EMMN: Emotional Motion Memory Network for Audio-driven Emotional Talking Face Generation (Supplementary Material)

Shuai Tan, Bin Ji , and Ye Pan *

Shanghai Jiao Tong University
{tanshuai0219,bin.ji,whitneypanye}@sjtu.edu.cn

In this supplementary material, we introduce more details about Network Architecture and Training Details. Then extensive experimental results are displayed to further verify the superiority and effectiveness of our method. Lastly, we present a discussion about our method.

A. Network Architecture and Training Details

A.1. Network Architecture

In the main paper, we describe the pipeline of our method and introduce Motion Memory Net in detail. Here, we present a more detailed architecture of Motion Reconstruction and Audio2Mouth.

Motion Reconstruction. Motion Reconstruction consists of Motion Disentanglement module and Motion Integration module. Mouth Encoder E_m and Expression Encoder E_e in Motion Integration component map the keypoints (p, j) into a high dimensional space via a positional encoding operation [13], and employ convolutional neural networks (CNN) followed by multi-layer perceptrons (MLP) to extract mouth motion feature $f_m \in \mathbb{R}^{256}$ and expression motion feature $f_e \in \mathbb{R}^{256}$. Then we feed features and keypoints (p^{neu}, j^{neu}) extracted from neutral image into Motion Decoder $D_{e,m}$, which is composed of 4-layers CNN-MLP structure to combine and project the features and keypoints (p^{neu}, j^{neu}) into reconstruction facial representation $(\hat{p}, \hat{j})$.

The detailed training process is shown in Fig. 1. For each batch, we randomly select two expression (i, j) and two content (a, b) to obtain four samples (input images: $x^{i,a}$,$x^{j,b}$ and ground truth: $y^{i,b}$,$y^{j,a}$). Expression motion features ($\hat{f_e^i}$, $\hat{f_e^j}$) and mouth motion features ($\hat{f_m^a}$, $\hat{f_m^b}$) disentangled from different inputs are crossly recombined to generate corresponding keypoints ($p^{\hat{i},b}$, $j^{\hat{i},b}$) and ($p^{\hat{j},a}$, $j^{\hat{j},a}$). The loss functions are concretely introduced in the main paper.

Audio2Mouth. Identity Encoder E_i comprises of 8 downsampling blocks to extract identity feature $f_i \in \mathbb{R}^{512}$. We adopt the same Content Encoder E_{con} and Emotion Encoder E_{emo} as EVP [9] to generate content embedding $e_c \in \mathbb{R}^{256}$ and emotion embedding $e_e \in \mathbb{R}^{128}$. By concatenating f_i and e_c along the channel dimension, Mouth Decoder D_m produces mouth motion feature $\hat{f_m}$ via a long short-term memory (LSTM) network [7] followed by several ResBlock and MLP.

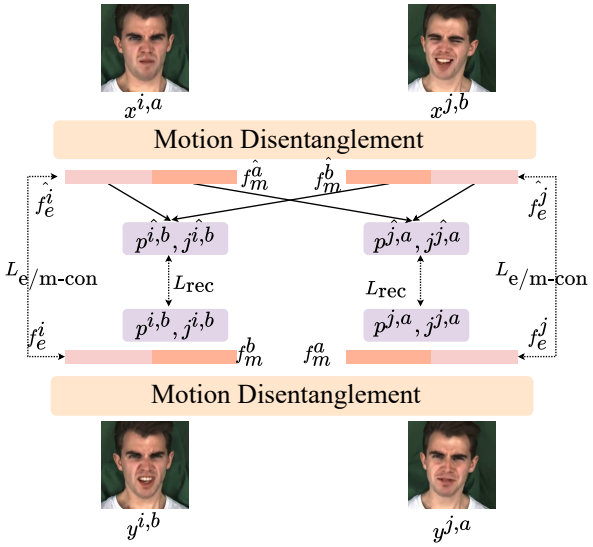


Figure 1: **Training process of Motion Reconstruction.** $L_{\text{e/m-con}}$ indicates two consistency losses, i.e, expression consistency loss $L_{\text{e-con}}$ and mouth consistency loss $L_{\text{m-con}}$.

A.2. Training Details

For each training video, we detect the face bounding box in the first frame and use it to crop the frame sequence

*Corresponding author.

in video. All videos are finally resized to 256×256 and converted to 25 frames per second. The audio is sampled at 16kHz frequency and aligned with the video frame by extracting 28×12 dim MFCC [11] features with the window size of 10 ms. Before training the whole framework, we pre-train the Content Encoder E_{con} and Emotion Encoder E_{emo} in Audio2Mouth module following EVP [9]. To achieve one-shot setting, the Identity Encoder E_i and Rendering module are pre-trained through audio-driven talking face task without emotion [17] on LRW [5], which contains sufficient speakers for better generalization performance. Moreover, we set the number of slots in Motion Memory Net as 128. When tuning the weights of different loss functions, we refer to previous works [17, 8] to help narrow down the available range of hyper-parameters in a random grid search. Finally, we set $\lambda_{\text{c-con}}$, $\lambda_{\text{e-con}}$, $\lambda_{\text{exp-mem}}$, λ_{align} and $\lambda_{\text{p,j}}$ in L_{stage1} as 0.01, 0.01, 0.1, 0.1 and 1, respectively.

B. More Results

B.1. Experiment Settings

Comparison Experiment Setting. We evaluate each method with their publicly available pre-trained models except for MEAD [16], which is a target-specific model, so we reproduce MEAD according to the descriptions in their original paper on test subjects in MEAD dataset [16]. For each comparing method, we feed them with the same audio and reference image for a fair comparison, while some methods require additional input for head pose and emotion control, we follow the original papers to pre-process data, achieving their respective optimal performance.

User Study Setting. For each emotion, we select 4 groups of inputs, where audios are randomly selected in MEAD dataset and reference images are selected in CREMA-D [3] and CFD dataset [12]. In addition, we also evaluate the performance of real data in MEAD dataset. Therefore, we generate 32 ($4\times$ 8 emotions) videos for each method and ground truth while generating 24 ($4\times$ 6 emotions) videos for ETK [6], which only considers 6 emotion categories. Videos are shuffled and randomly presented to participants. Participants are asked to classify the emotion perceived from the video and score for each video from 1 (worst) to 5 (best) on lip-synchronization, emotion naturalness and video quality.

B.2. More Experimental Results

Results of the Motion Reconstruction. To verify the effectiveness of Motion Reconstruction module, we randomly select images with different expressions and mouth shapes as expression sources and mouth sources. Then the images and the neutral image of the same subject are fed into Motion Reconstruction module for crossly reconstructing the results. For intuitive comparison, we generate the ground truth images via pseudo label generation strategy as described in Sec. 3.2. As displayed in Fig. 2, the results reconstructed by Motion Reconstruction module are similar to the ground truths with the same expressions and mouth shapes as expression sources and mouth sources, respectively. Please note that the mouth area of the results not only contains the same content information as mouth source, but also performs the same emotion as expression source in detail like mouth corners, as pointed out by red arrows. The results indicate that the Motion Reconstruction module effectively disentangles expression and mouth shape from inherently coupled face and globally integrate them to perform expression overall on the face.

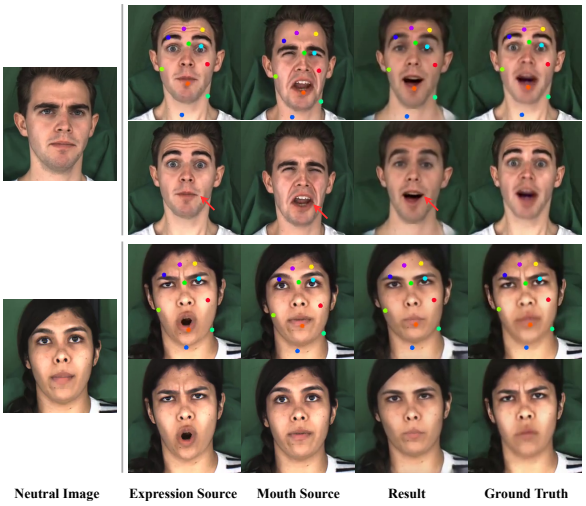


Figure 2: **Results of Motion Reconstruction.** We show two examples of our results with the same expression as the expression source and the same mouth shape as mouth source.

Emotion Manipulation. We achieve emotion manipulation by interpolating between emotion embedding extracted from audio with 'disgusted' and 'surprised' emotions. Specifically, we extract content embedding from the same audio and combine it with interpolated emotion embeddings as queries to calculate the value address in Motion Memory Network and generate emotional talking faces. As demonstrated in the top row of Fig. 3, the facial emotion dynamics are smoothly transited from disgust to surprise. Meanwhile, other facial factors related to emotion like head pose and mouth corner also transform with the change of interpolation weight α, which indicates that expression motion features contain all facial factors about expression as mentioned in the main paper. The bottom row of Fig. 3 demonstrates that the corresponding value addresses for

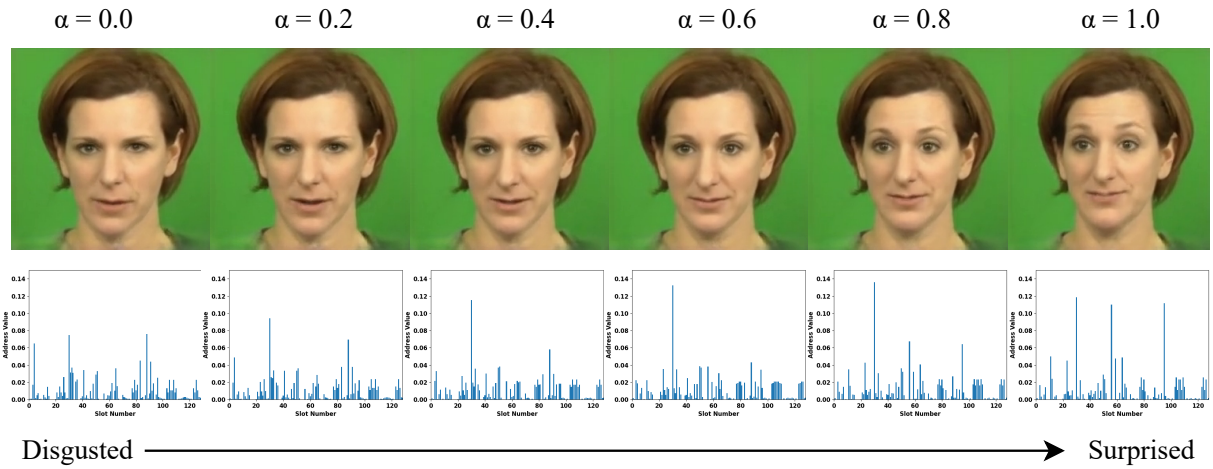


Figure 3: **Emotion interpolation and value addresses.** Top row shows the results generated by interpolating between 'disgusted' emotion embedding and 'surprised' emotion embedding, where α stands for interpolation weight. Bottom row presents the corresponding value address for each slot in Motion Memory Net.

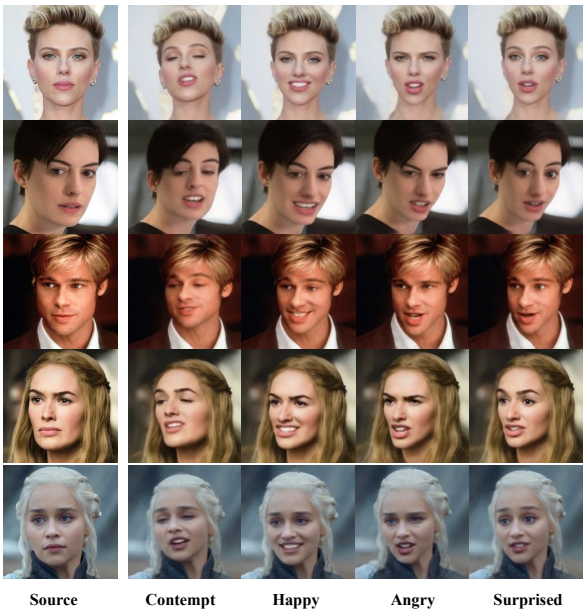


Figure 4: **More results on celebrities.**

each slot also linearly transit with the facial emotion dynamics transformation overall. The results suggest that the Motion Memory Net well stores expression motion features aligned with emo-mouth features to ensure the accuracy of emotion category and the consistency between expression and mouth shape. Besides, the results also suggest that more diverse expressions can be generated from the combination of the different expression features stored in memory. Moreover, we also test our method on celebrities shown in

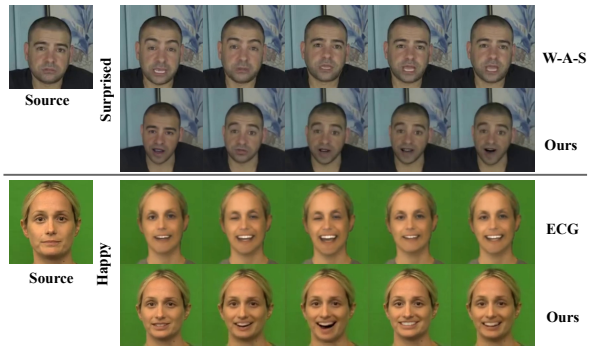
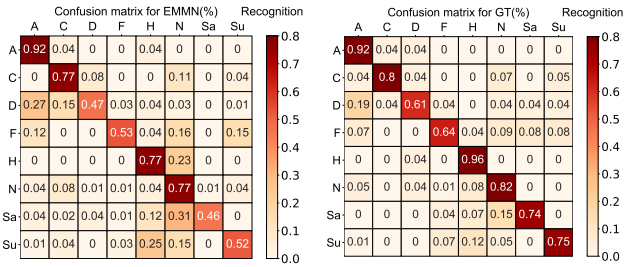


Figure 5: **Comparisons with person-specific and one-shot emotional talking face methods: Write-A-Speaker(W-A-S) and ECG.**

Fig. 4, which further verifies the effectiveness of the proposed method.

More Comparison Results. We further conduct a comparison between our method and two other emotional talking face methods: Target-specific Write-a-speaker [10] and one-shot ECG [15]. As the codes and pre-trained models for both methods have not been made public, we only can extract an emotional video clip from the provided demo video for comparison. The qualitative comparison results are presented in Fig. 5. Our method outperforms Write-a-speaker, a person-specific model which is trained on sufficient data of the target speaker, by generating more vivid emotional animation results. Moreover, unlike ECG, which fine-tunes their network on a single image of the target face,



(a) EMMN (b) GT

Figure 6: **Confusion matrices for perceived expression recognition (%) for 8 emotion categories.** The labels on the left side represent 'Angry', 'Contempt', 'Disgusted', 'Fear', 'Happy', 'Neutral', 'Sad' and 'Surprised', respectively.

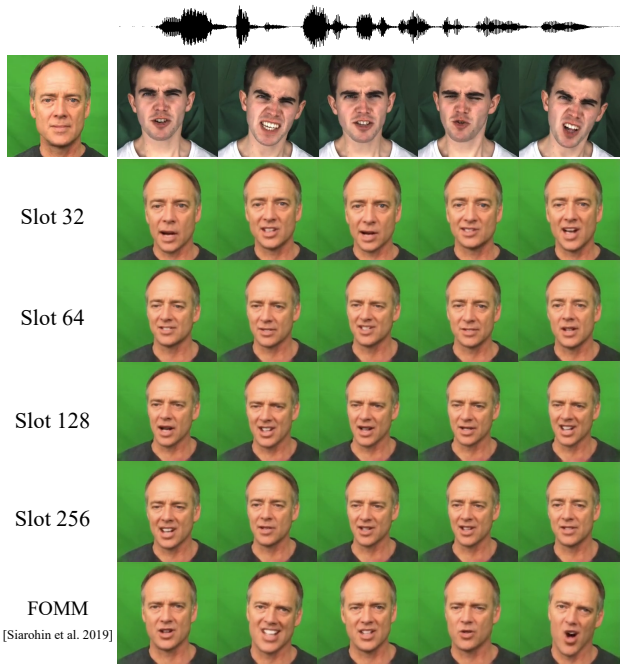


Figure 7: **Ablation study on the number of slots.** We explore the effect of different numbers of memory slot on the final performance.

our method synthesizes more realistic expressions and preserves more identity information without any fine-tuning. In addition, We offer more comparison to EAMM. Specifically, CSIM, FID and emotion accuracy are calculated to assess identity preservation, image quality and emotion representation of generated videos. Tab. 2 shows that our method outperforms EAMM among all metrics.

User Study. Due to various perceptions among individuals, it is challenging to achieve consensus on scores, especially when dealing with facial expression analysis [2]. To accommodate the variance of user opinions, we compute the mean and standard deviation of the scores and present an overview of the user study results in the main paper. Here, we additionally show the detailed emotion accuracy for our method and ground truth (GT). The confusion matrices for perceived expression recognition are illustrated in Fig. 6. In confusion matrix, each row denotes an emotion class and the values in this row represent the average probabilities that the videos with row emotion is recognized as the column emotions by the participants. We observe that we get high emotion accuracy on 'angry', 'contempt', 'happy' and 'neutral' emotions, while getting lower on 'disgusted' and 'fear'. Concretely, 27% videos with 'disgusted' emotion and 16% videos with 'fear' emotion are miscategorized as 'angry' and 'neutral'. Note that the results on GT also appear the similar inaccuracy on these emotion classes. We argue that it is extremely tough for people to recognize and perform 'disgusted' and 'fear' emotions.

Method/Metric	CSIM↑	FID↓	Emotion Accuracy↑
EAMM	0.699	83.396	63.36%
Ours	**0.753**	**59.718**	**72.13%**

Table 1: **More comparison to EAMM.**

Ablation Study. In this section, we conduct a set of experiments to investigate the effect of slot number selection on the final performance. Particularly, we set the slot number as 32, 64, 128 and 256 to conduct experiments, respectively. In addition, we drive the image of target subject with the video in MEAD through FOMM [14] as the ground truth for intuitive comparison. The qualitative results are given in Fig. 7, and the top row displays the reference image and ground truth frames of video in MEAD dataset. When the number of slot increases from 32 to 64, the expression become obvious but still unstable. We argue that the memory space is too limited to store sufficient emo-mouth feature and expression motion feature pairs. However, although a high expression performance can be achieved when the number of slots increases to 256, the mouth shape matching between results and ground truth gets worse, which may stem from excessive attention on expression but deficient on mouth shape. Accordingly, we empirically select the number of memory slots as 128 to obtain a relatively balanced trade-off between expressions and mouth shapes.

Comparison on computational cost. We conduct an analysis of the model parameters size and the average time required to generate a single frame. To ensure a fair comparison, we run all models on an NVIDIA GeForce GTX 3090 with 24GB memory and record the results in Tab. 2.

Metric/Method	ETK [6]	MEAD [16]	EAMM [8]	Ours
Computational cost (MB)	158.87	365.65	536.79	438.17
Inference time (S/Frame)	0.009	0.047	0.149	0.030

Table 2: **Computational cost and inference time for each method.**

Metric/Paper	ATVG [4]	Yi [18]	PC-AVS [19]	EAMM [8]
ATVG's SSIM	0.86	0.73	0.81	0.69

Table 3: **Inconsistent SSIM score reported by different papers on the same method (ATVG) and same dataset(LRW).**

As observed, ETK consumes the lowest computational resources, as it generates results with a resolution of 128×128 only. Both EAMM and our method are based on keypoint and dense flow field; however, EAMM employs an additional network to extract emotion feature from videos. In contrast, we store aligned emo-mouth features and expression features in a Memory Network, which consumes fewer computation resources and achieves faster inference time. Overall, our method is efficient and runs in reasonable time with reasonable cost.

C. Discussion

Quantitative comparison. We notice the inconsistent scores bewteen the displayed metics values in our paper and the one reported in the original papers. We evaluate our method and comparing methods following previous works [19, 8], which also appear inconsistent score cases in Tab. 3, where we list the ATVG's [4] SSIM score on LRW dataset reported by different papers. We attribute the inconsistent scores to several reasons: 1. **Different** train/test splits results in different videos for calculating the score, which inevitably introduce the errors. 2. **Randomly** selecting one frame per video as input for evaluation may lead to score variation. However, we consider these variations to be normal and negligible. This is because we verify that the current evaluation method [19, 8] is able to ensure that the scores fall within 95% confidence interval, when treating all frames of the test set as a whole, using z-test. 3. **Different** face cropping manners and metrics calculation codes unavoidably cause the inconsistency. We give an intuitive example in Fig 8, where we crop the generated videos and ground truth video into the resolution of 224×224 (Fig. 8 right) and 256×256 (Fig. 8 left), respectively. We calculate the SSIM of two pairs of images to verify score difference caused by different croppings: 0.655 (left), 0.688 (right). This indicates that different cropping methods indeed cause the inconsistent scores (higher when cropping lower resolution). To mitigate these errors, we use the same way to process the face images and calculate metrics when evaluating all methods.

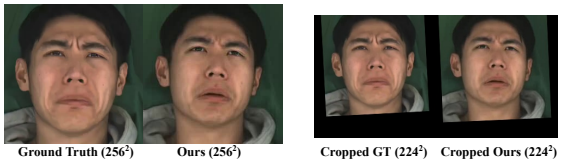


Figure 8: Source images and cropped images.

Despite slight differences in scores, the ranking of score between methods in our paper is consistent with original papers. Besides, we also conduct qualitative comparison and user study, which all validate the superiority of our method. To sum up, our experimental results are reliable.

Emotion representation. There are several ways to represent emotions including discrete emotion representation and dimensional emotion representation. Previous work [1] has suggested whether a dimensional or a discrete emotion representation is most appropriate based on Valence Focus and Arousal Focus theory. In this work, we utilize a dimensional $e_e \in \mathbb{R}^{128}$ to represent the emotion information extracted from audio [9]. Concretely, we infer expression features $\hat{f}_{e;em}$ from mouth features $\hat{f}_m$ and emotion embeddings e_e, which is a sequence. To make the subjects perform more diverse and smooth expressions and head poses, dimensional representation ($e_e \in \mathbb{R}^{128}$) is more suitable than discrete emotion representation (e.g., one-hot vector $v \in \mathbb{R}^8/\mathbb{R}^6$ in MEAD [16]/ECG [15]), which typically provides fixed features for one emotion category. However, during our user study experiment, the discrete emotion representation is more suitable for participants to identify which expression the video performs due to its high arousal focus requirement [1].

References

[1] Lisa Feldman Barrett. Discrete emotions or dimensions? the role of valence focus and arousal focus. *Cognition & Emotion*, 1998.

[2] Petra Saskia Bayerl and Karsten Ingmar Paul. What determines inter-coder agreement in manual annotations? a meta-analytic investigation. *Computational Linguistics*, 2011.

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

[4] Lele Chen, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7832–7841, 2019.

[5] Joon Son Chung and Andrew Zisserman. Lip reading in the wild. In *Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-*

24, 2016, Revised Selected Papers, Part II 13, pages 87–103. Springer, 2017.

[6] Sefik Emre Eskimez, You Zhang, and Zhiyao Duan. Speech driven talking face generation from a single image and an emotion condition. *IEEE Transactions on Multimedia*, 24:3480–3490, 2021.

[7] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 1997.

[8] 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. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, 2022.

[9] 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.

[10] Lincheng Li, Suzhen Wang, Zhimeng Zhang, Yu Ding, Yixing Zheng, Xin Yu, and Changjie Fan. Write-a-speaker: Text-based emotional and rhythmic talking-head generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1911–1920, 2021.

[11] Beth Logan. Mel frequency cepstral coefficients for music modeling. *international symposium/conference on music information retrieval*, 2000.

[12] Debbie S. Ma, Joshua Correll, and Bernd Wittenbrink. The chicago face database: A free stimulus set of faces and norming data. *Behavior Research Methods*, 2015.

[13] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

[14] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. *Advances in Neural Information Processing Systems*, 32, 2019.

[15] Sanjana Sinha, Sandika Biswas, Ravindra Yadav, and Brojeshwar Bhowmick. Emotion-controllable generalized talking face generation. In *International Joint Conference on Artificial Intelligence*. IJCAI, 2021.

[16] 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 *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI*, pages 700–717. Springer, 2020.

[17] S Wang, L Li, Y Ding, C Fan, and X Yu. Audio2head: Audio-driven one-shot talking-head generation with natural head motion. In *International Joint Conference on Artificial Intelligence*. IJCAI, 2021.

[18] Ran Yi, Zipeng Ye, Juyong Zhang, Hujun Bao, and Yong-Jin Liu. Audio-driven talking face video generation with learning-based personalized head pose. *arXiv preprint arXiv:2002.10137*, 2020.

[19] 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.