INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Emotional Talking Head Generation based on Memory-Sharing and Attention-Augmented Networks

Jianrong Wang[1] *Yaxin Zhao*[2] *Li Liu*[3(✉)] *Tianyi Xu*[1] *Qi Li*[4] *Sen Li*[1]

[1]College of Intelligence and Computing, Tianjin University, Tianjin, China
[2]Tianjin International Engineering Institute, Tianjin University, Tianjin, China
[3]The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China
[4]School of Electrical and Information Engineering, Tianjin University, Tianjin, China

avrillliu@hkust-gz.edu.cn

Abstract

Given an audio clip and a reference face image, the goal of the talking head generation is to generate a high-fidelity talking head video. Although some audio-driven methods of generating talking head videos have made some achievements in the past, most of them only focused on lip and audio synchronization and lack the ability to reproduce the facial expressions of the target person. To this end, we propose a talking head generation model consisting of a **M**emory-**S**haring **E**motion **F**eature extractor (MSEF) and an **A**ttention-**A**ugmented **T**ranslator based on **U**-net (AATU). Firstly, MSEF can extract implicit emotional auxiliary features from audio to estimate more accurate emotional face landmarks. Secondly, AATU acts as a translator between the estimated landmarks and the photo-realistic video frames. Extensive qualitative and quantitative experiments have shown the superiority of the proposed method to the previous works. Codes will be made publicly available.

Index Terms: Audio-driven Talking Head Generation, Emotion, Memory-sharing, U-net, Attention

1. Introduction

Audio-driven realistic talking head video generation plays a very important role in multiple applications, such as film making [1], video bandwidth reduction [2], virtual avatars animation [3, 4] and video conference [5], etc. According to the previous work [6, 7], an ideal realistic talking head video should satisfy the following requirements, *i.e.*, (1) the identity needs to be consistent with the target person, (2) the lip movements need to be synchronized with the audio content, (3) the videos should have natural facial expressions and head movements.

In the literature, some previous works have focused on generating lip-synchronized talking head videos [8–10], but they ignored the facial expressions modeling. In recent years, there has been some work on generating expression-controlled talking head videos. Blinking motions were added in [11, 12] to improve the realism by synthesizing talking head videos, but the results were still unsatisfactory, *i.e.*, the facial muscles were stiff. [13] relied on neutral video recordings of the target person to generate emotional talking head videos, but the facial expressiveness of the generated results was still insufficient. [6] designed a model to generate a talking head video that is emotionally consistent with an emotional source video by accepting four inputs, namely an identity reference image, an audio clip, a predefined pose video and the emotional source video. However, the video-driven based approach is limited by bandwidth, storage space, etc., and is not applicable in some application cases, such as bandwidth-constrained video conferencing.

Based on the above research and analysis, our work is to design an audio-driven talking head generation model that accepts two inputs, *i.e.*, an emotional audio clip and a reference facial image with the same emotion. The outputs are highly realistic videos of the target person. We believe that with the rapid development of photographic devices such as mobile phones and cameras, it should be easy to obtain such inputs. However, there are still two challenges in implementing such a model. Firstly, the facial pose of a person varies greatly across different emotional states. Secondly, rich facial expressions produce complex skin textures and facial shadows. To make the generated emotional talking head videos more realistic, we not only need to accurately predict the emotional facial landmarks, but also need to render the facial details during the regression from the landmarks to the images.

To solve these problems, we propose a new two-staged emotional talking head generation model. More precisely, in the first stage, because the emotional information in the audio is closely related to facial expressions, we explicitly extract the emotional features hidden in the audio as the auxiliary information. We train a *Memory-Sharing Emotional Feature extractor* (MSEF) in a supervised way, and propose a joint loss to change the optimization direction of the model to further improve the accuracy of the predicted landmarks. MSEF implicitly takes into account the relationship between different samples through the memory-sharing module with linear complexity, which is of great significance for extracting emotional features in audio. In the second stage, the predicted landmarks and the reference face image are fed into the *Attention-Augmented Translator* based on *U-net* (AATU) to generate photo-realistic talking head videos. AATU aims to focus on shallow details and important semantic features of the network simultaneously, reducing the loss of useful information and improving the model's performance, so that the output image can maintain more details such as skin texture and facial shadows of the target person.

To sum up, our contributions can be summarized as follows.

- We propose a novel model, a memory-sharing emotional feature extractor, to extract emotional features from audio signals. Using the extracted auxiliary features, the network can predict emotional face landmarks more accurately than previous works.
- An attention-augmented translator based on U-net is proposed to generate photo-realistic and emotional talking head video frames, *e.g.*, skin texture and facial shadows.
- Qualitative and quantitative experiments on the MEAD dataset show that the model achieves high-quality emotional talking head video generation, which is significantly superior

This work was supported by the National Natural Science Foundation of China (No. 61977049) and the National Natural Science Foundation of China (No. 62101351)

10.21437/Interspeech.2023-749

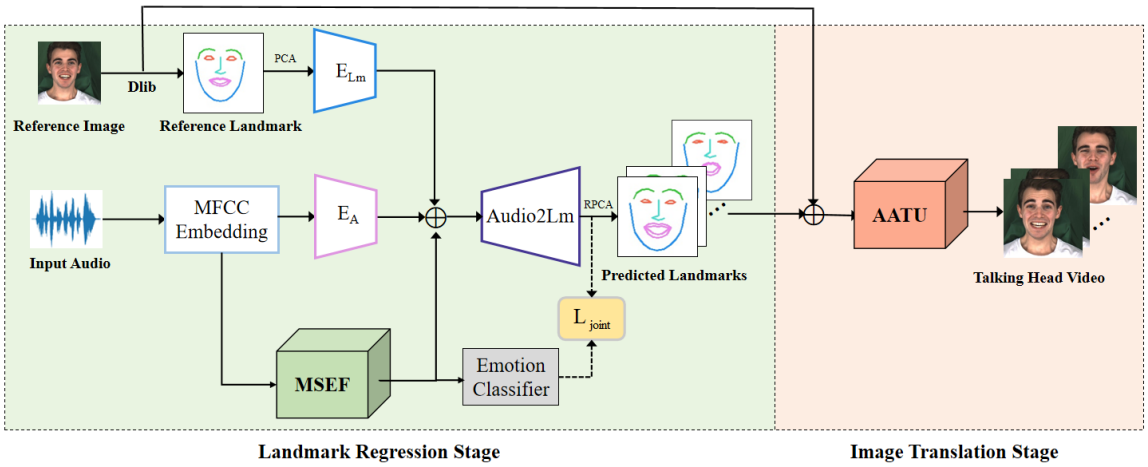

Figure 1: *Overview of our proposed talking head generation network. First of all, we extract implied emotional features from audio signals through MSEF module.* E_{Lm} *and* E_A *are landmark encoder and MFCC encoder, respectively. Secondly, we concatenate the landmark features, MFCC features and emotional features, and feed them into the* $Audio2Lm$ *module to predict dynamic face landmarks. Finally, we splice the predicted landmarks with the reference face image by channel, and render them into photo-realistic talking head video frames through the AATU module.*

to previous works.

2. Releted Work

In the literature, there are two general approaches to talking head video generation. One is the end-to-end mapping from audio to talking head video [14–16], and the other is the generation of talking head video through intermediate features, such as landmarks [17–19] and 3DMM parameters [20, 21].

The large number of parameters in the end-to-end talking head generation model often leads to the model overfitting the training data. Furthermore, the networks tend to be more concerned with lip and audio synchronization, making it difficult to generate face animations of various natural facial poses and expressions. To this end, Zhou *et al.* [22] first proposed a cascading model, *i.e.*, first mapping audio to landmarks and then converting them to images, which reduced the influence of non-audio related information in the video such as the angle of the shot. However, they only focus on the lip motion of the image. Song *et al.* [23], Thies *et al.* [24] and Zhang *et al.* [25] all regressed facial expression parameters of 3DMM model, but they all relied on the video of the target portrait, which is not applicable in most scenarios. Furthermore, the 3DMM parameter can only represent the geometry of a face and does not render a natural talking head video with high-quality skin textures.

Based on the current state of research on talking head generation, we propose a novel two-stage talking head generation model, which guarantees the expressiveness of facial emotions while rendering the detailed skin texture of the target person well.

3. Method

In this study, the model we proposed is shown in Figure 1, which mainly includes two stages. The first stage is to predict the lip-sync and emotional face landmarks from the audio and the reference face image. The second stage is an attention-augmented translator based on U-net, which takes our predicted landmarks and the reference face image as input to generate photo-realistic and emotional talking head video frames. In the following subsections, each module is described in detail.

3.1. Memory-Sharing Emotional Feature Extractor

In order to more accurately predict the face landmarks of the avatar, we extract the implied emotion auxiliary features from the audio signal through the MSEF module, the specific network structure is shown in Figure 2.

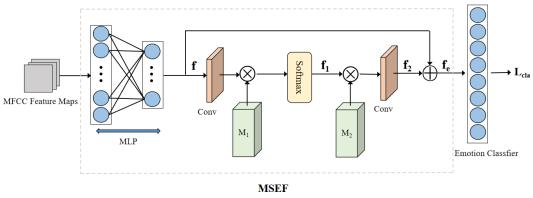

Figure 2: *The structure of memory-sharing emotional feature extractor.* M_1 *and* M_2 *are two memory units.* L_{ec} *is part of the joint loss.*

We encode MFCC features as 128-dimensional feature vectors (f) through MLP, and introduce two memory units with reference to the idea of [26]. The memory units are represented by linear layers, which can easily capture the global features in a single sample and can focus on the potential correlation between different samples. The emotional information in an audio clip is global features, and different audios may contain the same emotion. We believe that the memory-sharing units can explore the correlation between different samples and thus extract emotional features more accurately. Features that correspond to the same emotion with different audio, the treatment should be consistent. The specific formula is as follows:

$$f_e = f + g(softmax(g(f) \cdot M_1) \cdot M_2), \tag{1}$$

where g(·) represents the convolution operation, and the rest symbols are shown in Figure 2.

Then, we encode the emotional features into 8-dimensional feature vectors via an additional emotion classifier and introduce L_{ec} to supervise the training of this module. The loss function is formulated as follows:

$$L_{ec} = \frac{1}{N}\sum_{i=1}^{N} -[y_i \ln \hat{y}_i + (1-y_i)\ln(1-\hat{y}_i)], \quad (2)$$

where y represents the real emotion label of audio and $\hat{y}$ represents the predicted emotion category.

In addition, E_{Lm} and E_A are simple MLPs, encoding the landmark and MFCC into 512-dimension and 128-dimension feature vectors respectively. The $Audio2Lm$ module is composed of LSTM and a full connection layer.

In order to consider emotion without losing the accuracy of lips, we designed a joint loss function. In addition to the L_{ec} mentioned above, we add $L_{landmark}$ to the loss function to give the model the ability to regress face landmarks. At the same time, L_{lip} is added to make the model pay more attention to lips. The specific formulas are as follows:

$$L_{landmark} = \frac{1}{N}\sum_{i=1}^{N}(L_{real} - L_{fake})^2, \quad (3)$$

$$L_{joint} = L_{pca} + \alpha L_{landmark} + \beta L_{lip} + \gamma L_{ec}. \quad (4)$$

where the hyperparameter α, β and γ are the scaling factors, which we set to 10. L_{real} is the real face landmark and L_{fake} is the predicted face landmark. M_{real} is the real lip landmark, M_{fake} is the predicted lip landmark. L_{pca} and L_{lip} are calculated in a similar way to $L_{landmark}$, denoting the pca downscaling landmark and landmark of the lip region, respectively.

3.2. Attention-Augmented Translator based on U-net

In order to generate high-fidelity and emotional talking head video frames of the target person from the predicted landmarks, two challenges will be faced. Firstly, the photo-realistic talking head video frames need to pay attention to skin texture, and other details in order to better express emotions. Secondly, during the conversion from face landmarks to talking head video frames, a high degree of consistency with the target person's identity and a match to the predicted landmark facial contours and lip shape needs to be ensured.

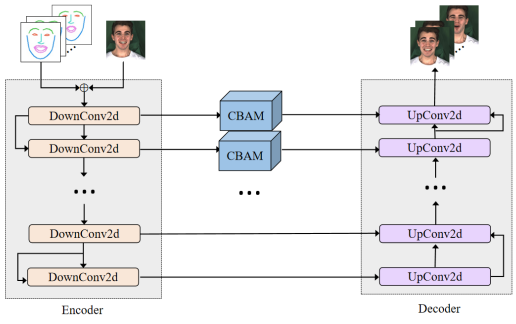

Figure 3: *The structure of attention-augmented translator based on U-net.*

In order to meet such challenges, we propose an attention-augmented translator based on U-net based on the MakeItTalk [18] framework. We propose AATU on this basis to further improve the quality of the generated video frames. As shown in Figure 3, we concatenate the predicted face landmarks with the reference face image by channel, and take it as the input of the encoder. The output of the decoder is photo-realistic and lip-sync talking head video frames. In the initial four layers of the encoder and decoder, we add a CBAM [27] module, respectively.

The CBAM consists of two sub-modules, spatial attention and channel attention, and implements a sequential attention structure from channel to space. We believe that in this task, spatial attention enables the neural network to pay more attention to the pixel areas in the image that play a determining role in facial expression and lip shape, while ignoring the unimportant areas. Channel attention is used to handle the distribution relationship of the feature map channels. Also, the distribution of attention over the two dimensions reinforced the impact of the attention mechanism on model performance. The shallow layer of the U-net structure effectively avoids the loss of spatial information caused by the fully connected layer, allowing the network to pay attention to skin texture and other details.

We use L1 loss as the loss function to supervise and train our network, and in order to enhance the quality of the generated talking head video frames, additional perceptual loss [28] is added. The specific formula are as follows:

$$L1 = \frac{1}{N}\sum_{i=1}^{N} ||f - \hat{f}||, \quad (5)$$

$$L_{per} = \frac{1}{N}\sum_{i=1}^{N} ||\phi_i(I) - \phi_i(\hat{I})||, \quad (6)$$

where I represents the real image, $\hat{I}$ represents the generated image. ϕ_i represents the layer i feature extraction layer of VGG-19 network [29].

4. Experiment And Result

4.1. Implement Details

1) Dateset and Setup. The dataset we use to evaluate the model is the same as EVP [13], *i.e.,* the MEAD dataset [30]. Other datasets, such as LRW [31] and VoxCeleb [32], are not suitable in our case, since they lack sentiment labels. And the CREMA-D [33] dataset does not distinguish much between various types of emotions. MEAD is a large-scale, high-quality emotional audio-visual dataset, which consists of 60 actors, including 8 basic emotions and 3 different emotional-intensity talking head videos. The training-test set is divided into a ratio of 8:2. We convert all talking head videos to 25fps and set the audio sample rate to 16KHz. For video 6 streams, we use Dlib to detect the face landmark of each frame. For audio streams, We extract MFCC at the window size of 25ms and hop size of 10ms. Our network is implemented using PyTorch. We use Adam optimizer, and the initial learning rate is set to 1e-4. We use the annealing strategy to adjust the learning rate through exponential decay.

2) Evaluation Metrics. In order to quantitatively evaluate different methods, we select common metrics in talking head generation. We used M-LMD and F-LMD to measure the accuracy of lip movements and facial contours. In addition, we use Structural Similarity Index Measure (SSIM) [34] and Peak Signal to Noise Ratio (PSNR) [35] to measure the quality of the generated talking head video frames.

3) Compared Methods. To the best of our knowledge, there are now open-source works that consider emotional information, such as EVP [13] and EAMM [6]. However, EAMM is speaker-independent when partitioning the dataset, and our approach is speaker-dependent in the same way as EVP. To be fair, we have compared our work with EVP, and our baseline model is based on ATVG [22] and MakeItTalk [18]. In addition, we also have compared with Audio2Head [36] , which is based on motion fields to generate talking head videos, and improved the realism of videos from the perspective of generating head movements.

4.2. Quantitative Result

Table 1: *Quantitative results on the MEAD test set.* ↑ *means the higher the better,* ↓ *means the lower the better.*

Method	F-LMD↓	M-LMD↓	SSIM↑	PSNR↑
Ground Truth	0.00	0.00	1.00	N/A
Baseline	2.39	3.38	0.69	32.38
EVP [13]	3.01	**2.45**	0.71	29.53
Audio2Head [36]	-	-	0.69	30.91
Ours w/o MSEF	2.38	3.06	0.72	33.30
Ours w/o AATU	2.35	3.02	0.72	33.29
Ours	**2.35**	3.02	**0.72**	**33.32**

"Ours w/o MSEF" represents only add AATU model, and "Ours w/o AATU" represents only add MSEF model. As can be seen from Table 1, when both MSEF and AATU are added, our model shows improvements in both emotion representation and image quality. Compared to EVP, our results show an increase of 0.66 in F-LMD and 3.79 in PSNR. The module proposed by EVP is helpful for lip accuracy. However, the module relies on long video drives of neutral emotions of the target person, which requires filtering audio pairs of the same content and different emotions, and is slightly weaker in terms of the intensity of emotional expression.

Because Audio2Head is not a landmark-based method, we have only compared the latter two metrics with it. Again, our method outperforms Baseline and Audio2Head in all metrics. The baseline lacks emotional information as an auxiliary secondary feature and is slightly less effective in emotional face fitting. Audio2Head generates pixel-level talking head video frames based on motion fields, losing some important information about the speaker and resulting in limited image quality generated by their method.

4.3. Qualitative Result

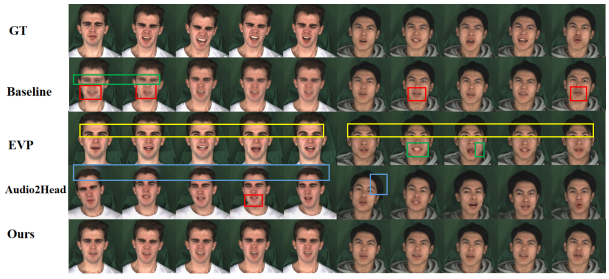

Figure 4: *Qualitative results on the MEAD test set.*

In order to visualize our comparison results, we also select some talking head video frames. As shown in Figure 4, our method generates high realistic talking head videos with strong emotions. Specifically, the yellow box locations are deficient in emotional expressiveness, the green box locations produce subtle artifacts, the red box locations have poor lip synchronization, and the blue box locations have poor identity consistency effects.

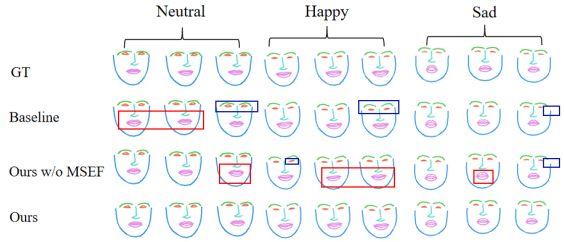

Figure 5: *Landmarks generated by different methods. Lip synchronization is inaccurate for the position marked by the red box, while landmarks of facial expressions are inaccurate for the position marked by the blue box (better view by zooming in).*

To further see the contribution of MSEF module to the accuracy of landmarks regression, we visualized the landmarks generated by different methods. It can be seen from Figure 5 that the landmarks generated after adding the MSEF module are closest to the ground truth. Specifically, the red box locations are inaccurate concerning the lip shapes, and the blue box locations are inaccurate concerning eye shapes and facial contours.

4.4. User Study

Table 2: *User study results of LS, EE and VPQ.*

Method	LS↑	EE↑	VPQ↑
Baseline	4.32	6.97	5.03
EVP [13]	**5.83**	5.55	5.29
Audio2Head [36]	5.07	5.90	5.59
Ours	5.25	**7.49**	**5.82**

In addition, we designed a detailed user study to assess the overall quality of the talking head videos. We used three metrics to measure video quality, *i.e.*, *Lip Synchronization* (LS), *Emotional Expressiveness* (EE) and *Video-Perceived Quality* (VPQ). A total of 30 participants completed our experimental questionnaire and they were asked to rate each video in the questionnaire from 1 (worst) to 10 (best). As Table 2 shows, although our lip sync is slightly worse than EVP, it is superior in terms of emotional expressiveness and video perception quality. This is because, EVP mainly focuses on lip synchronization, and our method works on whole facial modeling. Moreover, our method outperforms Baseline and Audio2Head on all metrics.

5. Conclusion

In this work, we propose a novel emotional talking head generation model, which consisted of a memory-sharing emotional feature extractor and an attention-augmented translator based on U-net. MSEF module is proposed to better predict the face landmarks in the talking head video. AATU module is proposed to better fit the facial details in the frames and improve the talking head video perception quality. Extensive experiments have proved that our method can generate lip-sync and emotional talking head videos. In the future, we will consider adding personalized head movements to the videos to further enhance the realism.

6. References

[1] H. Kim, M. Elgharib, M. Zollhöfer, H.-P. Seidel, T. Beeler, C. Richardt, and C. Theobalt, "Neural style-preserving visual dubbing," *ACM Transactions on Graphics (TOG)*, vol. 38, no. 6, pp. 1–13, 2019.

[2] T.-C. Wang, A. Mallya, and M.-Y. Liu, "One-shot free-view neural talking-head synthesis for video conferencing," in *Proceedings of the IEEE CVPR*, 2021, pp. 10 039–10 049.

[3] Y. Lu, J. Chai, and X. Cao, "Live speech portraits: real-time photorealistic talking-head animation," *TOG*, vol. 40, no. 6, pp. 1–17, 2021.

[4] J. Wang, Z. Tang, X. Li, M. Yu, Q. Fang, and L. Liu, "Cross-modal knowledge distillation method for automatic cued speech recognition," p. 2986–2990, 2021.

[5] Y. Guo, K. Chen, S. Liang, Y.-J. Liu, H. Bao, and J. Zhang, "Ad-nerf: Audio driven neural radiance fields for talking head synthesis," in *Proceedings of the IEEE CVPR*, 2021, pp. 5784–5794.

[6] X. Ji, H. Zhou, K. Wang, Q. Wu, W. Wu, F. Xu, and X. Cao, "Eamm: One-shot emotional talking face via audio-based emotion-aware motion model," in *ACM SIGGRAPH*, 2022, pp. 1–10.

[7] H. Zhou, Y. Sun, W. Wu, C. C. Loy, X. Wang, and Z. 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*, 2021, pp. 4176–4186.

[8] N. Sadoughi and C. Busso, "Speech-driven expressive talking lips with conditional sequential generative adversarial networks," *IEEE Transactions on Affective Computing*, vol. 12, no. 4, pp. 1031–1044, 2019.

[9] K. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, and C. Jawahar, "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*, 2020, pp. 484–492.

[10] J. Wang, Z. Wang, X. Hu, X. Li, Q. Fang, and L. Liu, "Residual-guided personalized speech synthesis based on face image," in *ICASSP*. IEEE, 2022, pp. 4743–4747.

[11] K. Vougioukas, S. Petridis, and M. Pantic, "Realistic speech-driven facial animation with gans," *ICCV*, vol. 128, no. 5, pp. 1398–1413, 2020.

[12] S. Sinha, S. Biswas, and B. Bhowmick, "Identity-preserving realistic talking face generation," in *2020 IJCNN*. IEEE, 2020, pp. 1–10.

[13] X. Ji, H. Zhou, K. Wang, W. Wu, C. C. Loy, X. Cao, and F. Xu, "Audio-driven emotional video portraits," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 14 080–14 089.

[14] B. Liang, Y. Pan, Z. Guo, H. Zhou, Z. Hong, X. Han, J. Han, J. Liu, E. Ding, and J. Wang, "Expressive talking head generation with granular audio-visual control," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3387–3396.

[15] Z. Ye, M. Xia, R. Yi, J. Zhang, Y.-K. Lai, X. Huang, G. Zhang, and Y.-j. Liu, "Audio-driven talking face video generation with dynamic convolution kernels," *IEEE Transactions on Multimedia*, 2022.

[16] J. Wang, Y. Zhao, H. Fan, T. Xu, Q. Li, S. Li, and L. Liu, "Memory-augmented contrastive learning for talking head generation," in *ICASSP*. IEEE, 2023, pp. 1–5.

[17] D. Das, S. Biswas, S. Sinha, and B. Bhowmick, "Speech-driven facial animation using cascaded gans for learning of motion and texture," in *European conference on computer vision*, 2020, pp. 408–424.

[18] Y. Zhou, X. Han, E. Shechtman, J. Echevarria, E. Kalogerakis, and D. Li, "Makelttalk: speaker-aware talking-head animation," *ACM Transactions on Graphics (TOG)*, vol. 39, no. 6, pp. 1–15, 2020.

[19] L. Liu, G. Feng, D. Beautemps, and X.-P. Zhang, "Re-synchronization using the hand preceding model for multi-modal fusion in automatic continuous cued speech recognition," *IEEE Transactions on Multimedia*, vol. 23, p. 292–305, 2020.

[20] C. Zhang, Y. Zhao, Y. Huang, M. Zeng, S. Ni, M. Budagavi, and X. Guo, "Facial: Synthesizing dynamic talking face with implicit attribute learning," in *Proceedings of the IEEE CVPR*, 2021, pp. 3867–3876.

[21] A. Richard, C. Lea, S. Ma, J. Gall, F. De la Torre, and Y. Sheikh, "Audio-and gaze-driven facial animation of codec avatars," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2021, pp. 41–50.

[22] L. Chen, R. K. Maddox, Z. Duan, and C. 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*, 2019, pp. 7832–7841.

[23] L. Song, W. Wu, C. Qian, R. He, and C. C. Loy, "Everybody's talkin': Let me talk as you want," *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 585–598, 2022.

[24] J. Thies, M. Elgharib, A. Tewari, C. Theobalt, and M. Nießner, "Neural voice puppetry: Audio-driven facial reenactment," in *European conference on computer vision*, 2020, pp. 716–731.

[25] Z. Zhang, L. Li, Y. Ding, and C. Fan, "Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset," in *Proceedings of the IEEE CVPR*, 2021, pp. 3661–3670.

[26] M.-H. Guo, Z.-N. Liu, T.-J. Mu, and S.-M. Hu, "Beyond self-attention: External attention using two linear layers for visual tasks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–13.

[27] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[28] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *ECCV*, 2016, pp. 694–711.

[29] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint:1409.1556*, 2014.

[30] K. Wang, Q. Wu, L. Song, Z. Yang, W. Wu, C. Qian, R. He, Y. Qiao, and C. C. Loy, "Mead: A large-scale audio-visual dataset for emotional talking-face generation," in *Computer Vision–ECCV 2020: 16th European Conference*, 2020, pp. 700–717.

[31] S. Yang, Y. Zhang, D. Feng, M. Yang, C. Wang, J. Xiao, K. Long, S. Shan, and X. Chen, "Lrw-1000: A naturally-distributed large-scale benchmark for lip reading in the wild," in *2019 14th IEEE international conference on automatic face & gesture recognition (FG 2019)*. IEEE, 2019, pp. 1–8.

[32] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," *arXiv preprint arXiv:1706.08612*, 2017.

[33] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "Crema-d: Crowd-sourced emotional multimodal actors dataset," *IEEE transactions on affective computing*, vol. 5, no. 4, pp. 377–390, 2014.

[34] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[35] N. D. Narvekar and L. J. Karam, "A no-reference perceptual image sharpness metric based on a cumulative probability of blur detection," in *2009 International Workshop on Quality of Multimedia Experience*. IEEE, 2009, pp. 87–91.

[36] S. Wang, L. Li, Y. Ding, C. Fan, and X. Yu, "Audio2head: Audio-driven one-shot talking-head generation with natural head motion," *arXiv preprint:2107.09293*, 2021.