



# Interactive Co-Learning with Cross-Modal Transformer for Audio-Visual Emotion Recognition

*Akihiko Takashima, Ryo Masumura, Atsushi Ando, Yoshihiro Yamazaki,  
Mihiro Uchida, Shota Orihashi*

NTT Corporation, Japan

akihiko.takashima.dg@hco.ntt.co.jp

## Abstract

This paper proposes a novel modeling method for audio-visual emotion recognition. Since human emotions are expressed multi-modally, jointly capturing audio and visual cues is a potentially promising approach. In conventional multi-modal modeling methods, a recognition model was trained from an audio-visual paired dataset so as to only enhance audio-visual emotion recognition performance. However, it fails to estimate emotions from single-modal inputs, which indicates they are degraded by overfitting the combinations of the individual modal features. Our supposition is that the ideal form of the emotion recognition is to accurately perform both audio-visual multi-modal processing and single-modal processing with a single model. This is expected to promote utilization of individual modal knowledge for improving audio-visual emotion recognition. Therefore, our proposed method employs a cross-modal transformer model that enables different types of inputs to be handled. In addition, we introduce a novel training method named interactive co-learning; it allows the model to learn knowledge from both and either of the modals. Experiments on a multi-label emotion recognition task demonstrate the effectiveness of the proposed method.

**Index Terms:** audio-visual emotion recognition, cross-modal transformer, interactive co-learning

## 1. Introduction

Audio-visual emotion recognition [1] is a technology that recognizes human emotions from both audio and visual cues, i.e. speech and facial information. Since human emotions are expressed multi-modally, it is more promising than those from speech [2, 3] or face recognition [4–6] alone. There are a wide range of potential applications such as intelligent tutoring systems, support for autism spectrum disorders, and cognitive load monitoring [7, 8]. Therefore, research and development of audio-visual emotion recognition technology is active in academic and industrial fields.

A large number of methods for audio-visual emotion recognition have been investigated. Most of the recent studies are based on neural networks. The emotional cues of each modality are extracted by modal-dependent encoders, and then integrated to estimate emotions [9–14]. One of the main topics of conventional studies is the integration of the emotional cues. Simple integration approaches such as the combinations of the emotional embedding vectors or just estimated probabilities of modal-dependent classifiers, i.e., early or late fusions, have been developed [15]. In addition, it has been reported cross-modal modeling based on the interaction of individual modals achieved better performance. Attention mechanisms or encoder-decoder style architectures are used to model the affection from source to target modals [14, 16, 17].

One of the problems of conventional cross-modal modeling is that they require all the modalities in the inference. In fact, the conventional models are trained from an audio-visual paired dataset so as to only enhance audio-visual emotion recognition performance. However, the conventional models fail to estimate emotions from single-modal inputs, which indicates they are degraded by overfitting the combinations of the individual modal features. We suppose that it is desired to recognize emotions from both multi-modal and single-modal inputs with the same recognition model as humans do. For example, a sample with a happiness label has smiling features in the visual cue, and laughing features in the audio cue. In this case, ideal form of emotion recognition works correctly even when either the smiling features or the laughing features is not available. Though there are several cross-modal modeling methods that evaluate the reliability of each modal and use the information of particular modals [18, 19], they require all of the modalities to estimate the reliabilities.

In this paper, we propose a novel audio-visual emotion recognition that uses the same model to process both multi-modal and single-modal information. Our key idea is to map all the representations of the audio and the visual cues into the same latent space to handle modal differences. To this end, a new recognition model and a new training scheme is introduced in the proposed method. Our proposed method employs a cross-modal transformer model that enables different types of inputs to be handled. This is motivated by recent successes in cross-modal representation learning [20, 21] where a unified model is utilized for both single-modal and multi-modal processing. In addition, we introduce a novel training method named interactive co-learning; it allows the model to learn knowledge from both and either of the modals. While co-learning is generally used for transferring knowledge attained from one modal into a different modal [22, 23], our interactive co-learning transfers knowledge not only from single-modal into multi-modal, but also from multi-modal into single-modal. Our interactive co-learning uses three types of inputs; audio-visual modality, audio modality alone and visual modality alone, and trains the cross-modal transformer model so as to enhance both single-modal and multi-modal emotion recognition performance. Experiments on a multi-label emotion recognition task demonstrate that the proposed method outperforms the conventional method with not only uni-modal inputs but also multi-modal inputs.

## 2. Audio-Visual Multi-Label Emotion Recognition

This section briefly describes audio-visual emotion recognition that recognizes human emotions from both audio cues and vi-

sual cues. Our emotion recognition task is a multi-label emotion recognition that jointly handles emotion-wise binary classification problems.

### 2.1. Definition

In audio-visual multi-label emotion recognition, emotion-wise labels  $\mathbf{L} = \{l_1, \dots, l_K\}$  are jointly estimated from audio features  $\mathbf{S} = \{s_1, \dots, s_M\}$  and its corresponding visual features  $\mathbf{C} = \{c_1, \dots, c_N\}$  where  $s_m$  is the  $m$ -th audio feature,  $c_n$  is the  $n$ -th visual feature.  $M$  is the number of audio features and  $N$  is the number of visual features. The  $k$ -th emotion  $l_k$  is represented as a binary label (0 or 1) and  $K$  is the number of target emotions. Audio features are generally extracted from speech information and visual features are extracted from facial RGB images. To model the audio-visual multi-label emotion recognition, we define emotion-wise conditional probabilities given the audio features  $\mathbf{S}$  and visual features  $\mathbf{V}$ . Thus, multi-label emotion recognition model jointly estimates emotion-wise conditional probabilities  $P(l_1|\mathbf{S}, \mathbf{C}, \Theta), \dots, P(l_K|\mathbf{S}, \mathbf{C}, \Theta)$  where  $\Theta$  represents the trainable model parameter set. In this case, the inference for the  $k$ -th emotion label  $\hat{l}$  is achieved by

$$\hat{l}_k = \operatorname{argmax}_{l_k \in \{0,1\}} P(l_k|\mathbf{S}, \mathbf{C}, \Theta). \quad (1)$$

To model this audio-visual multi-label emotion recognition problem, various network architectures [9–14, 16, 17, 24] can be used.

### 2.2. Conventional training method

The model parameter set  $\Theta$  is optimized on the training dataset  $\mathcal{D}$  composed of the audio/visual features with emotion-wise labels,

$$\mathcal{D} = \{(\mathbf{L}^t, \mathbf{S}^t, \mathbf{C}^t) \mid t \in \{1, \dots, T\}\}, \quad (2)$$

where  $T$  is the number of samples in the training dataset. In conventional studies, the model parameter set is optimized so as to enhance only audio-visual emotion recognition performance. Thus, the model parameter set is optimized by

$$\hat{\Theta} = \operatorname{argmin}_{\Theta} - \sum_{t=1}^T \sum_{k=1}^K \log P(l_k^t|\mathbf{S}^t, \mathbf{C}^t, \Theta). \quad (3)$$

The optimization is achieved by a mini-batch stochastic gradient decent algorithm. This means the model parameter set  $\Theta$  is optimized for the combination of  $\mathbf{S}^t$  and  $\mathbf{C}^t$ , not for each of them individually. In other words, there is no guarantee of estimating the emotion-wise conditional probabilities from  $\mathbf{S}^t$  or  $\mathbf{C}^t$  alone.

## 3. Proposed Method

This section details our proposed modeling method for audio-visual multi-label emotion recognition. We introduce a new recognition model and a new training scheme into the proposed modeling method.

### 3.1. Cross-modal transformer

The proposed modeling method employs a cross-modal transformer. The advantage of this approach is that different types of features can be handled by the same input method. Thus, this architecture can perform not only audio-visual emotion recognition but also audio emotion recognition and visual emotion recognition. Figure 1 shows how to perform these three types

of emotion recognition tasks using a unified architecture. Our architecture consists of four blocks; audio encoder, visual encoder, cross-modal encoder and multi-label classifier, see Fig. 1-(a).

**Audio encoder:** The audio encoder converts audio features  $\mathbf{S}$  into audio representations  $\mathbf{A}$  which are used in the cross-modal decoder. The audio representations are produced by

$$\mathbf{A}_{\text{co}} = \text{ConvolutionPooling}(\mathbf{S}; \theta_{\text{audio}}^{\text{co}}), \quad (4)$$

$$\mathbf{A}_{\text{po}} = \text{AddPosition}(\mathbf{A}_{\text{co}}), \quad (5)$$

$$\mathbf{A}_{\text{tr}} = \text{TransformerEnc}(\mathbf{A}_{\text{po}}; \theta_{\text{audio}}^{\text{tr}}), \quad (6)$$

$$\mathbf{A} = \text{AddAudioSegment}(\mathbf{A}_{\text{tr}}; \theta_{\text{audio}}^{\text{se}}), \quad (7)$$

where  $\{\theta_{\text{audio}}^{\text{co}}, \theta_{\text{audio}}^{\text{tr}}, \theta_{\text{audio}}^{\text{se}}\} \in \Theta$  is the trainable parameters of the audio encoder. ConvolutionPooling() is a function composed of convolution layers and pooling layers, AddPosition() is a function that adds a continuous vector in which position information is embedded, TransformerEnc() is a function of the transformer encoder blocks consisting of multi-head self-attention layers and position-wise feed-forward networks, and AddAudioSegment() is a function that adds a continuous vector in which speech segment information is embedded.

**Visual encoder:** The visual encoder converts visual features  $\mathbf{C}$  into visual representations  $\mathbf{V}$  which are used in the cross-modal decoder. The visual representations are produced by

$$\mathbf{V}_{\text{cnn}} = \text{CNN}(\mathbf{C}; \theta_{\text{visual}}^{\text{cnn}}), \quad (8)$$

$$\mathbf{V}_{\text{po}} = \text{AddPosition}(\mathbf{V}_{\text{cnn}}), \quad (9)$$

$$\mathbf{V}_{\text{tr}} = \text{TransformerEnc}(\mathbf{V}_{\text{po}}; \theta_{\text{visual}}^{\text{tr}}), \quad (10)$$

$$\mathbf{V} = \text{AddVisualSegment}(\mathbf{V}_{\text{tr}}; \theta_{\text{visual}}^{\text{se}}), \quad (11)$$

where  $\{\theta_{\text{visual}}^{\text{cnn}}, \theta_{\text{visual}}^{\text{tr}}, \theta_{\text{visual}}^{\text{se}}\} \in \Theta$  is the trainable parameters of the visual encoder. CNN() is a function that converts input RGB images into frame-by-frame features vectors. AddVisualSegment() is a function that adds a continuous vector in which visual segment information is embedded.

**Cross-modal encoder:** The cross-modal encoder handles outputs from the audio and visual encoders. It can flexibly switch whether to use both audio and visual features or either of them. The inputs for the cross-modal encoder  $\mathbf{Z}_0$  are

$$\mathbf{Z}_0 = \begin{cases} \text{Concat}(\mathbf{A}, \mathbf{V}) & \text{if audio-visual features are used,} \\ \mathbf{A} & \text{if audio features are only used,} \\ \mathbf{V} & \text{if visual features are only used,} \end{cases} \quad (12)$$

where Concat() is a function that concatenates inputs on the temporal axis. We produce cross-modal hidden representations  $\mathbf{Z}$  from

$$\mathbf{Z} = \text{TransformerEnc}(\mathbf{Z}_0; \theta_{\text{cross}}), \quad (13)$$

where  $\theta_{\text{cross}} \in \Theta$  is the trainable parameters of the cross-modal encoder. In this equation, the transformer encoder blocks perform cross-modal modeling of audio and visual representations if two modals are valid, while mapping the modal representations into cross-modal space if only the individual modals are available.

**Multi-label classifier:** The multi-label classifier computes emotion-wise conditional probabilities from the cross-modal representations  $\mathbf{Z}$ . They are simultaneously computed from

$$\mathbf{o} = \text{AttentivePooling}(\mathbf{Z}; \theta_{\text{label}}^{\text{att}}), \quad (14)$$

$$\bar{\mathbf{y}} = \text{Swish}(\mathbf{o}; \theta_{\text{label}}^{\text{sw}}), \quad (15)$$

$$\mathbf{y} = \text{Sigmoid}(\bar{\mathbf{y}}; \theta_{\text{label}}^{\text{sig}}), \quad (16)$$

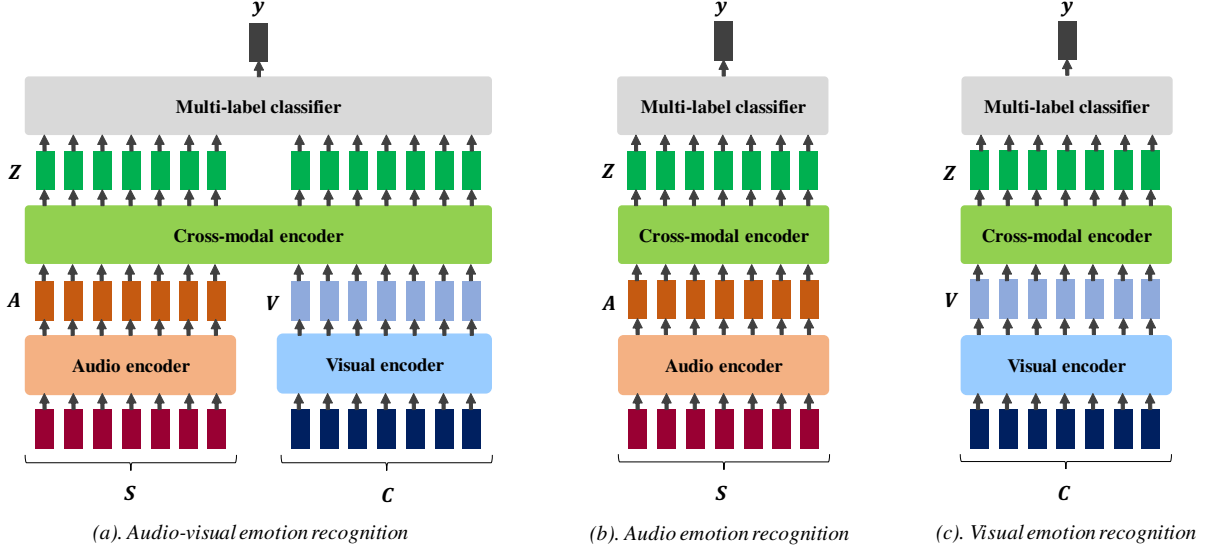


Figure 1: How to perform single-modal and multi-modal emotion recognition using cross-modal transformer.

where  $\{\theta_{\text{label}}^{\text{att}}, \theta_{\text{label}}^{\text{sw}}, \theta_{\text{label}}^{\text{sig}}\} \in \Theta$  is the trainable parameters of the label classifier.  $\text{AttentivePooling}()$  is the attentive pooling function,  $\text{Swish}()$  is a swish activation layer with a linear transformation and  $\text{Sigmoid}()$  is a sigmoid activation layer with a linear transformation. The  $k$ -th outputs in  $\mathbf{y}$  corresponds to  $P(l_k|\mathbf{S}, \mathbf{C}, \Theta)$  for the audio-visual emotion recognition,  $P(l_k|\mathbf{S}, \Theta)$  for the audio emotion recognition, and  $P(l_k|\mathbf{C}, \Theta)$  for the visual emotion recognition.

### 3.2. Interactive co-learning

To achieve the good performance in both audio-visual input and audio or visual input alone, the proposed method introduces a novel training scheme called interactive co-learning. The cross-modal model described in Section 3.1 is trained on three types of inputs; audio-visual features, audio features alone, and visual features alone. The last two are stimulated to drop features of the particular modal from  $\mathcal{D}$  in Eq. (2). In our interactive co-learning, we train model parameter set  $\Theta$  so as to enhance both single-modal and multi-modal emotion recognition performance. To this end, we define following functions:

$$\mathcal{L}_{AV}(\Theta) = - \sum_{t=1}^T \sum_{k=1}^K \log P(l_k^t | \mathbf{S}^t, \mathbf{C}^t, \Theta), \quad (17)$$

$$\mathcal{L}_A(\Theta) = - \sum_{t=1}^T \sum_{k=1}^K \log P(l_k^t | \mathbf{S}^t, \Theta), \quad (18)$$

$$\mathcal{L}_V(\Theta) = - \sum_{t=1}^T \sum_{k=1}^K \log P(l_k^t | \mathbf{C}^t, \Theta), \quad (19)$$

where  $\mathcal{L}_{AV}$ ,  $\mathcal{L}_A$ ,  $\mathcal{L}_V$  are the functions to evaluate the recognition performance for the audio-visual emotion recognition, audio emotion recognition, and visual emotion recognition, respectively. The interactive co-learning optimizes the model parameter set as

$$\hat{\Theta} = \underset{\Theta}{\text{argmin}} \{ \mathcal{L}_{AV}(\Theta) + \mathcal{L}_A(\Theta) + \mathcal{L}_V(\Theta) \}. \quad (20)$$

The trained parameters will be optimal in all cases of audio-visual input and audio or visual input alone. The optimization is achieved by mini-batch training where sample-level min-

Table 1: Distribution of CMU-MOSEI emotional categories

Emotion categories	Train	Validation	Test
<i>Happy</i>	8,735	1,005	2,505
<i>Sad</i>	4,269	520	1,129
<i>Anger</i>	3,526	338	1,071
<i>Surprise</i>	1,642	203	441
<i>Disgust</i>	2,955	281	805
<i>Fear</i>	1,331	176	385

batches are randomly sampled from one of audio-visual inputs, audio input alone and visual input alone.

## 4. Experiments

### 4.1. Dataset

For evaluation, we used CMU-MOSEI<sup>1</sup> [25]. It consists of approximately 23,000 annotated video clips made by more than 1,000 speakers. The task was the multi-label binary classification of the target emotions. The target emotions were *happiness*, *sadness*, *anger*, *surprise*, *disgust*, and *fear*. The dataset was split into training, validation, and test set following the official division<sup>2</sup>. The training, validation, and test sets held 16,327, 1,871, and 4,662 clips, respectively. Table 1 shows the number of samples by emotional category. *Happiness*, a positive emotion, tends to be the most common, while negative emotions tend to be the least common, a difference that is acceptable for machine learning tasks.

### 4.2. Setups

We constructed two baseline models, a conventional model and a proposed model. For the baseline models, we constructed an audio emotion recognition model and a visual emotion recognition model, each of which was trained from dataset with audio or visual features alone. They were composed from audio or visual encoder and multi-label classifier, described on sec-

<sup>1</sup>Although CMU-MOSEI provides manual transcriptions of audio cues, we do not use them as they are not available in actual situations. Thus, our setup is exactly audio-visual emotion recognition.

<sup>2</sup><https://github.com/A2Zadeh/CMU-multimodalSDK>

Table 2: Recognition performances for target emotions. “A”, “V”, and “A+V” represent the usage of the audio features alone, the visual features alone, and both of the audio and visual features. wF1 and mF1 are the weighted and macro F1s, respectively.

	Training	Inference	Happy		Sad		Anger		Surprise		Disgust		Fear		Average	
			wF1	mF1	wF1	mF1	wF1	mF1	wF1	mF1	wF1	mF1	wF1	mF1	wF1	mF1
Baseline	A	A	64.2	64.1	70.9	59.1	74.0	59.4	86.1	48.3	79.8	61.0	88.0	50.5	77.2	57.1
	V	V	60.5	60.4	70.2	57.0	74.9	61.0	85.9	48.0	78.2	57.0	87.8	49.8	76.3	55.5
Conventional	A+V	A	58.2	57.9	70.0	54.5	69.4	53.1	86.4	51.1	77.5	57.5	87.8	48.1	74.9	53.7
		V	61.9	61.7	69.9	57.0	73.7	61.6	86.1	49.8	76.4	58.4	87.4	50.7	75.9	56.5
		A+V	63.6	63.3	72.1	58.9	74.6	62.0	<b>86.2</b>	<b>50.3</b>	79.4	61.8	<b>88.0</b>	50.7	77.3	57.8
Proposed	Interactive co-learning	A	63.9	63.7	71.9	59.8	72.4	60.3	86.3	54.9	80.1	64.4	87.9	48.8	77.1	58.7
		V	60.0	60.2	68.5	56.0	74.2	61.6	86.1	48.0	77.8	58.0	87.5	53.9	75.7	56.3
		A+V	<b>66.0</b>	<b>66.0</b>	<b>72.2</b>	<b>62.2</b>	<b>75.5</b>	<b>63.8</b>	86.0	47.5	<b>81.0</b>	<b>66.5</b>	87.4	<b>55.0</b>	<b>78.0</b>	<b>60.2</b>

tion 3.1. For the conventional model, we constructed a cross-modal transformer model trained with Eq. (3). For the proposed model, we constructed a cross-modal transformer model trained with Eq. (20). The conventional and proposed models were evaluated on not only audio-visual emotion recognition but also audio or visual emotion recognition.

**Pre-processing:** We performed pre-processing to extract audio and visual features from video clips. For the acoustic features, we extracted 80 log Mel-scale filterbank coefficients appended with delta and acceleration coefficients as acoustic features. The frame shift was 10 ms. For the visual features, face regions in each input frame were detected with YOLOv3 [26] trained on the Wider Face dataset [27]. The face images were cropped and resized to  $128 \times 128$  and downsampled to 3 fps.

**Model configurations:** The model structures of acoustic encoder, visual encoder, cross-modal encoder and multi-label classifier were same among the baseline models, conventional model and proposed model. The configurations are as follows. For the audio encoder, audio features passed two convolution and max pooling layers with a stride of 2, so we down-sampled them to  $1/4$  along with the time axis. We stacked 6 transformer encoder blocks. For the visual encoder, the CNN function was composed from mobilenet-v3 [28]. After that, we stacked 2 transformer encoder blocks. For the cross-modal encoder, we stacked 2 transformer encoder blocks. For each transformer block, the dimensions of the output continuous representations were set to 128, dimensions of the inner outputs in the position-wise feed forward networks were set to 512, and number of heads in the multi-head attentions was set to 4. The Swish activation was used for the position-wise feed-forward networks. In the multi-label classifier, the outputs from the cross-modal encoder were weighted summed with the attention-weight obtained by the attention mechanism. After that, we introduced a fully-connected layer with swish activation function. The output layer was a fully-connected layer with the sigmoid activation function that outputs 6 classes of binaries.

**Training:** Before building multi-label emotion recognition models, some components were pre-trained with other datasets. First, all components in the audio encoder were pre-trained with end-to-end automatic speech recognition tasks using over 10K hours of speech. In addition, a CNN component in the visual encoder were pre-trained thorough two steps. It is pre-trained with a face recognition task using VGGFace2 [29] in the first step, and a still-image based facial expression recognition task using FER [30], RAF-DB [31], and AffectNet [32] datasets in the second step. Note that these pre-trained parameters were not fixed in the following main training. After these pre-training, multi-label emotion recognition models were trained. The mini-batch size was set to 16, and the dropout rate in the transformer blocks was set to 0.1. We used the RAdam [33] for optimization. The training steps were stopped based on early stopping using the validation set.

### 4.3. Results

Recognition performance for the target emotions are listed in Table 2. The evaluation metrics were weighted and macro F1 which are the weighted sum and the macro-average of F1 values, respectively.

The results for audio-visual inference show that the proposed method achieved better performance than the conventional model. The proposed model outperformed all other models except for *surprise* and *fear*. This may be due to the fact that *surprise* and *fear* have less data than the other categories, which causes a class imbalance problem. The proposed method does not use a special loss function to solve the class imbalance problem. However, it achieved the highest performance in the four emotional categories score and average score. These results indicate that by optimizing audio feature alone, visual features alone, and audio-visual features using interactive co-learning, more knowledge that contributes to emotion recognition was acquired than by the conventional method of optimizing only audio and visual features.

Comparing the average performances of the inferences with audio features alone, the proposed model improved the macro F1 from the baseline model, while the conventional model attained lower performance than baseline model. Comparing the average performances of the inferences with visual features alone, there is no significant difference between the proposed model and the conventional model, however the proposed model matched the performances of the baseline model. This indicates that the proposed method could utilize sufficient modality-specific knowledge of emotion recognition, while the baseline method results show that knowledge of individual modalities is not acquired by optimization on just the audio-visual paired features. In addition, comparing the proposed model by each inference data, the average performance is higher when audio-visual paired features are input than when audio or visual feature is input. This indicates that the proposed method yields better emotion recognition from the emotional cues of each modality when the input is multi-modal data.

## 5. Conclusion

This paper proposed a novel audio-visual emotion recognition method that allows the same model to utilizes both multi-modal and the single-modal information. We employed a cross-modal transformer model that can handle different types of inputs. In addition, we introduced a novel training method named interactive co-learning, which permits the model to acquire knowledge from both as well as either of the modals. Experiments showed that the proposed method provides better audio-visual emotion recognition performance than modeling that is trained so as to enhance only combined audio-visual processing. We also demonstrated that the proposed method achieved not only audio-visual multi-modal processing but also single-modal processing.

## 6. References

- [1] E. Avots, T. Sapiński, M. Bachmann, and D. Kamińska, “Audio-visual emotion recognition in wild,” *Machine Vision and Applications*, vol. 30, pp. 975–985, 2019.
- [2] K. Han, D. Yu, and I. Tashev, “Speech emotion recognition using deep neural network and extreme learning machine,” *In Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 223–227, 2014.
- [3] J. Lee and I. Tashev, “High-level feature representation using recurrent neural network for speech emotion recognition,” *In Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 1537–1540, 2015.
- [4] E. Sariyanidi, H. Gunes, and A. Cavallaro, “Automatic analysis of facial affect: A survey of registration, representation, and recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 6, pp. 1113–1133, 2015.
- [5] M. Pantic and L. J. M. Rothkrantz, “Automatic analysis of facial expressions: The state of the art,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, pp. 1424–1445, 2000.
- [6] Y. Li, J. Zeng, S. Shan, and X. Chen, “Occlusion aware facial expression recognition using cnn with attention mechanism,” *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2439–2450, 2019.
- [7] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor, “Emotion recognition in human-computer interaction,” *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32–80, 2001.
- [8] R. A. Calvo and S. D’Mello, “Affect detection: An interdisciplinary review of models, methods, and their applications,” *IEEE Transactions on Affective Computing*, vol. 1, no. 1, pp. 18–37, 2010.
- [9] D. Kollias and S. Zafeiriou, “Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arface,” *arXiv:1910.04855*, 2019.
- [10] Z. Sun, P. K. Sarma, W. Sethares, and E. P. Bucy, “Multi-modal sentiment analysis using deep canonical correlation analysis,” *In Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 1323–1327, 2019.
- [11] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou, “End-to-end multimodal emotion recognition using deep neural networks,” *IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING*, vol. 11, no. 8, pp. 1301–1309, 2017.
- [12] S. E. Kahou, X. Bouthillier, P. Lamblin, C. Gulcehre, V. Michalski, K. Konda, S. Jean, P. Froumenty, Y. Dauphin, N. Boulanger-Lewandowski *et al.*, “EmoNets: Multimodal deep learning approaches for emotion recognition in video,” *Journal on Multimodal User Interfaces*, vol. 10, no. 2, pp. 99–111, 2016.
- [13] V. Vielzeuf, S. Pateux, and F. Jurie, “Temporal multimodal fusion for video emotion classification in the wild,” *In Proc. ACM International Conference on Multimodal Interaction (ICMI)*, pp. 569–576, 2017.
- [14] A. Kumar and J. Vepa, “Gated mechanism for attention based multi modal sentiment analysis,” *In Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4477–4481, 2020.
- [15] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, “Multimodal machine learning: A survey and taxonomy,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2019.
- [16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, “Attention is All you Need,” *In Proc. Neural Information Processing Systems (NIPS)*, pp. 6000–6010, 2017.
- [17] H. Zhou, D. Meng, Y. Zhang, and X. Peng, “Exploring emotion features and fusion strategies for audio-video emotion recognition,” *In Proc. International Conference on Multimodal Interaction (ICMI)*, pp. 562–566, 2019.
- [18] T. Mittal, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, “M3ER: Multiplicative multimodal emotion recognition using facial, textual, and speech cues,” *In Proc. AAAI Conference on Artificial Intelligence (AAAI)*, pp. 1359–1367, 2020.
- [19] Y. Zhang, M. Chen, J. Shen, and C. Wang, “Tailor versatile multi-modal learning for multi-label emotion recognition,” *arXiv:2201.05834*, 2022.
- [20] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang, “VisualBERT: A simple and performant baseline for vision and language,” *arXiv:1908.03557*, 2019.
- [21] Z. Wang, J. Yu, A. W. Yu, Z. Dai, Y. Tsvetkov, and Y. Cao, “SIMVLM: Simple visual language model pretraining with weak supervision,” *arXiv:2108.10904*, 2021.
- [22] C. M. Christoudias, K. Saenko, L.-P. Morency, and T. Darrell, “Co-adaptation of audio-visual speech and gesture classifiers,” *In Proc. ACM International Conference on Multimodal Interaction (ICMI)*, pp. 84–91, 2006.
- [23] A. Zadeh, P. P. Liang, and L.-P. Morency, “Foundations of multimodal co-learning,” *Information Fusion*, vol. 64, pp. 188–193, 2020.
- [24] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Ng, “Multimodal deep learning,” *In Proc. International Conference on International Conference on Machine Learning (ICML)*, pp. 689–696, 2011.
- [25] A. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, “Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph,” *In Proc. Association for Computational Linguistics (ACL)*, pp. 2236–2246, 2018.
- [26] J. Redmon and A. Farhadi, “YOLOv3: An incremental improvement,” *arXiv:1804.02767*, 2018.
- [27] S. Yang, P. Luo, C. C. Loy, and X. Tang, “Wider face: A face detection benchmark,” *In Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5525–5533, 2016.
- [28] A. Howard, M. Sandler, B. Chen, W. Wang, L. Chen, M. Tan, G. Chu, V. Vasudevan, Y. Zhu, R. Pang, H. Adam, and Q. Le, “Searching for mobilenetv3,” *In Proc. International Conference on Computer Vision (ICCV)*, pp. 1314–1324, 2019.
- [29] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, “VG-Face2: A dataset for recognising face across pose and age,” *In Proc. IEEE International Conference on Automatic Face Gesture Recognition (FG)*, pp. 67–74, 2018.
- [30] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee, and *et al.*, “Challenges in representation learning: A report on three machine learning contests,” *In Proc. International Conference on Neural Information Processing (ICONIP)*, pp. 117–124, 2013.
- [31] L. Shan, D. Weihong, and D. JunPing, “Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild,” *In Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2584–2593, 2017.
- [32] A. Mollahosseini, B. Hasani, and M. H. Mahoor, “AffectNet: A database for facial expression, valence, and arousal computing in the wild,” *IEEE Transactions on Affective Computing*, vol. 10, pp. 18–31, 2019.
- [33] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and J. Han, “On the variance of the adaptive learning rate and beyond,” *In Proc. International Conference on Learning Representations (ICLR)*, 2020.