WildTalker: Talking Portrait Synthesis In the Wild

Seonghak Lee[1*], Jisoo Park[1*], and Junseok Kwon[1]

Chung-Ang University, Seoul, Korea
{tjdgkr6220,susiehome,jskwon}@cau.ac.kr

Abstract. We introduce WildTalker, a novel approach for synthesizing high-quality talking portraits that effectively addresses the challenges of real-world environments. Traditional methods often struggle with unpredictable movements and noisy audio. WildTalker overcomes these issues by integrating flow-guided temporal masking, which manages dynamic regions by capturing and de-emphasizing transient areas, and multi-scale spectral subtraction for robust audio denoising. This method allows WildTalker to excel in both controlled and variable scenarios, producing natural and synchronized talking portraits with accurate lip synchronization. Our experiments demonstrate that WildTalker significantly enhances the quality of audio-driven 3D talking portraits in dynamic settings, achieving superior lip synchronization under challenging audio conditions. These results highlight that our method outperforms existing approaches not only in real-world scenarios but also in controlled environments, underscoring its potential for practical applications.

Keywords: Neural Radiance Field · Talking Portrait Synthesis · Wild Scenario

1 Introduction

Reconstructing a talking portrait has been a long-standing focus of research in computer vision and computer graphics, with broad applications ranging augmented reality and visual avatars to film production and conversational agents.

Recently, AD-NeRF [17], a pioneer in applying NeRF to audio-driven talking portrait synthesis, has employed separate training for the head and torso NeRFs to accommodate varying degrees of motion. This approach has been widely adopted in subsequent studies [22, 26, 30, 35, 40] and has achieved impressive results in rendering realistic portraits, particularly in controlled environments. As a result, the synthesis of talking heads has reached a high level of realism, especially in scenarios where the subject remains relatively stationary.

However, despite these advancements in head synthesis, traditional methods [2, 4, 16, 17, 19, 22, 23, 34, 35, 40] still struggle significantly in real-world scenarios. These methods often falter when faced with unpredictable movements,

[*] These authors contributed equally to this work.

such as hand gestures, and ambient noise. While these approaches excel at rendering lip movements and facial expressions in stationary settings, they generally neglect the complexity of the entire "portrait", particularly the quality of the torso. The limitations of these traditional methods largely stem from their reliance on deriving torso movements from head poses. Although there is some correlation between head and torso movements, the irrelevant things (*e.g.* hand gestures) are challenging to current methods. In particular, in audio-driven tasks, where the input speech during inference may differ from the audio used during training, mismatch between speech and hand gestures can lead to unnatural results if hand movements do not align with the speech.

To address these challenges, we introduce a novel approach called *WildTalker*, which is designed to handle both disturbing movements and noisy audio while focusing on generating a realistic torso. We define high dynamic region with sudden, large movements as transient area and de-prioritize these areas during rendering. To achieve this, we utilize our Flow-guided temporal mask, which leverages optical flow to capture these areas. This approach allows us to effectively manage unpredictable movements while maintaining the visual coherence and realism of the overall portrait.

We also employ a Multi-scale Spectral Subtraction (MSS) denoiser to handle noisy audio, ensuring accurate lip synchronization even in challenging environments. Existing talking portrait synthesis methods [17, 22, 35] often struggle to accurately predict lip movements when the input audio is slightly corrupted. While deep learning methods often achieve high performance, they tend to be slower and heavily dependent on large training datasets. In contrast, traditional signal processing methods like spectral subtraction are more efficient, faster, and less reliant on extensive training data. Our MSS denoiser enhances the traditional spectral subtraction method by applying it across multiple scales, capturing a broader range of noise types and improving robustness in diverse scenarios.

Our comprehensive experiments demonstrate that the proposed WildTalker not only maintains high-quality synthesis in controlled settings but also effective in wild settings. As a result, it produces robust outcomes even when faced with diverse and unrefined real-world data.

2 Related Works

Talking Portrait Synthesis Audio-driven talking portrait synthesis focuses on reenacting realistic and expressive talking portrait from arbitrary speech audio. Traditional approaches range from image-based methods [10–12] to model-based methods [5, 13, 20, 31, 34, 37]. While effective in generating facial expressions and lip movements, they typically treat the torso as a static background, neglecting its dynamic movements.

Recently, NeRF-based methods [17, 22, 26, 30, 35, 40] have achieved photorealistic rendering. AD-NeRF [17] synthesized talking heads from audio inputs, bypassing the need for intermediate representations. RAD-NeRF [35] enhanced efficiency by decomposing the synthesis into spatial and temporal components. DFRF [30] enabled rapid personalization with minimal reference images. GeneFace [40] enhanced the fidelity using variational motion generator, domain-

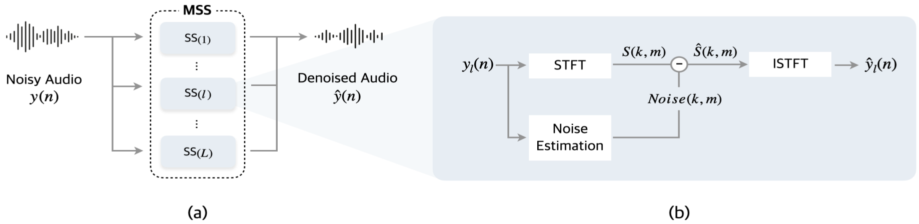

Fig. 1: Overview of Multi-scale Spectral Subtraction (MSS) audio denoiser. (a) shows the overall denoising process, where the noisy audio $y(n)$ is fed into the MSS denoiser. This module performs spectral subtraction (SS) at different scales, producing the denoised audio $\hat{y}(n)$. (b) details the internal workings of MSS for a single scale.

adaptive post-net, and head aware torso-NeRF to address head-torso separation. SyncTalk [26] improved synchronization by aligning lip movements with speech, accurately capturing expressions with a 3D blendshape model, and stabilizing head movements.

However, these methods excel mainly in controlled environments, such as videos of news anchors sitting upright with their hands still. They often struggle with more dynamic settings, particularly when rendering sudden movements like hand gestures, leading to blurry or unrealistic outputs in the torso region. Our approach builds on the strengths of NeRF while introducing mechanisms to handle dynamic torso movements effectively.

Unconstrained Scenes Uncertainty modeling in NeRF has become critical for applications in unstructured and unconstrained environments. This approach aims to enhance the reliability and robustness of scene representation, particularly when faced with the dynamics of real-world conditions. NeRF-W [24] extended the original NeRF to handle highly variable photographic collections by separating static and transient scene elements, effectively managing inconsistencies. Further advancements by D-NeRF [38] and RobustNeRF [29] focused on separating moving objects and their shadows from static backgrounds using dual NeRF architectures. This allows for more accurate modeling of dynamic elements without compromising the representation of the static scene.

Despite efforts to handle wild conditions, existing methods do not effectively address the synthesis of talking portraits in environments with dynamic movements and background noise, which degrade the output quality. In contrast, our method is specifically designed to compensate for diverse unpredictable elements in dynamic videos, resulting in higher-quality talking portraits.

3 Methods

Multi-scale Spectral Subtraction Denoiser The traditional spectral subtraction method [3] employs a Short-Time Fourier Transform (STFT) to reduce noise in audio signals. While effective in controlled environments, it often struggle with the varied and dynamic noise conditions in real-world scenarios. To improve robustness and adaptability in noise reduction, we propose Multi-scale Spectral Subtraction (MSS) denoiser.

Our MSS denoiser, as shown in Fig.1, processes the noisy audio signal $y(n)$ at each discrete time index n by transforming it through STFT to obtain the spectral representation $S(k, m)$ at each frequency bin k and frame m, defined as $S(k, m) = \mathrm{STFT}(y(n), k, m)$. Noise is then estimated and subtracted from the spectrum using spectral subtraction:

$$\hat{S}(k, m) = \max\left(|S(k, m)| - \alpha \cdot \mathrm{Noise}(k, m), 0\right), \tag{1}$$

where $\hat{S}(k, m)$ represents the magnitude spectrum of the estimated clean signal, α controls the level of noise suppression, and $\mathrm{Noise}(k, m)$ denotes the magnitude spectrum of the estimated noise signal. The noise estimation process assumes initial STFT frames primarily contain background noise, which is then subtracted from the noisy signal. After the noise reduction, the clean spectral components are converted back to the time domain using the Inverse Short-Time Fourier Transform (ISTFT), yielding the clean audio output $\hat{y}_l(n)$ for each scale. The final denoised audio signal, $\hat{y}(n)$, is obtained by averaging the outputs from all scales:

$$\hat{y}(n) = \frac{1}{L} \sum_{l=1}^{L} \hat{y}_l(n), \tag{2}$$

where L denotes the number of STFT window scales used. The scale parameters are number of Fast Fourier Transform components, hop length and window length. This multi-scale approach improves noise mitigation across a broad frequency range, addressing both high-frequency transient noises and low-frequency persistent disturbances. Then, $\hat{y}(n)$ is processed through a pre-trained Audio Feature Extractor (*e.g.* DeepSpeech [18]) to obtain the denoised lip feature l_{dn}.

Denoised Head Renderer Using the Face-Sync Controller and Head-Sync Stabilizer proposed in Synctalk [26], we extract the final lip feature f_l and expression feature f_e by utilizing the lip feature l_{dn} obtained from the MSS denoiser, along with the head pose parameters R and T. The Tri-Plane Hash Representation, first introduced in ER-NeRF [22], is used for high quality talking head synthesis. This method employs three distinct 2D multi-resolution hash grid encoders, $\mathcal{H}^{AB} : (a, b) \to f_{ab}^{AB}$, each with unique orientations to reduce hash collisions. Here, $\mathcal{H}^{AB}$ represents the multi-resolution hash encoder in $\mathbb{R}^{AB}$, where (a, b) denotes the projected coordinate. The three distinct 2D geometric features obtained from these encoders are concatenated to form a single geometric feature $f_{\mathbf{x}} = \mathcal{H}_{xy}^{XY} \odot \mathcal{H}_{yz}^{YZ} \odot \mathcal{H}_{xz}^{XZ}$. This geometric feature, along with lip features f_l, expression features f_e, 3D coordinates $\mathbf{x} = (x, y, z)$, and viewing direction $\mathbf{d} = (\theta, \phi)$, is passed through the implicit function $F : \mathrm{MLP}(\mathbf{x}, \mathbf{d}, f_l, f_e; f_{\mathbf{x}})$ to produce the final color $\mathbf{c}$ and density σ.

The colors and volume densities along the camera ray r, from the near bound t_n to the far bound t_f, are accumulated to calculate a pixel color $\hat{\mathcal{C}}(r)$:

$$\hat{\mathcal{C}}(r) = \int_{t_n}^{t_f} \sigma(r(t)) \cdot \mathbf{c}(r(t), \mathbf{d}) \cdot T(t) dt, \tag{3}$$

where $T(t)$ represents the accumulated transmittance along the ray ranging from t_n to t, as follows: $T(t) = \exp\left(-\int_{t_n}^{t} \sigma(r(s)) ds\right)$.

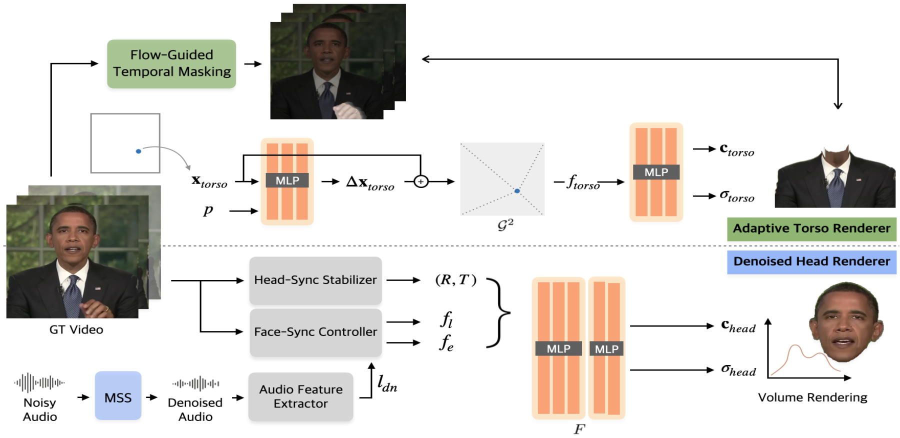

Fig. 2: WildTalker Network architecture. The figure shows two pipelines for talking portrait synthesis. The top demonstrates the Adaptive Torso Renderer, using flow-guided temporal masking. The bottom shows Denoised Head Renderer, with a Head-Sync Stabilizer and Face-Sync Controller, synchronized to denoised audio via Multi-scale Spectral Subtraction (MSS) denoiser.

Adaptive Torso Renderer To effectively address unpredictable regions such as suddenly appearing objects or abrupt hand gestures, we propose a flow-guided temporal masking that ensures efficient and realistic rendering of the torso. Our method leverages optical flow [39] to track dynamic objects and de-emphasize transient areas. By analyzing regions with high optical flow vectors, we can identify significant motion and classify transient objects. However, simply masking regions with large optical flow magnitudes is insufficient. This approach risks overlooking objects that persist beyond their initial appearance and can degrade the quality of torso rendering by failing to account for subtle or sustained movements. To address these issues comprehensively, we accumulate optical flow magnitudes over time, allowing us to robustly capture both abrupt motions and unwanted sustained objects. The temporal mask, M_i, for frame i is defined as:

$$M_i = \max\left(M_{i-1} \cdot \delta_i,\ \mathbf{1}(|V_i| > \theta)\right), \tag{4}$$

where M_{i-1} is the cumulative mask from the previous frame, δ_i is a decay factor dynamically adjusted based on the optical flow magnitude at frame i, and $\mathbf{1}(\cdot)$ generates a mask where the magnitude $|V_i|$ exceeds a threshold θ. This dynamic accumulation approach balances the need to capture transient objects while preserving important torso details. For torso rendering, we adopt a Pseudo-3D Deformable method [35]. The 2D coordinates $\mathbf{x}_{torso}$ and camera poses $\mathbf{p}$ are first deformed via $\Delta\mathbf{x}_{torso} = \text{MLP}(\mathbf{x}_{torso}, \mathbf{p})$. The deformation $\Delta\mathbf{x}_{torso}$ is then passed through a feature grid encoder $\mathcal{G}^2$ to obtain the torso feature f_{torso}, which is used by an implicit function to compute the color $\mathbf{c}_{torso}$ and density σ_{torso}. Fig. 2 illustrates the full network architecture.

Training Details We utilize Mean Squared Error (MSE) loss to measure the overall difference between the predicted color $\hat{\mathcal{C}}$ and the actual color $\mathcal{C}$:

$$\mathcal{L}_{recon}(r) = \sum_{i \in \mathcal{I}} ||\mathcal{C}_i(r) - \hat{\mathcal{C}}_i(r)||_2^2, \tag{5}$$

where the predicted color $\hat{\mathcal{C}}$ can be obtained by Eq.3 using $\mathbf{c}_{head}$ (or $\mathbf{c}_{torso}$) and σ_{head} (or σ_{torso}), and i and r denote individually the image index and ray.

We utilize the flow-guided temporal mask M_i, obtained from Eq.4, as the loss weight to modify the MSE loss in Eq.5:

$$\mathcal{L}_{torso} = M(r) \cdot \mathcal{L}_{recon}(r), \tag{6}$$

where $M(r)$ denotes the flow-guided temporal mask applied along the camera ray r, derived from M_i in Eq.4.

Due to the small size of the mouth, evaluating its movements with only the MSE loss in Eq.5 is insufficient, as it fails to capture the finer details of lip movements. To enhance the model's ability to learn these detailed features, we perform lip finetuning. We use patches $\mathcal{P}$ sampled from the image and incorporate LPIPS (Learned Perceptual Image Patch Similarity) [41] loss to further refine the training.

$$\mathcal{L}_{head} = \mathcal{L}_{recon}(r) \ + \ \lambda \text{LPIPS}(\mathcal{P}, \hat{\mathcal{P}}), \tag{7}$$

where $\hat{\mathcal{P}}$ represents the patches of the rendered image. As our training process is divided into head and torso pipelines, the head part is trained with Eq.7, and the torso part is trained with Eq.6.

4 Experiments

Experiments Details Unlike existing methods that primarily focus on the head part and evaluate performance after training only the head, we conducted experiments using a 5-minute video that fully captures the entire torso and hand movements. For a fair comparison, we excluded methods that do not handle the torso in the quantitative evaluation of torso. All experiments, except those using AD-NeRF [17] with a 450 × 450 resolution, were performed with images at a resolution of 512 × 512. To generate hand-absent videos, we synthesized the ground truth (GT) images without hands using LaMa [33]. The optical flow was used as input only during the training phase and was not utilized during testing. To test the method's denoising capability in real-world conditions, we generated noisy audio with various background noises sourced from YouTube.

Evaluation on Rendering Quality To evaluate the quality of the reconstructed portrait, we employed PSNR, LPIPS [41], and LMD [6]. Tab.1 presents a comparative analysis of our method against several state-of-the-art NeRF approaches. Notably, the proposed WildTalker achieved the lowest LPIPS and highest PSNR score, which underscores our method's ability to generate realistic images. Moreover, WildTalker also achieved the lowest LMD and AUE score, indicating its superior performance in accurately capturing lip movements.

Table 1: Quantitative Evaluation of Torso Reconstruction with dynamic movements and noisy audio. We highlighted the **best** and the second-best results. LMD for AD-NeRF could not be measured.

Methods	PSNR↑	LPIPS↓	LMD↓	AUE↓	LSE-C↑	LSE-D↓	Time(h)
AD-NeRF ICCV'21 [17]	22.742	0.228	-	0.467	0.323	12.071	36.4
RAD-NeRF arXiv'22 [35]	23.736	0.166	1.929	0.354	6.329	8.427	4.0
ER-NeRF ICCV'23 [22]	24.183	0.113	2.196	0.240	7.199	7.830	2.4
SyncTalk CVPR'24 [26]	24.580	0.101	2.100	0.211	6.096	8.402	**2.2**
WildTalker(Ours)	**25.602**	**0.079**	**1.881**	**0.176**	**8.752**	**6.283**	**2.2**

Table 2: Quantitative Evaluation of Lip Synchronization with both original and corrupted audio. The **best** and second-best results were highlighted.

		Original			Corrupted		
		AUE↓	LSE-C↑	LSE-D↓	AUE↓	LSE-C↑	LSE-D↓
2D	Wav2Lip ACM MM'20 [27]	**0.246**	8.447	6.241	0.313	**7.413**	**7.330**
	VideoReTalking SIGGRAPH Asia'22 [7]	0.270	8.066	7.075	0.290	7.014	7.953
	DINet AAAI'23 [42]	0.340	6.775	8.026	0.355	5.801	8.904
	TalkLip CVPR'23 [27]	0.300	6.219	8.378	0.304	5.325	8.502
	IP-LAP CVPR'23 [43]	0.294	5.571	8.975	0.298	3.845	10.632
NeRF	AD-NeRF ICCV'21 [17]	0.294	5.005	9.957	0.313	4.603	10.212
	RAD-NeRF arXiv'22 [35]	0.356	3.598	9.642	0.357	2.197	11.591
	GeneFace ICLR'23 [40]	0.266	6.876	7.076	0.318	4.587	9.243
	ER-NeRF ICCV'23 [22]	0.271	6.808	8.004	0.371	3.052	11.130
	SyncTalk CVPR'24 [26]	0.306	7.108	7.108	**0.278**	6.674	9.343
	WildTalker (Ours)	0.283	**8.928**	**6.126**	0.280	7.076	7.889

Furthermore, the LSE-C and LSE-D metrics [8] demonstrate the robustness of our approach in terms of synchronization and expression consistency, with our method showing substantial improvements, particularly in LSE-C. This suggests that WildTalker excels not only in rendering high-quality visuals but also in maintaining accurate lip synchronization and expression dynamics. Finally, we compared the training time to demonstrate the efficiency of our method.

Evaluation on Lip Synchronization Tab.2 shows the results of lip synchronization in head reconstruction, evaluated with both original and corrupted audio (synthesized with real-world noise). Our method outperforms state-of-the-art NeRF-based methods and provides results comparable to Wav2Lip [27], showing strong performance even with corrupted audio. To further assess the denoising performance of the MSS denoiser, we used VoiceBank-DEMAND [36] test sets and compared it to deep learning-based methods [1,9,14,15,21,25,28,32]. The test sets consist of 824 utterances from two speakers with four SNR levels (17.5, 12.5, 7.5, and 2.5 dB). Performance was evaluated using PESQ (speech quality), CSIG (signal distortion), CBAK (background noise), and COVL (overall speech quality). As shown in Tab.3, MSS achieved the highest scores in PESQ and CBAK, demonstrating strong denoising and noise reduction capabilities. Although it ranked second in CSIG and third in COVL, it remains highly competitive with SOTA methods [1], proving MSS's efficiency and robust performance.

Qualitative Evaluation Fig.3 demonstrates qualitative results under self-driven and cross-driven settings. We compared our method with RAD-NeRF [35], ER-NeRF [22], and SyncTalk [26], observing that existing methods often produce blurry outputs when handling dynamic movements. In the self-driven setting, our approach showed superior handling of consistent hand movements, outperforming previous methods, which frequently blur hand motions and degrade torso

Table 3: Quantitative Evaluation of Speech Enhancement compared with deep learning-based methods on the VoiceBank-DEMAND dataset. We highlighted the **best** and the second-best results.

Methods	PESQ↑	CSIG↑	CBAK↑	COVL↑
Noisy	1.97	3.35	2.44	2.63
SEGAN Interspeech'17 [25]	2.16	3.48	2.94	2.80
MMSE-GAN ICASSP'18 [32]	2.53	3.80	3.12	3.14
Metric-GAN ICML'19 [14]	2.86	3.99	3.18	3.42
HiFi-GAN NeurIPS'20 [21]	2.94	4.07	3.07	3.49
DEMUCS ICASSP'23 [28]	3.07	4.31	3.40	3.63
MetricGAN+ Interspeech'21 [15]	3.15	4.14	3.16	3.64
DPT-FSNET ICASSP'22 [9]	3.33	4.58	3.72	4.00
CMGAN ICASSP'24 [1]	3.41	**4.63**	3.94	**4.12**
MSS (Ours)	**3.87**	4.62	**4.16**	3.64

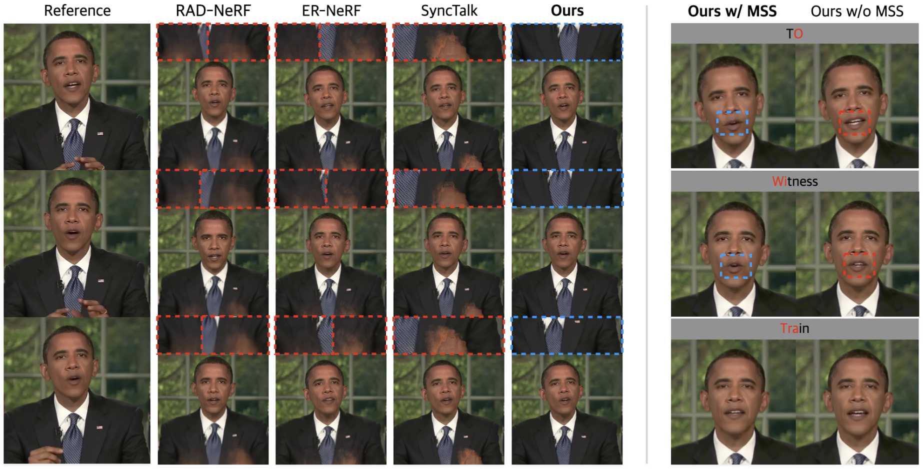

Fig. 3: Comparisons under self-driven(left) and cross-driven(right) settings. The left figure illustrates the results of rendering video with dynamic hand gestures. Our method effectively removed transient hand movements, producing a cleaner torso region, as highlighted in the zoomed-in hand areas. The right figure illustrates lip synchronization when inferring arbitrary noisy audio. MSS(blue boxes) yields more accurate results, while without MSS(red boxes), the synchronization is less precise.

quality. In the cross-driven setting, we evaluated lip synchronization with noisy audio inputs, demonstrating that incorporating the MSS denoiser enables more accurate and synchronized lip movements compared to our model without MSS. *More experiments can be found in the supplementary materials.*

5 Conclusion

In this paper, we present the WildTalker, a novel approach for synthesizing high-quality talking portraits in real-world environments. By integrating a flow-guided temporal mask and Multi-scale Spectral Subtraction (MSS) denoiser, WildTalker effectively handles unpredictable movements and noisy audio, achieving state-of-the-art results across various scenarios. Our method significantly outperforms existing approaches, as shown by improvements in perceptual quality metrics such as LPIPS and LSE, highlighting its ability to generate natural and synchronized talking portraits, even in challenging settings.

References

1. Abdulatif, S., Cao, R., Yang, B.: Cmgan: Conformer-based metric-gan for monaural speech enhancement. IEEE/ACM Transactions on Audio, Speech, and Language Processing (2024) 7, 8
2. An, S., Xu, H., Shi, Y., Song, G., Ogras, U.Y., Luo, L.: Panohead: Geometry-aware 3d full-head synthesis in 360deg. In: CVPR. pp. 20950–20959 (2023) 1
3. Boll, S.: Suppression of acoustic noise in speech using spectral subtraction. IEEE Transactions on Acoustics, Speech, and Signal Processing **27**(2), 113–120 (1979) 3
4. Chan, E.R., Lin, C.Z., Chan, M.A., Nagano, K., Pan, B., Mello, S.D., Gallo, O., Guibas, L., Tremblay, J., Khamis, S., Karras, T., Wetzstein, G.: Efficient geometry-aware 3D generative adversarial networks. In: arXiv preprint arXiv:2112.07945 (2021) 1
5. Chen, L., Cui, G., Liu, C., Li, Z., Kou, Z., Xu, Y., Xu, C.: Talking-head generation with rhythmic head motion. In: ECCV. pp. 35–51 (2020) 2
6. Chen, L., Li, Z., Maddox, R.K., Duan, Z., Xu, C.: Lip movements generation at a glance. In: ECCV. pp. 520–535 (2018) 6
7. Cheng, K., Cun, X., Zhang, Y., Xia, M., Yin, F., Zhu, M., Wang, X., Wang, J., Wang, N.: Videoretalking: Audio-based lip synchronization for talking head video editing in the wild. In: SIGGRAPH Asia. pp. 1–9 (2022) 7
8. Chung, J.S., Zisserman, A.: Out of time: automated lip sync in the wild. In: ACCV. pp. 251–263 (2017) 7
9. Dang, F., Chen, H., Zhang, P.: Dpt-fsnet: Dual-path transformer based full-band and sub-band fusion network for speech enhancement. In: ICASSP. pp. 6857–6861 (2022) 7, 8
10. Das, D., Biswas, S., Sinha, S., Bhowmick, B.: Speech-driven facial animation using cascaded gans for learning of motion and texture. In: ECCV. pp. 408–424 (2020) 2
11. Doukas, M.C., Zafeiriou, S., Sharmanska, V.: Headgan: Video-and-audio-driven talking head synthesis. arXiv preprint arXiv:2012.08261 **1**(2) (2020) 2
12. Eskimez, S.E., Zhang, Y., Duan, Z.: Speech driven talking face generation from a single image and an emotion condition. IEEE Transactions on Multimedia **24**, 3480–3490 (2021) 2
13. Fried, O., Tewari, A., Zollhöfer, M., Finkelstein, A., Shechtman, E., Goldman, D.B., Genova, K., Jin, Z., Theobalt, C., Agrawala, M.: Text-based editing of talking-head video. ACM Transactions on Graphics **38**(4), 1–14 (2019) 2
14. Fu, S.W., Liao, C.F., Tsao, Y., Lin, S.D.: Metricgan: Generative adversarial networks based black-box metric scores optimization for speech enhancement. In: ICML. pp. 2031–2041 (2019) 7, 8
15. Fu, S.W., Yu, C., Hsieh, T.A., Plantinga, P., Ravanelli, M., Lu, X., Tsao, Y.: MetricGAN+: An Improved Version of MetricGAN for Speech Enhancement. In: Interspeech. pp. 201–205 (2021) 7, 8
16. Guan, J., Zhang, Z., Zhou, H., HU, T., Wang, K., He, D., Feng, H., Liu, J., Ding, E., Liu, Z., Wang, J.: Stylesync: High-fidelity generalized and personalized lip sync in style-based generator. In: CVPR (2023) 1
17. Guo, Y., Chen, K., Liang, S., Liu, Y., Bao, H., Zhang, J.: Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In: ICCV (2021) 1, 2, 6, 7
18. Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., Prenger, R., Satheesh, S., Sengupta, S., Coates, A., et al.: Deep speech: Scaling up end-to-end speech recognition. arXiv preprint arXiv:1412.5567 (2014) 4

19. Ji, X., Zhou, H., Wang, K., Wu, W., Loy, C.C., Cao, X., Xu, F.: Audio-driven emotional video portraits. In: CVPR (2021) 1
20. Karras, T., Aila, T., Laine, S., Herva, A., Lehtinen, J.: Audio-driven facial animation by joint end-to-end learning of pose and emotion. ACM Transactions on Graphics **36**(4), 1–12 (2017) 2
21. Kong, J., Kim, J., Bae, J.: Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. In: NeurIPS (2020) 7, 8
22. Li, J., Zhang, J., Bai, X., Zhou, J., Gu, L.: Efficient region-aware neural radiance fields for high-fidelity talking portrait synthesis. In: ICCV. pp. 7568–7578 (2023) 1, 2, 4, 7
23. Lu, Y., Chai, J., Cao, X.: Live Speech Portraits: Real-time photorealistic talking-head animation. ACM Transactions on Graphics **40**(6) (2021) 1
24. Martin-Brualla, R., Radwan, N., Sajjadi, M.S.M., Barron, J.T., Dosovitskiy, A., Duckworth, D.: NeRF in the Wild: Neural Radiance Fields for Unconstrained Photo Collections. In: CVPR (2021) 3
25. Pascual, S., Bonafonte, A., Serrà, J.: SEGAN: Speech Enhancement Generative Adversarial Network. In: Interspeech. pp. 3642–3646 (2017) 7, 8
26. Peng, Z., Hu, W., Shi, Y., Zhu, X., Zhang, X., Zhao, H., He, J., Liu, H., Fan, Z.: Synctalk: The devil is in the synchronization for talking head synthesis. In: CVPR. pp. 666–676 (2024) 1, 2, 3, 4, 7
27. Prajwal, K.R., Mukhopadhyay, R., Namboodiri, V.P., Jawahar, C.: A lip sync expert is all you need for speech to lip generation in the wild. In: ACM Multimedia. p. 484–492 (2020) 7
28. Rouard, S., Massa, F., Défossez, A.: Hybrid transformers for music source separation. In: ICASSP (2023) 7, 8
29. Sabour, S., Vora, S., Duckworth, D., Krasin, I., Fleet, D.J., Tagliasacchi, A.: Robustnerf: Ignoring distractors with robust losses. In: CVPR. pp. 20626–20636 (2023) 3
30. Shen, S., Li, W., Zhu, Z., Duan, Y., Zhou, J., Lu, J.: Learning dynamic facial radiance fields for few-shot talking head synthesis. In: ECCV. pp. 666–682 (2022) 1, 2
31. Song, L., Wu, W., Qian, C., He, R., Loy, C.C.: Everybody's talkin': Let me talk as you want. IEEE Transactions on Information Forensics and Security **17**, 585–598 (2022) 2
32. Soni, M.H., Shah, N., Patil, H.A.: Time-frequency masking-based speech enhancement using generative adversarial network. In: ICASSP. pp. 5039–5043 (2018) 7, 8
33. Suvorov, R., Logacheva, E., Mashikhin, A., Remizova, A., Ashukha, A., Silvestrov, A., Kong, N., Goka, H., Park, K., Lempitsky, V.: Resolution-robust large mask inpainting with fourier convolutions. arXiv preprint arXiv:2109.07161 (2021) 6
34. Suwajanakorn, S., Seitz, S.M., Kemelmacher-Shlizerman, I.: Synthesizing obama: learning lip sync from audio. ACM Transactions on Graphics **36**(4), 1–13 (2017) 1, 2
35. Tang, J., Wang, K., Zhou, H., Chen, X., He, D., Hu, T., Liu, J., Zeng, G., Wang, J.: Real-time neural radiance talking portrait synthesis via audio-spatial decomposition. arXiv preprint arXiv:2211.12368 (2022) 1, 2, 5, 7
36. Thiemann, J., Ito, N., Vincent, E.: The diverse environments multi-channel acoustic noise database (demand): A database of multichannel environmental noise recordings. Journal of the Acoustical Society of America **133**, 3591–3591 (2013) 7
37. Thies, J., Elgharib, M., Tewari, A., Theobalt, C., Nießner, M.: Neural voice puppetry: Audio-driven facial reenactment. In: ECCV. pp. 716–731 (2020) 2

38. Wu, T., Zhong, F., Tagliasacchi, A., Cole, F., Oztireli, C.: Dˆ 2nerf: Self-supervised decoupling of dynamic and static objects from a monocular video. In: NeurIPS. pp. 32653–32666 (2022) 3
39. Xu, H., Zhang, J., Cai, J., Rezatofighi, H., Tao, D.: Gmflow: Learning optical flow via global matching. In: CVPR. pp. 8121–8130 (2022) 5
40. Ye, Z., Jiang, Z., Ren, Y., Liu, J., He, J., Zhao, Z.: Geneface: Generalized and high-fidelity audio-driven 3d talking face synthesis. arXiv preprint arXiv:2301.13430 (2023) 1, 2, 7
41. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018) 6
42. Zhang, Z., Hu, Z., Deng, W., Fan, C., Lv, T., Ding, Y.: Dinet: Deformation inpainting network for realistic face visually dubbing on high resolution video. In: AAAI. pp. 3543–3551 (2023) 7
43. Zhong, W., Fang, C., Cai, Y., Wei, P., Zhao, G., Lin, L., Li, G.: Identity-preserving talking face generation with landmark and appearance priors. In: CVPR. pp. 9729–9738 (2023) 7