NeRFInvertor: High Fidelity NeRF-GAN Inversion
for Single-shot Real Image Animation

Yu Yin[1], Kamran Ghasedi[2], HsiangTao Wu[2], Jiaolong Yang[2], Xin Tong[2], Yun Fu[1]

[1] Northeastern University

yin.yu1@northeastern.edu, yunfu@ece.neu.edu

[2] Microsoft

{kamrang,musclewu,jiaoyan,xtong}@microsoft.com

Abstract

Nerf-based Generative models have shown impressive capacity in generating high-quality images with consistent 3D geometry. Despite successful synthesis of fake identity images randomly sampled from latent space, adopting these models for generating face images of real subjects is still a challenging task due to its so-called inversion issue. In this paper, we propose a universal method to surgically fine-tune these NeRF-GAN models in order to achieve high-fidelity animation of real subjects only by a single image. Given the optimized latent code for an out-of-domain real image, we employ 2D loss functions on the rendered image to reduce the identity gap. Furthermore, our method leverages explicit and implicit 3D regularizations using the in-domain neighborhood samples around the optimized latent code to remove geometrical and visual artifacts. Our experiments confirm the effectiveness of our method in realistic, high-fidelity, and 3D consistent animation of real faces on multiple NeRF-GAN models across different datasets.

1. Introduction

Animating a human with a novel view and expression sequence from a single image opens the door to a wide range of creative applications, such as talking head synthesis [22, 34], augmented and virtual reality (AR/VR) [19], image manipulation [24, 32, 45], as well as data augmentation for training of deep models [25, 42, 43]. Early works of image animation mostly employed either 2D-based image generation models [14, 26, 31, 37], or 3D parametric models [4, 11, 40, 41] (*e.g.* 3DMM [6]), but they mostly suffer from artifacts, 3D inconsistencies or unrealistic visuals.

Representing scenes as Neural Radiance Fields (NeRF) [23] has recently emerged as a breakthrough

---

Project page: https://yuyin1.github.io/NeRFInvertor_Homepage/

Figure 1. **Image animation results of our method.** *NeRFInvertor* achieves 3D-consistent and ID-preserving animation (*i.e.* novel views and expressions) of real subjects given only a single image.

approach for generating high-quality images of a scene in novel views. However, the original NeRF models [5, 33, 44] only synthesize images of a static scene and require extensive multi-view data for training, restricting its application to novel view synthesis from a single image. Several studies have shown more recent advances in NeRFs by extending it to generate multi-view face images with single-shot data even with controllable expressions [7, 8, 12, 27, 38, 46]. These Nerf-based Generative models (NeRF-GANs) are able to embed attributes of training samples into their latent

variables, and synthesize new identity face images with different expressions and poses by sampling from their latent space.

While animatable synthesis of fake identity images is impressive, it is still challenging to generate 3D-consistent and identity-preserving images of real faces. Specifically, current Nerf-GANs have difficulties to accurately translate out-of-domain images into their latent space, and consequently change identity attributes and/or introduce artifacts when applied to most real-world images. In order to synthesize real faces, the conventional method applies optimization algorithms to invert the input image to a latent code in a smaller (*i.e.* $\mathcal{W}$) or an extended (*i.e.* $\mathcal{W}+$) NeRF-GAN latent space. However, they both either have ID-preserving or artifacts issues as shown in Figure 2. The $\mathcal{W}$ space inversion, in particular, generates realistic novel views and clean 3D geometries, but suffers from the identity gap between the real and synthesized images. In contrast, the $\mathcal{W}+$ space inversion well preserves the identity but commonly generates an inaccurate 3D geometry, resulting in visual artifacts when exhibited from new viewpoints. Hence, it remains as a trade-off to have a 3D-consistent geometry or preserve identity attributes when inverting face images out of latent space distribution.

In this paper, we present *NeRFInvertor* as a universal inversion method for NeRF-GAN models to achieve high-fidelity, 3D-consistent, and identity-preserving animation of real subjects given only a single image. Our method is applicable to most of NeRF-GANs trained for a *static* or *dynamic* scenes, and hence accomplishes synthesis of real images with both novel views and novel expressions (see Figure 1). Since the real images are mostly out of the domain of NeRF-GANs latent space, we surgically fine-tune their generator to enrich the latent space by leveraging the single input image without degrading the learned geometries.

In particular, given an optimized latent code for the input image, we first use image space supervision to narrow the identity gap between the synthesized and input images. Without a doubt, the fine-tuned model can be overfitted on the input image and well reconstruct the input in the original view. However, fine-tuning with just image space supervision produces erroneous 3D geometry due to the insufficient geometry and content information in a single image, resulting in visual artifacts in novel views. To overcome this issue, we introduce regularizations using the surrounding samples in the latent space, providing crucial guidance for the unobserved part in the image space. By sampling latent codes from the neighborhood of optimized latent variables with different poses and expressions, we enforce a novel geometric constraint on the density outputs of fine-tuned and original pretrained generators. We also further add regularizations on the rendered images of neighborhood samples obtained from the fine-tuned and pretrained genera-
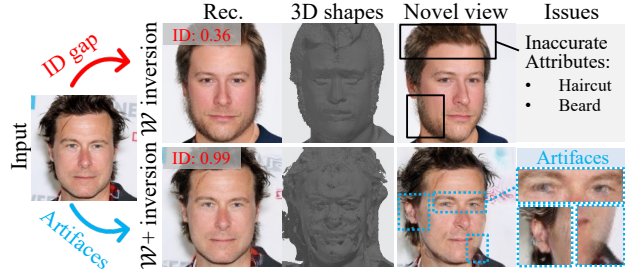

Figure 2. **Trade-off between ID-preserving and removing artifacts.** Optimizing latent variables of Nerf-GANs for synthesis of a real face leads to a trade-off between identity-preserving and geometrical and visual artifacts. Specifically, $\mathcal{W}$ space inversion results in clean geometry but identity gap between real and generated images, and $\mathcal{W}+$ space inversion causes preserving of identity attributes but inaccurate geometry and visual artifacts.

tors. These regularizations help us to leverage the geometry and content information of those in-domain neighborhood samples around the input. Our experiments validate the effectiveness of our method in realistic, high-fidelity, and 3D consistent animating of real face images.

The main contributions of this paper are as follows:

1. We proposed a universal method for inverting NeRF-GANs to achieve 3D-consistent, high-fidelity, and identity-preserving animation of real subjects given only a single image.

2. We introduce a novel geometric constraint by leveraging density outputs of in-domain samples around the input to provide crucial guidance for the unobserved part in the 2D space.

3. We demonstrate the effusiveness of our method on multiple NeRF-GAN models across different datasets.

2. Related Work

2.1. NeRF-GANs

Recently, the impressive performance of NeRF-GANs has demonstrated its potential as a promising research direction. GRAF [30] and Pi-GAN [9] are two early attempts that proposed generative models for radiance fields for 3D-aware image synthesis from the unstructured 2D images. Several recent studies (*e.g.* GRAM [12] and EG3D [8]) have enhanced synthesis quality with higher resolutions, better 3D geometry, and faster rendering. Furthermore, AniFaceGAN [38] introduced a deformable NeRF-GAN for dynamic scenes capable of synthesizing faces with controllable pose and expression. In this paper, we demonstrate the effectiveness of our method on NeRF-GANs for both static (GRAM and EG3D) and dynamic (AniFaceGAN) scenes.
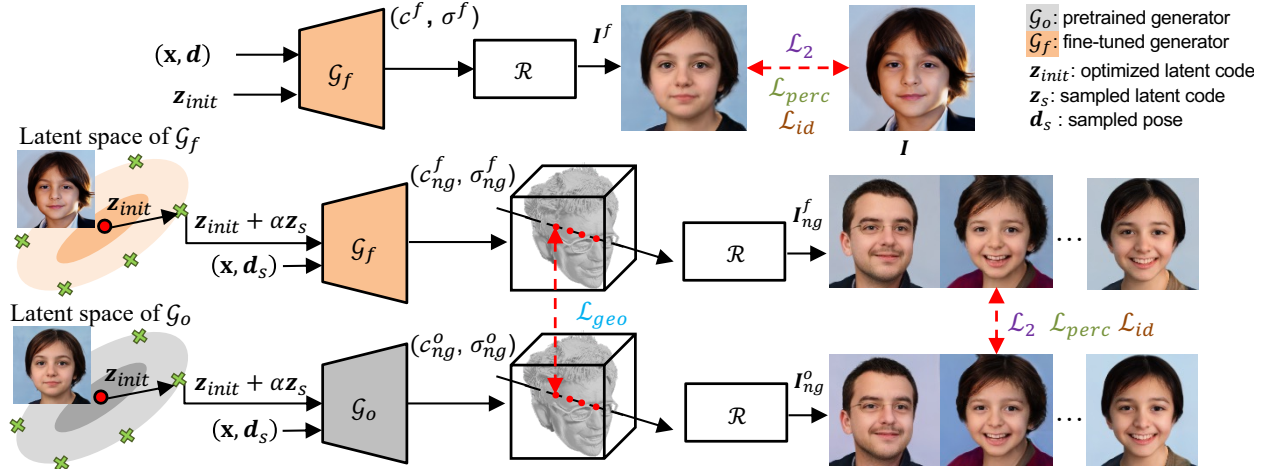
Figure 3. **Framework of *NeRFInvertor*.** Given the optimized latent code $\mathbf{z}_{init}$, we fine-tune the generator and first apply image space supervision to push the generated image to match the input image in the original view $\boldsymbol{d}$. To augment the NeRF-GAN manifold without worrying about visual artifacts in novel views, we then leverage the surrounding samples of the optimized latent code to regularize the realism and fidelity of the novel view and expression synthesis.

2.2. GAN Inversion

GAN priors can be beneficial in a variety of applications such as image editing [3,10,28] and face restoration [35,39]. To utilize the priors, the input image needs to be inverted into a latent code that optimally reconstructs the given image using a pretrained generator. The process is known as GAN inversion, and it is a well-known issue in 2D GANs (*e.g.* StyleGAN [16–18]). Abdal *et al.* [1,2] used a straight optimization method to transform an image into the latent space of StyleGAN and observed a distortion-editability trade-off between $\mathcal{W}$ and $\mathcal{W}+$ space when editing images. PTI [29] used an optimized latent code as an initial point and fine-tuned the generator, achieving good results in the process of reducing the distortion-editability trade-off.

To exploit NeRF-GAN priors for 3D tasks, recent attempts inverted the real image to NeRF-GAN latent space using the traditional optimization method [46], additive encoder [7,27], or PTI [8,20]. However, applying these 2D-based inversion methods directly to 3D-GAN models makes them insensitive to subtle geometric alterations in 3D space. To address this issue, we propose explicit geometrical regularization and implicit geometrical regularization to help produce better 3D geometry of the input.

2.3. Single-shot NeRFs

Due to the insufficient geometry and content information in a single image, single-shot NeRFs without additional supervision (*i.e.* multi-view images or 3D objects) remain challenging. Recently, Pix2NeRF [7] extended pi-GAN with the encoder to obtain a conditional single-shot

NeRF models. HeadNeRF [15] incorporated the NeRF into the parametric representation of the human head. Since the whole process is differentiable, the model can generate a NeRF representation from a single image using image fitting. However, there remains a notable identity gap between the synthesized results and input images, indicating the inefficiency of single-shot NeRFs for real image inversion.

3. Method

In this paper, we present *NeRFInvertor* as a universal NeRF-GAN inversion method to translate a single real image into a NeRF representation. Given an input image $\boldsymbol{I}$, the goal is to generate novel views or expressions of $\boldsymbol{I}$ using a pretrained Nerf-GAN $\mathcal{G}_o$. To do this, we first find an in-domain latent code[1] capable of generating an image $\boldsymbol{I}^o$ as close as possible to the input image. Given a NeRF-GAN generator $\mathcal{G}_{(\cdot)}$, the image $\boldsymbol{I}^{(\cdot)}$ can be synthesized as:

$$\boldsymbol{I}^{(\cdot)} = \mathcal{R}(\mathcal{G}_{(\cdot)}(\mathbf{z}, \mathbf{x}), \boldsymbol{d}) , \qquad (1)$$

where $\mathbf{x} \in \mathbb{R}^3$ is the a 3D location, $\mathbf{z}$ denotes a latent code, $\boldsymbol{d} \in \mathbb{R}^3$ is the camera pose, and $\mathcal{R}$ is the volume renderer discussed in [21, 23]. Therefore, the optimized latent code $\mathbf{z}_{init}$ can be computed as:

$$\mathbf{z}_{init} = \underset{\mathbf{z}}{\text{argmin}} \ \ \mathcal{L}_{perc}(\boldsymbol{I}, \boldsymbol{I}^o) + \lambda_0 \mathcal{L}_{pix}(\boldsymbol{I}, \boldsymbol{I}^o) , \qquad (2)$$

where $\mathcal{L}_{perc}$, $\mathcal{L}_{pix}$ represent perceptual and l2-norm pixel-wise loss functions, respectively.

---

[1]We optimize latent codes in $\mathcal{Z}$ or $\mathcal{W}$ space [18] for different NeRF-GAN models.

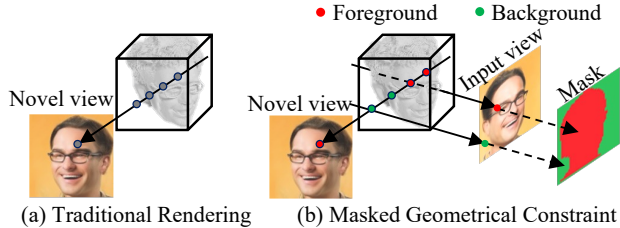(a) Traditional Rendering   (b) Masked Geometrical Constraint

Figure 4. **Masked Geometrical Constraint.** (a) It shows the traditional rendering process for a novel view image. (b) Masked geometrical constraint takes only the foreground points into account. To render the novel view, only the red dots are used for image rendering. The two green dots are ignored, as they are definitely from a background region.

Usually, there would be a gap between the generated and real images, since the real images are mostly the out-of-domain samples in NeRF-GANs. To solve this problem, we propose a fine-tuning process with novel regularizations in the following sections. Specifically, we fine-tune the generator with image space loss functions (Sec. 3.1) to reduce the identity gap. We also apply an explicit geometrical constraint (Sec. 3.2) and an implicit geometrical regularization (Sec. 3.3) to maintain the model's ability to produce high-quality and 3D-consistent images.

3.1. Image Space Supervision

Given the optimized latent code $\mathbf{z}_{init}$, we fine-tune the generator using image space supervision by pushing the generated image to match the input image in the original view $\boldsymbol{d}$. Denoting the fine-tuned generator as $\mathcal{G}_f$, it takes the optimized latent code $\mathbf{z}_{init}$ and a 3D location $\mathbf{x} \in \mathbb{R}^3$ as the inputs and outputs a color $c \in \mathbb{R}^3$ and a volume density $\sigma \in \mathbb{R}^1$ for each location. For the given view $\boldsymbol{d}$, we can then accumulate the colors and densities into a 2D image $\boldsymbol{I}^f$. Formally, the image $\boldsymbol{I}^f$ can be expressed as:

$$\boldsymbol{I}^f = \mathcal{R}(\mathcal{G}_f(\mathbf{z}_{init}, \mathbf{x}), \boldsymbol{d}) \tag{3}$$

We employ the following loss function as the image space supervision:

$$\mathcal{L}_{img} = \lambda_1 \mathcal{L}_{pix}(\boldsymbol{I}^f, \boldsymbol{I}) + \lambda_2 \mathcal{L}_{perc}(\boldsymbol{I}^f, \boldsymbol{I}) + \lambda_3 \mathcal{L}_{id}(\boldsymbol{I}^f, \boldsymbol{I}) , \tag{4}$$

where $\mathcal{L}_{perc}$, $\mathcal{L}_{pix}$ and $\mathcal{L}_{id}$ indicate perceptual, l2-norm pixel-wise and identity losses, respectively. λ_1, λ_2 and λ_3 are hyper-parameters of the losses. With the image space supervision, the fine-tuned model well reconstructs the input in the original view, but is prone to overfitting on the input image, causing artifacts in novel view synthesized images and inaccurate 3D geometry of the subject.

Figure 5. **The fogging artifacts around the hairline and/or in cheek region.**

3.2. Explicit Geometrical Regularization

To enrich the NeRF-GANs manifold using the input image attributes without worrying about visual artifacts in novel views, we relax the assumption of strict image space alignment described in Sec. 3.1. In order to regularize the model, we leverage the neighborhood samples around the optimized latent code to enhance the geometry, realism, and fidelity of the novel view and expression synthesis.

We first randomly sample different neighborhood latent codes $\mathbf{z}_{ng}$ surrounding the optimized latent code with various poses and expressions. The neighborhood latent codes can be obtained by:

$$\mathbf{z}_{ng} = \mathbf{z}_{init} + \alpha \frac{\mathbf{z}_{smp} - \mathbf{z}_{init}}{||\mathbf{z}_{smp} - \mathbf{z}_{init}||_2} , \tag{5}$$

where α is the interpolation distance between the randomly sampled $\mathbf{z}_{smp} \sim \mathcal{N}(\mathbf{0}, \mathbf{1})$ and optimized latent variable $\mathbf{z}_{init}$. To leverage the high-fidelity qualities of the original generator $\mathcal{G}_o$, we force the fine-tuned generator $\mathcal{G}_f$ to perform the same as $\mathcal{G}_o$ on the neighborhood latent codes, poses, and expressions. The geometrical constraint is defined based on both the color and density outputs of neighborhood samples on $\mathcal{G}_f$ and $\mathcal{G}_o$, that are expressed as:

$$\begin{aligned} c_{ng}^f, \sigma_{ng}^f &= \mathcal{G}_f(\mathbf{z}_{ng}, \mathbf{x}), \\ c_{ng}^o, \sigma_{ng}^o &= \mathcal{G}_o(\mathbf{z}_{ng}, \mathbf{x}), \end{aligned} \tag{6}$$

By reprojecting each ray (*i.e.* pixel) to 3D space according to its depth, we define two sets of point clouds S_o and S_f based on $\mathcal{G}_o$ and $\mathcal{G}_f$ images. We compare the similarity of two point clouds using the Chamfer distance, and define the geometrical constraint as follows:

$$\mathcal{L}_{exp} = \frac{1}{|S_o|} \sum_{p_o \in S_o} min_{p_f \in S_f} ||\sigma_{ng}^f(p_f) - \sigma_{ng}^o(p_o)||_2^2 +$$
$$||c_{ng}^f(p_f) - c_{ng}^o(p_o)||_2^2 \tag{7}$$

where p_o and p_f are the 3D locations in the point cloud sets S_o and S_f.

3.3. Implicit Geometrical Regularization

Moreover, we also add implicit geometrical regularizations on the rendering results of fine-tuned and the pre-
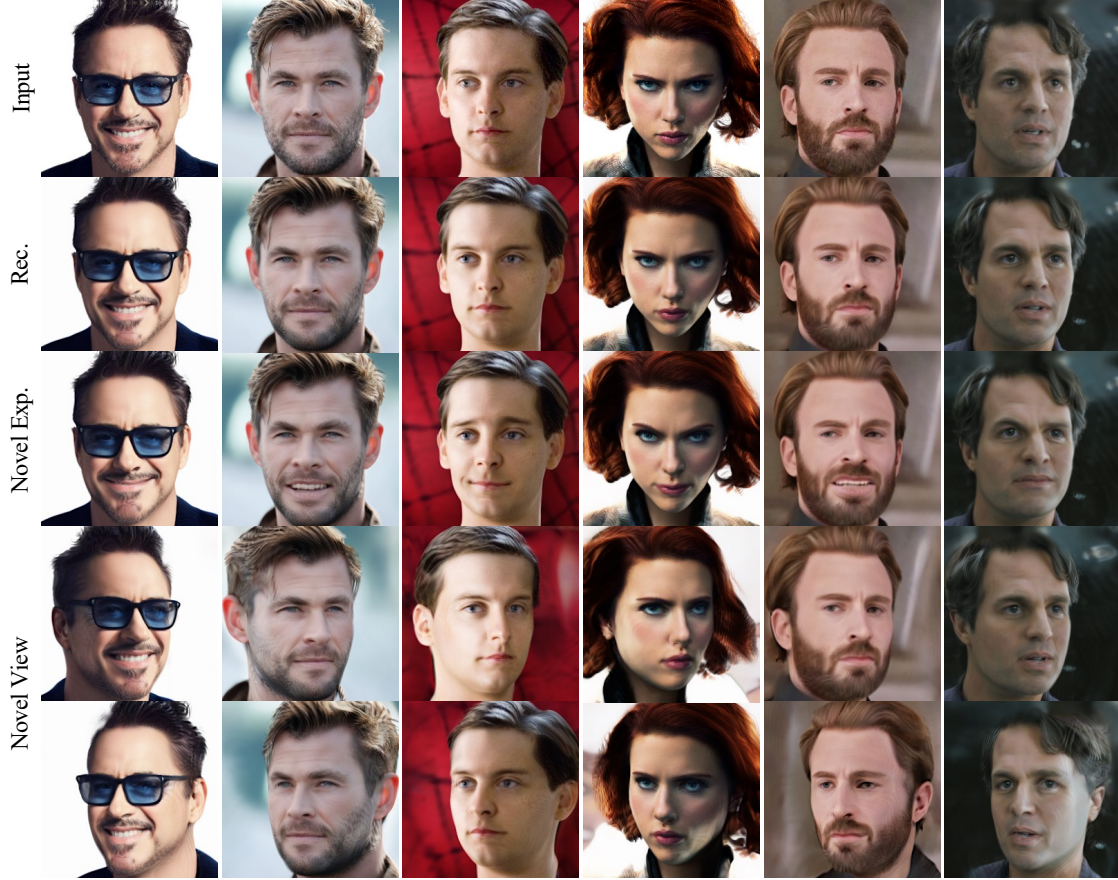
Figure 6. **Real image animation example on the "Avengers".** *NeRFInvertor* synthesizes realistic faces using the pretrained AniFaceGAN with controllable pose and expression sequences given a single training image. As can be seen, *NeRFInvertor* not only is capable of preserving identity attributes but also generates images with high quality and consistent appearance across different poses and expressions.

trained generator. Given a novel view $\boldsymbol{d}_s$, the rendered image of $\mathcal{G}_f$ and $\mathcal{G}_o$ can be expressed as:

$$
\begin{aligned}
\boldsymbol{I}_{ng}^f &= \mathcal{R}((c_{ng}^f, \sigma_{ng}^f), \boldsymbol{d}_s), \\
\boldsymbol{I}_{ng}^o &= \mathcal{R}((c_{ng}^o, \sigma_{ng}^o), \boldsymbol{d}_s).
\end{aligned}
\tag{8}
$$

We minimize the distance between the image generated by $\mathcal{G}_f$ and $\mathcal{G}_o$ using pixel-wise, perceptual, and identity losses. Hence, the overall loss can be expressed as:

$$
\begin{aligned}
\mathcal{L} =&\, \mathcal{L}_{img} + \lambda_4 \mathcal{L}_{exp} + \lambda_5 \mathcal{L}_{pix}(\boldsymbol{I}_{ng}^f, \boldsymbol{I}_{ng}^o) + \\
&\, \lambda_6 \mathcal{L}_{perc}(\boldsymbol{I}_{ng}^f, \boldsymbol{I}_{ng}^o) + \lambda_7 \mathcal{L}_{id}(\boldsymbol{I}_{ng}^f, \boldsymbol{I}_{ng}^o),
\end{aligned}
\tag{9}
$$

3.4. Masked Regularizations

Given a single-view image, we fine-tune the model with novel regularizations to achieve better 3D geometry and higher fidelity images in novel views. However, we still noticed some fogging parts around the hair or cheek as shown in Figure 5. In order to remove artifacts and get more accurate geometry, We enhance our geometrical and image reg-

ularizations by a mask, which is based on matting information on the input image. As shown in Figure 4, we predict the mask for the input view image based on the foreground and background regions shown by the red and green colors. In particular, if a shooting ray reaches the foreground region of the input image, we classify all the sample points on that ray as foreground points. Similarly, if a shooting ray reaches the background region of the image, we classify all the data points on that ray as background points. In the masked constraints, we take only the foreground (red) points into account for neighborhood density, color, and image rendering, and ignore the background (green) points.

4. Experiments

In this section, we compare our approach to existing inversion methods as well as single-shot NeRF models qualitatively and quantitatively. We validate our method on multiple NeRF-GANs trained for static (*i.e.* GRAM and EG3D) or dynamic scenes (*i.e.* AniFaceGAN). Evaluations are per-

Figure 7. **Comparison with prior inversion methods.**

Figure 8. **Comparison with single-shot NeRF methods.**

formed on reconstruction, novel view, and expression synthesis. Finally, we conduct ablation studies to disclose the contributions of the proposed components of our work.

4.1. Implementation Details

We use AniFaceGAN generator by default for the experiments unless otherwise specified. The original generators are pre-trained on the FFHQ and Cats datasets to synthesize animated human faces and cat faces, respectively. We evaluate face animation, including novel view and expression synthesis, on FFHQ, CelebA-HQ datasets, and a collection of notable individuals as out-of-domain samples to emphasize the *NeRFInvertor*'s identity-preserving capability.

We describe the training details of our *NeRFInvertor* applied on GRAM and EG3D in the Supplementary Material. For AniFaceGAN, the model is fine-tuned for 500 iterations. Hyper-parameters were set as follows: $\alpha = 5$, $\lambda_0 = 0.1$, $\lambda_1 = 1$, $\lambda_2 = 10$, $\lambda_3 = 0.1$, $\lambda_4 = 10$, $\lambda_5 = 1$, $\lambda_6 = 10$, $\lambda_7 = 0.1$. The model is trained on 2 Nvidia RTX GPUs at the resolution of 128×128. We used an ADAM optimizer with a learning rate of $2e^{-5}$. Our model takes approximately 30 minutes for training. We followed previous works (*i.e.* EG3D, GRAM) to use Deep3DFace [13] for estimating the camera pose of images.

4.2. Qualitative Evaluation

Comparison with inversion methods. We start by qualitatively comparing our approach to prior inversion methods including I2S ($\mathcal{W}$), and I2S ($\mathcal{W}+$) and PTI. I2S [2] employs
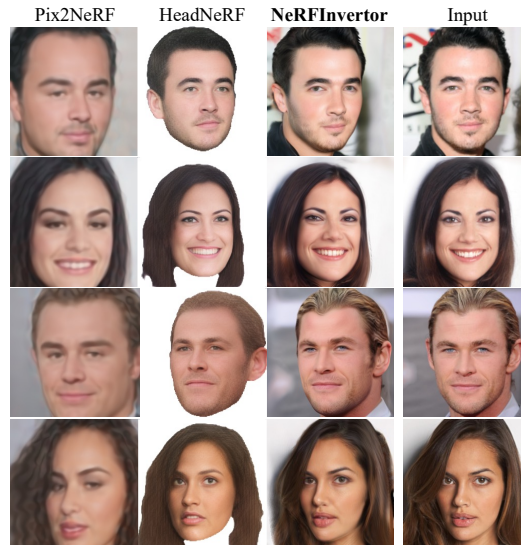
the conventional optimization strategy to invert real images to either the smaller $\mathcal{W}$ space or the extended $\mathcal{W}+$ space. PTI [29] is an inversion method for 2D-GANs with excellent performance. Figure 7 shows a qualitative comparison of these methods for novel view synthesis from a single input image. I2S ($\mathcal{W}$) is able to generate images with acceptable visual quality, however there is a notable identity gap with the input image. In contrast, I2S ($\mathcal{W}+$) keeps identity attributes but introduces lots of artifacts in novel views. PTI shows better performance than I2S ($\mathcal{W}$) and I2S ($\mathcal{W}+$) by fine-tuning the model with image space supervisions. However, since it lacks explicit constraints in 3D space, it fails to generate accurate geometry of the input subject and the hidden content in the original view. Our *NeRFInvertor* outperforms these methods in terms of visual quality and identity preservation. Furthermore, our method can generate novel view images with similar texture to the input image. For example, the visual details in Figure 7 such as wrinkle on the forehead (1st row) and dimples in the cheeks (3rd row) are well maintained in novel views. In the Supplementary Material, we demonstrate reconstruction results and show more novel views and expressions synthesis.

Comparison with single-shot NeRF methods. We compare our method with single-image view synthesis methods, Pix2NeRF and HeadNeRF, in Figure 8. Statistical comparisons are reported in the Supplementary Material. Pix2NeRF [7] proposed an encoder to translate images to the Pi-GAN [9] latent space and jointly trained the encoder and the Pi-GAN generator. In contrast to Pix2NeRF and our *NeRFInvertor*, which utilize GAN priors for novel view synthesis, HeadNeRF [15] proposed a NeRF-based parametric head model to synthesize images with various poses.
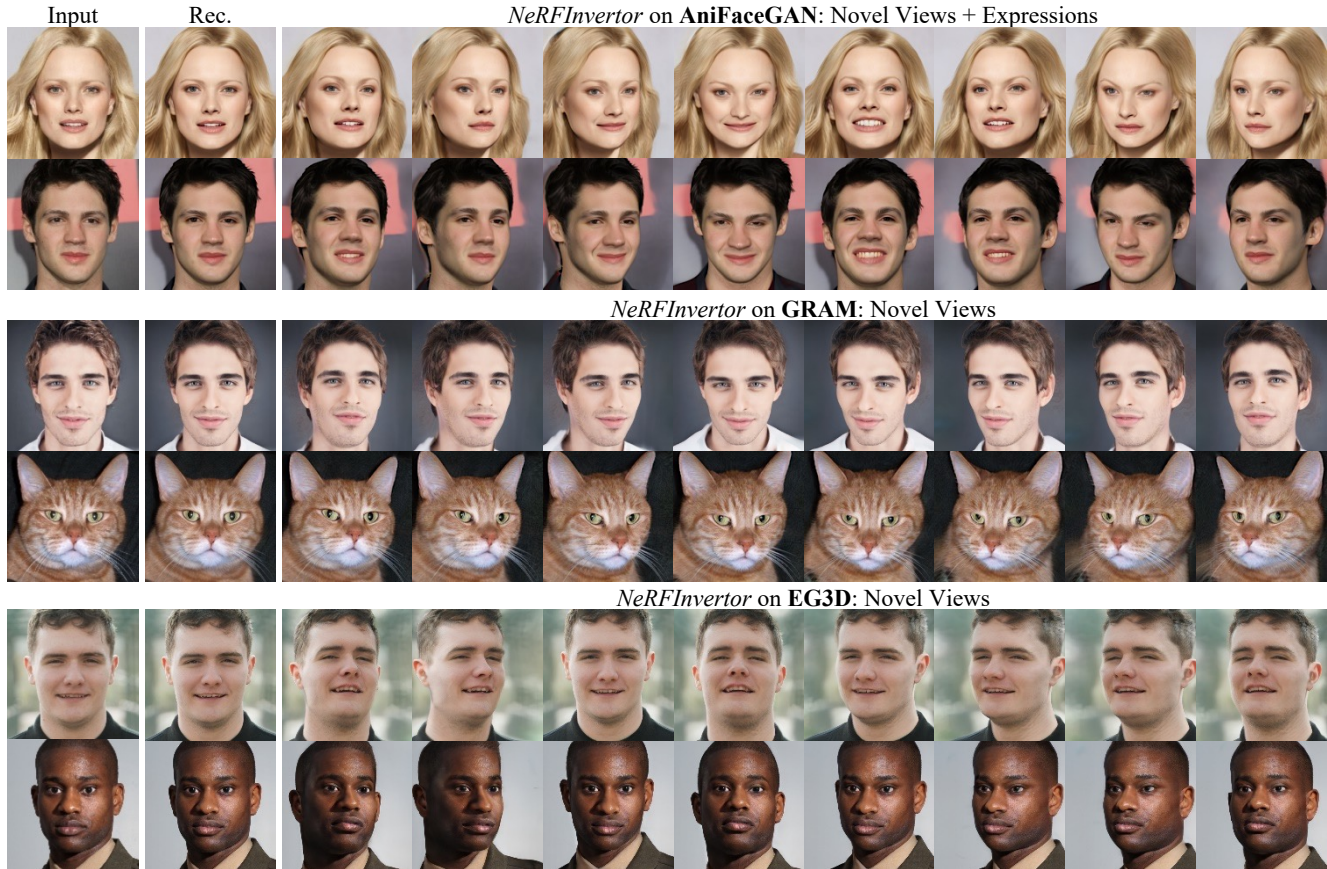
Figure 9. **Applying *NeRFInvertor* on multiple NeRF-GANs.** We show reconstruction, novel views and expressions synthesis. *NeRFInvertor* achieves high-fidelity, texture and 3D consistencies and ID-preserving synthesis across poses and expressions.

Table 1. **Quantitative comparisons on FFHQ and CelebA-HQ test set.** The **best**, and the second best scores are highlighted.

NeRF-GANs	Inversion Methods	FFHQ			CelebA-HQ		
		Rec. PSNR (↑)	Novel View FID (↓)	ID (↑)	Rec. PSNR (↑)	Novel View FID (↓)	ID (↑)
AniFace [38]	I2S [2] ($\mathcal{W}$)	15.62	58.37	0.33	16.50	42.53	0.25
	I2S [2] ($\mathcal{W}+$)	**25.44**	65.36	**0.83**	**26.86**	45.08	0.76
	PTI [29]	23.89	50.20	0.75	22.61	37.83	0.73
	NeRFInvertor	24.92	**45.89**	0.76	25.61	**34.07**	**0.77**
GRAM [12]	I2S [2] ($\mathcal{W}$)	16.74	49.65	0.39	17.57	**30.95**	0.19
	I2S [2] ($\mathcal{W}+$)	26.98	64.97	0.66	27.61	46.92	0.70
	PTI [29]	**28.90**	45.94	0.79	**29.26**	38.01	0.80
	NeRFInvertor	28.46	**43.58**	**0.80**	28.75	31.11	**0.81**

The results in Figure 8 show that Pix2NeRF is not a good candidate for real face synthesis, since it does not preserve the identity properly. HeadNeRF also shows a noticeable identity gap between the generated and input images. Furthermore, it also inherits the disadvantage of 3D parametric models that it produces some artificial visuals. Our approach achieves significantly higher fidelity and better identity preservation compared to these two methods.

Evaluation on multiple NeRF-GANs. We validate our

pipeline on a variety of NeRF-GANs, including AniFace-GAN [38], GRAM [12], and EG3D [8]. In Figure 9, we show the reconstruction, novel views, and expressions synthesis given a single input image. AniFaceGAN uses a deformable NeRF structure and is trained for a dynamic scene. We demonstrate that *NeRFInvertor* can faithfully translate a single image to a deformable NeRF representation, allowing us to generate realistic face editing with controllable poses and expressions (Figure 6 and 9). Given a fixed latent code, GRAM and EG3D generate the NeRF representation for a static scene. We also validate our method on these two methods, demonstrating that we can invert a single image to a traditional NeRF representation for static scenes.

4.3. Quantitative Evaluation

Quantitative experiments were performed on the first 150 samples from the CelebA-HQ test set and FFHQ dataset. For reconstruction results, we report PSNR in Table 1 and additional structural similarity (*i.e.* SSIM [36]) and identity similarity in the Supplementary Material. Since we do not have multi-view ground truth images, we report Frechet In-
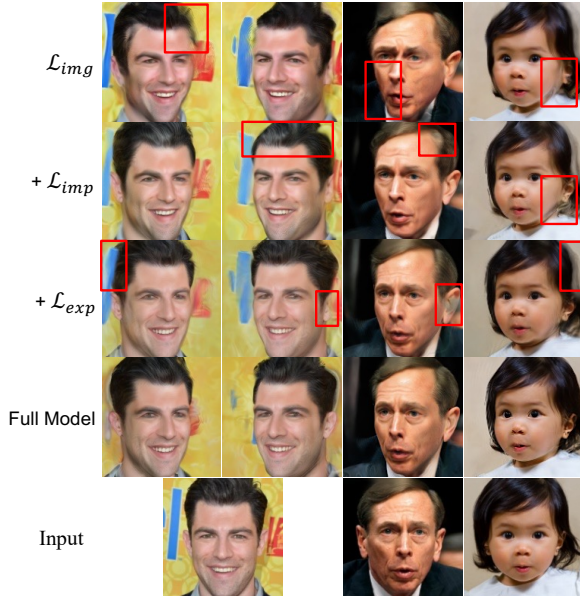
Figure 10. **Ablation study on different regularizations.**

Table 2. **Quantitative results of different regularizations.**

| | Rec. | Novel View | | Novel View+Exp. | |
	PSNR/SSIM ($\uparrow$)	FID ($\downarrow$)	ID ($\uparrow$)	FID ($\downarrow$)	ID ($\uparrow$)
$\mathcal{L}_{img}$	22.48 / 0.764	38.61	0.66	38.82	0.60
$+\mathcal{L}_{imp}$	22.61 / 0.761	37.83	0.73	36.99	0.66
$+\mathcal{L}_{exp}$	**25.64** / 0.828	34.68	0.76	34.55	0.66
Full Model	25.61 / **0.830**	**34.07**	**0.77**	**33.61**	**0.67**

ception Distance (*i.e.* FID) and identity similarity for novel view images following prior works [7,15]. The results align with our qualitative evaluation. I2S ($\mathcal{W}$) fails to preserve the identity of the input image and has poor ID scores. I2S ($\mathcal{W}+$) introduces artifacts in novel views and results in high FID scores. Compared to PTI, our method generate comparable reconstructions, but superior novel view synthesis in terms of visual quality (*i.e.* FID scores) and better identity preservation (*i.e.* ID scores). We achieve the best overall performance across all metrics. More quantitative results (*e.g.* novel view & expression evaluations, and user study) can be found in the Supplementary Material.

4.4. Ablation Study

Effectiveness of Regularization. We conduct an ablation study on the CelebA-HQ test set. Compared to existing image space losses $\mathcal{L}_{img}$, we show the effects of our proposed implicit geometrical regularization (*i.e.* $\mathcal{L}_{imp}$), explicit geometrical regularization (*i.e.* $\mathcal{L}_{exp}$), and masked regularizations (*i.e.* Full Model) in Figure 10 and Table 2. The results in Figure 10 indicate that the fine-tuned model with just image space losses $\mathcal{L}_{img}$ is prone to generating artifacts in novel-view images and inaccurate 3D geometry; the
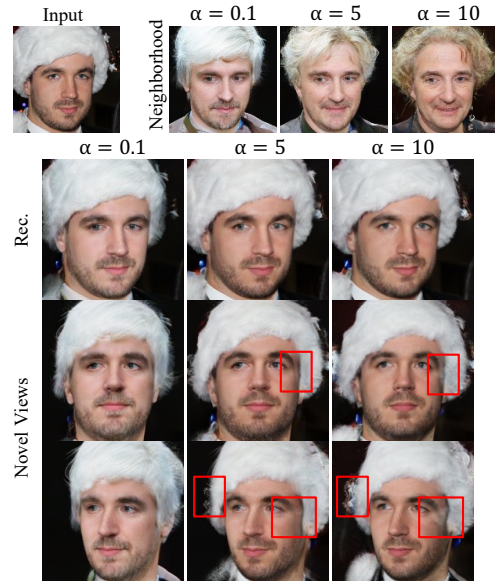

Figure 11. **Ablation study on distance of neighbor samples.**

implicit geometrical regularization $\mathcal{L}_{imp}$ helps to eliminate artifacts; the explicit geometrical regularization $\mathcal{L}_{exp}$ improves visual quality and the subject's 3D geometry; the full model with masked regularizations reduces fogging around the hair, ear, or cheek. The quantitative results in Table 2 are consistent with the qualitative results in Figure 10.

Neighborhood Selection. We empirically find out that the distance between the optimized and neighborhood latent codes affects the ID-preserving ability and geometrical constraints. As shown in Figure 11, if the selected neighborhood is too close to the optimized latent code (*i.e.* the distance $\alpha <= 1$), the model shows superior geometrical constraints but worse identity preservation. Conversely, increasing the distance too far (*i.e.*, $\alpha >= 10$) results in a model with superior ID-preserving ability but undesirable 3D shape generation. By selecting the value of α from the interval $[1, 10]$, we achieve a good compromise between regularization and supervision in image space. As a result, we chose to set the distance α to 5 in our experiments.

5. Conclusion

We introduced *NeRFInvertor* as a universal method for a single-shot inversion of real images on both static and dynamic NeRF-GAN models. We employed image space supervision to fine-tune NeRF-GANs generator for reducing identity gap, along with explicit and implicit geometrical constraints for removing artifacts from geometry and rendered images in novel views and expressions. Our experiments validate the importance of each component in our method for 3D consistent, ID-preserving, and high-fidelity animation of real face images.

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *CVPR*, pages 4432–4441, 2019. 3

[2] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan++: How to edit the embedded images? In *CVPR*, pages 8296–8305, 2020. 3, 6, 7

[3] Rameen Abdal, Peihao Zhu, Niloy J Mitra, and Peter Wonka. Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows. *ACM Transactions on Graphics (ToG)*, 40(3):1–21, 2021. 3

[4] Linchao Bao, Xiangkai Lin, Yajing Chen, Haoxian Zhang, Sheng Wang, Xuefei Zhe, Di Kang, Haozhi Huang, Xinwei Jiang, Jue Wang, et al. High-fidelity 3d digital human head creation from rgb-d selfies. *ACM Transactions on Graphics (TOG)*, 41(1):1–21, 2021. 1

[5] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5855–5864, 2021. 1

[6] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194, 1999. 1

[7] Shengqu Cai, Anton Obukhov, Dengxin Dai, and Luc Van Gool. Pix2nerf: Unsupervised conditional p-gan for single image to neural radiance fields translation. In *CVPR*, pages 3981–3990, 2022. 1, 3, 6, 8

[8] 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 *CVPR*, pages 16123–16133, 2022. 1, 2, 3, 7

[9] 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*, pages 5799–5809, 2021. 2, 6

[10] 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. 3

[11] Radek Daněček, Michael J Black, and Timo Bolkart. Emoca: Emotion driven monocular face capture and animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20311–20322, 2022. 1

[12] Yu Deng, Jiaolong Yang, Jianfeng Xiang, and Xin Tong. Gram: Generative radiance manifolds for 3d-aware image generation. In *CVPR*, pages 10673–10683, 2022. 1, 2, 7

[13] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *IEEE Computer Vision and Pattern Recognition Workshops*, 2019. 6

[14] Jiahao Geng, Tianjia Shao, Youyi Zheng, Yanlin Weng, and Kun Zhou. Warp-guided gans for single-photo facial animation. *ACM Transactions on Graphics (ToG)*, 37(6):1–12, 2018. 1

[15] Yang Hong, Bo Peng, Haiyao Xiao, Ligang Liu, and Juyong Zhang. Headnerf: A real-time nerf-based parametric head model. In *CVPR*, pages 20374–20384, 2022. 3, 6, 8

[16] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In *NeurIPS*, 2020. 3

[17] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 3

[18] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, pages 8110–8119, 2020. 3

[19] Zhong Li, Lele Chen, Celong Liu, Yu Gao, Yuanzhou Ha, Chenliang Xu, Shuxue Quan, and Yi Xu. 3d human avatar digitization from a single image. In *The 17th International Conference on Virtual-Reality Continuum and its Applications in Industry*, pages 1–8, 2019. 1

[20] 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. 3

[21] Nelson Max. Optical models for direct volume rendering. *IEEE Transactions on Visualization and Computer Graphics*, 1(2):99–108, 1995. 3

[22] Moustafa Meshry, Saksham Suri, Larry S Davis, and Abhinav Shrivastava. Learned spatial representations for few-shot talking-head synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13829–13838, 2021. 1

[23] 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. *Communications of the ACM*, 65(1):99–106, 2021. 1, 3

[24] Taesung Park, Jun-Yan Zhu, Oliver Wang, Jingwan Lu, Eli Shechtman, Alexei Efros, and Richard Zhang. Swapping autoencoder for deep image manipulation. *Advances in Neural Information Processing Systems*, 33:7198–7211, 2020. 1

[25] Swarup Pranjal and Adams Wai Kin Kong. Palmprint recognition using realistic animation aided data augmentation. In *2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–9. IEEE, 2019. 1

[26] Albert Pumarola, Antonio Agudo, Aleix M Martinez, Alberto Sanfeliu, and Francesc Moreno-Noguer. Ganimation: One-shot anatomically consistent facial animation. *International Journal of Computer Vision*, 128(3):698–713, 2020. 1

[27] Daniel Rebain, Mark Matthews, Kwang Moo Yi, Dmitry Lagun, and Andrea Tagliasacchi. Lolnerf: Learn from one look. In *CVPR*, pages 1558–1567, 2022. 1, 3

[28] 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. 3

[29] 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. 3, 6, 7

[30] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. Graf: Generative radiance fields for 3d-aware image synthesis. *NeurIPS*, 33:20154–20166, 2020. 2

[31] Ayush Tewari, Mohamed Elgharib, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zollhöfer, and Christian Theobalt. Pie: Portrait image embedding for semantic control. *ACM Transactions on Graphics (TOG)*, 39(6):1–14, 2020. 1

[32] 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. 1

[33] Liao Wang, Jiakai Zhang, Xinhang Liu, Fuqiang Zhao, Yanshun Zhang, Yingliang Zhang, Minye Wu, Jingyi Yu, and Lan Xu. Fourier plenoctrees for dynamic radiance field rendering in real-time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13524–13534, 2022. 1

[34] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10039–10049, 2021. 1

[35] Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. Towards real-world blind face restoration with generative facial prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9168–9178, 2021. 3

[36] 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. 7

[37] Xintian Wu, Qihang Zhang, Yiming Wu, Huanyu Wang, Songyuan Li, Lingyun Sun, and Xi Li. F^{3}a-gan: Facial flow for face animation with generative adversarial networks. *IEEE Transactions on Image Processing*, 30:8658–8670, 2021. 1

[38] Yue Wu, Yu Deng, Jiaolong Yang, Fangyun Wei, Qifeng Chen, and Xin Tong. Anifacegan: Animatable 3d-aware face image generation for video avatars. In *NeurIPS*, 2022. 1, 2, 7

[39] Tao Yang, Peiran Ren, Xuansong Xie, and Lei Zhang. Gan prior embedded network for blind face restoration in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 672–681, 2021. 3

[40] Dan Ye and Chiou-Shann Fuh. 3d morphable face model for face animation. *International Journal of Image and Graphics*, 20(01):2050003, 2020. 1

[41] Yuping Ye, Zhan Song, and Juan Zhao. High-fidelity 3d real-time facial animation using infrared structured light sensing system. *Computers & Graphics*, 104:46–58, 2022. 1

[42] Yu Yin, Songyao Jiang, Joseph P Robinson, and Yun Fu. Dual-attention gan for large-pose face frontalization. In *2020 15th IEEE international conference on automatic face and gesture recognition (FG 2020)*, pages 249–256. IEEE, 2020. 1

[43] Yu Yin, Joseph P Robinson, Songyao Jiang, Yue Bai, Can Qin, and Yun Fu. Superfront: From low-resolution to high-resolution frontal face synthesis. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1609–1617, 2021. 1

[44] Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. Plenoctrees for real-time rendering of neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5752–5761, 2021. 1

[45] Yin Yu, Ghasedi Kamran, Wu HsiangTao, Yang Jiaolong, Tong Xi, and Fu Yun. Expanding the latent space of stylegan for real face editing. *arXiv preprint arXiv:2204.12530*, 2022. 1

[46] Junzhe Zhang, Daxuan Ren, Zhongang Cai, Chai Kiat Yeo, Bo Dai, and Chen Change Loy. Monocular 3d object reconstruction with gan inversion. In *ECCV*, pages 673–689. Springer, 2022. 1, 3