Supplementary Material for 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

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 [10]. 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.

*This work was done during an internship at Alibaba.

A.2. Training Protocol

We employ Adam optimizer [5] 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 [1, 6].

A.3. Datasets

We conduct experiments on three widely-used high-resolution image datasets: CELEBA-HQ [3], FFHQ [4], and AFHQv2 [2].

CELEBA-HQ. CELEBA-HQ[1] [3] 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] [4] 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] [2] contains 15,000 high-quality animal face images at 512^2 resolution.

[1]https://github.com/tkarras/progressive_growing_of_gans
[2]https://github.com/NVlabs/ffhq-dataset
[3]https://github.com/clovaai/stargan-v2

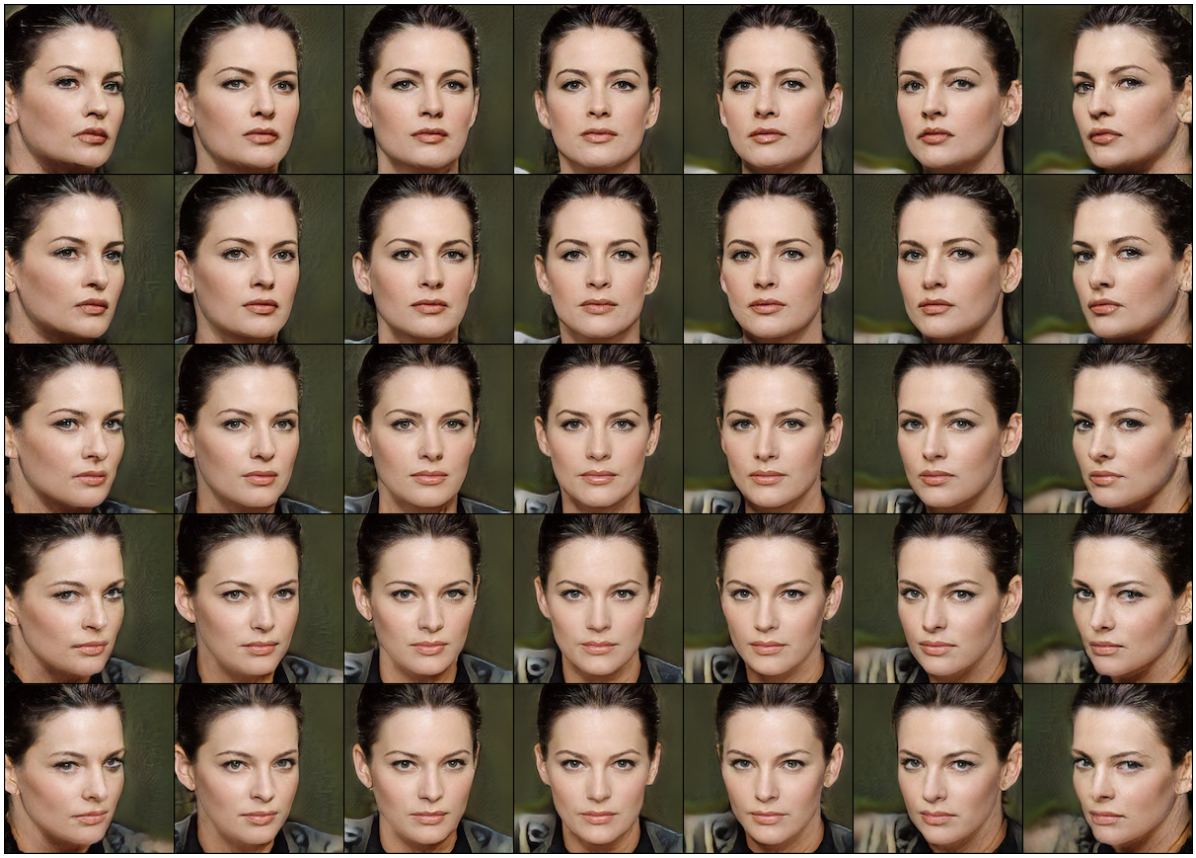

Figure 1. The images are rendered from 35 camera poses at resolution 256^2 .

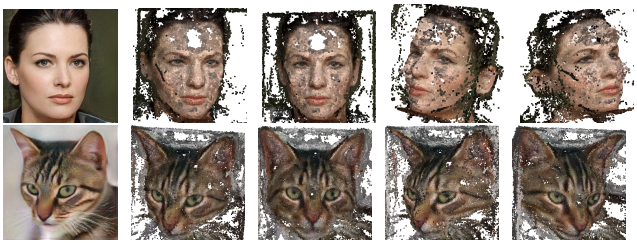

Figure 2. COLMAP reconstruction [8] from synthesized images at resolution 256^2.

The dataset has three categories: cat, dog, and wildlife, with each category providing 5,000 images. Following previous works [1, 7, 9], 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 [9], pi-GAN [1], and GIRAFFE [7].

GRAF. We use the official implementation[4] to train the model on CELEBA-HQ [3], FFHQ [4] and AFHQv2 [2] datasets.

pi-GAN. We adopt the author's implementation[5] of pi-GAN [1]. Following the practice in pi-GAN [1], 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 [4] on all datasets [2–4] 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 [8]. As shown in Fig. 1, we render images of a single instance from 35 views, and then perform dense 3D reconstruction by running COLMAP [8] with default parameters and no

[4] https://github.com/autonomousvision/graf
[5] https://github.com/marcoamonteiro/pi-GAN
[6] https://github.com/autonomousvision/giraffe

Figure 3. Images synthesized by MVCGAN on CELEBA-HQ [3] at resolution 512^2.

known camera poses. The results in Fig. 2 validate the correctness of the 3D shape learned by our model.

More Visualization Results. We provide more generated images in Fig. 3 and Fig. 4. Please also refer to the supple-

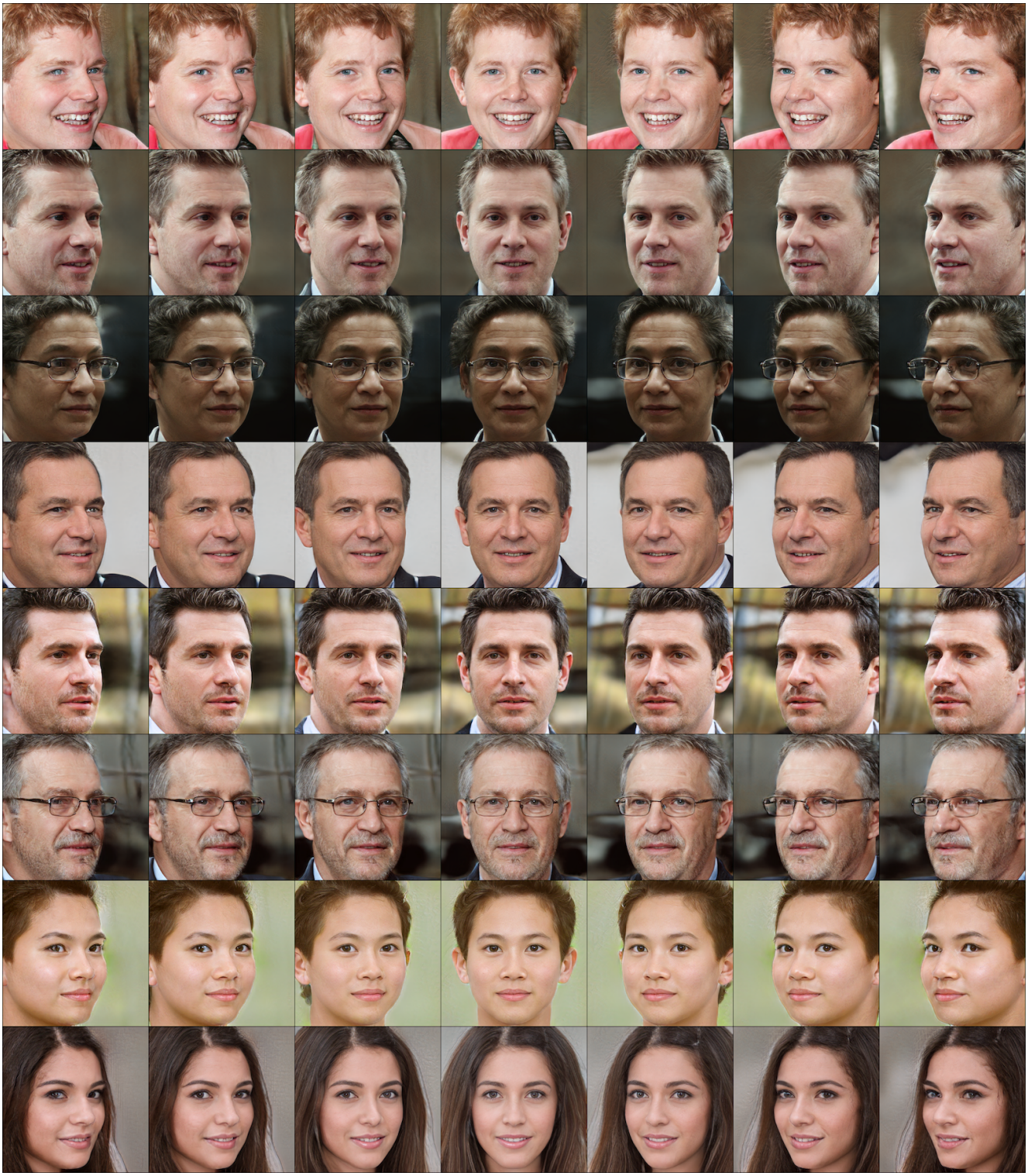

Figure 4. Images synthesized by MVCGAN on FFHQ [4] at resolution 512^2.

mentary video for more results.

References

[1] 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

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

[3] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *ICLR*, 2018. 1, 2, 3

[4] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 1, 2, 4

[5] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 1

[6] 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. 1

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

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

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

[10] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. In *NeurIPS*, 2020. 1