Learning 3D-aware Image Synthesis with Unknown Pose Distribution

Zifan Shi[†*1] Yujun Shen[†2] Yinghao Xu[*3] Sida Peng[4] Yiyi Liao[4] Sheng Guo[2]
Qifeng Chen[1] Dit-Yan Yeung[1]

[1]HKUST [2]Ant Group [3]CUHK [4]Zhejiang University

Abstract

*Existing methods for 3D-aware image synthesis largely depend on the 3D pose distribution pre-estimated on the training set. An inaccurate estimation may mislead the model into learning faulty geometry. This work proposes **PoF3D** that frees generative radiance fields from the requirements of 3D pose priors. We first equip the generator with an efficient pose learner, which is able to infer a pose from a latent code, to approximate the underlying true pose distribution automatically. We then assign the discriminator a task to learn pose distribution under the supervision of the generator and to differentiate real and synthesized images with the predicted pose as the condition. The pose-free generator and the pose-aware discriminator are jointly trained in an adversarial manner. Extensive results on a couple of datasets confirm that the performance of our approach, regarding both image quality and geometry quality, is on par with state of the art. To our best knowledge, PoF3D demonstrates the feasibility of learning high-quality 3D-aware image synthesis without using 3D pose priors for the first time. Project page can be found here.*

1. Introduction

3D-aware image generation has recently received growing attention due to its potential applications [3, 4, 9, 21, 30, 34, 47]. Compared with 2D synthesis, 3D-aware image synthesis requires the understanding of the geometry underlying 2D images, which is commonly achieved by incorporating 3D representations, such as neural radiance fields (NeRF) [2, 16, 17, 24, 25, 50], into generative models like generative adversarial networks (GANs) [8]. Such a formulation allows explicit camera control over the synthesized results, which fits better with our 3D world.

To enable 3D-aware image synthesis from 2D image collections, existing attempts usually rely on adequate camera pose priors [3, 4, 9, 20, 21, 30, 47] for training.

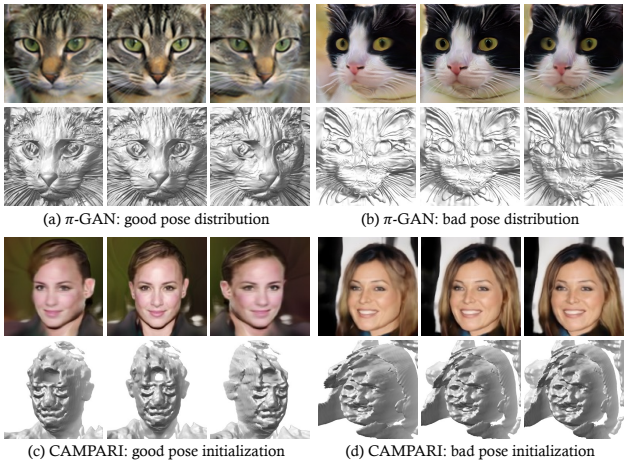

Figure 1. **Sensitivity to pose priors** in existing 3D-aware image synthesis approaches. π-GAN [4] works well given an adequate pose distribution in (a), but fails to learn decent geometries given a *slightly changed* distribution in (b). CAMPARI [20] delivers reasonable results relying on a good initial pose distribution in (c), but suffers from a wrongly estimated initialization in (d).

The priors are either estimated by conducting multiple pre-experiments [4, 9, 47], or obtained by borrowing external pose predictors or annotations [3]. Besides using a fixed distribution for pose sampling, some studies also propose to tune the pose priors together with the learning of image generation [20], but they are still dependent on the initial pose distribution. Consequently, although previous methods can produce satisfying images and geometry, their performance is highly sensitive to the given pose prior. For example, on the Cats dataset [51], π-GAN [4] works well with a uniform pose distribution [-0.5, 0.5] (Fig. 1a), but fails to generate a decent underlying geometry when changing the distribution to [-0.3, 0.3] (Fig. 1b). Similarly, on the CelebA dataset [15], CAMPARI [20] learns 3D rotation when using Gaussian distribution $\mathcal{N}(0, 0.24)$ as the initial prior (Fig. 1c), but loses the canonical space (*e.g.*, the middle image of Fig. 1d should be under the frontal view) when changing the initialization to $\mathcal{N}(0, 0.12)$ (Fig. 1d). Such a reliance on pose priors causes unexpected instability of 3D-aware image synthesis, which may burden this task with heavy experimental cost.

† indicates equal contribution.
* This work was done during an internship at Ant Group.

In this work, we present a new approach for 3D-aware image synthesis, which removes the requirements for pose priors. Typically, a latent code is bound to the 3D content alone, where the camera pose is independently sampled from a manually designed distribution. Our method, however, maps a latent code to an image, which is implicitly composed of a 3D representation (*i.e.*, neural radiance field) and a camera pose that can render that image. In this way, the camera pose is directly inferred from the latent code and jointly learned with the content, simplifying the input requirement. To facilitate the pose-free generator better capturing the underlying pose distribution, we re-design the discriminator to make it pose-aware. Concretely, we tailor the discriminator with a pose branch which is required to predict a camera pose from a given image. Then, the estimated pose is further treated as the conditional pseudo label when performing real/fake discrimination. The pose branch in the discriminator learns from the synthesized data and its corresponding pose that is encoded in the latent code, and in turn use the discrimination score to update the pose branch in the generator. With such a loop-back optimization process, two pose branches can align the fake data with the distribution of the dataset.

We evaluate our *pose-free* method, which we call **PoF3D** for short, on various datasets, including FFHQ [12], Cats [51], and Shapenet Cars [5]. Both qualitative and quantitative results demonstrate that PoF3D frees 3D-aware image synthesis from hyper-parameter tuning on pose distributions and dataset labeling, and achieves on par performance with state-of-the-art in terms of image quality and geometry quality.

2. Related Work

3D-aware Image Synthesis. 3D-aware image synthesis has achieved remarkable success recently [33, 43] and benefits lots of 3D-aware applications [39, 44, 49, 55] rather than the conventional 2D applications [1, 26, 27, 38, 40, 41, 54]. Different from 2D synthesis [11–13, 48], some works [18, 19, 56] propose to adopt voxels as 3D representation for image rendering but suffer from the poor image quality and consistency due to the voxel resolution restriction. Then a series of works [4, 6, 23, 30, 45] introduce neural implicit function [16, 17, 24] as the underlying 3D representation for image rendering. However, rendering high-resolution images with direct volume rendering is very heavy. Lots of works [3, 9, 21, 22, 31, 37, 46, 47, 52] resort to either 2D convolutional upsamplers [3, 9, 21, 22, 47], multi-plane image rendering [52], sparse-voxel [31] inference, patch-based training [37] for acceleration. Besides, some works [23, 35] focus on the improvement of geometry quality. Although these methods are able to synthesize high-quality 3D-aware images, they are restricted to strong pose prior, *i.e.*, manually tuned pose distribution, or well-annotated camera poses [3]. Our work instead naturally enables 3D-aware image synthesis without any pose prior via a pose-free generator, which gives a fruitful avenue for future work.

Camera Learning from 2D Images. Learning neural radiance fields (NeRF) [17] requires accurate pose annotations. To overcome this, some works [14, 42] try to optimize the annotated camera parameters or directly estimate camera poses with a small network, which is conceptually comparable to ours. On the other hand, our work differs from theirs in that our method focuses on learning pose distribution of a real dataset rather than exact camera poses. CAMPARI [20] and 3DGP [36] investigate a similar problem, aiming to estimate the pose distribution of an actual dataset. However, both of them need a manually designed prior, *i.e.*, uniform or Gaussian, and fail to learn appropriate 3D geometry when the prior is very distant from the native distribution. In contrast, our work abandons all the manually designed prior and allows the generator to learn pose distribution automatically via adversarial training.

3. Method

Our proposed method for 3D-aware image synthesis can be freed from the requirement of pose priors. To achieve this goal, we propose some new designs in both the generator and the discriminator in conventional 3D-aware GANs [3, 4, 9, 30]. Specifically, the pose-free generator is equipped with a pose learner that infers the camera pose from the latent code. The pose-aware discriminator extracts a pose from the given image and uses it as the conditional label when performing real/fake classification. The framework is shown in Fig. 2. Before going into details, we first briefly introduce the generative neural radiance field, which plays a crucial role in 3D-aware image synthesis.

3.1. Preliminary

A neural radiance field (NeRF) [17], $\mathrm{F}(\mathbf{x}, \mathbf{d}) \rightarrow (\mathbf{c}, \sigma)$, provides color $\mathbf{c} \in \mathbb{R}^3$ and volume density $\sigma \in \mathbb{R}$ from a coordinate $\mathbf{x} \in \mathbb{R}^3$ and a viewing direction $\mathbf{d} \in \mathbb{S}^2$, typically parameterized with multi-layer perceptron (MLP) networks. Then, the pixel values are accumulated from the colors and densities of the points on the sampled rays. However, NeRF is highly dependent on multi-view supervision, leading to the inability to learn from a single-view image collection. To enable random sampling from single-view captures, recent attempts propose to condition NeRF with a latent code $\mathbf{z}$, resulting in their generative forms [4, 30], $\mathrm{G}(\mathbf{x}, \mathbf{d}, \mathbf{z}) \rightarrow (\mathbf{c}, \sigma)$, to achieve diverse 3D-aware generation.

3.2. Pose-free Generator

Different from NeRF, the camera pose ξ deriving the spatial point $\mathbf{x}$ and view direction $\mathbf{d}$ in conventional 3D-aware

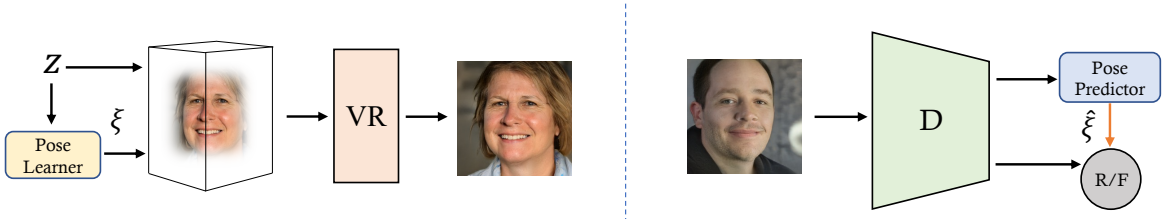

Figure 2. **Framework of PoF3D**, which consists of a pose-free generator and a pose-aware discriminator. The pose-free generator maps a latent code to a neural radiance field as well as a camera pose, followed by a volume renderer (VR) to output the final image. The pose-aware discriminator first predicts a camera pose from the given image and then use it as the pseudo label for conditional real/fake discrimination, indicated by the **orange** arrow.

generators is randomly sampled from a prior distribution p_ξ rather than well annotated by MVS [32] and SfM [29]. Such a prior distribution requires the knowledge of the pose distribution, and it should be tailored for different datasets, which introduces non-trivial hyperparameter tuning for model training. This generation process can be formulated as

$$\mathrm{G}(\mathbf{z}, \xi) = \mathbf{I}_f \sim p_\theta(\mathbf{I}_f | \mathbf{z}, \xi), \tag{1}$$

where the generator G synthesizes the image $\mathbf{I}_f$ by modeling the conditional probability given the latent code and the camera pose. Since the camera pose is independent of the latent code, we can rewrite Eq. (1) in the following form:

$$p_\theta(\mathbf{I}_f | \mathbf{z}, \xi) = \frac{p_\theta(\mathbf{I}_f, \mathbf{z}, \xi)}{p(\mathbf{z}, \xi)} = \frac{p_\theta(\mathbf{I}_f, \xi | \mathbf{z})}{p(\xi)} = \frac{p_\theta(\mathbf{I}_f, \xi | \mathbf{z})}{p_\xi}. \tag{2}$$

Obviously, if the pose prior does not align well with the real distribution, it is hard to correctly estimate the image distribution $p_\theta(\mathbf{I}_f | \mathbf{z}, \xi)$. As noticed by [9, 47], it always makes the training diverge if the prior does not align well with the real distribution. To free the generator from sampling poses from a prior distribution, we follow the formulation of conventional 2D GANs that synthesizes images from the latent code only:

$$\mathrm{G}(\mathbf{z}, \Psi(\mathbf{z})) = \mathbf{I}_f \sim p_\theta(\mathbf{I}_f | \mathbf{z}, \Psi(\mathbf{z})), \tag{3}$$

where the camera pose ξ is parameterized with a nonlinear function $\Psi(\cdot)$ that takes $\mathbf{z}$ as input. Concretely, we implement $\Psi(\cdot)$ by introducing an additional pose branch on the top of the generator. The estimated camera pose is further fed into a generative radiance field to render a 2D image. Compared with previous solutions, our generator aims at approximating the conditional probability only from the latent observation, which is prone to simulating the native data distribution without taking the pose prior into account.

3.3. Pose-aware Discriminator

As mentioned above, we remove the pose prior of a 3D-aware generator by parameterizing camera pose ξ from a latent code $\mathbf{z}$. The 3D representation leveraged in the generator is helpful in disentangling the camera factor. However, the conventional discriminator remains to differentiate real and fake images from 2D space, leading to inadequate supervision to factorize camera poses from latent codes. It easily makes the generator synthesize flat shapes and learn invalid camera distribution. Therefore, we propose to make the discriminator pose-aware.

Learning Pose-aware Discriminator. To let the discriminator become aware of poses, we assign a pose estimation task on *fake images* beyond bi-class domain classification, which is to derive the pose information from the given images. We introduce a pose branch $\Phi(\cdot)$ on top of the discriminator and optimize it with the following objective:

$$\hat{\xi} = \Phi(\mathbf{I}_f), \tag{4}$$

$$\mathcal{L}_{pose} = l_2(\hat{\xi}, \xi), \tag{5}$$

where $l_2(\cdot)$ denotes the function measuring l_2 distance between poses. ξ is the pose that is inferred from the same latent code as $\mathbf{I}_f$ by the generator.

Performing Pose-aware Discrimination. With the help of the pose branch, our discriminator can be adopted to infer camera poses for the given images. Even though the pose branch training is done on fake images, it is also generalizable on real images $\mathbf{I}_r$ from the dataset. In such cases, we can leverage the pose-aware discriminator to perform discrimination on real or fake images as well as differentiate whether the image is attached with an accurate pose or not. Concretely, we first ask the discriminator to extract a pose from the input image, and then the pose is treated as a pseudo label to perform conditional bi-class classification. Here, we take the real image part of the discriminator loss as an example:

$$\mathcal{L}_{\mathrm{D}}^{\mathbf{I}_r} = -\mathbb{E}[\log(\mathrm{D}(\mathbf{I}_r | \ \Phi(\mathbf{I}_r))]. \tag{6}$$

In this way, we facilitate the discriminator with awareness of pose cues, leading to better pose alignment across various synthesized samples. As a result, this alignment also

implicitly prevents the generator from degenerate solutions where a flat shape is usually generated with huge artifacts. It is worth noting that the discriminator learns poses from the generator and, in turn, uses the conditional discrimination score to update the pose branch in the generator. Therefore, the poses are learned in a loop-back manner without any annotations.

3.4. Training Objectives

Adversarial Loss. We use the standard adversarial loss for training following [8],

$$\begin{aligned}\mathcal{L}_{\mathrm{D}} = &- \mathbb{E}\big[\log\big(1 - \mathrm{D}(\mathbf{I}_f|\Phi(\mathbf{I}_f))\big)\big] \\ &- \mathbb{E}\big[\log\big(\mathrm{D}(\mathbf{I}_r|\Phi(\mathbf{I}_r))\big)\big] + \lambda\mathbb{E}\big[||\nabla_{\mathbf{I}_r}\mathrm{D}(\mathbf{I}_r)||_2^2\big],\end{aligned} \tag{7}$$

$$\mathcal{L}_{\mathrm{G}} = - \mathbb{E}[\log(\mathrm{D}(\mathbf{I}_f|\Phi(\mathbf{I}_f)))], \tag{8}$$

where $\mathbf{I}_r$ and $\mathbf{I}_f$ are real data and generated data, respectively. The third term in Eq. (7) is the gradient penalty, and λ denotes the weight for this term.

Symmetry Loss. Planar underlying shapes are easily generated when single-view images are synthesized for training. To avoid the trivial solution, we ask the network to synthesize the second image under another view, which we choose the symmetrical camera view ξ' regarding the yz-plane,

$$\mathbf{I}'_f = \mathrm{G}(\mathbf{z}, \xi'). \tag{9}$$

The novel view image will be used to calculate the adversarial loss as in Eq. (7) and Eq. (8), and get $\mathcal{L}_{\mathrm{D}'}$ and $\mathcal{L}_{\mathrm{G}'}$.

Pose Loss. As stated in Sec. 3.3, we also attach an auxiliary pose branch $\Phi(\cdot)$ in the discriminator to perform pose estimation on the given images. We use Eq. (5) for pose branch training.

Full Objectives. In summary, the pose-free generator and the pose-aware discriminator are jointly optimized with

$$\mathcal{L} = \mathcal{L}_{\mathrm{G}} + \mathcal{L}_{\mathrm{G}'} + \mathcal{L}_{\mathrm{D}} + \mathcal{L}_{\mathrm{D}'} + \gamma\mathcal{L}_{pose}, \tag{10}$$

where γ is the weight of pose loss.

3.5. Implementation Details

We build our PoF3D on the architecture of EG3D [3]. For the pose-aware generator, we do not use pose-conditioned generation. Instead, we instantiate the pose branch with two linear layers and a leaky ReLU activation in between. This branch takes in latent codes from w space and outputs camera poses. The camera pose consists of an azimuth angle and an elevation angle, which indicates the camera position for rendering. The triplane resolution is 256×256, and the rendering in the neural radiance field is conducted on 64×64 resolution. Both the feature map and the image are rendered. A super-resolution module then transforms the feature map into a high-resolution image. The pose-aware discriminator is inherited from the dual discriminator in EG3D [3]. We add the pose branch on the features before the last two fully-connected layers that output the realness score. The pose branch is composed of two fully-connected layers with a leaky ReLU as the activation function in the middle. The learning rate of the generator is $2.5e$-3 while that of its pose branch is set to $2.5e$-5. The discriminator's learning rate is $2e$-3. More details are available in the *Supplementary Material*.

4. Experiments

4.1. Experimental Settings

Datasets. We evaluate PoF3D on three datasets, including FFHQ [12], Cats [51], and Shapenet Cars [5]. FFHQ is a real-world dataset that contains unique 70K high-quality images of human faces. We follow [3] to align and crop the data. Cats includes 10K real-world cat images of various resolutions. The data is preprocessed following [6]. Shapenet Cars [5] is a synthetic dataset that contains different car models. We use the dataset rendered from [3] that has around 530K images. Unlike the face-forward datasets, the camera poses of it span the entire 360° azimuth and 180° elevation distributions. In the experiments, we use the resolution of 256×256 for FFHQ and Cats, and 128×128 for Shapenet Cars.

Baselines. We compare our approach against two methods: CAMPARI [20], the state-of-the-art pose learning method in 3D-aware image synthesis, and EG3D [3], the state-of-the-art in 3D-aware image synthesis. We also build a baseline method which is a combination of them, where we incorporate the pose learning method in CAMPARI into the framework of EG3D. More details regarding the baselines can be found in *Supplementary Material*.

Metrics. We use five metrics to evaluate the performance, including Fréchet Inception Distance (FID) [10], Depth Error [3], Pose Error [47], Reprojection Error (RE) [47], and Jensen–Shannon Divergence (JS). FID is measured between 50K generated images and all real images. Depth Error is used to assess the quality of geometry. We follow [3] and calculate the mean squared error (MSE) against pseudo-ground-truth depth estimated by [7] on 10K synthesized samples. We evaluate the pose accuracy on 10K generated samples as well. Given the synthesized images, we leverage the head pose estimator [53] and report the L1 distance between the estimated poses and the poses inferred from the generator. Note that we subtract the respective means of estimated and inferred pose distribution to compensate the canonical shift problem (see Fig. 4). Inter-view consistency is measured by the re-projection error following [47]. We render 5 views by sampling azimuth uniformly in the range of [-23°, 23°] and warp two consecutive views to each other

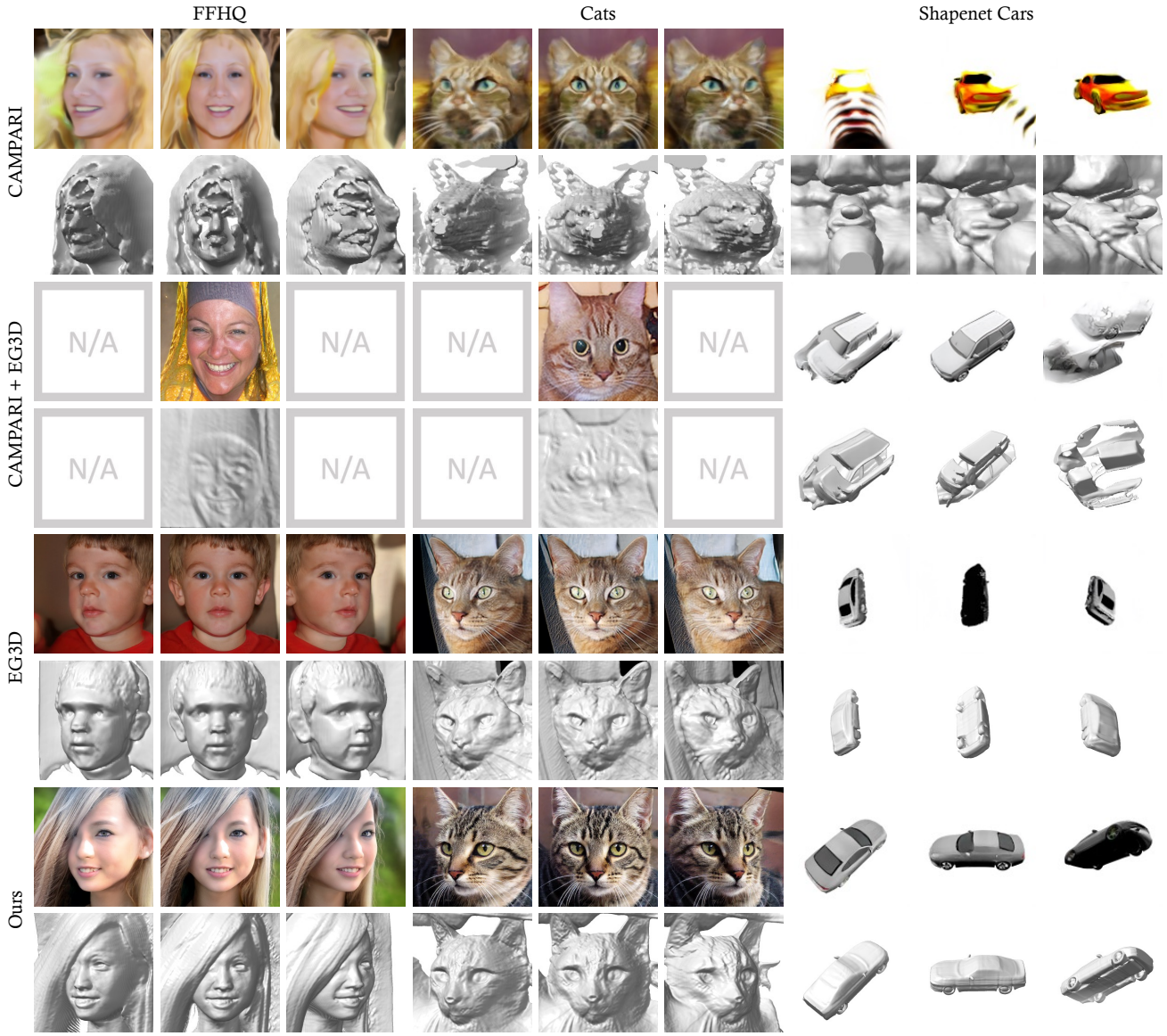

Figure 3. **Qualitative comparison between our approach and baselines.** CAMPARI [20] struggles to generate reasonable results. "CAMPARI + EG3D" suffers from mode collapse on the FFHQ and Cats datasets. EG3D [3] achieves impressive rendering and reconstruction results when camera poses of training images are given. In contrast, our approach can generate high-quality rendering and geometry without any pose priors.

to report the MSE. The images are normalized in the range of [-1, 1] for evaluation. Since the azimuth range learned by CAMPARI is much smaller, we use the learned range for sampling in the measurement of CAMPARI. To measure the quality of the learned pose distribution, we report Jensen–Shannon Divergence on 50K synthesized samples and all real images by averaging divergence values over azimuth and elevation.

4.2. Main Results

Qualitative Comparison. Fig. 3 shows the qualitative comparison against the baselines. CAMPARI learns faulty and very sharp shapes that span the entire space along the ray direction (the same phenomenon as observed in Fig. 1d). Therefore, even if the object is rotated with an extremely small angle (e.g., 2.5°), the caused visual effect is equivalent to that of the normal rotation of a decent shape with a much larger angle (e.g., 23°). We visualize CAMPARI in its own valid range of the horizontal angle (around [-2.5°, 2.5°] for FFHQ, [-5°, 5°] for Cats, and [-30°, 30°] for Shapenet Cars). For others, visualizations on FFHQ and Cats are conducted on the angle range [-23°, 23°]. Poses for cars are randomly sampled from the entire 360° azimuth and 180° elevation distributions. EG3D leverages ground-

Table 1. **Quantitative comparison** on FFHQ [12], Cats [51], and Shapenet Cars [5]. Our approach significantly outperforms CAMPARI [20] and "CAMPARI + EG3D" in terms of depth error, pose error, reprojection error, and Jensen-Shannon divergence score. Note that, EG3D [3] uses pose annotations, while our model is trained without any pose priors. * means we report the FID of EG3D on Shapenet Cars using its released model, which is inconsistent with the one reported in the paper.

Model	FFHQ [12]					Cats [51]			Shapenet Cars [5]	
	FID_{50k} ↓	Depth↓	Pose↓	RE↓	JS↓	FID_{50k} ↓	RE↓	JS↓	FID_{50k} ↓	JS↓
CAMPARI [20]	58.59	1.78	0.15	0.109	0.61	37.40	0.050	0.60	68.91	0.72
CAMPARI+EG3D	**3.25**	1.13	0.18	–	0.73	**4.83**	–	0.74	4.66	0.83
EG3D [3]	4.80	**0.29**	**0.07**	0.039	–	5.56	**0.042**	–	9.68*	–
Ours	4.99	**0.29**	0.10	**0.037**	**0.20**	5.46	0.046	**0.24**	**3.78**	**0.51**

truth camera poses for training and generates high-quality images and underlying geometry simultaneously. When incorporated with the pose learning method in CAMPARI, EG3D suffers from pose collapse and converges to one pose on FFHQ and Cats. Hence, novel views are not available in such cases, and the 3D-aware image synthesis problem is degraded to 2D image synthesis. On Shapenet Cars, the generated image and its corresponding shape are visually reasonable under a specific view but become completely unnatural when deviating from that view. Our method, without any pose priors, can generate high-fidelity images and shapes on par with EG3D.

Quantitative Comparison. We report the quantitative evaluation of baselines and our method in Tab. 1. CAMPARI fails to capture the underlying pose distribution and gets high Jensen–Shannon Divergence scores and pose errors. Consequently, the quality of images and shapes is unsatisfying as well, reflected by high FID scores and depth errors. CAMPARI+EG3D generates planar shapes and collapses to one pose or a small range of poses, which simplifies 3D-aware image synthesis to 2D image synthesis, resulting in lower FID scores but higher depth error, pose error and Jensen–Shannon Divergence score. EG3D, the method that uses pose annotations, exhibits good performance in terms of image quality and geometry quality. Our method yields results on par with EG3D and learns much better underlying pose distribution than the baselines. The quality of the geometry affects the multi-view consistency of the learned neural fields. Thus, EG3D and ours, which synthesize better shapes, can get better scores of reprojection errors. Note that our pose error is higher than that of EG3D, one of the reasons is that there exists the canonical view shift. As we do not provide any information on what the canonical view will be like, the model tends to learn a canonical view that leads to the best of its performance, which could be different from the standard. More discussions are in Sec. 4.4.

4.3. Ablation Study

We conduct ablation studies on FFHQ 256×256 to analyze the effectiveness of PoF3D. Evaluation results are shown in Tab. 2.

Table 2. **Ablation studies** on FFHQ dataset [12]. We analyze the influence of symmetry loss, pose-aware discriminator, and learning rate. See Sec. 4.3 for more details.

Model	FFHQ			
	FID_{50k} ↓	Depth↓	Pose↓	JS↓
w/o symmetry loss	4.50	0.41	0.12	0.20
w/o pose-aware D	**3.43**	1.30	0.19	0.21
w/o pose condition in D	3.51	1.48	0.19	0.26
lr $= 2.5e$-4	122.07	0.82	0.74	0.56
lr $= 2.5e$-6	10.20	0.70	0.16	0.38
Ours	4.99	**0.29**	**0.10**	**0.20**

Symmetry Loss. We render an image from the symmetrical camera view and leverage it to pose a constraint on the generated neural radiance field from another view. Without this constraint, although the network can synthesize visually pleasing images under the learned pose and capture a rough pose distribution, the underlying geometry has a tendency to flatten, resulting in higher depth error. The constraint from the symmetrical view can detect the flat shapes because the rendered image under the novel view deviates from the real distribution and shall be corrected by the model.

Pose-aware Discriminator. To verify the capability of the pose-aware discriminator, we compare it against the simple dual discriminator that removes the conditional label and purely performs real or fake discrimination. The simple dual discriminator does not provide any additional pose information to the generator other than the common realness score. Thus, the generator learns the pose freely and has difficulty aligning poses of data to share the same canonical view. For example, in the human face generation, one fake pose can correspond to images under different real poses. As a result, the pose error increases, and planar shapes are more frequently generated. We also report the results where only the pose conditioning discrimination is changed to the conventional pose-free discrimination. It drastically harms the geometry quality. With our pose-aware discriminator, the discriminator predicts the pose from both the fake and real data and uses it as the condition, which forces the generator to synthesize images under the same view as the conditional label, leading to better pose

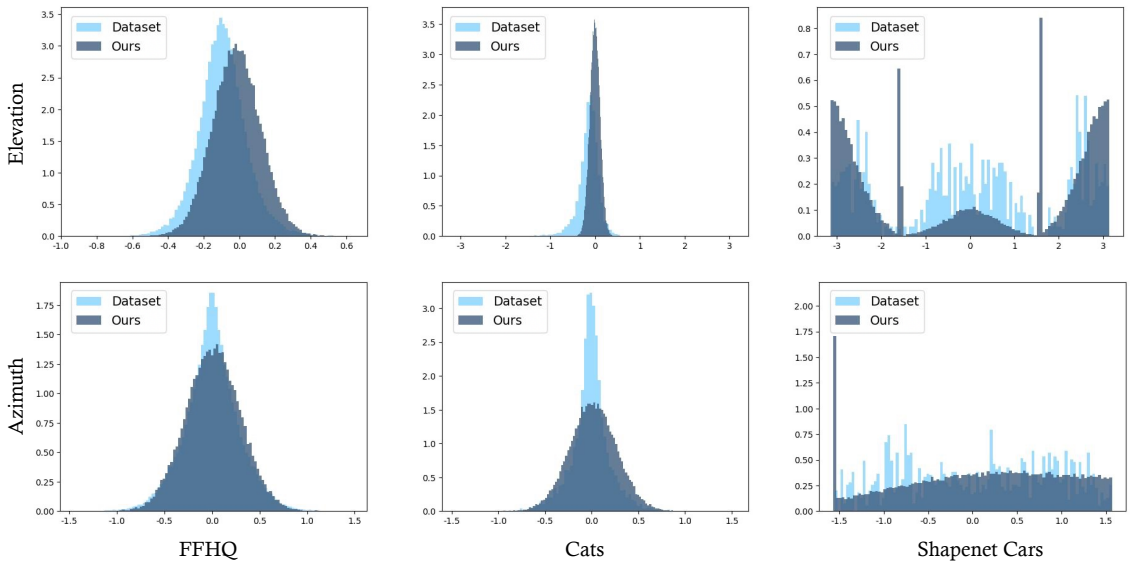

Figure 4. **Pose distribution learned by the pose-free generator.** The proposed approach well captures the pose distribution of the dataset. Note that, there is a slight shift of elevation distribution on the FFHQ dataset. The reason is that the model can automatically learn a canonical space for best performance during training.

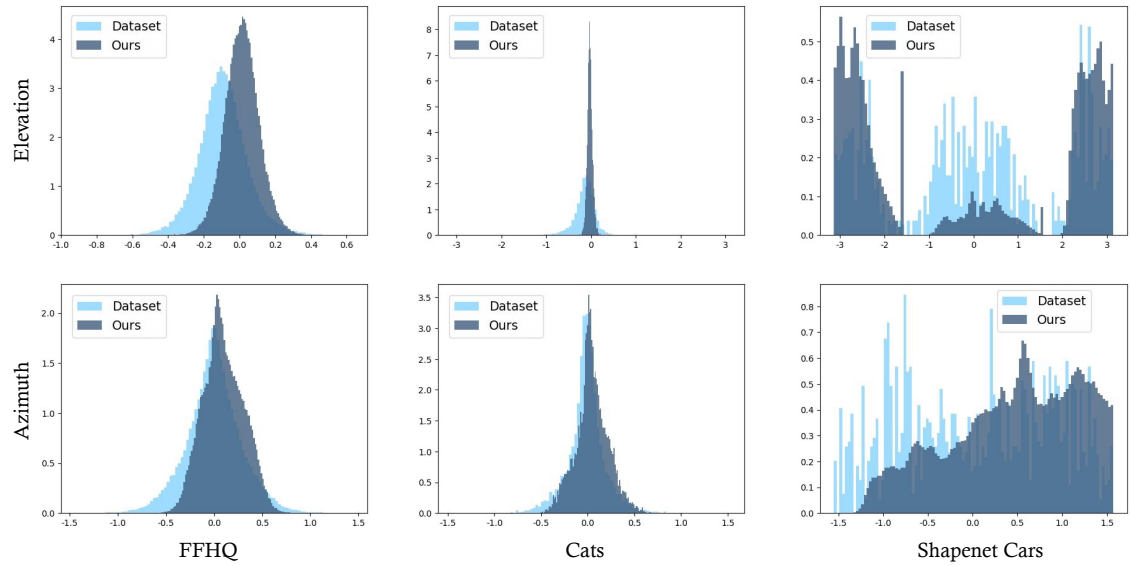

Figure 5. **Pose distribution learned by the pose-aware discriminator.** Results demonstrate that our discriminator can learn the pose distribution of the dataset. The issue of canonical space shift also exists from the discriminator perspective, same as Fig. 4.

alignment.

Learning Rate of Pose Learner. Learning rate of the pose branch in the generator matters a lot. With a large learning rate (*e.g.,* 2.5e-4 in Tab. 2), the generator always tries with new poses of large variation at the beginning and leaves the content learning behind, resulting in unstable training and higher tendency of mode collapse. As reported in Tab. 2, all the evaluation metrics experience a significant drop. On the contrary, if the learning rate is too small (*e.g.,* 2.5e-6 in Tab. 2), the generator cannot explore the pose distribution enough and easily gets stuck in a small range of pose distribution, putting more focus on the content learning. Consequently, the model fails to capture and understand the pose distribution, leading to worse shapes. A proper learning rate should balance the exploration of pose distribution and the learning of the content.

4.4. Pose Analysis

In this section, we analyze the pose distributions that are learned by our pose-free generator and pose-aware discriminator.

Poses in Pose-free Generator. To visualize the pose

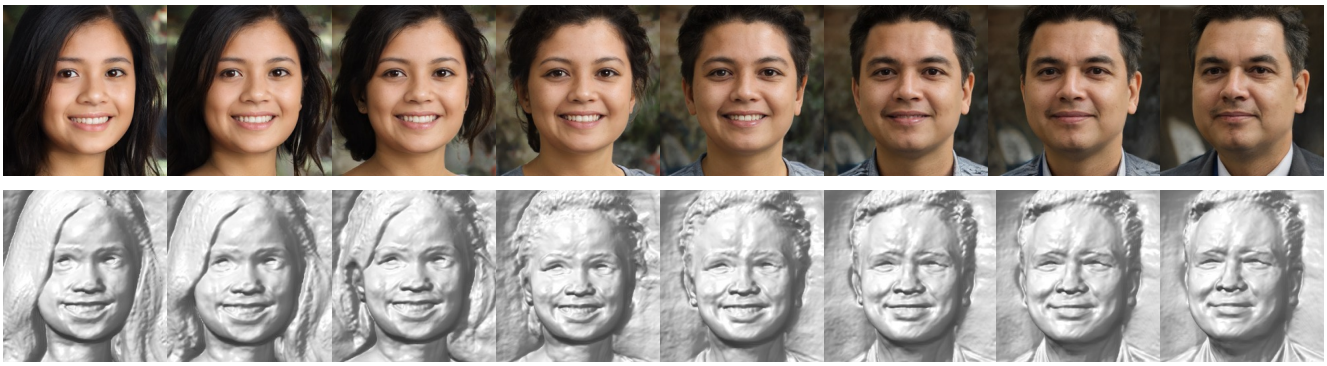

Figure 6. **Results of linear interpolation between two latent codes.** The appearance, the geometry, and the pose all change smoothly, indicating the latent space is semantically meaningful.

distribution learned in the pose-free generator, we randomly sample 50K latent codes and infer the poses from them. The pose distributions from the dataset and learned in the generator are shown in Fig. 4. Without any pose priors, our method can capture the pose range and distribution of the dataset in general. As observed in the figure for elevation distributions on the FFHQ dataset, there exists a distribution shift. The reason is that we do not release any information about the canonical view to the model, and therefore, the model can interpret the canonical view freely with different angles which might be different from the one defined in the dataset. Generally, the model will choose the one that leads to the best performance as its canonical view.

Poses in Pose-aware Discriminator. Since our discriminator is equipped with a pose predictor, we hire it to predict the poses for all data in the dataset and draw the distributions accordingly in Fig. 5. For FFHQ and Cats datasets, the distributions are well approximated, and the canonical view shift problem also exists in the discriminator. On Shapenet Cars, the distribution of azimuth is biased to one side. We conjecture the reason is that the model has difficulty distinguishing the front and rear of the car as they look very similar. As a result, the learned distribution shifts towards a hemisphere instead of the entire sphere.

4.5. Applications

Linearity in Latent Space. To demonstrate the latent space learned by PoF3D is semantically meaningful, we randomly sample two latent codes and linearly interpolate between them. The interpolation results are shown in Fig. 6. We can see that both the appearance and the underlying geometry are changing smoothly. Besides, as we infer the pose from the latent code, poses undergo smooth changes as well.

Real Image Inversion. One of the potential applications of PoF3D is to reconstruct the geometry from a real image and enable novel view synthesis. To achieve this goal, PTI [28], a GAN inversion method, is adopted to perform on our pose-free generator. In the previous methods [3] that demonstrate GAN inversion as an application, an off-the-shelf pose predictor is required to obtain the pose of the given image prior to performing GAN inversion. Ours, however, benefits from the changed formulation in which the camera pose is encoded in the latent space. Therefore, we can get not only the inverted neural radiance field but also the camera pose under which the given image is captured, when performing GAN inversion. The result is shown in Fig. 7. Our method can successfully invert the image and synthesize vivid images under novel views.

Figure 7. **Results of GAN inversion.** Thanks to the design of pose-free generator, we do not require pose estimation prior to performing inversion like previous works. Instead, we can directly invert the image to get the 3D representation and the corresponding pose. The resconstructed geometry is realistic and enables vivid novel view synthesis.

5. Conclusion

This work presents PoF3D, which learns 3D-aware image synthesis without using any pose priors. A pose-free generator is designed to infer camera poses directly from the latent space so that the requirement for pose priors in the previous works is removed. Besides, to help the generator better capture the underlying distribution, we make the discriminator pose-aware by asking it to first predict the camera pose and then use it as a condition when performing conditional real/fake discrimination. With such a design, experimental results demonstrate that our approach is able to synthesize high-quality images and high-fidelity shapes that are on par with state-of-the-art 3D synthesis methods without the need for manual pose tuning or dataset labeling.

Acknowledgement. This research has been made possible by funding support from the Research Grants Council of Hong Kong under the Research Impact Fund project R6003-21.

References

[1] Qingyan Bai, Yinghao Xu, Jiapeng Zhu, Weihao Xia, Yujiu Yang, and Yujun Shen. High-fidelity gan inversion with padding space. In *Eur. Conf. Comput. Vis.*, 2022. 2

[2] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Int. Conf. Comput. Vis.*, pages 5855–5864, 2021. 1

[3] 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 *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022. 1, 2, 4, 5, 6, 8

[4] 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 *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021. 1, 2

[5] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 2, 4, 6

[6] Yu Deng, Jiaolong Yang, Jianfeng Xiang, and Xin Tong. Gram: Generative radiance manifolds for 3d-aware image generation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022. 2, 4

[7] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, 2019. 4

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Adv. Neural Inform. Process. Syst.*, 2014. 1, 4

[9] 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. 1, 2, 3

[10] 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. In *Adv. Neural Inform. Process. Syst.*, 2017. 4

[11] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. In *Adv. Neural Inform. Process. Syst.*, 2021. 2

[12] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. 2, 4, 6

[13] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 2

[14] Zhengfei Kuang, Kyle Olszewski, Menglei Chai, Zeng Huang, Panos Achlioptas, and Sergey Tulyakov. Neroic: Neural rendering of objects from online image collections. *arXiv preprint arXiv:2201.02533*, 2022. 2

[15] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Int. Conf. Comput. Vis.*, 2015. 1

[16] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. 1, 2

[17] 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 *Eur. Conf. Comput. Vis.*, 2020. 1, 2

[18] Thu Nguyen-Phuoc, Chuan Li, Lucas Theis, Christian Richardt, and Yong-Liang Yang. Hologan: Unsupervised learning of 3d representations from natural images. In *Int. Conf. Comput. Vis.*, 2019. 2

[19] Thu H Nguyen-Phuoc, Christian Richardt, Long Mai, Yongliang Yang, and Niloy Mitra. Blockgan: Learning 3d object-aware scene representations from unlabelled images. *Adv. Neural Inform. Process. Syst.*, 2020. 2

[20] Michael Niemeyer and Andreas Geiger. Campari: Camera-aware decomposed generative neural radiance fields. In *International Conference on 3D Vision (3DV)*, 2021. 1, 2, 4, 5, 6

[21] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021. 1, 2

[22] 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 *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022. 2

[23] Xingang Pan, Xudong Xu, Chen Change Loy, Christian Theobalt, and Bo Dai. A shading-guided generative implicit model for shape-accurate 3d-aware image synthesis. In *Adv. Neural Inform. Process. Syst.*, 2021. 2

[24] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. 1, 2

[25] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 9054–9063, 2021. 1

[26] Abdal Rameen, Qin Yipeng, and Wonka Peter. Image2StyleGAN: How to embed images into the stylegan latent space? In *Int. Conf. Comput. Vis.*, 2019. 2

[27] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a StyleGAN encoder for image-to-image translation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021. 2

[28] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *ACM Trans. Graph.*, 2021. 8

[29] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016. 3

[30] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. Graf: Generative radiance fields for 3d-aware image synthesis. In *Adv. Neural Inform. Process. Syst.*, 2020. 1, 2

[31] Katja Schwarz, Axel Sauer, Michael Niemeyer, Yiyi Liao, and Andreas Geiger. Voxgraf: Fast 3d-aware image synthesis with sparse voxel grids. *Adv. Neural Inform. Process. Syst.*, 2022. 2

[32] Steven M Seitz, Brian Curless, James Diebel, Daniel Scharstein, and Richard Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2006. 3

[33] Zifan Shi, Sida Peng, Yinghao Xu, Yiyi Liao, and Yujun Shen. Deep generative models on 3d representations: A survey. *arXiv preprint arXiv:2210.15663*, 2022. 2

[34] Zifan Shi, Yujun Shen, Jiapeng Zhu, Dit-Yan Yeung, and Qifeng Chen. 3d-aware indoor scene synthesis with depth priors. In *Eur. Conf. Comput. Vis.*, 2022. 1

[35] Zifan Shi, Yinghao Xu, Yujun Shen, Deli Zhao, Qifeng Chen, and Dit-Yan Yeung. Improving 3d-aware image synthesis with a geometry-aware discriminator. In *Adv. Neural Inform. Process. Syst.*, 2022. 2

[36] Ivan Skorokhodov, Aliaksandr Siarohin, Yinghao Xu, Jian Ren, Hsin-Ying Lee, Peter Wonka, and Sergey Tulyakov. 3d generation on imagenet. In *Int. Conf. Learn. Represent.*, 2023. 2

[37] Ivan Skorokhodov, Sergey Tulyakov, Yiqun Wang, and Peter Wonka. Epigraf: Rethinking training of 3d gans. In *Adv. Neural Inform. Process. Syst.*, 2022. 2

[38] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for StyleGAN image manipulation. *ACM Trans. Graph.*, 2021. 2

[39] Tengfei Wang, Bo Zhang, Ting Zhang, Shuyang Gu, Jianmin Bao, Tadas Baltrusaitis, Jingjing Shen, Dong Chen, Fang Wen, Qifeng Chen, et al. Rodin: A generative model for sculpting 3d digital avatars using diffusion. *arXiv preprint arXiv:2212.06135*, 2022. 2

[40] Tengfei Wang, Ting Zhang, Bo Zhang, Hao Ouyang, Dong Chen, Qifeng Chen, and Fang Wen. Pretraining is all you need for image-to-image translation. *arXiv preprint arXiv:2205.12952*, 2022. 2

[41] Tengfei Wang, Yong Zhang, Yanbo Fan, Jue Wang, and Qifeng Chen. High-fidelity gan inversion for image attribute editing. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022. 2

[42] Zirui Wang, Shangzhe Wu, Weidi Xie, Min Chen, and Victor Adrian Prisacariu. NeRF−−: Neural radiance fields without known camera parameters. *arXiv preprint arXiv:2102.07064*, 2021. 2

[43] Weihao Xia and Jing-Hao Xue. A survey on 3d-aware image synthesis. *arXiv preprint arXiv:2210.14267*, 2022. 2

[44] Jiaxin Xie, Hao Ouyang, Jingtan Piao, Chenyang Lei, and Qifeng Chen. High-fidelity 3d gan inversion by pseudo-multi-view optimization. *arXiv preprint arXiv:2211.15662*, 2022. 2

[45] Xudong Xu, Xingang Pan, Dahua Lin, and Bo Dai. Generative occupancy fields for 3d surface-aware image synthesis. In *Adv. Neural Inform. Process. Syst.*, 2021. 2

[46] Yinghao Xu, Menglei Chai, Zifan Shi, Sida Peng, Ivan Skorokhodov, Aliaksandr Siarohin, Ceyuan Yang, Yujun Shen, Hsin-Ying Lee, Bolei Zhou, et al. Discoscene: Spatially disentangled generative radiance fields for controllable 3d-aware scene synthesis. *arXiv preprint arXiv:2212.11984*, 2022. 2

[47] Yinghao Xu, Sida Peng, Ceyuan Yang, Yujun Shen, and Bolei Zhou. 3d-aware image synthesis via learning structural and textural representations. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022. 1, 2, 3, 4

[48] Yinghao Xu, Yujun Shen, Jiapeng Zhu, Ceyuan Yang, and Bolei Zhou. Generative hierarchical features from synthesizing images. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021. 2

[49] Fei Yin, Yong Zhang, Xuan Wang, Tengfei Wang, Xiaoyu Li, Yuan Gong, Yanbo Fan, Xiaodong Cun, Ying Shan, Cengiz Oztireli, et al. 3d gan inversion with facial symmetry prior. *arXiv preprint arXiv:2211.16927*, 2022. 2

[50] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. Nerf++: Analyzing and improving neural radiance fields. *arXiv preprint arXiv:2010.07492*, 2020. 1

[51] Weiwei Zhang, Jian Sun, and Xiaoou Tang. Cat head detection-how to effectively exploit shape and texture features. In *Eur. Conf. Comput. Vis.*, 2008. 1, 2, 4, 6

[52] Xiaoming Zhao, Fangchang Ma, David Güera, Zhile Ren, Alexander G Schwing, and Alex Colburn. Generative multiplane images: Making a 2d gan 3d-aware. In *Eur. Conf. Comput. Vis.*, 2022. 2

[53] Yijun Zhou and James Gregson. Whenet: Real-time fine-grained estimation for wide range head pose. In *Brit. Mach. Vis. Conf.*, 2020. 4

[54] Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou. In-domain GAN inversion for real image editing. In *Eur. Conf. Comput. Vis.*, 2020. 2

[55] Jiapeng Zhu, Ceyuan Yang, Yujun Shen, Zifan Shi, Deli Zhao, and Qifeng Chen. Linkgan: Linking gan latents to pixels for controllable image synthesis. *arXiv preprint arXiv:2301.04604*, 2023. 2

[56] Jun-Yan Zhu, Zhoutong Zhang, Chengkai Zhang, Jiajun Wu, Antonio Torralba, Joshua B. Tenenbaum, and William T. Freeman. Visual object networks: Image generation with disentangled 3D representations. In *Adv. Neural Inform. Process. Syst.*, 2018. 2