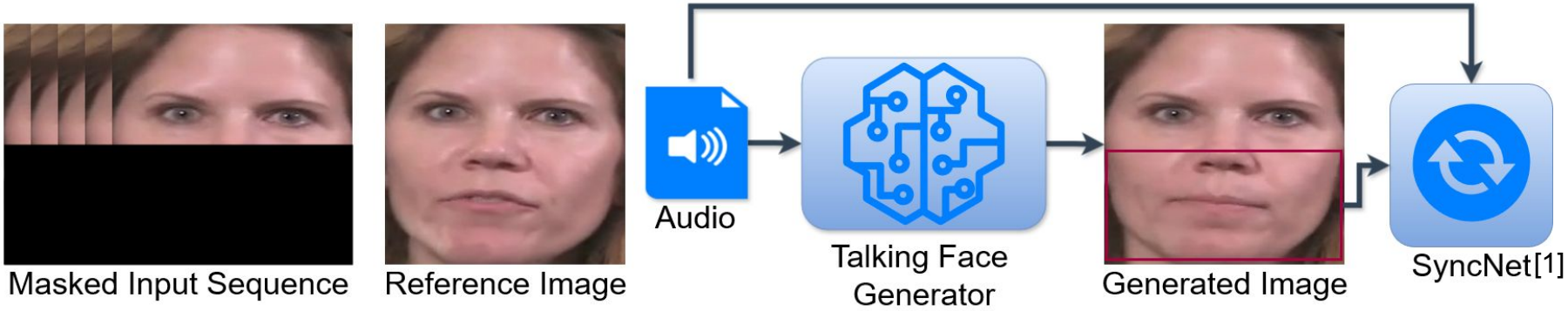
Audio-driven Talking Face Generation with Stabilized Synchronization Loss

Dogucan Yaman[1], Fevziye Irem Eyiokur[1], Leonard Bärmann[1],
Hazim Kemal Ekenel[2], Alexander Waibel[1,3]

[1]Karlsruhe Institute of Technology, [2]Istanbul Technical University, [3]Carnegie Mellon University

Introduction

- 2D audio-driven talking face generation (a.k.a. face dubbing)

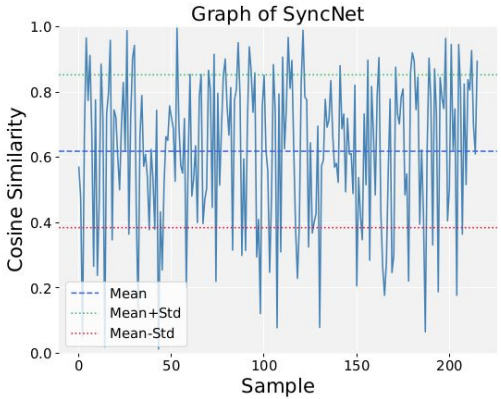


[1] Prajwal, K. R., et al. "A lip sync expert is all you need for speech to lip generation in the wild." *Proceedings of the 28th ACM international conference on multimedia*. 2020. (Modified SyncNet)

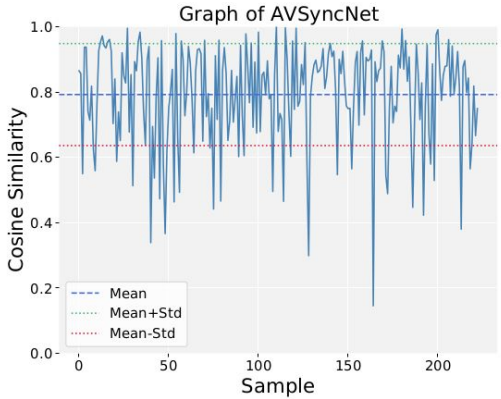
Original SyncNet: Chung, Joon Son, and Andrew Zisserman. "Out of time: automated lip sync in the wild." *Computer Vision–ACCV 2016 Workshops: ACCV 2016 International Workshops, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II 13*. Springer International Publishing, 2017.

Introduction

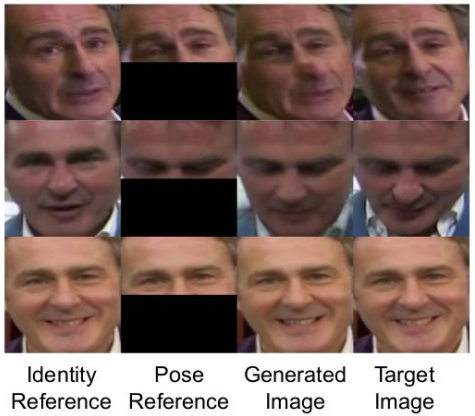
- **Problems:**
  - SyncNet instability
  - Lip leaking from ID ref.
  - Unstable training



(a) SyncNet



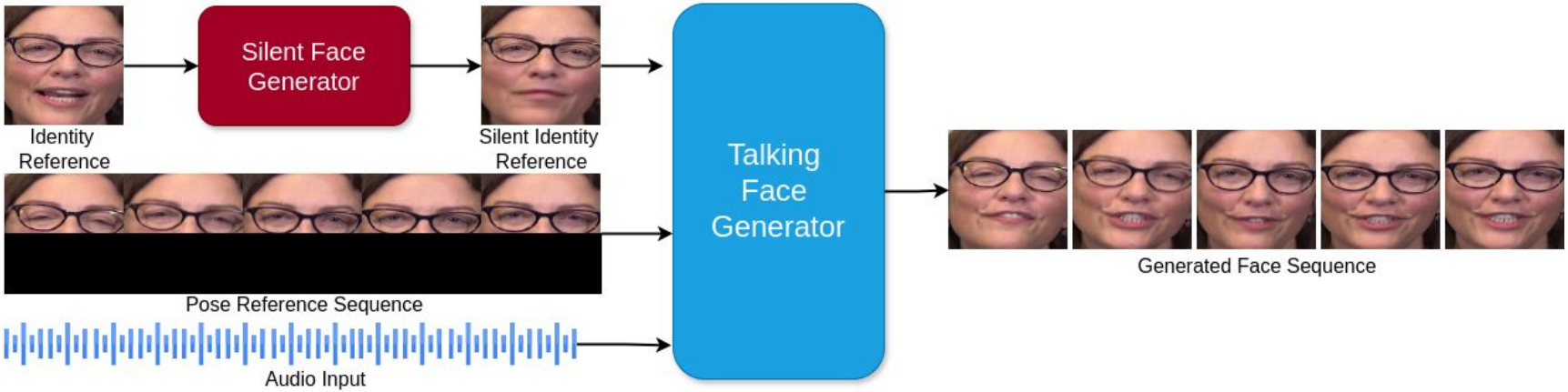
(b) AVSyncNet (ours)



(c) Leaking problems

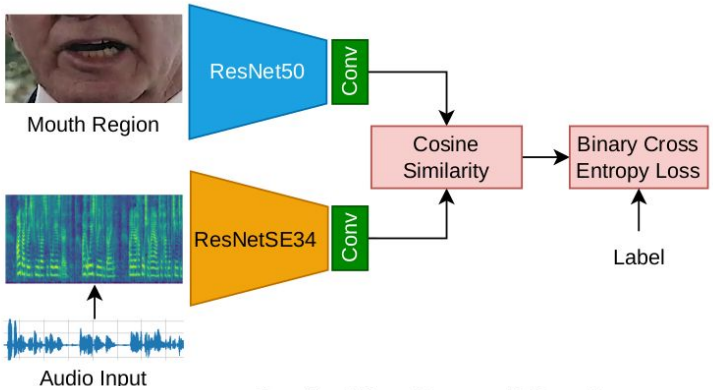
Contributions

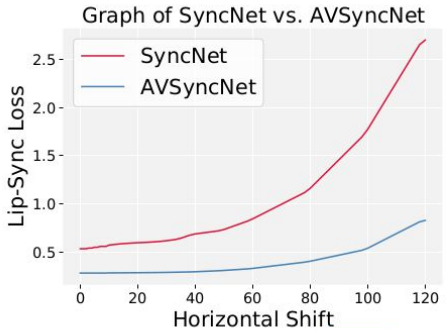
- **AVSyncNet:** Robust and shift-invariant version of SyncNet.
- **Stabilized synchronization loss:** Relative distance to alleviate AVSyncNet instability further.
- **Silent-lip generator:** Modify lips of the identity reference to mitigate lip leaking
- Identifying and analyse fundamental issues that harm lip-sync & visual quality



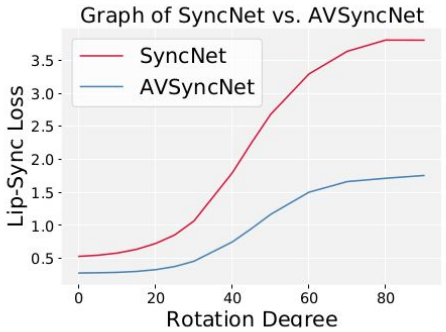
AVSyncNet

- Similar learning strategy with SyncNet.
- More robust
- Shift-invariant

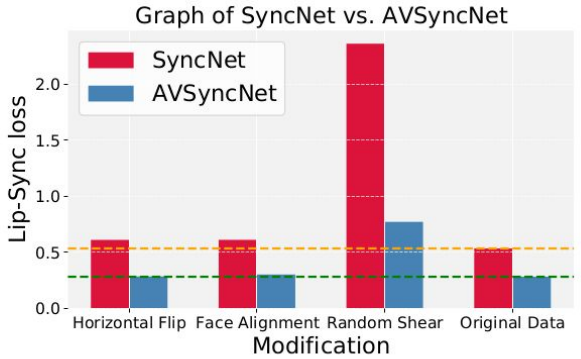




(a) Shifting

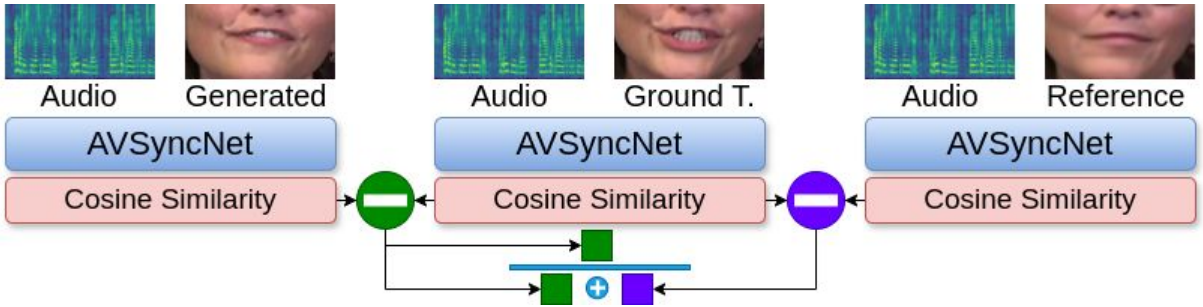


(b) Rotation.



(c) Further analysis.

Stabilized Synchronization Loss



$$P_{sync} = \frac{F_I \cdot F_A}{max(||F_I||_2 \cdot ||F_A||_2, \epsilon)}$$

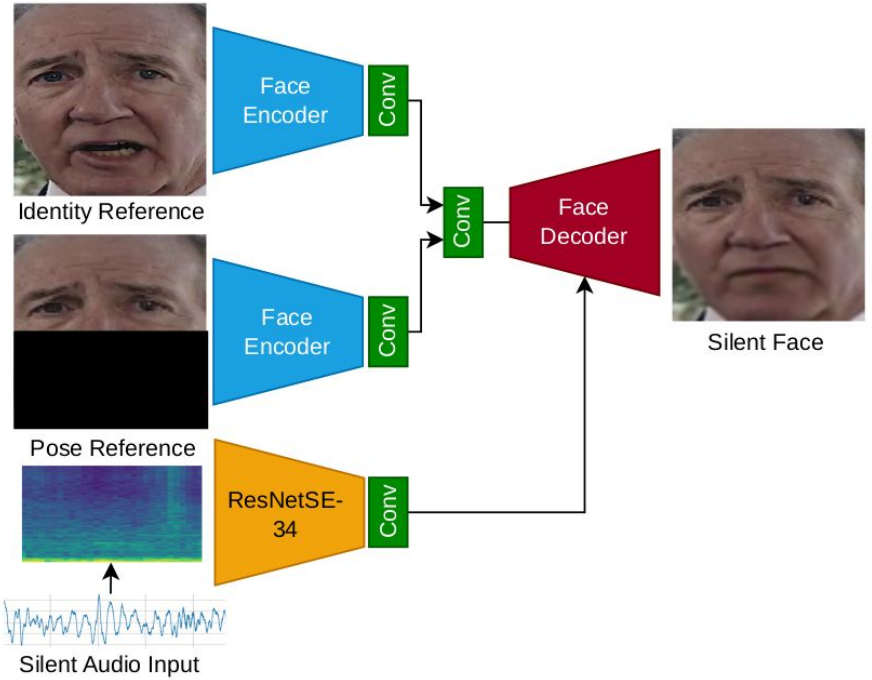
$$L_{sync} = \frac{1}{N} \sum_{1}^{N} -log(P_{sync}^{i})$$

$$L_{ss} = -\log\left(1 - \frac{|x-y| + \epsilon}{|x-y| + |y-d| + \epsilon}\right)$$

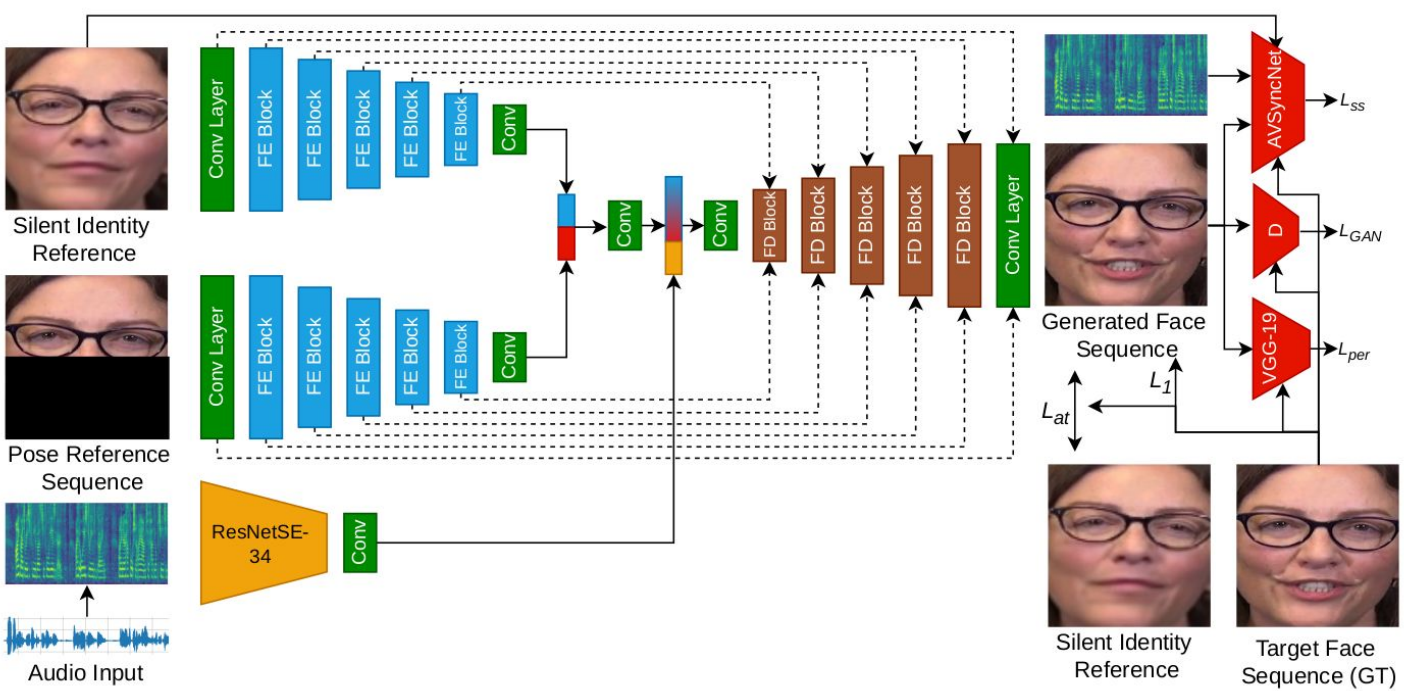
$$x = \text{AVSIM}\left(I',\ A\right), \quad y = \text{AVSIM}\left(I^{GT},\ A\right), \quad d = \text{AVSIM}\left(I^{R},\ A\right)$$

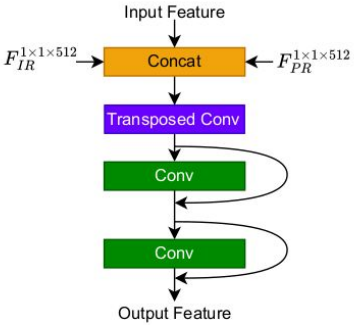
Silent-Lip Generation

- Implicit learning
- Like talking face generation
- Without synchronization loss
- **Inference:** Silent audio as input



Talking Face Generation





Quantitative Results

	LRS2							LRW						
Method	SSIM ↑	PSNR ↑	FID ↓	IFC ↓	LMD ↓	LSE-C ↑	LSE-D ↓	SSIM ↑	PSNR ↑	FID ↓	IFC ↓	LMD ↓	LSE-C ↑	LSE-D ↓
Wav2Lip	0.86	26.53	7.05	0.21	2.38	7.59	6.75	0.85	25.14	6.81	0.20	2.14	7.49	6.51
PC-AVS	0.73	28.24	18.40	0.46	1.93	6.41	7.52	0.81	32.25	14.27	0.38	1.42	6.53	7.15
VideoReTalking w/ FR	0.84	25.58	9.28	0.22	2.61	7.49	6.82	0.87	27.11	5.30	0.23	2.39	6.59	7.12
DINet	0.78	24.35	4.26	0.25	2.30	5.37	8.37	0.88	27.50	8.17	0.22	1.96	5.24	9.09
TalkLip	0.86	26.11	4.94	0.24	2.34	8.53	6.08	0.86	26.34	15.73	0.26	1.83	7.28	6.48
IPLAP	0.87	29.67	4.10	0.20	2.11	6.49	7.16	0.91	30.45	8.40	0.21	1.64	5.94	7.76
Ours w/o FR	0.95	32.64	3.83	0.16	1.13	8.41	6.03	0.92	31.45	4.46	0.18	1.22	7.86	6.24
Ours w/ FR (VQFR)	0.90	31.80	5.23	0.27	1.36	8.52	5.83	0.90	30.21	7.05	0.21	1.41	7.92	6.00

Conclusion

- Identified and analysed fundamental issues.
- Improved audio-driven talking face generation.
- Silent-lip generator to alleviate lip leaking
- AVSyncNet to improve lip-sync
- Stabilized synchronization loss to improve the lip-sync further
- SOTA results in most of the metrics

Limitations:

- AVSyncNet's unstable nature must be investigated further.
- Teeth are invisible in the identity reference due to the silent face generator, causes suboptimal teeth generation.

Thank You! - Questions?

Paper

Acknowledgement: This work was supported in part by the European Commission project Meetween (101135798) under the call HORIZON-CL4-2023-HUMAN-01-03

https://yamand16.github.io/TalkingFaceGeneration/

Webpage

Qualitative Results

Ablation Study

Ablation	Setup	Method	PSNR ↑	SSIM ↑	FID ↓	LMD ↓	LSE-C ↑	LSE-D ↓	IFC ↓
Components	A	$G_L + \mathcal{L}_s$	26.349	0.853	12.25	2.408	7.116 ± 1.92	7.396 ± 1.03	0.221
	B	A + $E_{a,S}$	26.614	0.868	9.82	2.325	7.271 ± 1.76	7.106 ± 0.98	0.223
	C	A + $E_{a,W}$	26.590	0.869	10.56	2.278	7.220 ± 1.75	7.158 ±0.99	0.228
	D	B + G_S	27.180	0.872	8.16	1.741	7.752 ± 1.71	6.413 ± 0.95	0.221
	E	$G_L + E_{a,S} + G_S + \mathcal{L}_{ss}$	31.166	0.925	5.27	1.140	8.370 ± 1.16	**6.032** ± 0.59	0.174
	F	E + $\mathcal{L}_t$	30.658	0.917	6.24	1.250	8.260 ± 1.34	6.176 ± 0.64	0.183
	G	E + $\mathcal{L}_{at}$	**32.755**	0.949	4.02	1.135	8.382 ± 1.16	6.057 ± 0.61	0.163
	H	G w/ AVSyncNet	32.640	**0.952**	**3.83**	**1.130**	**8.410** ± 0.97	6.037 ± 0.55	**0.160**
Post-processing	FR1	Setup H + GPEN	28.991	**0.919**	58.77	**1.197**	7.625	6.457	**0.192**
	FR2	Setup H + GFPGAN	31.169	0.916	13.07	1.219	7.624	6.496	0.214
	FR3	Setup H + VQFR: full model	**31.806**	0.905	**5.23**	1.365	**8.528**	**5.838**	0.278
Silent face generation	VRT-S	VideoReTalking silent data	22.124	0.646	33.60	-	-	-	0.463
	Ours-S	Our silent data (G_S)	**33.328**	**0.951**	**4.41**	-	-	-	**0.141**