High-Fidelity Talking Portrait Synthesis with Personalized 3D Generative Prior

Anonymous ICCV submission

Paper ID *****

Abstract

*Recent audio-driven talking head synthesis methods optimize neural radiance fields (NeRF) on monocular videos but struggle with incomplete face geometry reconstruction due to limited 3D information. We introduce **Talk3D**, a novel framework that reconstructs plausible facial geometries by adopting pre-trained 3D-aware generative priors through generator personalization. Our audio-guided attention U-Net architecture predicts dynamic face variations in NeRF space driven by input audio, with conditioning tokens that disentangle scene variations unrelated to audio. Talk3D excels at generating realistic frames under extreme head poses, demonstrating superior performance compared to existing methods in extensive quantitative and qualitative evaluations.*

1. Introduction

Audio-driven talking portrait synthesis [17, 28, 29] aims to synthesize facial videos with lip movements synchronized to input audio. This task poses challenges including accurately capturing phonemes, generating realistic facial dynamics, and achieving high-fidelity synthesis. Early approaches utilized 2D generative models [17, 29] but exhibited limitations in head pose control. To address this, some works employed explicit structural priors using 2D landmarks or 3D facial models [19, 21, 28], but these often struggle with consistent pose control and coherent deformations due to errors in intermediate representations.

Recent studies have utilized neural radiance fields (NeRF) [16] for talking head generation, leveraging NeRF's multi-view consistency and pose controllability. These approaches either directly condition NeRF on audio features [10, 14, 15, 20] or use intermediate representations [23, 24]. However, constructing dynamic facial NeRF from monocular videos remains challenging due to limited head poses and 3D information, resulting in poor visual quality from unseen viewpoints and implausible depth artifacts at extreme poses (see Fig. 1).

To address these challenges, we introduce **Talk3D**, a framework for synthesizing plausible talking portraits at unseen viewpoints by leveraging 3D-aware generative adversarial networks (3D-GANs) [1, 4]. Our method adopts a personalization strategy [9] to fine-tune the generator and obtain personalized triplanes. Our U-Net architecture predicts triplane offsets (deltaplanes) modulated by audio features, representing precise lip movements in NeRF space. Additionally, our attention-based module with conditioning tokens disentangles subtle variations (torso, background, eye movements) from lip movement, enhancing reconstruction quality and lip-sync accuracy.

ER-NeRF [14] **Talk3D (Ours)**

Figure 1. **Comparison of generated talking heads by NeRF-based ER-NeRF [14], and Talk3D rendered at extreme camera poses.** Talk3D robustly generates high-fidelity realistic geometry of talking heads at unseen poses during training.

2. Preliminary

2.1. NeRF-based 3D-aware GANs

While conventional NeRF [16] aims to be optimized for a single static scene, NeRF-based 3D-aware GANs [1, 4] achieved explicitly pose-controllable image generation by conditioning their NeRF space with random-sampled latent code $\mathbf{w}$. Among these works, EG3D [4] demonstrates its superior performance using three stages. First, EG3D employs a plane generator $\mathcal{G}(\cdot;\theta_{\mathcal{G}})$ parametrized by $\theta_{\mathcal{G}}$ that efficiently synthesizes low resolution feature plane $\mathbf{P}$ such that $\mathbf{P} = \mathcal{G}(\mathbf{w};\theta_{\mathcal{G}})$. This feature plane is reshaped into three orthogonal feature planes, $\{\mathbf{P}^{\text{xy}}, \mathbf{P}^{\text{yz}}, \mathbf{P}^{\text{zx}}\}$. EG3D then utilizes an MLP that takes features aggregated from the

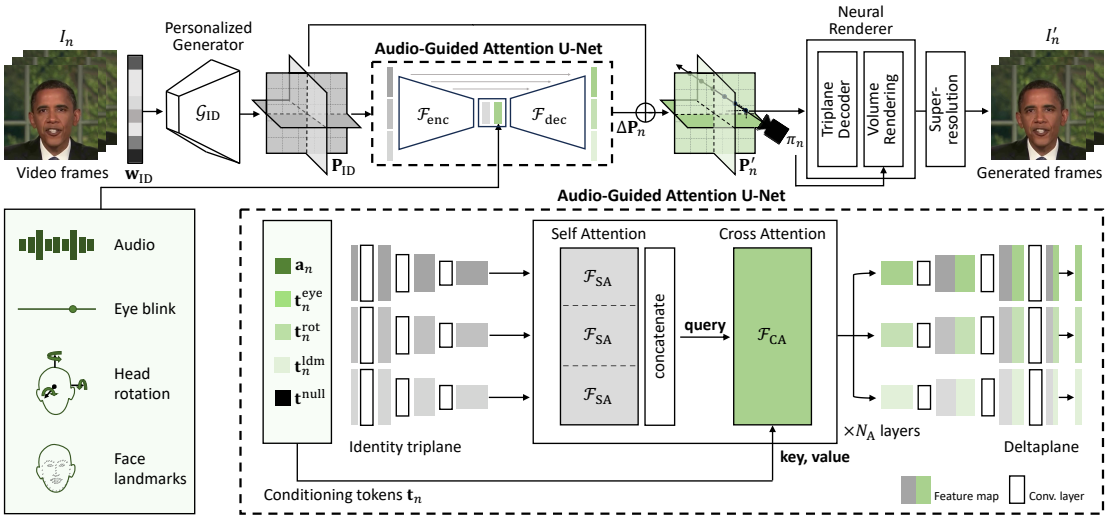

Figure 2. **Overview of our Talk3D framework and audio-guided attention U-Net.**

orthogonal planes and maps it to volume density σ and feature $\mathbf{f}$. This feature field is rendered to a low resolution 2D feature map $\mathbf{F}$, Finally, the produced feature map $\mathbf{F}$ undergoes processing in a 2D super-resolution module comprised of several convolutional layers to generate the final image I. We denote $\mathcal{R}(\cdot;\theta_{\mathcal{R}})$ as this sequential process involving volume rendering and super-resolution module. Given $\theta_{\mathcal{R}}$ as the learnable parameters, the final synthesized image can be formulated as: $I = \mathcal{R}(\mathbf{P}, \pi; \theta_{\mathcal{R}})$.

3. Methodology

3.1. Problem Formulation and Overview

We describe the main components of **Talk3D**, which enables pose-controllable audio-driven talking portrait synthesis. Given N video frames for a specific identity, $\mathcal{V} = \{I_n\}$, our model takes n-th frame image I_n with corresponding audio feature $\mathbf{a}_n$ and camera parameter π_n. We formulate the audio-driven rendering process as:

$$\begin{aligned} \mathbf{P} &= \mathcal{G}(\mathbf{w}; \theta_{\mathcal{G}}), \\ I'_n &= \mathcal{R}(\mathbf{P}, \pi_n, \mathbf{a}_n; \theta_{\mathcal{R}}). \end{aligned} \tag{1}$$

To attain the rendered portrait image I'_n that best replicates the lip movement of the frame I_n, our model aims to find the optimal EG3D parameters denoted as $\{\theta^*_{\mathcal{G}}, \theta^*_{\mathcal{R}}\}$, and the optimal triplane $\mathbf{P}^*$ which encapsulates the appropriate scene encodings. At inference, given new audio $\mathbf{a}_n^{\text{novel}}$, we reformulate (1) as:

$$\begin{aligned} \mathbf{P}_{\text{ID}} &= \mathcal{G}(\mathbf{w}_{\text{ID}}; \theta^*_{\mathcal{G}}), \\ I_n^{\text{novel}} &= \mathcal{R}(\mathbf{P}_{\text{ID}}, \pi_n, \mathbf{a}_n^{\text{novel}}; \theta^*_{\mathcal{R}}), \end{aligned} \tag{2}$$

where $\mathbf{w}_{\text{ID}}$ denotes an identity latent code that corresponds to a specific person's facial identity. Then, such a personalized generator generates $\mathbf{P}_{\text{ID}}$, namely identity triplane.

To formulate this, we first train a personalized generator that gives $\mathbf{w}_{\text{ID}}$ and $\{\theta^*_{\mathcal{G}}, \theta^*_{\mathcal{R}}\}$. In the renderer $\mathcal{R}(\cdot)$, to condition $\mathbf{a}_n^{\text{novel}}$, we propose a deltaplane generator that generates a new plane $\Delta\mathbf{P}_n^{\text{novel}}$ from $\mathbf{a}_n^{\text{novel}}$ to manipulate the identity plane $\mathbf{P}_{\text{ID}}$ such that $\mathbf{P}' = \mathbf{P}_{\text{ID}} + \Delta\mathbf{P}_n^{\text{novel}}$. Then our final renderer is defined as follows:

$$I_n^{\text{novel}} = \mathcal{R}(\mathbf{P}', \pi_n; \theta^*_{\mathcal{R}}) = \mathcal{R}(\mathbf{P}_{\text{ID}} + \Delta\mathbf{P}_n^{\text{novel}}, \pi_n; \theta^*_{\mathcal{R}}). \tag{3}$$

3.2. Personalized Generator

3D-aware GANs trained on extensive facial datasets like FFHQ [13] generate diverse identities but may not be optimal for single-person monocular videos. We adopt VIVE3D [9], a fine-tuning strategy for single-identity generation. This strategy uses pivotal tuning [18], inverting selected frames to find optimal latent vectors in $\mathbf{w}$-space, then jointly fine-tuning generator $\mathcal{G}$ and $\mathcal{R}$. They optimize latent vectors $\mathbf{w}_{\text{ID}} + \mathbf{o}_m$ for M frames I_m, where $\mathbf{o}_m$ captures local variants like expressions. After fine-tuning on I_m with fixed optimal latents, they conduct frame-by-frame inversion on fine-tuned generator $\mathcal{G}_{\text{ID}}$ for all N frames, predicting offsets $\mathbf{o}_n$ and camera parameters π_n.

3.3. Audio-Guided Attention U-Net

Through the inversion process, we obtain personalized generator $\mathcal{G}_{\text{ID}}$, global identity $\mathbf{w}_{\text{ID}}$, and camera parameters π_n. We derive identity triplane $\mathbf{P}_{\text{ID}} = \mathcal{G}_{\text{ID}}(\mathbf{w}_{\text{ID}}; \theta^*_{\mathcal{G}})$ for our training framework. Our goal is modulating the generator with audio features for NeRF space conditioning. While predicting latent vectors within the generator's manifold [2] is straightforward, we found this may not be optimal. Alternatively, we introduce a training method that focuses on the direct prediction of a triplane grid rather than the $\mathbf{w}$-space

latent vector. In the following, we explain how to manipulate the triplane with given condition $\mathbf{a}_n$.

As depicted in Fig. 2, the U-Net-based architecture $\mathcal{F}$ is employed, where identity triplane serves as input, yielding an offset triplane grid $\Delta\mathbf{P}_n$ such that: $\Delta\mathbf{P}_n = \mathcal{F}(\mathbf{P}_{\mathrm{ID}}, \mathbf{a}_n; \theta)$, where $\mathbf{a}$ denotes a given audio feature. This offset grid $\Delta\mathbf{P}_n$, which we call deltaplane is further combined with $\mathbf{P}_{\mathrm{ID}}$ through summation. This training strategy offers several distinct advantages compared to manipulation in the GAN latent space. Editing in the GAN latent space cannot represent the disentangled lip movement due to the high-dimensionality of GAN latent space. This obstacle leads to undesired movements within the predicted scene, such as flickering in the background or torso area. Furthermore, the triplane grid directly represents the 3D grid structure of the NeRF space which guides the model to understand and manipulate the spatial relationships within the scene. Lastly, the triplane grid is basically a 2D feature map returned from convolutional networks, which enables leveraging existing 2D-based network architectures.

Attention design. In an ideal setting, the deltaplane should seamlessly amalgamate temporal motion signals with the identity triplane, ensuring that the signals are appropriately synchronized with the relevant facial segments. This becomes imperative for audio, as their impact on the entirety of the facial movements is not uniform. We incorporate cross-attention at the deepest hidden layer of U-Net architecture to effectively capture localized facial dynamics during the generation of the deltaplane. Specifically, the U-Net encoder $\mathcal{F}_{\mathrm{enc}}$ encodes $\mathbf{P}$ into a low-resolution feature map as $\mathbf{E} = \mathcal{F}_{\mathrm{enc}}(\mathbf{P})$. Consequently, this feature map passes through a N_{A} number of attention layers, each composed of self-attention layer (SA) and cross-attention (CA) layer, which we denote as: $\mathcal{F}_{\mathrm{SA}}$ and $\mathcal{F}_{\mathrm{CA}}$. Specifically, SA and CA can be defined as:

$$\begin{aligned} \mathbf{e} &= \mathcal{F}_{\mathrm{SA}}(\mathrm{flatten}(\mathbf{E} + \mathbf{E}^{\mathrm{pos}})). \\ \mathbf{E}_n^{\mathrm{out}} &= \mathcal{F}_{\mathrm{CA}}(\mathbf{e}, \mathbf{a}_n), \end{aligned} \tag{4}$$

where $\mathbf{E}^{\mathrm{pos}}$ denotes 3D positional encoding. Finally, the U-Net decoder yields deltaplane by $\Delta\mathbf{P}_n = \mathcal{F}_{\mathrm{dec}}(\mathbf{E}_n^{\mathrm{out}})$.

Split-convolution. The original EG3D [4] employs a single convolution network to generate the triplane, where each plane, $\mathbf{P}^{\mathrm{xy}}$, $\mathbf{P}^{\mathrm{yz}}$, and $\mathbf{P}^{\mathrm{zx}}$, is channel-wise concatenated. However, we observed a performance decline when utilizing the $\mathcal{F}_{\mathrm{enc}}$ structure as a singular model. This degradation stems from the orthogonality of each plane within the NeRF space, and the channel-wise concatenation hinders the 3D-awareness of the triplane. To address this issue, our architecture processes each plane independently to maintain each plane's attributes. Nevertheless, since each plane's features equally contribute to the query sampled points by concatenation, the aforementioned split convolution structure hinders the learning of the correlation between each plane. Therefore, we incorporate the roll-out method [22] to appropriately blend features from each plane.

Augmenting condition. Due to the image cropping process in the utilization of the EG3D [4], our pre-processed video data has variations in image crop regions. Consequently, a specific challenge arises, wherein alterations to the crop area may give the appearance of unnecessary movement between the background and the torso's position, which interferes with the learning of audio features. To mitigate this challenge, we encode additional signals with causal relationships to the torso and background movements. Features capturing independent actions, such as background motion (inferred from facial landmarks), and torso dynamics (head rotation), are tokenized as $\mathbf{t}^{\mathrm{rot}}$ and $\mathbf{t}^{\mathrm{ldm}}$ and then incorporated through cross-attention layers. The intuition here lies in the effectiveness of our model's cross-attention layer, allowing diverse tokens to be efficiently learned for local editing within the triplane. Especially for $\mathbf{t}^{\mathrm{ldm}}$, we select only portions of landmarks, since the landmarks near the lip regions tend to degrade the lip-sync accuracy when novel audio features are given. Following ER-NeRF [14], we also employ the AU45 [8] features to predict eye movements, which also be tokenized into $\mathbf{t}^{\mathrm{eye}}$. Additionally, a single null-token $\mathbf{t}^{\mathrm{null}}$ is incorporated uniformly across all frames to encode global scene representation across video frames. Again, (4) can be reformulated with $\mathbf{t}_{\mathrm{n}}$ which denotes the concatenation of all tokens:

$$\mathbf{E}_n^{\mathrm{out}} = \mathcal{F}_{\mathrm{CA}}(\mathbf{e}, \mathbf{t}_n), \quad \mathbf{t}_n = \{\mathbf{a}_n, \mathbf{t}_n^{\mathrm{eye}}, \mathbf{t}_n^{\mathrm{rot}}, \mathbf{t}_n^{\mathrm{ldm}}, \mathbf{t}^{\mathrm{null}}\}. \tag{5}$$

3.4. Loss Functions

To train the model, we mainly adopt $L1$ loss $\mathcal{L}_{\mathrm{L1}}$ and LPIPS loss $\mathcal{L}_{\mathrm{lpips}}$ [27] to reconstruct given input frame I. Let $\mathcal{L}_{\mathrm{rec}}$ denotes the combination of the above reconstruction loss as: $\mathcal{L}_{\mathrm{rec}} = \mathcal{L}_{\mathrm{L1}} + \lambda_{\mathrm{lpips}}\mathcal{L}_{\mathrm{lpips}}$. We give additional reconstruction loss on lip segment $S_{\mathrm{lip}}(I)$, extracted using BiSeNet [25, 26], to enhance the reconstruction loss on the local image area. Moreover, we adopt ID similarity loss $\mathcal{L}_{\mathrm{id}}$ and syncnet loss $\mathcal{L}_{\mathrm{sync}}$ [6] to further optimize the generation results. We additionally take a few epochs to update the super-resolution module, in order to boost the model to generate sharper image results.

4. Experiments

4.1. Experimental Settings

We compare against the standard NeRF-based models: AD-NeRF [10], RAD-NeRF [20], and ER-NeRF [14]. We conduct two evaluation settings: *novel-view synthesis* assessing viewpoint robustness by rendering from diverse angles, and

Head angle (yaw, pitch)	$(-30^\circ, -20^\circ)$			$(-15^\circ, -10^\circ)$			$(+15^\circ, +10^\circ)$			$(+30^\circ, +20^\circ)$		
	Sync↑	FID↓	IDSIM↑	Sync↑	FID↓	IDSIM↑	Sync↑	FID↓	IDSIM↑	Sync↑	FID↓	IDSIM↑
AD-NeRF	2.24	212.85	0.07	3.47	175.98	0.28	3.82	152.02	0.48	2.52	193.34	0.03
RAD-NeRF	4.94	167.83	0.19	5.54	123.92	0.38	6.83	94.67	0.61	5.45	185.72	0.28
ER-NeRF	4.77	198.29	0.23	7.34	87.59	0.58	6.65	80.56	0.50	2.70	141.63	0.02
Talk3D	**7.20**	**81.11**	**0.61**	**7.93**	**37.77**	**0.77**	**8.14**	**39.97**	**0.80**	**7.77**	**68.68**	**0.64**

Table 1. **Quantitative comparison under the novel view synthesis setting.** The head poses are selected with 15° yaw intervals and 10° pitch intervals. The best results are shown in **bold**.

Methods	PSNR ↑	LPIPS ↓	FID ↓	LMD ↓	AUE ↓	Sync ↑	IDSIM ↑
Ground Truth	N/A	0	0	0	0	9.077	1
AD-NeRF	27.611	0.049	20.243	5.692	2.331	5.692	0.904
RAD-NeRF	28.797	0.038	14.218	3.467	2.163	6.316	0.921
ER-NeRF	29.284	0.032	11.860	3.417	2.025	6.724	0.940
Talk3D(Ours)	**30.185**	**0.027**	**8.626**	**2.932**	**1.920**	**7.383**	**0.944**

Table 2. **Quantitative results of the audio-driven setting**. The best results are shown in **bold**.

audio-driven evaluation using ground-truth video camera viewpoints with original audio. We evaluate reconstruction quality using PSNR and LPIPS. Additional metrics include FID [11], landmark distance LMD [5], SyncNet confidence score (Sync) [7], action units error (AUE) [3], and identity similarity (IDSIM) [12].

4.2. Novel View Synthesis

We evaluate our method's robustness to extreme viewpoints by rendering from diverse novel angles (yaw: $-30^\circ \sim 30^\circ$, pitch: $-20^\circ \sim 20^\circ$). As shown in Tab. 1, while most methods exhibit comparable performance in frontal view rendering, other NeRF-based techniques show notable score decline at extreme angles. Our method demonstrates consistently high scores across all metrics, maintaining both generation quality and lip-sync accuracy across diverse camera viewpoints. The qualitative results in Fig. 3 reveal that previous NeRF-based methods suffer from performance degradation at extreme camera angles. RAD-NeRF and ER-NeRF experience substantial quality decline due to their pseudo-3D deformable module, showing irregular torso boundaries and geometry deterioration. AD-NeRF, with independently learned head and torso volumes, produces disembodied heads at side angles. In contrast, Talk3D maintains realistic facial geometry and consistent quality across all viewpoints.

4.3. Audio-driven Synthesis

Our audio-driven evaluation (tested with unseen audio data during training) demonstrates superior performance across most image quality metrics while achieving the best lip synchronization among NeRF-based methods. As shown in Tab. 2, Talk3D outperforms in PSNR, LPIPS, and IDSIM, demonstrating high-fidelity detail reconstruction and facial identity preservation. The superior FID score indicates advantages of utilizing generative priors, while improved Sync, LMD and AUE scores show enhanced facial dynamics accuracy. We also provide qualitative comparison in Fig. 4, where Talk3D demonstrates robust results through its unified generation process.

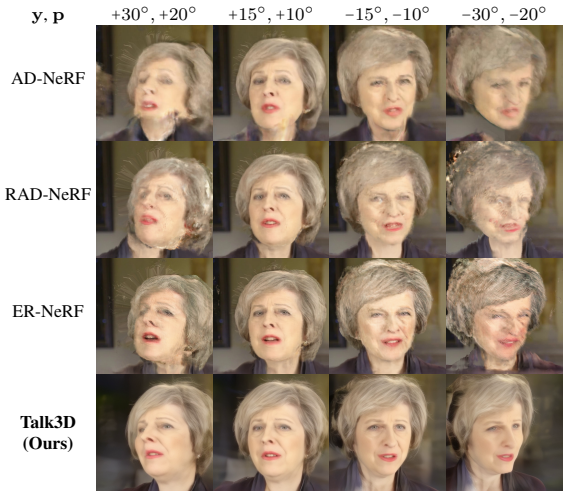

Figure 3. **Synthesized portraits from head poses unseen during training.** We show a randomly selected frame from synthesized talking portraits rendered at various yaw and pitch ($\mathbf{y}$, $\mathbf{p}$) angles.

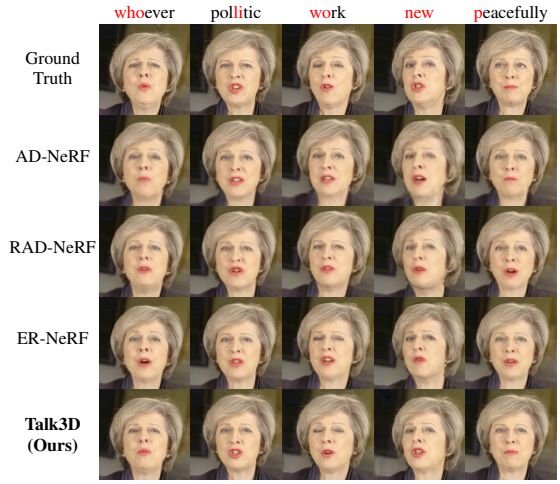

Figure 4. **The keyframe comparison of generated portraits.** We show visualizations of our method and previous methods generated using ground-truth head poses and audio from the test set. Best viewed in zoom.

5. Conclusion

In this work, we introduced Talk3D, a unified framework for audio-driven talking head synthesis that leverages 3D-aware generative priors and direct NeRF space manipulation. Our approach achieves robust lip-sync accuracy and high-fidelity facial detail reconstruction, even under challenging novel-view and audio-driven scenarios. Extensive experiments demonstrate that Talk3D delivers consistent and realistic results, advancing the quality of controllable talking portrait generation.

References

[1] Sizhe An, Hongyi Xu, Yichun Shi, Guoxian Song, Umit Y. Ogras, and Linjie Luo. Panohead: Geometry-aware 3d full-head synthesis in 360deg. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 20950–20959, 2023. 1

[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 (CVPR), pages 4541–4551, 2023. 2

[3] Tadas Baltrušaitis, Marwa Mahmoud, and Peter Robinson. Cross-dataset learning and person-specific normalisation for automatic action unit detection. In 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), pages 1–6. IEEE, 2015. 4

[4] 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, 3

[5] Lele Chen, Zhiheng Li, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. Lip movements generation at a glance. In Computer Vision–ECCV 2018: 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part VII 15, pages 538–553. Springer, 2018. 4

[6] Joon Son Chung and Andrew Zisserman. Lip reading in the wild. In Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II 13, pages 87–103. Springer, 2017. 3

[7] Joon Son Chung and Andrew Zisserman. Out of time: Automated lip sync in the wild. In Computer Vision–ACCV 2016 Workshops: ACCV 2016 International Workshops, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II 13, pages 251–263. Springer, 2017. 4

[8] Paul Ekman and Wallace V. Friesen. Facial Action Coding System: Manual. Palo Alto: Consulting Psychologists Press, 1978. 3

[9] Anna Frühstück, Nikolaos Sarafianos, Yuanlu Xu, Peter Wonka, and Tony Tung. Vive3d: Viewpoint-independent video editing using 3d-aware gans. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 4446–4455, 2023. 1, 2

[10] Yudong Guo, Keyu Chen, Sen Liang, Yong-Jin Liu, Hujun Bao, and Juyong Zhang. Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 5784–5794, 2021. 1, 3

[11] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems, 30, 2017. 4

[12] Yuge Huang, Yuhan Wang, Ying Tai, Xiaoming Liu, Pengcheng Shen, Shaoxin Li, Jilin Li, and Feiyue Huang. Curricularface: Adaptive curriculum learning loss for deep face recognition. In CVPR, 2020. 4

[13] 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. 2

[14] Jiahe Li, Jiawei Zhang, Xiao Bai, Jun Zhou, and Lin Gu. Efficient region-aware neural radiance fields for high-fidelity talking portrait synthesis. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pages 7568–7578, 2023. 1, 3

[15] Xian Liu, Yinghao Xu, Qianyi Wu, Hang Zhou, Wayne Wu, and Bolei Zhou. Semantic-aware implicit neural audio-driven video portrait generation. In Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVII, pages 106–125. Springer, 2022. 1

[16] 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 ECCV, 2020. 1

[17] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In Proceedings of the 28th ACM International Conference on Multimedia, pages 484–492, 2020. 1

[18] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. ACM Trans. Graph., 2021. 2

[19] Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Synthesizing obama: learning lip sync from audio. ACM Transactions on Graphics (ToG), 36(4):1–13, 2017. 1

[20] Jiaxiang Tang, Kaisiyuan Wang, Hang Zhou, Xiaokang Chen, Dongliang He, Tianshu Hu, Jingtuo Liu, Gang Zeng, and Jingdong Wang. Real-time neural radiance talking portrait synthesis via audio-spatial decomposition. arXiv preprint arXiv:2211.12368, 2022. 1, 3

[21] Justus Thies, Mohamed Elgharib, Ayush Tewari, Christian Theobalt, and Matthias Nießner. Neural voice puppetry: Audio-driven facial reenactment. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16, pages 716–731. Springer, 2020. 1

[22] Tengfei Wang, Bo Zhang, Ting Zhang, Shuyang Gu, Jianmin Bao, Tadas Baltrusaitis, Jingjing Shen, Dong Chen, Fang Wen, Qifeng Chen, et al. Rodin: A generative model for sculpting 3d digital avatars using diffusion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 4563–4573, 2023. 3

[23] Zhenhui Ye, Ziyue Jiang, Yi Ren, Jinglin Liu, Jinzheng He, and Zhou Zhao. Geneface: Generalized and high-fidelity audio-driven 3d talking face synthesis. In The Eleventh International Conference on Learning Representations, 2022. 1

[24] Zhenhui Ye, Jinzheng He, Ziyue Jiang, Rongjie Huang, Jiawei Huang, Jinglin Liu, Yi Ren, Xiang Yin, Zejun Ma, and Zhou Zhao. Geneface++: Generalized and stable real-time audio-driven 3d talking face generation. arXiv preprint arXiv:2305.00787, 2023. 1

[25] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In Proceedings of the European conference on computer vision (ECCV), pages 325–341, 2018. 3

[26] Changqian Yu, Changxin Gao, Jingbo Wang, Gang Yu, Chunhua Shen, and Nong Sang. Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation. International Journal of Computer Vision, 129: 3051–3068, 2021. 3

[27] 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. 3

[28] Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, and Fei Wang. Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 8652–8661, 2023. 1

[29] Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li. Makelttalk: speaker-aware talking-head animation. ACM Transactions On Graphics (TOG), 39(6):1–15, 2020. 1