G3FA: Geometry-guided GAN for Face Animation

Alireza Javanmardi
alireza.javanmardi@dfki.de
Alain Pagani
alain.pagani@dfki.de
Didier Stricker
didier.stricker@dfki.de

German Research Center for Artificial Intelligence (DFKI)
Kaiserslautern, Germany

Abstract

Animating human face images aims to synthesize a desired source identity in a natural-looking way mimicking a driving video's facial movements. In this context, Generative Adversarial Networks have demonstrated remarkable potential in real-time face reenactment using a single source image, yet are constrained by limited geometry consistency compared to graphic-based approaches. In this paper, we introduce Geometry-guided GAN for Face Animation (G3FA) to tackle this limitation. Our novel approach empowers the face animation model to incorporate 3D information using only 2D images, improving the image generation capabilities of the talking head synthesis model. We integrate inverse rendering techniques to extract 3D facial geometry properties, improving the feedback loop to the generator through a weighted average ensemble of discriminators. In our face reenactment model, we leverage 2D motion warping to capture motion dynamics along with orthogonal ray sampling and volume rendering techniques to produce the ultimate visual output. To evaluate the performance of our G3FA, we conducted comprehensive experiments using various evaluation protocols on VoxCeleb2 and TalkingHead benchmarks to demonstrate the effectiveness of our proposed framework compared to the state-of-the-art real-time face animation methods. Our code is available at github.com/dfki-av/G3FA.

1 Introduction

Talking head generation involves the task of re-rendering a source face image with a new pose and expression, often controlled by either the same individual or a different one. This technology finds particular application in video conferencing tools by enabling significant bandwidth reduction through keypoints transmission or by displaying alternative appearance representations. This not only reduces transmission costs but also mitigates potential biases in scenarios like job interviews while preserving a sense of presence for participants.
While this challenge was initially tackled by computer graphics researchers, new advances in deep generative modeling led to significant progress in this field [8, 29, 36, 42, 50].
Approaches centered around graphics-based head synthesis demonstrate impressive geometric capabilities as they aim to fully reconstruct and animate human heads [38].

Typically, these approaches leverage parametric models [3, 20] to reconstruct the geometry of the human's head, subsequently mapping the source identity's texture to render the animation [38]. Recent advances in this domain, particularly by the introduction of 3D Gaussian Splitting technique [16] are moving towards increased photorealism level and also rendering speed [46]. Nonetheless, this approach necessitates multiple images captured from various viewpoints of the subject's head and entails some processing time for head reconstruction.

In parallel, researchers have been harnessing deep generative models to enhance face reenactment techniques. This is typically achieved by extracting features from a desired source image and warping them based on a driving video [22, 36, 51] or utilizing a pre-trained StyleGAN2 model [15] and navigation in the latent space [43, 47]. Compared to graphics-based approaches, these methods demonstrate higher generalization capabilities, having been trained on thousands of video clips and capable of animating diverse human face images. They can start rendering instantaneously using a single shot of the source identity, producing photorealistic outcomes including fine details and accessories like glasses. However, since these methods are trained on 2D data, they are generally less robust in terms of geometry reconstruction. This leads to poor reconstruction quality in the case of head rotations far from the frontal view - typically when the driving source turns the head left or right.

To overcome these limitations, this paper introduces a novel face animation method based on Generative Adversarial Networks (GAN) [8] that incorporates 3D supervision to enhance the generator's awareness of the true distribution of the human's head while synthesizing novel head poses. To address this challenge, an inverse rendering approach is employed to extract geometry properties including depth and normal maps from RGB frames. Our method uses a weighted combination of discriminators, to inject this information into the generator as a core module for image synthesis. The proposed framework can be easily applied to most adversarial-based face reenactment models, leveraging off-the-shelf pre-trained inverse rendering models.

The contributions of this paper can be summarized as follows:

- We propose implicit 3D supervision leveraging prior human head information for face reenactment models to enhance geometry consistency without affecting inference time.
- We introduce a novel integration of 3D properties into a GAN framework for face animation tasks, without imposing significant computational overhead during training.
- We conduct extensive experiments across various scenarios on benchmark datasets, demonstrating a superior realism compared to recent state-of-the-art approaches, both quantitatively and qualitatively, particularly in the case of extreme head pose variations.

2 Related Works

2.1 Face Reenactment

Recent reconstruction-based methodologies typically employ the FLAME 3D face model [20] and rendering algorithms, such as rasterization, to achieve real-time reenactment.

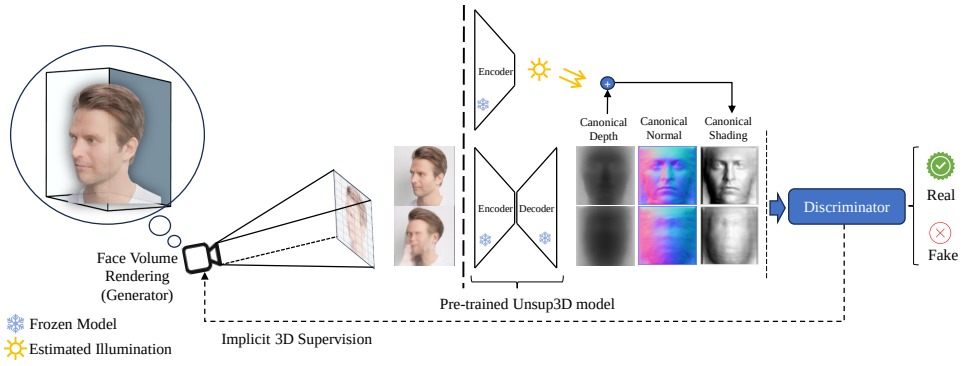

Figure 1: Implicit 3D supervision: This figure shows how an inverse rendering module can guide the generator to generate more geometry-consistent output. We utilized canonical shading here to better visualize the differences between two cases and it is not used in the model's pipeline.

These approaches involve reconstructing the avatar's head and subsequently animating it by manipulating the morphable model's parameters. However, these techniques exhibit limitations in rendering fine details and entail significant training time when reconstructing heads from 2D-captured data. [9, 38]. Generative models, including GANs and Diffusion models [11], have demonstrated significant success in this field. In the real-time domain, GAN-based techniques initiate by estimating 2D or 3D keypoints and deriving the optical flow of these points between the source and driving frames. Following this, they deform the features of the source image and utilize a generator model to produce the animated output [12, 36, 42, 49]. Expanding on the concept of GAN inversion [34], another research direction, aims to animate face images based on latent space navigation of StyleGAN model [2, 43, 47]. On the other hand, Diffusion models provide face animation approaches that can even work with other modalities like audio information [33, 39, 45]. These models are typically used for offline tasks where rendering time is not crucial. Additionally, some researchers have explored the application of neural rendering techniques within an adversarial framework, capitalizing on their ability to generate photorealistic results [30, 48, 51].

2.2 Generative Models

Generative adversarial networks (GANs) [8] serve as the primary engine for generating high-fidelity samples in the domain of talking head synthesis. Researchers have endeavored to expedite convergence and enhance the stability of GAN models by optimizing their architectures [13, 25] or loss functions [1, 10]. In this regard, certain works have explored the use of two discriminators, whereby forward and backward KL divergence is computed, resulting in improved performance compared to the standard framework [28]. Additionally, the combination of pre-trained discriminators with varying architectures has been employed to enhance the quality of the generated images [19, 27]. Some studies have also tried to guide the generator through more effective supervision by incorporating geometry information during the model's training process [12, 35, 40].

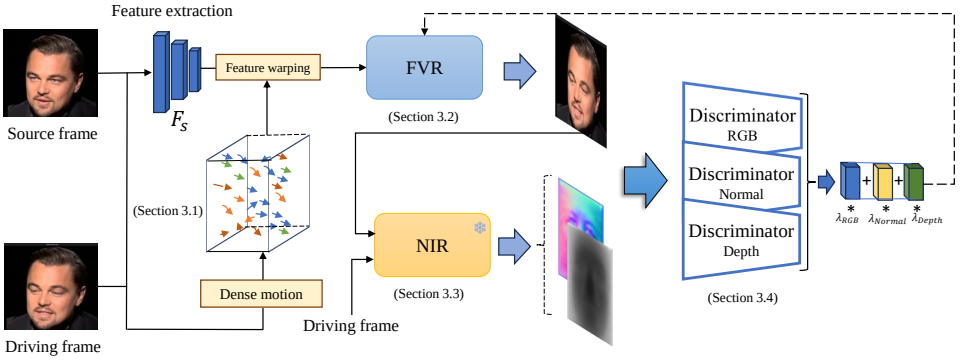

Figure 2: Face animation pipeline: Capturing facial expression and pose based on keypoints, followed by an implicit 3D supervision using inverse rendering and an ensemble of discriminators. NIR stands for Neural Inverse Rendering which is a pre-trained model and FVR is our Face Volume Rendering module.

3 Method

The global architecture of our method is shown in Fig. 2. We leverage the face volume rendering (FVR) technique [23] in conjunction with an adversarial training procedure. To empower the model in generating photorealistic outcomes mainly based on RGB training data, we integrated a neural inverse rendering module and two additional discriminators to process the extracted 3D attributes. Our objective is to synthesize images with the appearance of a given source image I_S while incorporating the pose and facial expressions from a set of driving frames $\{I_{D_1}, I_{D_2}, ..., I_{D_n}\}$, which could be sourced from videos or live webcam streams. In the subsequent sections, we provide comprehensive information on each individual module.

3.1 2D Motion Estimation

We start by extracting features from the source image f_s using a dedicated self-supervised feature extraction module F_s. To capture the facial dynamics of the driving frames, we employ a motion estimation module. Within this module, we detect and extract a set of $k = 15$ 2D landmarks from both the source and driving frames, leading to pairs $\{q_{s,k}, q_{d,k}\} \in \mathbb{R}^2$. Subsequently, to better emulate facial changes, we consider the neighboring points as a 2×2 Jacobian matrix $\{J_{s,k}, J_{d,k}\} \in \mathbb{R}^{2\times 2}$ for each landmark. Using first-order Taylor expansion, we can obtain an affine approximation of motion field modeled by $\tau_{S\leftarrow D}$ based on backward optical flow as in [36]:

$$\tau_{S\leftarrow D,k}(z) \approx q_{s,k} + J_{s,k}{J_{d,k}}^{-1}(z - q_{d,k}). \quad (1)$$

Although some studies have advocated for the utilization of 3D keypoints and the elimination of the Jacobian matrix [17], we did not observe significant differences. Moreover, adopting 3D keypoints introduces a substantial number of parameters to the model. Leveraging a dense motion field for each keypoint [36], we combine them with a weight factor as a mask, ranging from 0 to k $\{M_0, M_0, ..., M_K\}$, where the initial one represents the background. The dense motion estimation module also produces an occlusion map O to generate the occluded

parts using Hadamard product while synthesizing novel poses:

$$\hat{\tau}_{S\leftarrow D}(z) = M_0 z + \sum_{k=1}^{K} M_k \tau_{S\leftarrow D,k}(z), \tag{2}$$

$$F_w = O \odot f_w(f_s, \hat{\tau}_{S\leftarrow D,k}). \tag{3}$$

Finally, in this part of the framework, the extracted features from the source image are warped using information obtained from the dense motion estimation module. This allows the warped features to be utilized by the face volume rendering component.

3.2 Rendering

To render the final animated face results, we use the face volume rendering architecture proposed by [51]. This architecture, compared to other approaches utilizing a single U-Net model [31] as the generator, offers a reduced parameter count and the ability to generate photorealistic outcomes through orthogonal ray sampling and volume rendering. It is important to note that our proposed framework seamlessly integrates with other architectures that follow an adversarial training procedure.

To employ the face volume rendering module, which consists of several parts, we initially feed the warped features F_w obtained from the previous stage into two networks: the 3D shape extractor ϕ_σ and the color extractor ϕ_{color} similar to [51]. These networks serve to both separate these two types of information from the features and elevate them to a higher-dimensional space. In order to calculate the sampled color and density, we leverage the proposed orthogonal ray-sampling module, which employs f_θ; a multi-layer perceptron (MLP) to estimate the color field F_{color} and voxel probability F_σ in an adaptive manner [51], instead of adhering to the ray-sampling strategy proposed in NeRF [24]:

$$F_\sigma = \phi_\sigma(F_w), \tag{4}$$

$$F_{color} = \phi_{color}(F_w), \tag{5}$$

$$p_\sigma, p_{color} = f_\theta(F_\sigma, F_{color}). \tag{6}$$

We confine the camera pose to the frontal view and employ an adaptive ray sampling method, which significantly enhances the rendering speed. In our framework, we did not employ a 3D face reconstruction component to supervise the rendering model. However, by employing geometry guidance based on 3D geometric properties and an ensemble of discriminators, we can provide implicit supervision to the rendering model. To aggregate the color p_{color} and the density p_σ obtained from the adaptive ray sampling segment, we utilize a volume rendering algorithm, albeit distinct from [23]:

$$F_{r,i} = \sum_{j=1}^{D} \tau_j (1 - exp(-p_{\sigma,i,j})) p_{color,i,j}, \tag{7}$$

$$\tau_j = exp(\sum_{k=1}^{j-1} -p_{\sigma,i,k}), \tag{8}$$

where τ_j is the transmittance term. This algorithm enables the derivation of the final RGB value. Furthermore, we have observed that concatenating the extracted features from the source image f_s with the output of the volume rendering module F_r and passing them through a shallow decoder with SPADE layers improves the fidelity of the results, as suggested in [51]. A detailed comparison of computational costs is provided in the supplementary material.

3.3 Neural Inverse Rendering

As depicted in Fig. 1, we utilized a well-established Unsup3D model [44] to obtain 3D geometric properties, such as canonical depth and normal map, from a single RGB image. These properties are subsequently employed to enhance the quality of the rendered results. By regarding the RGB image I as a function $\Omega \rightarrow \mathbb{R}^3$ that maps coordinate $\{0,...,W-1\} \times \{0,...,H-1\}$ in a 2D space to tensor in $\mathbb{R}^{3\times W\times H}$, and under the assumption that the image is centered and relatively symmetric, we adopt an autoencoder architecture model, as in [44], to obtain 3D properties, including depth map d, global light direction l, viewpoint w, and albedo image a, using an analysis-by-synthesis approach making $\hat{I} \approx I$:

$$\hat{I} = \prod(\psi(a,d,l),d,w), \tag{9}$$

Here, ψ synthesizes a version of the object from a canonical viewpoint, then the reprojection function $\prod$ tries to simulate the viewpoint effect to the model using shaded canonical image $\psi(a,d,l)$ and also depth map d. However, in our framework, we utilize a pre-trained Unsup3D model trained on the CelebA dataset [21], extracting only depth information, from which we subsequently derive the normal map.

3.4 Ensemble of Discriminators

To incorporate the extracted 3D information, obtained through prior knowledge of the human face, into our talking head synthesis pipeline, we introduce two additional discriminators. For this purpose, we feed each discriminator with the RGB image, canonical depth, and normal map, respectively. All of these discriminators employ a multi-scale architecture [41] with spectral normalization [26]. To ensure the convergence of the model is unaffected, we assign different weights λ_i to each discriminator [18] and aggregate them based on the min-max formulation to obtain the final result:

$$\mathcal{L}_{\text{GAN}}(G,D_{total}) = \mathbb{E}_{\mathbf{x}\sim p_{\text{real}}(\mathbf{x})}[\log D_{total}(\mathbf{x})] + \mathbb{E}_{\mathbf{x}'\sim p_{\mathbf{fake}}(\mathbf{x}')}[\log(1-D_{total}(\mathbf{x}'))], \tag{10}$$

$$D_{total} = \sum_{i\in S} \lambda_i D_i(x_i), \tag{11}$$

where $S = \{\text{RGB, depth, normal}\}$ and we assume $\sum_i \lambda_i = 1$. By incorporating the 3D face geometry as additional supervision, the generative model can better estimate the true data distribution. The photorealistic results generated by the Unsup3D model produce meaningful depth and normals, thereby reducing the adversarial loss, rendering them close to real images. This integration of 3D information into a 2D GAN pipeline, in contrast to [35] and [12], preserves the inference time efficiency, and when combined with more powerful inverse rendering techniques, can yield even more effective outcomes.

4 Experiment

The G3FA framework is devised to operate in a fully self-supervised manner during the training process. It randomly extracts two frames from a video: the first frame serves as the source, while the second frame acts as the driving frame. The framework then animates the source image based on the changes exhibited by the driving frame, while simultaneously incorporating the ground-truth animated frame for comparison. Adversarial and perceptual losses [14] are employed to ensure the photorealism of the generated samples.

Table 1: Comparison with prior works on same-identity reconstruction on VoxCeleb2 [5]. Bold values indicate the best performance, while underlined values represent the second-best.

Method	L1 ↓	LPIPS ↓	PSNR ↑	SSIM ↑	FID ↓	AKD ↓
FOMM *(NeurIPS 2019)*	12.31	0.109	23.52	0.71	24.59	1.89
Face vid2vid *(CVPR 2021)*	<u>11.13</u>	0.125	24.15	0.84	21.78	1.72
DaGAN *(CVPR 2022)*	11.22	0.117	**25.64**	<u>0.88</u>	22.83	<u>0.91</u>
FNeVR *(NeurIPS 2022)*	11.16	<u>0.094</u>	24.18	0.77	21.11	0.95
LIA *(ICLR 2022)*	12.02	0.106	<u>25.57</u>	0.84	**16.47**	1.12
HyperReenact *(ICCV 2023)*	13.42	0.111	22.51	0.69	28.87	1.28
G3FA (ours)	**10.87**	**0.081**	24.51	**0.91**	<u>18.79</u>	**0.80**

Table 2: Quantitative comparison of cross-identity reenactment on VoxCeleb2[5] and TK[42]. Bold values indicate the best performance, while underlined values represent the second-best.

	VoxCeleb2		TK	
Method	**FID ↓**	**CSIM ↑**	**FID ↓**	**CSIM ↑**
FOMM	142.18	0.5219	130.78	0.5402
Face vid2vid	139.74	0.5971	121.44	0.6114
DaGAN	130.77	0.6021	<u>120.94</u>	0.6264
FNeVR	132.36	0.5408	122.47	0.6021
LIA	**122.26**	<u>0.6078</u>	**118.59**	<u>0.6338</u>
HyperReenact	152.94	0.5144	147.63	0.4923
G3FA (ours)	<u>127.12</u>	**0.6274**	122.83	**0.6455**

The adversarial loss encompasses the combined predictions of all discriminators, while the perceptual loss [14] leverages a pre-trained VGG-19 model [37] trained on ImageNet [32] to extract features. Additionally, to ensure the consistency of detected keypoints in the face, the framework incorporates an equivariance loss [36]. The keypoint extraction module extracts keypoints and their jacobians, all of which are attained in a fully self-supervised manner:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_P(x,x') + \mathcal{L}_{\text{GAN}}(\{x_i,x_i'\}_{i\in S}) + \mathcal{L}_E(\{x_{d,k}\}) \tag{12}$$

4.1 Implementation Details

Datasets: Our experiments were conducted using two well-known datasets, VoxCeleb2 [5] and TK (TalkingHead-1KH) [42]. VoxCeleb2 [5] comprises over 1 million videos encompassing approximately 6K distinct identities, while TK offers about 180K high-quality samples. As part of the preprocessing stage, we traced and cropped face images from the videos, resizing them to a standardized dimension of 256×256, as in [36]. This preprocessing methodology holds particular significance for our neural inverse rendering module, which was also trained on centered images, ensuring the acquisition of meaningful geometric properties.

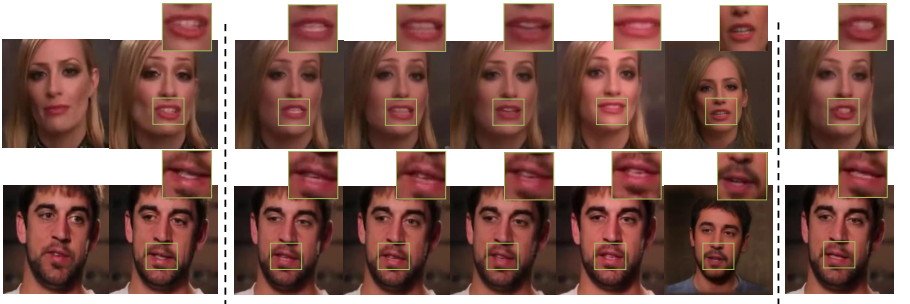

Figure 3: Same-identity reconstruction: Our method exhibits superior performance in terms of both photorealism image generation and precise synthesis of fine details on VoxCeleb2 [5].

4.2 Training details

Our G3FA model was trained on the pre-defined training set of VoxCeleb2 [5]. Moreover, randomly selected 90% of the videos from the TK dataset were included for training. For network optimization, we employed the Adam optimizer [17] with the same parameters across all modules $\eta = 2 \times 10^{-4}$, $\beta_1 = 0.5$ and $\beta_2 = 0.9$. After conducting several experiments, we determined the weights of the three discriminators as follows: $\lambda_{RGB} = 50\%$, and $\lambda_{depth} = \lambda_{normal} = 25\%$ for depth and normal. These weights remained fixed throughout the training process. Ablation studies investigating various values of λ are detailed in the appendix.

4.3 Evaluation Metrics

To gauge the reconstruction quality, we employed the L1 loss, Learned Perceptual Image Patch Similarity (LPIPS), Structural Similarity Index Metric (SSIM), Peak Signal-to-Noise Ratio (PSNR), Average Keypoint Distance (AKD) [7] using MTCNN [52] and to evaluate the identity preservation, we utilized cosine similarity metric (CSIM). Additionally, to compare the distributions and assess the quality and diversity of real and generated samples, we utilized the Fréchet Inception Distance (FID) metric.

4.4 Comparison with State-of-the-Art

We conducted a comparative analysis of G3FA against six prominent models: First Order Motion Model (FOMM) [36], One-Shot Free-View Neural Talking-Head Synthesis for Video Conferencing (face vid2vid) [42], Depth-aware Generative Adversarial Network for Talking Head Video Generation (DaGAN) [12], Neural Volume Rendering for Face Animation (FNeVR) [51], Latent Image Animator (LIA) [43] and One-Shot Reenactment via Jointly Learning to Refine and Retarget Faces (HyperReenact) [4] on the VoxCeleb2 dataset [5]. We used the official implementations of the models, except for face vid2vid, where we relied on an unofficial yet well-known implementation.

Figure 4: Cross-identity reenactment: demonstrating our method's superiority in geometry reconstruction and photorealistic face generation through a Qualitative Comparison on the TK Dataset[42].

4.5 Same-identity Reconstruction

Same-identity reconstruction involves reconstructing the face of a single individual using both the source image and driving frames. In Table 1, we show the results of our method compared to recent state-of-the-art approaches. As demonstrated, our framework exhibits substantial improvements across the majority of metrics, yielding highly refined samples. Notably, even without incorporating the face reconstruction component, our model, when coupled with the FNeVR architecture-based talking head model, achieves superior results. Fig. 3 shows a qualitative evaluation of our method for this scenario on two examples. Our method produces the most realistic images, providing the closest resemblance to the driving faces.

4.6 Cross-identity Reenactment

In this section, we compare the performance of our method with state-of-the-art approaches on both the VoxCeleb2 [5] and TK [42] datasets. In this scenario, we evaluate the generalization of our face animation model by utilizing two distinct identities for the source and driving frames. As indicated in the Table 2, our method attains the highest CSIM value compared to all other approaches which shows the better identity preservation of our work. In addition, our G3FA achieves the highest FID score among the majority, closely approaching LIA. This work leverages a pretrained StyleGAN2 model, enhancing FID values. However, qualitative evaluation reveals shortcomings in accurately mimicking head rotation, a factor not assessed in FID calculation due to the predominance of frontal face samples. Notably, as depicted in Fig. 4, our model consistently produces geometrically consistent results in extreme head poses compared to other approaches.

5 Conclusion

We presented a novel integration of 3D information derived from neural inverse rendering into an adversarial learning framework, employing an ensemble of discriminators for one shot talking head synthesis. Our framework takes advantage of the intrinsic characteristics of 3D geometry to enhance the synthesis process outperforming current state-of-the-art face reenactment models. Importantly, our method can easily be integrated with existing face animation architectures based on Generative Adversarial Networks (GANs), without requiring any modifications to their fundamental structure. Leveraging off-the-shelf geometry extraction modules and discriminators to provide feedback to the generator preserves inference speed while enhancing the quality of generated samples.

6 Acknowledgments

This research has been partially funded by the EU project CORTEX² (GA: Nr 101070192).

References

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

[2] Yunpeng Bai, Yanbo Fan, Xuan Wang, Yong Zhang, Jingxiang Sun, Chun Yuan, and Ying Shan. High-fidelity facial avatar reconstruction from monocular video with generative priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4541–4551, 2023.

[3] 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.

[4] Stella Bounareli, Christos Tzelepis, Vasileios Argyriou, Ioannis Patras, and Georgios Tzimiropoulos. Hyperreenact: one-shot reenactment via jointly learning to refine and retarget faces. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7149–7159, 2023.

[5] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. VoxCeleb2: Deep Speaker Recognition. In *Proc. Interspeech 2018*, pages 1086–1090, 2018. doi: 10.21437/Interspeech.2018-1929.

[6] Nikita Drobyshev, Jenya Chelishev, Taras Khakhulin, Aleksei Ivakhnenko, Victor Lempitsky, and Egor Zakharov. Megaportraits: One-shot megapixel neural head avatars. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 2663–2671, 2022.

[7] Michael Gashler. Waffles: A machine learning toolkit. *Journal of Machine Learning Research*, 12(69):2383–2387, 2011.

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[9] Philip-William Grassal, Malte Prinzler, Titus Leistner, Carsten Rother, Matthias Nießner, and Justus Thies. Neural head avatars from monocular rgb videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18653–18664, June 2022.

[10] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. *Advances in neural information processing systems*, 30, 2017.

[11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[12] Fa-Ting Hong, Longhao Zhang, Li Shen, and Dan Xu. Depth-aware generative adversarial network for talking head video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3397–3406, 2022.

[13] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[14] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution, 2016.

[15] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8110–8119, 2020.

[16] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4):1–14, 2023.

[17] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

[18] J. Kittler, M. Hatef, R.P.W. Duin, and J. Matas. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239, 1998. doi: 10.1109/34.667881.

[19] Nupur Kumari, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Ensembling off-the-shelf models for gan training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10651–10662, 2022.

[20] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph.*, 36(6):194–1, 2017.

[21] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild, 2015.

[22] Arun Mallya, Ting-Chun Wang, and Ming-Yu Liu. Implicit warping for animation with image sets. *Advances in Neural Information Processing Systems*, 35:22438–22450, 2022.

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

[24] 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.

[25] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[26] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018.

[27] Gonçalo Mordido, Haojin Yang, and Christoph Meinel. Dropout-gan: Learning from a dynamic ensemble of discriminators. *arXiv preprint arXiv:1807.11346*, 2018.

[28] Tu Nguyen, Trung Le, Hung Vu, and Dinh Phung. Dual discriminator generative adversarial nets. *Advances in Neural Information Processing Systems*, 30, 2017.

[29] Youxin Pang, Yong Zhang, Weize Quan, Yanbo Fan, Xiaodong Cun, Ying Shan, and Dong-ming Yan. Dpe: Disentanglement of pose and expression for general video portrait editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 427–436, 2023.

[30] Yurui Ren, Ge Li, Yuanqi Chen, Thomas H Li, and Shan Liu. Pirenderer: Controllable portrait image generation via semantic neural rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13759–13768, 2021.

[31] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[32] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge, 2015.

[33] Shuai Shen, Wenliang Zhao, Zibin Meng, Wanhua Li, Zheng Zhu, Jie Zhou, and Jiwen Lu. Difftalk: Crafting diffusion models for generalized audio-driven portraits animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1982–1991, 2023.

[34] Yujun Shen and Bolei Zhou. Closed-form factorization of latent semantics in gans. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1532–1540, 2021.

[35] Zifan Shi, Yinghao Xu, Yujun Shen, Deli Zhao, Qifeng Chen, and Dit-Yan Yeung. Improving 3d-aware image synthesis with a geometry-aware discriminator. *Advances in Neural Information Processing Systems*, 35:7921–7932, 2022.

[36] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. *Advances in Neural Information Processing Systems*, 32, 2019.

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

[38] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Niessner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2016.

[39] Linrui Tian, Qi Wang, Bang Zhang, and Liefeng Bo. Emo: Emote portrait alive-generating expressive portrait videos with audio2video diffusion model under weak conditions. *arXiv preprint arXiv:2402.17485*, 2024.

[40] Qiulin Wang, Lu Zhang, and Bo Li. Safa: Structure aware face animation. In *2021 International Conference on 3D Vision*, pages 679–688. IEEE, 2021.

[41] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2018.

[42] 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.

[43] Yaohui Wang, Di Yang, Francois Bremond, and Antitza Dantcheva. Latent image animator: Learning to animate images via latent space navigation. In *International Conference on Learning Representations*, 2022.

[44] Shangzhe Wu, Christian Rupprecht, and Andrea Vedaldi. Unsupervised learning of probably symmetric deformable 3d objects from images in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1–10, 2020.

[45] Sicheng Xu, Guojun Chen, Yu-Xiao Guo, Jiaolong Yang, Chong Li, Zhenyu Zang, Yizhong Zhang, Xin Tong, and Baining Guo. Vasa-1: Lifelike audio-driven talking faces generated in real time. *arXiv preprint arXiv:2404.10667*, 2024.

[46] Yuelang Xu, Benwang Chen, Zhe Li, Hongwen Zhang, Lizhen Wang, Zerong Zheng, and Yebin Liu. Gaussian head avatar: Ultra high-fidelity head avatar via dynamic gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[47] Fei Yin, Yong Zhang, Xiaodong Cun, Mingdeng Cao, Yanbo Fan, Xuan Wang, Qingyan Bai, Baoyuan Wu, Jue Wang, and Yujiu Yang. Styleheat: One-shot high-resolution editable talking face generation via pre-trained stylegan. In *European conference on computer vision*, pages 85–101. Springer, 2022.

[48] Yu Yin, Kamran Ghasedi, HsiangTao Wu, Jiaolong Yang, Xin Tong, and Yun Fu. Nerfinvertor: High fidelity nerf-gan inversion for single-shot real image animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8539–8548, 2023.

[49] Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. Few-shot adversarial learning of realistic neural talking head models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9459–9468, 2019.

[50] Egor Zakharov, Aleksei Ivakhnenko, Aliaksandra Shysheya, and Victor Lempitsky. Fast bi-layer neural synthesis of one-shot realistic head avatars. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, pages 524–540. Springer, 2020.

[51] Bohan Zeng, Boyu Liu, Hong Li, Xuhui Liu, Jianzhuang Liu, Dapeng Chen, Wei Peng, and Baochang Zhang. Fnevr: Neural volume rendering for face animation. *Advances in Neural Information Processing Systems*, 35:22451–22462, 2022.

[52] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23(10):1499–1503, 2016.