Interspeech 2024
1-5 September 2024, Kos, Greece

Participant-Pair-Wise Bottleneck Transformer for Engagement Estimation from Video Conversation

Keita Suzuki, Nobukatsu Hojo, Kazutoshi Shinoda, Saki Mizuno, Ryo Masumura

NTT Corporation, Japan
keitaxs.suzuki@ntt.com

Abstract

This study investigates the task of estimating the engagement of a target participant from video and audio during a multi-person conversation. For this task, interaction should be modeled effectively, considering the redundancy of video and audio across frames among multiple participants. Conventional Transformer-based methods in multimodal sentiment analysis succeeded in such efficient modeling by constraining the attention across multimodal data streams to go through only a small set of latent fusion units ("global tokens") that form an attention bottleneck. However, performance can be limited in the multi-person model because it needs to model interaction among a larger number of data streams based on only a single global token sequence. To address this problem, we propose a participant-pair-wise bottleneck transformer (PPBT) that involves multiple global token sequences, each of which is dedicated to a particular pair of participants and demonstrates its effect.

Index Terms: engagement, global token, multi-person conversation

1. Introduction

The use of online conferencing is becoming a tool of modern work. In particular, the COVID-19 pandemic triggered an increase in the number of companies utilizing remote and hybrid work, hence, the number of opportunities for online communication is on the rise. Many companies have continued to communicate online even after the end of COVID-19. However, there are drawbacks, such as the difficulty of maintaining participant motivation in situations where not everyone is face-to-face, increasing the need for engagement estimation and providing appropriate feedback.

In general, for an engagement estimation task, audio and video information is important, such as changes in tone and pitch of voice, facial expressions, and gestures [1]. It has also been shown that not only information about the target participant but also information about interactions with other participants is important [2].

To model the interaction among participants, previous studies proposed cross-person transformers (CPT) [2, 3]. By inputting feature token sequences from two different participants into a cross-attention transformer, they have shown that participant interaction could be effectively captured. On the other hand, it has been pointed out that cross-attention has challenges in representing relationships between data streams, such as video and audio [4, 5]. This is because video and audio have high redundancy across multiple frames, which makes a free attention flow between all the time steps of two different modalities in a cross-attention excessively complex. In this regard, in multimodal sentiment analysis, researchers proposed a multimodal bottleneck transformer (MBT) [4] to model the interaction between different input modalities. This method allows free attention flow within a modality but forces the model to condense information from each modality before sharing it. In particular, they introduced a small set of latent fusion units ("global tokens") that form an attention bottleneck through which cross-modal interactions within a layer must pass. This method enables efficient modeling because it does not take attention directly between modalities but takes attention through the mediation of defined global tokens.

This approach could also be useful not only for conventional multimodal models but also for multi-person models for engagement estimation. However, the performance can be limited in the multi-person model because it needs to model interaction among a larger number of data streams. In multimodal sentiment analysis, previous studies used global tokens for two (video and speech [5]) or three (text in addition [4]) data streams. In contrast, in the multi-person model, the global token needs to represent the interaction among the number of participants (five in this study). There is a concern that it is difficult to capture the interaction among such a large number of data streams based on only a single global token sequence.

To address this problem, we propose a new approach, participant-pair-wise bottleneck transformer (PPBT). PPBT uses multiple global token sequences, each of which is dedicated to a particular pair of participants. For example, for a five-party conversation, ten different global tokens are defined because ten pairs of two-party conversations can be made. Since each global token sequence models the interaction of two specific participants, the model training becomes easier. The contributions of this research are as follows;

- **Introduction of global tokens for engagement estimation from multi-party conversation:** This is a new method for capturing relationships among multiple participants. By using multiple global tokens, the dynamic dependencies among multiple participants can be modeled efficiently.
- **Proof of effect:** We showed that the introduction of multiple global tokens can achieve higher accuracy than conventional methods in an engagement estimation experiment based on the publicly available RoomReader corpus.

With these contributions, this research has set a new direction in predicting multi-person engagement and is expected to contribute to the development of future research.

10.21437/Interspeech.2024-1329

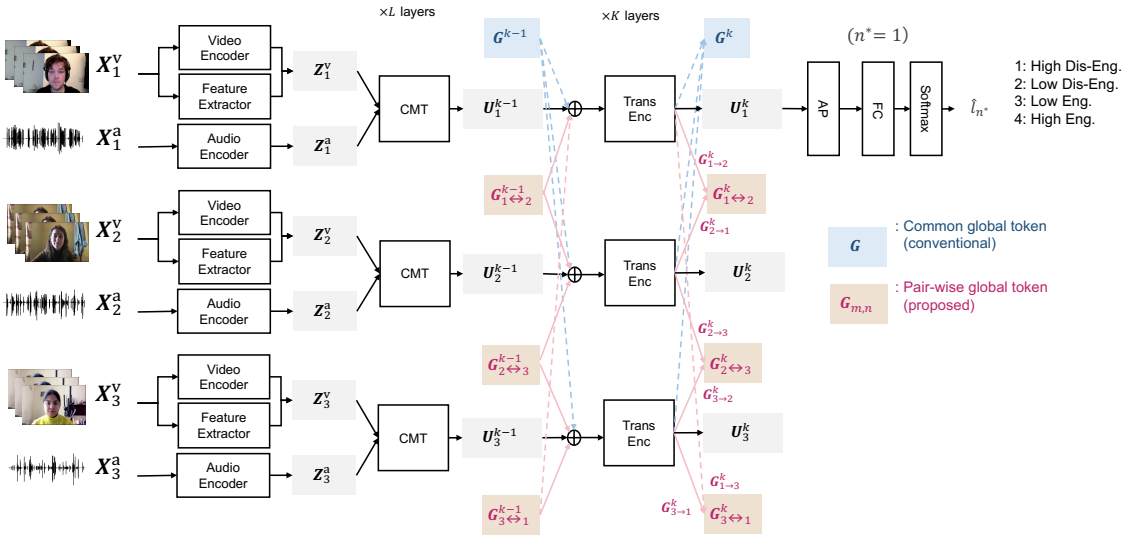


Figure 1: *Schematic diagram of the engagement estimation task and proposed method. "CMT", "Trans Enc", "AP" and "FC" denote a cross-modal Transformer, a Transformer Encoder, an attention pooling and fully-connected layer, respectively.*

2. Related Works

2.1. Engagement Estimation

Previous studies estimated engagement based on video and speech information of the target participant using CNN-Transformer and CNN-LSTM [6, 7]. The landscape of engagement estimation is changing, with recent years witnessing the adoption of bootstrapping and ensembling strategies, as evidenced by the introduction of models such as BOOT and ENSMODEL [8, 9]. These methods aggregate from several models to mitigate individual model biases and variances, which, in turn, result in more robust and reliable engagement predictions. Further, HTMIL leverages a Bi-LSTM with multi-scale attention along with clip-level and video-level objectives, while TEMMA adopts a Resnet-Transformer model [10, 11]. In contrast to our work, these methods don't consider the group setting in the model and rather focus on individual modeling. In this study, we focus on multi-person models that account for interactions.

2.2. Corpus for Engagement Estimation

The primary datasets traditionally utilized for engagement estimation within the scope of human-computer interaction research have predominantly been the RECOLA and Noxi corpus [12, 13]. These collections are foundational in the field, offering extensive insights into the dynamics of only the interaction of the two corresponding participants through meticulously captured dyadic conversations. The RECOLA corpus and the Noxi corpus are resources for researchers aiming to decode the subtle nuances of engagement between two individuals. In an effort to expand the horizons of engagement estimation beyond the confines of dyadic interactions, our study introduces the utilization of the RoomReader corpus as an approach to understanding engagement in more complex, multi-person conversation settings, as in the previous study [2].

3. Method

3.1. Task

Figure 1 shows an overview of the engagement estimation task. Let N be the number of all participants in a conversation. We estimate the engagement of a target participant $n^* \in \{1, \cdots, N\}$ at a given time t from the conversation data. In particular, we estimate engagement given a time interval $[t - D, t]$, i.e., a video and audio clip of length in D seconds are used as a context. In the following, we omit time t and denote $\boldsymbol{X}_n^{\mathrm{v}}$ and $\boldsymbol{X}_n^{\mathrm{a}}$ as the video and audio in the clip of speaker n. Also, let l_{n^*} denote the engagement label of the target speaker. In this study, we define the problem as a four-class classification problem with $l_n \in \{1$ (High Dis-Engagement), 2 (Low Dis-Engagement), 3 (Low Engagement), 4 (High Engagement)$\}$. The input information $\boldsymbol{X}$ is

$$\boldsymbol{X} = \{\boldsymbol{X}_1^{\mathrm{v}}, \boldsymbol{X}_1^{\mathrm{a}}, \cdots, \boldsymbol{X}_N^{\mathrm{v}}, \boldsymbol{X}_N^{\mathrm{a}}\}, \tag{1}$$

then, the engagement estimation problem is

$$\hat{l_{n^*}} = f(\boldsymbol{X}, n^*; \Theta), \tag{2}$$

where $f(\cdot)$ is a classification function defined by the model and Θ is its parameter. This setting is based on the previous study [2], except that we also use audio data streams.

3.2. Baseline Multi-person Bottleneck Transformer (MPBT)

The baseline MPBT model is a direct extension of a multimodal bottleneck transformer architecture to a multi-person model. It first uses pre-trained encoders to extract audio and visual features from the input information for each speaker $n \in \{1, \cdots, N\}$ using audio and video encoders.

$$\boldsymbol{Z}_n^{\mathrm{venc}} = \texttt{VideoEncoder}(\boldsymbol{X}_n^{\mathrm{v}}; \theta_{\mathrm{venc}}), \tag{3}$$

$$\boldsymbol{Z}_n^{\mathrm{a}} = \texttt{SpeechEncoder}(\boldsymbol{X}_n^{\mathrm{a}}; \theta_{\mathrm{a}}), \tag{4}$$

where $\texttt{VideoEncoder}(\cdot)$ and $\texttt{SpeechEncoder}(\cdot)$ are a projection function from data to the feature vector for video and

speech, respectively. θ_{venc} and θ_{a} are parameters of the encoders. $\boldsymbol{Z}_n^m \in \mathbb{R}^{D_{\mathrm{model}} \times T_m}$ denotes the feature vectors for the modality $m \in \{\mathrm{venc}, \mathrm{a}\}$, where D_{model} is the feature dimension and T_m is its time length. As in Multipar-T, we also obtain features such as head movements from video using OpenFace [14], which are added to the output of the video encoder.

$$\boldsymbol{Z}_n^{\mathrm{vfeat}} = \texttt{VideoFeatureExtractor}(\boldsymbol{X}_n^{\mathrm{v}}; \theta_{\mathrm{vfeat}}), \quad (5)$$

$$\boldsymbol{Z}_n^{\mathrm{v}} = \boldsymbol{Z}_n^{\mathrm{venc}} + \texttt{FC}(\boldsymbol{Z}_n^{\mathrm{vfeat}}; \theta_{\mathrm{FC}_1}), \quad (6)$$

where $\texttt{VideoFeatureExtractor}(\cdot)$ and θ_{vfeat} are a projection function from data to the feature vectors and its parameters, respectively. $\boldsymbol{Z}_n^{\mathrm{vfeat}} \in \mathbb{R}^{D_{\mathrm{vfeat}} \times T_{\mathrm{venc}}}$ denotes the feature vectors where D_{vfeat} is the feature dimension. $\texttt{FC}(\cdot)$ and θ_{FC_1} are a fully-connected layer and its parameters, respectively. The $\boldsymbol{Z}_n^{\mathrm{v}}$ and $\boldsymbol{Z}_n^{\mathrm{a}}$ are input to the cross-modal transformer (CMT), referring to [3].

$$\boldsymbol{S}_n^{m,0} = \boldsymbol{Z}_n^m \ (m = \mathrm{a}, \mathrm{v}), \quad (7)$$

$$\boldsymbol{S}_n^{\mathrm{a},l} = \texttt{TransformerEnc}(\boldsymbol{S}_n^{\mathrm{a},0}, \boldsymbol{S}_n^{\mathrm{v},l-1}; \theta_{\mathrm{a}\to\mathrm{v}}^{l-1}), \quad (8)$$

$$\boldsymbol{S}_n^{\mathrm{v},l} = \texttt{TransformerEnc}(\boldsymbol{S}_n^{\mathrm{v},0}, \boldsymbol{S}_n^{\mathrm{a},l-1}; \theta_{\mathrm{v}\to\mathrm{a}}^{l-1}), \quad (9)$$

where $\texttt{TransformerEnc}(\cdot)$ is the Transformer encoder block [15]. The first argument of the $\texttt{TransformerEnc}(\cdot)$ is used as a query, and the second argument is used as a key and value. $\theta_{\mathrm{a}\to\mathrm{v}}^l$ and $\theta_{\mathrm{v}\to\mathrm{a}}^l$ are model parameters of the l-th layer of the Transformer encoder block.

Next, the interaction of the participants is modeled by a global token. The outputs of CMTs are concatenated to form a feature vector $\boldsymbol{U}_n^0 \in \mathbb{R}^{D_{\mathrm{model}} \times (T_{\mathrm{v}} + T_{\mathrm{a}})}$ for each participant. To tame the quadratic complexity of attention, we next introduce a small set of bottleneck global tokens. We denote the global token used in the input layer as $\boldsymbol{G}^{(0)} = \theta_{\mathrm{g}} \in \mathbb{R}^{D_{\mathrm{model}} \times B}$, where D_{model}, B is the dimension and the length of the bottleneck tokens.

$$\boldsymbol{U}_n^0 = [\boldsymbol{S}_n^{\mathrm{a},L} || \boldsymbol{S}_n^{\mathrm{v},L}], \quad (10)$$

$$[\boldsymbol{U}_n^k || \boldsymbol{G}_n^k] = \texttt{TransformerEnc}([\boldsymbol{U}_n^{k-1} || \boldsymbol{G}^{k-1}]; \theta_{\mathrm{MPBT}}^{k-1}), \quad (11)$$

$$\boldsymbol{G}^k = \sum_n \boldsymbol{G}_n^k, \quad (12)$$

where $[\cdot||\cdot]$ is a matrix concatenation function along time axis and θ_{MPBT}^k is a model parameter of the k-th layer of the Transformer encoder block. The model first outputs global tokens for each participant $\boldsymbol{G}_n^k$, and then adds them together for all participants to update global tokens $\boldsymbol{G}^k$ to represent the interaction among all participants. Finally, pooling is performed on the output of the final layer of the Transformer encoder to obtain the posterior probabilities of the labels.

$$P(l_{n^*} | \boldsymbol{X}, n^*, \Theta) = \texttt{softmax}(\texttt{FC}(\texttt{AP}(\boldsymbol{U}_{n^*}^K; \theta_{\mathrm{AP}}); \theta_{\mathrm{FC}_2})), \quad (13)$$

where $\mathrm{softmax}(\cdot)$, $\mathrm{AP}(\cdot)$, θ_{FC_2} are a softmax, an attention pooling layer, fully connected parameters, respectively.

The model parameters Θ are optimized by minimizing cross-entropy loss using training data $\mathcal{D}$,

$$\Theta = \{\theta_{\mathrm{FC}_1}, \{\theta_{\mathrm{a}\to\mathrm{v}}^l, \theta_{\mathrm{v}\to\mathrm{a}}^l\}, \theta_{\mathrm{g}}, \{\theta_{\mathrm{MPBT}}^k\}, \theta_{\mathrm{AP}}, \theta_{\mathrm{FC}_2}\} \quad (14)$$

$$\mathcal{L} = \sum_{\boldsymbol{X}, n^*, l_{n^*} \in \mathcal{D}} -\log P(l_{n*} | \boldsymbol{X}, n^*, \Theta). \quad (15)$$

Note that the encoder parameters $\theta_{\mathrm{venc}}, \theta_{\mathrm{a}}$ and θ_{vfeat} are pretrained and frozen while training.

3.3. Proposed Participant-pair-wise Bottleneck Transformer (PPBT)

To precisely model the interaction between each pair of participants, the proposed PPBT model defines multiple global tokens, each of which corresponds to a single pair of participants. First, let $I = \{1, 2, \cdots, N\}$ denote the set of participant indices. A global token for each participant pair $\{m, n\}$ is denoted as $\boldsymbol{G}_{m\leftrightarrow n}$. We initialize each pair-wise global token used in the input layer as $\boldsymbol{G}_{m\leftrightarrow n}^{(0)} = \theta_{\mathrm{g}} \in \mathbb{R}^{D_{\mathrm{model}} \times B}$, where D_{model}, B is the dimension and the length of the bottleneck tokens. Instead of equation (11) in the baseline method, for each participant n, the proposed method concatenates the global token associated with n and inputs to a Transformer encoder layer,

$$[\boldsymbol{U}_n^k || \bigoplus_{m \in I \setminus n} \boldsymbol{G}_{n\to m}^k] = \texttt{TransformerEnc}([\boldsymbol{U}_n^{k-1} || \bigoplus_{m \in I \setminus n} \boldsymbol{G}_{m\leftrightarrow n}^{k-1}]; \theta_{\mathrm{PPBT}}^{k-1}), \quad (16)$$

where $\bigoplus$ denotes the vector concatenation ($\bigoplus_{i=1,2} A_i = [A_1 || A_2]$). θ_{PPBT}^k denotes the parameter of the k-th layer in the Transformer encoder block. The output $\boldsymbol{G}_{n\to m}^k$ denotes a variable that represents dependency from participant n to m. Each global token is updated by adding up variables representing the dependencies of the participants in both directions.

$$\boldsymbol{G}_{m\leftrightarrow n}^k = \boldsymbol{G}_{n\to m}^k + \boldsymbol{G}_{m\to n}^k \quad (17)$$

$$\boldsymbol{G}_{n\leftrightarrow m}^k = \boldsymbol{G}_{n\to m}^k + \boldsymbol{G}_{m\to n}^k. \quad (18)$$

The process of calculating the posterior probabilities from the output of the final layer is the same as that of the conventional MPBT.

4. Experiments

4.1. Experimetal Dataset

We used the RoomReader dataset [16]. The dataset includes multimodal, multi-party conversational interactions in which participants followed a collaborative online student-tutor scenario designed to elicit spontaneous speech. This dataset is processed to separate audio and video for each participant and synchronize video and audio. The resolution of the video is 2560 × 1440, the frame rate is 60 fps, and the sampling frequency of the audio is 32 kHz in 16-bit quantization. In the experiment, the frame rate was reduced to 8 fps. It also provides a continuous annotation for the engagement. This data set is labeled every second, and the label at the last second for each clip was used as a target. The labels are in the [-2, 2] range. Instead of regression, we define the task as a 4-class classification, where labels between (1, 2] refer to high engagement, (0, 1]: low engagement, (-1, 0]: low disengagement, [-2, -1]: high disengagement. We trained on 24 groups' data and tested on 6 groups. We used only clips that were successfully pre-processed for face region detection and all feature extraction (video feature extraction,

Table 1: *Frequency of each label in the training and test data.*

label	Train	Test	ratio (Train)	ratio (Test)
High Dis-Eng.	217	45	0.004	0.004
Low Dis-Eng.	815	368	0.015	0.029
Low Eng.	10910	1377	0.205	0.108
High Eng.	41250	10966	0.775	0.860

Table 2: *Evaluation results.*

Model	Bottleneck tokens		All Joint Engagement Classes			High Dis-Eng.	Low Dis-Eng.	Low Eng.	High Eng.
	B (Each Token)	Total	Accuracy	Weighted F1	Macro F1	F1	F1	F1	F1
CPT (Conventional)	-	-	0.513	0.603	0.223	0.000	0.030	0.186	0.677
MPBT (Baseline)	4	4	0.545	0.627	0.236	0.000	0.030	**0.213**	0.701
MPBT (Baseline)	40	40	0.623	0.677	0.237	0.000	0.031	0.145	0.768
PPBT (Proposed)	4	40	**0.766**	**0.780**	**0.313**	0.000	**0.171**	0.205	**0.876**

audio feature extraction, OpenFace feature extraction). Note that this resulted in a smaller number of clips used in the experiment than in the previous study. The number of video clips was 53,192 for training and 12,756 for testing.

4.2. Setups

Pre-processing: For video inputs, we detected face regions in each input frame with YOLOv3 [17] trained on the Wider Face dataset [18]. The size of each image is resized to 128 $\times$ 128. Table 1 shows the frequency of each label in the training and test set. We can see that there is a severe class imbalance. To counter the effects of class imbalance, we oversampled the infrequent class to balance the frequency. We did not use Focal Loss [19] because it was not effective in our experimental condition.

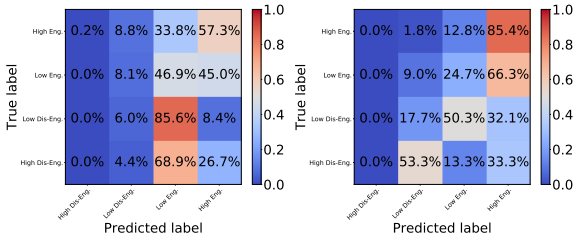
Encoder Configurations: For the video encoder, we used Resnet-50 [20]. For video features, we utilize the normalized eye gaze direction, location of the head, location of 3D landmarks, and facial action units extracted via OpenFace [14]. The xlsr-53 features from the final layer were used as audio features [21]. The feature dimensions were D_{venc}=2048, D_{a}=1024, D_{vfeat}=709. The length of each features was T_{venc}, T_{vfeat}=64, T_{a}=799.

Methods: We evaluated conventional CPT, MPBT and proposed PPBT. We first describe the setup common to all three models. We set the number of participants N as 5, the number of Transformer Encoder blocks l and k as 2. We used 8 seconds worth of video context information, i.e., D=8. The number of multi-head attention was 4. We used the relu activation function. We set D_{model} as 256. The batch size is 4, the learning rate is 0.0001, the optimizer is Radam, and Early stopping is applied [22].

For the Conventional CPT, instead of eq. (11) in the baseline method, we used a Transformer based on cross-attention and input the target participant's feature as keys and values and another participant's feature as queries, referring to the previous study [2]. The outputs of the five CPTs are combined in the feature dimension direction and used as $\boldsymbol{U}_{n^*}^{K}$ in eq. (13). For MPBT and PPBT, we set the length of the global tokens B as 4. To confirm that the effect of the proposed method is not simply due to the total length of the global token series being longer, we also set B as 40 for MPBT so that its total length is equal to the proposed method.

4.3. Results

Table 2 shows the results. The proposed method showed generally higher Accuracy, Weighted F1, and Macro F1 than the baseline and conventional methods, indicating that the proposed method of modeling the interaction among participants with multiple global tokens is effective. In particular, the proposed method improved F1 for Low Dis-eng. by 0.141 and High Eng. by 0.175, compared with the conventional and baseline method. The proposed method was also superior to the MPBT ($B = 40$). This demonstrates that the improvement by the proposed method was not simply due to the longer total length of the global tokens but to the use of a different global token for each participant pair. We also observe that all methods failed to estimate High Low-Eng in our experimental condition, unlike in the previous study [2]. This may be due to changes in the distribution of labels in the dataset caused by differences in data preprocessing and feature extraction methods, as well as differences in the details of upsampling. In Figure 2, the confusion matrix shows a tendency toward misclassification, and the proposed method is more accurate in classifying low disengagement as low engagement even if it cannot classify low disengagement. In addition, high engagement is less likely to be misclassified as low engagement or low disengagement in the high engagement category.



(a) *The baseline method.* (b) *The proposed method.*

Figure 2: *Confusion matrices.*

5. Conclusion

In this study, we presented the participant-pair-wise Bottleneck Transformer (PPBT) to capture information about how multiple participants are interacting. Our method captured the interaction between two corresponding participants via dedicated global tokens, which outperforms current systems that make use of a single global token common to all participants. Although this study focused on modeling the cross-person portion, it remains to be seen what would happen if the cross-modal portion were also modeled using global tokens. In the future, it is expected that the PPBT concept will be further developed and applied to a wider variety of scenarios and applications to further deepen our understanding of human interaction.

6. References

[1] H. Salam, O. Celiktutan, H. Gunes, and M. Chetouani, "Automatic Context-Aware Inference of Engagement in HMI: A Survey," *IEEE Transactions on Affective Computing*, pp. 1–20, 2023.

[2] D. W. Lee, Y. Kim, R. W. Picard, C. Breazeal, and H. W. Park, "Multipar-T: Multiparty-Transformer for Capturing Contingent Behaviors in Group Conversations," in *Proceedings of the International Joint Conference on Artificial Intelligence*, 2023, pp. 3893–3901.

[3] Y. Kim, D. W. Lee, P. P. Liang, S. Alghowinem, C. Breazeal, and H. W. Park, "HIINT: Historical, Intra-and Inter-personal Dynamics Modeling with Cross-person Memory Transformer," in *Proceedings of the ACM International Conference on Multimodal Interaction*, 2023, pp. 314–325.

[4] L. Sun, Z. Lian, B. Liu, and J. Tao, "Efficient Multimodal Transformer with Dual-Level Feature Restoration for Robust Multimodal Sentiment Analysis," *IEEE Transactions on Affective Computing*, 2023.

[5] A. Nagrani, S. Yang, A. Arnab, A. Jansen, C. Schmid, and C. Sun, "Attention Bottlenecks for Multimodal Fusion," in *Proceedings of the International Conference on Neural Information Processing Systems*, 2021, pp. 14 200–14 213.

[6] Y.-Y. Li and Y.-P. Hung, "Feature Fusion of Face and Body for Engagement Intensity Detection," in *Proceedings of the IEEE International Conference on Image Processing*, 2019, pp. 3312–3316.

[7] Y. Xiong, G. Xinya, and J. Xu, "CNN-Transformer: A deep learning method for automatically identifying learning engagement," *Education and Information Technologies*, pp. 1–20, 2023.

[8] K. Wang, J. Yang, D. Guo, K. Zhang, X. Peng, and Y. Qiao, "Bootstrap Model Ensemble and Rank Loss for Engagement Intensity Regression," in *In proceedings of the ACM International Conference on Multimodal Interaction*, 2019, pp. 551–556.

[9] V. Thong Huynh, S.-H. Kim, G.-S. Lee, and H.-J. Yang, "Engagement Intensity Prediction withFacial Behavior Features," in *In proceedings of the ACM International Conference on Multimodal Interaction*, 2019, pp. 567–571.

[10] J. Ma, X. Jiang, S. Xu, and X. Qin, "Hierarchical Temporal Multi-Instance Learning for Video-based Student Learning Engagement Assessment," in *In Proceedings of the International Joint Conference on Artificial Intelligence*, 2021, pp. 2782–2789.

[11] H. Chen, D. Jiang, and H. Sahli, "Transformer Encoder With Multi-Modal Multi-Head Attention for Continuous Affect Recognition," *IEEE Transactions on Multimedia*, vol. 23, pp. 4171–4183, 2020.

[12] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions," in *Proceedings of the IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, 2013, pp. 1–8.

[13] A. Cafaro, J. Wagner, T. Baur, S. Dermouche, M. Torres Torres, C. Pelachaud, E. André, and M. Valstar, "The Noxi database: multimodal recordings of mediated novice-expert interactions," in *Proceedings of the ACM International Conference on Multimodal Interaction*, 2017, pp. 350–359.

[14] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "OpenFace: An open source facial behavior analysis toolkit," in *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, 2016, pp. 1–10.

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is All you Need," in *Proceedings of the International Conference on Neural Information Processing Systems*, 2017, pp. 6000–6010.

[16] J. Reverdy, S. O'Connor Russell, L. Duquenne, D. Garaialde, B. R. Cowan, and N. Harte, "RoomReader: A multimodal Corpus of Online Multiparty Conversational Interactions," in *Proceedings of the International Conference on Language Resources and Evaluation Conference*, 2022, pp. 2517–2527.

[17] J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[18] S. Yang, P. Luo, C.-C. Loy, and X. Tang, "WIDER FACE:A Face Detection Benchmark," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5525–5533.

[19] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal Loss for Dense Object Detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2980–2988.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[21] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised Cross-Lingual Representation Learning for Speech Recognition," in *Proceedings of the Annual Conference of the International Speech Communication Association*, 2021, pp. 2426–2430.

[22] 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 *Proceedings of the International Conference on Learning Representations*, 2020.