High-fidelity Generalized Emotional Talking Face Generation with Multi-modal Emotion Space Learning

Chao Xu[1] Junwei Zhu[2] Jiangning Zhang[2] Yue Han[1] Wenqing Chu[2]
Ying Tai[2] Chengjie Wang[2,4*] Zhifeng Xie[3] Yong Liu[1*]
[1] APRIL Lab, Zhejiang University [2]Youtu Lab, Tencent [3]Shanghai University [4]Shanghai Jiao Tong University
{21832066, 22132041}@zju.edu.cn, yongliu@iipc.zju.edu.cn, wqchu16@gmail.com
{junweizhu, vtzhang, yingtai, jasoncjwang}@tencent.com, zhifeng_xie@shu.edu.cn

We supplement the following contents, which are not presented in the paper due to space limitations:

- Details of model architecture
- Additional experiments
- Limitations
- Societal impacts

1. Details of Model Architecture

1.1. CLIP Audio Encoder

We adopt the Transformer as our CLIP audio encoder, which could capture short- and long-term information, and adaptively focus on the emotion-related timestamps to precisely disentangle emotion style from entangled audio clips. In practice, we prepend the `Cls` token as ViT [2] for pool purpose, which outputs the emotion code e_{audio} after Transformer encoding and a liner layer for dimensional projection. The layer number is 4, and the head number is 8, and the latent feature dimension is 256. The framework is depicted in Fig. 1(a).

1.2. Transformer Φ in EAC

We use the Transformer $\mathbf{\Phi}$ with 8 layers, 8 heads, and 256 latent feature dimensions to project emotion-audio correlated inputs to expression coefficients.

1.3. StyConv in HEF

Fig. 1(b) shows details of the StyConv in HEF block. The definitions of `A`, `Mod`, and `Demod` are the learned affine transform, modulation, and demodulation layers, respectively. For the details, please refer to StyleGAN2 [5].

*Corresponding authors

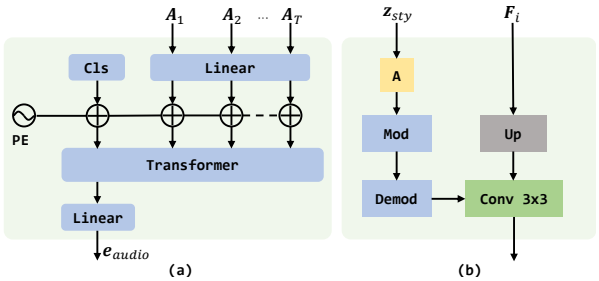

Figure 1. (a) The architecture of the CLIP audio encoder. (b) The architecture of the StyConv in HEF block.

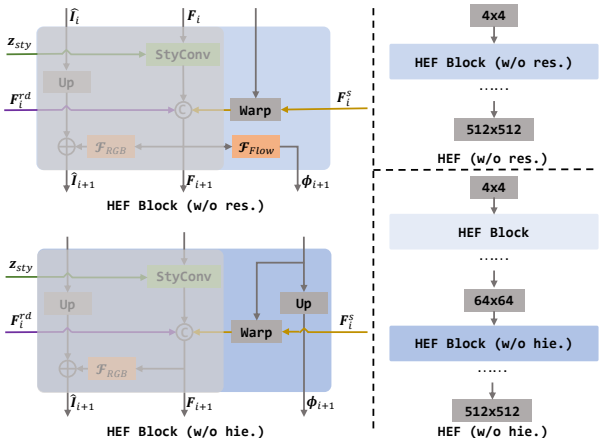

Figure 2. The architecture of HEF w/o res. and w/o hie., and corresponding block architecture. We highlight the modification of the flow operation.

1.4. Two Variants in Ablation Study

In Fig. 2, we show the architectures of two HEF variants used in Sec. 4.4. Please pay attention to the flow operation. For the original HEF block, please refer to the main manuscript.

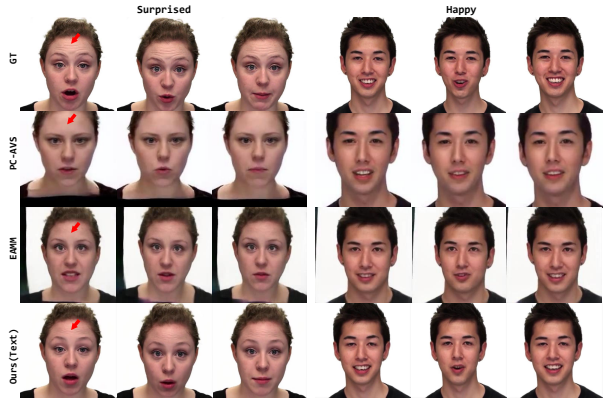

Figure 3. Qualitative results on RAVDESS dataset. Different columns mean several sampled timestamps.

2. Additional Experiments

2.1. The Comparisons on RAVDESS

We perform comparisons with a common audio-driven method, PC-AVS [9], and the one considering emotion style, EAMM [3], on RAVDESS [7]. This database contains 24 identities (12 female, 12 male) with neutral, calm, happy, sad, angry, fearful, disgust, and surprised expressions at two intensity levels. In this experiment, we use all identities for testing. As shown in Fig. 3, PC-AVS and EAMM receive aligned faces as input, while our method is not sensitive to the crop manner. It can be seen that our results achieve better image quality, more precise emotion and lip movements than the two competitors. For example, the results of the female could generate forehead wrinkles when under the surprised emotion, which is more faithful to the ground truth. Besides, we can draw the same conclusions from the Tab. 1.

2.2. User Study on MEAD

We conduct a user study to evaluate the performance of three emotional talking face generation methods with officially released codes, *i.e.*, MEAD [8], EVP [4], and EAMM [3]. Following the GC-AVT [6], the participants are required to evaluate the given videos from four dimensions: 1) Lip synchronization; 2) Expression accuracy and realness; 3) Overall video quality; 4) Identity consistency with the source face, and rate scores from 1 to 5. The results reported in Tab. 2 are based on the answers from 15 users with the given 50 videos of M003, showing that our method significantly surpassed the other methods.

Besides, among recent SOTA methods with released codes, only MEAD describes how to handle compound styles, and EAMM can generalize to arbitrary identities. Thus we add a user study compared with MEAD to evaluate the *unseen style generalization* on 10 compound emotions of the subject M003, and another one compared with EAMM to evaluate the *unseen identity generalization* on 10 subjects from VoxCeleb2 [1] repeated by 5 basic emotions. Our method achieves better performance than two competitors, *i.e.*, *4.76 vs. 2.19* in unseen style accuracy, and *4.53 vs. 3.57* in unseen identity consistency.

Furthermore, we report the evaluation on temporal consistency. Our method achieves higher score than PC-AVS [9] and EAMM due to the accurate flow estimation, *i.e.*, *4.54 vs. 3.27 vs. 3.80*.

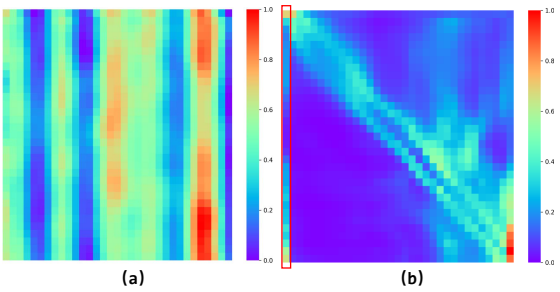

Figure 4. (a) Attention visualization of the CLIP audio encoder. (b) Attention visualization of the Transformer $\mathbf{\Phi}$.

Method	EFD ↓	LMD ↓	Sync ↑	CSIM ↑	FID ↓	PSNR ↑	SSIM ↑
PC-AVS	0.127	2.90	3.01	0.78	37.83	27.98	0.66
EAMM	0.108	2.67	3.09	0.69	40.57	28.12	0.71
Ours-T	**0.088**	**2.49**	**3.23**	**0.82**	**17.06**	**29.19**	**0.83**

Table 1. Quantitative comparison on RAVDESS dataset. Ours-T means taking text as emotion condition.

Method	Lip Sync. ↑	Emo Acc. ↑	Video Quality ↑	ID Cons. ↑
MEAD	3.67	4.11	3.09	4.27
EVP	3.78	3.84	3.94	3.12
EAMM	3.83	4.09	3.25	3.34
Ours	**4.34**	**4.52**	**4.28**	**4.69**

Table 2. The results of user study on MEAD dataset.

2.3. Visualization of Attention Weights

Fig. 4 visualizes the average attention weights across all heads of the CLIP audio encoder and Transformer $\mathbf{\Phi}$. We observe that the former focuses on some emotion-related timestamps to obtain precise style code, while the latter focuses on the local cues, which are more likely to influence the current expression. The relatively high weights of the first column in (b) indicate the prepended token effectively models the emotion intensity.

3. Limitations

Since the 3DMM does not cover the tooth region and our HEF is only trained on the MEAD dataset, the texture generator lacks rich facial priors to generate realistic mouth regions. Besides, like previous works [3, 6], although EAC uses the Transformer to model temporal information, HEF is an image-based texture generator, which inevitably intro-

duces temporal inconsistency. These will be the focus of our future research.

4. Societal Impacts

The development of face generation has attracted widespread attention and is used for many ethical and reasonable uses, such as films and computer games. However, like any technology, it can be used for good or be abused. Our intention to boost the performance of emotional talking face generation is to maximize its influence as a tool for learning and experimenting. We will have zero tolerance for anyone using our work for unethical purposes and actively discourage any such use.

References

[1] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622*, 2018. 2

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1

[3] Xinya Ji, Hang Zhou, Kaisiyuan Wang, Qianyi Wu, Wayne Wu, Feng Xu, and Xun Cao. Eamm: One-shot emotional talking face via audio-based emotion-aware motion model. *arXiv preprint arXiv:2205.15278*, 2022. 2

[4] Xinya Ji, Hang Zhou, Kaisiyuan Wang, Wayne Wu, Chen Change Loy, Xun Cao, and Feng Xu. Audio-driven emotional video portraits. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14080–14089, 2021. 2

[5] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 1

[6] Borong Liang, Yan Pan, Zhizhi Guo, Hang Zhou, Zhibin Hong, Xiaoguang Han, Junyu Han, Jingtuo Liu, Errui Ding, and Jingdong Wang. Expressive talking head generation with granular audio-visual control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3387–3396, 2022. 2

[7] Steven R Livingstone and Frank A Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5):e0196391, 2018. 2

[8] Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change Loy. Mead: A large-scale audio-visual dataset for emotional talking-face generation. In *European Conference on Computer Vision*, pages 700–717. Springer, 2020. 2

[9] Hang Zhou, Yasheng Sun, Wayne Wu, Chen Change Loy, Xiaogang Wang, and Ziwei Liu. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4176–4186, 2021. 2