Audio-Visual Fusion Network Based on Conformer for Multimodal Emotion Recognition

Peini Guo[1,2], Zhengyan Chen[1], Yidi Li[1], and Hong Liu[1(✉)]

[1] Key Laboratory of Machine Perception, Peking University, Shenzhen Graduate School, Shenzhen, China
guopeini@stu.pku.edu.cn, {chenzhengyan,yidili,hongliu}@pku.edu.cn
[2] Shanghai University, Shanghai, China

Abstract. Audio-visual emotion recognition aims to integrate audio and visual information for accurate emotion prediction, which is widely used in real application scenarios. However, most existing methods lack fully exploiting complementary information within modalities to obtain rich feature representations related to emotions. Recently, Transformer and CNN-based models achieve remarkable results in the field of automatic speech recognition. Motivated by this, we propose a novel audio-visual fusion network based on 3D-CNN and Convolution-augmented Transformer (Conformer) for multimodal emotion recognition. Firstly, the 3D-CNN is employed to process face sequences extracted from the video, and the 1D-CNN is used to process MFCC features of audio signals. Secondly, the visual and audio features are fed into a feature fusion module, which contains a set of convolutional layers for extracting local features and the self-attention mechanism for capturing global interactions of multimodal information. Finally, the fused features are input into linear layers to obtain the prediction results. To verify the effectiveness of the proposed method, experiments are performed on RAVDESS and a newly collected dataset named PKU-ER. The experimental results show that the proposed model achieves state-of-the-art performance in audio-only, video-only, and audio-visual fusion experiments.

Keywords: Multimodal emotion recognition · Audio-visual fusion · Convolutional Neural Network · Transformer

1 Introduction

In recent years, emotion recognition has received a great deal of attention for its robust performance and high application value, which has a wide range of applications such as intelligent transportation and human-computer interaction. Emotion recognition is a challenging task in the real world, as the way emotions

Supported by National Natural Science Foundation of China (No. 62073004), Science and Technology Plan of Shenzhen (No. JCYJ20200109140410340).

L. Fang et al. (Eds.): CICAI 2022, LNAI 13605, pp. 315–326, 2022.
https://doi.org/10.1007/978-3-031-20500-2_26

are expressed often varies across individuals and cultures [1]. Some recent works focus on multimodal emotion recognition that utilizes multiple modalities simultaneously for prediction, which can compensate for the lack of information in a single modality and has better performance than unimodal emotion recognition models. Modalities related to emotions usually include vision, audio, text, EEG, body gestures, etc. [2]. Considering that the visual and audio modalities are the most widely used due to their simplicity of sampling and good expressiveness, this paper focuses on the fusion of visual and audio modalities for emotion recognition.

Existing works of audio-visual fusion can be classified into three types according to the way features are fused: early fusion, late fusion, and model fusion [3]. Early fusion usually directly concatenates the video features and audio features after extracting them, and then feeds the fused features into fully connected layers for prediction. Deng et al. proposed a multimodal neural network structure that used Long Short-Term Memory (LSTM) to obtain time-varying visual information and performed emotional analysis by feature-level fusion of visual, audio, and textual information [4]. Kumar et al. proposed an improved multimodal emotion recognition method based on the attention mechanism and Bidirectional Gated Recurrent Unit (BiGRU) [5]. The implementation of early fusion is relatively simple, and it considers the correlation between the lower-level features of the modalities. However, it does not make full use of the complementarity of intra- and inter-modal information, resulting in little performance improvement over unimodal models. Late fusion usually models audio and visual features separately, and then joints the predictions to obtain the final predictions. Yu et al. used Convolutional Neural Network (CNN) and Deep Neural Network (DNN) to analyze textual and visual information, respectively, and then fused the prediction results of the two modalities by using the averaging strategy and weight [6]. Huang et al. firstly used image and text modalities for emotion classification separately, then proposed a multimodal model based on the fusion of two modalities, and finally fused the results of three models to obtain the final prediction results [7]. Because separate models are required to model and classify visual and audio modalities, the model structure of late fusion is more complex, and late fusion does not effectively exploit complementary information between modalities. Model fusion means that the extracted visual and audio features are fused by machine learning methods, deep neural networks, etc. Model fusion can leverage the complementary information between visual and audio modalities, which can bring significant improvement over unimodal models. In recent years, there are a lot of audio-visual fusion works based on the model fusion approach. Petridis et al. introduced a model based on residual networks and BiGRU, which is the first end-to-end audio-visual fusion model that simultaneously extracts features from raw pixels and audio waveforms and trains them on the large publicly available dataset [8]. Other methods based on model fusion using Recurrent Neural Network (RNN) for audio-visual speech recognition are proposed [9,10]. Due to the great results of the Transformer architecture based on the attention mechanism in the fields of computer vision and natural language processing,

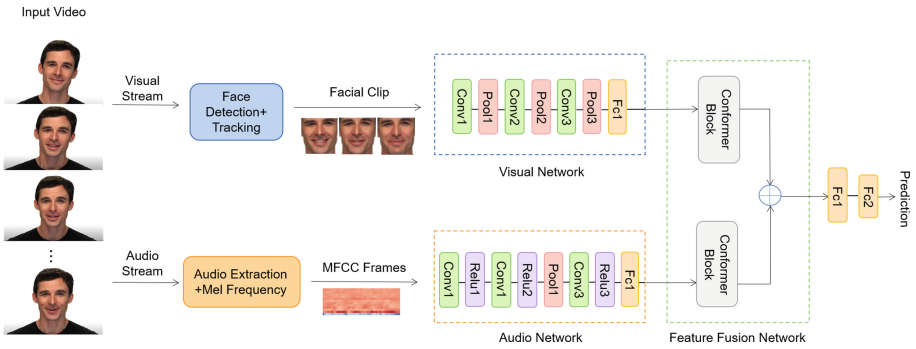


Fig. 1. The overall architecture of the proposed model, including the visual network, audio network, feature fusion network, and prediction head.

an increasing number of researchers have applied the attention mechanism in model fusion work [11]. Fu et al. proposed a feature fusion block based on the self-attention mechanism and residual structure for audio-visual fusion emotion recognition [12]. Zhang et al. proposed a feature fusion module based on the leader-follower attention mechanism [13]. The multi-modal perception attention network is proposed in [14] to learn the perception weights by using the complementarity across the audio-visual cues. Moreover, several audio-visual fusion works based on model fusion exploited the attention mechanism, with favorable results [15–17].

However, most of the existing audio-visual fusion methods focus on the design of feature fusion module, ignoring the significant impact of feature extraction on the model performance. Most existing methods only extract inter-modal complementary information, but do not effectively obtain intra-modal complementary information. Moreover, existing methods cannot guarantee the integrity of information during feature extraction and fusion, which may lose important semantic information and thus reduce the performance of the model.

Recently, models based on Transformer and CNN architectures have achieved promising results in the field of audio-visual fusion emotion recognition. To utilize the advantages of both Transformer and CNN in feature extraction, Gulati et al. proposed the Convolution-augmented Transformer (Conformer) to extract features with stronger generalization [18]. It was shown that the Conformer significantly outperformed the previous Transformer and CNN-based models, achieving state-of-the-art accuracy. Inspired by this, we propose a novel audio-visual fusion network based on Conformer architecture for emotion recognition. Specifically, the MFCC features extracted from audios are processed by 1D-CNN and the expression and action information included in consecutive frames of videos are extracted by 3D-CNN. Secondly, audio features and video features are fed into the feature fusion network, which learns location-based local features by convolution mechanism and content-based global dependencies by self-attention mechanism to effectively extract complementary information within

two modalities, and achieves effective feature fusion. Finally, the fused features are fed into fully connected layers to obtain the prediction results. To verify the effectiveness of proposed method, experiments are performed on RAVDESS [19] and PKU-ER multimodal emotion recognition datasets, and the experimental results show that our model outperforms other advanced methods and achieves state-of-the-art performance.

2 Methodology

As shown in Fig. 1, a new audio-visual fusion network for emotion recognition is proposed. Firstly, 3D-CNN is used to extract features from consecutive face frames in temporal and spatial dimensions, and 1D-CNN is used to process the MFCC features extracted from original audio signals. Secondly, the audio features and visual features are further processed using Conformer to make full use of complementary information within the modalities and obtain representations with strong generalization. The features are fused as an enhanced audio-visual feature representation in this step. Finally, the fused features are fed to the fully connected layers to obtain prediction results. The process is described in detail below.

2.1 Visual Network

The input of the visual network is consecutive face frames extracted from videos. Considering the video data are time- and space-correlated, the 3D convolution is utilized that can extract face expressions and movements simultaneously. The visual modality feature extraction network mainly consists of the 3D convolution layers, the max-pooling layers and the batch normalization layers. The group mechanism is incorporated to 3D convolution layers, which can effectively reduce the number of parameters for convolution and improve the model performance. To avoid the overfitting phenomenon, dropout layers are added to the model with the probability of 0.2. The final layer of the network compresses the feature dimensions from the original five dimensions (B, C, D, H, W) to three dimensions (B, C, N), where B denotes the batch size, C denotes the number of channels, D denotes the sequence depth, H denotes the height, W denotes the width, and N denotes the compressed dimension with the size of $\mathrm{D} \times \mathrm{H} \times \mathrm{W}$. For input v_i, the equations of output process in the visual network are as follows:

$$v_i' = \mathrm{Maxpool}\left(\mathrm{Conv}\left(v_i\right)\right), \tag{1}$$

$$v_i'' = \mathrm{Maxpool}\left(\mathrm{Conv}\left(v_i'\right)\right), \tag{2}$$

$$v_i''' = \mathrm{Maxpool}\left(\mathrm{Conv}\left(v_i''\right)\right), \tag{3}$$

$$v_o = \mathrm{Fc}\left(v_i'''\right). \tag{4}$$

where Conv denotes the 3D convolution layer, Maxpool denotes the max-pooling layer, and Fc denotes the fully connected layer.

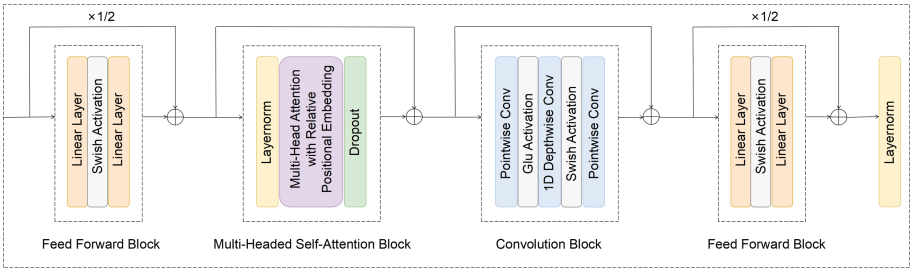


Fig. 2. Model architecture of Conformer block. It consists of two feed forward blocks sandwiching a multi-headed self-attention block and a convolution block.

2.2 Audio Network

MFCC is a kind of feature that has been proved to significantly improve the performance of speech recognition systems in recent years. MFCC is based on human auditorily perceptual characteristics, and simulates the human's audio system to the greatest extent, which is one of the most common and effective speech feature extraction algorithms. MFCC is the inverse spectral coefficient extracted in the frequency domain of the Mel scale, which describes the nonlinear nature of frequency perception by the human's ear. Therefore, the MFCC features extracted from original audios are used as the input to the network. The input passes through two 1D convolution layers to extract the local correlation of the sequence features, after which the features are compressed by a max-pooling layer to remove redundant information, and finally the features are fed to a 1D convolution layer to extract high-order semantic features. The output of intermediate layers is activated by Relu function. To avoid overfitting phenomenon, dropout layers are added to the model with the probability of 0.2. In the last layer, a linear function is used to align the last dimension of audio features with visual features to facilitate feature fusion. For input a_i, the equations of output process in the audio network are as follows:

$$a_i' = \mathrm{Conv}\left(\mathrm{Relu}\left(a_i\right)\right), \tag{5}$$

$$a_i'' = \mathrm{Conv}\left(\mathrm{Relu}\left(a_i'\right)\right), \tag{6}$$

$$a_i''' = \mathrm{Maxpool}\left(a_i''\right), \tag{7}$$

$$a_i'''' = \mathrm{Conv}\left(\mathrm{Relu}\left(a_i'''\right)\right), \tag{8}$$

$$a_o = \mathrm{Fc}\left(a_i''''\right). \tag{9}$$

where Conv denotes the 1D convolution layer, Maxpool denotes the max-pooling layer, and Fc denotes the fully connected layer.

2.3 Feature Fusion Network

A novel Conformer-based feature fusion network is introduced in this part, which aims to achieve efficient feature fusion by making full use of complementary information within multi-modalities. As shwon in Fig. 2, a Conformer block consists

of two half-step feed forward blocks with a multi-headed self-attention block and a convolution block in the middle [18]. This structure is inspired by Macaron-Net, which suggests replacing the original feed forward layer in the Transformer with two half-step feed forward layers, while using half-step residual connections in both feed forward layers [20]. The Conformer utilizes both convolution and self-attention mechanisms, which can simultaneously learn location-based local dependencies and content-based global representations to extract features with stronger generalization ability. Mathematically, for Conformer block i, with input x_i, the output y_i of the block is:

$$x_i = x_i + \frac{1}{2}\mathrm{FFN}\left(x_i\right), \tag{10}$$

$$x_i' = \tilde{x}_i + \mathrm{MHSA}\left(\tilde{x}_i\right), \tag{11}$$

$$x_i'' = x_i' + \mathrm{Conv}\left(x_i'\right), \tag{12}$$

$$y_i = \mathrm{Layernorm}\left(x_i'' + \frac{1}{2}\mathrm{FFN}\left(x_i''\right)\right), \tag{13}$$

where FFN denotes feed forward block, MHSA denotes multi-headed self-attention block, and Conv denotes convolution block.

The visual and audio features processed by Conformer are added up in the channel dimension, which ensures the integrity of the fused information. When the feature of audio modality is y_a and the feature of visual modality is y_v, the fused feature is:

$$I = y_a \oplus y_v. \tag{14}$$

where $\oplus$ denotes the element-wise addition.

2.4 Classfication

The fused features are fed into fully connected layers to obtain prediction results. The loss function is the cross-entropy function, which has great performance in multi-classification problems. The specific equations are as follows:

$$\text{prediction} = WI + b \in \mathbb{R}^N, \tag{15}$$

$$\mathrm{L} = -\sum_i y_i \log\left(\widehat{y}_i\right). \tag{16}$$

where N denotes the dimension of the output vector, I denotes the fused feature, W denotes the weight tensor, $b \in \mathbb{R}^N$ denotes the bias, $y = \{y_1, y_2, y_3 \ldots y_n\}^{\mathrm{T}}$ denotes the one-hot vector of emotion labels, $\widehat{y} = \{\hat{y}_1, \hat{y}_2, \hat{y}_3 \ldots \hat{y}_n\}$ denotes the probability distribution of predictions, and n denotes the number of emotion categories.

3 Experiments and Discussions

3.1 Datasets

For the emotion recognition task of audio-visual fusion, a good video emotion classification dataset is crucial. The metrics used to evaluate the dataset include

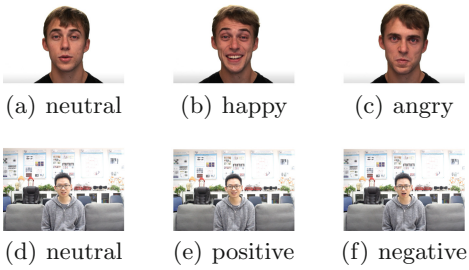


Fig. 3. Samples from datasets. Above: RAVDESS. Below: PKU-ER.

the clarity of the video/audio, the degree of the character's facial expression/lip movement changes, the obviousness of the pitch changes, etc. Based on the above metrics, a publicly available dataset and a self-made dataset are used to evaluate the performance of the model simultaneously.

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) contains 7356 files of utterances containing emotional polarity delivered by 24 professional actors in neutral North American accents [19]. Speechs include eight types of emotions: neutral, calm, happy, sad, angry, fearful, disgusted and surprised, and songs include five types of emotions: calm, happy, sad, angry and fearful. Each emotion contains two emotional intensities (normal, strong). All conditions are available in three formats: audio-only (16bit, 48 KHz, .wav), video-only (no sound), and audio-video (720p H.264, AAC 48 KHz, .mp4). The audio-video format of speech part is selected, with total 1440 files. The PKU-Emotion Recognition Dataset (PKU-ER) contains total 896 Chinese emotion video files recorded by 14 experimenters in human-computer interaction scenes. The emotions are classified into three categories: positive, negative, and neutral. Each emotion contains only one emotional intensity (normal). All files are in audio-video format. The above datasets facilitate the study of specific characteristics related to emotions, avoiding possible prejudices in the expression of emotions caused by cultural differences. Moreover, the number of documents corresponding to each emotion is nearly equal, obviating the problems that may result from unbalanced data samples. Some examples of the two datasets are shown in Fig. 3.

3.2 Implementation Details

For the visual modality, OpenFace [21] is utilized to extract 68 face landmarks in each frame of the video, thus cropping each frame to a face ROI with the resolution of 224×224. To alleviate the problem of small sample size of the dataset, data augmentation is performed on images, including random cropping, random horizontal flipping and normalization. For the audio modality, since there is usually no sound in the first 0.5 s, 2.45 s duration of the audio starting from the 0.5 s are extracted, with the sampling rate 44100 Hz. The first 13 MFCC features are selected for each cropped audio clip.

Table 1. Impact of model architecture on RAVDESS and PKU-ER datasets.

Model	Accuracy (on RAVDESS)	Accuracy (on PKU-ER)
Ours	78.45	90.53
w/o group convolution	77.63	89.25
w/o feature fusion block	72.57	84.61
w/o conformer encoder (A)	75.82	87.74
w/o conformer encoder (V)	74.50	86.37

The 6-fold cross-validation is performed on datasets to provide more robust results, i.e., the datasets are divided into training and validation sets in the ratio of 5:1, without dividing the separate test set. The model is trained using Adam optimizer with the initial learning rate of 0.001. The loss function is the cross-entropy function. To avoid overfitting, earlystopping is used, i.e., the accuracy is tested on validation sets after each epoch, and training is stopped in advance if the accuracy does not rise for n consecutive times, in experiments $n = 20$. All parts of the model are trained simultaneously. The training process of the model is done on a single NVIDIA RTX3090. The final accuracy reported is the average accuracy over 6 folds.

3.3 Ablation Study

Impact of Model Architecture. To verify the effectiveness of model, the impact of model architecture on performance is explored. The experimental results are shown in Table 1. When group convolution in the visual network is not used, the accuracy of the model decreases by 0.82% and 1.28% on RAVDESS and PKU-ER datasets, respectively, indicating that group convolution can effectively reduce the number of calculation parameters and improve the model performance. To verify the impact of the feature fusion network, the extracted features are stitched directly and fed into linear layers to obtain prediction results, compared with the case of using the feature fusion network. The experimental results show that when not using the feature fusion network, the classification accuracy of the model decreases by 5.88% and 5.92% on RAVDESS and PKU-ER datasets, respectively. This illustrates that feature fusion network can make full use of complementary information within and across modalities and remove redundant information, facilitating accurate predictions of emotion. When the Conformer for audio modality is removed, the classification accuracy of the model decreases by 2.63% and 2.79% on RAVDESS and PKU-ER datasets, respectively. When the Conformer for video modality is removed, the classification accuracy of the model decreases by 3.95% and 4.16% on RAVDESS and PKU-ER datasets, respectively. This indicates that Conformer can effectively learn both location-based local features and content-based global interactions to obtain representations with stronger generalization. Furthermore, Conformer has more influence on video modality than audio modality, suggesting that features within the video modality are more relevant.

Table 2. Comparison of the classification accuracy with different inputs.

Stream	Accuracy (on RAVDESS)	Accuracy (on PKU-ER)
A	62.35	56.84
V	70.03	81.58
A + V	**78.45**	**90.53**

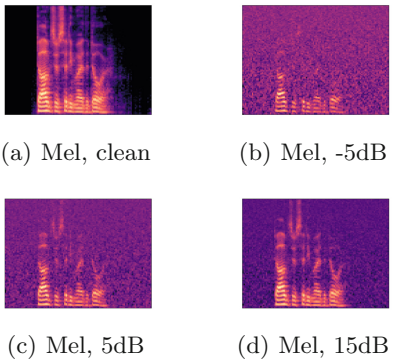
Impact of Different Inputs. The classification accuracies of model on two datasets with different inputs (audio-only, video-only, audio-video) are shown in Table 2. When the input contains both audio and video, the accuracy of the model on datasets is 78.45% and 90.53%, respectively. When the input is only audio or video, the accuracy decreases significantly, and the accuracy is lower when the input is audio than when the input is video. This indicates that the video modality provides more information than the audio modality, in addition, audio and video modalities can effectively provide complementary information.

Impact of Noise. The random noise with Signal-to-Noise Ratios(SNR) of −5, 0, 5, 10, 15, 20 dB is added to the audio data of RAVDESS dataset to investigate the impact of different noise intensities on model performance. Figure 4 shows some visualization results of Mel-spectrogram in the case of clean and different intensities of noise. It can be seen that the features in the Mel-spectrogram become clearer as the noise intensity decreases. The experimental results are shown in Fig. 5, where the horizontal axis denotes the SNR, the vertical axis denotes the classification accuracy of the model on RAVDESS dataset, and three broken lines denote the model under three types of inputs (A for audio-only, V for video-only, and A+V for audio-video). When using video frames as input, the accuracy remains constant 70.03% as visual information is not affected by acoustic noises. In the case of high noise, the performance of audio-only model degrades dramatically, and the accuracy of audio-visual fusion model decreases more slowly, indicating that audio-visual fusion model is more robust to noise than audio-only model. In contrast, the audio-visual fusion model outperforms other models in the low-noise case, suggesting that the video and audio modalities efficiently provide complementary information which enhances the performance of audio-visual fusion model.

3.4 Comparison with the State-of-the-Art

The model is compared with the following methods:

1. MMTM [22]: A feature fusion model that allows feature fusion in convolutional layers of different spatial dimensions and feature selection using the self-attention mechanism.
2. MSAF [23]: A feature fusion model where each modal features are segmented in the channel direction into feature blocks with equal number of channels and a joint representation is created.



(a) Mel, clean (b) Mel, -5dB

(c) Mel, 5dB (d) Mel, 15dB

Fig. 4. Visualization of Mel-spectrogram under clean and different intensities of noise.

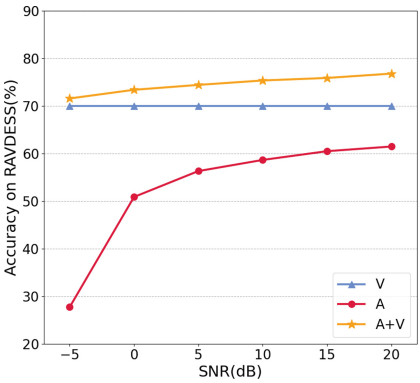


Fig. 5. The impact of noise on model.

Table 3. Comparison with state-of-the-arts on RAVDESS and PKU-ER datasets.

Model	Accuracy (on RAVDESS)	Accuracy (on PKU-ER)	Params
MMTM [22]	73.12	83.58	31.97M
MSAF [23]	74.86	83.12	25.94M
ERANNs [24]	75.23	86.94	23.60M
Ours	**78.45**	**90.53**	**18.40M**

3. ERANNs [24]: A new CNN structure proposed for audio-visual fusion emotion recognition, achieving state-of-the-art performance.

The experiment results are shown in Table 3. On both datasets, our proposed model outperforms the above methods and achieves state-of-the-art. The proposed model improves 5.33% over MMTM, 3.59% over MSAF, and 3.22% over ERANNs on the RAVDESS dataset. Moreover, the number of parameters in our model is much less than above methods, with only 18.4 millions. This indicates that the proposed model based on Conformer sufficiently obtains audio-visual features with high generalization and enables effective feature fusion and decision classification, which can achieve more advanced performance with the high parameter efficiency.

4 Conclusion

In this paper, we focus on how to extract relevant information within the audio and visual modalities and achieve effective feature fusion. To solve the problem, an audio-visual fusion network based on convolution-augmented Transformer is proposed for multimodal emotion recognition. The strategy of combining

convolution with self-attention mechanism can simultaneously learn location-based local features and content-based global dependencies, obtaining sufficient complementary information within modalities. The element-wise addition operation ensures the integrity of the fused features. The experimental results on RAVDESS and PKU-ER datasets show that the proposed model achieves efficient feature extraction and fusion, achieving state-of-the-art performance. Our future work will extend the proposed method to fine-grained emotion classification, and introduce text information for multimodal emotion recognition.

References

1. Praveen, R.G., et al.: A Joint Cross-Attention Model for Audio-Visual Fusion in Dimensional Emotion Recognition. arXiv preprint arXiv:2203.14779 (2022)
2. Praveen, R.G., Granger, E., Cardinal, P.: Cross attentional audio-visual fusion for dimensional emotion recognition. In: IEEE International Conference on Automatic Face and Gesture Recognition, pp. 1–8 (2021)
3. Wu, C.H., Lin, J.C., Wei, W.L.: Survey on audio-visual emotion recognition: databases, features, and data fusion strategies. APSIPA Trans. Signal Inf. Process. **3**(1) (2014)
4. Deng, D., Zhou, Y., Pi, J., Shi, B.E.: Multimodal utterance-level affect analysis using visual, audio and text features. arXiv preprint arXiv:1805.00625 (2018)
5. Kumar, A., Vepa, J.: Gated mechanism for attention based multimodal sentiment analysis. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 4477–4481 (2020)
6. Yu, Y., Lin, H., Meng, J., Zhao, Z.: Visual and textual sentiment analysis of a microblog using deep convolutional neural networks. Algorithms **9**(2), 41 (2016)
7. Huang, F., Zhang, X., Zhao, Z., Xu, J., Li, Z.: Image-text sentiment analysis via deep multimodal attentive fusion. Knowl.-Based Syst. **167**, 26–37 (2019)
8. Petridis, S., Stafylakis, T., Ma, P., Cai, F.: End-to-end audiovisual speech recognition. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 6548–6552 (2018)
9. Liu, H., Chen, Z., Yang, B.: Lip graph assisted audio-visual speech recognition using bidirectional synchronous fusion. In: Conference of the International Speech Communication Association, pp. 3520–3524 (2020)
10. Liu, H., Xu, W., Yang, B.: Audio-visual speech recognition using a two-step feature fusion strategy. In: International Conference on Pattern Recognition, pp. 1896–1903 (2021)
11. Vaswani, A., et al.: Attention is all you need. In: Annual Conference on Neural Information Processing Systems, vol. 30, pp. 6000–6010 (2017)
12. Fu, Z., et al.: A Cross-Modal Fusion Network Based on Self-attention and Residual Structure for Multimodal Emotion Recognition. arXiv preprint arXiv:2111.02172 (2021)
13. Zhang, S., Ding, Y., Wei, Z., Guan, C.: Continuous emotion recognition with audio-visual leader-follower attentive fusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3567–3574 (2021)
14. Li, Y., Liu, H., Tang, H.: Multi-modal perception attention network with self-supervised learning for audio-visual speaker tracking. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 1456–1463 (2022)

15. Serdyuk, D., Braga, O., Siohan, O.: Audio-visual speech recognition is worth $32 \times 32 \times 8$ voxels. In: IEEE Automatic Speech Recognition and Understanding Workshop, pp. 796–802 (2021)
16. Tran, M., Soleymani, M.: A pre-trained audio-visual transformer for emotion recognition. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 4698–4702 (2022)
17. Chang, F.J., Radfar, M., Mouchtaris, A., King, B., Kunzmann, S.: End-to-end multi-channel transformer for speech recognition. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 5884–5888 (2021)
18. Gulati, A., et al.: Conformer: convolution-augmented transformer for speech recognition. In: Conference of the International Speech Communication Association, pp. 5036–5040 (2020)
19. Livingstone, S.R., Russo, F.A.: 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 Computational Linguistics, pp. 2978–2988 (2019)
20. Lu, Y., et al.: Understanding and improving transformer from a multi-particle dynamic system point of view. In: Workshop on Integration of Deep Neural Models and Differential Equations (2020)
21. Baltrusaitis, T., Zadeh, A., Lim, Y.C., Morency, L.P.: Openface 2.0: facial behavior analysis toolkit. In: IEEE International Conference on Automatic Face and Gesture Recognition, pp. 59–66 (2018)
22. Joze, H.R.V., Shaban, A., Iuzzolino, M.L., Koishida, K.: MMTM: multimodal transfer module for CNN fusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13289–13299 (2020)
23. Su, L., Hu, C., Li, G., Cao, D.: MSAF: Multimodal Split Attention Fusion. arXiv preprint arXiv:2012.07175 (2020)
24. Verbitskiy, S., Berikov, V., Vyshegorodtsev, V.: ERANNs: Efficient Residual Audio Neural Networks for Audio Pattern Recognition. arXiv preprint arXiv:2106.01621 (2021)