High-fidelity 3D GAN Inversion by Pseudo-multi-view Optimization

Jiaxin Xie[1*] Hao Ouyang[1*] Jingtan Piao[2†] Chenyang Lei[3] Qifeng Chen[1†]
[1]HKUST [2]MMLab, CUHK [3]CAIR, HKISI-CAS

Abstract

We present a high-fidelity 3D generative adversarial network (GAN) inversion framework that can synthesize photo-realistic novel views while preserving specific details of the input image. High-fidelity 3D GAN inversion is inherently challenging due to the geometry-texture trade-off, where overfitting to a single view input image often damages the estimated geometry during the latent optimization. To solve this challenge, we propose a novel pipeline that builds on the pseudo-multi-view estimation with visibility analysis. We keep the original textures for the visible parts and utilize generative priors for the occluded parts. Extensive experiments show that our approach achieves advantageous reconstruction and novel view synthesis quality over prior work, even for images with out-of-distribution textures. The proposed pipeline also enables image attribute editing with the inverted latent code and 3D-aware texture modification. Our approach enables high-fidelity 3D rendering from a single image, which is promising for various applications of AI-generated 3D content. The source code is at https://github.com/jiaxinxie97/HFGI3D/.

1. Introduction

Real-world 3D-aware editing with *a single 2D image* fascinates various essential applications in computer graphics, such as virtual reality (**VR**), augmented reality (**AR**), and immersive meetings. Recent advancement in 3D GANs [12, 13, 21, 45] has achieved photo-realistic 3D-consistent image generation. With the GAN inversion approaches [19, 51, 72], which can map the images to the latent space of the pretrained 3D-aware model, high-fidelity 3D-aware editing becomes promising.

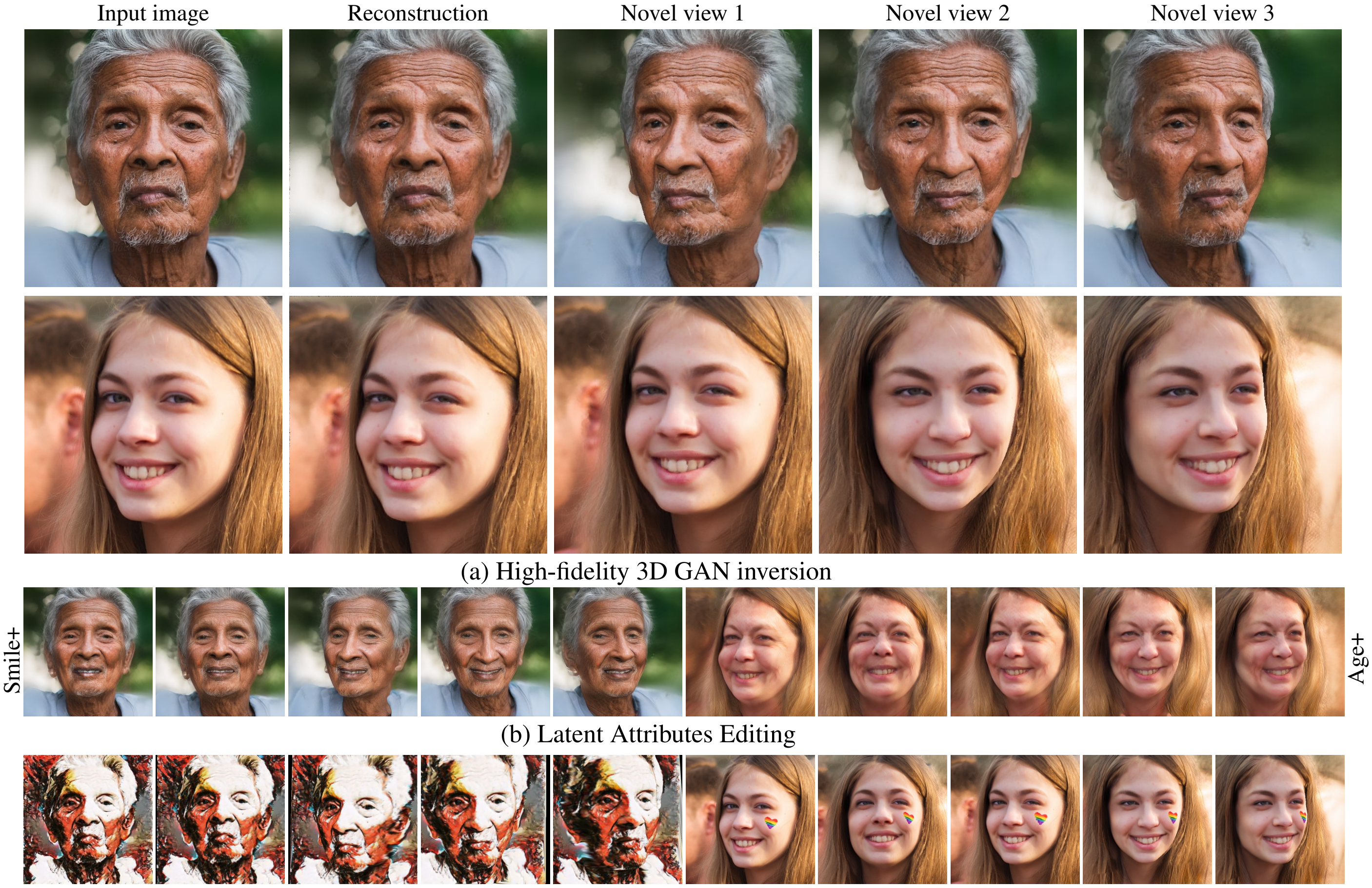
High-fidelity 3D-aware inversion aims to generate novel views with high-quality reconstruction and 3D consistency, but existing methods can hardly meet these two goals simultaneously. Although current GAN inversion methods based on 2D GANs [1, 53, 54, 69] can perform 3D-related attributes (e.g., head pose) editing, the generated view is inconsistent due to the lack of the underlying 3D representation. When applying the existing optimization-based inversion approaches on 3D-aware GANs [12, 36], we can retrieve high-fidelity reconstruction by overfitting to a single input image. However, different from 2D GAN inversion, the reconstruction quality of 3D GAN depends not only on the input view's faithfulness but also on the quality of the synthesized novel views. During the optimization process, the obvious artifacts in the synthesized novel views occur with the appearance of high-fidelity details in the input view as analyzed in Sec. 3. As only a single image is available in the optimization process, the reconstruction suffers from extreme ambiguity: infinite combinations of color and density can reconstruct the single input image, especially with out-of-distribution textures.

Based on the above observation, we propose our 3D-aware inversion pipeline by optimizing the reconstruction not only on the input image but also on a set of pseudo-multi-views. The pseudo views provide additional regularization, and thus the ambiguity is greatly reduced. Estimating the pseudo views is non-trivial as it requires maintaining the texture details while also generating the occluded parts in a plausible manner, based on the input view. We first estimate an initial geometry and conduct a visibility analysis to solve these challenges. We directly utilize the textures from the input image for the visible parts to preserve the texture details. For the occluded parts, we use a pretrained generator to synthesize the reasonable inpainted regions. With the additional supervision from the pseudo-multi-views, our approach achieves high-fidelity reconstruction results with the correct 3D geometry.

Our approach enables two types of editing: latent attributes editing and 3D-aware texture modification. We follow the previous work [55] and calculate the attribute direction in the latent code space. By modifying the inverted latent code in a specific direction, we can control the general attribute (e.g., smile, ages for portraits as in Figure 1(b)). Since the proposed pipeline enables inversion with out-of-distribution textures, we can achieve compelling 3D consistent editing by only modifying the textures of the input images (e.g., stylization or adding a tattoo in Figure 1(c)).

In summary, we propose a high-fidelity 3D GAN in-

[*] Joint first authors
[†] Joint corresponding authors



(a) High-fidelity 3D GAN inversion

(b) Latent Attributes Editing

(c) 3D-aware Textures Modification

Figure 1. High-fidelity 3D GAN inversion results on real-world images with two types of editing ability. Our method preserves compelling details and achieves high 3D consistency.

version method by pseudo-multi-view optimization given an input image. Our approach can synthesize compelling 3D-consistent novel views that are visually and geometrically consistent with the input image. We perform extensive quantitative and qualitative experiments, which demonstrate that our 3D GAN inversion approach outperforms other 2D/3D GAN inversion baselines in both photorealism and faithfulness.

2. Related Work

2.1. GAN Inversion

GAN inversion [19, 51, 72, 82] retrieves the latent code that can faithfully reconstruct the input image given a pretrained generative model. The recovered latent code facilitates various applications such as image editing [1,2,55,81], interpolation [1, 44], and restoration [1, 48]. The existing inversion methods can be classified into three categories: optimization-based [1,2,18,27,30,71,83], encoder-based [3, 7, 22, 53, 64, 69], and hybrid [5, 11, 54, 81]. The optimization-based methods reduce the reconstruction error to optimize a latent code directly, whereas the encoder-based approaches involve training an encoder to map from the image space to the latent space. A hybrid scheme combines the above methods by using the encoder for initialization and refining the latent code. Although recent 2D GAN inversion [4, 69] has achieved faithful reconstruction with high editing capability, the editing related to 3D attributes (e.g., control camera angle and head pose) still suffers inevitable inconsistency and severe flickering as the pretrained generator is not 3D-aware. With the rapid development of 3D-aware GANs [12, 13], 3D consistent editing becomes promising with GAN inversion techniques. Unlike the concept of fidelity of 2D GAN inversion, which considers only the pixel-wise difference of the input view, the 3D inversion additionally involves the quality of the synthesized novel views. With the optimization-based inversion scheme, even though the details in the input view are

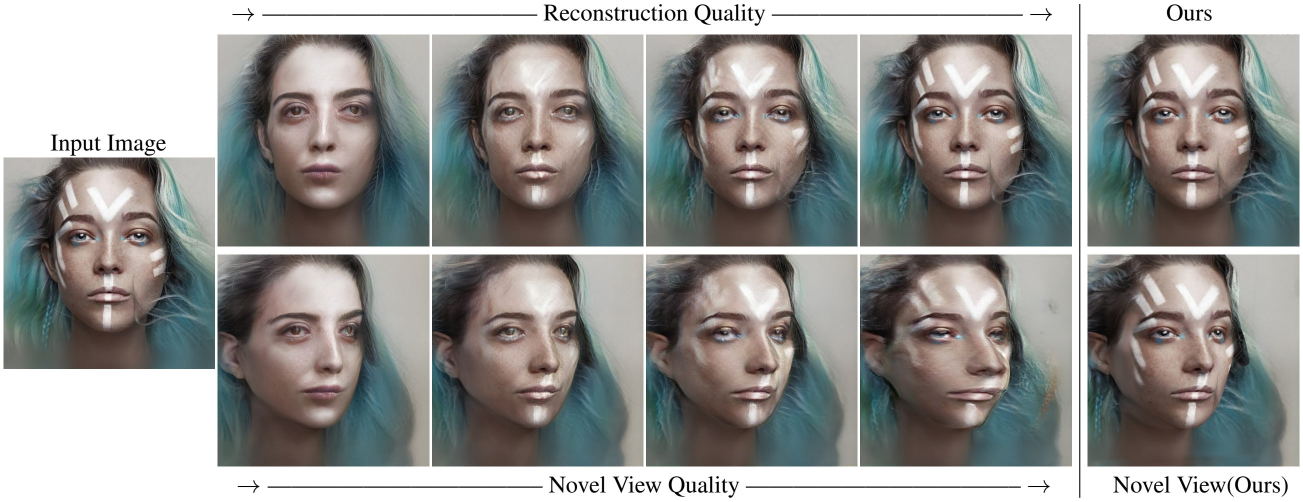


Figure 2. Reconstruction quality vs. novel view quality during the optimization process. The reconstruction quality improves as the iteration increases, but the novel view quality decreases if we optimize the generator only using loss with the single input image. Our method achieves both highly preserved details with compelling novel view synthesis.

well-preserved, the overall fidelity can still be low as the synthesized view contains severe artifacts because of the issues such as entangled texture and geometry. This paper focuses on solving these issues and achieving high-fidelity 3D-aware reconstruction.

In addition to GAN inversion, there is another emerging field of research focused on reconstructing 3D representations from a single image, specifically depth [73,75]. While these approaches solely focus on reconstruction, inversion-based methods also enable editing capabilities.

2.2. 3D-aware GANs

Geometrically consistent GANs [24] have recently become a trending research topic. Early works explore generating 3D consistent images with different 3D representations such as mesh [16,23,28,31,40], voxels [29,41,42,65], multi-plane images [47,79] and point clouds [63]. The generated images usually suffer from blurry details since the 3D resolution is low, considering the memory cost. Although a learned neural rendering module [41, 42] can increase the generation quality, it damages the view consistency and results in inconsistent novel views. The neural 3D representation [6, 9, 16, 25, 37–39, 49, 50, 60, 66] (Neural Radiance Fields [39] especially) achieves stunning photorealism for novel view synthesis and can serve as the underlying representation for the 3D-aware generation. The recently proposed 3D-aware GANs [12, 13, 21, 43, 45, 56–58, 80] rely on implicit 3D representations and render high-resolution outputs with impressive details and excellent 3D consistency. Our works adopt EG3D [12] as the pretrained architecture as it generates photorealistic 3D consistent images comparable to StyleGAN [32, 33] while maintaining high computational efficiency. Note that the proposed inversion pipeline can be easily adapted to other 3D-aware GANs by replacing the underlying triplane model with other 3D representations.

Researchers have started early exploration in 3D GAN inversion [8, 34, 36, 52, 61, 62, 70]. Lin *et al.* [36] explores the latent space of EG3D and enables animating a portrait with a single image. IDE-3D [61] and FENerf [62] retrains EG3D and PiGAN in a semantic-aware way, achieving semantic-conditioned editing with a hybrid and optimization inversion scheme respectively. Note that NARRATE [70] utilizes inversion to retrieve the 3D shapes, but their goals are to estimate the normal for novel view portrait editing. Different from these works, which mainly focus on the downstream applications of 3D GAN inversion, our work aims to improve the faithfulness of the 3D inversion while preserving the editing ability.

3. Method

3.1. Overview

3D GAN Inversion. Given the generator G of a pretrained GAN that maps from latent space $\mathcal{W}$ to image space $\mathcal{X}$, GAN inversion aims at mapping from $\mathcal{X}$ back to $\mathcal{W}$, where the latent w can faithfully reconstruct the input image x_0. As the 3D-aware GANs involve the physics-based rendering stage, the additional input camera pose p_0 is involved in synthesizing 3D consistent images. Formally, we formulate the 3D GAN inversion problem as follows:

$$w^* = \arg\min_{w} \mathcal{L}(G(w, p_0), x_0), \tag{1}$$

where $\mathcal{L}(\cdot)$ is the optimization loss that represents image distance in image or feature spaces.

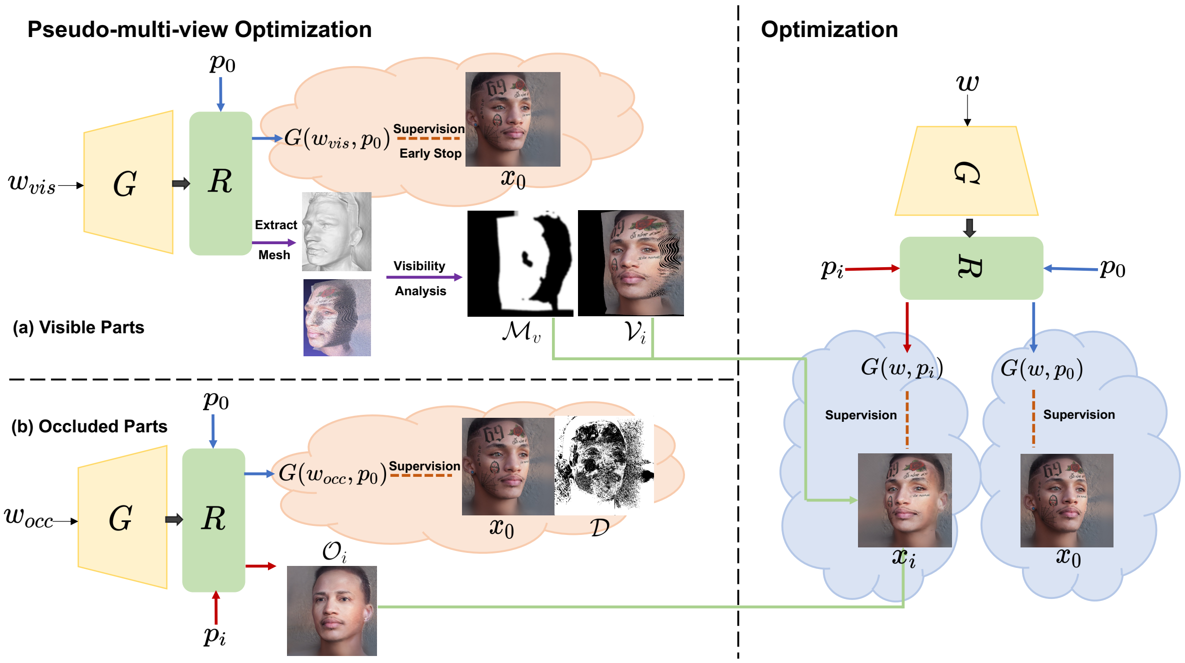


Figure 3. The pipeline of the proposed 3D-aware inversion framework. G and R are the pretrained 3D-aware generator and the corresponding renderer, respectively. To improve the inversion quality, we synthesize the pseudo-view with camera pose p_i to regularize the optimization. The pseudo-multi-view estimation consists of two steps: the visible image $\mathcal{V}_i$ warping (upper-left) and the occlusion image $\mathcal{O}_i$ generation (lower-left), combined with the estimated visibility mask $\mathcal{M}_v$. With the additional supervision of the pseudo-views, the pipeline achieves high-fidelity 3D-consistent inversion.

Analysis. Researchers have conducted some attempts in 3D GAN inversion by directly applying optimization or hybrid methods for portrait reconstruction. Although the general facial features of the input images can be reconstructed with these approaches, the image-specific details, such as the freckles, wrinkles, and tattoos, are lost, and the fidelity is thus degraded. We observe that the optimization procedure in these methods is "early stopped" at a certain iteration before the input image x_0 is "overfitted."

However, if we trivially increase the max optimization iteration, as shown in Figure 2, while the inversion preserves more image-specific details in the input image, obvious artifacts appear in the synthesized novel views. We further find that as the optimization iterations increase, although the reconstruction quality of the input image increases, the 3D consistency, on the contrary, decreases(more quantitative results are shown in the *supplement*). Our experiments demonstrate that "overfitting" to the input image that considers only the input camera pose damages the geometry of 3D GAN inversion. As the latent code is optimized to generate the image-specific textures, the corresponding geometry goes out of the distribution and leads to visually-unpleasant novel view results. To achieve high-fidelity inversion with high-quality novel view synthesis, we need a more delicate optimization scheme to handle the texture-geometry trade-off.

Design. As analyzed above, even though the optimized latent w can render an image similar to the input image x_0 given the input camera pose p_0, w does not guarantee reasonable rendering under other camera poses. In this work, we improve the inversion performance by regularizing the outputs of additional camera poses $\{p_1, p_2, ..., p_n\}$ with the corresponding pseudo-multi-views $\{x_1, x_2, ..., x_n\}$. To this end, we can reformulate the objective of the 3D GAN inversion as follows:

$$w^* = \arg\min_w \mathcal{L}(G(w, p_0), x_0) + \alpha \sum_i^n \mathcal{L}(G(w, p_i), x_i), \tag{2}$$

where α controls the strength of the regularization from the pseudo-views. For a pseudo-view x_i with camera pose p_i, to increase the quality of the inversion, we should have the following properties: (1) if the texture from x_0 is visible under the camera pose p_i, the texture should be preserved; (2) the occluded parts should reasonably inpainted. Next, we will describe in detail how to synthesize the reasonable pseudo-views that satisfy the above properties.

Note that we have explored other alternative regularization strategies, such as regularizing the density or keeping the coarse geometry unchanged during the optimization while the reconstruction loss is only calculated on a single input, and the details are provided in the *supplement*.

3.2. Pseudo-multi-view Optimization

We divide the pseudo-multi-view estimation into two steps. Given a novel view, the synthesized image is expected to contain some parts visible from the input image. As only a single input view is provided, certain parts are occluded in the novel view. For the visible parts, we can directly warp the original textures from the input image. For the occluded parts, the pretrained generator G is able to synthesize various photo-realistic 3D consistent images, which can be utilized to inpaint the occluded areas.

3.2.1 Visible Part Reconstruction

From the analysis in Figure 2, although the early stage of the optimization fails to reconstruct faithful textures, the coarse geometry generally matches the input image. Thus we can utilize the initially estimated geometry with the early stop to conduct the visibility analysis. We denote the optimized latent code as w_{vis}, and then we can reproject the input color image onto a 3D mesh derived from w_{vis}, as shown in Figure 3. For a new camera pose p_i, we project the mesh to the image plane [10] and denote the image regions with projected mesh as visible parts $\mathcal{M}_v$ (other regions as occluded parts $\mathcal{M}_o$). There are abrupt changes in the geometry along the boundary in $\mathcal{M}_v$, so we erode $\mathcal{M}_v$ with blur kernels around the boundary regions. The detailed description of each step is included in the *supplement*.

With the estimated geometry, we can warp the texture from the input image to the novel view. For the novel pose p_i, we obtain the pseudo textures $\mathcal{V}_i$ by projecting the color mesh. As shown in Figure 3, $\mathcal{V}_i$ preserves the details from the input image but contains large missing parts that are occluded. Next, we will focus on inpaint these occluded parts.

3.2.2 Occluded Part Reconstruction

The inpainted textures for the occluded parts should generate reasonable shapes, be consistent with $\mathcal{V}_i$, and be 3D consistent for different views. One possible choice is to directly inpaint the missing regions with the pre-trained image or video inpainting pipelines [46, 68, 68, 74, 76, 78]. While inpainting for a single view can be reasonable in 2D, the 3D consistency suffers due to the lack of underlying 3D representations. Note that the pretrained 3D-aware generator G can synthesize photo-realistic 3D consistent images, and thus we propose to inpaint the occluded parts with G. However, as in Figure 2, the novel view quality degrades because of the out-of-distribution textures. To increase the robustness of the inpainting, we exclude the out-of-distribution textures for the inversion and then synthesize reasonable novel views to inpaint the occluded parts.

As shown in Figure 3, given a camera pose p_i, the GAN inversion produces a reconstructed image $\mathcal{O}_i$, which keeps the general appearance of the input image with inpainted textures and shapes. To exclude the out-of-distribution textures, we calculate a difference map $\mathcal{D}$ as$||x_0 - G(w_{vis}, p_0)||_2$. Then we binarize $\mathcal{D}$ by setting the pixel value of $\mathcal{D}$ to 0 if the difference is greater than a threshold θ; otherwise, its value is set to 1.

Finally, we can optimize a latent code w_{occ} for occluded parts based on the binary difference map $\mathcal{D}$ with the camera pose p_0:

$$w_{occ} = \arg\min_{w} \mathcal{L}(G(w, p_0)\mathcal{D}, x_0\mathcal{D}), \tag{3}$$

$$\mathcal{O}_i = G(w_{occ}, p_i). \tag{4}$$

With w_{occ}, we can generate any $\mathcal{O}_i$ for the camera pose p_i.

3.2.3 Optimization

With the estimated pseudo-multi-views as additional supervision, we perform optimization to retrieve the latent code. We first perform optimization in the $\mathcal{W}+$ space, then unfreeze and finetune the pretrained generator G, following Roich et al. [54]. For each iteration, we randomly select an auxiliary camera pose p_i for optimization. Considering the visible parts and the occlusion parts, we can represent the loss in each gradient descent step as

$$\mathcal{L}_{rec}(G(w, p_0), x_0) + \mathcal{L}_{rec}(\mathcal{M}_o G(w, p_i), \mathcal{M}_o \mathcal{O}_i) + \mathcal{L}_{rec}(\mathcal{M}_v G(w, p_i), \mathcal{M}_v \mathcal{V}_i), \tag{5}$$

where the reconstruction loss $\mathcal{L}_{rec}$ is the weighted sum of the $\mathcal{L}_2$ loss and the perceptual loss LPIPS with the features extracted from the VGG network [15, 59, 77]. With the perceptual loss (LPIPS), we achieve higher quality for the perceptual details such as the hairs. After the fitting, the optimized latent code can be used for synthesizing novel views or various attribute editing.

4. Experiments

4.1. Experimental Setup

For the initial visibility estimation stage, we set the learning rate at 5e-3 and the training iteration at 1000. We set the learning rate for the optimization stage at 3e-4 and the training iteration at 3000. We choose EG3D [12] as the 3D-aware generator G as it synthesizes photo-realistic images with high 3D consistency. We utilize a pretrained EG3D model trained on the FFHQ [32] dataset for optimization and then evaluate the performance on CelebA-HQ [35] dataset. All experiments are done on a single NVIDIA RTX 3090 GPU. We attach more detailed settings in the *supplement*.

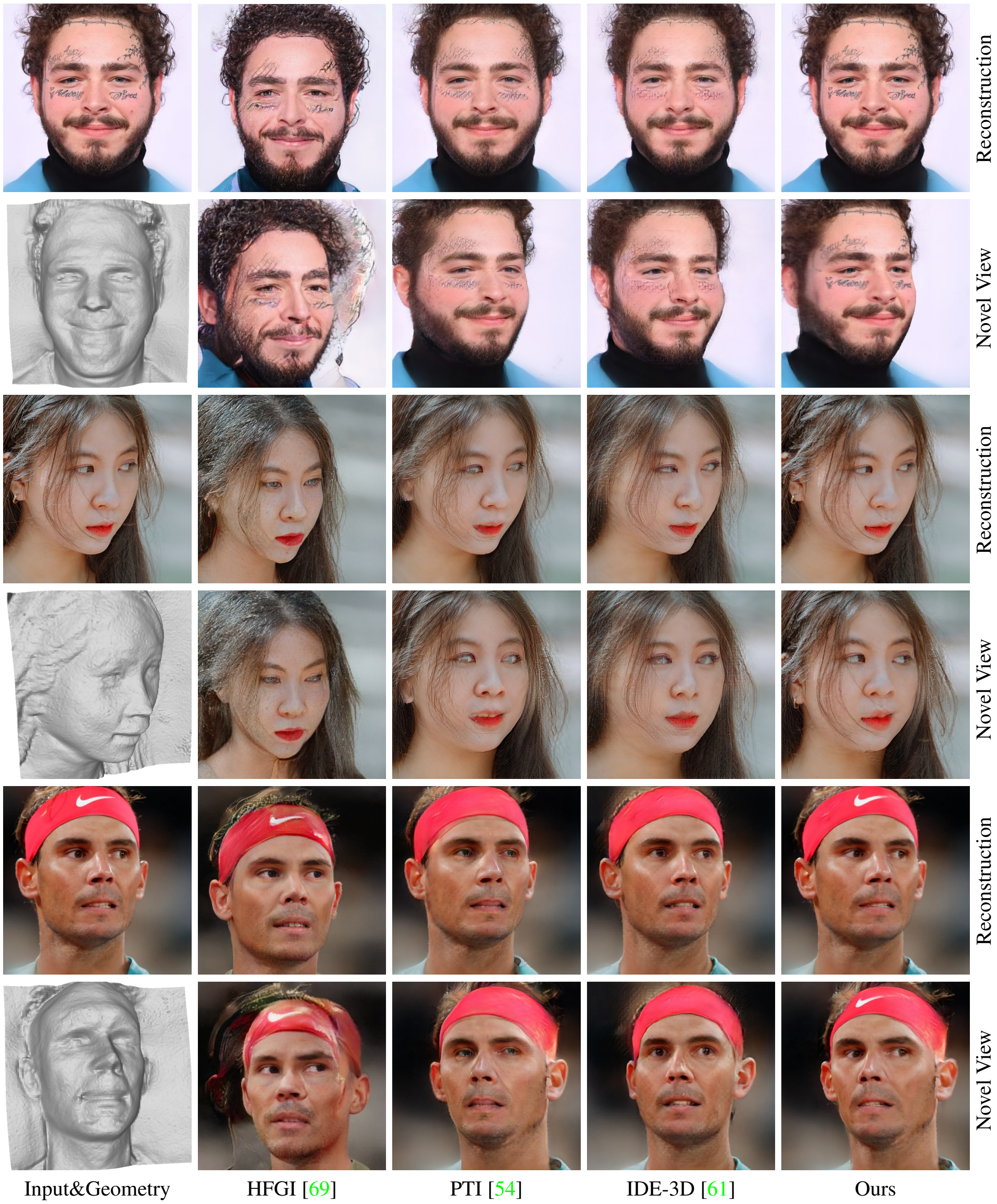


Figure 4. Qualitative comparison with baselines. More results are attached in the *Supplement*.

4.2. Evaluation

We perform both qualitative and quantitative evaluations for the proposed approach in terms of faithfulness and 3D consistency. We compare our method with three state-of-the-art inversion methods, HFGI [69], PTI [54] and IDE-3D [61]. HFGI [69] is the state-of-the-art encoder-based 2D GAN inversion method that achieves high-fidelity image reconstruction. Although PTI [54] was originally proposed for inversing 2D GAN, the method has been proven to achieve reasonable performance on 3D GAN inversion [12, 36]. IDE-3D [61] proposes a hybrid inversion approach on 3D GAN and trains an encoder that maps from the input image to the latent for initialization.

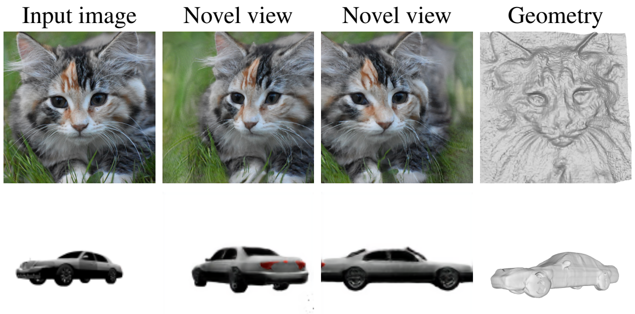


Figure 5. Inversion results on AFHQ-cats and ShapeNet-cars.

Method	PSNR↑	SSIM↑	Lpips↓	3D Consistency↑	ID ↑
PTI [54]	26.64	0.879	0.271	21.20	0.657
IDE-3D [61]	26.45	0.878	0.273	20.69	0.671
HFGI [68]	22.51	0.772	0.268	*N/A*	*N/A*
Ours	**29.43**	**0.918**	**0.172**	**21.69**	**0.744**

Table 1. Quantitative evaluation of different GAN inversion methods.

	Ours > PTI [54]	Ours > IDE-3D [61]
Preference rate	90.7	92.5

Table 2. The result of the user study.

4.2.1 Qualitative Analysis

We demonstrate the visual comparisons in Figure 4. Note that HFGI is for 2D GAN inversion, and thus it has a slightly different viewpoint and is inconsistent in 3D-aware editing. The proposed approach is robust to the out-of-distribution textures such as the tattoos and can faithfully reconstruct the image-specific details, while PTI and IDE-3D fail to keep the identity from the source for the novel views. We also provide additional inversion results on the AFHQ-cats [17] and ShapeNet-cars [14] datasets in Figure 5, which shows our method also works on other datasets.

4.2.2 Quantitative Analysis

Table 1 presents the quantitative comparison between our method and baselines. For test data, We randomly select 1500 images from the CelebA-HQ dataset. For reconstruction fidelity, we adopt PSNR, MS-SSIM, and LPIPS [77] on input image as the evaluation metrics. Our approach obtains the best scores on all the evaluated metrics compared with baselines, which indicates that our method reconstructs high-fidelity details of input images.

In terms of the 3D consistency, we adopt the evaluation setting from [26]. Specifically, we use 5 synthesized novel views near the input camera pose to predict the input image using the IBRNet [67] and calculate the difference with PSNR. The reported metrics in the novel view synthesis prove that more stable 3D consistency is achieved by our approach. Also, We randomly select one novel view for every CelebA-HQ test image and compute the mean Arcface [20] cosine similarity between rendered novel view images and input images to evaluate our 3D consistency. Our ID loss exceeds other baselines by a large margin. Note that the quantitative evaluation of the 3D consistency is still an open question, and we report more metrics in the *supplement*. We recommend readers watch our result videos for a comprehensive evaluation of 3D consistency.

4.2.3 User Study

We conduct a user study to perceptually evaluate the reconstruction quality of our approach. We use 14 random images from our test dataset and perform the 3D inversion. For each vote, the user is provided with the input image, the video rendered with a sphere camera trajectory that looks at the center of the face of our approach, and the video of baselines rendered with the same trajectory. We ask the participant to compare which video is preferred in the following two aspects: keeping the best identity of the input image and inducing the least flicker. As in Table 2, from the collected 1120 votes from 40 participants, the proposed method outperforms other baselines by a large margin.

4.2.4 Ablation Analysis

We conduct the ablation analysis to show the effectiveness of our design. Specifically, we implement two ablated models: (1) Without occluded parts, we remove our occluded part reconstruction and apply bilinear interpolation instead of generative priors for invisible pixels; (2) Without original textures, we use the generated texture instead of the original textures for the visible parts. Figure 6 show the qualitative comparison. We spot the following findings: (1) with only interpolation, the synthesized novel view contains obvious artifacts near the face boundary. The generated invisible part leads to obvious improvement in synthesizing the reasonable face shape; (2) with only the generated textures, the high-frequency details from the input images are lost.

4.3. Applications

Inversion of 3D GANs enables many applications, such as 3D-aware editing. We demonstrate two types of editing: latent-based attribute editing and texture-based image editing. For attribute editing, we follow the pipeline in [55] to calculate the moving direction in the latent space (Details attached in the *supplement*). We show transferring gender, changing ages, smiling, and wearing glasses in Figure 7(a) and render the corresponding 3D-consistent views. As our approach can handle out-of-distribution textures, we can

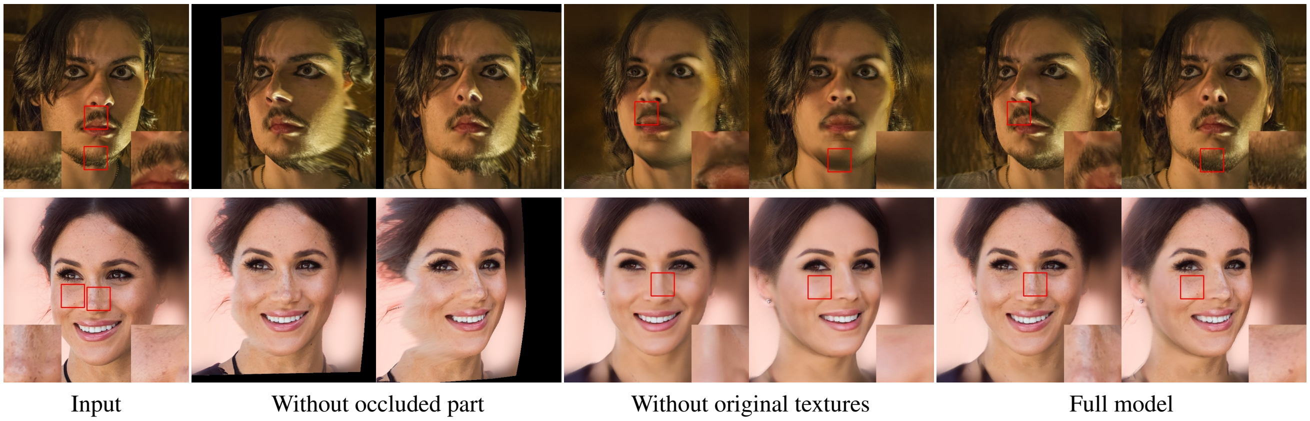
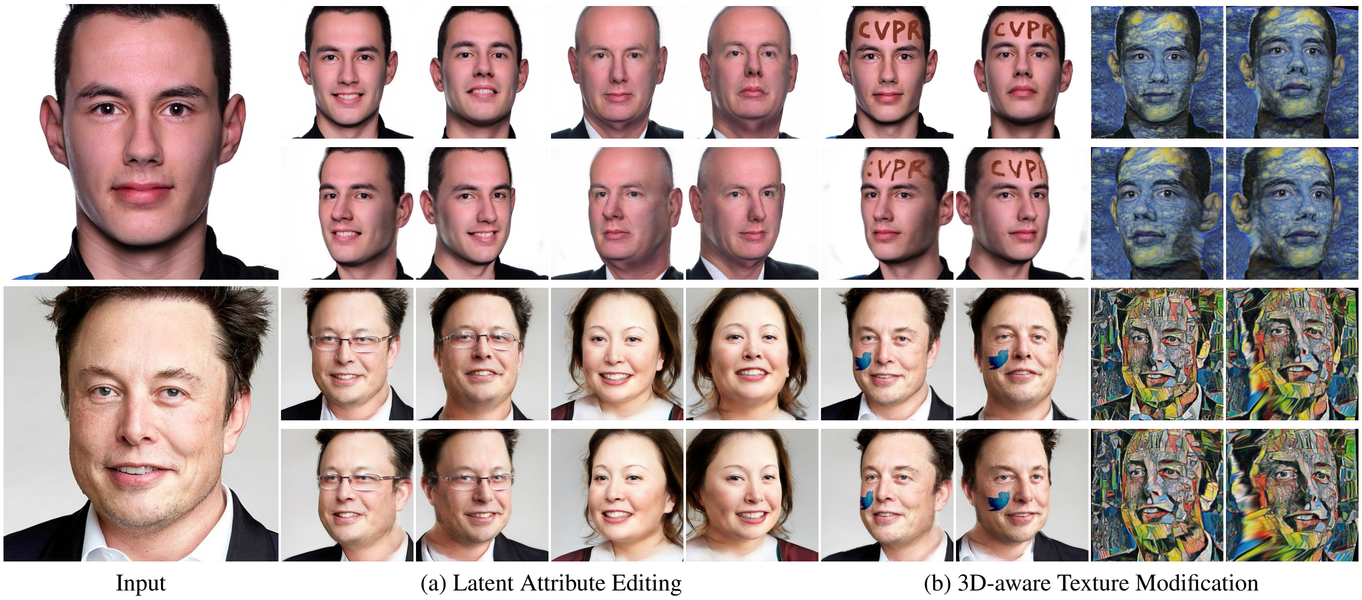

Input Without occluded part Without original textures Full model

Figure 6. Ablation study analysis. Without the occluded part reconstruction, the synthesized face shape is incorrect. Without the visible textures, the image-specific details, such as the beard and skin textures, are not preserved. Zoom in for details.



Input (a) Latent Attribute Editing (b) 3D-aware Texture Modification

Figure 7. 3D-aware editing results on real-life photos. Our approach enables two types of editing: (a) latent attribute editing where we can modify specific attributes such as smile, age, glasses, and gender. (b)3D-aware texture modification: by editing the textures on the input images such as adding logos or stylization.

perform texture editing on the input image and synthesize the novel view for the modified images. As in Figure 7(b), we can paint the desired textures (e.g., CVPR on the forehead and logo on the face) or apply different styles on the input view. We can generate 3D-consistent views for the edited input images.

5. Conclusion

This work studies high-fidelity 3D GAN inversion, which enables latent-based attribute editing and texture-based image editing. Extensive experiments demonstrate that our method can robustly synthesize the novel view of the input image with excellent detail preservation. The proposed pipeline is general and easy to apply as we can conduct the visibility analysis and the pseudo-multi-view generation for 3D-aware GANs. Still, our approach suffers from several limitations. One primary limitation is the difficulty in reconstructing the geometry of the out-of-distribution objects (e.g., trinkets and hands). As a result of the initially incorrect geometry, the following operations inevitably fail to synthesize reasonable results. Additionally, the estimated geometry for input with extreme poses may suffer from distortions. We attach examples of the failure cases in the *supplement*. Nevertheless, the proposed approach is promising to serve as a practical solution for 3D-aware reconstruction and editing with only a single input, and we expect future works to solve the remaining issues.

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4432–4441, 2019. 1, 2

[2] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan++: How to edit the embedded images? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8296–8305, 2020. 2

[3] Yuval Alaluf, Or Patashnik, and Daniel Cohen-Or. Restyle: A residual-based stylegan encoder via iterative refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6711–6720, 2021. 2

[4] Yuval Alaluf, Or Patashnik, Zongze Wu, Asif Zamir, Eli Shechtman, Dani Lischinski, and Daniel Cohen-Or. Third time's the charm? image and video editing with stylegan3. *arXiv preprint arXiv:2201.13433*, 2022. 2

[5] Yuval Alaluf, Omer Tov, Ron Mokady, Rinon Gal, and Amit Bermano. Hyperstyle: Stylegan inversion with hypernetworks for real image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18511–18521, 2022. 2

[6] Matan Atzmon and Yaron Lipman. SAL: Sign agnostic learning of shapes from raw data. In *Proc. CVPR*, 2020. 3

[7] Qingyan Bai, Yinghao Xu, Jiapeng Zhu, Weihao Xia, Yujiu Yang, and Yujun Shen. High-fidelity gan inversion with padding space. *arXiv preprint arXiv:2203.11105*, 2022. 2

[8] Shengqu Cai, Anton Obukhov, Dengxin Dai, and Luc Van Gool. Pix2nerf: Unsupervised conditional p-gan for single image to neural radiance fields translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3981–3990, 2022. 3

[9] Ang Cao and Justin Johnson. Hexplane: A fast representation for dynamic scenes. *arXiv preprint arXiv:2301.09632*, 2023. 3

[10] Ang Cao, Chris Rockwell, and Justin Johnson. Fwd: Real-time novel view synthesis with forward warping and depth. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15713–15724, 2022. 5

[11] Lucy Chai, Jun-Yan Zhu, Eli Shechtman, Phillip Isola, and Richard Zhang. Ensembling with deep generative views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14997–15007, 2021. 2

[12] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16123–16133, 2022. 1, 2, 3, 5, 6

[13] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5799–5809, 2021. 1, 2, 3

[14] Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. ShapeNet: An Information-Rich 3D Model Repository. Technical Report arXiv:1512.03012 [cs.GR], Stanford University — Princeton University — Toyota Technological Institute at Chicago, 2015. 7

[15] Qifeng Chen and Vladlen Koltun. Photographic image synthesis with cascaded refinement networks. In *Proc. ICCV*, 2017. 5

[16] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *Proc. CVPR*, 2019. 3

[17] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 7

[18] Edo Collins, Raja Bala, Bob Price, and Sabine Susstrunk. Editing in style: Uncovering the local semantics of gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5771–5780, 2020. 2

[19] Antonia Creswell and Anil Anthony Bharath. Inverting the generator of a generative adversarial network. *IEEE transactions on neural networks and learning systems*, 30(7):1967–1974, 2018. 1, 2

[20] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019. 7

[21] Yu Deng, Jiaolong Yang, Jianfeng Xiang, and Xin Tong. Gram: Generative radiance manifolds for 3d-aware image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10673–10683, 2022. 1, 3

[22] Tan M Dinh, Anh Tuan Tran, Rang Nguyen, and Binh-Son Hua. Hyperinverter: Improving stylegan inversion via hypernetwork. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11389–11398, 2022. 2

[23] Shubham Goel, Angjoo Kanazawa, and Jitendra Malik. Shape and viewpoint without keypoints. In *Proc. ECCV*, 2020. 3

[24] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Proc. NeurIPS*, 2014. 3

[25] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. In *Proc. ICML*, 2020. 3

[26] 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. 7

[27] Jinjin Gu, Yujun Shen, and Bolei Zhou. Image processing using multi-code gan prior. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3012–3021, 2020. 2

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

[29] Philipp Henzler, Niloy J Mitra, and Tobias Ritschel. Escaping plato's cave: 3d shape from adversarial rendering. In *Proc. ICCV*, 2019. 3

[30] Minyoung Huh, Richard Zhang, Jun-Yan Zhu, Sylvain Paris, and Aaron Hertzmann. Transforming and projecting images into class-conditional generative networks. In *European Conference on Computer Vision*, pages 17–34. Springer, 2020. 2

[31] Angjoo Kanazawa, Shubham Tulsiani, Alexei A. Efros, and Jitendra Malik. Learning category-specific mesh reconstruction from image collections. In *Proc. ECCV*, 2018. 3

[32] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proc. CVPR*, 2019. 3, 5

[33] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *Proc. CVPR*, 2020. 3

[34] Jaehoon Ko, Kyusun Cho, Daewon Choi, Kwangrok Ryoo, and Seungryong Kim. 3d gan inversion with pose optimization. *arXiv preprint arXiv:2210.07301*, 2022. 3

[35] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5549–5558, 2020. 5

[36] Connor Z Lin, David B Lindell, Eric R Chan, and Gordon Wetzstein. 3d gan inversion for controllable portrait image animation. *arXiv preprint arXiv:2203.13441*, 2022. 1, 3, 6

[37] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proc. CVPR*, 2019. 3

[38] Mateusz Michalkiewicz, Jhony K Pontes, Dominic Jack, Mahsa Baktashmotlagh, and Anders Eriksson. Implicit surface representations as layers in neural networks. In *Proc. ICCV*, 2019. 3

[39] 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 *Proc. ECCV*, 2020. 3

[40] Siva Karthik Mustikovela, Varun Jampani, Shalini De Mello, Sifei Liu, Umar Iqbal, Carsten Rother, and Jan Kautz. Self-supervised viewpoint learning from image collections. In *Proc. CVPR*, 2020. 3

[41] Thu Nguyen-Phuoc, Chuan Li, Lucas Theis, Christian Richardt, and Yong-Liang Yang. Hologan: Unsupervised learning of 3d representations from natural images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7588–7597, 2019. 3

[42] Thu H Nguyen-Phuoc, Christian Richardt, Long Mai, Yongliang Yang, and Niloy Mitra. Blockgan: Learning 3d object-aware scene representations from unlabelled images. *Advances in Neural Information Processing Systems*, 33:6767–6778, 2020. 3

[43] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11453–11464, 2021. 3

[44] Yotam Nitzan, Amit Bermano, Yangyan Li, and Daniel Cohen-Or. Face identity disentanglement via latent space mapping. *arXiv preprint arXiv:2005.07728*, 2020. 2

[45] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13503–13513, 2022. 1, 3

[46] Hao Ouyang, Tengfei Wang, and Qifeng Chen. Internal video inpainting by implicit long-range propagation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14579–14588, 2021. 5

[47] Hao Ouyang, Bo Zhang, Pan Zhang, Hao Yang, Jiaolong Yang, Dong Chen, Qifeng Chen, and Fang Wen. Real-time neural character rendering with pose-guided multiplane images. *arXiv preprint arXiv:2204.11820*, 2022. 3

[48] Xingang Pan, Xiaohang Zhan, Bo Dai, Dahua Lin, Chen Change Loy, and Ping Luo. Exploiting deep generative prior for versatile image restoration and manipulation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 2

[49] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proc. CVPR*, 2019. 3

[50] Songyou Peng, Michael Niemeyer, Lars Mescheder, Marc Pollefeys, and Andreas Geiger. Convolutional occupancy networks. In *Proc. ECCV*, 2020. 3

[51] Guim Perarnau, Joost Van De Weijer, Bogdan Raducanu, and Jose M Álvarez. Invertible conditional gans for image editing. *arXiv preprint arXiv:1611.06355*, 2016. 1, 2

[52] Pierluigi Zama Ramirez, Diego Martin Arroyo, Alessio Tonioni, and Federico Tombari. Unsupervised novel view synthesis from a single image. *arXiv preprint arXiv:2102.03285*, 2021. 3

[53] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2287–2296, 2021. 1, 2

[54] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *ACM Transactions on Graphics (TOG)*, 42(1):1–13, 2022. 1, 2, 5, 6, 7

[55] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9243–9252, 2020. 1, 2, 7

[56] 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. 3

[57] Zifan Shi, Yujun Shen, Yinghao Xu, Sida Peng, Yiyi Liao, Sheng Guo, Qifeng Chen, and Dit-Yan Yeung. Learning 3d-aware image synthesis with unknown pose distribution. *arXiv preprint arXiv:2301.07702*, 2023. 3

[58] Zifan Shi, Yinghao Xu, Yujun Shen, Deli Zhao, Qifeng Chen, and Dit-Yan Yeung. Improving 3d-aware image synthesis with a geometry-aware discriminator. *arXiv preprint arXiv:2209.15637*, 2022. 3

[59] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5

[60] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. In *Proc. NeurIPS 2019*, 2019. 3

[61] Jingxiang Sun, Xuan Wang, Yichun Shi, Lizhen Wang, Jue Wang, and Yebin Liu. Ide-3d: Interactive disentangled editing for high-resolution 3d-aware portrait synthesis. *arXiv preprint arXiv:2205.15517*, 2022. 3, 6, 7

[62] Jingxiang Sun, Xuan Wang, Yong Zhang, Xiaoyu Li, Qi Zhang, Yebin Liu, and Jue Wang. Fenerf: Face editing in neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7672–7682, 2022. 3

[63] Maxim Tatarchenko, Alexey Dosovitskiy, and Thomas Brox. Multi-view 3D models from single images with a convolutional network. In *Proc. ECCV*, 2016. 3

[64] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. *ACM Transactions on Graphics (TOG)*, 40(4):1–14, 2021. 2

[65] Shubham Tulsiani, Hao Su, Leonidas J. Guibas, Alexei A. Efros, and Jitendra Malik. Learning shape abstractions by assembling volumetric primitives. In *Proc. CVPR*, 2017. 3

[66] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021. 3

[67] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P Srinivasan, Howard Zhou, Jonathan T Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2021. 7

[68] Tengfei Wang, Hao Ouyang, and Qifeng Chen. Image inpainting with external-internal learning and monochromic bottleneck. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5120–5129, 2021. 5, 7

[69] Tengfei Wang, Yong Zhang, Yanbo Fan, Jue Wang, and Qifeng Chen. High-fidelity gan inversion for image attribute editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11379–11388, 2022. 1, 2, 6

[70] Youjia Wang, Teng Xu, Yiwen Wu, Minzhang Li, Wenzheng Chen, Lan Xu, and Jingyi Yu. Narrate: A normal assisted free-view portrait stylizer. *arXiv preprint arXiv:2207.00974*, 2022. 3

[71] Zongze Wu, Dani Lischinski, and Eli Shechtman. Stylespace analysis: Disentangled controls for stylegan image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12863–12872, 2021. 2

[72] Weihao Xia, Yulun Zhang, Yujiu Yang, Jing-Hao Xue, Bolei Zhou, and Ming-Hsuan Yang. Gan inversion: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 1, 2

[73] Jiaxin Xie, Chenyang Lei, Zhuwen Li, Li Erran Li, and Qifeng Chen. Video depth estimation by fusing flow-to-depth proposals. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10100–10107. IEEE, 2020. 3

[74] Rui Xu, Xiaoxiao Li, Bolei Zhou, and Chen Change Loy. Deep flow-guided video inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3723–3732, 2019. 5

[75] Sicheng Xu, Jiaolong Yang, Dong Chen, Fang Wen, Yu Deng, Yunde Jia, and Tong Xin. Deep 3d portrait from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7710–7720, 2020. 3

[76] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4471–4480, 2019. 5

[77] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 5, 7

[78] Shengyu Zhao, Jonathan Cui, Yilun Sheng, Yue Dong, Xiao Liang, Eric I Chang, and Yan Xu. Large scale image completion via co-modulated generative adversarial networks. *arXiv preprint arXiv:2103.10428*, 2021. 5

[79] Xiaoming Zhao, Fangchang Ma, David Güera, Zhile Ren, Alexander G Schwing, and Alex Colburn. Generative multiplane images: Making a 2d gan 3d-aware. In *European Conference on Computer Vision*, pages 18–35. Springer, 2022. 3

[80] Peng Zhou, Lingxi Xie, Bingbing Ni, and Qi Tian. Cips-3d: A 3d-aware generator of gans based on conditionally-independent pixel synthesis. *arXiv preprint arXiv:2110.09788*, 2021. 3

[81] Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou. In-domain gan inversion for real image editing. In *European conference on computer vision*, pages 592–608. Springer, 2020. 2

[82] Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A Efros. Generative visual manipulation on the natural image manifold. In *European conference on computer vision*, pages 597–613. Springer, 2016. 2

[83] Peihao Zhu, Rameen Abdal, Yipeng Qin, John Femiani, and Peter Wonka. Improved stylegan embedding: Where are the good latents? *arXiv preprint arXiv:2012.09036*, 2020. 2