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

We sincerely thank all reviewers for their time and efforts. We will carefully revise our paper according to the comments and suggestions. Below please find the responses to some specific comments.

Response to Reviewer 1

Q1: Inferior to StyleGAN (2D GAN). 1. Yes, our method is inferior to StyleGAN in terms of FID score, but it is worth noting that our method is orthology to StyleGAN. In this work, we mainly focus on addressing multi-view inconsistency by introducing geometry constraints. In fact, it is possible to adopt any 2D GAN backbone as the 2D decoder of our model. In the future, we will integrate the design of StyleGAN into our framework to further improve the image quality. **2.** Our method can also generate images at 1024^2 resolution by adding one more layer into the 2D decoder. We will add the high-resolution results in the revision. Thanks a lot for the suggestion.

Q2: How to improve. Please refer to **Q4 of Reviewer 2**.

Q3: Extend to other datasets and video generation. 1. Yes, our method can be extended to datasets such as the *Chairs* and *Cars* datasets used in GRAF [46]. **2.** Yes, for video generation, we can first synthesize images from a set of consecutive viewpoints, and then turn the image sequence into a video (see the supplementary video).

Response to Reviewer 2

Q4: More discussion on limitations and future works. We really appreciate your great suggestion. In this paper, our method mainly focuses on single-object scenes with simple backgrounds. To extend to the scenario with complex background and multiple objects, one possible way is to learn a compositional radiance field that can represent the scene as a composition of foreground and background [64]. Specifically, we can use two NeRF branches to model foreground objects and background surroundings separately. To render the whole scene, the geometry relationships between foreground objects and the background can be established by combing depth maps and occlusion maps. We will add more discussion in the revision.

Response to Reviewer 3

Q5: Computational efficiency. Thanks for pointing out this issue. Here we provide a quantitative comparison of computation complexity (GFLOPs) and time efficiency (rendering speed). For example, to render an image with 512^2 resolution, pi-GAN [4] costs 3314.6 GFLOPs and 3393.3 ms on a single RTX 8000 GPU. In contrast, our method has lower computation complexity (249.9 GFLOPs) and a faster rendering speed (62.9 ms / per image).

Q6: Effectiveness of feature-level optimization. The effectiveness of feature-level optimization can be demonstrated from two aspects. **1. Quantitative comparison.** As we stated in Line 794 and 798, the feature-level optimization achieves better performance (FID=13.7) than the image-level optimization (FID=22.5). **2. Qualitative comparison.** When zooming in Fig. 10, we observe that the images generated by feature-level optimization have more delicate details, such as clear wrinkles, the highlight on the forehead, and the shadow of the cheeks.

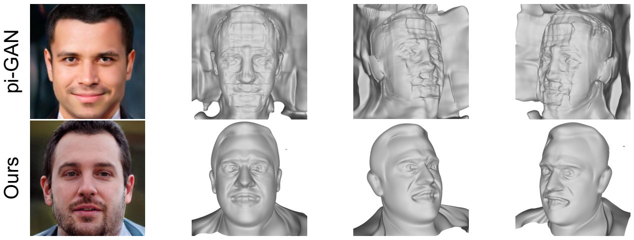

Figure 12. Visualization of extracted 3D meshes. Please zoom in for better visualization results.

Q7: Explanation of η in Eq. 6. Actually, η is not a fixed hyper-parameter, but a dynamic number randomly sampled in every training iteration. We will clarify this in the revision. The stereo mixup module is inspired by *mixup*, a data augmentation method for image classification [65]. By constructing a linear combination of $\mathcal{I}_{pri}$ and $\mathcal{I}_{warp}$ with dynamic weights, the stereo mixup mechanism can optimize the two images with only one discriminator.

Q8: Explanation for better performance under large pose variations. Existing approaches suffer from collapsed results under large pose variations because the models fail to learn an appropriate 3D shape in the absence of geometry constraints (see Sec. 3.2.1). dIn contrast, we incorporate geometry constraints into training to enhance the geometric reasoning ability of the generative model. To better illustrate the learned 3D representation, we extract the underlying geometry of generated images using marching cubes. As shown in Fig. 12, our model learns a more accurate 3D shape and thus can achieve better performance under large view variations.

Q9: More discussion on limitations. Thanks for pointing out this issue, please refer to **Q4 of Reviewer 2**.

Q10: Occlusion mask. Yes, I think incorporating occlusion maps can better handle multi-object scenes with severe occluded regions.

References

[64] 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. 1

[65] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*, 2018. 1