Multi-View Consistent Generative Adversarial Networks for 3D-aware Image Synthesis

Xuanmeng Zhang[1,2] * Zhedong Zheng[1] Daiheng Gao[2]
Bang Zhang[2] Pan Pan[2] Yi Yang[3]
[1]ReLER, AAII, University of Technology Sydney
[2]DAMO Academy, Alibaba Group [3]Zhejiang University
{zhangxuanmeng.zxm, zdzheng12}@gmail.com
{daiheng.gdh, zhangbang.zb, panpan.pp}@alibaba-inc.com yangyics@zju.edu.cn

Abstract

*3D-aware image synthesis aims to generate images of objects from multiple views by learning a 3D representation. However, one key challenge remains: existing approaches lack geometry constraints, hence usually fail to generate multi-view consistent images. To address this challenge, we propose **M**ulti-**V**iew **C**onsistent **G**enerative **A**dversarial **N**etworks (**MVCGAN**) for high-quality 3D-aware image synthesis with geometry constraints. By leveraging the underlying 3D geometry information of generated images, i.e., depth and camera transformation matrix, we explicitly establish stereo correspondence between views to perform multi-view joint optimization. In particular, we enforce the photometric consistency between pairs of views and integrate a stereo mixup mechanism into the training process, encouraging the model to reason about the correct 3D shape. Besides, we design a two-stage training strategy with feature-level multi-view joint optimization to improve the image quality. Extensive experiments on three datasets demonstrate that MVCGAN achieves the state-of-the-art performance for 3D-aware image synthesis.*

1. Introduction

We study the problem of 3D-aware image synthesis, aiming at generating images with explicit control over the camera pose. Generating photorealistic and editable image content is a long-standing problem in computer vision and graphics. In the past years, generative adversarial networks (GAN) [19] have demonstrated impressive results in synthesizing high-resolution images of high quality from unstructured image collections [3, 8, 9, 22, 24–26, 64, 66]. Despite the tremendous success, most of the methods typically only learn the manifold of 2D images while ignoring the 3D representation of the scene.

Figure 1. Images synthesized by MVCGAN on the CELEBA-HQ [24] dataset.

Several works consider the task of 3D-aware image synthesis [1, 13, 20, 30, 38, 39, 67], which can generate images of objects from multiple views by learning a 3D-aware generative model. Different from 2D generative adversarial networks, 3D-aware image synthesis models learn 3D scene representations from images, such as voxels [38, 39], intermediate 3D primitives [30], and neural radiance fields (NeRF) [4, 13, 40, 46]. Among these approaches, NeRF-based approaches [4, 13, 40, 46] have gained a surge of interest due to the extraordinary performance for high-fidelity view synthesis. However, one key challenge remains in existing approaches [4, 40, 46]: they do not guarantee geometry constraints between views, hence usually fail to generate multi-view consistent images in some views.

In this paper, we address this problem by proposing

*This work was done during an internship at Alibaba.

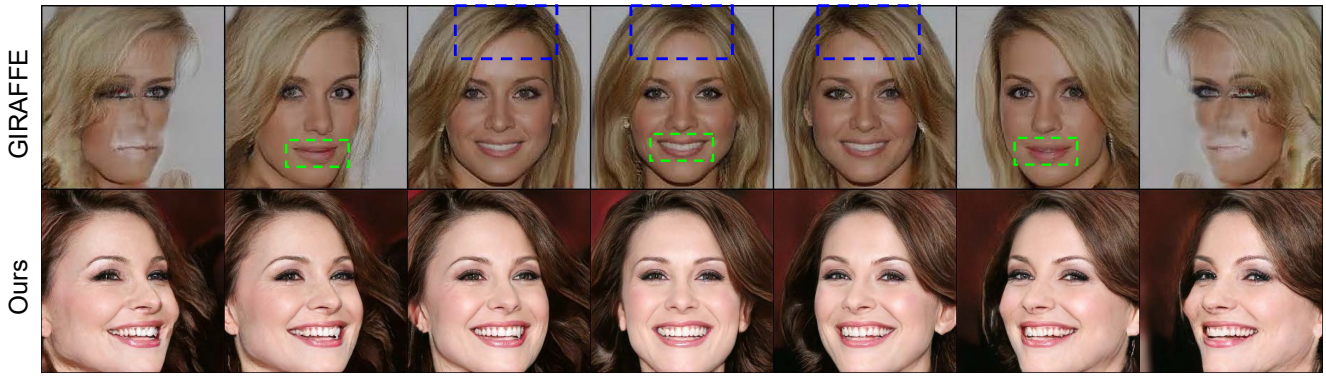

Figure 2. **Typical failure cases.** Taking a representative method GIRAFFE [40] as an example, the generated images in the first row have obvious appearance inconsistent artifacts between views, such as the direction of hair (blue box) and the opening mouth (green box). Besides, we notice that GIRAFFE [40] suffers from collapsed results under large pose variations (see the leftmost and rightmost pictures in the first row), which indicates that the model does not learn an appropriate 3D shape. In contrast, our method generates high-quality images with multi-view consistency (see the second row).

MVCGAN, a multi-view consistent generative model for high-quality 3D-aware image synthesis with geometry constraints (see Fig. 1). We first present typical failure cases of existing approach [40] in Fig. 2. Then we identify the cause of the inconsistent phenomenon between views: previous methods optimize a single view of the generated image independently while ignoring the geometry constraints between views (see Sec. 3.2.1). To tackle this problem, we take inspiration from the classical multi-view geometry methods [2, 6, 11, 18, 47, 65] and jointly optimize multiple views with geometry constraints. By leveraging the underlying 3D geometry information, we explicitly establish stereo correspondence between views through projective geometry. To encourage the network to reason about the correct 3D shape, we perform multi-view joint optimization by enforcing the photometric consistency between pairs of views with re-projection loss and integrating a stereo mixup mechanism into the training process. Therefore, the generator not only learns the manifold of 2D images, but also ensures the correctness of the underlying 3D shape.

Besides, we notice that NeRF-based generative approaches [4, 40, 46] typically struggle to render high-resolution images with fine details due to the huge computational complexity of NeRF model [36]. Existing methods [4, 40, 46] adopt different strategies to synthesize high-resolution images. However, they all have limitations. GRAF [46] introduces a multi-scale patch-based discriminator, which causes uneven image quality and local overfitting to the last batch. pi-GAN [46] increases the resolution of the generator by sampling rays more densely, which still requires intensive memory consumption. GIRAFFE [40] combines the 3D representation with a neural rendering pipeline, which suffers from collapsed results under large pose variations. In this paper, we adopt a hybrid MLP-CNN architecture to disentangle the geometry of 3D shape from the fine details of 2D appearance. In particular, the MLP-based NeRF model [36] renders the geometry of 3D shape, and the CNN-based decoder produces fine details for 2D appearance. The structure can generate photorealistic high-resolution images while alleviating the computation-intensive problem.

Overall, our contributions are summarized as follows:

1. We identify one challenging problem of missing the geometry constraints in 3D-aware image synthesis, which leads to inconsistent images across views.

2. We propose a multi-view consistent generative model (MVCGAN) for high-quality 3D-aware image synthesis. By establishing the geometry constraints, we jointly optimize multiple views to guarantee the geometry consistency between views. Besides, we design a two-stage training strategy with the feature-level multi-view joint optimization to further improve the image quality.

3. We demonstrate the effectiveness of the proposed approach through evaluating on various datasets, *i.e.*, CELEBA-HQ [24], FFHQ [25], and AFHQv2 [9]. Extensive experiments substantiate that MVCGAN achieves the state-of-the-art performance for 3D-aware image synthesis.

2. Related Work

Multi-view Geometry. A large number of approaches reconstruct 3D structure with multi-view geometry constraints as supervision signals, such as COLMAP [45] and ORB-SLAM [37]. In recent years, Some deep learning techniques [18, 57, 65] also combine traditional approaches [6, 10, 49] to address 3D vision problems. Inspired by the classical multi-view geometry methods [2, 6, 11, 18,

47,65], we explicitly involve the geometry constraints in the training process for learning a reasonable 3D shape.

Neural Radiance Fields. Recently, using volumetric rendering and implicit function to synthesize novel views of a scene has gained a surge of interest. Mildenhall *et al.* [36] represent complex scenes as Neural Radiance Fields (NeRF) for novel view synthesis by optimizing an implicit continuous volumetric scene function. Due to the simplicity and extraordinary performance, NeRF [36] has been extended to plenty of variants, *e.g.*, faster inference [17, 32, 43, 43, 44], pose estimation [23, 31, 34, 53, 58], generalization [5, 7, 46, 50, 59], video [16, 28, 29, 41, 55], and depth estimation [54].

3D-aware Image Synthesis. Several recent works have investigated how to incorporate 3D representation into generative models [1, 13, 20, 30, 38, 39, 67]. Nguyen *et al.* [38] combine a strong inductive bias about the 3D world with deep generative models to learn disentangled representations of 3D objects. HoloGAN [38] provides control over the pose of generated objects through rigid-body transformations of the learned 3D features. Schwarz *et al.* [46] propose GRAF, a generative model radiance fields model for 3D-aware image synthesis from unposed 2D images. pi-GAN [4] adopts a SIREN-based neural implicit representation with periodic activation functions as the backbone of the generator. By representing scenes as compositional generative neural feature fields, GIRAFFE [40] disentangles individual objects from the background. However, these methods optimize a single view of the generated scene independently and ignore the geometry constraints between views.

3. Method

Our goal is to generate photorealistic high-resolution images with explicit control over the camera pose while maintaining multi-view consistency. We now present the main components of the proposed method. First, we briefly review the background of NeRF-based generative adversarial networks [4, 40, 46] and identify the limitations of previous methods (see Sec. 3.1). Second, we analyze the cause of the multi-view inconsistency problem and propose the image-level multi-view joint optimization to address this problem (see Sec. 3.2.1). Besides, we design a two-stage training strategy that extends multi-view optimization to the feature level to generate high-resolution images with fine details. (see Sec. 3.2.2). Finally, we describe the training details in Sec. 3.3. Fig. 4 shows the framework of the proposed method.

3.1. Preliminaries

Neural Radiance Fields. Neural radiance field (NeRF) synthesizes novel views of the scene by optimizing a fully-connected network using a set of input views. The MLP network maps a continuous 5D coordinate (3D location $\mathbf{x}$ and 2D viewing direction $\mathbf{d}$) to an emitted color $\mathbf{c}$ and volume density σ [36]:

$$(\gamma(\mathbf{x}), \gamma(\mathbf{d})) \rightarrow (\mathbf{c}, \sigma), \tag{1}$$

where γ indicates the positional encoding mapping function. To render the neural radiance field from a viewpoint, Mildenhall *et al.* [36] use classic volume rendering to accumulate the output colors $\mathbf{c}$ and densities σ into an image.

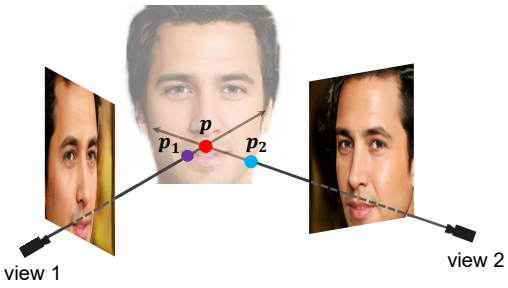

Figure 3. **Visualization of shape-radiance ambiguity.** For illustration, we assume the p (the red dot) is the location of correct geometry, and p_1 (the violet dot) and p_2 (the blue dot) are incorrect geometries. In the absence of geometry constraints, the model can fit to incorrect geometry p_1 in view 1 and p_2 in view 2 independently to simulate the effect of the correct geometry p.

Generative Radiance Fields. Generative neural radiance fields aim to learn a model for synthesizing novel scenes by training on unposed 2D images. Schwarz *et al.* [46] adopt an adversarial framework to train a generative model for radiance fields (GRAF). The generative radiance field is conditioned on a shape code z_s and an appearance code z_a:

$$(\gamma(\mathbf{x}), \gamma(\mathbf{d}), z_s, z_a)) \rightarrow (\mathbf{c}, \sigma). \tag{2}$$

Following GRAF [46], Niemeyer *et al.* [40] introduce a compositional generative neural feature field (GIRAFFE). Inspired by StyleGAN [25], Chan *et al.* [4] instead propose periodic implicit generative adversarial networks (pi-GAN) with feature-wise linear modulation (FiLM) conditioning.

Limitations. We notice two limitations of existing approaches [4, 40, 46]. First, they do not guarantee geometry constraints between different views. Consequently, they usually suffer from collapsed results under large pose variations or have obvious inconsistent artifacts across views. Second, the rendered high-resolution images typically lack realism and fine details due to the huge computational cost of NeRF model.

3.2. Multi-view Joint Optimization

3.2.1 Image-level Multi-view Joint Optimization

Shape-radiance Ambiguity. In this part, we analyze the cause of multi-view inconsistency problem in NeRF-based generative models. We observe that optimizing the radiance

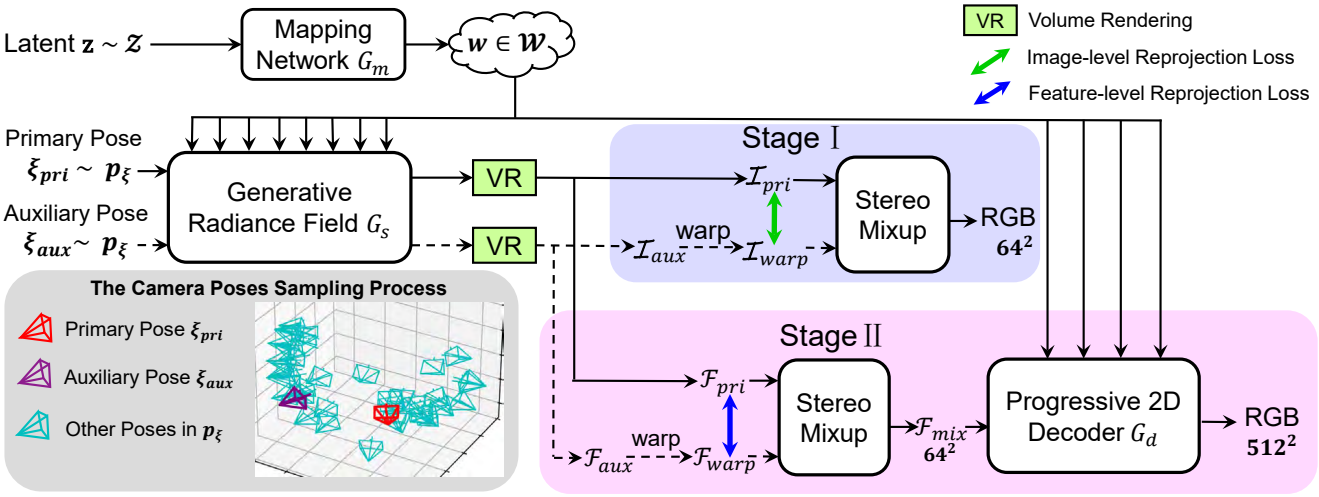

Figure 4. **The structure of generator G_θ.** During training, the generative radiance field network G_s takes primary pose ξ_{pri} and auxiliary pose ξ_{aux} as input. The mapping network G_m maps the input latent z to intermediate latent w, which conditions both the generative radiance field network G_s and the progressive 2D decoder G_d. In Stage I, we directly render primary image $\mathcal{I}_{pri}$ and auxiliary image $\mathcal{I}_{aux}$ with the color and density output from G_s. Then we perform image-level multi-view joint optimization and output a low-resolution RGB image (64^2). In Stage II, we instead use volume rendering to accumulate 2D feature maps at low resolution (64^2), and then perform multi-view optimization at the feature level. The progressive 2D decoder G_d upsamples 2D feature map $\mathcal{F}_{mix}$ to a high-resolution RGB image (128^2, 256^2, 512^2) for fine 2D details. During inference, only the primary pose is required without auxiliary pose (the dotted lines do not participate in inference).

fields from a set of 2D training images can encounter critical degenerate solutions in the absence of geometry constraints. This phenomenon is referred to as shape-radiance ambiguity [61], in which the model can fit the training images with inaccurate 3D shape by a suitable choice of radiance field at each surface point (see Fig 3). To better illustrate the shape-radiance ambiguity [61], we warp the rendered images from view 1 to view 2 based on the underlying depth and camera transformation matrix $[R, t]$ (see the details of warping process in Fig. 5 and Eq. 4). We find the warped image shows a wrong appearance, which verifies the assumption of degenerate solutions to the learned 3D shape. To avoid the shape-radiance ambiguity [61], NeRF [36] requires a large number of posed training images from different input views for the scene. However, generative radiance fields have neither annotated camera poses nor sufficient multi-view images in the training dataset. Consequently, the generative model can synthesize reasonable images in some views but produce poor renderings in other views (see Fig. 2).

Warping Process. To alleviate the shape-radiance ambiguity [61], we propose to establish multi-view geometry constraints [2, 6, 11, 18, 47, 65] via the warping process between views. First, following pi-GAN [4], we adopt a style-based generator which contains a synthesis network G_s (a SIREN-based [4, 48] generative radiance field) and a mapping network G_m (a simple MLP network with ReLU) (see Fig. 4). Given a latent code $z \in \mathbb{R}^{256}$ in the input latent space $\mathcal{Z}$, the mapping network G_m:$\mathcal{Z} \rightarrow \mathcal{W}$ can produce the intermidiate latent $w \in \mathbb{R}^{256}$, which controls the synthesis network G_s at each layer. Second, instead of only optimizing a single view independently, we aim to optimize multiple views jointly to maintain the 3D consistency across views. As shown in the left of Fig. 4, we randomly sample two camera poses, *i.e.*, the primary pose ξ_{pri} and the auxiliary pose ξ_{aux}, from the pose distribution p_ξ. Taking ξ_{pri} and ξ_{aux} as input, the generative model G_s synthesizes two views of the generated images separately: the primary image $\mathcal{I}_{pri}$ and the auxiliary image $\mathcal{I}_{aux}$. Then we can build geometry constraints between ξ_{pri} and ξ_{aux} via image warping, which reconstructs the primary view by sampling pixels from the auxiliary image $\mathcal{I}_{aux}$. Specifically, for each point v_{pri} in the

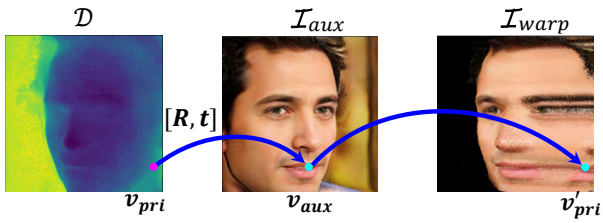

Figure 5. **Illustration of the warping process.** For each pixel v_{pri} in the primary image $\mathcal{I}_{pri}$, we first calculate the location of v_{aux} (the corresponding pixel of v_{pri} in the auxiliary image $\mathcal{I}_{aux}$) based on the depth value $\mathcal{D}(v_{pri})$ and camera transformation matrix $[R, t]$. Then we can reconstruct the pixel v'_{pri} of the warped image $\mathcal{I}_{warp}$ from the primary view using the value of pixel v_{aux}. We observe that the warped image has a wrong appearance, which verifies the incorrect geometry shape learned by model.

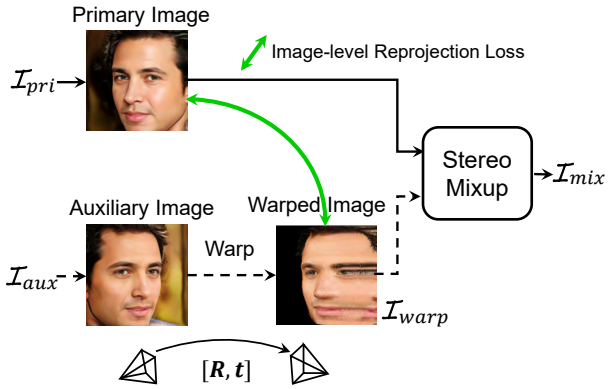

Figure 6. **Image-level multi-view joint optimization.** We enforce the photometric consistency between the primary image and the warped image by minimizing the image-level re-projection loss. Besides, we integrate a stereo mixup module to encourage the warped image to be similar to a real image. The dotted line does not participate in the inference stage.

primary image $\mathcal{I}_{pri}$, we first find the corresponding pixel v_{aux} in the auxiliary image $\mathcal{I}_{aux}$ through the stereo correspondence, and then reconstruct the pixel v'_{pri} of the warped image $\mathcal{I}_{warp}$ in primary view using the value of v_{aux} (see Fig. 5). Next, we present a detailed calculation procedure of the warping process. The stereo correspondence is calculated based on the depth map $\mathcal{D}$ of the primary image and camera transformation matrix from ξ_{pri} to ξ_{aux}. The depth can be rendered in a similar way as rendering the color image [12, 36]. Given the pixel v_{pri} from the primary view, the depth value $\mathcal{D}(v_{pri})$ is formulated as:

$$\begin{aligned} &\mathcal{D}(v_{pri}) = \sum_{i=1}^{N} T_i(1 - exp(-\sigma_i\delta_i))d_i, \\ &\text{where } T_i = exp(-\sum_{j=1}^{i} \sigma_j\delta_j), \end{aligned} \tag{3}$$

where N is the number of samples in the camera ray, $\delta_i = d_{i+1} - d_i$ is the distance between adjacent sample points and σ_i is the volume density of sample i (refer to [12, 36] to see more details). With the depth value $\mathcal{D}(v_{pri})$, we can obtain the homogeneous coordinates h_{pri} of pixel v_{pri} in the primary camera coordinate system through perspective projection. Then the projected coordinates h_{aux} in the auxiliary view can be calculated as:

$$h_{aux} = K[R, t]\mathcal{D}(v_{pri})K^{-1}h_{pri}, \tag{4}$$

where the camera intrinsics K are known parameters and the camera transformation matrix $[R, t]$ can be calculated from the primary pose ξ_{pri} and the auxiliary pose ξ_{aux}. Finally, we can reconstruct the pixel v'_{pri} in the warped image $\mathcal{I}_{warp}$ from the primary view using the value of pixel v_{aux} (located in h_{aux} of $\mathcal{I}_{aux}$).

Image-level Joint Optimization. After obtaining the warped image $\mathcal{I}_{warp}$, we perform image-level multi-view joint optimization by enforcing the photometric consistency and employing a stereo mixup module (see Fig. 6). To satisfy the geometry constraints between views, we enforce the photometric consistency across views by minimizing the re-projection loss between the primary image $\mathcal{I}_{pri}$ and the warped image $\mathcal{I}_{warp}$. Following the common practice in image reconstruction [18, 33, 42, 52, 62, 65], we formulate the image-level re-projection loss as the combination of L1 [62] and SSIM [52]:

$$L_{ir} = (1-\mu)||\mathcal{I}_{pri} - \mathcal{I}_{warp}||_1 + \frac{\mu}{2}(1 - SSIM(\mathcal{I}_{pri}, \mathcal{I}_{warp})), \tag{5}$$

where SSIM is a perceptual metric of image structural similarity and $\mu = 0.85$ empirically. In addition to being similar to the primary image, the warped image should also look like a real image. A straightforward method is introducing two discriminators. One is to compare the warped image $\mathcal{I}_{warp}$ with an arbitrary real image sampled from the training dataset, and the other one compares the primary image $\mathcal{I}_{pri}$. However, introducing extra modules can increase the computation complexity. Inspired by the *mixup* strategy [60], we instead propose a stereo mixup module to optimize both $\mathcal{I}_{pri}$ and $\mathcal{I}_{warp}$ by constructing a virtual mixed image:

$$\mathcal{I}_{mix} = \eta\mathcal{I}_{pri} + (1 - \eta)\mathcal{I}_{warp}, \tag{6}$$

where η is a dynamic number randomly sampled from the range of $[0, 1]$ in every training iteration, and $\mathcal{I}_{mix}$ is the input of discriminator. It is worth noting that the auxiliary pose is introduced to construct the geometry constraints, and thus only required in the training process. In the inference stage, the generative model only takes the primary pose ξ_{pri} and latent code z as input to generate the primary image directly.

3.2.2 Feature-level Multi-view Joint Optimization

In practice, we also encounter one practical challenge: NeRF-based generative models [4, 40, 46] typically struggle to render high-resolution images with fine details due to the huge computational of NeRF [36] model. To render images with both fine 2D details and correct 3D shape, we design a two-stage training strategy and extend multi-view optimization to the feature level. We begin training at a low resolution (64^2) in Stage I, and then increase to high resolutions (128^2, 256^2, 512^2) in Stage II (see Fig. 4). In Stage I, we directly render primary and auxiliary images with the color and density output from the generative radiance field network G_s. With the guidance of geometry constraints, we perform image-level multi-view joint optimization to enhance the geometric reasoning ability of the model. In Stage

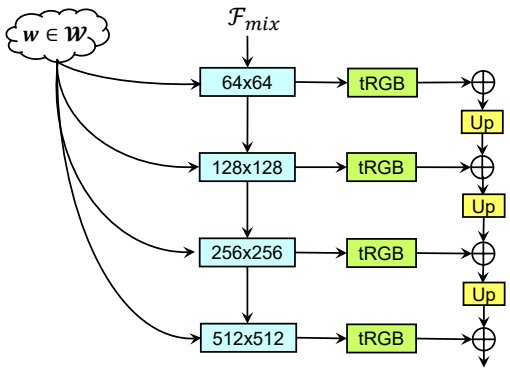

Figure 7. **Progressive 2D decoder G_d.** During training, the decoder takes the stereo mixup feature $\mathcal{F}_{mix}$ (produced by $\mathcal{F}_{pri}$ and $\mathcal{F}_{warp}$) as input at low resolution (64^2). Then the intermediate latent w conditions the decoder at each layer. Here tRGB denotes the 1x1 convolutions which convert the high-dimensional features to RGB images, and Up denotes the bilinear upsampling operation.

II, to alleviate the computation-intensive problem of rendering high-resolution images, we instead train the model via feature-level multi-view optimization for better visual quality. First, we adopt a hybrid MLP-CNN architecture to disentangle the geometry of 3D shape from fine details of 2D appearance. Then we generalize volume rendering [40] to the feature level by rendering 2D primary feature map $\mathcal{F}_{mix}$ at low resolution (64^2):

$$\mathcal{F}_{pri} = \sum_{i=1}^{N} T_i(1 - exp(-\sigma_i \delta_i)) f_i, \tag{7}$$

where $f_i \in \mathbb{R}^{256}$ is the feature before the final layer of G_s, and other symbols are defined in Eq. 3. The auxiliary feature map $\mathcal{F}_{aux}$ is rendered in the same way as $\mathcal{F}_{pri}$, and the warped feature map $\mathcal{F}_{warp}$ can be obtained through the warping process. Second, we perform multi-view feature-level joint optimization on low-resolution feature maps (64^2). To enforce the geometry consistency in the feature space, we take the implicit diversified Markov Random Fields (MRF) loss [51] as the feature-level reprojection loss:

$$L_{fr} = L_{mrf}(\mathcal{F}_{pri}, \mathcal{F}_{warp}), \tag{8}$$

which can encourage the model to capture high-frequency geometry details [15]. Then the stereo mixup mechanism is also applied to the 2D feature maps: $\mathcal{F}_{mix} = \eta \mathcal{F}_{pri} + (1 - \eta)\mathcal{F}_{warp}$. Third, we increase the resolution with a style-based 2D decoder [25] G_d, which takes $\mathcal{F}_{mix}$ as input and then upsamples to high-resolution RGB image (see Fig. 7). The 2D decoder G_d is conditioned by the mapping network G_m through adaptive instance normalization (AdaIN) [14, 22, 25]. As training progresses, we adopt the progressive growing strategy to grow the generator for higher resolution [24]. When new layers are added to G_d, we use skip connections to fade the inserted layers in smoothly to stabilize and speed up the training process [24, 26].

	CELEBA-HQ		FFHQ		AFHQv2
	256^2	512^2	256^2	512^2	256^2
GRAF [46]	47.5	57.7	67.2	71.2	75.8
pi-GAN [4]	39.7	41.8	38.1	39.9	42.0
GIRAFFE [40]	36.0	36.2	34.6	37.7	29.2
Ours	**11.8**	**12.9**	**13.7**	**13.4**	**17.1**

Table 1. Quantitative comparison. We calculate FID between 20,000 generated and real images.

3.3. Training Details

We use a progressive growing convolutional discriminator D_ϕ to compare the fake image produced by generator G_θ and real image $\mathcal{I}$ sampled from the training data with distribution $p_\mathcal{D}$. We train MVCGAN using a non-saturating GAN objective with R_1 gradient penalty [35] and the proposed geometry-constrained objective L_{re} as the total loss:

$$\begin{aligned}\mathcal{V}(\theta, \phi) = \mathbf{E}_{z \sim \mathcal{Z}, \xi_{pri} \sim p_\xi, \xi_{aux} \sim p_\xi}[f(D_\phi(G_\theta(z, \xi_{pri}, \xi_{aux}))] \\ + \mathbf{E}_{\mathcal{I} \sim p_\mathcal{D}}[f(-D_\phi(\mathcal{I})) - \lambda ||\nabla D_\phi(\mathcal{I})||^2] + L_{re},\end{aligned} \tag{9}$$

where $f(t) = -log(1 + exp(-t))$, $L_{re} = L_{ir}$ for Stage I (see Eq. 5), $L_{re} = L_{fr}$ for Stage II (see Eq. 8), and $\lambda = 10$. More implementation details can be found in the supplementary material.

4. Experiments

4.1. Experimental Settings

Datasets. We conduct experiments on three widely-used high-resolution image datasets: CELEBA-HQ [24], FFHQ [25], and AFHQv2 [9]. We choose the cat face images in the AFHQv2 [9] dataset to conduct experiments for a fair comparison with previous works [4, 40, 46].

4.2. Comparison with SOTA

For quantitative comparison, we report Frechet Inception Distance (FID) [21] to evaluate image quality. We compare our approach against three state-of-the-art 3D-aware image synthesis methods: GRAF [46], GIRAFFE [40] and pi-GAN [4]. As shown in Tab. 1, our method consistently outperforms other methods [4, 40, 46] on all datasets [9, 24, 25] by a large margin. We also visualize the generated images on FFHQ [25] and AFHQv2 [9] datasets for qualitative comparison. As illustrated in Fig. 8, we render images from a wide range of viewpoints. We observe that GRAF [46],

(a) Results on FFHQ [25].

(b) Results on AFHQv2 [9].

Figure 8. Qualitative comparison at 512^2 resolution with GRAF [46], GIRAFFE [40], and pi-GAN [4].

GIRAFFE [40] and pi-GAN [4] either fail to synthesize reasonable results under large view variations or have obvious multi-view inconsistent artifacts. By comparison, our method achieves the best performance both in visual quality and multi-view consistency. Please refer to the supplementary material for more visualization results.

4.3. Ablation Studies

Image-level and Feature-level Optimization. We conduct ablation studies to help understand the individual contributions of image-level and feature-level multi-view joint optimization. From Fig. 10 (a), we observe that the generated

Figure 9. **Style interpolation.** We perform linear interpolation both in the intermediate latent and camera pose space.

(a) With image-level multi-view join optimization (FID=22.5).

(b) With feature-level multi-view join optimization (FID=13.7).

Figure 10. Ablation study on FFHQ [25] at 256^2 resolution.

images maintain the multi-view consistency under poses variations (FID=22.5), indicating the image-level optimization can guide the model to learn a reasonable 3D shape. With feature-level optimization (see Fig. 10 (b)), our approach can further improve the visual quality of generated images with fine 2D details (FID=13.7).

Shape-detail Disentanglement. Besides, we design a style mixing experiment to study what kinds of representations the generative radiance field G_s and progressive 2D decoder G_d learned respectively. Specifically, we input two latent codes z_A and z_B into the mapping network G_m, and obtain the corresponding intermediate latent w_A, w_B in $\mathcal{W}$ space. Then we can generate style mixing images by applying w_A and w_B to control the different parts of the generator (G_s and G_d). As shown in Fig. 11, we observe that controlling G_s changes the 3D shape (identity and pose) while controlling G_d changes 2D appearance details (colors of skins, hair, and beard). The results verify that the hybrid MLP-CNN architecture can disentangle the geometry of 3D shape from fine details of 2D appearance.

Style Interpolation. We also conduct the style interpolation experiments to investigate the intermediate latent w learned by the mapping network G_m. Given two generated images, we perform linear interpolation both in the intermediate latent space $\mathcal{W}$ and the camera pose space. As illustrated in Fig. 9, the smooth transition of both pose and appearance demonstrates that our model learns semantically meaningful intermediate latent space $\mathcal{W}$.

5. Conclusion and Discussion

We present a multi-view consistent generative model (MVCGAN) for 3D-aware image synthesis. The key idea underpinning the proposed method is to enhance the geometric reasoning ability of the generative model by introducing geometry constraints. Extensive experiments

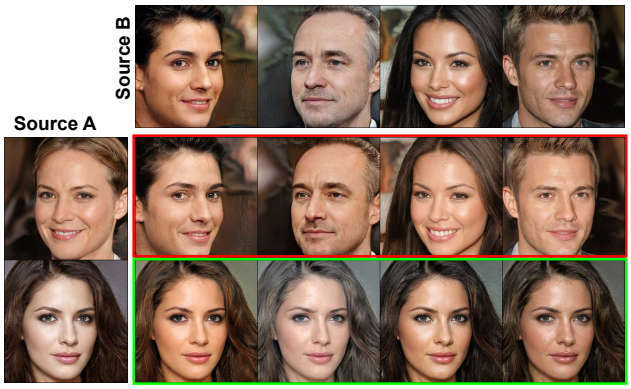

Figure 11. **Style mixing.** The source A and B images are generated from their input latent codes z_A and z_B. The images in the red box are generated by applying the w_B (the intermediate latent corresponding to z_B) to G_s and w_A (corresponding to z_A) to G_d. The images in the green box are generated by applying the w_A to G_s and w_B to G_d.

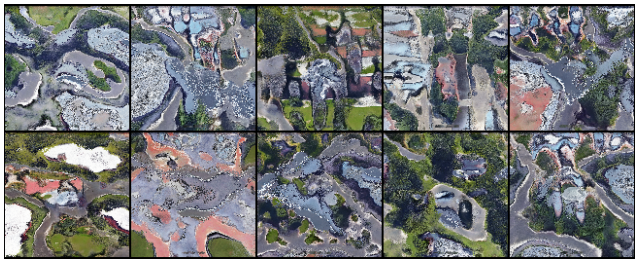

Figure 12. **Failure cases.** Our method does not perform well in scenarios with multiple objects and complex backgrounds. For example, our model fails to synthesize high-quality images on the University-1652 dataset [63].

demonstrate that MVCGAN achieves the state-of-the-art performance for 3D-aware image synthesis.

Limitations and future work. In this paper, our method mainly focuses on single-object scenes with simple backgrounds, and does not work well in multi-object and complex background-attached scenes (see Fig. 12). To extend to the scenarios with complex background and multiple objects, one possible way is to learn a compositional radiance field that can model the foreground and background separately [56]. To render the whole scene, the geometry relationships between foreground objects and the background can be established by combing depth maps and occlusion maps. In the future, we will incorporate extra image annotations to handle more complex real-world scenarios.

References

[1] Hassan Abu Alhaija, Siva Karthik Mustikovela, Andreas Geiger, and Carsten Rother. Geometric image synthesis. In *ACCV*, 2018. 1, 3

[2] Alex M Andrew. Multiple view geometry in computer vision. *Kybernetes*, 2001. 2, 4

[3] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. In *ICLR*, 2018. 1

[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 *CVPR*, 2021. 1, 2, 3, 4, 5, 6, 7, 11

[5] Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. *arXiv preprint arXiv:2103.15595*, 2021. 3

[6] Shenchang Eric Chen and Lance Williams. View interpolation for image synthesis. In *Conference on Computer graphics and interactive techniques*, 1993. 2, 4

[7] Julian Chibane, Aayush Bansal, Verica Lazova, and Gerard Pons-Moll. Stereo radiance fields (srf): Learning view synthesis for sparse views of novel scenes. In *CVPR*, 2021. 3

[8] 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 *CVPR*, 2018. 1

[9] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *CVPR*, 2020. 1, 2, 6, 7, 11

[10] Robert T Collins. A space-sweep approach to true multi-image matching. In *CVPR*, 1996. 2

[11] Paul E Debevec, Camillo J Taylor, and Jitendra Malik. Modeling and rendering architecture from photographs: A hybrid geometry-and image-based approach. In *Conference on Computer graphics and interactive techniques*, 1996. 2, 4

[12] Kangle Deng, Andrew Liua, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised nerf: Fewer views and faster training for free. *arXiv preprint arXiv:2107.02791*, 2021. 5

[13] Terrance DeVries, Miguel Angel Bautista, Nitish Srivastava, Graham W. Taylor, and Joshua M. Susskind. Unconstrained scene generation with locally conditioned radiance fields. In *ICCV*, 2021. 1, 3

[14] Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur. In *A learned representation for artistic style*, 2020. 6

[15] 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. 6

[16] Chen Gao, Ayush Saraf, Johannes Kopf, and Jia-Bin Huang. Dynamic view synthesis from dynamic monocular video. *arXiv preprint arXiv:2105.06468*, 2021. 3

[17] Stephan J Garbin, Marek Kowalski, Matthew Johnson, Jamie Shotton, and Julien Valentin. Fastnerf: High-fidelity neural rendering at 200fps. *arXiv preprint arXiv:2103.10380*, 2021. 3

[18] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *ICCV*, 2019. 2, 4, 5

[19] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014. 1

[20] Paul Henderson, Vagia Tsiminaki, and Christoph H Lampert. Leveraging 2d data to learn textured 3d mesh generation. In *CVPR*, 2020. 1, 3

[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. In *NeurIPS*, 2017. 6

[22] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, 2017. 1, 6

[23] Yoonwoo Jeong, Seokjun Ahn, Christopehr Choy, Animashree Anandkumar, Minsu Cho, and Jaesik Park. Self-calibrating neural radiance fields. In *ICCV*, 2021. 3

[24] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *ICLR*, 2018. 1, 2, 6, 11, 13

[25] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 1, 2, 3, 6, 7, 8, 11, 14

[26] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, 2020. 1, 6

[27] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 11

[28] Tianye Li, Mira Slavcheva, Michael Zollhoefer, Simon Green, Christoph Lassner, Changil Kim, Tanner Schmidt, Steven Lovegrove, Michael Goesele, and Zhaoyang Lv. Neural 3d video synthesis, 2021. 3

[29] Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. Neural scene flow fields for space-time view synthesis of dynamic scenes. *https://arxiv.org/abs/2011.13084*, 2020. 3

[30] Yiyi Liao, Katja Schwarz, Lars Mescheder, and Andreas Geiger. Towards unsupervised learning of generative models for 3d controllable image synthesis. In *CVPR*, 2020. 1, 3

[31] Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. Barf: Bundle-adjusting neural radiance fields. In *ICCV*, 2021. 3

[32] David B Lindell, Julien NP Martel, and Gordon Wetzstein. Autoint: Automatic integration for fast neural volume rendering. In *CVPR*, 2021. 3

[33] Xiaoyang Lyu, Liang Liu, Mengmeng Wang, Xin Kong, Lina Liu, Yong Liu, Xinxin Chen, and Yi Yuan. Hr-depth: High resolution self-supervised monocular depth estimation. In *AAAI*, 2021. 5

[34] Quan Meng, Anpei Chen, Haimin Luo, Minye Wu, Hao Su, Lan Xu, Xuming He, and Jingyi Yu. Gnerf: Gan-based neural radiance field without posed camera. *arXiv preprint arXiv:2103.15606*, 2021. 3

[35] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for gans do actually converge? In *ICML*, 2018. 6

[36] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 2, 3, 4, 5, 11

[37] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. Orb-slam: a versatile and accurate monocular slam system. *IEEE Transactions on Robotics*, 31(5):1147–1163, 2015. 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 *CVPR*, 2019. 1, 3

[39] Thu Nguyen-Phuoc, Christian Richardt, Long Mai, Yong-Liang Yang, and Niloy J Mitra. Blockgan: Learning 3d object-aware scene representations from unlabelled images. In *NeurIPS*, 2020. 1, 3

[40] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *CVPR*, 2021. 1, 2, 3, 5, 6, 7, 11

[41] 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 *CVPR*, 2021. 3

[42] Sudeep Pillai, Rareş Ambruş, and Adrien Gaidon. Superdepth: Self-supervised, super-resolved monocular depth estimation. In *ICRA*, 2019. 5

[43] Daniel Rebain, Wei Jiang, Soroosh Yazdani, Ke Li, Kwang Moo Yi, and Andrea Tagliasacchi. Derf: Decomposed radiance fields. In *CVPR*, 2021. 3

[44] Christian Reiser, Songyou Peng, Yiyi Liao, and Andreas Geiger. Kilonerf: Speeding up neural radiance fields with thousands of tiny mlps. *arXiv preprint arXiv:2103.13744*, 2021. 3

[45] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016. 2, 11, 12

[46] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. Graf: Generative radiance fields for 3d-aware image synthesis. In *NeurIPS*, 2020. 1, 2, 3, 5, 6, 7, 11

[47] Steven M Seitz and Charles R Dyer. View morphing. In *Conference on Computer graphics and interactive techniques*, 1996. 2, 4

[48] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. In *NeurIPS*, 2020. 4, 11

[49] Richard Szeliski and Polina Golland. Stereo matching with transparency and matting. In *ICCV*, 1998. 2

[50] Alex Trevithick and Bo Yang. Grf: Learning a general radiance field for 3d scene representation and rendering. In *ICCV*, 2021. 3

[51] Yi Wang, Xin Tao, Xiaojuan Qi, Xiaoyong Shen, and Jiaya Jia. Image inpainting via generative multi-column convolutional neural networks. In *NeurIPS*, 2018. 6

[52] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 5

[53] Zirui Wang, Shangzhe Wu, Weidi Xie, Min Chen, and Victor Adrian Prisacariu. NeRF–: Neural radiance fields without known camera parameters. *https://arxiv.org/abs/2102.07064*, 2021. 3

[54] Yi Wei, Shaohui Liu, Yongming Rao, Wang Zhao, Jiwen Lu, and Jie Zhou. Nerfingmvs: Guided optimization of neural radiance fields for indoor multi-view stereo. In *ICCV*, 2021. 3

[55] Wenqi Xian, Jia-Bin Huang, Johannes Kopf, and Changil Kim. Space-time neural irradiance fields for free-viewpoint video. *https://arxiv.org/abs/2011.12950*, 2020. 3

[56] Bangbang Yang, Yinda Zhang, Yinghao Xu, Yijin Li, Han Zhou, Hujun Bao, Guofeng Zhang, and Zhaopeng Cui. Learning object-compositional neural radiance field for editable scene rendering. In *ICCV*, 2021. 8

[57] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *ECCV*, 2018. 2

[58] Lin Yen-Chen, Pete Florence, Jonathan T. Barron, Alberto Rodriguez, Phillip Isola, and Tsung-Yi Lin. iNeRF: Inverting neural radiance fields for pose estimation. *https://arxiv.org/abs/2012.05877*, 2020. 3

[59] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *CVPR*, 2021. 3

[60] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*, 2018. 5

[61] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. Nerf++: Analyzing and improving neural radiance fields. *arXiv preprint arXiv:2010.07492*, 2020. 4

[62] Hang Zhao, Orazio Gallo, Iuri Frosio, and Jan Kautz. Loss functions for image restoration with neural networks. *IEEE Transactions on Computational Imaging*, 3(1):47–57, 2016. 5

[63] Zhedong Zheng, Yunchao Wei, and Yi Yang. University-1652: A multi-view multi-source benchmark for drone-based geo-localization. In *ACM MM*, 2020. 8

[64] Zhedong Zheng, Xiaodong Yang, Zhiding Yu, Liang Zheng, Yi Yang, and Jan Kautz. Joint discriminative and generative learning for person re-identification. In *CVPR*, 2019. 1

[65] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *CVPR*, 2017. 2, 4, 5

[66] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017. 1

[67] Jun-Yan Zhu, Zhoutong Zhang, Chengkai Zhang, Jiajun Wu, Antonio Torralba, Josh Tenenbaum, and Bill Freeman. Visual object networks: Image generation with disentangled 3d representations. 2018. 1, 3

Appendix

In the supplementary document, we first present the implementation details in Sec. A. Next, we provide additional visualization results in Sec. B.

A. Implementation Details

In this section, we first present the network architectures of the generative radiance field G_s, the mapping network G_m, the progressive 2D decoder G_d, and the discriminator D_ϕ in Sec. A.1. Second, we discuss the training protocol in Sec. A.2. Third, we describe the datasets used in experiments (see Sec. A.3). Finally, we provide the details of compared methods in Sec. A.4.

A.1. Network Architectures

Generative Radiance Field. The generative radiance field network G_s is a 8-layer SIREN-based MLP with periodic activation functions [48]. The dimension of the hidden layers is 256.

Mapping Network. The mapping network G_m is a 4-layer MLP network with leakyReLU as the activation function. The dimension of the hidden layers is 256. We sample the input latent code z from a 256-dimensional standard Gaussian distribution.

Progressive 2D Decoder. The progressive 2D decoder G_d is a fully-convolution neural network, which decreases the feature dimension from 256 (at 64^2) to 32 (at 512^2).

Discriminator. The discriminator D_ϕ is a progressive growing convolutional network, which uses eight layers for 64^2 and fourteen layers for 512^2.

A.2. Training Protocol

We employ Adam optimizer [27] with $\beta_1 = 0$, $\beta_2 = 0.9$, and the batch size of 56 for optimization. The initial learning rate is set to 6.0×10^{-5} for the generator and 2.0×10^{-4} for the discriminator, and decay over training to 1.5×10^{-5} and 5.0×5^{-5} respectively. We use 12 samples per ray for all datasets without hierarchical sampling strategy [4, 36].

A.3. Datasets

We conduct experiments on three widely-used high-resolution image datasets: CELEBA-HQ [24], FFHQ [25], and AFHQv2 [9].

CELEBA-HQ. CELEBA-HQ[1] [24] consists of 30,000 high-quality images of human face at 1024^2 resolution. During training, we sample the pitch and yaw of the camera pose from a Gaussian distribution with the horizontal standard deviation of 0.3 radians and the vertical standard deviation of 0.155 radians.

FFHQ. Flickr-Faces-HQ (FFHQ)[2] [25] is a large scale human face dataset which contains 70,000 high-quality images at 1024^2 resolution. The images contain various styles with different ages, ethnicity, and background. Besides, the humans in the images wear different accessories such as earrings, sunglasses, hats, and eyeglasses. In the training stage, we sample the pitch and yaw of the camera pose from a Gaussian distribution with the horizontal standard deviation of 0.3 radians and the vertical standard deviation of 0.155 radians.

AFHQv2. Animal Faces-HQ (AFHQv2)[3] [9] contains 15,000 high-quality animal face images at 512^2 resolution. The dataset has three categories: cat, dog, and wildlife, with each category providing 5,000 images. Following previous works [4, 40, 46], we conduct experiments on the cat face images to make a fair comparison. During training, we sample the pitch and yaw of the camera pose from a uniform distribution with the horizontal standard deviation of 0.4 radians and the vertical standard deviation of 0.2 radians.

A.4. Competitive Methods

We compare our approach against three state-of-the-art 3D-aware image synthesis methods: GRAF [46], pi-GAN [4], and GIRAFFE [40].

GRAF. We use the official implementation[4] to train the model on CELEBA-HQ [24], FFHQ [25] and AFHQv2 [9] datasets.

pi-GAN. We adopt the author's implementation[5] of pi-GAN [4]. Following the practice in pi-GAN [4], we begin training at 32^2 and gradually increase to 128^2 during training. The high-resolution images are rendered by sampling rays more densely (256^2, 512^2).

GIRAFFE. We train GIRAFFE [25] on all datasets [9, 24, 25] with the official implementation[6].

B. Additional Results

We provide additional results to show the multi-view consistency and the quality of the generated images.

3D Reconstruction. To further demonstrate the multi-view consistency of our method, we recover the 3D shape from generated images with the 3D reconstruction method [45]. As shown in Fig. 13, we render images of a single instance from 35 views, and then perform dense 3D reconstruction by running COLMAP [45] with default parameters and no known camera poses. The results in Fig. 14 validate the correctness of the 3D shape learned by our model.

More Visualization Results. We provide more generated

[1] https : / / github . com / tkarras / progressive _ growing_of_gans

[2] https://github.com/NVlabs/ffhq-dataset

[3] https://github.com/clovaai/stargan-v2

[4] https://github.com/autonomousvision/graf

[5] https://github.com/marcoamonteiro/pi-GAN

[6] https://github.com/autonomousvision/giraffe

Figure 13. The images are rendered from 35 camera poses at resolution 256^2 .

Figure 14. COLMAP reconstruction [45] from synthesized images at resolution 256^2.

images in Fig. 15 and Fig. 16. Please also refer to the supplementary video for more results.

Figure 15. Images synthesized by MVCGAN on CELEBA-HQ [24] at resolution 512^2.

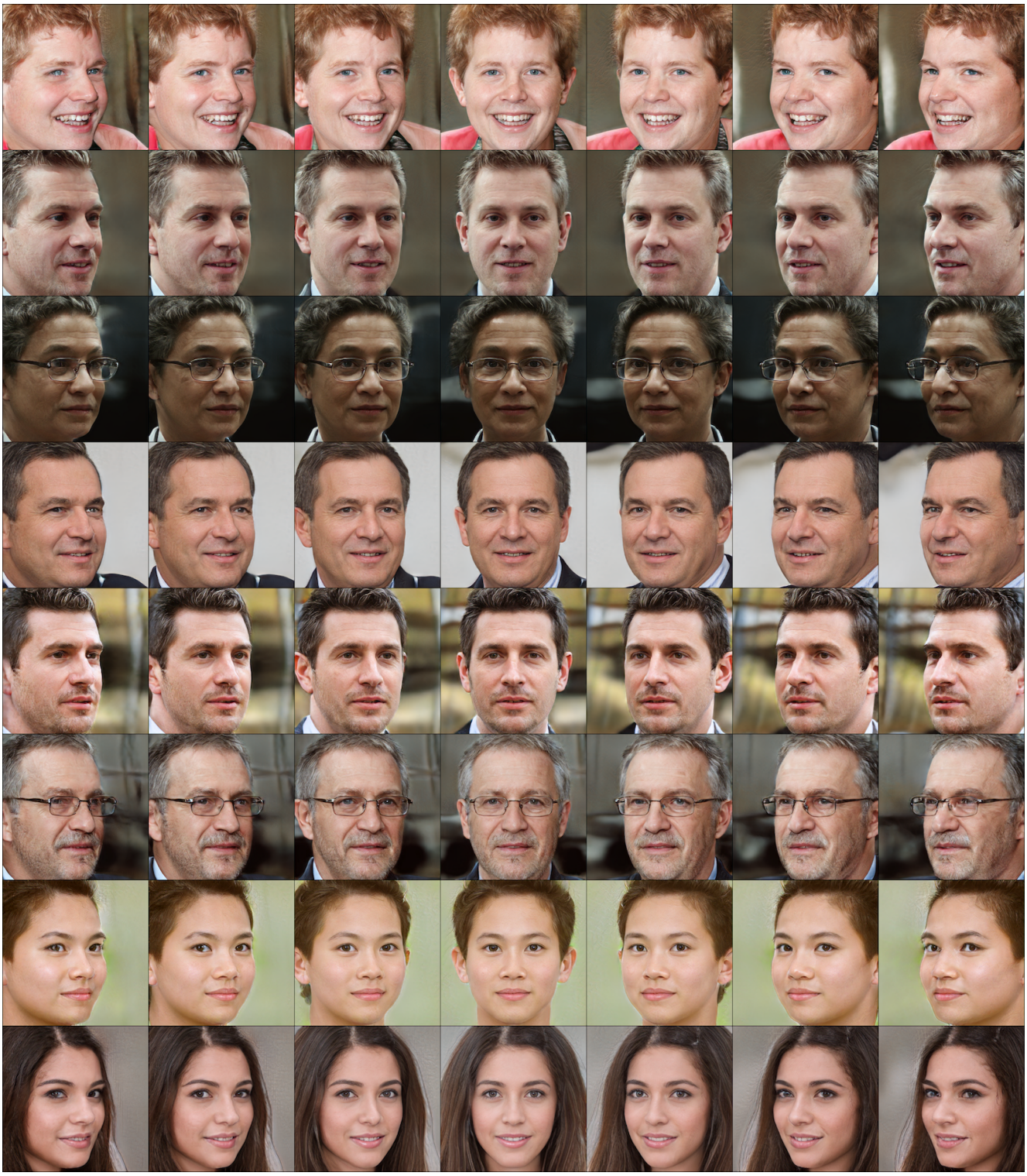

Figure 16. Images synthesized by MVCGAN on FFHQ [25] at resolution 512^2.