EasyChair Preprint

№ 14673

# Routing Mixture-of-Experts Based on Ensembled Spatiotemporal Representations for Engagement Estimation in Online Courses

Hongqiang Shen, Jun Wang, Juncheng Li and Jun Shi



September 3, 2024

# Routing Mixture-of-Experts based on Ensembled spatiotemporal Representations for Engagement Estimation in Online Courses

Hongqiang Shen [1], Jun Wang [1], Juncheng Li [1], and Jun Shi [1]

[1] School of Communication and Information Engineering, Shanghai University, Shanghai 200444, China, wangjun_shu@shu.edu.cn

**Abstract.** Automatic estimation of students' engagement provides real-time feedback to the teachers in online courses. Although some deep learning methods have shown success in engagement estimation, most of them are developed based on convolutional neural networks (CNNs), which fail to capture long-range spatial and temporal dependencies in video data. Even when both temporal and spatial representations are extracted, they are not fully utilized, decreasing the accuracy of engagement estimation. To address these issues, we propose a novel Mixture-of-Experts (MoE) method that effectively ensembles spatial and temporal representations. Specifically, we introduce a Routing Mixture-of-Experts (RMoE) method designed to capture comprehensive and discriminative spatiotemporal representations. The method uses a routing mechanism to dynamically select the most relevant experts for a given input, ensuring accurately capture both spatial and temporal representations. We evaluated the effectiveness of our model using the Dataset for Affective States in E-Environments (DAiSEE). Experimental results show that our model significantly outperforms several state-of-the-art methods, highlighting its potential to improve the accuracy of student engagement estimation in online learning environments.

**Keywords:** Online Courses, Engagement Estimation, Ensemble Learning, Spatiotemporal Representations

## 1 Introduction

With the rapid expansion of education and electronic learning (e-learning) [1], maintaining student engagement in online courses has become a great challenge for educators [2]. Previous studies have indicated that many students could not immerge themselves in online courses [3]. Consequently, automatic engagement estimation has become essential, as students' engagement levels enables educators to enhance learning efficiency [4]. Engagement is defined as the state of being either immersed or not immersed in a task [5]. Based on the learner's interest and attentiveness[6], engagement can be classified into four levels: very low, low, high, and very high [7].

Recently, engagement estimation has become a research hotspot across various fields [5]. Its goal is to directly monitor students and maintain a high level of

engagement in online courses. Multiple modalities have been utilized for engagement estimation including images[8, 9], videos [10-12], audio [13], and Electrocardiogram (ECG) [14]. Due to its ubiquitous, cost-effective, and non-intrusive nature [4], computer vision (CV) has shown significant potential in engagement estimation [15]. CV methods can be divided into spatial representation-based and spatiotemporal representation-based methods. The former focuses on individual images, as spatial information has proven reliable for predicting levels of engagement. For example, Gupta et al. [8] employed Inception-v3 to predict various affective states based on single frames in videos, while Batra et al. [9] demonstrated that ResNet-18 outperformed DenseNet-121 and MobileNet-v1 in single-frame engagement estimation.

The limitation of the above methods is that they are developed based on static images or isolated frames. However, engagement focuses on spatiotemporal affective states and varies over time, and it cannot be completely described by static images or isolated frames. Some researcher [16] argue that assessing students' engagement requires evaluating their status at finer time intervals, taking into account that the temporal correlations between frames. Therefore, it is more appropriate to incorporate temporal information for more accurate engagement estimation. Representative network backbones in this category includes Convolutional 3D (C3D) [11], Long-term Recurrent Convolutional Networks (LRCN) [17] and Inflated 3D (I3D) [10]. For instance, Geng et al. [11] utilized a C3D classifier to classify engagement levels; Zhang et al. [10] introduced a modified Inflated 3D (I3D) model to estimate engagement levels; Abedi et al. [12] proposed a hybrid network ResTCN that combined Residual Network (ResNet) and Temporal Convolutional Network (TCN). However, these methods are developed based on convolutional neural networks (CNNs) that primarily rely on convolutional filters to extract local representation from videos, limiting their ability to capture long-range spatial and temporal dependencies. In addition, these methods predominantly focus on temporal information without effectively integrating both spatial and temporal data for engagement estimation.

To this end, we propose a Routing Mixture-of-Experts (RMoE) method based on adaptive mixtures of local experts [18] that aims to effectively learn discriminative and ensembled spatiotemporal representations for engagement estimation. The main contributions of this work are as follows:

1) We propose a RMoE that encodes spatial and temporal representations. It not only incorporates temporal information based on spatial information, but also effectively integrates both spatial and temporal information for more accurate engagement estimation.

2) We introduce Transformers as spatial and temporal encoders to handle the long-range spatial and temporal dependencies in videos.

The remainder of the paper is organized as follows: Section 2 proposes RMoE, highlighting the advantages and innovations of our method; Section 3 presents the experimental results, which includes information on the dataset, comparison experiments, and ablation studies. Section 4 draws the conclusion that provides a summary of the key insights from our research.

## 2 Method

In this section, we propose a RMoE method that learns ensembled spatiotemporal representations through spatial and temporal encoders. This scheme differs from previous CNNs-based methods. In the following sections, we first illustrate the process of frame uniformity and then introduce our network.

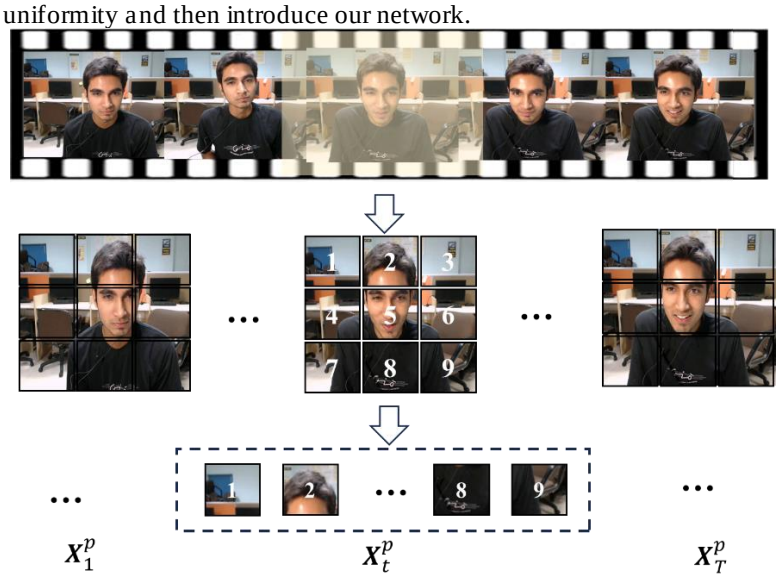


**Fig. 1.** Data preprocessing pipeline. *T* frames are uniformly sampled from the video clip, with each 2D frame independently embedded.

### 2.1 Data preprocessing

The video clips used in this study are denoted as $\boldsymbol{X} \in \mathbb{R}^{T\times H\times W\times C}$, where $T$ is the number of the video frames, $H \times W$ is the resolution of a single frame, and $C$ is the number of channels. Each video frame at time $t$ is cropped into a sequence of $P \times P$ patches, i.e. $\boldsymbol{X}_t^p = \left[\boldsymbol{x}_t^{p\,(1)}; \ldots \boldsymbol{x}_{t,}^{p(N)}\right] \in \mathbb{R}^{N\times(P^2C)}$, for $t$ = 1,2,……$T$, where $N = \frac{HW}{P^2}$ is the number of patches. Fig.1. shows the data preprocessing pipeline.

### 2.2 RMoE framework

In this section, we propose a RMoE framework that includes four components: the token embedding module, the expert networks, the Routing network, and the tower network. Fig.2 shows the architecture of the proposed Routing Mixture-of-Expert.

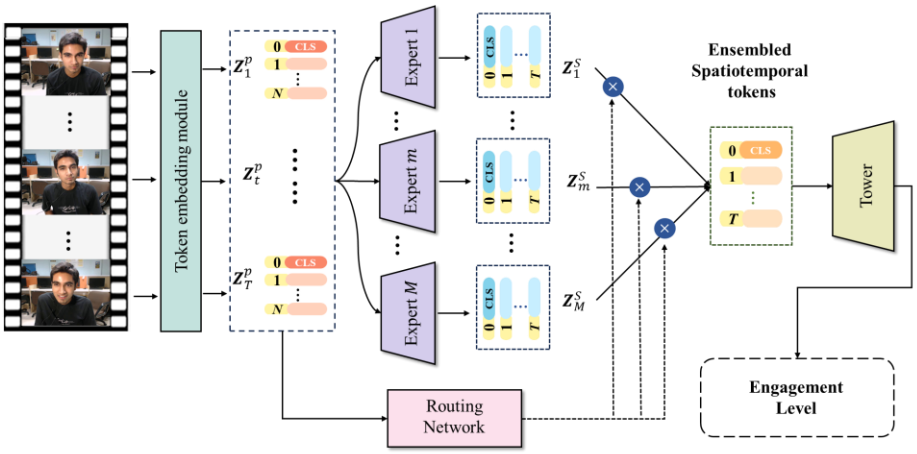


Fig. 2. Architecture of Routing Mixture-of-Experts. The model includes token embedding module, expert networks, the routing network and the tower network.

Given an input video, the token embedding module transforms the input data into token embeddings. The expert networks then extract the spatial representations from the token embeddings and the routing network selects the output of a subset of experts for subsequent fusion of spatiotemporal representations. Finally, the tower network extracts spatiotemporal representations from a sequence of spatial representations across all time points and predicts the results using an MLP head.

**Token embedding module.**

The token embedding module is used to transform the input patches into token embeddings by a trainable linear transformation. Given a sequence of patches representing the corresponding flattened patches of the video frame $\boldsymbol{X}_t^{\square}$, these patches are projected into a $D$-dimensional representation space using a trainable linear transformation $\boldsymbol{E}^S \in \mathbb{R}^{(P^2C)\times D}$. This transformation ensures that the expert networks can process the patch embeddings with a consistent size across all layers.

For the video frame at time point $t$, the sequence of the patch embeddings is denoted as follows:

$$\boldsymbol{Z}_t^p = \left[\boldsymbol{z}_{\boldsymbol{CLS}}; \boldsymbol{x}_t^{p\,(1)}\boldsymbol{E}^S; \cdots; \boldsymbol{x}_t^{p(N)}\boldsymbol{E}^S\right] + \boldsymbol{P}_t^S \tag{1}$$

where $\boldsymbol{z}_{\boldsymbol{CLS}} \in \mathbb{R}^D$ is a learnable position embedding to preserve the spatial information of each video frame in the video, and $\boldsymbol{P}_t^S \in \mathbb{R}^{(N+1)\times D}$ is a learnable spatial positional embedding added to the token embeddings to maintain spatial information.

**Expert networks.**

The expert network is designed to learn the spatial representations from the token embeddings generated by the token embedding module. The RMoE framework

includes $M$ expert networks, each including $T$ spatial Transformer encoders. Each encoder is responsible for generating a spatial representation for an individual video frame. The architecture features alternating layers of multi-headed self-attention (MSA) and feed-forward network (FFN) blocks. Layer Normalization (LN) [19] is implemented to accelerate convergence of the training process, and the residual connections are employed to enhance the information flow for improved performance [20]. The FFN consists of two layers, both using Gaussian Error Linear Unit (GELU) as the activation function. Fig.3 shows the architecture of an expert network.

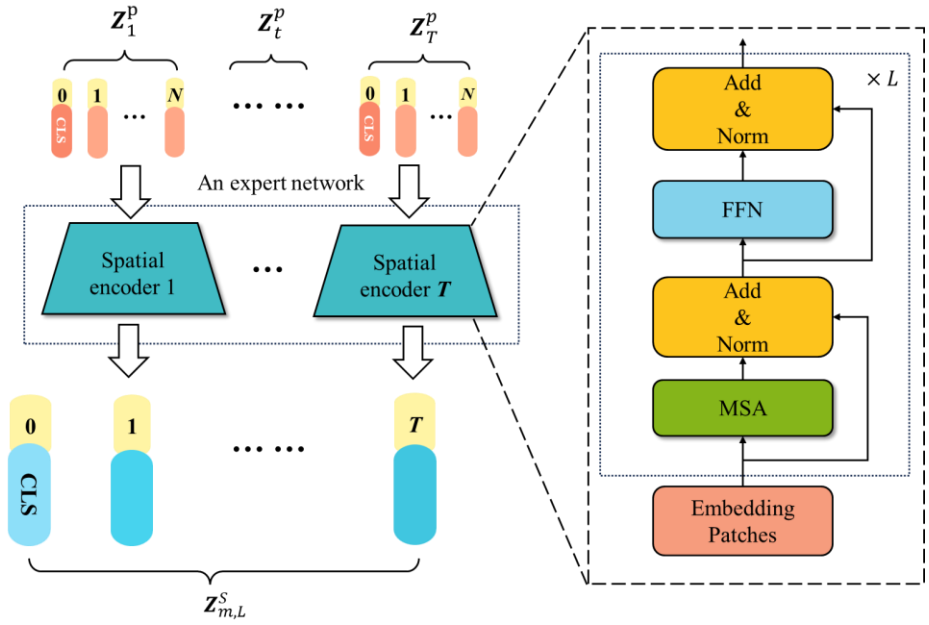


**Fig. 3.** An illustration that an expert network learns the spatial representations.

Consider $\boldsymbol{Z}_t^p$ as a sequence of patch embeddings input into the $t$-th spatial encoder within an expert network, producing a series of spatial embedding $\boldsymbol{Z}_m^S$ for $m = 1, \cdots, M$. Specifically, each spatial encoder in the expert network captures the interaction between the token embeddings extracted from video frame at the same time point, thereby generating $T$ frame-level representations $\boldsymbol{Z}_{m,t}^S$ for $t = 1, \cdots, T$. To simplify the notation, we denote the output of each expert as $\boldsymbol{Z}_m^s \in \mathbb{R}^{T \times D}$, Given the outputs of $M$ experts, we concatenate them as Eq. (2):

$$\boldsymbol{Z}^S = [\boldsymbol{Z}_1^s, ; \boldsymbol{Z}_2^s ; \ldots ; \boldsymbol{Z}_M^s] \tag{2}$$

**Routing network.**

The Routing network "selects" several experts for subsequent fusion of spatiotemporal representations. This allows the network to flexibly obtain the most effective representations from the spatial encoders. Specifically, a probability distribution over the experts is generated by the routing network based on token embeddings $\boldsymbol{Z}_t^p$ generated

from the input patches $\boldsymbol{X}_t^p$ by the token embedding module. The final output is a weighted combination of the outputs of all experts:

$$f(\{\boldsymbol{Z}_t^p\}) = \sum_{m=1}^{M} g(\{\boldsymbol{Z}_t^p\})_m f_m(\{\boldsymbol{Z}_t^p\}) \tag{3}$$

where $R(\cdot)$ represents the routing network, $R(\{\boldsymbol{Z}_t^p\})_m$ denotes the $m$-th logit of the $g(\cdot)$ output, $f_m(\cdot)$ is the $m$-th expert network, and $M$ is the number of experts.

In this work, we formulate the routing network as a linear transformation of the input through the *Softmax* function:

$$R(\boldsymbol{X}) = softmax(\boldsymbol{W}_r \boldsymbol{X}) \tag{4}$$

where $\boldsymbol{W}_r \in \mathbb{R}^{M \times D}$ is a trainable matrix, $\boldsymbol{X}$ is the input sequence of patches; $M$ and $D$ are the expert number and representation dimension, respectively.

**Tower network.**

The tower network models the interactions base on $T$-time-step spatial representations and outputs the final classification results, i.e.

$$\hat{\boldsymbol{y}} = h\left(f(\{\boldsymbol{Z}_t^p\})\right) \tag{5}$$

where $\boldsymbol{Z}_t^p$ are token embeddings, $h(\cdot)$ represents the tower network, and $f(\cdot)$ is the routing network computed using Eq. (4).

The tower network includes a temporal Transformer encoder and a classification head with $K$ output nodes. The temporal encoder is designed to capture temporal interactions within the token embeddings of length $T$. The output of the $m$-th expert $\boldsymbol{Z}_m^s$ is expressed as:

$$\boldsymbol{Z}_m^s = [\boldsymbol{z}_{\boldsymbol{CLS}}; \boldsymbol{z}_{m,1}^s; \boldsymbol{z}_{m,2}^s; \cdots; \boldsymbol{z}_{m,T}^s] + \boldsymbol{P}^{\boldsymbol{T}} \tag{6}$$

where $m = 1, \cdots, M$, and $\boldsymbol{P}^{\boldsymbol{T}} \in \mathbb{R}^{(T+1) \times D}$ is a learnable spatiotemporal embedding to preserve spatiotemporal information. The spatial tokens output by the expert networks are ensembled by the routing network and subsequently processed through a temporal Transformer encoder. After the spatiotemporal representations are ensembled, a classification head is employed to perform $K$-class classification. Each output node of the classification head corresponds to a specific engagement level.

### 2.3 Optimization

The overall architecture is optimized using the focal loss [21], which is defined as follows:

$$\mathcal{L} = -\sum_{i=1}^{N}\sum_{k=1}^{K} y_{ik}\,\alpha_{ik}(1-\hat{y}_{ik})^{\gamma}\log\hat{y}_{ik} \tag{7}$$

where $y_{ik} = \{0,1\}$ is the true label indicating whether sample $i$ belongs to the $k$-th class, and it takes 0 when sample $i$ does not belong to the $k$-th class, and $\hat{y}_{ik} \in (0,1)$ is the predicted probability for the $k$-th class of sample $i$. To address the class imbalance issue, $\alpha_{ik}$ is included to denote the proportion of $k$-th class in the loss function. Specifically, $\alpha_{ik}$ is set to the inverse of the number of classes, ensuring that each class contributes equally to the loss function, regardless of its representation in the dataset. However, this alone does not differentiate between easy and hard samples. Therefore, a modulation factor $(1 - \hat{y}_{ik})^{\gamma}$ is added, where $\gamma \geq 0$ is a tunable parameter that reduces the influence of easy samples, while enhancing the contribution of hard samples, thereby focusing the training on difficult instances.

# 3 Experiments

## 3.1 Details of experiments

**Dataset.**

The DAiSEE dataset [8] is used in this study. It includes 9,068 videos taken from 112 students in online courses. Four types of emotional states such as boredom, confusion, engagement, and frustration are used to describe the students. Since this study focuses on student engagement, we use the engagement as the labels (0: very low, 1: low, 2: high, 3: very high). Each video lasts a few seconds, with a frame rate of 30 fps and a resolution of 640×480 pixels. Table 1 provides the details of the training, validation, and test sets used in our study. Fig.4 shows 16 video frames of four students participating in online courses, with their engagement levels ranging from 0 to 3, respectively, from top to bottom.

**Table 1.** The details of the training, validation, and test sets in the DAiSEE dataset.

| Levels | 0 (very low) | 1 (low) | 2 (high) | 3 (very high) |
|---|---|---|---|---|
| Training | 34 | 213 | 2617 | 2494 |
| Validation | 23 | 143 | 813 | 450 |
| Testing | 4 | 84 | 882 | 814 |
| Total | 61 | 440 | 4312 | 3758 |

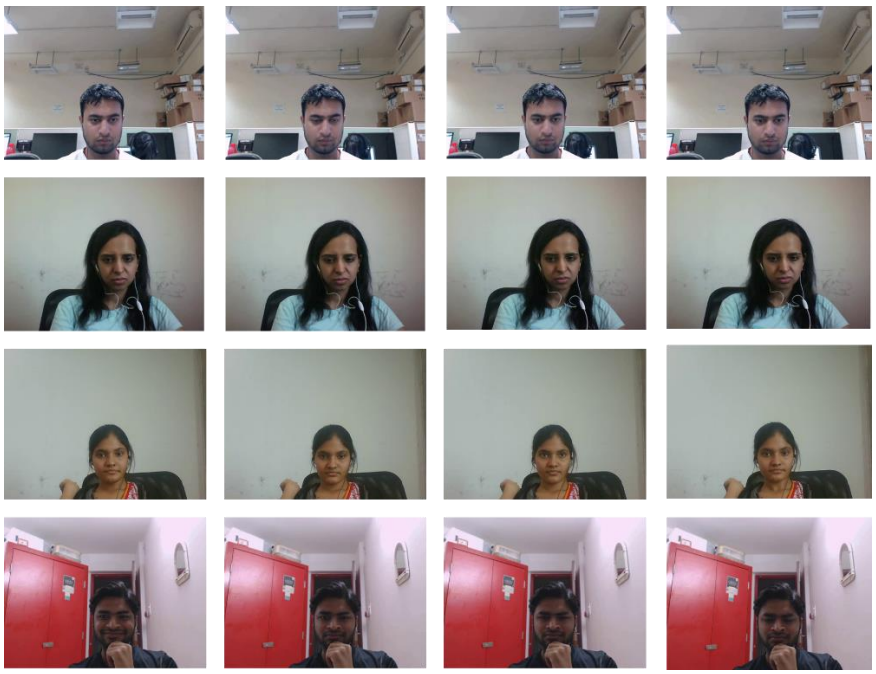

**Fig. 4.** Four students participating in online courses, with their engagement levels ranging from 0 to 3, respectively, from top to bottom.

**Experiment settings.**

In alignment with previous research methodologies, the training and validation set videos are used to develop and fine-tune the architecture, while a separate set of 1,784 test videos is utilized for the final performance evaluation. This approach ensures that the model's performance is assessed on unseen data, providing a robust measure of its generalization capability.

One of the challenges encountered in this study is the significant class imbalance present in the dataset. In scenarios where certain classes are heavily underrepresented, traditional evaluation metrics like accuracy can be misleading, as they tend to favor the majority class. For instance, a model could achieve high accuracy by simply predicting the majority class in most cases, despite performing poorly on the minority class. To address this issue, we employ evaluation metrics specifically designed to handle imbalanced classification tasks, which offer a more accurate reflection of the model's performance across all classes.

The key metrics used in this evaluation are Precision, Recall, the macro F1 Score (F1-macro), and the weighted F1 Score (F1-weighted). These metrics provide a comprehensive understanding of the model's ability to correctly identify each class. Let *TP*, *TN*, *FP*, and *FN* denote the number of true positives, true negatives, false positives, and false negatives, respectively. Precision, which measures the proportion of correctly identified positive instances out of all instances predicted as positive, defined as

TP/(FP+TP). Recall, which measures the proportion of actual positive instances that were correctly identified by the model, is defined as TP/(TP+FN). The macro F1 Score (F1-macro) and weighted F1 Score (F1-weighted) [22] are computed accordingly:

$$\begin{cases} F1-macro = \dfrac{\sum_{i=1}^{N} F1_i}{N} \\ F1-weighted = \dfrac{\sum_{i=1}^{N} F1_i \times w_i}{\sum_{i=1}^{N} w_i} \\ F1 = \dfrac{2 \times Precision \times Recall}{Precision + Recall} \end{cases} \tag{8}$$

In addition to the F1 scores, the Area Under the Curve (AUC) was also employed as a key evaluation metric [23, 24]. The model was trained on the training and validation sets for 100 epochs, and the weights yielding the best AUC were used for testing on the test set.

### 3.2 Comparison with State-of-the-Art methods

In this subsection, we compare the classification performance of RMoE with classical deep learning methods, specifically CNN-based and transformer-based methods. CNN-based methods include C3D [25], I3D [26], and ResTCN [27], where raw frames of videos are used for engagement estimation. Transformer-based methods include TimeSformer [27], ViViT [28] and video swin transformer [29], involve cropping video into patches before processing. Although these methods are not specifically designed for engagement estimation, they are popular computer vision models applicable to this task.

**Table 2.** Results (%) of the proposed and other classical methods.

| Methods | F1-macro | F1-weighted | AUC |
|---|---|---|---|
| ResTCN [12] | 23.05 | 46.01 | 57.17 |
| C3D [25] | 24.16 | 46.45 | 57.34 |
| I3D [26] | 22.95 | 41.08 | 55.75 |
| TimeSformer [27] | 19.91 | 38.87 | 54.13 |
| ViViT [28] | 16.50 | 32.54 | 54.63 |
| Video Swin Transformer[29] | 21.46 | 42.06 | 55.14 |
| RMoE (Ours) | 26.4 | 49.2 | 60.44 |

As shown in Table 2, our method consistently demonstrates superior performance, surpassing all competing models in F1-macro, F1-weighted, and AUC metrics. Compared to traditional convolutional methods such as ResTCN, C3D, and I3D, our method shows significant average improvements. It also outperforms transformer-based

methods like TimeSformer, ViViT, and Video Swin Transformer. Overall, the proposed RMoE excels over both convolutional and transformer-based methods across all evaluated metrics.

### 3.3 Ablation Study

To evalute the contribution of each element within the RMoE framework for engagement estimation, we developed the following RMoE variants for comparative analysis:

**RMoE w/ ResNet***:* In this variant, the ResNet [30] is used as the spatial encoder, is which is designed to capture intricate spatial representations through deep residual learning.

**RMoE w/ TCN**: TCN is effective in handling sequential data and modeling long-range dependencies This variant leverage TCN [31] as the temporal encoder.

**RMoE w/ LSTM**: This variant incorporates LSTM [32] as the temporal encoder, which is adept at processing sequences by retaining long-term dependencies

**RMoE w/o transformer encoder:** This variant excludes transformer encoder, instead utilizing ResNet as spatial encoder and TCN as temporal encoder. As a result, the variant is equal to ResTCN.

**RMoE w/o MoE framework:** This variant does not utilize our proposed the RMoE framework, and just use a spatial encoder and a temporal encoder, both of which are transformer encoder. The variant is equal to ViViT.

**Table 3.** Results (%) of the proposed method and ablation experiments

| Variants | F1-macro | F1-weighted | AUC |
|---|---|---|---|
| RMoE w/ ResNet | 23.98 | 41.35 | 54.45 |
| RMoE w/ TCN | 24.26 | 45.68 | 55.4 |
| RMoE w/ LSTM | 24.9 | 45.68 | 56.59 |
| RMoE w/o transformer encoder | 23.05 | 46.01 | 57.17 |
| RMoE w/o MoE framework | 16.50 | 32.54 | 54.63 |
| RMoE | 26.4 | 49.2 | 60.44 |

As shown in Table 3, the proposed RMoE framework achieves the highest performance across all evaluation metrics, outperforming all other variants. The results indicate that the variants employing ResNet, TCN, and LSTM as encoder perform poorly across all metrics due to their limitations in capturing long-term dependencies. Therefore, RMoE w/ ResNet, RMoE w/ TCN, and RMoE w/ LSTM exhibit poor performance. These observations motivate us to integrate the transformer encoder in our method, as its effectively address these limitations by capturing both spatial and temporal dependencies with greater accuracy. Consequently, the integration of the Transformer within the RMoE framework leads to superior engagement estimation, as evidenced by the significant improvements in F1-macro, F1-weighted, and AUC scores.

## 4 Conclusion

Most deep models for engagement estimation are developed based on CNNs, which fail to provide long-range spatial and temporal dependencies. To this end, we propose a novel ensemble learning method RMoE which ensembles multiple spatial encoders and one temporal encoder to generate more discriminative spatiotemporal representations, ultimately aiming at more accurate engagement estimation. In addition, Transformer is introduced as spatial and temporal encoders to capture more discriminative representations for engagement estimation. We evaluate the RMoE method on the DAiSEE dataset and compare its performance with several popular methods. The experimental results indicate that our method outperform these methods.

This study provides a promising solution to the challenges of maintaining and assessing student engagement on digital education platforms. Future work will explore further refinements of the model and its application to other domains where engagement estimation is critical.

**Acknowledgments.** This work was supported in part by the National Natural Science Foundation of China under Grant 62272289, and in part by 111 Project under Grant D20031. (Corresponding authors: Jun Wang).

## References

1. R. Das and S. Dev, "Enhancing frame-level student engagement classification through knowledge transfer techniques," *Applied Intelligence,* pp. 1-16, 2024.
2. H. Sharif-Nia *et al.*, "University Student Engagement Inventory: Validation in the Indian Online Learning Context," *Measurement and Evaluation in Counseling and Development,* pp. 1-13, 2024.
3. M. E. Nicholls, K. M. Loveless, N. A. Thomas, T. Loetscher, and O. Churches, "Some participants may be better than others: Sustained attention and motivation are higher early in semester," *Quarterly journal of experimental psychology,* vol. 68, no. 1, pp. 10-18, 2015.
4. A. Nguyen, M. Kremantzis, A. Essien, I. Petrounias, and S. Hosseini, "Enhancing Student Engagement Through Artificial Intelligence (AI): Understanding the Basics, Opportunities, and Challenges," *Journal of University Teaching and Learning Practice,* vol. 21, no. 06, 2024.
5. C. Oertel *et al.*, "Engagement in human-agent interaction: An overview," *Frontiers in Robotics and AI,* vol. 7, p. 92, 2020.
6. D. Wilson and L. Summers, "The importance of teaching assistant support and interactions in student engagement," *Authorea Preprints,* 2024.
7. A. W. Fazil, M. Hakimi, A. K. Shahidzay, and A. Hasas, "Exploring the Broad Impact of AI Technologies on Student Engagement and Academic Performance in University Settings in Afghanistan," *RIGGS: Journal of Artificial Intelligence and Digital Business,* vol. 2, no. 2, pp. 56-63, 2024.
8. A. Gupta, A. D'Cunha, K. Awasthi, and V. Balasubramanian, "Daisee: Towards user engagement recognition in the wild," *arXiv preprint arXiv:1609.01885,* 2016.
9. S. Batra *et al.*, "DMCNet: Diversified model combination network for understanding engagement from video screengrabs," *Systems and Soft Computing,* vol. 4, p. 200039, 2022.

10. H. Zhang, X. Xiao, T. Huang, S. Liu, Y. Xia, and J. Li, "An novel end-to-end network for automatic student engagement recognition," in *2019 IEEE 9th International Conference on Electronics Information and Emergency Communication (ICEIEC)*, 2019: IEEE, pp. 342-345.
11. L. Geng, M. Xu, Z. Wei, and X. Zhou, "Learning deep spatiotemporal feature for engagement recognition of online courses," in *2019 IEEE Symposium Series on Computational Intelligence (SSCI)*, 2019: IEEE, pp. 442-447.
12. A. Abedi and S. S. Khan, "Improving state-of-the-art in detecting student engagement with resnet and tcn hybrid network," in *2021 18th Conference on Robots and Vision (CRV)*, 2021: IEEE, pp. 151-157.
13. A. Dhall, G. Sharma, R. Goecke, and T. Gedeon, "Emotiw 2020: Driver gaze, group emotion, student engagement and physiological signal based challenges," in *Proceedings of the 2020 International Conference on Multimodal Interaction*, 2020, pp. 784-789.
14. K. Doherty and G. Doherty, "Engagement in HCI: conception, theory and measurement," *ACM Computing Surveys (CSUR)*, vol. 51, no. 5, pp. 1-39, 2018.
15. S. S. Panda and B. Malviya, "Investigating factors affecting studengt engagement among studentspursuing management programme," *International Journal of Innovation Studies*, vol. 8, no. 1, pp. 201-220, 2024.
16. A. Kaur, A. Mustafa, L. Mehta, and A. Dhall, "Prediction and localization of student engagement in the wild," in *2018 Digital Image Computing: Techniques and Applications (DICTA)*, 2018: IEEE, pp. 1-8.
17. J. Donahue *et al.*, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2625-2634.
18. R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton, "Adaptive mixtures of local experts," *Neural computation*, vol. 3, no. 1, pp. 79-87, 1991.
19. J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.
20. Q. Wang *et al.*, "Learning deep transformer models for machine translation," *arXiv preprint arXiv:1906.01787*, 2019.
21. 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.
22. N. Q. K. Le and B. P. Nguyen, "Prediction of FMN binding sites in electron transport chains based on 2-D CNN and PSSM profiles," *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 18, no. 6, pp. 2189-2197, 2019.
23. J. Huang and C. X. Ling, "Using AUC and accuracy in evaluating learning algorithms," *IEEE Transactions on knowledge and Data Engineering*, vol. 17, no. 3, pp. 299-310, 2005.
24. N. Q. K. Le, Q.-T. Ho, V.-N. Nguyen, and J.-S. Chang, "BERT-Promoter: An improved sequence-based predictor of DNA promoter using BERT pre-trained model and SHAP feature selection," *Computational Biology and Chemistry*, vol. 99, p. 107732, 2022.
25. K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6546-6555.
26. J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299-6308.
27. G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?," in *ICML*, 2021, vol. 2, no. 3, p. 4.

28. A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "Vivit: A video vision transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 6836-6846.
29. Z. Liu *et al.*, "Video swin transformer," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 3202-3211.
30. C. Szegedy *et al.*, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1-9.
31. S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *arXiv preprint arXiv:1803.01271*, 2018.
32. Y. Yu, X. Si, C. Hu, and J. Zhang, "A review of recurrent neural networks: LSTM cells and network architectures," *Neural Computation*, vol. 31, no. 7, pp. 1235-1270, 2019.