DOI 10.15622/ia.24.2.10

A. NAVEEN, I. JACOB, A. MANDAVA

DETECTION OF STUDENT ENGAGEMENT VIA TRANSFORMER-ENHANCED FEATURE PYRAMID NETWORKS ON CHANNEL-SPATIAL ATTENTION

Naveen A., Jacob I., Mandava A. **Detection of Student Engagement via Transformer-Enhanced Feature Pyramid Networks on Channel-Spatial Attention.**

Abstract. One of the most important aspects of contemporary educational systems is student engagement detection, which involves determining how involved, attentive, and active students are in class activities. For educators, this approach is essential as it provides insights into students' learning experiences, enabling tailored interventions and instructional enhancements. Traditional techniques for evaluating student engagement are often time-consuming and subjective. This study proposes a novel real-time detection framework that leverages Transformer-enhanced Feature Pyramid Networks (FPN) with Channel-Spatial Attention (CSA), referred to as BiusFPN_CSA. The proposed approach automatically analyses student engagement patterns, such as body posture, eye contact, and head position, from visual data streams by integrating cutting-edge deep learning and computer vision techniques. By integrating the attention mechanism of CSA with the hierarchical feature representation capabilities of FPN, the model can accurately detect student engagement levels by capturing contextual and spatial information in the input data. Additionally, by incorporating the Transformer architecture, the model achieves better overall performance by effectively capturing long-range dependencies and semantic relationships within the input sequences. Evaluation using the WACV dataset demonstrates that the proposed model outperforms baseline techniques in terms of accuracy. Specifically, in terms of accuracy, the FPN_CSA_Trans_EH variant of the proposed model outperforms FPN_CSA by 3.28% and 4.98%, respectively. These findings underscore the efficacy of the BiusFPN_CSA framework in real-time student engagement detection, offering educators a valuable tool for enhancing instructional quality, fostering active learning environments, and ultimately improving student outcomes.

Keywords: Feature Pyramid Network (FPN), Channel-Spatial Attention (CSA), student engagement detection, Transformer.

1. Introduction. When students are actively involved in their educational assignments and activities, this is referred to as student engagement. This involvement not only directly impacts on school improvements, such as enhancing teachers' professional identities and fostering a welcoming school environment [1], but it also appears to boost the academic performance of underperforming students, reduce dropout rates, and decrease dissatisfaction. Because of this, scholars have remained highly interested in student engagement and its various implications over the past 20 years. Academic success has always been considered a crucial result of student engagement. With the rise of network technology, computer technology, and other advancements, online learning has emerged. It emphasises communication between students and the accessibility of the learning resources. A significant number of students

now participate in online learning, which has become a predominant learning method. However, due to limited interaction between teachers and students, communication is often insufficient, student participation in online learning tends to be suboptimal, inconsistent, and inefficient. Student engagement is essential for learning and significantly impacts online learning as well.

As the phrase "student engagement" can have different meanings for different people, the method of assessing student engagement utilised by researchers in their studies is akin to selecting a specific conceptualization of the construct. Prior to choosing a method for assessing student engagement, it is important to define the term precisely. Early researchers frequently operationalized student engagement in terms of observable behaviours, such as the level of participation in various tasks and the time required to complete them. Various aspects, such as facial expression recognition, head pose detection, and body language analysis, can be used to assess student attention in class. Facial expressions have been used in studies to analyse student participation in both in-person and online classes, and results have demonstrated that this method is effective for determining student engagement levels. This method has been used in studies to evaluate student participation in face-to-face and online classes, and it has proven effective in determining levels of engagement. For instance, the authors in [19] proposed a system that employs facial expression recognition to measure engagement levels in real time, and they demonstrated its effectiveness in a classroom setting.

In order to support timely intervention, it is essential to assess and study student engagement in online learning, help teachers understand student engagement, enable students to reflect on their learning, and encourage participation in the learning process. The quantification of student behaviour, cognitive engagement, and emotional engagement are all aspects of measuring student engagement. Currently, research has focused on studying student participation through theoretical models, explicit behavioural data, influencing factors, effect analysis, and the lack of accurate measurement of student engagement. As a result, the research developed a reliable and quantifiable model of student engagement, examining students' cognitive engagement, emotional engagement, and engagement patterns. Researchers are exploring ways to incorporate transformers into computer vision applications due to their powerful representational capabilities. Transformer-based models outperform convolutional and recurrent neural networks across various visual benchmarks, with some models achieving superior performance. The computer vision community is increasingly focusing on transformers due to

their strong performance and reduced need for inductive bias tailored to specific types of vision. Transformer, a type of deep neural network primarily based on the self-attention mechanism, was first applied in natural language processing. In this study, an enhanced transformer is presented to identify student engagement. The transformer is applied for the first time in engagement recognition. The findings reveal a positive correlation between the final exam scores and the level of student participation in the online class sessions.

The main contributions of this work are summarized as follows:

1. We propose a Feature Pyramid Network model with a location and channel-aware attention module to effectively learn facial representations during online class sessions.
2. The transformer module is also integrated with this model to analyse the global context features along with local convolutional features.
3. The encoded FPN-based significant features are combined with the eye and head movement-based features to enhance the performance of the framework.
4. The performance of the proposed FPN_CSA_Trans is analyzed with the help of DAiSEE and WACV datasets.

2. Related Work. Facial expression analysis requires more advanced expertise in the field of computer vision. In recent years, there has been growing interest in using technology to track and assess students' facial expressions in order to better understand and enhance their engagement in class. The emotions of students can be inferred from their facial expressions, which are a nonverbal form of communication [1]. Analysis of body language was employed by some researchers to gauge levels of engagement. This method has been used in studies to assess student participation in both traditional and online classroom settings, demonstrating its usefulness in determining levels of engagement. This may involve analysing a student's body in various positions, such as sitting or standing, as well as their head and gaze. According to research in this field, body position evaluation may be employed to determine levels of engagement.

For example, the authors in [2] used body posture and movement analysis to assess engagement levels in real time, proving its efficacy in a classroom context, but it lacked facial-based features. Numerous other publications [3, 4] also proposed using body language analysis to gauge student engagement during lectures. According to the authors in [3], who used a sample of 800 students, the accuracy rate for identifying engagement levels was 89.3%. Keystroke dynamics, or the analysis of typing habits, such as speed and errors, has been widely used by researchers to identify levels of engagement. According to research, keystroke dynamics can be

used to gauge student interest levels and enhance the effectiveness of instruction. Keystroke dynamics may vary depending on the scenario. Another method for determining engagement levels involves analysing body language and head position. This may involve analysing body posture, eye contact, and head position. According to research in this field, head posture and body language can be used to gauge student engagement levels and improve instructional efficacy. The use of head and body posture alone lacks the ability to extract features from the facial region.

The authors in [5, 6] proposed a real-time student engagement detection method. Students' eye tracking and head movements were recorded using a depth camera, and machine learning techniques were used to classify their engagement levels. They used a depth camera to capture students' eye tracking and head movements, and applied machine learning algorithms to classify engagement levels.

Another approach involves using head pose and body movements to detect engagement levels. This may include analysing the position of head, eye gaze, and body posture [20]. Research in this area has shown that head pose and body movements can be used to detect engagement levels and improve the effectiveness of teaching. To address occlusion, researchers in [7 – 10] used texture features or reconstructed geometric features. To recover a lost or drifted facial point, an improved Kanade-Lucas tracker [7] was proposed. PCA-based approaches were used for missing point reconstruction [8, 9]. Another method for identifying facial expressions, known as the modified transferable belief model, was proposed in [10]. The performance of the facial expression analysis mechanism can be affected by facial poses. To address pose variations, the authors recommended training with a single classifier [11]. Adversarial feature learning [12] was used by researchers for the same purpose. Study [13] proposed using the k-means algorithm to group students based on 12 engagement measures divided into two categories: interaction-related and effort-related. Quantitative analysis is used to identify students who are not engaged and may require assistance. In order to identify student participation in a classroom setting, the authors in [14] proposed a real-time facial expression detection system. They recorded students' faces using a camera and then applied machine learning algorithms to classify their facial expressions into different levels of engagement. The authors in [15] proposed a deep learning-based method for identifying student interest in video-based online classes, using aspects of facial expression, head pose, and gaze. They found that their approach achieved high accuracy in determining engagement levels. Using a Histogram of Oriented Gradients, further trained by a CNN, the authors in [16] were able to extract facial features. Since the histogram of oriented

gradients extracts information from spatial gradients, this study achieved better performance. Studies [17, 18] also used variables alongside facial expression to determine student interest in the classroom. The authors in [19] used this method to analyse student participation in online classes, demonstrating its effectiveness in determining engagement levels. According to another study [20], engagement levels can be determined using a combination of head pose, facial expression, and gaze analysis. The study found that different modalities are used for student engagement analysis. Each modality has its own strengths. Therefore, in this paper, features learned from face, eye, and head movements are used for training and testing.

The analysis of various studies shows that student engagement through modalities, such as facial expressions, head movements, gaze tracking, body posture, and keystroke dynamics, has yielded promising results in both traditional and online classrooms. Each modality has its strengths but also presents limitations. For example, methods focusing on body posture or keystroke dynamics often overlook finer details of facial expressions, while facial expression analysis may struggle with variations caused by occlusion, lighting conditions, and pose differences.

3. Methodology

3.1. The proposed FPN_CSA model for Engagement Analysis. The current study aims to investigate three approaches to applying facial recognition technology in classroom engagement analysis. By leveraging the Transformer mechanism and the BiusFPN with Inter-Cross Coordinate Self-Attention model, a person's engagement based on facial expression during online sessions can be recognised. Frames are extracted from the real-time video stream and used for face detection. Eye tracking is performed after face detection, and head rotation is also analysed. These three metrics are used to determine the engagement level. The outcome is classified as disengaged, partially engaged, and fully engaged. This method effectively identifies specific facial expressions associated with varying degrees of engagement. The overall architecture of the proposed model is depicted in Figure 1.

Using Resnet-18 as the backbone, the Feature Pyramid Network (FPN) is employed in the design of the proposed framework for engagement analysis. For a single flow-down sampling process, this bi-directional Feature Pyramid Network (FPN) manages two upsampling operations.

This model integrates traditional channel and spatial attention at the second level of the upsampling process, along with an Inter-Cross Coordinate Self-Attention model. Figure 2 illustrates the general architecture of the proposed framework.

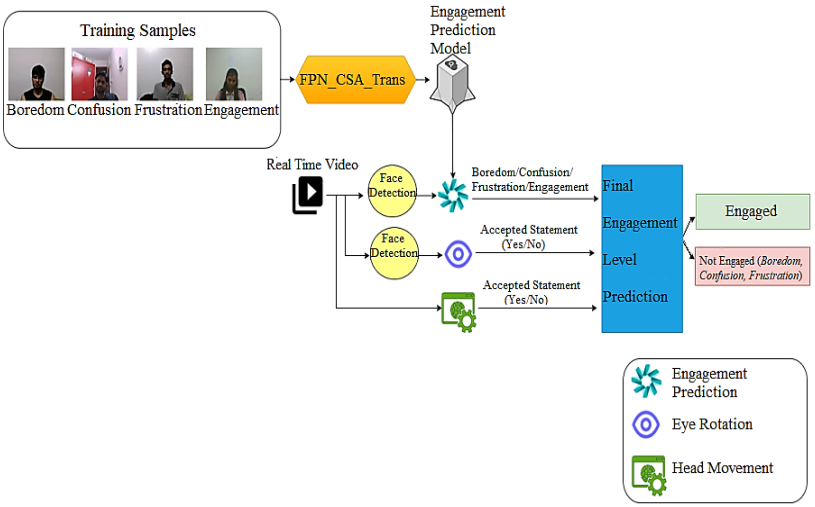

Fig. 1. Overall architecture of the proposed work

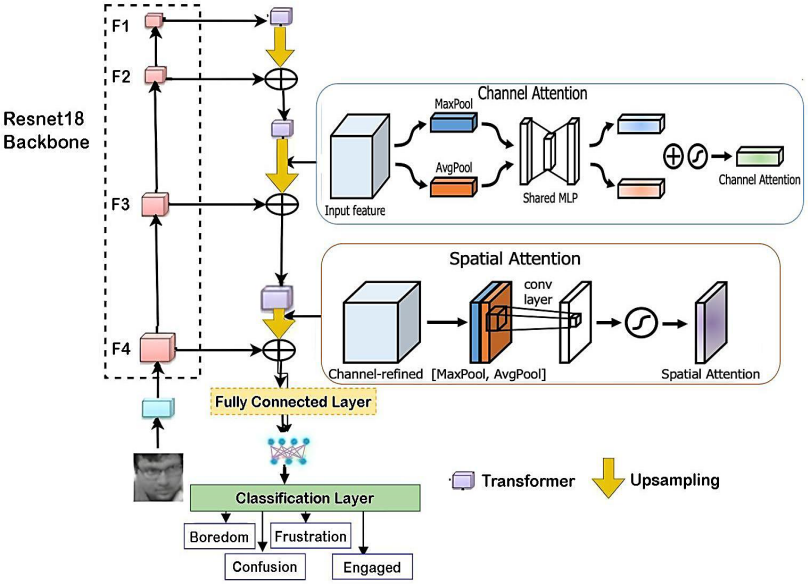

Fig. 2. Architecture of the proposed FPN_CSA_Trans model

3.1.1. Resnet-18. The term "ResNet" refers to the 18-layer Convolutional Neural Network introduced by [21]. Designed to facilitate the effective operation of multiple convolutional layers, ResNet-18 is a 72-layer architecture with 18 deep layers, including residual blocks, as described in [22]. However, as a network is expanded with multiple deep layers, the output performance usually deteriorates. The vanishing gradient problem is addressed by neural networks using gradient descent to determine the weights that minimise the loss function during backpropagation training. The gradient "vanishes," leading to network saturation or even performance loss due to repeated multiplication across multiple layers. Residual Blocks in ResNet-18 utilise skip connections to address the vanishing gradient issue. The skip connection bypasses a few intermediate levels in order to connect layer activations to subsequent layers. As a result, the residual block remains intact. The approach used in this system allows the residual mapping to fit the system, rather than requiring the layers to learn the underlying mapping directly. Figure 3 illustrates the skip connection mechanism within a residual block.

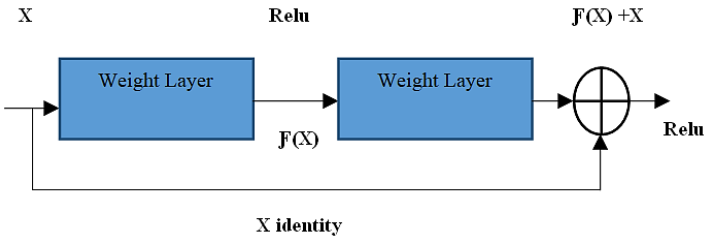

Fig. 3. Skip connections

The benefit of skip connections is that they prevent any layer from degrading the network's performance. Thus, vanishing or exploding gradients do not pose problems when training very deep neural networks. The backbone network for the proposed model is the Resnet-18.

3.1.2. Transformer with Multi-Head Attention. Numerous advances in deep learning tasks have resulted from Transformers [Vaswani et al., 2017; Devlin et al., 2019; Velickovic et al., 2018b]. The Transformer stands out due to its ability to combine all computations in the same layer and its lack of recurrent connections, which improves scalability, effectiveness, and efficiency. The Transformer only uses the attention mechanism to determine the dependencies between input tokens, eliminating the need for recurrent connections. To be more precise, the Transformer utilises a novel multi-head attention module designed to more effectively recognise the dependencies between input tokens.

It has been noted that a key factor in the Transformer's success is its multi-head attention module. Recurrent neural networks (RNNs) have been shown to outperform Transformers on machine translation benchmarks when both utilise multi-head encoder-decoder attention. In contrast, Transformers perform worse when not utilising multi-head attention [5]. In addition to the Transformer, multi-head attention has been implemented in RNNs [5], Graph Attention Networks [Velickovic et al., 2018a], Convolutional Neural Networks [Xiao et al., 2020; Fang et al., 2019], and other architectures.

The overarching belief is that multi-head attention distinguishes itself by attending to multiple positions concurrently, whereas a conventional attention module can only focus on one position in a single layer. Multi-head attention specifically performs multiple attention computations in parallel and projects the input data into multiple distinct subspaces (Figure 4).

The combination of ResNet-18, FPN, and Transformers leverages the strengths of each architecture, making it a powerful choice for analysing visual data where both spatial and sequential information must be effectively captured, such as in assessing student engagement levels from video feeds. This setup addresses both the efficiency and depth of feature extraction required for accurate and real-time predictions, outperforming the narrower focus and limitations of RNNs or standalone CNNs in such applications.

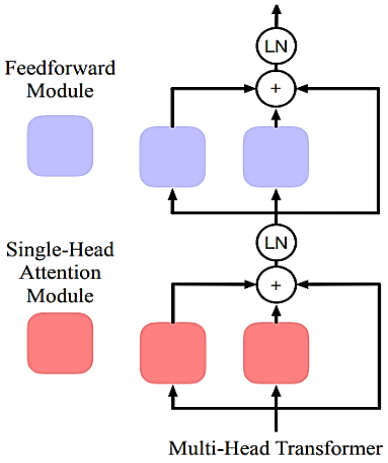

Fig. 4. Multi-Head Transformer

3.1.3. Feature Pyramid. An essential component of recognition algorithms is the feature pyramid, which is used to detect objects at different scales. An FPN is a multi-scale deep learning framework that builds feature pyramids with minimal additional computational cost. The design is widely integrated to create high-level semantic maps at all scales. In various applications, this performs significantly better as a general feature extractor. In recognition tasks, convolutional networks (ConvNets) have largely replaced handcrafted features. ConvNets are more robust in representing higher-level semantics with scale invariance that can identify computed features from a single input scale. Multi-scale feature extraction is achieved by characterising the high-resolution features at each pyramid level. Feature pyramids are constructed using top-down, bottom-up, and lateral connections [23].

The semantically stronger features, which are subsampled fewer times and thus have more accurate localisation information, are combined with features from earlier levels in the architecture developed by the Feature Pyramid Network (FPN). When the FPN, which serves as a feature extractor, was developed, the precision and speed of the pyramid concept were key considerations. Unlike detectors such as the extractor for object recognition in Faster R-CNN, the FPN generates multiple feature map layers with higher-quality data compared to conventional feature pyramids. The use of multi-scale feature maps from multiple layers computed during the forward pass makes it computationally efficient. Given its numerous advantages, our proposed model, which uses ResNet-18 as the backbone, incorporates the feature pyramid.

We used a 100×100 grayscale input image containing only detectable faces in our model. Initially, a convolutional layer with 64 filters and a 3×3 kernel size was applied to the input image. The convolutional layer produces an output of size 100×100×64. Subsequently, these feature maps are passed to the ResNet-18. In ResNet-18, we utilised four different types of convolutional layers. Four feature maps, designated as F1, F2, F3, and F4, were produced. The first convolutional layer in ResNet-18, corresponding to F4, has 64 filters, a kernel size of 3×3, and a stride of 1. As a result, the F4 feature map retains the original size of 100 x 100 with 64 filters. The second convolutional layer in ResNet-18, corresponding to F3, has 128 filters, a kernel size of 3×3, and a stride of 2. As a result, the F3 feature map has dimensions of 50×50×64. The third convolutional layer in ResNet-18, corresponding to F2, uses 256 filters, a kernel size of 3x3, and a stride of 2. Consequently, the F2 feature map has dimensions of 25×25×256. The final convolutional layer in ResNet-18, corresponding to F1, has 512 filters, a kernel size of 3×3, and a stride of 2. As a result, the F1

feature map has dimensions of 13×13×512. The ResNet-18 processing is now complete. Next, we proceed to the pyramidal structures in the Feature Pyramid Network (FPN) using upsampling and addition.

Subsequent processing is applied to each of the feature maps: F1, F2, F3, and F4. Following the F1 feature map, a 2D convolutional layer with 256 filters, a kernel size of 1x1, a stride of 1, dilation of 1, groups of 1, and ReLU activation is applied. It is then passed through a transformer layer and upsampled before being merged with F2. The result of the first addition, add1, has dimensions of 25×25×256. The output of add1 is then passed through a convolutional layer with 128 filters. After passing through a transformer layer, it is upsampled and merged with F3. The result of the second addition, add2, has dimensions of 50×50×128. The output of add2 is then passed through a convolutional layer with 64 filters. After that, it is upsampled and merged with F4. The result of the third addition, add3, has dimensions of 100×100×64. To provide input for the second stage of upsampling, the output from the first upsampling in the FPN is passed through three distinct attention mechanisms: Channel Attention, Spatial Attention, and the proposed ICCSA.

3.1.4. Channel and Spatial Attention. A channel attention module [24] is utilised in convolutional neural networks to provide channel-based attention. Figure 5 illustrates the channel attention architecture. An attention mechanism is introduced to create a channel attention map by leveraging the relationships between features across channels. Given an input image, the channels of a feature map act as feature detectors, and thus, channel attention focuses on "what" is important. Effective channel attention computation requires a reduction of the spatial dimension of the input feature map. Using both average-pooling and max-pooling processes, the spatial information of a feature map is aggregated to produce two unique spatial context descriptors, F_{avg}^{c} and F_{max}^{C}, representing average-pooled features and max-pooled features, respectively. Subsequently, both descriptors are passed through a shared network, which generates the channel attention map $M_c \in R^{cX1X1}$. Here, C represents the number of channels. The shared network consists of multi-layer perceptrons (MLPs) with a single hidden layer. To minimise parameter overhead, the hidden activation size is set to $R^{c/rX1X1}$, where r is the reduction ratio. After processing each descriptor through the shared network, the output feature vectors are combined element-wise. In summary, channel attention is computed as follows: here, F represents the input feature, where AvgPool and MaxPool denote the average and max pooling operations, as shown in Figure 5. Equation (1) represents the overall process of channel attention, as illustrated in Figure 5, with the corresponding module notations.

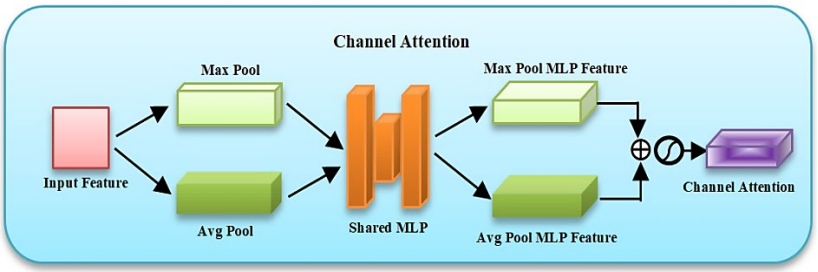

Fig. 5. Channel Attention

$$M_c(F) = \sigma(MLP(AvgPool(F) + MLP(MaxPool(F))),$$
$$M_c(F) = \sigma(W_1\left(W_0\left(F_{avg}^c\right)\right) + W_1(W_0(F_{max}^C))), \quad (1)$$

where $W_0 \in R^{c/rXc}$, $W_1 \in R^{cXc/r}$, and σ denotes the sigmoid function. Note that the MLP weights W_0 and W_1 for both inputs are followed by the ReLU activation function.

The spatial attention module is another component in convolutional neural networks [24]. It generates a spatial attention map by leveraging the spatial relationships among features. Spatial attention focuses on "where" information is located, in contrast to channel attention, which focuses on "what" is informative. Before computing spatial attention, average-pooling and max-pooling operations are performed along the channel axis, and the results are concatenated to produce an effective feature descriptor. The spatial attention map $M_s(\mathrm{F}) \in R^{HXW}$, which encodes where to emphasise or suppress, is generated by applying a convolution layer to the concatenated feature descriptor. Two 2D maps, $F_{avg}^s \in R^{1XHXW}$ and $F_{max}^s \in R^{1XHXW}$, are generated by pooling the channel information in a feature map using two different methods. Each map represents the max-pooled and average-pooled features of the channel, respectively. These maps are concatenated and convolved using a standard convolution layer to generate the 2D spatial attention map. In summary, the computation of spatial attention is described in Equation 2, with the corresponding module notations and mathematical representations.

$$M_s(F) = \sigma\left(f^{7x7}([AvgPool(F); MaxPool(F)])\right),$$
$$M_s(F) = \sigma\left(f^{7x7}\left(\left[F_{avg}^s; F_{max}^s\right]\right)\right), \quad (2)$$

where the convolution operation with a 7 × 7 filter size is denoted by f^{7x7}, and σ represents the sigmoid function. Figure 6 illustrates the spatial attention architecture.

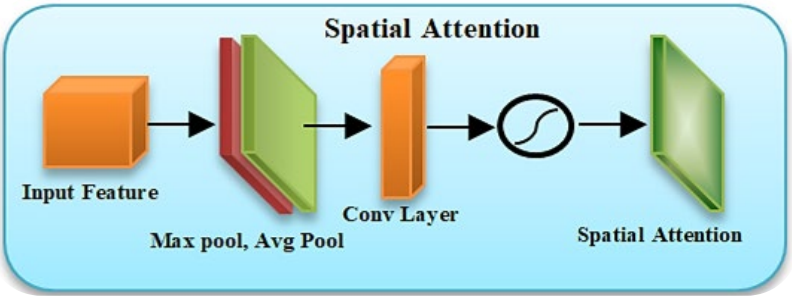

Fig. 6. Spatial Attention

In the decoding upsampling flow, the input is first processed by the transformer block, as shown in Figure 2. After processing through the transformer block, the output is upsampled and combined with the corresponding next-level encoded output. In the second level, the process is repeated, but channel attention features are added to extract more channel-oriented information. In the third level of FPN decoding, the transformer is incorporated, followed by upsampling and addition with the spatial attention module. The final third-level addition block produces a feature map of size 100×100×64. After the fully connected layer, the output is passed to a final fully connected layer with four classes and a filter size of 1,086. The classification layer classifies the input image into one of the categories: Disengaged, Partially Engaged, or Engaged.

4. Experimental Results. The proposed FPN_CSA_Trans_EH method achieved an accuracy of 71.02% on the DAiSEE dataset and 88.57% on the WACV dataset.

4.1. Dataset Description

4.1.1. DAiSEE dataset. The DAiSEE dataset [25] is publicly available. It consists of video recordings of participants in an e-learning environment, annotated with publicly sourced labels for engagement, frustration, confusion, and boredom. The dataset, made publicly available along with unique crowd-sourced annotations, captures real-world "in the wild" environments. The methodologies for data collection, annotation, and vote aggregation are described below. The DAiSEE dataset includes 9,068 clips from 112 students taking online courses. The four states of people watching online courses – boredom, confusion, frustration, and engagement – were annotated on the videos. Each state is assigned one of four ordinal

levels: 0 (very low), 1 (low), 2 (high), and 3 (very high). This work focuses solely on the classification of engagement levels. The clips are 10 seconds long, recorded at 30 frames per second (fps), with a resolution of 640×480 pixels (Figure 7).

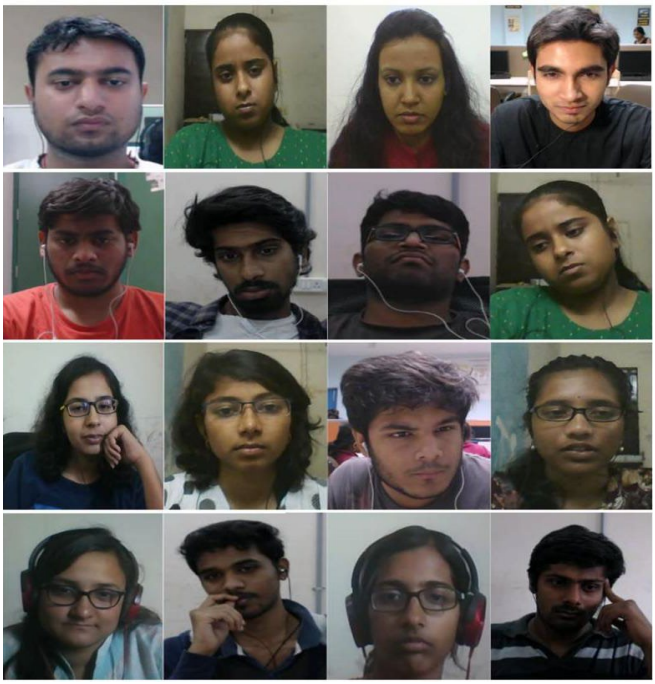

Fig. 7. Samples of the DAiSEE dataset:
Engagement (first row), Boredom (second row), Confusion (third row),
and Frustration (bottom row)

4.1.2. WACV dataset. This section describes the evaluation of student engagement levels using the WACV dataset. We used the open-source WACV dataset [28] for our research. The dataset contains three distinct classes: disengaged, partially engaged, and engaged. The dataset consists of 4,424 RGB images of varying sizes. All images were resized to a uniform shape of 100×100×3. The dataset is not balanced, with 412 images in the "disengaged" class, 2,247 images in the "partially engaged" class, and 1,765 images in the "engaged" class. To create a balanced set of 412 photos for each class, we randomly selected 412 images from the "partially engaged" class and 412 images from the "engaged" class.

We divided this data into training and testing sets (80% and 20%, respectively). The figures below show class-wise examples from the WACV and the DAiSEE datasets.

Figure 8 illustrates the Disengaged, Partially Engaged, and Engaged samples of the WACV dataset; and Figure 7 illustrates the Boredom, Confusion, Frustration, and Engagement samples of the DAiSEE dataset.

Fig. 8. Engaged (top), partially engaged (middle), and disengaged (bottom) samples of the WACV dataset

4.2. Evaluation metrics.

A. *Accuracy.* Accuracy measures the overall performance of the classifier. The model's performance is evaluated by comparing the percentage of accurate predictions to all cases. Accuracy is calculated using the following formula:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}. \quad (3)$$

B. AUC. The ROC curve's summary, Area Under the Curve (AUC), measures how successfully a classifier can distinguish between classes. A higher AUC indicates better performance in distinguishing between positive and negative classes.

C. *Gini Index.* The Gini Index is calculated by subtracting the sum of the squared probabilities of each class from one. It is simple to compute and tends to favour larger segments. In simple terms, it measures the

probability that a randomly selected feature is misclassified. The Gini index is calculated as follows:

$$GI = 1 - \sum_{i=1}^{n}(P_i)^2. \quad (4)$$

D. *AGF*. The AGF metric is an enhanced version of the F1 score that can accurately assess the performance of our algorithm even with imbalanced data. A high AGF value indicates that class imbalance had minimal or no impact on the results. The AGF is calculated using the following formula:

$$\text{AGF} = \sqrt{\text{inv}F_{0.5} \times F_2}. \quad (5)$$

5. Results and Discussion

5.1. WACV Dataset. The graphical representation of the accuracy comparison between the WACV dataset and existing methods is provided in Table 1. In terms of Accuracy, FPN_CSA_Trans_EH outperforms FPN_CSA_Trans by 3.28%, FPN_CSA by 4.98%, ResNet-18 [28] by 8.57%, DenseNet-121 [28] by 10.57%, MobileNetV1 [28] by 22.57%, HOG+SVM [28] by 30.57%, and CNN [27] by 51.57%.

Table 1. Accuracy of different methods for the WACV dataset

WACV	Accuracy(%)
CNN[27]	37
HOG+SVM[28]	58
MobileNetV1[28]	66
DenseNet-121[28]	78
ResNet-18[28]	80
FPN_CSA	83.59
FPN_CSA_Trans	85.29
FPN_CSA_Trans_EH	**88.57**

Figure 9 illustrates the class-wise accuracy comparison between ResNet-18 and FPN_CSA_Trans_EH (Table 2). In terms of accuracy, FPN_CSA_Trans_EH outperforms ResNet-18 by 6.1% for the disengaged class, 9.58% for the partially engaged class, and 7.37% for the engaged class.

Table 2. Comparison of metrics for ResNet-18 and FPN_CSA_Trans_EH

Engaged	**Partially engaged**	**Disengaged**	**Classes**	
83.56%	76.83%	84.14%	Accuracy	ResNet18[28]
90.09%	84.3%	90.03%	AUC	
80.18%	68.6%	80.07%	Gini Index	
88.49%	83.31%	87.08%	AGF	
86.11%	81.73%	82.92%	Accuracy	FPN_CSA
88.35%	86.73%	87.78%	AUC	
76.7%	73.46%	75.57%	Gini Index	
88.36%	84%	85.13%	AGF	
87.53%	83.07%	87.8%	Accuracy	FPN_CSA_Trans
89.24%	88.77%	90.28%	AUC	
78.49%	77.55%	80.52%	Gini Index	
89.4%	85.63%	87.51%	AGF	
90.93%	**86.41%**	**90.24%**	**Accuracy**	**FPN_CSA_Trans_EH**
92.17%	**91.25%**	**92.06%**	**AUC**	
84.34%	**82.5%**	**84.13%**	**Gini Index**	
92.29%	**88.51%**	**89.61%**	**AGF**	

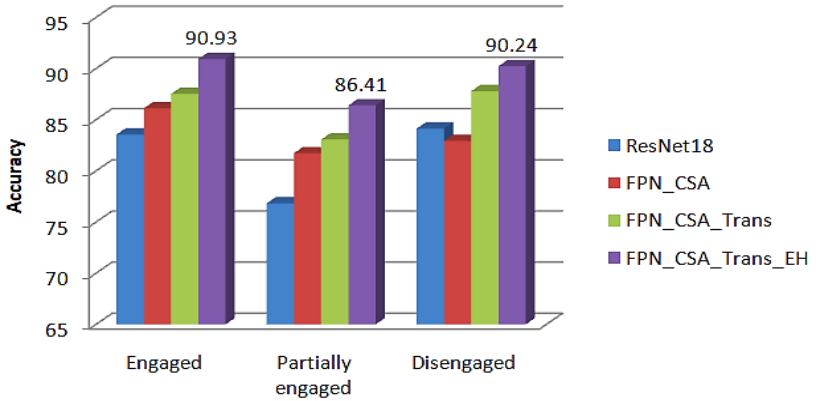

Fig. 9. Class-wise accuracy comparison of ResNet-18 and FPN_CSA_EH Trans on the WACV dataset

Class-wise AUC comparison between ResNet-18 and FPN_CSA_Trans_EH. In terms of AUC, FPN_CSA_Trans_EH outperforms ResNet-18 by 2.03% for the disengaged class, 6.95% for the partially engaged class, and 2.08% for the engaged class.

Class-wise Gini Index comparison between ResNet-18 and FPN_CSA_Trans_EH. In terms of the Gini Index, FPN_CSA_Trans_EH outperforms ResNet-18 by 4.06% for the disengaged class, 13.9% for the partially engaged class, and 4.16% for the engaged class.

Class-wise AGF comparison between ResNet-18 and FPN_CSA_Trans_EH. In terms of AGF, FPN_CSA_Trans_EH outperforms ResNet-18 by 2.53% for the disengaged class, 5.2% for the partially engaged class, and 3.8% for the engaged class.

5.2. DAiSEE Dataset. A graphical comparison of accuracy between the proposed method and existing methods on the DAiSEE dataset is provided in Table 3. In terms of accuracy, FPN_CSA_Trans_EH outperforms FPN_CSA_Trans by 2.19%, FPN_CSA by 4.55%, ResNet-18 [28] by 4.38%, Neural Turing Machine [32] by 9.72%, DERN [31] by 11.02%, DFSTN [30] by 12.22%, C3D+LSTM [29] by 14.42%, 13D [27] by 18.62%, and C3D [25] by 22.92%.

Table 3. Accuracy comparison of different methods on the DAiSEE dataset

DAiSEE	**Accuracy(%)**
C3D [25]	48.1
I3D [27]	52.4
C3D + LSTM [29]	56.6
DFSTN [30]	58.8
DERN [31]	60
Neural Turing Machine [32]	61.3
ResNet-18[28]	66.64
FPN_CSA	66.47
FPN_CSA_Trans	68.83
FPN_CSA_Trans_EH	**71.02**

Figure 10 illustrates the class-wise accuracy comparison between ResNet-18 and FPN_CSA_Trans_EH (Table 4). For the boredom class, ResNet-18 and FPN_CSA_Trans_EH achieve similar accuracy. FPN_CSA_Trans_EH outperforms ResNet-18 by 14.28% for the confusion class, 3.51% for the frustration class, and 4.3% for the engagement class.

Table 4. Comparison of metrics for ResNet-18 and FPN_CSA

Engagement	Frustration	Confusion	Boredom	Classes	
70.76%	61.45%	80.95%	75%	Accuracy	ResNet-18 [28]
70.07%	69.69%	88.09%	86.99%	AUC	
40.14%	39.38%	76.18%	73.98%	Gini Index	
71.32%	66.31%	82.98%	63.59%	AGF	
71.86%	60.09%	82.14%	50%	Accuracy	FPN_CSA
70.21%	69.29%	88.51%	87.19%	AUC	
40.42%	38.58%	77.02%	74.38%	Gini Index	
71.83%	65.5%	82.98%	70.65%	AGF	
72.72%	63.37%	86.9%	75%	Accuracy	FPN_CSA_Trans
71.67%	71.26%	91.33%	87.21%	AUC	
43.34%	42.53%	82.66%	74.43%	Gini Index	
73.03%	68.03%	86.49%	71.86%	AGF	
75.06%	**64.96%**	**95.23%**	**75%**	Accuracy	**FPN_CSA_Trans_EH**
73.61%	**72.83%**	**95.73%**	**87.24%**	AUC	
73.61%	**45.67%**	**91.47%**	**74.49%**	Gini Index	
75.1%	**69.5%**	**90.84%**	**73.13%**	AGF	

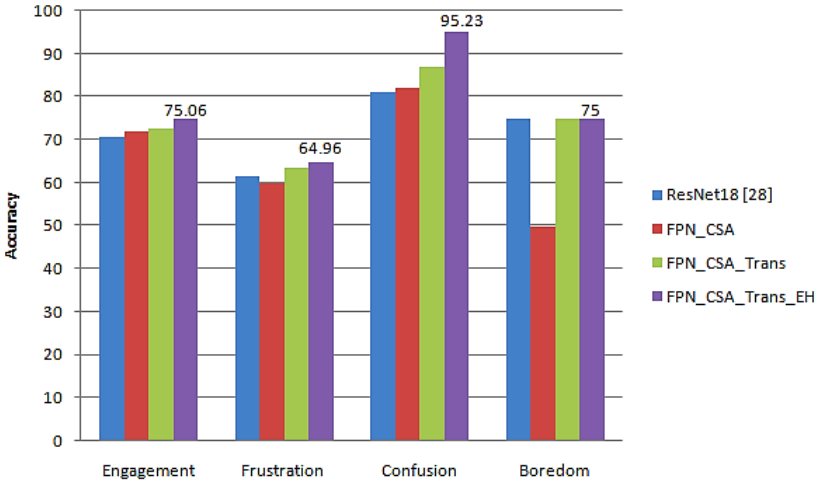

Fig. 10. Class-wise accuracy comparison of ResNet-18 and FPN_CSA on the DAiSEE dataset

Class-wise AUC comparison between ResNet-18 and FPN_CSA_Trans_EH. In terms of AUC, FPN_CSA_Trans_EH outperforms ResNet-18 by 0.25% for the boredom class, 7.64% for the

confusion class, 3.14% for the frustration class, and 3.54% for the engagement class.

Class-wise Gini Index comparison between ResNet-18 and FPN_CSA_Trans_EH. In terms of the Gini Index, FPN_CSA_Trans_EH outperforms ResNet-18 by 0.51% for the boredom class, 15.29% for the confusion class, 6.29% for the frustration class, and 2.29% for the engagement class.

Class-wise AGF comparison between ResNet-18 and FPN_CSA_Trans_EH. In terms of AGF, FPN_CSA_Trans_EH outperforms ResNet-18 by 9.54% for the boredom class, 7.86% for the confusion class, 3.19% for the frustration class, and 3.78% for the engagement class.

The performance of FPN_CSA_Trans_EH on the DAiSEE and WACV datasets underscores the model's strengths and identifies areas for improvement. While the model achieves exceptional performance on the WACV dataset, the challenges with the DAiSEE dataset offer valuable insights for further improvement.

5. Conclusion and Future Work. This study introduces FPN_CSA_Trans_EH, a novel framework for real-time identification of student engagement in educational environments. The proposed approach integrates Channel-Spatial Attention (CSA) with Transformer-enhanced Feature Pyramid Networks (FPN), offering a robust method for automatically detecting patterns of student engagement in visual data streams. By combining attention mechanisms with hierarchical feature representation, FPN_CSA_Trans_EH efficiently captures spatial and contextual information, enabling accurate determination of student engagement levels. The Transformer architecture enhances the model's ability to recognise long-range dependencies and semantic relationships within input sequences. The proposed framework outperforms baseline methods on the WACV dataset, demonstrating its potential for practical applications in educational settings. Future research will explore multiple avenues for enhancement and expansion. Initial efforts will focus on improving the model's efficiency and scalability to handle larger datasets and real-world implementation scenarios. Additionally, incorporating multimodal data sources, such as text and audio, could enhance the model's understanding of student interactions and behaviour. Finally, field tests and longitudinal studies will be conducted to evaluate the model's effectiveness in real educational settings and its impact on teaching and learning outcomes. Our goal is to advance student engagement detection techniques and contribute to the development of inclusive and effective educational technologies. Future work could extend this model by integrating audio modalities alongside visual features to predict student engagement in online settings.

References

1. Marks H.M. Student engagement in instructional activity: Patterns in the elementary, middle, and high school years. American Educational Research Journal. 2000. vol. 37. pp. 153–184. DOI: 10.3102/00028312037001153.
2. Nomura K., Iwata M., Augereau O., Kise K. Estimation of student's engagement based on the posture. Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the ACM International Symposium on Wearable Computers. ACM, 2019. pp. 164–167. DOI: 10.1145/3341162.3343767.
3. Kaur A., Mustafa A., Mehta L., Dhall A. Prediction and Localization of Student Engagement in the Wild. Digital Image Computing: Techniques and Applications (DICTA). Australia, Canberra: ACT, 2018. pp. 1–8. DOI: 10.1109/DICTA.2018.8615851.
4. Hatori Y., Nakajima T., Watabe S. Body Posture Analysis for the Classification of Classroom Scenes. Interdisciplinary Information Sciences. 2022. vol. 28(1). pp. 55–62. DOI: 10.4036/iis.2022.a.05.
5. Liu Y., Chen J., Zhang M., Rao C. Student engagement study based on multi-cue detection and recognition in an intelligent learning environment. Multimedia Tools Appl. 2018. vol. 77(21). pp. 28749–28775.
6. Sharma P., Joshi S., Gautam S., Maharjan S., Khanal S.R., Reis M.C., Barroso J., de Jesus Filipe V.M. Student Engagement Detection Using Emotion Analysis, Eye Tracking and Head Movement with Machine Learning. Technology and Innovation in Learning, Teaching and Education. TECH-EDU 2022. Communications in Computer and Information Science. vol. 1720. pp. 52–68. DOI: 10.1007/978-3-031-22918-3_5.
7. Bourel F., Chibelushi C. Low A. Recognition of Facial Expressions in the Presence of Occlusion. 2001. vol. 1. DOI: 10.5244/C.15.23.
8. Mao X., Xue Y., Li Z., Huang K., Lv S. Robust facial expression recognition based on RPCA and AdaBoost. 10th Workshop on Image Analysis for Multimedia Interactive Services. 2009. pp. 113–116. DOI: 10.1109/WIAMIS.2009.5031445.
9. Jiang B., Jia Kb. Research of Robust Facial Expression Recognition under Facial Occlusion Condition. Active Media Technology, 2011. pp. 92–100. DOI: 10.1007/978-3-642-23620-4_13.
10. Hammal Z., Arguin M., Gosselin F. Comparing a novel model based on the transferable belief model with humans during the recognition of partially occluded facial expressions. Journal of vision. 2009. vol. 9. DOI: 10.1167/9.2.22.
11. Zhang F., Zhang T., Mao Q., Xu C. Joint Pose and Expression Modeling for Facial Expression Recognition. IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018. pp. 3359–3368. DOI: 10.1109/CVPR.2018.00354.
12. Wang C., Wang S., Liang G. Identity- and Pose-Robust Facial Expression Recognition through Adversarial Feature Learning. 2019. pp. 238–246. DOI: 10.1145/3343031.3350872.
13. Moubayed A., Injadat M., Shami A., Lutfiyya H. Student Engagement Level in e-Learning Environment: Clustering Using K-means. American Journal of Distance Education. 2020. vol. 34(2). pp. 137–156. DOI: 10.1080/08923647.2020.1696140.
14. Gupta S., Kumar P., Tekchandani R. Facial emotion recognition based real-time learner engagement detection system in online learning context using deep learning models. Multimedia Tools and Applications. 2022. vol. 82. pp. 11365–11394. DOI: 10.1007/s11042-022-13558-9.
15. Bhardwaj P., Gupta P., Panwar H., Siddiqui M.K., Morales-Menendez R., Bhaik A. Application of Deep Learning on Student Engagement in e-learning environments. Computers & Electrical Engineering. 2021. vol. 93. DOI: 10.1016/j.compeleceng.2021.107277.

16. Fakhar S., Baber J., Bazai S., Marjan S., Jasiński M., Jasińska E., Chaudhry M.U., Leonowicz Z., Hussain S. Smart Classroom Monitoring Using Novel Real-Time Facial Expression Recognition System. Applied Sciences. 2022. vol. 12(23). DOI: 10.3390/app122312134.
17. Sümer Ö., Goldberg P., D'Mello S., Gerjets P., Trautwein U., Kasneci E., Multimodal Engagement Analysis From Facial Videos in the Classroom. IEEE Transactions on Affective Computing. 2023. vol. 14. no. 2. pp. 1012–1027. DOI: 10.1109/TAFFC.2021.3127692.
18. Psaltis A., Apostolakis K.C., Dimitropoulos K., Daras P. Multimodal Student Engagement Recognition in Prosocial Games. IEEE Transactions on Games. 2018. vol. 10. no. 3. pp. 292–303. DOI: 10.1109/TCIAIG.2017.2743341.
19. Mohamad Nezami O., Dras M., Hamey L., Richards D., Wan S., Paris C. Automatic Recognition of Student Engagement Using Deep Learning and Facial Expression. Lecture Notes in Computer Science. Springer, Cham. 2020. vol. 11908. pp. 273–289. DOI: 10.1007/978-3-030-46133-1_17.
20. Yu H., Gupta A., Lee W., Arroyo I., Betke M., Allesio D., Murray T., Magee J., Woolf B.P. Measuring and integrating facial expressions and head pose as indicators of engagement and affect in tutoring systems Adaptive Instructional Systems. Adaptation Strategies and Methods. Cham Springer. 2021. pp. 219–233.
21. He K., Zhang X., Ren S., Sun J. Deep Residual Learning for Image Recognition. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016. pp. 770–778. DOI: 10.1109/CVPR.2016.90.
22. Gao M., Song P., Wang F., Liu J., Mandelis A., Qi D. A Novel Deep Convolutional Neural Network Based on ResNet-18 and Transfer Learning for Detection of Wood Knot Defects. Journal of Sensors. 2021. pp. 1–16. DOI: 10.1155/2021/4428964.
23. Lin T.-Y., Dollár P., Girshick R., He K., Hariharan B., Belongie S. Feature Pyramid Networks for Object Detection. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017. pp. 936–944. DOI: 10.1109/CVPR.2017.106.
24. Woo S., Park J., Lee J.Y., Kweon I.S. CBAM: Convolutional Block Attention Module. Computer Vision – ECCV 2018. Lecture Notes in Computer Science. Springer, Cham. 2018. vol. 11211. pp. 3–19. DOI: 10.1007/978-3-030-01234-2_1.
25. Gupta A., DCunha A., Awasthi K., Balasubramanian V. DAiSEE: Towards User Engagement Recognition in the Wild. 2016. arXiv preprint: arXiv:1609.01885.
26. Islam M., Hossain E. Foreign Exchange Currency Rate Prediction using a GRU-LSTM Hybrid Network. Soft Computing Letters. 2021. vol. 3. DOI: 10.1016/j.socl.2020.100009.
27. Zhang H., Xiao X., Huang T., Liu S., Xia Y., Li J. An Novel End-to-end Network for Automatic Student Engagement Recognition. 2019 IEEE 9th International Conference on Electronics Information and Emergency Communication (ICEIEC). 2019. pp. 342–345. DOI: 10.1109/ICEIEC.2019.8784507.
28. Batra S., Wang S., Nag A., Brodeur P., Checkley M., Klinkert A., Dev S. DMCNet: Diversified Model Combination Network for Understanding Engagement from Video Screengrabs. 2022. arXiv preprint: arXiv: 2204.06454. DOI: 10.48550/arXiv.2204.06454.
29. Abedi A., Khan S.S. Improving state-of-the-art in Detecting Student Engagement with Resnet and TCN Hybrid Network. 18th Conference on Robots and Vision (CRV). 2021. pp. 151–157. DOI: 10.1109/CRV52889.2021.00028.
30. Liao J., Liang Y., Pan J. Deep facial spatiotemporal network for engagement prediction in online learning. Applied Intelligence. 2021. vol. 51. pp. 6609–6621. DOI: 10.1007/s10489-020-02139-8.
31. Huang T., Mei Y., Zhang H., Liu S., Yang H. Fine-grained Engagement Recognition in Online Learning Environment. IEEE 9th International Conference on Electronics

Information and Emergency Communication (ICEIEC). 2019. pp. 338–341. DOI: 10.1109/ICEIEC.2019.8784559.

32. Ma X., Xu M., Dong Y., Sun Z. Automatic Student Engagement in Online Learning Environment Based on Neural Turing Machine. International Journal of Information and Education Technology. 2021. vol. 11(3). pp. 107–111. DOI: 10.18178/ijiet.2021.11.3.1497.

Naveen A. — Researcher, Department of computer science and engineering, Gitam University-Bengaluru Campus. Research interests: computer science, engineering. The number of publications — 0. a.naveen21@gmail.com; 207, Nagadenehalli Doddaballapur, Taluk, 561203, Bengaluru, Karnataka, India; office phone: +91(80)2809-8000.

Jacob I. Jeena — Professor, Associate professor, Department of computer science and engineering, Gitam University-Bengaluru Campus. Research interests: computer vision, deep learning. The number of publications — 53. ijacob@gitam.edu; 207, Nagadenehalli Doddaballapur, Taluk, 561203, Bengaluru, Karnataka, India; office phone: +91(80)2809-8000.

Mandava Ajay Kumar — Associate professor, Department of electrical, electronics and communication engineering, Gitam University-Bengaluru Campus. Research interests: computer vision, pattern recognition, signal and image processing. The number of publications — 46. amandava@gitam.edu; 207, Nagadenehalli Doddaballapur, Taluk, 561203, Bengaluru, Karnataka, India; office phone: +91(80)2809-8000.

УДК 004 **DOI** 10.15622/ia.24.2.10

А. НАВИН, И. ДЖЕЙКОБ, А. МАНДАВА

ОПРЕДЕЛЕНИЕ ВОВЛЕЧЕННОСТИ УЧАЩИХСЯ С ПОМОЩЬЮ СЕТЕЙ ПИРАМИДАЛЬНЫХ ПРИЗНАКОВ, УЛУЧШЕННЫХ ТРАНСФОРМЕРОМ, С КАНАЛЬНО-ПРОСТРАНСТВЕННЫМ ВНИМАНИЕМ

Навин А., Джейкоб И., Мандава А. **Определение вовлеченности учащихся с помощью сетей пирамидальных признаков, улучшенных трансформером, с канально-пространственным вниманием.**

Аннотация. Одним из важнейших аспектов современных образовательных систем является определение вовлеченности учащихся, которое включает выявление того, насколько вовлечены, внимательны и активны учащиеся на занятиях в классе. Для преподавателей этот подход имеет важное значение, поскольку он дает представление об опыте обучения учащихся, позволяя адаптировать подходы в обучении и улучшать качество обучения. Традиционные методы оценки вовлеченности учащихся часто являются трудоемкими и субъективными. В этом исследовании предлагается новая система определения степени вовлеченности учащихся в реальном времени, которая использует сети пирамидальных признаков (FPN), улучшенные с помощью архитектуры Трансформера, с канально-пространственным вниманием (CSA), называемая BiusFPN_CSA. Предлагаемый подход автоматически анализирует модели вовлеченности учащихся, такие как поза тела, зрительный контакт и положение головы, из визуальных потоков данных путем интеграции передовых методов глубокого обучения и компьютерного зрения. За счет интеграции механизма внимания CSA с возможностями иерархического представления признаков FPN, модель может точно определять уровни вовлеченности учащихся, улавливая контекстную и пространственную информацию во входных данных. Кроме того, благодаря внедрению архитектуры Трансформера, модель достигает лучшей общей производительности за счет эффективного учета долгосрочных зависимостей и семантических связей во входных последовательностях. Оценка с использованием набора данных WACV показывает, что предлагаемая модель превосходит базовые методы с точки зрения точности. В частности, вариант FPN_CSA_Trans_EH предлагаемой модели превосходит FPN_CSA на 3,28% и 4,98% соответственно. Эти результаты подчеркивают эффективность структуры BiusFPN_CSA в определении вовлеченности учащихся в реальном времени, предлагая преподавателям ценный инструмент для повышения качества обучения, создания активной среды обучения и, в конечном итоге, улучшения результатов учащихся.

Ключевые слова: сеть пирамидальных признаков (FPN), канально-пространственное внимание (CSA), определение вовлеченности учащихся, трансформер.

Литература

1. Marks H.M. Student engagement in instructional activity: Patterns in the elementary, middle, and high school years. American Educational Research Journal. 2000. vol. 37. pp. 153–184. DOI: 10.3102/00028312037001153.
2. Nomura K., Iwata M., Augereau O., Kise K. Estimation of student's engagement based on the posture. Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the ACM International Symposium on Wearable Computers. ACM, 2019. pp. 164–167. DOI: 10.1145/3341162.3343767.

3. Kaur A., Mustafa A., Mehta L., Dhall A. Prediction and Localization of Student Engagement in the Wild. Digital Image Computing: Techniques and Applications (DICTA). Australia, Canberra: ACT, 2018. pp. 1–8. DOI: 10.1109/DICTA.2018.8615851.
4. Hatori Y., Nakajima T., Watabe S. Body Posture Analysis for the Classification of Classroom Scenes. Interdisciplinary Information Sciences. 2022. vol. 28(1). pp. 55–62. DOI: 10.4036/iis.2022.a.05.
5. Liu Y., Chen J., Zhang M., Rao C. Student engagement study based on multi-cue detection and recognition in an intelligent learning environment. Multimedia Tools Appl. 2018. vol. 77(21). pp. 28749–28775.
6. Sharma P., Joshi S., Gautam S., Maharjan S., Khanal S.R., Reis M.C., Barroso J., de Jesus Filipe V.M. Student Engagement Detection Using Emotion Analysis, Eye Tracking and Head Movement with Machine Learning. Technology and Innovation in Learning, Teaching and Education. TECH-EDU 2022. Communications in Computer and Information Science. vol. 1720. pp. 52–68. DOI: 10.1007/978-3-031-22918-3_5.
7. Bourel F., Chibelushi C. Low A. Recognition of Facial Expressions in the Presence of Occlusion. 2001. vol. 1. DOI: 10.5244/C.15.23.
8. Mao X., Xue Y., Li Z., Huang K., Lv S. Robust facial expression recognition based on RPCA and AdaBoost. 10th Workshop on Image Analysis for Multimedia Interactive Services. 2009. pp. 113–116. DOI: 10.1109/WIAMIS.2009.5031445.
9. Jiang B., Jia Kb. Research of Robust Facial Expression Recognition under Facial Occlusion Condition. Active Media Technology, 2011. pp. 92–100. DOI: 10.1007/978-3-642-23620-4_13.
10. Hammal Z., Arguin M., Gosselin F. Comparing a novel model based on the transferable belief model with humans during the recognition of partially occluded facial expressions. Journal of vision. 2009. vol. 9. DOI: 10.1167/9.2.22.
11. Zhang F., Zhang T., Mao Q., Xu C. Joint Pose and Expression Modeling for Facial Expression Recognition. IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018. pp. 3359–3368. DOI: 10.1109/CVPR.2018.00354.
12. Wang C., Wang S., Liang G. Identity- and Pose-Robust Facial Expression Recognition through Adversarial Feature Learning. 2019. pp. 238–246. DOI: 10.1145/3343031.3350872.
13. Moubayed A., Injadat M., Shami A., Lutfiyya H. Student Engagement Level in e-Learning Environment: Clustering Using K-means. American Journal of Distance Education. 2020. vol. 34(2). pp. 137–156. DOI: 10.1080/08923647.2020.1696140.
14. Gupta S., Kumar P., Tekchandani R. Facial emotion recognition based real-time learner engagement detection system in online learning context using deep learning models. Multimedia Tools and Applications. 2022. vol. 82. pp. 11365–11394. DOI: 10.1007/s11042-022-13558-9.
15. Bhardwaj P., Gupta P., Panwar H., Siddiqui M.K., Morales-Menendez R., Bhaik A. Application of Deep Learning on Student Engagement in e-learning environments. Computers & Electrical Engineering. 2021. vol. 93. DOI: 10.1016/j.compeleceng.2021.107277.
16. Fakhar S., Baber J., Bazai S., Marjan S., Jasiński M., Jasińska E., Chaudhry M.U., Leonowicz Z., Hussain S. Smart Classroom Monitoring Using Novel Real-Time Facial Expression Recognition System. Applied Sciences. 2022. vol. 12(23). DOI: 10.3390/app122312134.
17. Sümer Ö., Goldberg P., D'Mello S., Gerjets P., Trautwein U., Kasneci E., Multimodal Engagement Analysis From Facial Videos in the Classroom. IEEE Transactions on Affective Computing. 2023. vol. 14. no. 2. pp. 1012–1027. DOI: 10.1109/TAFFC.2021.3127692.

18. Psaltis A., Apostolakis K.C., Dimitropoulos K., Daras P. Multimodal Student Engagement Recognition in Prosocial Games. IEEE Transactions on Games. 2018. vol. 10. no. 3. pp. 292–303. DOI: 10.1109/TCIAIG.2017.2743341.
19. Mohamad Nezami O., Dras M., Hamey L., Richards D., Wan S., Paris C. Automatic Recognition of Student Engagement Using Deep Learning and Facial Expression. Lecture Notes in Computer Science. Springer, Cham. 2020. vol. 11908. pp. 273–289. DOI: 10.1007/978-3-030-46133-1_17.
20. Yu H., Gupta A., Lee W., Arroyo I., Betke M., Allesio D., Murray T., Magee J., Woolf B.P. Measuring and integrating facial expressions and head pose as indicators of engagement and affect in tutoring systems Adaptive Instructional Systems. Adaptation Strategies and Methods. Cham Springer. 2021. pp. 219–233.
21. He K., Zhang X., Ren S., Sun J. Deep Residual Learning for Image Recognition. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016. pp. 770–778. DOI: 10.1109/CVPR.2016.90.
22. Gao M., Song P., Wang F., Liu J., Mandelis A., Qi D. A Novel Deep Convolutional Neural Network Based on ResNet-18 and Transfer Learning for Detection of Wood Knot Defects. Journal of Sensors. 2021. pp. 1–16. DOI: 10.1155/2021/4428964.
23. Lin T.-Y., Dollár P., Girshick R., He K., Hariharan B., Belongie S. Feature Pyramid Networks for Object Detection. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017. pp. 936–944. DOI: 10.1109/CVPR.2017.106.
24. Woo S., Park J., Lee J.Y., Kweon I.S. CBAM: Convolutional Block Attention Module. Computer Vision – ECCV 2018. Lecture Notes in Computer Science. Springer, Cham. 2018. vol. 11211. pp. 3–19. DOI: 10.1007/978-3-030-01234-2_1.
25. Gupta A., DCunha A., Awasthi K., Balasubramanian V. DAiSEE: Towards User Engagement Recognition in the Wild. 2016. arXiv preprint: arXiv:1609.01885.
26. Islam M., Hossain E. Foreign Exchange Currency Rate Prediction using a GRU-LSTM Hybrid Network. Soft Computing Letters. 2021. vol. 3. DOI: 10.1016/j.socl.2020.100009.
27. Zhang H., Xiao X., Huang T., Liu S., Xia Y., Li J. An Novel End-to-end Network for Automatic Student Engagement Recognition. 2019 IEEE 9th International Conference on Electronics Information and Emergency Communication (ICEIEC). 2019. pp. 342–345. DOI: 10.1109/ICEIEC.2019.8784507.
28. Batra S., Wang S., Nag A., Brodeur P., Checkley M., Klinkert A., Dev S. DMCNet: Diversified Model Combination Network for Understanding Engagement from Video Screengrabs. 2022. arXiv preprint: arXiv: 2204.06454. DOI: 10.48550/arXiv.2204.06454.
29. Abedi A., Khan S.S. Improving state-of-the-art in Detecting Student Engagement with Resnet and TCN Hybrid Network. 18th Conference on Robots and Vision (CRV). 2021. pp. 151–157. DOI: 10.1109/CRV52889.2021.00028.
30. Liao J., Liang Y., Pan J. Deep facial spatiotemporal network for engagement prediction in online learning. Applied Intelligence. 2021. vol. 51. pp. 6609–6621. DOI: 10.1007/s10489-020-02139-8.
31. Huang T., Mei Y., Zhang H., Liu S., Yang H. Fine-grained Engagement Recognition in Online Learning Environment. IEEE 9th International Conference on Electronics Information and Emergency Communication (ICEIEC). 2019. pp. 338–341. DOI: 10.1109/ICEIEC.2019.8784559.
32. Ma X., Xu M., Dong Y., Sun Z. Automatic Student Engagement in Online Learning Environment Based on Neural Turing Machine. International Journal of Information and Education Technology. 2021. vol. 11(3). pp. 107–111. DOI: 10.18178/ijiet.2021.11.3.1497.

Навин А. — научный сотрудник, факультет компьютерных наук и инженерии, Университет Гитам – кампус в Бангалоре. Область научных интересов: компьютерные науки, инженерия. Число научных публикаций — 0. a.naveen21@gmail.com; Нагаденехалли Доддабаллапур, Талук, 207, 561203, Бангалор, Карнатака, Индия; р.т.: +91(80)2809-8000.

Джейкоб И. Джина — профессор, доцент, факультет компьютерных наук и инженерии, Университет Гитам – кампус в Бангалоре. Область научных интересов: компьютерное зрение, глубокое обучение. Число научных публикаций — 53. ijacob@gitam.edu; Нагаденехалли Доддабаллапур, Талук, 207, 561203, Бангалор, Карнатака, Индия; р.т.: +91(80)2809-8000.

Мандава Аджай Кумар — доцент, кафедра электротехники, электроники и средств связи, Университет Гитам – кампус в Бангалоре. Область научных интересов: компьютерное зрение, распознавание образов, обработка сигналов и изображений. Число научных публикаций — 46. amandava@gitam.edu; Нагаденехалли Доддабаллапур, Талук, 207, 561203, Бангалор, Карнатака, Индия; р.т.: +91(80)2809-8000.