Dual Mapping of 2D StyleGAN for 3D-Aware Image Generation and Manipulation (Student Abstract)

Zhuo Chen[1], Haimei Zhao[2], Chaoyue Wang[2], Bo Yuan[3], Xiu Li[1]
[1]Shenzhen International Graduate School, Tsinghua University
[2]University of Sydney
[3]University of Queensland
z-chen17@mails.tsinghua.edu.cn, hzha7798@uni.sydney.edu.au, chaoyue.wang@outlook.com, boyuan@ieee.org, li.xiu@sz.tsinghua.edu.cn

Abstract

3D-aware GANs successfully solve the problem of 3D-consistency generation and furthermore provide a 3D shape of the generated object. However, the application of the volume renderer disturbs the disentanglement of the latent space, which makes it difficult to manipulate 3D-aware GANs and lowers the image quality of style-based generators. In this work, we devise a dual-mapping framework to make the generated images of pretrained 2D StyleGAN consistent in 3D space. We utilize a tri-plane representation to estimate the 3D shape of the generated object and two mapping networks to bridge the latent space of StyleGAN and the 3D tri-plane space. Our method does not alter the parameters of the pretrained generator, which means the interpretability of latent space is preserved for various image manipulations. Experiments show that our method lifts the 3D awareness of pretrained 2D StyleGAN to 3D-aware GANs and outperforms the 3D-aware GANs in controllability and image quality.

Introduction

Style-based generators (Karras, Laine, and Aila 2019; Karras et al. 2020) not only improved the reality and fidelity of the synthesized images but also disentangled the latent space semantically, enabling image attribute editing and animation. However, style-based generators have been observed to struggle with producing consistent images when adjusting the head pose or camera view. The development of Nerf (Mildenhall, Srinivasan et al. 2020) has sparked the concept of unsupervised learning of 3D object priors from 2D images (Chan et al. 2022; Gu et al. 2022), which enables generative networks to synthesize 3D-consistent images.

However, the usage of 3D representation during training the generator disturbs the semantic disentanglement of StyleGAN latent space and decreases the image quality. For instance, the latent space of a StyleGAN which is trained on human facial photos are semantically disentangled. Linear interpolation in the latent space can edit the facial attributes such as smiling, glasses and gender of the generated images . Although most 3D-aware GANs also utilize style-based generators, the aggregation of features in 3D space during the training makes it more challenging to semantically disentangle the latent space. As a result, current 3D-aware GANs can hardly control most attributes of the generated images except for the camera view.

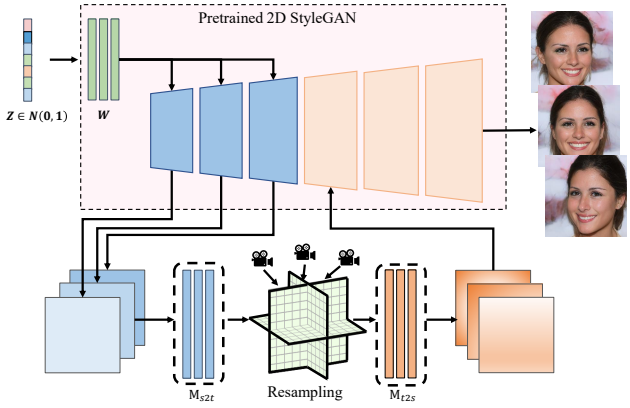

Figure 1: The overall framework of our method, consisting of a pretrained 2D StyleGAN and Dual-mapping Networks.

We propose that 2D StyleGANs also have the potential to model the 3D shape of the generated object within the latent space. We introduce tri-plane (Chan et al. 2022) to represent the 3D feature space and utilize two mapping networks to transform the features between style space and 3D space.

Our method lifts 2D StyleGAN for 3D-aware generation and manipulation dispense with re-training the whole network. Compared to other 3D-aware GANs, our method reserves the semantically disentangled latent space of 2D StyleGAN. Former image attribute editing methods such as GANSpace (Härkönen et al. 2020) can also be applied to manipulate the generated images based on our framework.

Method

As shown in Figure 1, our framework consists of a pretrained 2D style-based generator G and two mapping networks M_{s2t} and M_{t2s}. The parameters of G are frozen all the time. One mapping network M_{s2t} is used to transform the 2D style features taken from the G to tri-plane space. The other one M_{t2s} can transform the re-sampled tri-plane features back into the space of StyleGAN.

Specifically, given a randomly sampled vector $z \in N(0,1)$, we pass it to the fixed generator G to collect the the temperate features $\{F_1^s, F_2^s, ...F_m^s\}$ from the shal-

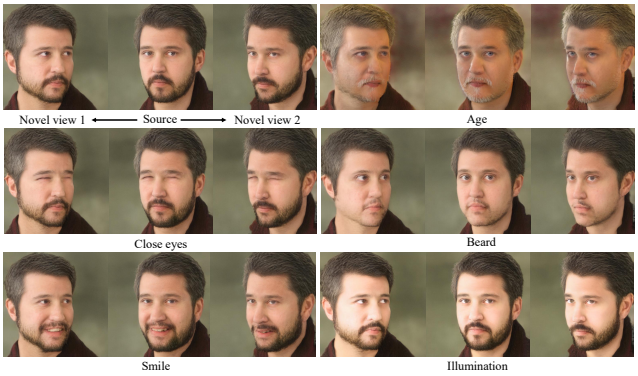

Figure 2: Examples of facial attribute editing with different 3D view points.

low layers. We then resize the collected features to the same size and concentrate them through the channel dimension to feature $F^s_{[1...m]}$. By transforming the feature at each pixel, we transform $F^s_{[1...m]}$ to tri-plane space, which formulates the 3D shape in a simple and efficient mode. Then we can resample from the tri-plane at target camera view points $\{\theta_1, \theta_2..., \theta_n\}$ to get responding features donated as $\{T^{\theta_1}, T^{\theta_2}, ..., T^{\theta_n}\}$. The other mapping network M_{t2s} is used to transform the re-sampled features back to StyleGAN space as $\{F^{\theta_1}, F^{\theta_2}, ..., F^{\theta_n}\}$, which can be passed to the latter layers of G to produce high-resolution images.

Our framework can manipulate the generated images by interpolating the features in W space. As the generator is fixed, the attribute edition methods which were designed for 2D StyleGAN are also applicable for our framework. Furthermore, we can control the camera views of the generated images by resampling from the tri-plane space at the given angles. Therefore, the proposed dual-mapping architecture ensures both editability and 3D consistency.

Evaluation

We show the results of image manipulation in Figure 2. The interpolation directions are found by GANSpace (Härkönen et al. 2020). Our method is able to manipulate various facial attributes such as age, beard, illumination, and expression while keeping the 3D consistency of edited images. The generated images with different facial attributes and 3D view points are similar to the source image in reality and fidelity.

In the quantitative experiment, We compare our method with StyleGAN2 (Karras et al. 2020) and a 3D-aware GAN EG3D (Chan et al. 2022) on high-resolution image datasets CelebA-HQ (Lee et al. 2020) and FFHQ (Karras, Laine, and Aila 2019). For each comparison, we randomly select 2000 images as the sources and another 2000 images as the targets. We animate the source image to be with the same head pose and expression as the target. We use Frechet Inception Distance (FID) to describe the reality and quality of the generated images. To measure the identity proximity of the source and the animated face, we computed the cosine similarity (CSIM) between extracted ArcFace features. We evaluate the accuracy of facial expression (Exp) and head pose

Methods	CelebA-HQ				FFHQ			
	FID↓	CSIM↑	Exp↓	Pose↓	FID↓	CSIM↑	Exp↓	Pose↓
StyleGAN2	**46.42**	0.57	3.64	1.25	**54.75**	0.56	5.10	1.31
EG3D	75.33	0.54	5.81	1.08	83.14	0.52	6.57	1.12
Ours	50.96	**0.59**	**3.44**	**1.04**	58.38	**0.59**	**4.70**	**1.08**

Table 1: Quantitative experiments on CelebA-HQ (Lee et al. 2020) and FFHQ (Karras, Laine, and Aila 2019).

(Pose) by measuring the normalized mean error between the facial landmarks and head Euler angles of the animated and target images. As reported in Table 1, our method outperforms EG3D in both image quality and animation accuracy. Meanwhile, the proposed method fixes the head pose control problem of StyleGAN2.

Conclusion

Our study explores the extended utility of pretrained 2D StyleGANs. We establish that the latent space of StyleGAN exhibits promise for capturing the 3D shape characteristics of generated objects, even in the absence of explicit 3D considerations during training the generator. The proposed dual-mapping framework empowers 2D StyleGANs to generate 3D-consistent images by establishing a connection between the latent space and the tri-plane space through two mapping networks. In contrast to the 3D-aware GANs which integrate the 3D feature during training the generator, our model preserves semantic editability for various image manipulation. Experimental results substantiate that our approach enhances the efficacy of 3D-aware image manipulation techniques. We anticipate that our research will provide valuable insights for future endeavors in leveraging pretrained GANs for various applications and advancements.

Acknowledgments

This research was partly supported by Shenzhen Key Laboratory of next generation interactive media innovative technology (Grant No: ZDSYS20210623092001004).

References

Chan, E. R.; et al. 2022. Efficient geometry-aware 3D generative adversarial networks. In *CVPR*, 16123–16133.

Gu, J.; et al. 2022. StyleNeRF: A Style-based 3D Aware Generator for High-resolution Image Synthesis. In *ICLR*.

Härkönen, E.; et al. 2020. GANSpace: Discovering Interpretable GAN Controls. In *NeurIPS*, 9841–9850.

Karras, T.; Laine, S.; and Aila, T. 2019. A style-based generator architecture for generative adversarial networks. In *CVPR*, 4401–4410.

Karras, T.; et al. 2020. Analyzing and improving the image quality of stylegan. In *CVPR*, 8110–8119.

Lee, C.-H.; Liu, Z.; Wu, L.; and Luo, P. 2020. MaskGAN: Towards Diverse and Interactive Facial Image Manipulation. In *CVPR*, 5549–5558.

Mildenhall, B.; Srinivasan, P. P.; et al. 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *ECCV*, 405–421.