

Random Patch Masking with ConvNeXt V2 for Robust Student Engagement Recognition in Online Learning

Chih-Hsien Hsia,^{1,2*} Ting-Hsuan (Megan) Chen,³ Hui-Zhen Lai,¹ and Liang-Ying Ke⁴

¹Department of Computer Science and Information Engineering, National Ilan University, Yilan 260007, Taiwan

²Office of Research and Industry-Academia Collaboration, Chaoyang University of Technology,
Taichung 413310, Taiwan

³Taipei American School, Taipei 11152, Taiwan

⁴Department of Engineering Science, National Cheng Kung University, Tainan 701401, Taiwan

(Received May 11, 2026; accepted August 19, 2026)

Keywords: deep learning, data augmentation, student engagement, patch random masking, online learning

The rapid proliferation of online learning platforms has made student engagement a crucial metric for evaluating learning effectiveness. However, existing computer-vision-based methods often inadequately address facial occlusion and pose variations caused by changes in camera angle and viewing distance, resulting in unstable feature representations and limiting recognition performance. Moreover, because partial information loss may occur during camera-based facial image acquisition, in this study, we propose a student engagement recognition framework that integrates conditional random masking (CRM) with ConvNeXt-V2 to mitigate the adverse effects of incomplete or disturbed visual sensing signals. The proposed CRM mechanism leverages facial heatmaps to identify key regions and incorporates a random masking strategy, thereby enhancing the model's ability to learn critical engagement features while improving robustness to occlusion and pose variations. Experimental results demonstrate that the proposed method achieves accuracies of 98.15 and 87.97% on the Student Engagement Dataset (SED) and Spontaneous Visual Database for Online Learning (SVD-OL) datasets, respectively. Compared with existing approaches, the proposed method improves overall recognition performance and demonstrates its effectiveness on datasets containing head-pose variations and partial facial occlusion.

1. Introduction

Student participation refers to the extent to which students engage in the learning process.⁽¹⁾ It serves as a critical indicator of learning progress and is closely associated with academic performance, learning motivation, and mental health, encompassing students' cognitive behaviors, emotional interactions, and motivational displays. Student participation is commonly conceptualized as comprising three dimensions: behavioral, emotional, and cognitive engagement.⁽²⁾ Academic emotions play a pivotal role in student participation. These emotions

*Corresponding author: e-mail: hsiach@niu.edu.tw
<https://doi.org/10.18494/SAM6428>

not only directly affect students' involvement in learning activities but also dynamically interact with academic performance.⁽³⁾ Such emotions including happiness, engagement, confusion, and frustration exert a significant effect on student participation through their positive or negative valence. For instance, positive academic emotions, such as the anticipation of success or enthusiasm for a subject, can stimulate students' intrinsic motivation and prompt them to adopt more effective learning strategies, thereby enhancing their focus and participation in the learning process. Moreover, through emotional contagion, positive emotions can shape the behaviors of both teachers and students on online learning platforms, further enriching the learning experience and engagement. In contrast, negative academic emotions, such as the fear of failure or boredom with a topic, may weaken learning motivation, reduce participation, and potentially lead to learning burnout or even dropout.

In online learning environments, student participation across the three aforementioned dimensions is highly susceptible to interference. The absence of real-time teacher supervision renders students who learn at home more prone to distraction, while feelings of isolation may further reduce their learning motivation, ultimately adversely affecting learning outcomes. Moreover, determining an appropriate video length is inherently challenging, as students' fluctuating levels of focus and distraction are unpredictable. This often leads to inconsistent video durations, which in turn compromises analytical accuracy. Additionally, students exhibit a diverse range of emotional expressions, and different academic emotions frequently co-occur within a single video clip.⁽⁴⁾ For instance, brief moments of distraction may arise even in an otherwise attentive video, making it difficult to determine whether such transient distractions meaningfully impact overall participation. Finally, the substantial volume of video data imposes increased demands on model training, processing time, and computational resources.⁽⁵⁾ Image feature extraction lies at the core of facial emotion recognition,⁽⁶⁾ as facial features serve as the primary medium through which humans express emotions and psychological states, thereby clearly reflecting students' academic emotions. According to Yu *et al.*,⁽⁷⁾ language and speech contribute 7 and 38% to emotional information, respectively, whereas facial expressions account for as much as 55%, underscoring their importance in emotion recognition. Furthermore, compared with body movement recognition, the face conveys richer information and can more accurately reflect variations in student participation.⁽⁸⁾ Despite its significance, challenges in student participation recognition persist, particularly in the presence of facial occlusion, head posture variation, image blur, and data imbalance, all of which affect the accuracy and stability of recognition models. Although numerous studies have proposed solutions to various challenges in emotion recognition, facial occlusion and head posture remain topics requiring further improvement. Variations in camera angle, viewing distance, head pose, and facial occlusion may cause partial information loss or interference in the acquired visual sensing data, thereby affecting the reliability of student engagement recognition.

Accordingly, in this study, we aim to enhance the accuracy of student participation recognition, especially in scenarios involving facial occlusion or changes in head posture. How to accurately identify changes in student participation under such conditions constitutes a critical challenge, as shown in Fig. 1. Facial occlusion refers to the partial obstruction of the face by one's own or external objects, resulting in the loss of feature information. Factors such as a

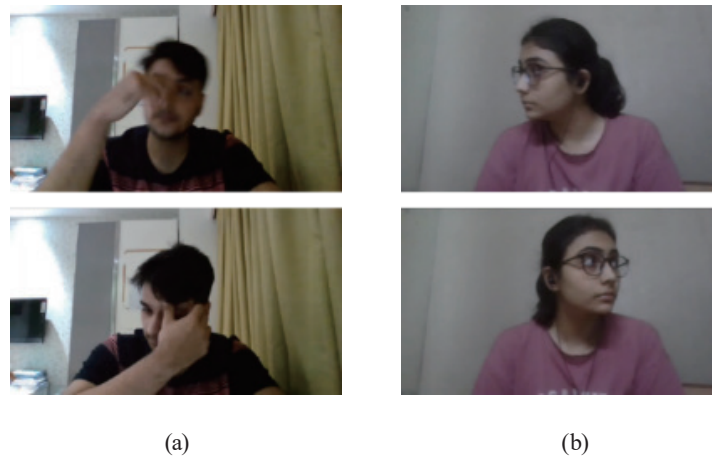


Fig. 1. (Color online) Challenges in student engagement identification: (a) facial occlusion and (b) changes in head posture.

student's hair, hand gestures, wearing glasses, or using a mobile phone can all impede complete facial feature extraction. Such occlusion compromises the model's ability to extract emotional features, rendering facial expression information incomplete or even misleading the model, thereby reducing emotion recognition accuracy. Head posture involves variations in the angle and position of the head relative to the camera, for example, frontal view, side view, head tilt, or inclined head. The frontal view provides complete facial features and is the most readily processed by the model, whereas side views or other non-frontal angles may obscure certain facial features, increasing the difficulty of feature extraction and recognition. Moreover, the asymmetry of features between the left and right sides of the face further complicates the model's handling of different postures. The accurate recognition of student participation is essential for understanding students' learning status and for providing teachers with timely, effective feedback to help them adjust their teaching strategies, thereby enhancing the system's practicality and overall teaching effectiveness.

2. Literature Review

Early studies relied on handcrafted features and traditional machine learning classifiers as basic components to recognize student engagement. These studies were primarily based on small datasets and encountered difficulties in accurately identifying student engagement when confronted with complex and noisy datasets.^(2,9) In contrast, convolutional neural networks (CNNs) exhibit the ability to learn salient features by progressively capturing the spatial structure and patterns of images through convolutional layers. Perumal *et al.*⁽¹⁰⁾ proposed a transfer learning approach utilizing CNNs to automatically detect and classify student engagement, achieving 93% accuracy on the Student Engagement Dataset (SED). However, this study was tested only on a specific dataset, and its generalization ability to other datasets remained poor. Anny *et al.*⁽¹¹⁾ developed a hybrid CNN-based end-to-end framework for face

recognition and engagement measurement in online learning platforms. The framework employs two specialized CNN models: CNN-I for recognizing learners' faces, facial expressions, eye gaze, and head movements, and CNN-II for detecting student engagement. It achieved 95.25% accuracy on the SED dataset. Nevertheless, the model was evaluated solely on a single dataset, and its generalization ability has yet to be verified.

According to previous research,⁽¹²⁾ the human face perception mechanism simultaneously extracts information from holistic and local facial features. By learning facial features from multiple angles, this mechanism achieves higher performance under occlusion and pose variations. Drawing on this insight, Li and Yao⁽¹³⁾ proposed a deep learning (DL)-based facial expression recognition technique specifically designed to detect student participation in physical classrooms. Their model enhances extraction capability through an improved hybrid attention mechanism, integrating a convolutional block attention module and a skip connection structure to optimize convergence speed and model stability. However, the introduction of this improved hybrid attention mechanism also increases the model's computational complexity, making real-time deployment in resource-constrained environments challenging. Xu *et al.*⁽¹⁴⁾ proposed an emotion recognition network based on class activation regularization for identifying student participation. This network generates class activation maps (CAMs) via a class activation dropout module, thereby obtaining class-specific attention and randomly discarding a portion of class-related features to further create class uncertainty regions. In the hard sample mining module, the network leverages model prediction confidence and an adaptive threshold within the class uncertainty regions to mine hard samples, applying a regularization loss to these samples to promote uncertainty learning in sentiment semantics. Nevertheless, the dual-branch structure of this approach imposes substantial memory resource demands, particularly during training, which may pose a challenge in resource-limited environments.

Sukumaran and Manoharan⁽¹⁵⁾ developed a MobileNetV2-based model that incorporates facial expressions, head posture, and gaze to assess student participation. They also designed a participation index algorithm and validated its effectiveness and accuracy using post-class test scores. However, a notable discrepancy was observed between the training and testing set accuracies in facial emotion recognition, suggesting that the model may suffer from overfitting, thereby resulting in poor generalization to new and unseen data. Sowjanya and Krithiga⁽¹⁶⁾ enhanced the local texture features of facial images through data augmentation and a unified local binary pattern, employing a CNN-based emotion detection model to focus on student participation and behavior detection. Although the model achieved satisfactory performance in student participation recognition tasks, its effectiveness and applicability remain to be further verified when dealing with more complex academic emotional expressions or more challenging datasets.

These reviews indicate that while significant progress has been made in DL methods that integrate various image features, numerous challenges persist, particularly in handling occlusion and head pose variations. Therefore, to further improve the stability and accuracy of emotion recognition systems, more diverse feature fusion strategies and deeper exploration are necessary.

3. Proposed Method

3.1 Model architecture

In this study, we propose employing the ConvNeXt-V2 architecture as the backbone network. A pretrained facial masked autoencoder (FMAE)⁽¹⁷⁾ model combined with gradient-weighted class activation mapping (Grad-CAM)⁽¹⁸⁾ is used to generate the corresponding heatmaps of input images, thereby identifying important attention regions and non-attention regions on the face. Subsequently, the proposed conditional random masking (CRM) module is applied separately to these regions to ensure that information from certain attention regions is preserved, thus avoiding the complete masking of any attention region. This design enhances the model's recognition capability by enabling it to focus more effectively on important regions when processing student participation recognition while preventing over-reliance on features from non-attention regions. Consequently, both the accuracy and robustness of recognition are improved.

First, the original image is fed into the CRM module, where a heatmap is generated using the FMAE model and Grad-CAM technology to identify important regions of interest (ROIs) on the face. Subsequently, a threshold is applied such that regions in the heatmap exceeding this threshold are designated as ROIs, thereby partitioning the heatmap into ROIs and non-ROIs. On the basis of this partition, random occlusion processing is performed separately on both ROIs and non-ROIs according to a predefined masking ratio. Next, the masks generated from the ROIs and non-ROIs are summed and then elementwise multiplied with the original image to produce the final masked face image. Finally, these masked face images are input into the ConvNeXt-V2 model for student engagement recognition, as shown in Fig. 2. The innovation of this method lies in its effective integration of dynamic occlusion processing applied separately to ROIs and non-ROIs, which preserves certain features of the ROIs. This design simultaneously improves student engagement recognition accuracy and enhances model interpretability.

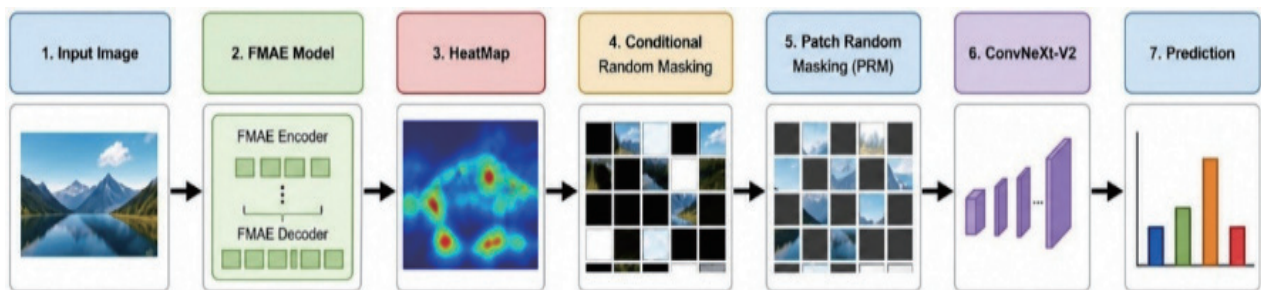


Fig. 2. (Color online) Block diagram of the proposed method.

3.2 Patch random masking

3.2.1 Facial mask autoencoder

In this study, we employed a pretrained FMAE model for image recognition. Grad-CAM was applied to the model to generate a heatmap corresponding to the image, thereby identifying key ROIs within the facial area. The primary motivation of the FMAE model is to improve the accuracy of action unit (AU)⁽¹⁹⁾ detection and reduce the model's reliance on identity features. To address the issue of identity learning bias, the FMAE model employs an innovative approach that integrates identity adversarial training (IAT) with the pretrained FMAE architecture. This method aims to enhance the model's ability to learn AU features while mitigating the effects of identity features on model outputs, thereby preventing the phenomenon of identity shortcut learning. By focusing on features closely related to AUs, the model not only effectively improves AU detection accuracy but also enhances robustness in emotion recognition tasks, demonstrating superior performance.

The FMAE model first applies RM to the input facial image, retaining only a portion of the area as the visible region. This allows the model to concentrate on learning useful features within the image while reducing reliance on background or redundant information. Subsequently, a vision transformer (ViT)-based encoder extracts high-level feature representations from the unmasked regions to capture their semantic information. These features, along with the masking information, are fed into the decoder to attempt to reconstruct the masked parts of the image and restore the complete original content. The model is then optimized using a reconstruction loss function, which measures the pixel-level differences between the decoder's output and the original image to learn the image's global contextual structure and details. Subsequently, the model incorporates IAT to introduce adversarial regularization during training, reducing the model's over-reliance on identity features and enhancing its ability to learn AU-related features.⁽¹⁷⁾ FMAE is pretrained on the large-scale Face9M dataset to acquire general facial features and then fine-tuned on specific task datasets to further improve performance. The feature representations ultimately output by the model can be used for facial analysis or emotion recognition, demonstrating excellent accuracy and robustness.

In addition, the FMAE model demonstrates superior performance in addressing occlusion and head pose variations in facial expression recognition, significantly improving the accuracy of emotion detection. The FMAE model has achieved superior performance on a public dataset containing occlusions and head pose variations, known as the real-world affective faces database (RAF-DB).^(17,20) It can accurately extract facial affective features and is further applied to the task of precisely identifying student emotions and engagement, providing robust support for affect analysis in real-world educational settings. Furthermore, in this study, we incorporated noise during the training process, which helps prevent overfitting and thereby enhances the model's generalization ability. This feature enables the model to demonstrate higher stability and reliability in real-world scenarios. The FMAE model utilizes a mask-based learning strategy, enabling it to capture key facial features even when the face is partially occluded, thereby avoiding reliance on irrelevant regions and effectively addressing the issue of facial occlusion.

3.2.2 Heatmap

To distinguish between important ROI and non-interest regions on the face, we employed Grad-CAM⁽¹⁸⁾ to generate a heatmap of the image. Since FMAE adopts a ViT-based encoder rather than a convolutional backbone, the feature maps used by Grad-CAM in this study are extracted from the output of the last Transformer block of the encoder, in which the patch tokens are reshaped into a spatial grid according to their positional order, and the gradients are then calculated via backpropagation.

These gradients are processed via global average pooling (GAP) and then multiplied by the corresponding feature maps, ultimately generating a heatmap that represents the model's ROI. The Grad-CAM architecture overlooks the subtle structural differences among models designed for specific tasks. In the feature layer, weights, representing the importance of each feature, are calculated using the backpropagated gradients and the feature map vectors (values obtained after applying an activation function). Subsequently, CAM is used to generate a preliminary heatmap to highlight the regions deemed important by the model while simultaneously utilizing guided gradient propagation to extract detailed information from the image. Finally, the heatmap is combined with this detailed information to produce a visualization that reflects the model's classification rationale while retaining rich details. Figure 3 shows the heatmap obtained after image recognition using the FMAE model, followed by the application of Grad-CAM.

3.2.3 Conditional random masking

Existing studies provide relevant theoretical and empirical support for the masking-based feature learning strategy adopted in this study. Zhong *et al.*⁽²¹⁾ generated synthetic occlusions by randomly removing local image regions, thereby reducing overfitting and improving the model's adaptability to partial information loss. Similar block-based augmentations, such as Cutout, Hide-and-Seek, and GridMask, remove either a contiguous rectangular region or whole patches of a uniform grid, while attention-guided dropout instead erases the entire most discriminative

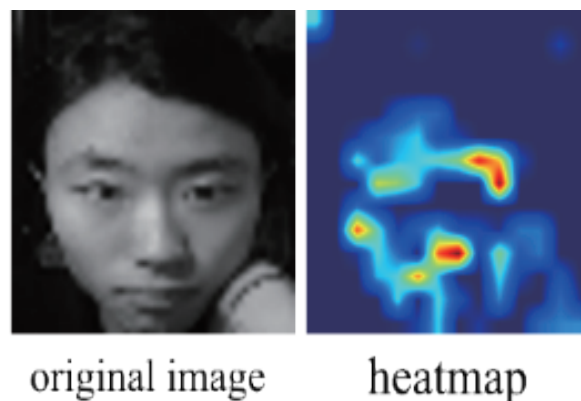


Fig. 3. (Color online) Example of heatmap application.

region; in all of these methods, a discriminative region is either fully retained or fully removed. Gong *et al.*⁽²²⁾ further demonstrated that saliency maps can be used to identify and preserve discriminative image regions, preventing conventional random masking from removing critical information and consequently improving the effectiveness of data augmentation. In addition, Wang *et al.*⁽²³⁾ employed a region attention mechanism to adaptively assess the importance of different facial regions, demonstrating that region-aware feature learning improves facial expression recognition under occlusion and pose variations. These studies support the rationale for the facial heatmap-guided regional masking strategy adopted in this study.

Compared with traditional RM methods, the CRM module proposed in this study is more effective at preserving important ROIs within the facial area. The original image is fed into the FMAE model and Grad-CAM, and the results are then passed as input to the CRM module. First, the CRM module generates a corresponding grid mask based on the heatmap using a preset window size. It calculates the height and width of the heatmap, divides each by the window size to determine the numbers of vertical and horizontal windows, respectively, and then initializes a mask matrix with zeros. Next, the heatmap is binarized to fill the grid mask. Specifically, the heatmap is divided into fixed-size windows, and the average value within each window is calculated. When the average of a window exceeds the set threshold of 0.5, the corresponding mask area is set to 1, and that area is marked as a focus region; when the average of a window is less than the threshold of 0.5, the corresponding mask area is set to 0, and that area is marked as a non-focus region.

The CRM module then applies the RM operation to the focus and non-focus regions in the heatmap based on the set masking ratio of 0.6. Note that the masking is performed within each of the two regions separately, rather than removing the ROI information entirely. Consequently, the model still retains part of the key facial features while reducing its excessive reliance on any single salient region, which encourages it to learn complementary and stable features from different ROIs. By simulating partial visual information loss, this design reduces the model's dependence on individual facial regions and encourages it to learn more stable and discriminative student engagement features under variations in head pose and viewing angle. Moreover, the same masking ratio is applied to both region types to control the masking-strength variable and avoid introducing additional region-specific hyperparameters, which would otherwise increase the complexity of interpreting the results and obscure the contribution of the CRM mechanism itself. For each masked region, the total number of nonzero pixels is first calculated and multiplied by the masking ratio to determine the number of pixels to be masked. A range of pixel indices is then generated, and pixel indices to be masked are randomly selected; the selected pixels are set to 0, effectively masking them. After performing the above operation on the focus region, a focus mask is generated, and after performing the same operation on the non-focus region, a non-focus mask is generated. To ensure that the final mask matches the dimensions of the original image, the CRM module remaps the focus and non-focus masks to the same size as the original window. The two masks are then summed to generate a composite mask. Finally, the composite mask is multiplied elementwise with the original image to produce the final image that preserves the critical facial regions. This result fully retains the important ROIs on the face and effectively separates occluded from non-occluded areas, as shown in Fig. 4.

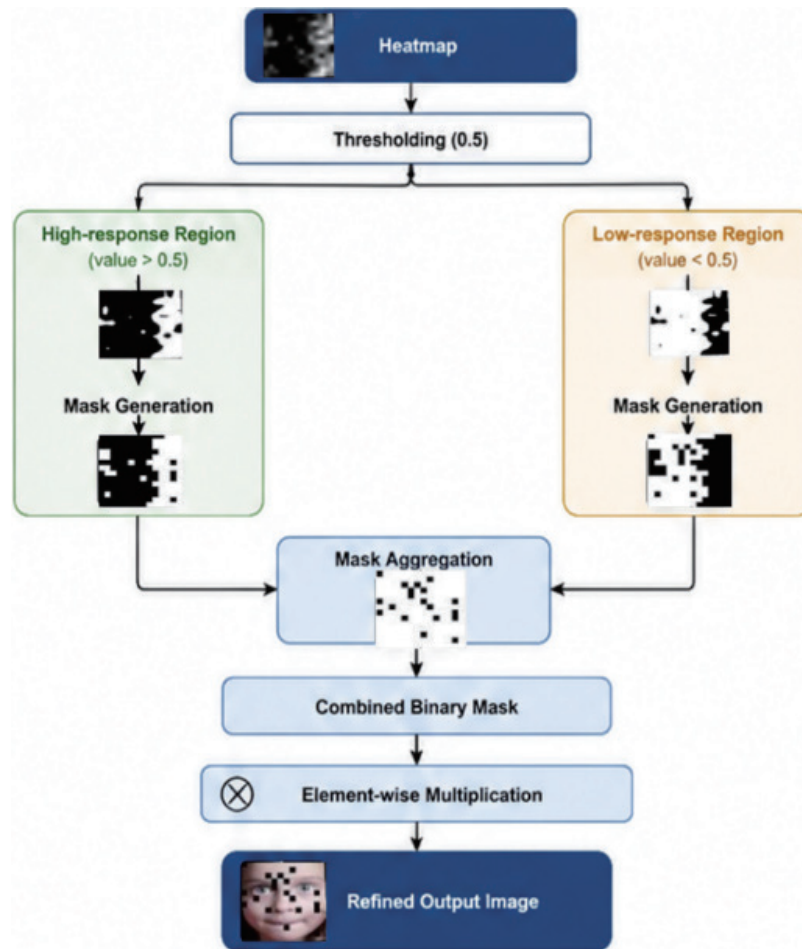


Fig. 4. (Color online) Flowchart of the CRM module.

3.3 ConvNeXt-V2

In this study, we propose the ConvNeXt-V2 model for classifying student engagement. The face-masked images obtained from the CRM module in the previous section are fed into the ConvNeXt-V2 model for classification. ConvNeXt-V2 is a CNN model based on ConvNeXt.⁽²⁴⁾ ConvNeXt-V2 incorporates the concept of a masked autoencoder (MAE) into ConvNeXt; however, since this concept was specifically designed for ViT, it is incompatible with CNN architectures.⁽²⁵⁾ To enable the MAE to operate within a ConvNet architecture, the fully convolutional masked autoencoder (FCMAE) framework was designed. The FCMAE framework uses a ConvNeXt based on sparse convolution as the encoder and a lightweight ConvNeXt block as the decoder. The FCMAE framework is asymmetric: the encoder processes only the unmasked regions, whereas the decoder reconstructs the image by utilizing information from the masked regions and mask tokens, and calculates the loss solely on the basis of the error in the masked regions.

During the training of the FCMAE framework, it was discovered that the multilayer perceptron (MLP) in the ConvNeXt model suffered from feature collapse. To address this issue, the ConvNeXt-V2 model was improved by introducing a global response normalization (GRN) layer to emphasize contrast and selectivity across different channels, thereby resolving the feature collapse problem. In the ConvNeXt-V2 block, the number of input channels remains at 96. Feature extraction is performed using 7×7 depthwise separable convolutions, allowing the model to effectively preserve local spatial information. This is followed by feature normalization via LN, then channel expansion to 384 via 1×1 convolutions, and the application of the GELU activation function to enhance nonlinear expressiveness while simultaneously using a GRN layer to enhance contrast between channels and improve feature selectivity, thereby mitigating the impact of feature collapse and helping the model learn more discriminative feature representations during training. Finally, another 1×1 convolution layer compresses the number of channels back to 96, and the processed features are combined with the input via a residual connection to form the final output.⁽²⁵⁾

4. Experimental Results and Analysis

4.1 Experimental dataset

As shown in Fig. 5, the SED dataset was obtained from Kaggle and consists of six types of academic emotion and 2,120 images. The six academic emotions are further categorized into two engagement levels, namely, engaged and disengaged, providing rich data to advance research on student engagement recognition, as shown in Table 1. The dataset was collected by three different participants to capture the academic emotions they exhibited in the classroom

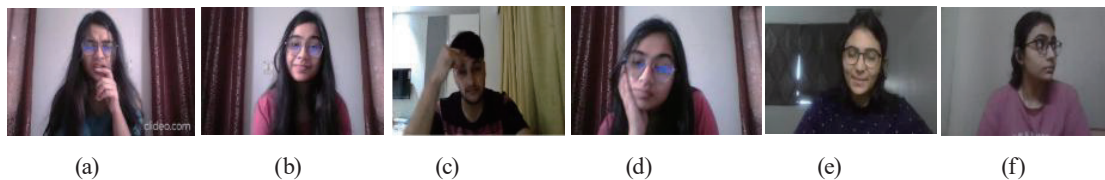


Fig. 5. (Color online) SED dataset: (a) confused, (b) engaged, (c) frustrated, (d) bored, (e) drowsy, and (f) looking away.

Table 1
Academic emotion labels and image counts in the SED dataset.

Engagement level	Category	Academic emotion	No. of images
Engaged	0	Confused	369
	1	Engaged	347
	2	Frustrated	360
	3	Bored	358
Non-engaged	4	Drowsy	263
	5	Looking away	423

and is provided in image format. Since the SED dataset retains its original format, the need for preprocessing in this study is significantly reduced, making it convenient for direct use. The SED dataset was split into a 70% training set and a 30% validation set to train and evaluate the model, with classification based on academic emotions.

Another publicly available dataset is Spontaneous Visual Database for Online Learning (SVD-OL), which consists of six types of academic emotion and 21,632 images. It was collected by 78 college students who were asked to focus on capturing their nonverbal academic emotions and behaviors during online learning activities. The dataset includes spontaneous expressions of academic emotions, bounding boxes for facial and hand localization, and data collected in uncontrolled environments, demonstrating variations in occlusion, lighting, and background. The dataset is split into training and validation sets and classified by academic emotion, as shown in Table 2. Figure 6 shows examples of the six academic emotions in the SVD-OL dataset.

4.2 Experimental settings

The experiments were conducted using the PyTorch framework on a system equipped with an Intel Core i7-12700 CPU and an NVIDIA RTX 3060 GPU. During data preprocessing, all images were resized to 224×224 pixels, the batch size was set to 64, and the random seed was fixed at 123. Data augmentation consists of random horizontal flipping with a probability of 50% and random rotation within a range of $\pm 10^\circ$, using bilinear interpolation. The images were subsequently normalized using the ImageNet mean and standard deviation. The model was trained for 100 epochs using the AdamW optimizer with an initial learning rate of 0.0001 and a cosine annealing learning-rate scheduler.

Table 2
Academic emotion labels and image counts in the SVD-OL dataset.

Category	Academic emotion	No. of images
0	Engaged	6156
1	Happy	3777
2	Neutral	5656
3	Confused	2864
4	Frustrated	885
5	Sleepy	2295

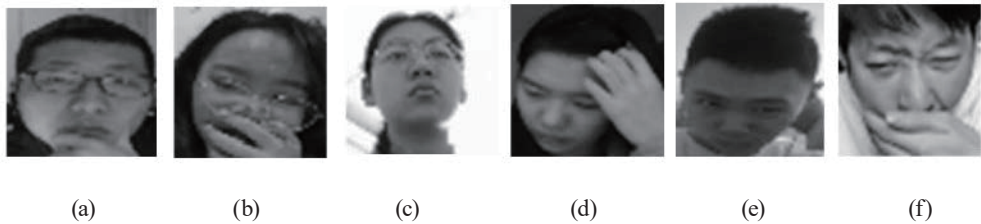


Fig. 6. SVD-OL dataset: (a) engaged, (b) happy, (c) neutral, (d) confused, (e) frustrated, and (f) sleepy.

4.3 Comparison and analysis of results

In this study, we evaluated model performance using three metrics: accuracy, F1-score, and AUC. We conducted a comprehensive comparison of our model's performance in identifying student engagement against other methods reported in the literature on two public datasets: SED and SVD-OL. On the SED dataset, our model achieved an accuracy of 98.15%, which outperforms the accuracy reported in recent related studies. Furthermore, our model achieved an F1-score of 98.16% and an AUC of 99.80%, both of which outperformed classical models. The superior performance of the F1-score and AUC highlights the practical application value of the method, as shown in Table 3. Furthermore, in student engagement recognition experiments conducted on the SVD-OL dataset, the proposed CRM method achieved an accuracy of 87.97%, significantly outperforming the 79.00% accuracy reported in Ref. 14. By retaining the facial attention region and reducing interference from non-attention regions, this method streamlines feature extraction to accurately identify student engagement. Compared with classical models, this study demonstrates superior performance in terms of accuracy, F1-score, and AUC, indicating its reliability in practical applications, as shown in Table 4.

4.4 Ablation study

To evaluate the effectiveness of different data augmentation strategies, ablation experiments are conducted on the SED and SVD-OL datasets, as shown in Table 5. When CRM is applied alone, the model achieves an accuracy of 97.76%, an F1-score of 97.78%, and an AUC of 99.79% on the SED dataset, outperforming RM by 0.29, 0.27, and 0.04%, respectively. On the SVD-OL

Table 3
Comparison of accuracy, F1-score, and AUC for student engagement recognition on the SED dataset.

Method	Accuracy (%)	F1-score (%)	AUC (%)
Perumal <i>et al.</i> ⁽¹⁰⁾	93.00	—	—
Anny <i>et al.</i> ⁽¹¹⁾	95.25	—	—
A-ViT ⁽²⁶⁾	97.03	—	—
POSTER-V2 ⁽²⁷⁾	97.29	—	—
DenseNet ⁽²⁸⁾	97.30	97.32	99.55
EfficientNet ⁽²⁹⁾	96.96	96.98	99.34
DeiT ⁽³⁰⁾	97.64	97.66	99.77
This work	98.15	98.16	99.80

Table 4
Comparison of accuracy, F1-score, and AUC for student engagement recognition on the SVD-OL dataset.

Method	Accuracy (%)	F1-score (%)	AUC (%)
Xu <i>et al.</i> ⁽¹⁴⁾	79.00	—	—
DenseNet ⁽²⁸⁾	87.49	87.85	97.59
EfficientNet ⁽²⁹⁾	87.38	87.72	97.53
DeiT ⁽³⁰⁾	87.93	88.12	97.70
This work	87.97	88.41	97.74

Table 5
Ablation study of different masking strategies on the SED and SVD-OL datasets.

Method	Data augmentation	SED			SVD-OL		
		Acc. (%)	F1 (%)	AUC (%)	Acc. (%)	F1 (%)	AUC (%)
This work	N/A	97.47	97.51	99.75	86.59	87.09	96.65
	RM	97.47	97.51	99.75	83.97	83.90	93.51
	RR	97.66	97.68	99.78	86.08	86.05	96.40
	RHF	97.32	97.35	99.56	83.98	83.92	93.61
	CRM	97.76	97.78	99.79	86.97	86.89	96.76
	RHF + RR	97.66	97.68	99.78	86.51	86.49	97.64
	CRM + RHF + RR	98.15	98.16	99.80	87.97	88.41	97.74

dataset, CRM achieves an accuracy of 86.97%, an F1-score of 86.89%, and an AUC of 96.76%, representing improvements of 3.00, 2.99, and 3.25% over RM, respectively. These results indicate that CRM provides higher overall recognition performance than RM on datasets containing facial occlusion and pose variations. To determine the optimal data augmentation configuration, additional combinations of augmentation strategies are evaluated. When CRM is combined with random horizontal flipping (RHF) and random rotation (RR), the model achieves an accuracy of 98.15%, an F1-score of 98.16%, and an AUC of 99.80% on the SED dataset. On the SVD-OL dataset, it achieves an accuracy of 87.97%, an F1-score of 88.41%, and an AUC of 97.74%, outperforming the other augmentation configurations. Therefore, CRM, RHF, and RR are adopted as the final data augmentation strategy.

To determine the optimal masking ratio, an ablation study is conducted using different masking ratios on the SED and SVD-OL datasets, as shown in Table 6. The results demonstrate that a masking ratio of 0.6 achieves the highest performance on both datasets. Specifically, the model obtains an accuracy of 98.15%, an F1-score of 98.16%, and an AUC of 99.80% on the SED dataset. On the SVD-OL dataset, it achieves an accuracy of 87.97%, an F1-score of 88.41%, and an AUC of 97.74%. Therefore, the masking ratio of CRM is set to 0.6, enabling the model to preserve partial information from critical regions while reducing its dependence on individual salient features, thereby improving the accuracy of student engagement recognition.

To investigate the effects of different window sizes on recognition performance, an ablation study is conducted using different window sizes within CRM, as illustrated in Table 7. When the window size is set to 4, the model achieves an accuracy of 98.15%, an F1-score of 98.16%, and an AUC of 99.80% on the SED dataset. On the SVD-OL dataset, it achieves an accuracy of 87.97%, an F1-score of 88.41%, and an AUC of 97.74%, representing the highest performance on both datasets. In contrast, a window size of 8 produces the lowest performance on both datasets. These results indicate that a smaller window enables a finer partitioning of facial regions and is more suitable when the face occupies a relatively small portion of the image. Therefore, the window size of CRM is set to 4.

To evaluate the effects of different facial-region guidance sources on CRM, we employed the heatmap generated by FMAE and the facial landmarks detected by Dlib⁽³¹⁾ for ROI localization

Table 6
Ablation study of different masking ratios on the SED and SVD-OL datasets.

Method	Mask ratio	SED			SVD-OL		
		Acc. (%)	F1 (%)	AUC (%)	Acc. (%)	F1 (%)	AUC (%)
This work	0.2	97.47	97.51	99.75	86.57	86.56	96.63
	0.4	97.66	97.68	99.78	86.10	86.06	96.43
	0.6	98.15	98.16	99.80	87.97	88.41	97.74
	0.8	97.49	97.53	99.76	86.26	86.27	96.55

Table 7
Ablation study of different window sizes on the SED and SVD-OL datasets.

Method	Window size	SED			SVD-OL		
		Acc. (%)	F1 (%)	AUC (%)	Acc. (%)	F1 (%)	AUC (%)
This work	4	98.15	98.16	99.80	87.97	88.41	97.74
	8	97.47	97.51	99.75	86.99	87.22	97.14
	16	97.81	97.83	99.79	87.09	87.45	97.23

Table 8
Accuracy comparison using different facial-region guidance sources on the SED dataset.

Method	Guidance source	Accuracy
This work	Dlib	87.69 %
	FMAE	97.47 %

and region partitioning. FMAE uses Grad-CAM to generate a heatmap that highlights the regions attended by the model, whereas Dlib constructs ROI by applying morphological dilation to the detected facial landmarks. The experimental results showed that using FMAE as the facial-region guidance source achieves an accuracy of 97.47%, outperforming the 87.69% obtained using Dlib. One possible reason is that facial occlusion and head-pose variations may prevent Dlib from accurately detecting all facial landmarks, thereby reducing the reliability of ROI localization. In contrast, FMAE combined with Grad-CAM identifies discriminative regions based on the learned facial representations. Therefore, we selected FMAE as the facial-region guidance source for CRM, as shown in Table 8.

5. Conclusions

In this study, we proposed a novel CRM method to enhance the model's capability to capture key facial regions and improve robustness. The main contributions are summarized as follows: (1) A region-decomposition-based CRM framework was introduced, which partitions the face

into ROIs and non-interest regions to explicitly model their distinct contributions to emotion recognition. By simulating partial information loss in camera-acquired facial data, CRM encourages the model to extract engagement-related features from incomplete visual sensing signals. (2) A conditional PM module was designed to perform region-specific masked representation learning, followed by adaptive fusion to generate discriminative facial representations. (3) A robust feature learning mechanism was developed, enabling accurate recognition using a limited set of key features, even under occlusion and pose variations. Extensive experiments on the SED and SVD-OL datasets demonstrated that, in terms of accuracy, F1-score, and AUC, the proposed method consistently outperforms state-of-the-art approaches. It also exhibits superior stability under diverse facial expressions and complex environmental conditions, thereby validating its effectiveness, robustness, and generalization capability in real-world learning scenarios.

References

- 1 C.-H. Hsia, B. Chiang, L.-Y. Ke, Z.-Y. Ciou, and C.-F. Lai: IET Int. Conf. Engineering Technologies and Applications (2022) 1–2.
- 2 J. Whitehill, Z. Serpell, Y. C. Lin, A. Foster, and J. R. Movellan: IEEE Trans. Affective Computing **5** (2014) 86.
- 3 R. Pekrun and L. L. Garcia: Handbook of Research on Student Engagement (2012) pp. 259–282.
- 4 H. Wang, B. Li, S. Wu, S. Shen, F. Liu, S. Ding, and A. Zhou: IEEE/CVF Conf. Computer Vision and Pattern Recognition (2023) 17958–17968.
- 5 C. Y. Wu, R. Girshick, K. He, C. Feichtenhofer, and P. Krahenbuhl: IEEE/CVF Conf. Computer Vision and Pattern Recognition (2020) 153–162.
- 6 Q. You, J. Luo, H. Jin, and J. Yang: AAAI Conf. Artificial Intelligence **30** (2016).
- 7 M. Yu, H. Zheng, Z. Peng, J. Dong, and H. Du: Pattern Recognit. Lett. **131** (2020) 166.
- 8 T. S. Ashwin and R. M. R. Guddeti: IEEE Access **7** (2019) 150693.
- 9 Z. Zhang, Z. Li, H. Liu, T. Cao, and S. Liu: J. Educ. Comput. Res. **58** (2020) 63.
- 10 B. Perumal, P. Nagaraj, T. S. N. Charan, Y. V. S. SaiDeepak, C. V. V. Reddy, and S. Nagendra: Int. Conf. Communication and Electronics Systems (2023) 1233–1238.
- 11 T. N. Anny, J. I. Chowdhury, T. Ahsan, I. Zahur, M. J. A. Patwary, and M. H. Miraz: Int. Conf. Computing, Electronics & Communications Engineering (2023) 56–61.
- 12 G. Yovel and B. Duchaine: J. Cognitive Neurosci. **18** (2006) 580.
- 13 L. Li and D. Yao: IEEE Access **11** (2023) 143050.
- 14 L. Xu, Y. Gan, and Y. Jin: IEEE Trans. Affective Comput. **16** (2025) 1044.
- 15 A. Sukumaran and A. Manoharan: IEEE Access **12** (2024) 25228.
- 16 U. L. Sowjanya and R. Krithiga: IEEE Access **12** (2024) 106273.
- 17 N. Mang, A. A. Salah, and I. O. Ertugrul: arXiv preprint arXiv:2407.11243 (2024).
- 18 R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra: IEEE Int. Conf. Computer Vision (2017) 618–626.
- 19 J.-M. Guo, C.-H. Hsia, L.-W. Hsiao, and C.-C. Yao: J. Inf. Sci. Eng. **37** (2021) 793.
- 20 S. Li, W. Deng, and J. Du: IEEE Conf. Computer Vision and Pattern Recognition (2017) 2852–2861.
- 21 Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang: AAAI Conf. Artificial Intelligence **34** (2020) 13001.
- 22 C. Gong, D. Wang, M. Li, V. Chandra, and Q. Liu: IEEE/CVF Conf. Computer Vision and Pattern Recognition (2021) 1055–1064.
- 23 K. Wang, X. Peng, J. Yang, D. Meng, and Y. Qiao: IEEE Trans. Image Process. **29** (2020) 4057.
- 24 Z. Liu, H. Mao, C. Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie: IEEE/CVF Conf. Computer Vision and Pattern Recognition (2022) 11976–11986.
- 25 S. Woo, S. Debnath, R. Hu, X. Chen, Z. Liu, I. S. Kweon, and S. Xie: IEEE/CVF Conf. Computer Vision and Pattern Recognition (2023) 16133–16142.
- 26 H. Yin, A. Vahdat, J. M. Alvarez, A. Mallya, J. Kautz, and P. Molchanov: IEEE/CVF Conf. Computer Vision and Pattern Recognition (2022) 10799–10808.

- 27 J. Mao, R. Xu, X. Yin, Y. Chang, B. Nie, A. Huang, and Y. Wang: *Pattern Recognit.* **157** (2024) 110951.
- 28 G. Huang, Z. Liu, L. v. d. Maaten, and K. Q. Weinberger: *IEEE Conf. Computer Vision and Pattern Recognition* (2017) 4700–4708.
- 29 M. Tan and Q. Le: *Int. Conf. Machine Learning* (2019) 6105–6114.
- 30 H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jegou: *Int. Conf. Machine Learning* (2021) 10347–10357.
- 31 D. King: *Dlib: a toolkit for machine learning* (2009) <http://dlib.net/>