, the pose-matching loss is computed between a pair of an image and a camera pose. The pose-matching loss considers both positive and negative pairs of an image and a camera pose to more strongly guide the generator to produce pose-consistent images. In the case of a positive pair whose image and camera pose are supposed to agree with each other, the pose-matching loss penalizes the generator if the image does not agree with the pose. On the other hand, in the case of a negative pair whose image and camera pose are supposed to not agree, the pose-matching loss penalizes the generator if the image agrees with the pose.

Formally, we define the pose-matching loss $\mathcal{L}^{\text{gen}}_{\text{pose}}$ for the generator as:

$$\begin{aligned}\mathcal{L}^{\text{gen}}_{\text{pose}}(\theta) &= \mathcal{L}^{\text{gen},+}_{\text{pose}}(\theta) + \mathcal{L}^{\text{gen},-}_{\text{pose}}(\theta) \\ &= \mathbb{E}_{\boldsymbol{\xi}^+\sim p_\xi}[h(-(D^{\text{sp}}_\phi(\hat{\mathbf{I}}) \otimes E_\phi(\boldsymbol{\xi}^+)))] \\ &+ \mathbb{E}_{\boldsymbol{\xi}^-\sim p_\xi}[h(D^{\text{sp}}_\phi(\hat{\mathbf{I}}) \otimes E_\phi(\boldsymbol{\xi}^-))],\end{aligned} \tag{1}$$

where $\otimes$ is an element-wise multiplication, $D^{\text{sp}}_\phi(\cdot) = D^{\text{p}}_\phi(D^{\text{s}}_\phi(\cdot))$, $\hat{\mathbf{I}} = G_\theta(\mathbf{z}, \boldsymbol{\xi}^+)$, and $\mathbf{z} = (\mathbf{z}_{\text{fg}}, \mathbf{z}_{\text{bg}})$. h is the softplus activation function and p_ξ is the pose distribution, whose details will be given in Sec. 4.3 A negative pose $\boldsymbol{\xi}^-$ is randomly sampled so as not to be the same as the positive pose $\boldsymbol{\xi}^+$. For a generated image $\hat{\mathbf{I}}$, its positive pose $\boldsymbol{\xi}^+$ is the rendering pose used in the generator.

We also define a pose-matching loss to train the discriminator as:

$$\mathcal{L}^{\text{dis}}_{\text{pose}}(\phi) = \mathcal{L}^{\text{dis},+}_{\text{pose}}(\phi) + \mathcal{L}^{\text{dis},-}_{\text{pose}}(\phi), \tag{2}$$

where the terms on the right-hand-side are computed using positive and negative pairs, respectively. Both $\mathcal{L}^{\text{dis},+}_{\text{pose}}(\phi)$ and $\mathcal{L}^{\text{dis},-}_{\text{pose}}(\phi)$ are defined using both real and synthesized images for positive and negative pairs. Specifically, $\mathcal{L}^{\text{dis},+}_{\text{pose}}$ is defined as:

$$\begin{aligned}\mathcal{L}^{\text{dis},+}_{\text{pose}}(\phi) = \mathbb{E}_{(\mathbf{I},\boldsymbol{\xi}^+)\sim(p_r,p_\xi)}[h(-(D^{\text{sp}}_\phi(\mathbf{I}) \otimes E_\phi(\boldsymbol{\xi}^+)))] \\ + \mathbb{E}_{\boldsymbol{\xi}^+\sim p_\xi}[h(-(D^{\text{sp}}_\phi(\hat{\mathbf{I}}) \otimes E_\phi(\boldsymbol{\xi}^+)))],\end{aligned} \tag{3}$$

where p_r is the distribution of real images and $\mathbf{I}$ is a real image. The first and second terms on the right-hand side use real and synthesized pairs as positive pairs, respectively. For the first term, we sample a real image $\mathbf{I}$ and its corresponding ground-truth pose $\boldsymbol{\xi}^+$ as a positive sample. The pose-matching loss $\mathcal{L}^{\text{dis},-}_{\text{pose}}(\phi)$ for a negative pose is defined as:

$$\begin{aligned}\mathcal{L}^{\text{dis},-}_{\text{pose}}(\phi) = \mathbb{E}_{\mathbf{I}\sim p_r,\boldsymbol{\xi}^-\sim p_\xi}[h(D^{\text{sp}}_\phi(\mathbf{I}) \otimes E_\phi(\boldsymbol{\xi}^-))] \\ + \mathbb{E}_{\boldsymbol{\xi}^-\sim p_\xi}[h(D^{\text{sp}}_\phi(\hat{\mathbf{I}}) \otimes E_\phi(\boldsymbol{\xi}^-))].\end{aligned} \tag{4}$$

Note that both positive and negative pairs of the pose-matching loss for the discriminator are defined using both real and synthesized images. Thanks to this, the pose branch of the discriminator is trained to focus only on the pose consistency regardless of whether an image looks real or fake, and subsequently, resulting in the generator being trained to produce pose-consistent images.

4.2. Final Loss

In addition to the pose-matching loss, we adopt other loss terms in our final loss as described in the following.

Non-Saturating GAN Loss. In the dual-branched discriminator, the image branch D^{i}_ϕ is optimized by a non-saturating GAN loss to learn the entire target image distribution. The non-saturating GAN loss for the generator is defined as

$$\mathcal{L}^{\text{gen}}_{\text{adv}}(\theta) = \mathbb{E}_{\mathbf{z}\sim p_z,\boldsymbol{\xi}^+\sim p_\xi}[h(-D^{\text{si}}_\phi(G_\theta(\mathbf{z},\boldsymbol{\xi}^+)))], \tag{5}$$

where $D^{\text{si}}_\phi(\cdot) = D^{\text{i}}_\phi(D^{\text{s}}_\phi(\cdot))$. The non-saturating GAN loss for the discriminator with $R1$ regularization [1] is defined as

$$\begin{aligned}\mathcal{L}^{\text{dis}}_{\text{adv}}(\phi) = \mathbb{E}_{\mathbf{z}\sim p_z,\boldsymbol{\xi}^+\sim p_\xi}[h(D^{\text{si}}_\phi(G_\theta(\mathbf{z},\boldsymbol{\xi}^+)))] \\ + \mathbb{E}_{\mathbf{I}\sim p_r}[h(-D^{\text{si}}_\phi(\mathbf{I})) + \lambda_{R1}|\nabla D^{\text{si}}_\phi(\mathbf{I})|^2],\end{aligned} \tag{6}$$

where λ_{R1} is a balancing weight.

Identity Regularization. To encourage the generator to create semantically various images, we train the generator with an additional identity regularization term $\mathcal{L}_{\text{id}} = \lambda_{\text{z}}\mathcal{L}_{\text{z}} + \lambda_{\text{c}}\mathcal{L}_{\text{c}}$, where $\mathcal{L}_{\text{z}}$ is a loss term to promote images with diverse identities, and $\mathcal{L}_{\text{c}}$ is a term to prevent the identity of a generated image from being affected by the camera parameter $\boldsymbol{\xi}$. λ_{z} and λ_{c} are balancing weights. $\mathcal{L}_{\text{z}}$ is defined as

$$\mathcal{L}_{\text{z}}(\theta) = \mathbb{E}_{\mathbf{z}_1,\mathbf{z}_2\sim p_z,\boldsymbol{\xi}^+\sim p_\xi}[\langle E_{id}(\hat{\mathbf{I}}_1), E_{id}(\hat{\mathbf{I}}_2)\rangle], \tag{7}$$

where $\hat{\mathbf{I}}_1 = G_\theta(\mathbf{z}_1,\boldsymbol{\xi}^+)$, $\hat{\mathbf{I}}_2 = G_\theta(\mathbf{z}_2,\boldsymbol{\xi}^+)$, and E_{id} is a face identity network [4]. $\langle\cdot,\cdot\rangle$ calculates the cosine similarity. $\mathcal{L}_{\text{c}}$ is defined as:

$$\mathcal{L}_{\text{c}}(\theta) = \mathbb{E}_{\mathbf{z}\sim p_z,\boldsymbol{\xi}^+_1,\boldsymbol{\xi}^+_2\sim p_\xi}\left[\frac{1 - \langle E_{\text{id}}(\hat{\mathbf{I}}_1), E_{id}(\hat{\mathbf{I}}_2)\rangle}{\parallel \hat{\mathbf{I}}_1 - \hat{\mathbf{I}}_2 \parallel_1}\right], \tag{8}$$

where $\hat{\mathbf{I}}_1 = G_\theta(\mathbf{z},\boldsymbol{\xi}^+_1)$, and $\hat{\mathbf{I}}_2 = G_\theta(\mathbf{z},\boldsymbol{\xi}^+_2)$. The identity regularization $\mathcal{L}_{\text{id}}$ helps the generator faithfully learn semantic information from the dataset, enabling image synthesis with high fidelity (Sec. 5.2).

Final Loss. The final losses for training the generator and the discriminator are then defined as:

$$\begin{aligned}\mathcal{L}^{\text{gen}}_{\text{total}} &= \mathcal{L}^{\text{gen}}_{\text{adv}} + \lambda_{\text{pose}}\mathcal{L}^{\text{gen}}_{\text{pose}} + \mathcal{L}_{\text{id}} + \lambda_{\text{d}}\mathcal{L}_{\text{d}}, \quad \text{and} \\ \mathcal{L}^{\text{dis}}_{\text{total}} &= \mathcal{L}^{\text{dis}}_{\text{adv}} + \lambda_{\text{pose}}\mathcal{L}^{\text{dis}}_{\text{pose}},\end{aligned} \tag{9}$$

where $\mathcal{L}_{\text{d}}$ is an additional L^1-based density regularization term [24]. λ_{pose} and λ_{d} are weights to balance the terms. More details on the losses can be found in Sec. 5.

4.3. Additional Uniform Pose Sampling

As previous methods mostly focus on learning frontal-view images in their training because of the pose-imbalanced dataset, they lack opportunities for learning side-view images, resulting in degenerate side-view image quality. To increase the opportunities of learning side-view images in pose-imbalanced datasets, our AUPS strategy samples camera poses for rendering fake images from the training dataset like EG3D [1] and additionally sample poses from a uniform distribution in training. Specifically, for computing the non-saturating GAN loss with the image branch of the discriminator, we use camera poses sampled from the training dataset and the uniform distribution together. For computing the pose-matching loss and the identity regularization with the pose branch of the dual-branched discriminator, on the other hand, we simply use camera poses sampled only from the training dataset like EG3D as we found that the pose branch can already be effectively trained without AUPS.

While sampling camera poses solely from the uniform distribution may seem straightforward to increase the learning opportunities at steep angles, it can lead to a significant discrepancy between the real and fake image distributions,

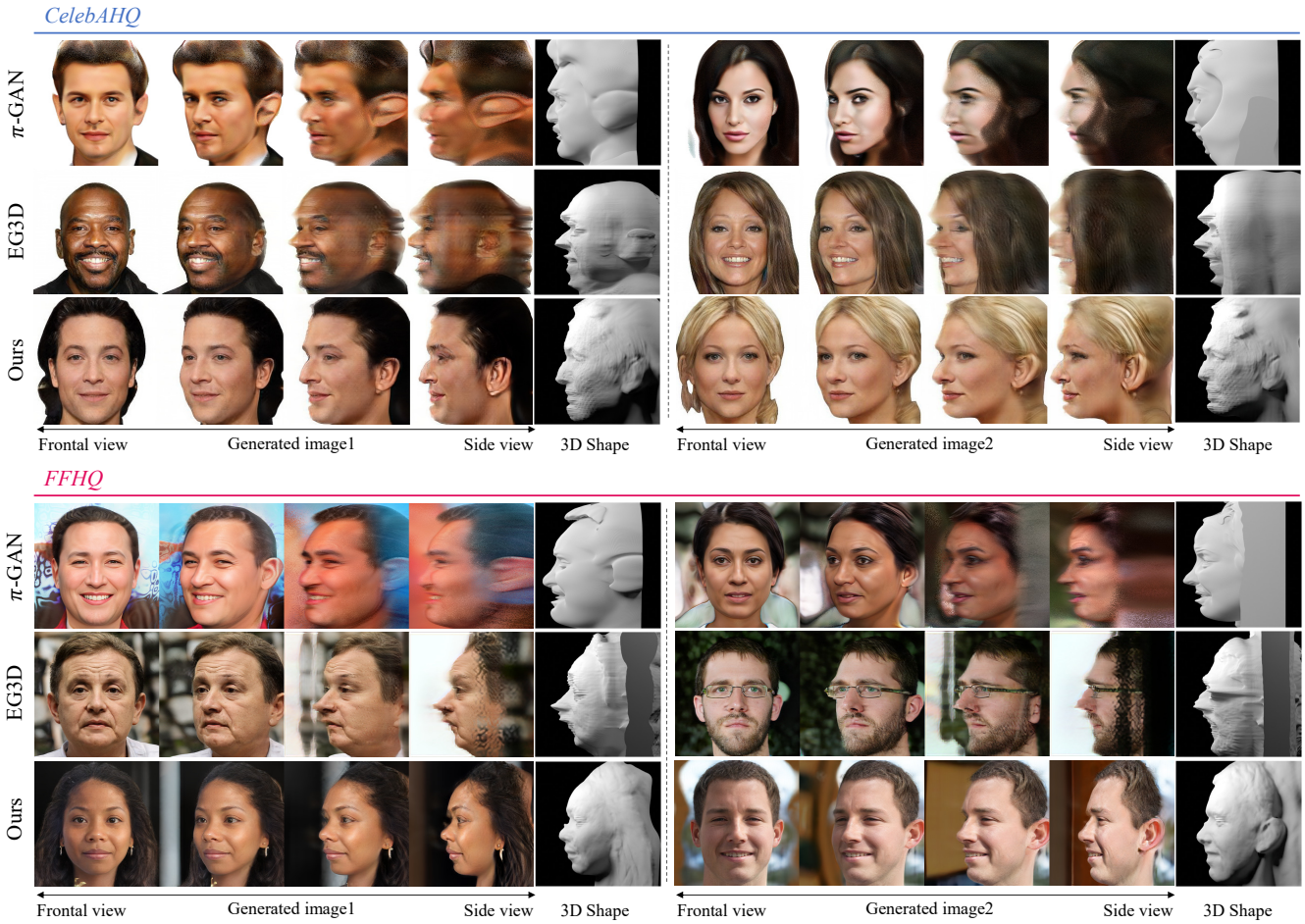

Figure 4: Qualitative comparison among π-GAN [2], EG3D [1] and ours. All the models are trained without transfer learning. Unlike blurry images and noisy geometry of baselines at the steep pose, our method generates high-quality images and shapes on the target datasets. (Columns 1-4, 6-9 : the results of a 30-degree rotation from the frontal to the side view. Columns 5, 10 : the side views of the shape obtained using the marching cube.)

which may harm the training process. To mitigate this, we use both pose distribution of the training dataset and uniform distribution together to decrease the distribution discrepancy while increasing learning opportunities for steep angles. More details on the AUPS can be found in Sec. B.3.

5. Experiments

Implementation Details. Most of the experimental setups and preprocessing methods are the same as those of EG3D [1] except for the following. We set the dimensions of the background latent vector $\mathbf{z}_{bg}$ to 512. The final image resolution of our model is 256×256 and the neural rendering resolution is fixed as 64×64. The neural rendering result is bilinearly upsampled to 128×128 and fed to the super-resolution module in the image generator. The batch size is set to 64 in all the experiments. The balancing weights for the loss terms are set as follows: $\lambda_{\text{pose}} = 1$, $\lambda_{\text{z}} = 0.5$, $\lambda_{\text{c}} = 0.25$, $\lambda_{\text{d}} = 0.25$ and $\lambda_{R1} = 1$.

Datasets. We validate our method on real-world human face datasets (CelebAHQ [12] and FFHQ [13]) and a real-world cat face dataset (AFHQ Cats [3]). To show results both with and without background regions, we remove the background regions of the CelebAHQ dataset using the ground-truth segmentation masks, but keep the background regions of the FFHQ dataset in our experiments. We obtain the ground-truth poses of real images using pre-trained camera pose estimation models [8, 15].

Transfer learning. As used in previous 3D GANs for compensating for the small dataset size, we optionally adopt transfer learning to improve the quality of side-view image synthesis [1, 10]. To be specific, we pre-train a generator with a pose-balanced synthetic dataset and fine-tune it with a pose-imbalanced in-the-wild dataset to compensate for insufficient knowledge for side-view images in in-the-wild datasets. Specifically, we use the FaceSynthetics dataset [26] for pre-training. We also remove the

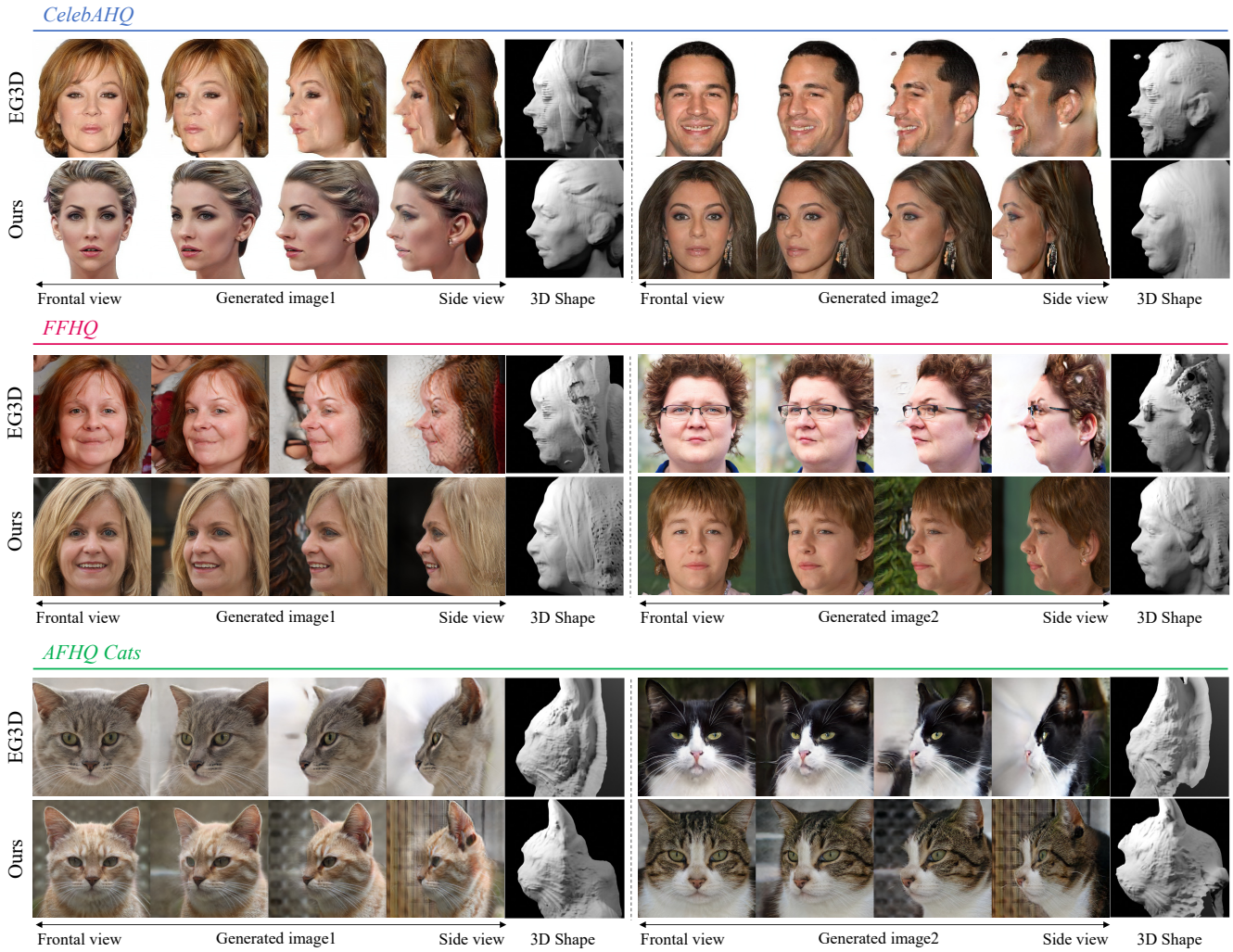

Figure 5: Qualitative comparison between EG3D [1] and ours. Both models are trained with transfer learning. Unlike unnatural images and geometry of the baseline at the steep pose, our method generates high-quality images and shapes on the target datasets. (Columns 1-4, 6-9 : the results of a 30-degree rotation from the frontal to the side view. Columns 5, 10 : the side views of the shape obtained using the marching cube.)

background regions of the FaceSynthetics dataset with the ground-truth segmentation masks to accurately learn 3D geometries. We use the training strategy of EG3D [1] to pretrain models.

5.1. Comparison

We first conduct qualitative and quantitative comparisons of SideGAN and previous 3D GANs (π-GAN [2] and EG3D [1]) on different datasets (CelebAHQ [12], FFHQ [13] and AFHQ Cats [3]) both with and without transfer learning.

Qualitative Comparison. Fig. 4 shows a qualitative comparison between our method and the previous 3D GANs. In this comparison, all the models are trained from scratch without transfer learning. The AFHQ Cats dataset [3] is not included in this comparison as the dataset is too small to train a generator without transfer learning. For all the real-world human face datasets, π-GAN and EG3D generate blurry images for steep angles compared to realistic frontal images. In contrast, SideGAN robustly generates high-quality images irrespective of camera pose. Fig. 5 shows another qualitative comparison where we adopt transfer learn-

	FID↓			Depth error↓	
Method \Dataset	CelebAHQ	FFHQ	AFHQ(Cats)	CelebAHQ	FFHQ
π-GAN	80.372	120.991	-	2.438	1.365
EG3D	40.760	35.348	-	0.760	0.921
Ours	**37.417**	**22.174**	-	**0.580**	**0.649**
EG3D+transfer learning	28.912	26.627	15.639	0.606	0.864
Ours+transfer learning	**22.219**	**24.571**	**10.134**	**0.549**	**0.657**

Table 1: Quantitative comparison of the image and shape quality with baselines.

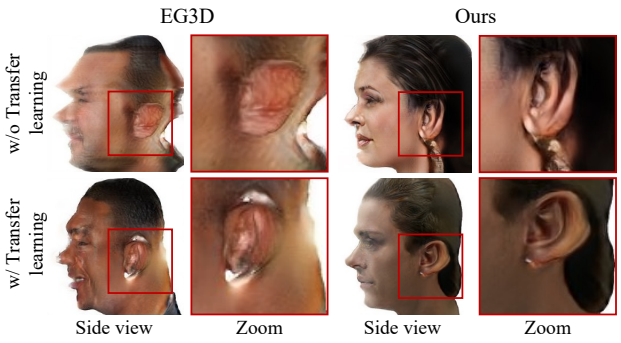

Figure 6: Visual comparison of side-view images on CelebAHQ [12] on the setting with or without transfer learning. Without transfer learning, our proposed method outperform the baseline (EG3D [1]), which shows noisy facial boundaries. With transfer learning, our method also outperforms the baseline, which generates holes.

ing. For all the datasets, EG3D generates unnatural images for steep angles compared to realistic frontal images. On the other hand, SideGAN robustly generates high-quality images irrespective of camera pose. These results indicate that our method is effective in learning to synthesize high-quality images at all camera poses in both cases with and without transfer learning. Additional results are in Sec. D.

Fig. 6 shows zoomed-in patches of side-view images of SideGAN and EG3D [1] to compare the quality of synthesized details. As the figure shows, SideGAN produces more realistic details for side-view images with much less artifacts than EG3D regardless of transfer learning. In addition, the figure also shows that transfer learning helps both models generate clearer images as it provides additional information on side views of human faces. Nevertheless, the result of EG3D with transfer learning still suffers from severe artifacts such as holes due to its pose-sensitive training process.

Quantitative Comparison. We conduct a quantitative evaluation on the image and shape quality. To evaluate the pose-irrespective performance of the models, we generate images and shapes at randomly sampled camera poses from a uniform distribution. Refer to Sec. C.5. for more details regarding the pose sampling strategy used in this experiment. Tab. 1 shows the quantitative comparison. As the table shows, in both cases with and without transfer learning, SideGAN outperforms all the other baselines in terms of image quality based on FID [11] thanks to our effective training method.

Due to the absence of 3D geometries corresponding to synthesized images, we evaluate the shape quality with pseudo-ground-truth shapes, which are estimated from synthesized images using an off-the-shelf 3D reconstruction model [8], as done in EG3D [1]. We measure depth error by calculating MSE between generated depth from our model and rendered depth from the estimated geometry. Tab. 1 shows that SideGAN achieves the best depth accuracy than the other baselines for the shape quality both with and without transfer learning. This remarkable improvement can also be shown in Fig. 4 and Fig. 5, where generated shapes from SideGAN show high-fidelity 3D geometries compared to those of the other methods.

	AUPS	Dual-branched discriminator & Pose-matching loss	Identity regularization ($\mathcal{L}_{\text{id}}$)	FID↓
EG3D				28.912
	✓			30.553
SideGAN	✓	✓		23.106
(Ours)	✓	✓	✓	**22.219**

Table 2: Ablation study for key components of the proposed method on CelebAHQ [12].

	FID↓	Depth error↓
Ours w/ pose-regression loss	30.069	0.624
Ours w/ pose-matching loss	**22.219**	**0.549**

Table 3: Comparison between the pose-matching loss and the pose-regression loss [6] on CelebAHQ [12].

Figure 7: Additional visual results of ablation study. (b) While AUPS helps improve side-view image quality, artifacts still remain. (d) $\mathcal{L}_{\text{id}}$ results in a slightly clearer side-view image than (c). (e) Instead of our pose-matching loss, our model with the pose-regression loss leads to a flattened shape.

5.2. Ablation Studies

We conduct ablation studies to evaluate the benefits of three components in our framework: 1) the dual-branched discriminator (Sec. 3.2), 2) the pose-matching loss (Sec. 4.1), 3) AUPS (Sec. 4.3), and 4) the identity regularization $\mathcal{L}_{\text{id}}$ (Sec. 4.2). The ablation studies are conducted using the CelebAHQ dataset [12].

Tab. 2 and Fig. 7 report the ablation study result. In Tab. 2, the pose-matching loss and the dual-branched discriminator are applied together since supervision is needed for the pose branch of the discriminator. With AUPS, the image quality of side-view is improved in EG3D [1] (Fig. 7 (b)) since AUPS increases the learning opportunities at steep angles. However, the side-view images still have artifacts and the FID of EG3D deteriorates. This is because EG3D learns the real/fake distribution in a pose-wise manner through a pose-conditional GAN loss, which is unstable under the misalignment between two distributions, caused by AUPS as mentioned in Sec. 4.3. Unlike

EG3D, our framework with AUPS improves the FID as each component is added, proving the benefit of each component. This is because SideGAN's GAN loss is more robust to the mismatch of the pose distribution than EG3D and our model learns photo-realism and pose-consistency separately through the dual-branched discriminator and the pose-matching loss.

To evaluate the effectiveness of the pose-matching loss in learning side-view images and 3D geometries, we compare our pose-matching loss with the pose-regression loss of GRAM [6] both quantitatively and qualitatively. As shown in Tab. 3, our pose-matching loss results in a significantly lower FID score and depth error, owing to the fact that our binary-classification-based pose-matching loss allows for easier training. We also provide visual comparison in Fig. 7. Compared to our model trained with the pose-matching loss (d), the model trained with the pose-regression loss (e) produces a flattened shape, demonstrating the advantages of the pose-matching loss.

5.3. Effects on the Steep and Extrapolated Angles

Finally, we conduct a more detailed quantitative analysis of SideGAN for different camera poses by measuring the FID scores of synthesized images for frontal, steep, and extrapolated angles. Measuring FID scores requires a sufficient amount of ground-truth images for each camera pose, which is not the case for the in-the-wild datasets. Thus, we conduct our analysis using the FaceSynthetic dataset [26], which is pose-balanced and provides a larger number of images for a wider range of camera poses compared to the in-the-wild datasets. Specifically, we first construct a pose-imbalanced training dataset from FaceSynthetics by randomly sampling images within the pose range from $-50°$ to $50°$ to have a Gaussian pose distribution like in-the-wild datasets. Then, we train our model with the dataset, and evaluate its FID scores for different camera poses using the original FaceSynthetics dataset. For comparison, we also evaluate the FID scores of EG3D [1]. In this experiment, we did not apply transfer learning.

Fig. 8 shows the evaluation result for different camera poses. In the figure, the near-frontal angles are $(-30°, 30°)$, and the steep angles are $(-50°, -30°) \bigcup (30°, 50°)$. The extrapolated angles indicate angles smaller or larger than $-50°$ and $50°$, which are outside the training distribution. As shown in the figure, our model performs comparably to EG3D at near-frontal angles, and as the angle gets larger, our model performs significantly better than EG3D, proving the effectiveness of our approach.

6. Conclusion

In this paper, we proposed SideGAN, a novel 3D GAN training method to generate high-quality images irrespective of the camera pose. Our method is based on the key idea that decomposes the originally challenging problem into two easier subproblems, each of which promotes pose consistency and photo-realism, respectively. Based on this, we propose a novel dual-branched discriminator and a pose-matching loss. We also presented AUPS to increase the learning opportunities for improving the synthesis quality at a side viewpoint.

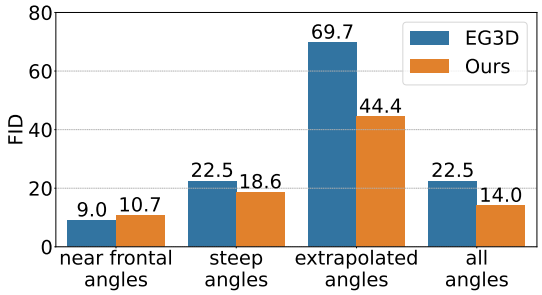

Figure 8: Comparison on image quality (FID↓) with regard to the range of the camera angles. We limit the FaceSynthetic dataset [26] not to have any images within the range of extrapolated angles. SideGAN outperforms EG3D [1] in image quality except for the range of the frontal view, which shows even competitive result. All angles are from -90 to 90 degrees based on the frontal view.

Our experimental results show that our method can synthesize photo-realistic images irrespective of the camera pose on human and animal face datasets. Especially, even only with pose-imbalanced in-the-wild datasets, our model can generate details of side-view images such as ears, unlike blurry images from the baselines.

Our method is not free from limitations. For animal faces, we found that black spot-like artifacts appear behind the ear, which might be due to the lack of knowledge about the back of the ear since we conduct transfer learning from synthetic human face to animal face. Also, despite the background network, the background region is sometimes not clearly separated. However, we expect that a more advanced background separation scheme such as [22] would be able to resolve this.

Acknowledgement This work was supported by the National Research Foundation of Korea (NRF) grant (NRF-2018R1A5A1060031), the Institute of Information & communications Technology Planning & Evaluation (IITP) grant (No.2019-0-01906, Artificial Intelligence Graduate School Program(POSTECH)) funded by the Korea government (MSIT), Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2019-0-00075, Artificial Intelligence Graduate School Program (KAIST)), and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2022R1A2B5B02001913).

References

[1] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *Proc. of IEEE conference on computer vision and pattern recognition (CVPR)*, pages 16123–16133, 2022.

[2] 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 *Proc. of IEEE conference on computer vision and pattern recognition (CVPR)*, pages 5799–5809, 2021.

[3] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proc. of IEEE conference on computer vision and pattern recognition (CVPR)*, pages 8188–8197, 2020.

[4] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proc. of IEEE conference on computer vision and pattern recognition (CVPR)*, pages 4690–4699, 2019.

[5] Yu Deng, Jiaolong Yang, Dong Chen, Fang Wen, and Xin Tong. Disentangled and controllable face image generation via 3d imitative-contrastive learning. In *Proc. of IEEE conference on computer vision and pattern recognition (CVPR)*, pages 5154–5163, 2020.

[6] Yu Deng, Jiaolong Yang, Jianfeng Xiang, and Xin Tong. Gram: Generative radiance manifolds for 3d-aware image generation. In *Proc. of IEEE conference on computer vision and pattern recognition (CVPR)*, pages 10673–10683, 2022.

[7] Shi et al. 3d-aware indoor scene synthesis with depth priors. In *ECCV*, pages 406–422, 2022.

[8] Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. Learning an animatable detailed 3d face model from in-the-wild images. *ACM Transactions on Graphics (ToG)*, 40(4):1–13, 2021.

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[10] 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.

[11] 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. *Proc. of the Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2017.

[12] 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.

[13] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proc. of IEEE conference on computer vision and pattern recognition (CVPR)*, pages 4401–4410, 2019.

[14] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *Proc. of IEEE conference on computer vision and pattern recognition (CVPR)*, 2020.

[15] Taehee Brad Lee. Cat hipsterizer. `https://github.com/kairess/cat_hipsterizer`, 2018. Accessed: 2022-11-08.

[16] 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. *Communications of the ACM*, 65(1):99–106, 2021.

[17] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *Proc. of IEEE conference on computer vision and pattern recognition (CVPR)*, pages 11453–11464, 2021.

[18] Roy Or-El, Xuan Luo, Mengyi Shan, Eli Shechtman, Jeong Joon Park, and Ira Kemelmacher-Shlizerman. Stylesdf: High-resolution 3d-consistent image and geometry generation. In *Proc. of IEEE conference on computer vision and pattern recognition (CVPR)*, pages 13503–13513, 2022.

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

[20] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *Proc. of IEEE conference on computer vision and pattern recognition (CVPR)*, pages 9243–9252, 2020.

[21] Yujun Shen and Bolei Zhou. Closed-form factorization of latent semantics in gans. In *Proc. of IEEE conference on computer vision and pattern recognition (CVPR)*, pages 1532–1540, 2021.

[22] Minjung Shin, Yunji Seo, Jeongmin Bae, Young Sun Choi, Hyunsu Kim, Hyeran Byun, and Youngjung Uh. Ballgan: 3d-aware image synthesis with a spherical background. *arXiv preprint arXiv:2301.09091*, 2023.

[23] Ivan Skorokhodov, Sergey Tulyakov, Yiqun Wang, and Peter Wonka. Epigraf: Rethinking training of 3d gans. *arXiv preprint arXiv:2206.10535*, 2022.

[24] Jingxiang Sun, Xuan Wang, Yichun Shi, Lizhen Wang, Jue Wang, and Yebin Liu. Ide-3d: Interactive disentangled editing for high-resolution 3d-aware portrait synthesis. *arXiv preprint arXiv:2205.15517*, 2022.

[25] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6142–6151, 2020.

[26] Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Sebastian Dziadzio, Matthew Johnson, Virginia Estellers, Thomas J. Cashman, and Jamie Shotton. Fake it till you make it: Face analysis in the wild using synthetic data alone, 2021.

[27] Yinghao Xu, Sida Peng, Ceyuan Yang, Yujun Shen, and Bolei Zhou. 3d-aware image synthesis via learning structural and textural representations. In *Proc. of IEEE conference on computer vision and pattern recognition (CVPR)*, pages 18430–18439, 2022.

[28] 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.