TCA-NET: TRIPLET CONCATENATED-ATTENTIONAL NETWORK FOR MULTIMODAL ENGAGEMENT ESTIMATION

Hongyuan He[⋆†] *Daming Wang*[⋆†] *Md Rakibul Hasan*[‡] *Tom Gedeon*[†‡] *Md Zakir Hossain*[†‡]

[†] Australian National University, Canberra, ACT 2601, Australia
[‡] Curtin University, Perth, WA 6102, Australia

2024 IEEE International Conference on Image Processing (ICIP)

ABSTRACT

Human social interactions involve intricate social signals that artificial intelligence and machine learning models aim to decipher, particularly in the context of artificial mediators that can enhance human interactions across domains like education and healthcare. Engagement, a key aspect of these interactions, relies heavily on multimodal information like facial expressions, voice and posture. Recently, many deep learning methods have been deployed in engagement estimation. Still, they often focus on unimodality or bimodality, leading to the results lacking robustness and adaptability due to factors like noise and varying individual responses. To address this challenge, we introduce a novel modality fusion framework named Triplet Concatenated-Attentional Net (*TCA-Net*). This framework takes three distinct types of data modality (video, audio and Kinect) as inputs and delivers a prediction score as output. Within this network, a specially designed concatenated-attention fusion mechanism serves the purpose of modality fusion and preserves the intra-modal features. Experimental results validate the efficiency of our *TCA-Net* in enhancing the accuracy and reliability of engagement estimation across diverse scenarios, with a test set Concordance Correlation Coefficient (CCC) of 0.75. We release our code at `https://github.com/Daming-W/Multimodal_Engagement_Estimation`.

***Index Terms*—** Engagement estimation, attention network, multimodal fusion, human interaction, deep learning

1. INTRODUCTION

Engagement is the process through which multiple participants initiate, sustain and end their perceived connection. The capacity of artificial mediators to perceive and estimate users' engagement is vital for facilitating prompt, lifelike and emotion-aware interactions with users, making them companions for educational and therapeutic purposes [3]. Estimation of user engagement enables systems to achieve real-time intervention and interaction [4]. Since it is a multidimensional concept, the definition, annotation and automated prediction are focal points of research. Traditional methods have used nonverbal engagement cues, including facial expression, gaze patterns, body posture, proxemics and task-related behaviours to construct non-parametric classifiers for engagement states [5, 6]. When big gaps exist between and within target people, these models often fail to fit everyone.

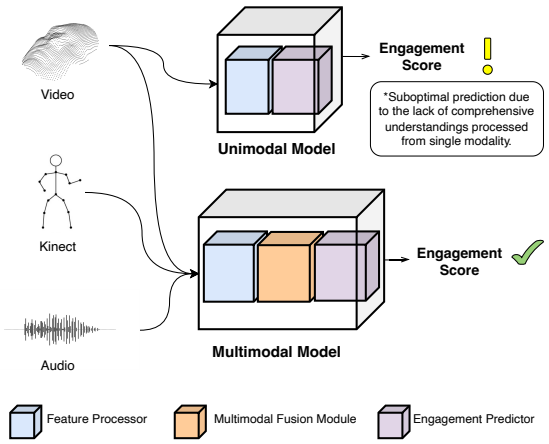

Fig. 1. **Top**: Previous unimodal engagement estimation methods [1, 2], mainly adopting the facial feature extracted from captured video clips. **Bottom**: Proposed end-to-end multimodal engagement estimation method which allows inputting and fusing multiple modality features and computing corresponding engagement scores.

Deep learning models provide state-of-the-art performance in many applications, including object detection, emotion recognition, image classification [7] and computational social science such as empathy [8] and personality [9]. Inspired by their performance, some methods apply deep learning architecture to challenge the difficulties in engagement estimation. Active learning and reinforcement learning are employed to mitigate the person-specific styles of engagement expressions in [10]. A recurrent neural network is applied to capture the temporal dynamics in [2]. These studies demonstrate efficient improvement in the personalised estimation of engagement, but they still lack robustness across diverse individuals [11].

[⋆]HONGYUAN HE AND DAMING WANG ARE CO-FIRST AUTHORS.

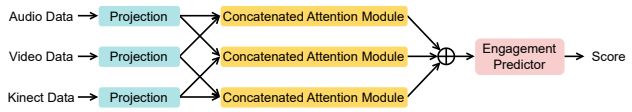

Fig. 2. An overview of the proposed *TCA-Net* model: the embeddings of three modalities are standardised in scale and proceeded modality-specific learning through projectors. The computed modalities are fused pairwise using the proposed concatenated-attention modules. The fused results, once concatenated, are then passed into the predictor for prediction.

Human behaviours can be expressed through various modalities, such as voice, action, face, text and physiology, which carry complementary information. Processing only a single modality leads to a lack of comprehensive understanding and can be easily affected by noisy data. In contrast, multimodal systems have emerged as a promising solution to these challenges. Figure 1 illustrates the difference between unimodal and multimodal models in the engagement estimation task. Integrating information from various sources makes the systems more robust to noise, offering higher accuracy and adaptability to different contexts and individuals. Studies such as [12, 11, 13, 14] have proved the notable advancements of modality fusion framework for engagement estimation and other human behaviours.

This paper proposes a novel modality fusion framework, *TCA-Net*, to address this challenge. *TCA-Net* accepts multimodal feature representations to estimate engagement scores. The core component of *TCA-Net* – the concatenated attention fusion module – learns both intra- and inter-modality correlations and computes the fused embeddings. This work presents that the *TCA-Net* improves engagement estimation performance on the NOvice eXpert Interaction (NoXi) dataset [15]. Besides, the ablation study investigates the importance of each involved modality feature in this task. Our major contributions include: **(1)** A novel modality fusion framework, *TCA-Net*, to estimate participants' engagement, **(2)** A concatenated-attention fusion module to combine different modalities and **(3)** A study on modality significance in the engagement estimation task.

2. RELATED WORK

Engagement estimation is a task to predict user engagement or interest in a given context. It is crucial in various applications, such as recommendation systems, personalised advertising and user behaviour analysis. Early works primarily adopt unimodal approaches, only using one data type. For example, [6] used body posture to predict engagement, [16] studied gaze data to recognise conversational engagement and [10] combined Long Short-Term Memory network with reinforcement learning to estimate a child's engagement level. A vision transformer was applied to predict student engagement in [17]. Due to the limitation of only focusing on one data type, these models often lack accuracy and robustness.

Recent works have explored multimodal data, combining information from various sources, such as visual, auditory and textual cues, to enhance robustness and accuracy. For instance, [14] combined learning-based and rule-based approaches to evaluate the estimation based on multimodalities. [12] implemented Hybrid Majority Voting by fusing three modalities – appearance, context and mouse movements – to detect students' behavioural engagement. Some studies have experimented with different multimodal approaches on the NoXi datasets. For example, [18] proposed the Dilated Convolutional Transformer Model, which processed three data types with a dilated convolution module and merged by attention-based or gated-based fusion. [19] used a feedforward fully connected (FC) network with engineered video and audio features. Finally, [20] adopted the subsequence feature to learn engagement representation with Seq2seq modelling.

Multimodal fusion integrates information from multiple modalities into a stable representation. Bilinear pooling fuses visual and textual feature vectors to obtain a joint representation space by calculating the outer product and pooling [21]. However, operations like outer product and pooling may result in some loss of information. An attention mechanism, on the other hand, is widely adopted for flexible and task-specific multimodal fusion [13]. This paper proposes a novel attention-based fusion model that can learn both inter- and intra-modality correlations.

3. METHOD

3.1. Task Formulation

The engagement estimation task aims to continuously predict each participant's conversational engagements, the confidential score $\boldsymbol{S}$. The score is computed from multiple data sources, including head pose, body pose and human voice.

Estimation of participant engagement in the task involves three main challenges. Firstly, the information fusion $\hat{X}$ from different modalities $X_i \in \mathbb{R}^{N \times L}$ presents a complex problem, where N represents the number of input feature representations and L denotes the embedding length of each feature vector. There may exist heterogeneity and inconsistency between different modalities, necessitating an effective integration and accurate estimate of participants' engagement. Secondly, the data from different modalities may be imbalanced, meaning certain modalities may have significantly more feature information than others. Lastly, due to the specificity of multimodal data, multimodal deep models require a balance between model complexity and performance, along with the selection of appropriate training techniques to avoid overfitting and training difficulties. To address these challenges, this paper proposes a novel multimodal model based on the attention-based fusion mechanism, termed *TCA-Net*.

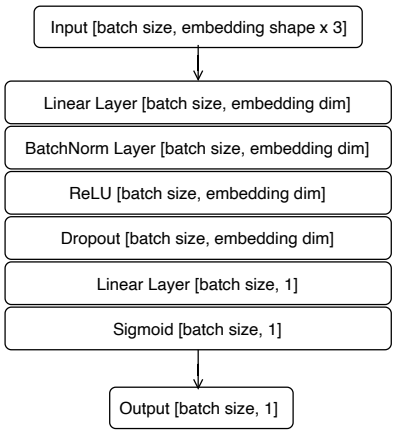

Fig. 3. The *TCA-Net* employs a feedforward MLP as the predictor, which accepts fused embeddings and computes the engagement score. Additionally, batch normalisation and Dropout are applied to avoid potential overfitting.

3.2. Overall TCA-Net

This work proposed a multimodal deep-learning network to address the engagement estimation task. Based on the specificity of the multimodal dataset and to enhance the integration and utilisation of informative data, the proposed model first embeds the input from multiple modalities with audio $X_a \in \mathbb{R}^{d_a}$, video $X_v \in \mathbb{R}^{d_v}$ into corresponding projection layers to align the dimensions of each modality input, which can be represented as below where $P_a = P_v = P_k$:

$$P_a = \text{Projection}_{\text{audio}}(d_a) \quad (1)$$

$$P_v = \text{Projection}_{\text{video}}(d_v) \quad (2)$$

$$P_k = \text{Projection}_{\text{kinect}}(d_k) \quad (3)$$

The proposed *TCA-Net*, as shown in Figure 2, performs pairwise learning on the available multimodal information, which employs the concatenated-attention fusion module as the core component to fuse the three types of input data.

The concatenated-attention fusion module performs attention learning on the two input modalities and outputs the fusion results. Therefore, the entire network utilises three modules to conduct parallel learning for different modality pairs. The outputs of the three fusion modules are concatenated to represent the joint representation of all modalities.

The fused joint feature embeddings are then passed to the last prediction stage, where the engagement score encoder consists of a shallow feedforward network, as shown in Figure 3. Dropout is used in the middle to prevent overfitting, possibly caused by multiple stacked FC. Finally, a Sigmoid activation function is used for output activation, mapping to the 0–1 range.

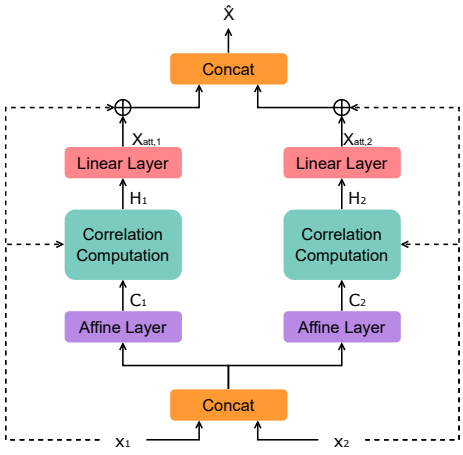

Fig. 4. Proposed concatenated-attention fusion module. After concatenating the features together, it is separated into different affine layers to generate the correlation matrices $\boldsymbol{C_1}$ and $\boldsymbol{C_2}$. These matrices are then integrated with their initial features to calculate the modality correlations $\boldsymbol{H_1}$ and $\boldsymbol{H_2}$ using specialised correlation computation blocks. Finally, applying the linear layers and adding the original modality feature, the formed attended features $\boldsymbol{X_{att,1}}$ and $\boldsymbol{X_{att,2}}$ are concatenated to produce $\hat{\boldsymbol{X}}$, which serves as the output of this module.

3.3. Concatenated-Attention Fusion

The deep features of the three input modalities – audio, video and Kinect – were pre-extracted or recorded. Each modality contains participants' perspectives and physical information. Video modality conveys rich appearance information about the engagement, audio modality carries the energy relevant to the intensity of the engagement, and Kinect modality contains information on the location of various joints.

These three modalities provide relevant information at different levels for a certain sequence. Multiple modalities convey more diverse and comprehensive information for engagement than a single modality. A concatenated-attention fusion mechanism is proposed to reliably fuse these modalities, which can effectively encode the intermodal data and preserve intra-modal features. In this mechanism, the joint representations are concatenated by any two types of modality features, such as video-audio, video-Kinect and audio-Kinect. In this way, the fused dual-modality feature includes both individual modality information and the captured relationship between the two modalities. This proposed concatenated-attention fusion module is shown in Figure 4.

The proposed concatenated-attention fusion module takes feature vectors from two identical-scaled modalities. After analysing the joint embedding using affine layers, correlations are captured with each input modality. These inter-modality correlations are computed and further transformed as attention weights through FC layers and combined with the corresponding input modality embeddings. At last, the merged

results are the produced fusion feature, which will be applied for further prediction.

Let $\boldsymbol{X_1}$ and $\boldsymbol{X_2}$ represent two sets of deep features extracted from two different modalities, where $\boldsymbol{X_1} = \{\boldsymbol{x_1^1}, \boldsymbol{x_1^2}, \cdots, \boldsymbol{x_1^N}\} \in \mathbb{R}^{N\times L}$ and $\boldsymbol{X_2} = \{\boldsymbol{x_2^1}, \boldsymbol{x_2^2}, \cdots, \boldsymbol{x_2^N}\} \in \mathbb{R}^{N\times L}$. N represents the number of the input feature representations, L denotes the embedding length of each feature vector, x_1^n and x_2^n represent the feature vectors extracted from the two modalities under consideration, respectively, for $n = 1, 2, \cdots, N$ samples. A compact concatenated attention module is proposed here to make the model consider and utilise more comprehensive information from both modalities and alleviate the heterogeneity between modalities. The joint representation $\boldsymbol{J}$ is obtained by concatenating two input feature representations:

$$\boldsymbol{J} = [\boldsymbol{X_1}, \boldsymbol{X_2}] \in \mathbb{R}^{N\times 2L}. \tag{4}$$

Now the joint representation $\boldsymbol{J}$ gets through the affine layers to obtain the joint correlation matrices $\boldsymbol{C_1}$ and $\boldsymbol{C_2}$, respectively, which are given by:

$$\boldsymbol{C_1} = \tanh(\frac{\boldsymbol{JW_{j1}}}{\sqrt{2L}}), \tag{5}$$

$$\boldsymbol{C_2} = \tanh(\frac{\boldsymbol{JW_{j2}}}{\sqrt{2L}}), \tag{6}$$

where $\boldsymbol{W_{j1}}, \boldsymbol{W_{j2}} \in \mathbb{R}^{2L\times L}$ represent learnable weight matrices, enabling the model to learn the most useful representations automatically.

The joint correlation matrices $\boldsymbol{C_1}, \boldsymbol{C_2} \in \mathbb{R}^{N\times L}$ provide a semantic measure of relevance across modalities. This type of measurement makes it better to understand and utilise the inter- and intra-modal relationships. Meanwhile, the higher correlation coefficients within the joint correlation matrices suggest that the associated samples strongly correlate across different modalities.

To enhance the expressiveness, the original deep features and the joint correlations are combined through the non-linear transformation to compute the attention weights of modalities by crossing the correlation computation layers. For the modality feature $\boldsymbol{X_1}$, it is combined with its corresponding joint correlation matrix $\boldsymbol{C_1}$ by using the learnable weight matrix $\boldsymbol{W_1}$ and $\boldsymbol{W_{c1}}$, which can be written as:

$$\boldsymbol{H_1} = \mathrm{ReLU}(\boldsymbol{X_1W_1} + \boldsymbol{C_1W_{c1}}), \tag{7}$$

where $\boldsymbol{W_1} \in \mathbb{R}^{L\times L}$, $\boldsymbol{W_{c1}} \in \mathbb{R}^{L\times L}$, and $\boldsymbol{H_1}$ indicates the attention map of the corresponding modality. The ReLU activation function helps to capture complex patterns in the data.

Similarly, the attention map $\boldsymbol{H_2}$ of the other modality is given by:

$$\boldsymbol{H_2} = \mathrm{ReLU}(\boldsymbol{X_2W_2} + \boldsymbol{C_2W_{c2}}), \tag{8}$$

where $\boldsymbol{W_2} \in \mathbb{R}^{L\times L}$ and $\boldsymbol{W_{c2}} \in \mathbb{R}^{L\times L}$.

Table 1. Modalities and Dimensions of the NoXi Dataset

Modality	Dimension
Audio (GeMAPS)	58
Audio (SoundNet)	256
Video (OpenFace2)	673
Video (OpenPose)	350
Kinect (Skeleton)	350
Kinect (AU)	17

Finally, after employing the learnable weighted matrices $\boldsymbol{W_{h1}}$ and $\boldsymbol{W_{h2}}$ through the linear layers, the attended features of these two modalities are given by:

$$\boldsymbol{X_{att,1}} = \boldsymbol{H_1W_{h1}} + \boldsymbol{X_1}, \tag{9}$$

$$\boldsymbol{X_{att,2}} = \boldsymbol{H_2W_{h2}} + \boldsymbol{X_2}, \tag{10}$$

where $\boldsymbol{W_{h1}} \in \mathbb{R}^{L\times L}$ and $\boldsymbol{W_{h2}} \in \mathbb{R}^{L\times L}$.

The attended features $\boldsymbol{X_{att,1}}$ and $\boldsymbol{X_{att,1}}$ are further concatenated as:

$$\hat{\boldsymbol{X}} = [\boldsymbol{X_{att,1}}, \boldsymbol{X_{att,2}}] \in \mathbb{R}^{N\times 2L}. \tag{11}$$

The fused feature embeddings $\hat{\boldsymbol{X}}$ will be joined with other concatenated-attention modules' outputs and feedforward to the predictor, as mentioned in the previous section.

4. EXPERIMENTS

4.1. Dataset

To demonstrate the performance of the *TCA-Net* and its feature fusion capabilities, this work conducts experiments with a recent dataset named NOvice eXpert Interaction (NoXi) [15]. This dataset was employed in the 2023 MultiMediate engagement estimation challenge[1] [22]. The confidential score (degree of engagement) $\boldsymbol{S}$ in this dataset ranges from 0 (lowest) to 1 (highest).

The NoXi dataset recorded multimodal information contained in 64 conversation sessions by experts and novices, which were 25 hours and 18 minutes long. The record's content simultaneously possesses characteristics of multiple languages, various topics (58 topics in diverse domains) and multiple features. The providers applied semi-automated solutions and proposed a novel annotation tool, NOVA. The NoXi dataset has 2,502,433 frame annotations in total with both visual and audio data. Furthermore, the recorded data can be counted as the signal types with specific embedding dimensions shown in Table 1.

The engagement estimation challenge with the NoXi Dataset employed the Concordance Correlation Coefficient (CCC) as the official evaluation metric. CCC is a statistical measure that assesses the agreement between two sets of

[1] https://multimediate-challenge.org/

Thanks to the challenge organiser and the NoXi dataset provider.

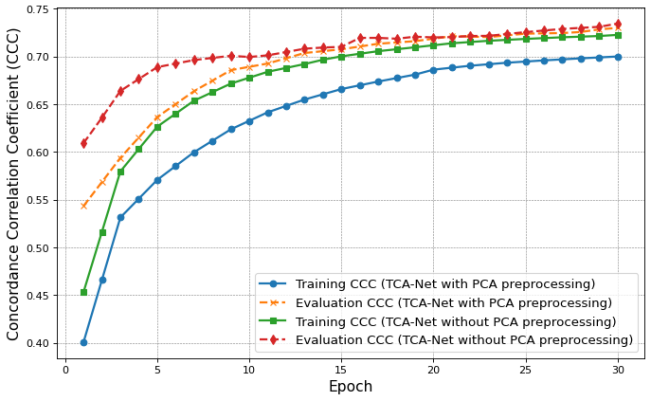

Fig. 5. Comparison of training and evaluation CCC for *TCA-Net* with and without PCA preprocessing methods.

continuous variables. It reflects the accuracy of the model's predictions and emphasises the consistency between the predicted values and the actual data.

4.2. Implementation Details

The processed modal embeddings have the following feature dimensions: 314 for audio features, 1023 for video features and 367 for Kinect features. Additionally, for computational convenience, the *TCA-Net* utilises projection layers for each of the three modal features to obtain embeddings of the same dimension with 256 for projection embeddings size, which then serve as inputs to the concatenated-attention module.

We implement *TCA-Net* by using Python 3.8.5 and PyTorch 1.9.0 framework. The *TCA-Net* is trained from scratch for 30 epochs with a batch size of 256. The motivation behind the training process is to minimise the Mean Squared Error (MSE) loss function. During training, the AdamW optimiser is utilised, along with a cosine annealing learning rate scheduler, which suppresses from 5×10^{-6} to 5×10^{-7} to achieve better convergence of the loss. To control potential overfitting, a weight decay of 1×10^{-4} is set, and a Dropout mechanism with a 0.25 ratio is used to randomly select fully-connected layers' neurons. We employed two NVIDIA V100 GPUs to train the proposed *TCA-Net* on the NoXi dataset.

4.3. Results and Analysis

We compare *TCA-Net* 's performance with the baseline provided by the 2023 MultiMediate challenge organiser [22] and the models reported by challenge participants [18, 19, 20] (Table 2). The baseline approach fuses the features through a straightforward linear fusion method after Principal Component Analysis (PCA) preprocessing. The engagement score is then trained (supervised) using a feedforward neural network. To align with the settings of the baseline study [22], we maintained the same PCA operation during data preprocessing to

Table 2. Comparison of CCC Scores on Validation and Test Sets

Method/Reference	Validation CCC	Test CCC
Head [AUs] [22]	0.31	0.22
Body [OpenPose] [22]	0.53	0.43
Voice [GeMAPS] [22]	0.58	0.55
Baseline with PCA [22]	0.71	0.59
DCTM [18]	0.75	0.66
FC [19]	0.74	0.70
Seq2seq [20]	–	0.71
TCA-Net (ours)	**0.73**	**0.75**
TCA-Net with PCA (ours)	**0.73**	**0.74**

cope with the validation bias that might be brought from PCA. This PCA operation was applied to reduce the dimensionality of all modal feature embeddings. Consequently, in this group experiments, adjustments were made to the projection layers of the *TCA-Net* to modify the input dimensions and maintain the rectified output dimensions of the multimodal features consistent with the aforementioned setup. The details of training and evaluation CCC for both methods, shown in Figure 5, present the proper learning progress of our proposed model. Because of the application of anti-overfitting techniques, such as Dropout and batch normalisation, there are gaps between the training and evaluation curves.

By comparing the results of *TCA-Net* trained with modality embeddings that have undergone PCA dimensionality reduction and those that have not, it is evident that there is still a slight decline in performance after PCA processing (Figure 5). While PCA removes noise and redundant information from the data, it also leads to the loss of information valuable for engagement prediction. Direct dimension mapping through the projectors also enhances the model's capability to capture the intricate structure within each modality data.

Our network's test set performance outperforms the baseline method and some visible challenge participants of the 2023 MultiMediate challenge (Table 2). It can be inferred that our concatenated attention module plays a significant facilitative role in integrating modalities. Through pairwise learning, the obtained inter-modality correlations serve as attention weights, enabling module design to achieve the desired fusion effect in multiple modalities.

The complexity of our model is 160.865 GFLOPS. Therefore, it can run on common devices like the RTX 1080Ti (having 11.34 rated TFLOPS), which took an inference time of 0.01418 seconds in our experiment.

From the perspective of network structure design, the *TCA-Net* adopts several FC layers for inter-modality correlation computations. Dropout and batch normalisation are, therefore, widely used to solve the risk of overfitting, which might not be an elegant solution. Future work can consider optimising the structure by replacing the FC layers with another alternative, such as the 1×1 convolution layer, global pooling layer, etc. In addition, the design of the engagement

Table 3. Engagement Estimation Performance of *TCA-Net* on Different Modality Types of the NoXi Dataset

Modality Types	Validation CCC	Test CCC
Audio + Video	0.59	0.61
Audio + Kinect	0.65	0.65
Video + Kinect	0.47	0.50

prediction head can also be upgraded to a deeper decoder.

4.4. Ablation Study

We investigate the contributions of each modality by validating the proposed network by training with two of the three types of modal feature embeddings in the NoXi dataset. Simplifying *TCA-Net* to incorporate only two modal data for engagement estimation, we assess the individual contributions of each modality to this task.

The setup for ablation experiments remained consistent with previous work. The correlation between engagement estimation and modality representation was analysed by comparing the CCC of bimodal fusion predictions (Table 3).

The comparison of performance from bimodal fusion shows that the audio modality significantly contributes to the engagement estimation task, as the performance in the experimental group lacking audio is inferior to the others. The above results also potentially indicate an overlap of the information in video and Kinect modalities, with limited complementarity. This observation could account for the lower performance of the experimental group (video and Kinect).

By observing Table 2 in the preceding experimental results section, similar viewpoints to this ablation study can be derived. In the baseline method, researchers evaluated using individual modalities for engagement estimation. The evaluation results also demonstrated that using the voice modality [GeMAPS] alone yielded significantly better results compared to using only the Head modality [AUs] or only the Body modality [OpenPose] alone. This further indicates that voice should be given more significant consideration for the engagement estimation task.

From another perspective, considering the modality complementary property in multimodal learning, some modalities might exhibit suboptimal performance when predicting independently but could contribute valuable supplementary information when combined with other modalities. Even though the audio modality may appear to contribute most significantly, combining video and Kinect modalities could potentially provide crucial contextual information.

5. CONCLUSIONS

This study explored the intricacies of estimating human engagement using artificial mediators in diverse contexts. The proposed *TCA-Net* effectively integrates multiple modalities like audio, video and Kinect data through concatenated-attention fusion modules, aiming to capture and merge the nuanced information spread across these channels. The overall *TCA-Net* model presents outstanding performance on the multimodal engagement estimation task. The proposed network achieves 0.75 CCC on the NoXi dataset, outperforming the linear fusion baseline and other participants in the 2023 MultiMediate challenge using the same dataset. This work also investigates the modality significance in the engagement estimation task, which presents that audio plays a critical role in predicting engagement scores. Such an engagement estimation model can be integrated into predicting other aspects of human communication, such as empathy, emotion and social dynamics. Furthermore, our proposed multimodal *TCA-Net* can be extended to other tasks, such as visual question answering, video content understanding, and cross-modal retrieval.

6. REFERENCES

[1] Carl Vondrick and Deva Ramanan, "Video annotation and tracking with active learning," *Advances in Neural Information Processing Systems*, vol. 24, 2011.

[2] Panagiotis Tzirakis, George Trigeorgis, Mihalis A Nicolaou, Björn W Schuller, and Stefanos 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.

[3] Hae Won Park, Ishaan Grover, Samuel Spaulding, Louis Gomez, and Cynthia Breazeal, "A model-free affective reinforcement learning approach to personalization of an autonomous social robot companion for early literacy education," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, vol. 33, pp. 687–694.

[4] Stefan Glasauer, Markus Huber, Patrizia Basili, Alois Knoll, and Thomas Brandt, "Interacting in time and space: Investigating human-human and human-robot joint action," in *19th international symposium in robot and human interactive communication*. IEEE, 2010, pp. 252–257.

[5] Charles Rich, Brett Ponsler, Aaron Holroyd, and Candace L Sidner, "Recognizing engagement in human-robot interaction," in *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2010, pp. 375–382.

[6] Jyotirmay Sanghvi, Ginevra Castellano, Iolanda Leite, André Pereira, Peter W McOwan, and Ana Paiva, "Automatic analysis of affective postures and body motion to detect engagement with a game companion," in *Proceedings of the 6th international conference on Human-robot interaction*, 2011, pp. 305–312.

[7] Athanasios Voulodimos, Nikolaos Doulamis, Anastasios Doulamis, Eftychios Protopapadakis, et al., "Deep learning for computer vision: A brief review," *Computational intelligence and neuroscience*, vol. 2018, 2018.

[8] Md Rakibul Hasan, Md Zakir Hossain, Tom Gedeon, and Shafin Rahman, "LLM-GEm: Large language model-guided prediction of people's empathy levels towards newspaper article," in *Findings of the Association for Computational Linguistics: EACL 2024*, Yvette Graham and Matthew Purver, Eds. 2024, pp. 2215–2231, Association for Computational Linguistics.

[9] Md Rakibul Hasan, Md Zakir Hossain, Tom Gedeon, Susannah Soon, and Shafin Rahman, "Curtin OCAI at WASSA 2023 empathy, emotion and personality shared task: Demographic-aware prediction using multiple transformers," in *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*. 2023, pp. 536–541, Association for Computational Linguistics.

[10] Hae Won Park, John Busche, Bjorn Schuller, Cynthia Breazeal, Rosalind W Picard, et al., "Personalized estimation of engagement from videos using active learning with deep reinforcement learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.

[11] Ömer Sümer, Patricia Goldberg, Sidney D'Mello, Peter Gerjets, Ulrich Trautwein, and Enkelejda Kasneci, "Multimodal engagement analysis from facial videos in the classroom," *IEEE Transactions on Affective Computing*, 2021.

[12] Nese Alyuz, Eda Okur, Utku Genc, Sinem Aslan, Cagri Tanriover, and Asli Arslan Esme, "An unobtrusive and multimodal approach for behavioral engagement detection of students," in *Proceedings of the 1st ACM SIGCHI International Workshop on Multimodal Interaction for Education*, 2017, pp. 26–32.

[13] Yuanchao Li, Tianyu Zhao, and Xun Shen, "Attention-based multimodal fusion for estimating human emotion in real-world hri," in *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, 2020, pp. 340–342.

[14] Ahmed A Abdelrahman, Dominykas Strazdas, Aly Khalifa, Jan Hintz, Thorsten Hempel, and Ayoub Al-Hamadi, "Multimodal engagement prediction in multi-person human–robot interaction," *IEEE Access*, vol. 10, pp. 61980–61991, 2022.

[15] Angelo Cafaro, Johannes Wagner, Tobias Baur, Soumia Dermouche, Mercedes Torres Torres, Catherine Pelachaud, Elisabeth André, and Michel Valstar, "The NoXi database: Multimodal recordings of mediated novice-expert interactions," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. 2017, p. 350–359, Association for Computing Machinery.

[16] Roman Bednarik, Shahram Eivazi, and Michal Hradis, "Gaze and conversational engagement in multiparty video conversation: an annotation scheme and classification of high and low levels of engagement," in *Proceedings of the 4th workshop on eye gaze in intelligent human machine interaction*, 2012, pp. 1–6.

[17] Sandeep Mandia, Kuldeep Singh, and Rajendra Mitharwal, "Vision transformer for automatic student engagement estimation," in *2022 IEEE 5th International Conference on Image Processing Applications and Systems (IPAS)*. IEEE, 2022, pp. 1–6.

[18] Vu Ngoc Tu, Van Thong Huynh, Hyung-Jeong Yang, Soo-Hyung Kim, Shah Nawaz, Karthik Nandakumar, and M. Zaigham Zaheer, "Dctm: Dilated convolutional transformer model for multimodal engagement estimation in conversation," in *Proceedings of the 31st ACM International Conference on Multimedia*, New York, NY, USA, 2023, pp. 9521–9525, Association for Computing Machinery.

[19] Chunxi Yang, Kangzhong Wang, Peter Q. Chen, MK Michael Cheung, Youqian Zhang, Eugene Yujun Fu, and Grace Ngai, "Multimediate 2023: Engagement level detection using audio and video features," in *Proceedings of the 31st ACM International Conference on Multimedia*, New York, NY, USA, 2023, p. 9601–9605, Association for Computing Machinery.

[20] Jun Yu, Keda Lu, Mohan Jing, Ziqi Liang, Bingyuan Zhang, Jianqing Sun, and Jiaen Liang, "Sliding window seq2seq modeling for engagement estimation," New York, NY, USA, 2023, MM '23, p. 9496–9500, Association for Computing Machinery.

[21] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach, "Multimodal compact bilinear pooling for visual question answering and visual grounding," *arXiv preprint arXiv:1606.01847*, 2016.

[22] Philipp Müller, Michal Balazia, Tobias Baur, Michael Dietz, Alexander Heimerl, Dominik Schiller, Mohammed Guermal, Dominike Thomas, François Brémond, Jan Alexandersson, Elisabeth André, and Andreas Bulling, "Multimediate '23: Engagement estimation and bodily behaviour recognition in social interactions," in *Proceedings of the 31st ACM International Conference on Multimedia*. 2023, p. 9640–9645, Association for Computing Machinery.