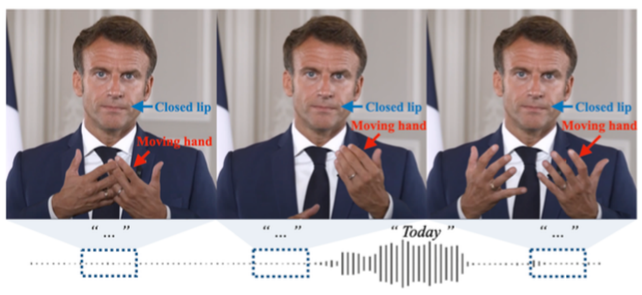
WildTalker: Talking Portrait Synthesis In the Wild

Seonghak Lee* (tjdkgr6220@cau.ac.kr) Jisoo Park* (susiehome@cau.ac.kr)
Junseok Kwon † (jskwon@cau.ac.kr)
Chung-Ang University, Seoul, Korea

* These authors contributed equally to this work.

Motivation

- We address new in-the-wild challenges for 3D talking portrait synthesis, focusing on scenarios that have not been fully explored in previous research.
- We propose leveraging optical flow to identify and de-prioritize transient areas with sudden, large movements, thereby improving the audio-visual coherence of talking portraits.
- We introduce a Multi-scale Spectral Subtraction(MSS) denoiser that efficiently handles real-world noise, enhancing audio-lip synchronization without requiring extensive training datasets.



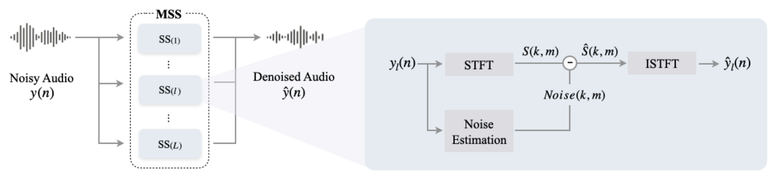
In audio-driven tasks, where the input speech during inference differs from the training audio, mismatches between speech and hand gestures can lead to unnatural outcomes if hand movements do not align with the speech.

Introduction

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 transient movements and noisy audio. WildTalker overcomes these issues by integrating flow-guided temporal masking, which adeptly processes dynamic regions by capturing and de-emphasizing transient areas that could disrupt visual coherence. Additionally, WildTalker employs multi-scale spectral subtraction for robust audio denoising, ensuring accurate and natural lip synchronization even under challenging auditory conditions. This comprehensive approach allows WildTalker to excel in both controlled and variable scenarios, producing highly realistic and synchronized talking portraits. 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.

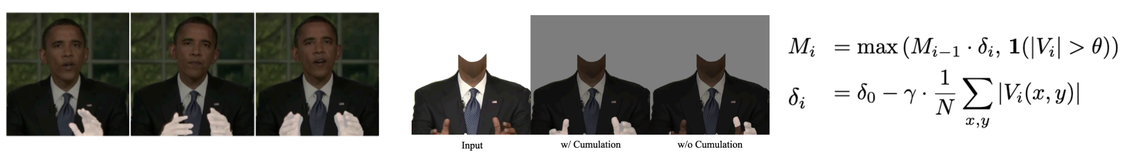
Proposed Method

- MSS Denoiser



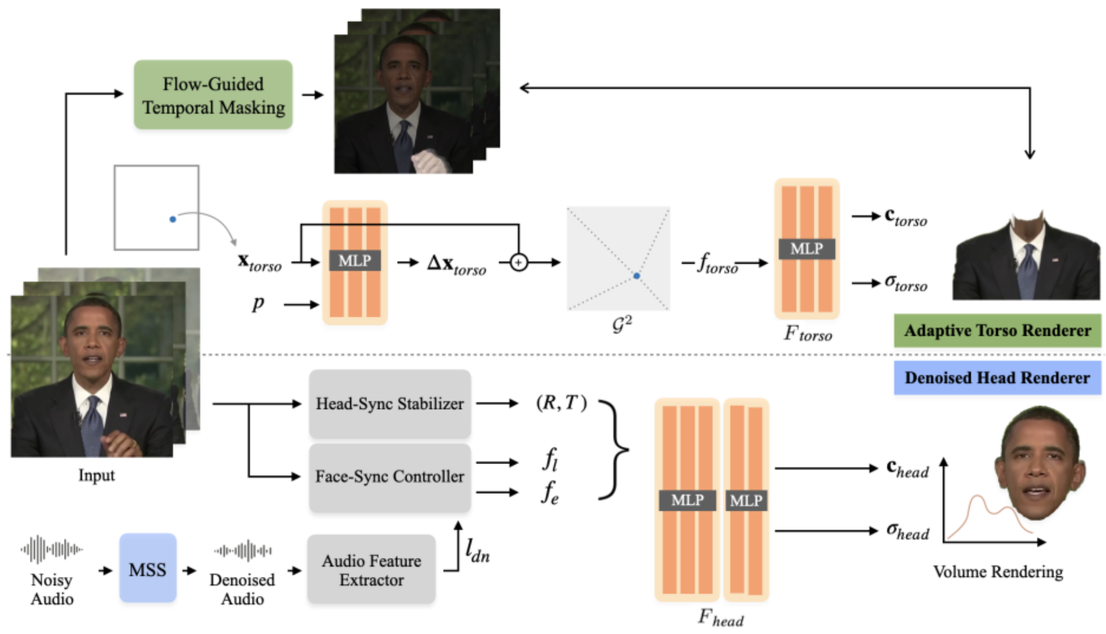
$$\hat{S}(k,m) = \max\left(|S(k,m)| - \alpha \cdot \text{Noise}(k,m), 0\right)$$
$$\hat{y}(n) = \frac{1}{L}\sum_{l=1}^{L}\hat{y}_l(n)$$

- Flow Guided Temporal Masking



$$M_i = \max\left(M_{i-1} \cdot \delta_i,\ \mathbf{1}(|V_i| > \theta)\right)$$
$$\delta_i = \delta_0 - \gamma \cdot \frac{1}{N}\sum_{x,y}|V_i(x,y)|$$

- Network Architecture

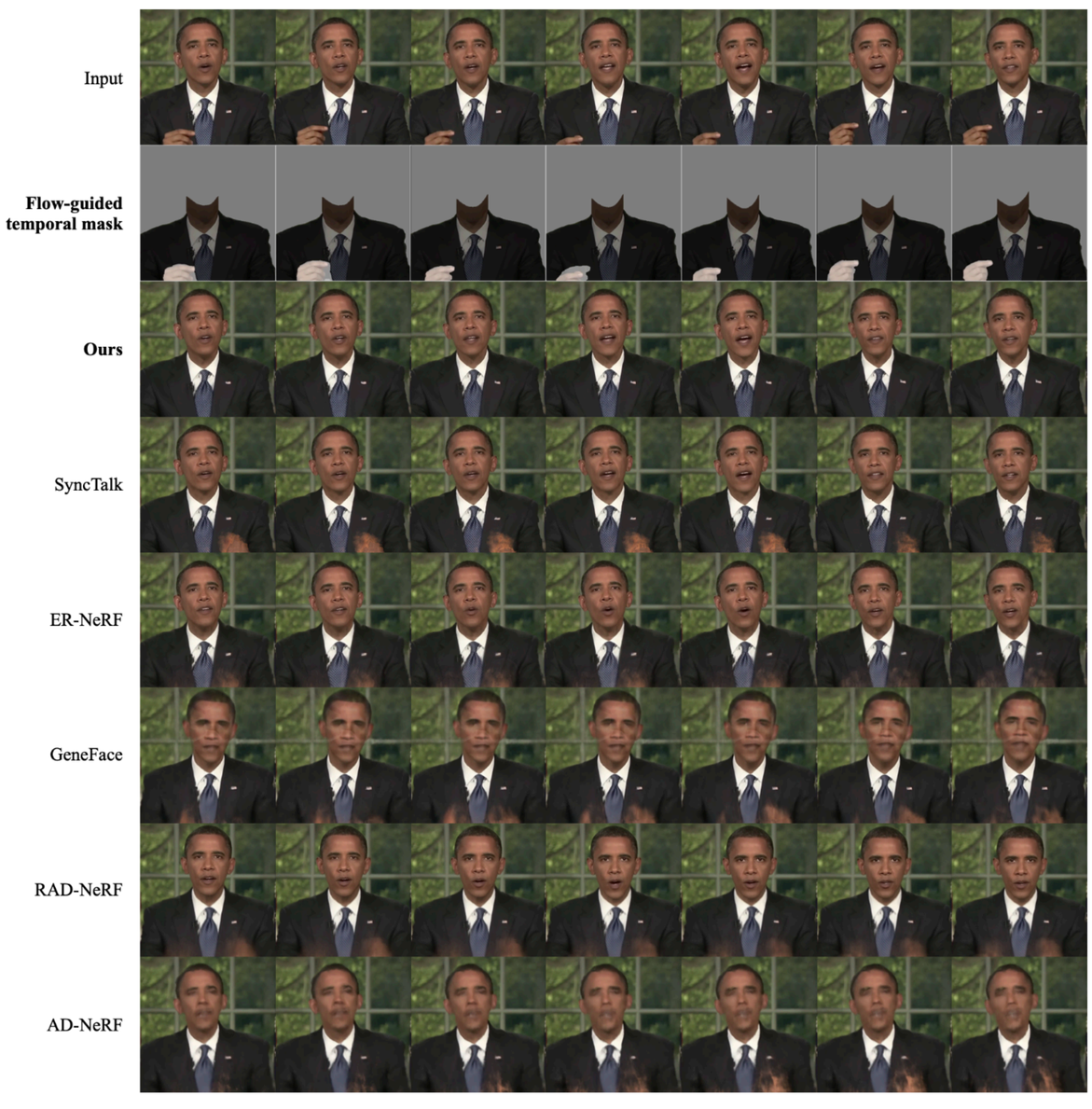


- Training Details

$$\hat{\mathcal{C}}(r) = \int_{t_n}^{t_f}\sigma(r(t))\cdot\mathbf{c}(r(t),\mathbf{d})\cdot T(t)dt$$
$$\mathcal{L}_{recon}(r) = \sum_{i\in\mathcal{I}}||\mathcal{C}_i(r) - \hat{\mathcal{C}}_i(r)||_2^2$$
$$\mathcal{L}_{head} = \mathcal{L}_{recon}(r) + \lambda\text{LPIPS}(\mathcal{P},\hat{\mathcal{P}}) \quad \text{(Head Training)}$$
$$\mathcal{L}_{torso} = M(r)\cdot\mathcal{L}_{recon}(r) \quad \text{(Torso Training)}$$

Experiements:Torso Reconstruction

- Torso Reconstruction Performance



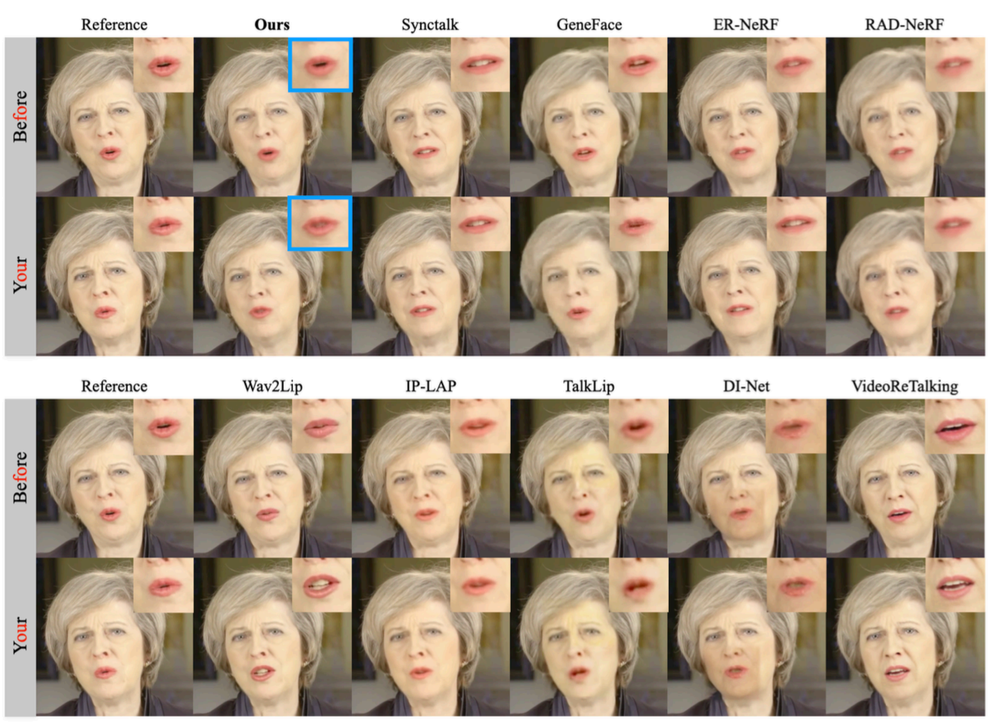
Methods	PSNR↑	LPIPS↓	LMD↓	AUE↓	LSE-C↑	LSE-D↓	Time(h)↓
AD-NeRF ICCV'21 [18]	22.742	0.228	-	0.467	0.323	12.071	36.4
RAD-NeRF arXiv'22 [36]	23.736	0.166	1.929	0.354	6.329	8.427	4.0
GeneFace ICLR'23 [43]	24.173	0.108	2.961	0.534	0.211	13.552	37.0
ER-NeRF ICCV'23 [23]	24.183	0.113	2.196	0.240	7.199	7.830	2.4
SyncTalk CVPR'24 [27]	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**

Experiements: Lip Synchronization

- Audio Denoising Performance of MSS Denoiser

Methods	PESQ↑	CSIG↑	CBAK↑	COVL↑
Noisy	1.97	3.35	2.44	2.63
SEGAN Interspeech'17 [26]	2.16	3.48	2.94	2.80
MMSE-GAN ICASSP'18 [33]	2.53	3.80	3.12	3.14
Metric-GAN ICML'19 [15]	2.86	3.99	3.18	3.42
HiFi-GAN NeurIPS'20 [22]	2.94	4.07	3.07	3.49
DEMUCS ICASSP'23 [29]	3.07	4.31	3.40	3.63
MetricGAN+ Interspeech'21 [16]	3.15	4.14	3.16	3.64
DPT-FSNET ICASSP'22 [10]	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

- Lip Synchronization Performance



		Original			Corrupted		
		AUE↓	LSE-C↑	LSE-D↓	AUE↓	LSE-C↑	LSE-D↓
2D	Wav2Lip ACM MM'20 [28]	**0.246**	8.447	6.241	0.313	**7.413**	**7.330**
2D	VideoReTalking SIGGRAPH Asia'22 [8]	0.270	8.066	7.075	0.290	7.014	7.953
2D	DINet AAAI'23 [45]	0.340	6.775	8.026	0.355	5.801	8.904
2D	TalkLip CVPR'23 [40]	0.300	6.219	8.378	0.304	5.325	8.502
2D	IP-LAP CVPR'23 [46]	0.294	5.571	8.975	0.298	3.845	10.632
NeRF	AD-NeRF ICCV'21 [18]	0.294	5.005	9.957	0.313	4.603	10.212
NeRF	RAD-NeRF arXiv'22 [36]	0.356	3.598	9.642	0.357	2.197	11.591
NeRF	GeneFace ICLR'23 [43]	0.266	6.876	7.076	0.318	4.587	9.243
NeRF	ER-NeRF ICCV'23 [23]	0.271	6.808	8.004	0.371	3.052	11.130
NeRF	SyncTalk CVPR'24 [27]	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