Semantic-Aware Implicit Neural Audio-Driven Video Portrait Generation

Supplementary Material

Xian Liu[1], Yinghao Xu[1], Qianyi Wu[2], Hang Zhou[1], Wayne Wu[3], and Bolei Zhou[1]

[1] Multimedia Laboratory, The Chinese University of Hong Kong
[2] Monash University
[3] SenseTime Research
alvinliu@ie.cuhk.edu.hk

In the supplemental document, we introduce additional quantitative comparisons with baselines (Sec. A); model architecture details (Sec. B); analysis on the generalize ability to unseen persons (Sec. C); additional qualitative results (Sec. D); user study details (Sec. E); ethical considerations with possible prevention measures against potential negative social impact (Sec. F). Note that due to the Youtube video copyright issues, we can not directly release the raw videos. The Youtube link of portrait videos we used in this work are recorded in Sec. G. The licenses of existing assets involved in this paper are listed in Sec. H.

A Additional Quantitative Comparisons

We conduct additional quantitative evaluations on the Testset A in terms of: **CSIM**, which firstly extracts the embedding vectors from generated results through a pretrained face recognition model [5], then measures the cosine similarity to account for the identity preserving performance; **CPBD**, which measures the sharpness of synthesized images. The results are reported in Table 1. We can see that both our method and AD-NeRF [7] achieve high performance on two metrics. This mainly derives from the ability of implicit function to generate full-resolution images of higher quality, while the explicit model-based methods [3,13,18,17] generate comparatively low-quality images. Such low-fidelity generation makes results blurry and difficult to preserve the speaker's identity.

B Architecture Details

Deformation Implicit Function. The deformation implicit function F_{Φ}^{deform} takes 3D spatial coordinate $\mathbf{x}$, time t, head pose $\mathbf{p}_h(t)$ and canonical pose $\mathbf{p}_c$ as input, and outputs the 3D coordinate's displacement $\Delta\mathbf{x} = (\Delta x, \Delta y, \Delta z)$. Note that both $\mathbf{x}$ and t are firstly processed by positional encoding of function $\gamma(q) =< (\sin(2^l\pi q), \cos(2^l\pi q)) >_0^L$, where we use $L = 10$ for $\mathbf{x}$, and $L = 4$ for t. Then all the inputs are concatenated together and fed into an 8-layer MLP of hidden size 128 and ReLU activation.

Table 1. Additional quantitative comparisons on Testset A. We compare CSIM and CPBD metrics to baselines [3,13,18,17,7].

Methods	CSIM ↑	CPBD ↑
ATVG [3]	0.907	0.109
Wav2Lip [13]	0.963	0.172
MakeitTalk [18]	0.958	0.179
PC-AVS [17]	0.912	0.182
AD-NeRF [7]	**0.989**	0.188
SSP-NeRF (Ours)	**0.990**	**0.191**

Table 2. Architecture of SparseConvNet adapted from [12].

	Layer Description	Output Dim.
	Input volume	D×H×W×4
1-2	$(3 \times 3 \times 3$ conv, 4 features, stride 1) $\times$ 2	D×H×W×4
3	$3 \times 3 \times 3$ conv, 8 features, stride 2	½D×½H×½W×8
4-5	$(3 \times 3 \times 3$ conv, 8 features, stride 1) $\times$ 2	½D×½H×½W×8
6	$3 \times 3 \times 3$ conv, 16 features, stride 2	¼D×¼H×¼W×16
7-9	$(3 \times 3 \times 3$ conv, 16 features, stride 1) $\times$ 3	¼D×¼H×¼W×16
10	$3 \times 3 \times 3$ conv, 32 features, stride 2	⅛D×⅛H×⅛W×32
11-13	$(3 \times 3 \times 3$ conv, 32 features, stride 1) $\times$ 3	⅛D×⅛H×⅛W×32
14	$3 \times 3 \times 3$ conv, 32 features, stride 2	1/16D×1/16H×1/16W×32
15-17	$(3 \times 3 \times 3$ conv, 32 features, stride 1) $\times$ 3	1/16D×1/16H×1/16W×32

Semantic-Aware Implicit Function. The semantic-aware implicit function $F_{\Theta}^{\text{semantic}}$ takes the *deformed* 3D coordinate $\mathbf{x} + \Delta\mathbf{x}$, view direction $\mathbf{d}$, audio feature $\mathbf{a}$ and 3D latent code $\mathbf{f}$ as input, then outputs the semantic logits $\mathbf{s}$, density σ and color $\mathbf{c}$ of the scene. Similarly, positional encoding is applied to 3D coordinate and view direction by $L = 10$ for $\mathbf{x}$, and $L = 4$ for $\mathbf{d}$. The network also consists of an 8-layer MLP of hidden size 128 and ReLU activation. Notably, since the semantic information of the scene is view-invariant and we use totally 11 semantic categories in this work, an additional FC layer (128, 11) is used to output semantic logits, *i.e.*, the semantic output is derived from the same branch as volume density σ due to the view-invariant reason, just with one more FC layer.

Structured 3D Encoder. The extraction of 3D latent code $\mathbf{f}$ follows below steps: **(1)** We first estimate a rough 3DMM [2] mesh with mean expression parameters to only provide rough facial geometry information. **(2)** We assign a latent code of dimension 4 to each mesh point. Specifically, the 3DMM mesh we use in this work contain totally 34650 points. An embedding layer of (34650, 4) is leveraged to assign latent code information. **(3)** Since the structured latent codes are sparse in the 3D scene, we have to find an effective way to process them. In particular, we follow [12] to compute the 3D bounding box of 3DMM mesh and then divide into voxels with size of $5mm \times 5mm \times 5mm$. The latent code of a

Fig. 1. Additional Qualitative Results. We visualize some additional rendered results to show the robustness and superiority of our method.

non-empty voxel is assigned to the mean of latent codes inside itself. Afterwards, the SparseConvNet adapted from Peng *et al.* [12] is utilized to extract the 3D feature volume as well as diffuse the 3D information to the neighborhood regions in the 3D scene, with details in Table 2. **(4)** Finally, when we input a specific 3D coordinate $\mathbf{x}$ to the implicit function, we can query its latent code $\mathbf{f}$ by trilinear interpolation in the 3D feature volume derived from last step.

C Analysis on the Generalize Ability to Unseen Persons

1) Since speaker's appearance and motion are highly-entangled in the implicit function, NeRF-based model is identity-specific. Decoupling portrait appearance and movements remains unsolved in current setting. **2)** The problem is strongly correlated with NeRF generalization in static scenes, which is not perfectly solved. It would be more difficult in the dynamic speaking portrait scene. **3)** The balance between person-agnostic and person-specific methods has long been discussed. While person-agnostic methods enjoy strong generalization ability, they require more training clips and leads to poor visual quality. On the other hand, our identity-specific method generates much more realistic results on shorter training clips. The ultimate goal is general person-agnostic result with high quality.

D Additional Qualitative Results

In Fig. 1, we show some additional rendered results. Notably, among the five speakers, some people have longer and more non-rigid hair, which is more difficult to learn. Our method manages to generate realistic speaking portrait results, which can demonstrate the robustness and superiority of our proposed approach.

E User Study Details

The study involves 18 participants, including 9 females and 9 males with age range of 20-30 years old. Specifically, the users are unaware of which generated result corresponds to which method for fair comparison. The participants are asked to judge the three perspectives of the generated portraits: (1) *Lip-sync*

Accuracy; (2) *Video Realness*; (3) *Image Quality*. Before they rate the quality of synthesized results, we will first show them the Ground Truth (original raw video) to help them make more accurate judgement. Note that *Image Quality* refers to whether the image is blurry, while *Video Realness* focuses on whether the presented portrait looks smooth and real.

F Ethical Considerations

Our method could animate high-fidelity talking portraits, which is envisioned to facilitate extensive applications like digital human, film-making and virtual video conferences. On the other hand, however, such techniques could be misused for malicious purposes such as identity theft, media manipulation and deepfake generation. Recent studies [14,15,16,6,10,11,4] in digital media forensics have shown promising results on detecting deepfakes. However, the lack of large-corpus and realistic portrait data limits their performance and generalizeability. As part of our responsibility, we strongly advocate all safeguarding measures against malicious use of facial images and feel obliged to share our generated results with the deepfake detection community to improve the method's robustness. We believe that the proper use of this technique will enhance positive societal development on both machine learning research and human's daily entertainment. Here we list some possible measures to fight against malicious use of speaking portraits generation:

- **Build large-corpus deepfake data for detection.** Since the quantity of data for deepfake detection matters a lot for the model's robustness and generalizability. Hence we suggest to incorporate realistic talking face or speaking video portrait data into the large-corpus dataset to improve the performance of detection algorithm. Besides, such data should be carefully managed to avoid from negative impact and ought to be used for non-profit purpose.
- **Legislation.** Despite the enormous potential harms of deepfakes, few laws have been proposed to regulate the use of this technique. Therefore, we advocate that the laws which explicitly state how to legally make use of talking face techniques should be established.
- **Make the public be aware of how to fight against malicious use of deepfake.** The harm of malicious use of deepfakes often enlarges when public unconsciously spread them on the social media. They may have no malicious purpose, but unconsciously cause detrimental influences. So we argue that we should make the public be aware of how to legally and positively use this technique to enrich their entertainments.

G Dataset Details

Due to the copyright issues, we can not directly share the raw video files. Hence we provide the Youtube link list of all portraits videos involved in our paper in Table 3.

Table 3. Youtube links of the portrait videos used in our paper.

Videos	Portrait Video Youtube Link
Biden	https://www.youtube.com/watch?v=BR17SwEbp1o
Cameron-clip1	https://www.youtube.com/watch?v=q-g_DECdl3M
Cameron-clip2	https://www.youtube.com/watch?v=G1qo1gd4klc
Cameron-clip3	https://www.youtube.com/watch?v=9wZTd0fWjBA
Leonardo	https://www.youtube.com/watch?v=QWYCtbho-vY
Macron-clip1	https://www.youtube.com/watch?v=GI8kwPe6Uek
Macron-clip2	https://www.youtube.com/watch?v=uAx0tkejzbg
May	https://www.youtube.com/watch?v=nOj49CzODEU
Obama-clip1	https://www.youtube.com/watch?v=Gh76oepKFc8
Obama-clip2	https://www.youtube.com/watch?v=koy-KasFhFM
Obama-clip3	https://www.youtube.com/watch?v=zxk9mxURChw

Table 4. Licenses of existing assets we have used in this work.

Asset	License Link
Face Parsing [8]	https://github.com/zllrunning/face-parsing.PyTorch/blob/master/LICENSE
NeRF [9]	https://github.com/yenchenlin/nerf-pytorch/blob/master/LICENSE
DeepSpeech [1]	https://github.com/mozilla/DeepSpeech/blob/master/LICENSE
AD-NeRF [7]	https://github.com/YudongGuo/AD-NeRF/blob/master/LICENSE

H Assets License

In the Table 4, we list the licenses of all the existing assets we have used in this work.

References

1. Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Cheng, Q., Chen, G., et al.: Deep speech 2: End-to-end speech recognition in english and mandarin. In: International conference on machine learning. pp. 173–182. PMLR (2016)
2. Blanz, V., Vetter, T.: A morphable model for the synthesis of 3d faces. In: Proceedings of the 26th annual conference on Computer graphics and interactive techniques. pp. 187–194 (1999)
3. Chen, L., Maddox, R.K., Duan, Z., Xu, C.: Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In: CVPR (2019)
4. Cozzolino, D., Rossler, A., Thies, J., Nießner, M., Verdoliva, L.: Id-reveal: Identity-aware deepfake video detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15108–15117 (2021)
5. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4690–4699 (2019)
6. Dolhansky, B., Howes, R., Pflaum, B., Baram, N., Ferrer, C.C.: The deepfake detection challenge (dfdc) preview dataset. arXiv preprint arXiv:1910.08854 (2019)

7. Guo, Y., Chen, K., Liang, S., Liu, Y., Bao, H., Zhang, J.: Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In: IEEE/CVF International Conference on Computer Vision (ICCV) (2021)
8. Lee, C.H., Liu, Z., Wu, L., Luo, P.: MaskGAN: Towards diverse and interactive facial image manipulation. In: CVPR (2020)
9. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: European conference on computer vision. pp. 405–421. Springer (2020)
10. Nguyen, H.H., Fang, F., Yamagishi, J., Echizen, I.: Multi-task learning for detecting and segmenting manipulated facial images and videos. arXiv preprint arXiv:1906.06876 (2019)
11. Nirkin, Y., Wolf, L., Keller, Y., Hassner, T.: Deepfake detection based on discrepancies between faces and their context. IEEE Transactions on Pattern Analysis and Machine Intelligence (2021)
12. Peng, S., Zhang, Y., Xu, Y., Wang, Q., Shuai, Q., Bao, H., Zhou, X.: Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9054–9063 (2021)
13. Prajwal, K., Mukhopadhyay, R., Namboodiri, V.P., Jawahar, C.: 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. pp. 484–492 (2020)
14. Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Nießner, M.: Faceforensics: A large-scale video dataset for forgery detection in human faces. arXiv preprint arXiv:1803.09179 (2018)
15. Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Nießner, M.: Faceforensics++: Learning to detect manipulated facial images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1–11 (2019)
16. Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., Ortega-Garcia, J.: Deepfakes and beyond: A survey of face manipulation and fake detection. Information Fusion **64**, 131–148 (2020)
17. Zhou, H., Sun, Y., Wu, W., Loy, C.C., Wang, X., Liu, Z.: Pose-controllable talking face generation by implicitly modularized audio-visual representation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4176–4186 (2021)
18. Zhou, Y., Han, X., Shechtman, E., Echevarria, J., Kalogerakis, E., Li, D.: Makelttalk: speaker-aware talking-head animation. ACM Transactions on Graphics (TOG) **39**(6), 1–15 (2020)