EAMM: One-Shot Emotional Talking Face via Audio-Based Emotion-Aware Motion Model

Xinya Ji
Nanjing University
Nanjing, China
xinya@smail.nju.edu.cn

Hang Zhou
The Chinese University of Hong Kong
Hong Kong, China
zhouhang@link.cuhk.edu.hk

Kaisiyuan Wang
University of Sydney
Sydney, Australia
kaisiyuan.wang@sydney.edu.au

Qianyi Wu
Monash University
Melbourne, Australia
qianyi.wu@monash.edu

Wayne Wu*
SenseTime Research
Shanghai, China
wuwenyan@sensetime.com

Feng Xu*
Tsinghua University
Beijing, China
xufeng2003@gmail.com

Xun Cao*
Nanjing University
Nanjing, China
caoxun@nju.edu.cn

ACM Reference Format:
Xinya Ji, Hang Zhou, Kaisiyuan Wang, Qianyi Wu, Wayne Wu, Feng Xu, and Xun Cao. 2022. EAMM: One-Shot Emotional Talking Face via Audio-Based Emotion-Aware Motion Model. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Proceedings (SIGGRAPH '22 Conference Proceedings), August 7–11, 2022, Vancouver, BC, Canada.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3528233.3530745

In this supplementary material, we provide more details about the network architecture and loss functions. More information on our experimental settings and more results of our method are also provided. We strongly recommend watching the supplementary video.

Figure 1: Emotion interpolation. We make interpolation between different emotion features and the pose is set to be static for better visualization. Please zoom in to see more details. Natural faces from Voxceleb dataset [Nagrani et al. 2020] ©*Visual Geometry Group* (CC BY).

A NETWORK ARCHITECTURE AND TRAINING DETAILS

We describe more implementation details on the network architecture used in Sec. A.1 and the training settings in Sec. A.2.

A.1 Network Architecture

Here, we present the details of the network architecture as shown in Figure 2. Note that we use the same Keypoint Detector $\boldsymbol{E}_k$, Flow Estimator $\boldsymbol{F}$ and Image Generator $\boldsymbol{G}$ as in [Siarohin et al. 2019]. Please refer to [Siarohin et al. 2019] for more details. The architecture of other networks are described below.

- *Identity encoder* $\boldsymbol{E}_I$. This network extracts identity information from the source image I. We use a number of down-sampling blocks to produce the identity feature $\mathbf{f}_I$ with the channel dimension of 512.
- *Audio encoder* $\boldsymbol{E}_a$. By taking the 28×12-dim audio features as input, we apply convolutional neural networks (CNN) followed by multi-layer perceptrons (MLP) to obtain the 256-dim audio feature $\mathbf{f}_a$.
- *Pose encoder* $\boldsymbol{E}_p$. The pose encoder $\boldsymbol{E}_p$ is composed of a 2-layer multi-layer perceptrons (MLP) that project the 6-dim pose vector into the 256-dim pose feature $\mathbf{f}_p$.
- *Decoder* $\boldsymbol{D}$. After concatenating the features extracted from three different inputs (*i.e.*, $\mathbf{f}_I$, $\mathbf{f}_a$ and $\mathbf{f}_p$), we employ a long short-term memory (LSTM) network followed by convolutional neural networks (CNN) to predict the unsupervised key-points $\boldsymbol{x}^a_{1:T}$ and jacobians $\boldsymbol{J}^a_{1:T}$. In this way, the sequential relationship between audio signals and motion representations can be better captured.
- *Emotion Extractor* $\boldsymbol{E}_e$. The network takes the emotion source frames as input and aims at extracting the disentangled emotion information. We borrow a part of the architecture from $\boldsymbol{E}_k$ in [Siarohin et al. 2019] which consists of down sampling and an Hourglass network to extract the high-level information from the input frames. The subsequent ResBlock and MLP are used to decouple the emotion feature $\mathbf{f}_e$ with the dimension of 512.
- *Displacement predictor* $\boldsymbol{P}_d$. The goal of this component is to predict emotional displacement for the motion representations based on the extracted emotion feature and the neutral key-poinsts $\boldsymbol{x}^a$ and jacobians $\boldsymbol{J}^a$. We first perform positional

*Corresponding authors

SIGGRAPH '22 Conference Proceedings, August 7–11, 2022, Vancouver, BC, Canada

ACM ISBN 978-1-4503-9337-9/22/08…$15.00
https://doi.org/10.1145/3528233.3530745

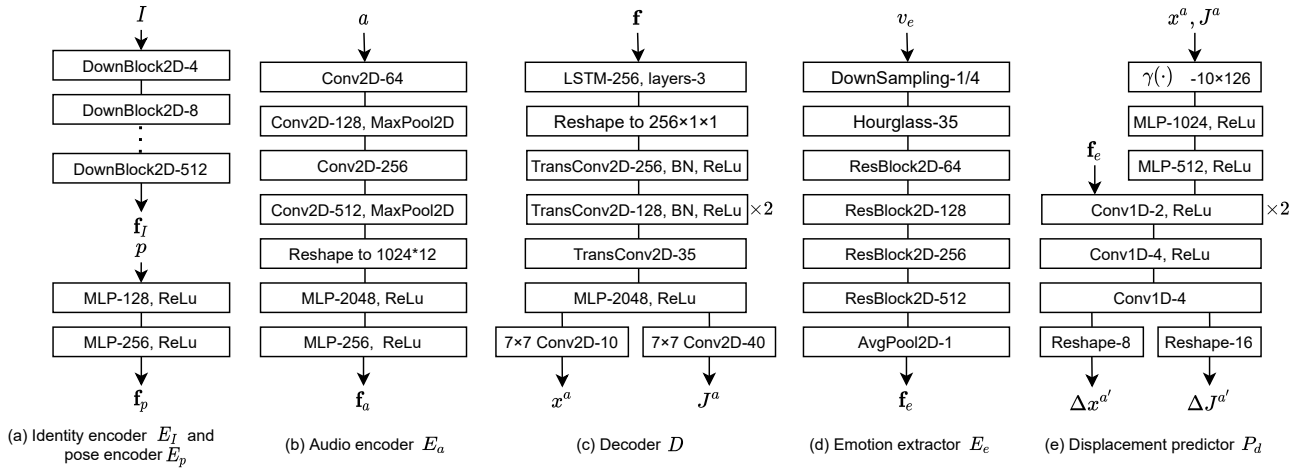

Figure 2: The network architectures of different components in our Emotion-Aware Motion Model.

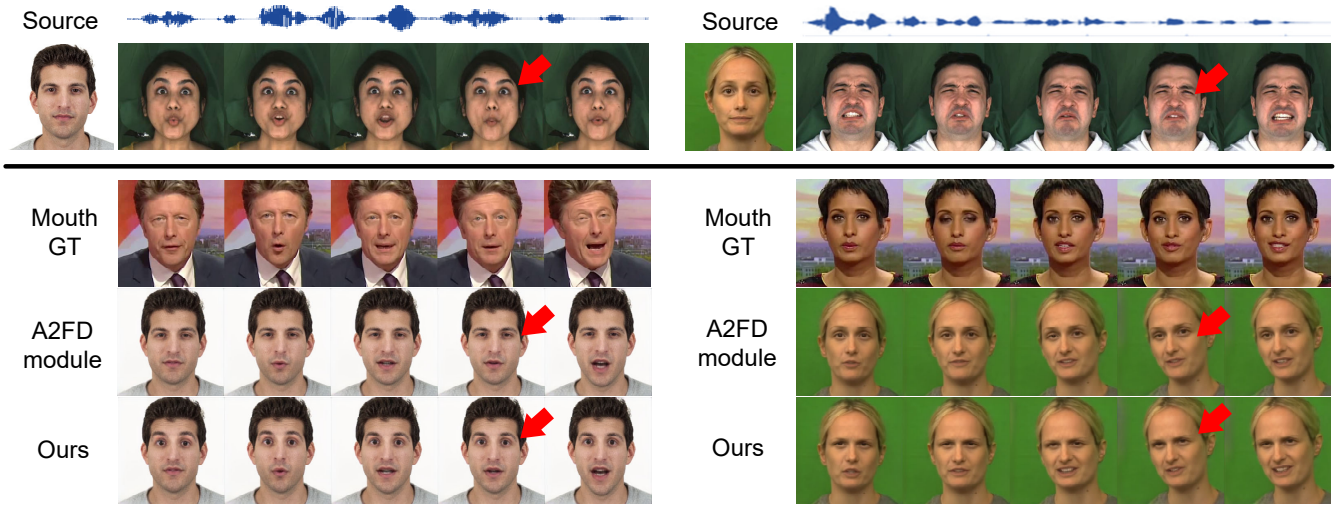

Figure 3: Results of the Audio2Facial-Dynamics Module on two cases. The top row shows the identity, speech content audio, and emotion source clips. The second row shows the corresponding frames of the audio source, i.e., ground truth for mouth shapes. Please zoom in to see more details.

encoding on the key-points and jacobians to capture high-frequency details, in which the dimension of the sinusoid is set to be 10. Then a 2-layer MLP is utilized to project the neutral embedding into a 512-dim feature. We combine the neutral feature with the emotion feature and feed them into a 4-layer MLP to obtain the final key-points $\Delta \boldsymbol{x}^{a'}$ and jacobians $\Delta \boldsymbol{J}^{a'}$ displacements.

A.2 Training Details

For the loss function of *Audio2Facial-Dynamics Module*, we follow [Siarohin et al. 2019] to add another loss term $\lambda_h \|\boldsymbol{h}_t^a - \boldsymbol{h}_t^v\|_1$ into L_{kp} in Section 3.1, where $\boldsymbol{h}_t^v$ denotes the intermediate heatmap produced by $\boldsymbol{E}_k$ from the training video clip $\mathcal{V}$ and λ_h is the weight for the heatmap loss. Experiments show that this heatmap loss is helpful for training convergence and we empirically set λ_h as 10. In addition, we set the weight of perceptual loss term λ_{per} as 0.1 during training.

We follow the self-supervised training strategy and the training procedure of our approach is performed in a progressive manner. Specifically, we first train our Audio2Facial-Dynamics module on the LRW dataset. Then we freeze this part and train our Implicit Emotion Displacement Learner with the MEAD dataset. Note that different from LRW, here we use a randomly selected image from neutral videos of the same speaker as the identity image in MEAD so that neutral motion representations can be generated from the A2DF module. All experiments are implemented on PyTorch using

Figure 4: Results of the Audio2Facial-Dynamics Module on two cases with emotional input. The upper part shows results with emotional source image input while the lower part shows results with emotional audio input. Natural videos from LRW dataset [Chung and Zisserman 2016] ©*BBC* and MEAD dataset [Wang et al. 2020] ©*SenseTime*. Natural face (top) from CREMA-D dataset [Cao et al. 2014] (ODbL). Natural face (bottom) from Voxceleb dataset [Nagrani et al. 2020] ©*Visual Geometry Group* (CC BY).

Figure 5: Comparisons with person-specific emotional talking face methods.

Adam optimizer with initial learning rate of 2×10^{-4}, which linearly decays to 2×10^{-5}. The two parts require 3 and 2 days for training on 4 NVIDIA 1080Ti GPUs, respectively.

B EXPERIMENT SETTINGS AND RESULTS

B.1 Experiment Settings

Quantitative Experiment Setting. The quantitative experiments are conducted in a self-driving setting. For the LRW dataset which has no emotion, we only use the A2FD Module to generate neutral results from the source image, the audio source and the pose source. The audio and pose source are directly taken from the test video, and the source image is randomly selected from the test video. While for the MEAD dataset with emotion, we use the full pipeline to generate emotional results, in which an additional emotion source is required. The audio of the test video is used as the audio source, while the source image is randomly selected from a neutral video of the same speaker as in the test video. However, since our method involves an additional video to control the emotion, we adopt a fair setting as in [Zhou et al. 2021] for emotion source acquisition by not using the test video as emotion source directly. Specifically, we first select another image with a different identity and then drive this image with the test video in MEAD through FOMM [Siarohin et al. 2019]. The generated emotional video shares the same facial expression and head motion but has a different identity with the test video. Thus can serve as the emotion source and pose source.

Qualitative Experiment Setting. Different from the self-driving setting in quantitative experiments, our qualitative experiments aim to evaluate our method under real scenarios. Therefore, in the qualitative experiments, we use in-the-wild data as the source images (e.g., celebrity portraits from the Internet). In order to evaluate the capability of our method, we randomly select different audio and pose sources from the LRW dataset. Note that the audio source and the pose source can come from different videos. In terms of the emotion source videos, we randomly select them from the test set in the Mead dataset.

B.2 More Experimental Results and Analysis

Results of the A2FD module. We further explore the results of the A2FD module by providing different kinds of input. We first show the results of our A2FD module in Figure 3. Our A2FD module generates talking face animation with accurate mouth motions and head movements in a neutral expression. The comparison between the results of the A2FD module and ours also demonstrates that the emotion displacements learned from the Implicit Emotion Displacement Learner are disentangled with the speech content and facial structure information of the emotion source video. More comparisons results can be seen in the accompanying video.

Considering that one may also be interested in whether the generated results of our A2FD module are generally neutralized, since the input signals (i.e., source image and speech source) can also have emotions, we also conduct an experiment by replacing either the source image or the speech audio with an emotional one to further evaluate our A2FD module. The results are shown in Figure 4. When using emotional source image, the generated frames just maintain the same facial expressions as the emotional source image (see the first row), which results in an unnatural static upper face. While no obvious changes are made to the facial expressions of synthesized results when using emotional speech audio (see the bottom row). The results demonstrate that the A2FD module mainly models the facial dynamics related to the input pose and speech content, while hardly synthesizes emotion modifications on the source image.

Notably, our objective is to achieve emotion manipulation for one-shot talking face setting, which focuses on generating emotional results based on a *neutral source image* and an *emotional audio input*. Our EAMM decomposes this problem into two subproblems, lip-sync and emotion editing, where the A2FD module is designed to tackle with the former. Therefore, we only train the A2FD module with an emotion-free dataset LRW [Chung and Zisserman 2016] (see Sec 3.3 in the main paper) in order to produce

Table 1: Comparison of identity preservation. We show cosine similarities (the maximum value is 1) of the identity features under eight emotions.

Method/Emotion	Angry	Contempt	Disgusted	Fear	Happy	Neutral	Sad	Surprised	Mean
ATVG [2019]	0.77	0.81	0.80	0.81	0.80	0.82	0.80	0.80	0.80
SDA [2018]	0.40	0.42	0.39	0.37	0.41	0.38	0.42	0.41	0.40
Wav2Lip [2020]	**0.85**	**0.85**	**0.85**	**0.85**	**0.85**	0.84	**0.85**	**0.85**	**0.85**
MakeItTalk [2020]	0.81	0.82	0.82	0.81	0.82	0.82	0.82	0.82	0.82
PC-AVS [2021]	0.57	0.51	0.59	0.58	0.58	0.57	0.53	0.59	0.56
Real	0.68	0.64	0.58	0.72	0.71	**0.87**	0.67	0.76	0.70
Ours	0.70	0.63	0.70	0.68	0.74	0.81	0.71	0.69	0.70

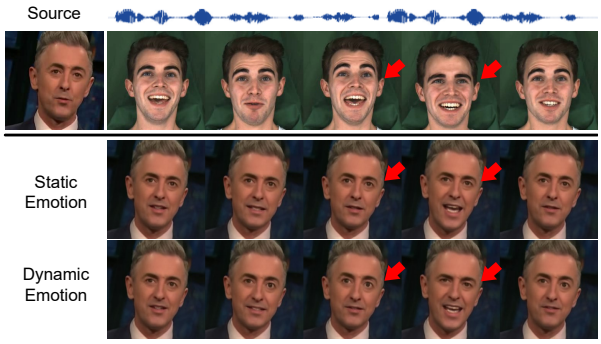

Figure 6: Results with static and dynamic emotion source. The top row shows the identity, audio and dynamic emotion source. We take the first frame in dynamic source as the static emotion source. Natural face from Voxceleb dataset [Nagrani et al. 2020] ©*Visual Geometry Group* (CC BY).

accurate mouth shapes, and that's why the A2FD module cannot generate emotional results even with emotional audio inputs.

More Comparison Results. We further make comparisons with person-specific emotional talking face methods EVP [Ji et al. 2021] and Write-a-speaker [Li et al. 2021]. When comparing with Write-a-speaker, we first crop a neutral image with the same identity as the source image for our method. Then we randomly select a video in MEAD dataset with the same emotion type as the emotion source. Since the code of Write-a-speaker has not been fully released, we can only crop an emotional video clip from the provided demo video accordingly as the counterpart. When comparing with EVP, we select a video with the same identity and emotion type as the emotion source for our method. In terms of pose source, we directly use the input video of EVP as our pose source video for fairness. The qualitative comparison results are provided in Figure 5. Our method can generate vivid emotional animation results though we lack clarity and facial details like wrinkles compared with the other two works. The reason is that they are both person-specific methods that require a long reference video of the target speaker for training, while our method only requires one reference image.

Identity Preservation. We compare the identity preservation ability of our network with other state-of-the-art methods by leveraging an off-the-shelf face recognition network [Deng et al. 2019]. The experiments are conducted on MEAD dataset. Specifically, we first use the recognition network to extract the deep identity feature for each frame and then compute the cosine similarities between the features of the generated frames and the input neutral source image. The results are illustrated in Table 1. We show the mean similarities under eight different emotions. Note that only our method can generate results with obvious emotions. Wav2Lip achieves the highest score on almost all emotions because it only edits mouth motion areas while keeping other areas unaltered. Moreover, we find that emotion can affect the identity similarities, since the value of other emotion categories are lower than neutral for real data. Considering these two factors, we claim that our method achieves comparable identity preservation with other methods since our value under neutral expressions is high and we get mean values close to the real data.

Dynamic Emotion Generation. By extracting emotion from an emotion source video, we attempt to generate emotion dynamics on neutral faces. We perform another experiment to compare the facial expressions synthesized from a static emotion source and a dynamic emotion source. Concretely, we randomly select one frame from the dynamic emotion video as the static emotion source and the comparison results are shown in Figure 6. Our method can capture the motion trends in video, e.g., the movement of eyebrows (see the red arrows), while results with static emotion input keep the same expression. Thus using an emotion video input can bring facial expression dynamics to the animation results.

Emotion Manipulation. By representing emotion in the emotion source video as a high-dimensional feature, we construct a continuous emotion latent space, where features of the same type are clustered. Thus we are able to manipulate the emotion by interpolating between the features from different emotion categories. The results are shown in Figure 1. The transformation is smooth and other factors like the head pose and identity are preserved well in the process, which indicates that our work is able to achieve emotion manipulation.

Limitations. We claim limitations of our work in Section 4.4 and here we show a simple case for each limitation in Figure 7. The

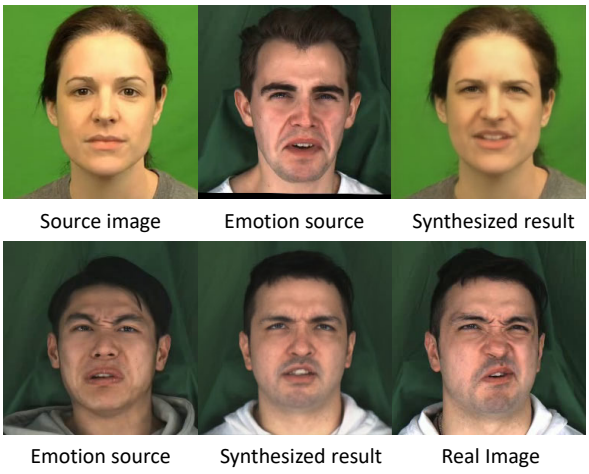

Figure 7: Two cases for limitations. Natural face from CREMA-D dataset [Cao et al. 2014] (ODbL).

first limitation is our method cannot generate satisfactory emotion dynamics related to the mouth region due to the mouth occlusion operation in the data augmentation process. As illustrated in the top row of Figure 7, the emotion traits in the mouth region, e.g. the dropping mouth corner in the synthesized result are not obvious compared with the real data. The second limitation is the emotion pattern extracted from a character sometimes seems to be unnatural after being transferred into another one. The bottom row shows the comparison between our synthesized emotion and the real emotion of the same identity. We can observe that the emotion pattern transferred from the emotion source video looks less natural when compared with the real one. Thus how to generate personalized emotion of a certain character remains an open challenge. Moreover, since we resort to a video for emotion retrieval, we have not considered the correlation between audio and facial emotion thoughtfully in our work. For example, the emotion categories of the audio and video can be different. Users have to pay more attention to the emotion type of the input to avoid the inconsistency.

REFERENCES

Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma. 2014. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing* 5, 4 (2014), 377–390.

Lele Chen, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. 2019. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7832–7841.

Joon Son Chung and Andrew Zisserman. 2016. Lip reading in the wild. In *Asian conference on computer vision*. Springer, 87–103.

Jiankang Deng, Jia Guo, Xue Niannan, and Stefanos Zafeiriou. 2019. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In *CVPR*.

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

Lincheng Li, Suzhen Wang, Zhimeng Zhang, Yu Ding, Yixing Zheng, Xin Yu, and Changjie Fan. 2021. Write-a-speaker: Text-based Emotional and Rhythmic Talking-head Generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 1911–1920.

Arsha Nagrani, Joon Son Chung, Weidi Xie, and Andrew Zisserman. 2020. Voxceleb: Large-scale speaker verification in the wild. *Computer Speech & Language* 60 (2020), 101027.

KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. 2020. 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*. 484–492.

Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. 2019. First order motion model for image animation. *Advances in Neural Information Processing Systems* 32 (2019), 7137–7147.

Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. 2018. End-to-end speech-driven facial animation with temporal gans. *arXiv preprint arXiv:1805.09313* (2018).

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

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

Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li. 2020. MakeItTalk: speaker-aware talking-head animation. *ACM Transactions on Graphics (TOG)* 39, 6 (2020), 1–15.