Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.054982

ARTICLE

Re-Distributing Facial Features for Engagement Prediction with ModernTCN

Xi Li[1,2], Weiwei Zhu[2], Qian Li[3,*], Changhui Hou[1,*] and Yaozong Zhang[1]

[1]College of Information and Artificial Intelligence, Nanchang Institute of Science and Technology, Nanchang, 330108, China

[2]School of Electrical and Information Engineering, Wuhan Institute of Technology, Wuhan, 430205, China

[3]School of Electronic Information Engineering, Wuhan Donghu University, Wuhan, 430212, China

*Corresponding Authors: Qian Li. Email: liqian@wdu.edu.cn; Changhui Hou. Email: houchanghuinist@163.com

Received: 13 June 2024 Accepted: 22 August 2024 Published: 15 October 2024

ABSTRACT

Automatically detecting learners' engagement levels helps to develop more effective online teaching and assessment programs, allowing teachers to provide timely feedback and make personalized adjustments based on students' needs to enhance teaching effectiveness. Traditional approaches mainly rely on single-frame multimodal facial spatial information, neglecting temporal emotional and behavioural features, with accuracy affected by significant pose variations. Additionally, convolutional padding can erode feature maps, affecting feature extraction's representational capacity. To address these issues, we propose a hybrid neural network architecture, the redistributing facial features and temporal convolutional network (RefEIP). This network consists of three key components: first, utilizing the spatial attention mechanism large kernel attention (LKA) to automatically capture local patches and mitigate the effects of pose variations; second, employing the feature organization and weight distribution (FOWD) module to redistribute feature weights and eliminate the impact of white features and enhancing representation in facial feature maps. Finally, we analyse the temporal changes in video frames through the modern temporal convolutional network (ModernTCN) module to detect engagement levels. We constructed a near-infrared engagement video dataset (NEVD) to better validate the efficiency of the RefEIP network. Through extensive experiments and in-depth studies, we evaluated these methods on the NEVD and the Database for Affect in Situations of Elicitation (DAiSEE), achieving an accuracy of 90.8% on NEVD and 61.2% on DAiSEE in the four-class classification task, indicating significant advantages in addressing engagement video analysis problems.

KEYWORDS

Engagement prediction; spatiotemporal network; re-distributing facial features; temporal convolutional network

1 Introduction

Given the rise of the internet in recent years, online education has become a hot topic in the field of education. While schools have successfully transitioned from traditional face-to-face teaching to online teaching, teachers are now facing a new challenge: assessing student engagement in the courses they teach is difficult, a problem that is exacerbated by large numbers of students. The development of automated engagement predictions is critical for teachers to quantify the quality of online education and improve learning effectiveness.

The automatic detection of student engagement involves various modalities, including images, videos, audio, and electrocardiograms (ECGs) [1–3] video cameras and webcams are commonly used to capture students' behaviour in learning environments, and they are extensively employed to assess student engagement in online classrooms. Despite significant recent advancements in attention recognition, this task remains challenging for two primary reasons: (1) the complexity of data collection and annotation and (2) the spatiotemporal complexity of engagement prediction. Regarding the former, accurately measuring attention typically requires high-quality annotated data, which is difficult to achieve on a large scale in educational settings because the annotation process is both time-consuming and prone to errors. Regarding the latter, engagement encompasses not only the state at a specific moment but also the dynamic changes throughout the learning process, necessitating the analysis of both static information in individual video frames and dynamic variations across consecutive frames.

Computer vision-based methods for detecting student engagement can be categorized into two types: image-based methods and video-based methods. The former relies solely on extracting multimodal features from a single image to assess engagement. However, this approach is limited to spatial information from a single frame, making it susceptible to variations in lighting, pose, and facial expressions, which can decrease accuracy and fail to fully capture students' spatiotemporal emotional behaviours [4]. To address these limitations, numerous video-based methods have been proposed for predicting engagement [5]. Although these methods require the annotation of each segment, the annotation workload is smaller than that of image-based methods. Nevertheless, several challenges remain: (1) When the face is affected by challenging factors, such as significant lighting changes, large pose variations, and severe occlusions, the extraction of facial feature key points may become inaccurate. (2) Insufficient samples, difficulties in obtaining annotated data, and inadequate generalization capabilities limit the accuracy and reliability of video-based engagement detection. (3) The models are typically large, demanding high computational and storage resources, which makes them unsuitable for real-time online prediction.

Additionally, we observed some inherent characteristics of convolutional neural networks that limit their performance and application to some extent. Convolutional kernels have a weaker perception of image edges and corners because their operation tends to focus more on the central region of the image, undervaluing the importance of edge information. This phenomenon is known as perception bias. Convolutional neural networks typically use padding to balance the perception frequency of each pixel by adding zero values to the periphery of the image. However, as the network depth increases, the number of channels in the feature map grows exponentially, while the size of the feature map continues to shrink. Edge pixels occupy a larger proportion of the feature map, which leads to sparse and refined pixel information, making it more difficult for deeper convolutional layers to capture edge information. To address this issue, networks often continue to use padding to ensure the capture of critical information, but this process can result in the so-called padding trap. Excessive padding can cause severe information distortion, affecting the multiplication and summation operations between each area's pixel values and the kernel pixel values. As shown in Fig. 1, the severely eroded pixels are primarily concentrated in the edge regions of the feature map, with the peripheral areas becoming particularly prominent due to information blurring. As padding gradually moves inwards, each layer's padding operation progressively erodes the feature map, with untreated white features directly affecting the representational capability of the subsequent layers.

To achieve the aforementioned goals, Lu et al. [6] combined multiple behavioural features, such as head pose, gaze direction, and blink rate, with facial appearance proposed to predict engagement. Although this method improved the prediction accuracy, it also resulted in high computational complexity and long processing times. Liao et al. [7] employed an attention mechanism, however, it may overreact to similar facial regions while neglecting other distinguishing features that might be important for engagement recognition. This issue is particularly severe when the face is occluded or exhibits significant pose variations, as some facial features may become invisible. Class Attention in Video Transformer (CavT) [8] predicts engagement intensity through self-attention between patches and class attention between class tokens and patches. It uses a second-order representative sampling method (BorS) to add multiple video sequences for each video, enhancing the training set and addressing the sample insufficiency problem. However, this network has high redundancy, difficult hyperparameter tuning, sensitivity to input sequence length, and high computational resource requirements.

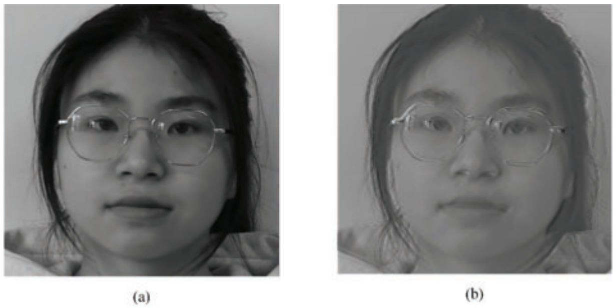

Figure 1: (a) shows the original face image, while (b) shows the appearance after one fill convolution. The edges of the face are blurred or eroded by the filler

In response to the aforementioned issues, this paper proposes a novel end-to-end residual network and temporal convolutional network hybrid neural network architecture—RefEIP—to predict engagement. As shown in Fig. 2, first, for a given attention video dataset, the Multitask Convolutional Neural Network (MTCNN) [9] is used to extract facial images, in the spatial module, we use the initial part of the Residual Network (ResNet-18) [10] to extract features from each image, up to the global average pooling (GAP) layer. Next, we introduce the redistributing facial features module, which consists of three parts. The large kernel attention (LKA) [11] spatial attention mechanism is used to automatically capture local patches, focusing not only on the most discriminative local patches but also on adaptively exploring global local patches. To address the feature erosion issue, we propose the feature organization and weight distribution (FOWD) module, which organizes the original feature maps into a single channel and efficiently increases efficiency by cleverly reducing the weight of the white features. Inspired by the nature of engagement intensity, we delved into the changes in facial features and found a certain correlation between adjacent frames. We categorize the extracted features into common features and unique features, simplifying feature extraction. High-quality feature vectors extracted from consecutive frames are regarded as multidimensional inputs extending the continuous time steps of the ModernTCN [12] module to simulate temporal information in the video. Finally, a projection layer activated by softmax maps the final representation to the ultimate classification result.

Furthermore, existing datasets for attention detection often focus on the visible light spectrum, making them susceptible to changes in lighting conditions, such as dim or completely dark environments. In contrast, near-infrared (NIR) imaging technology provides a more reliable solution because its imaging is less sensitive to changes in lighting conditions. To further validate the performance of the RefEIP architecture, we collected a dataset called the Near Infrared Engagement Video Dataset (NEVD). We evaluated the performance of the RefEIP architecture on both the NEVD and the publicly available DAiSEE [13]. Our method achieved a state-of-the-art detection accuracy of 90.8% on the NEVD and 61.2% on the DAiSEE, outperforming most other competing methods.

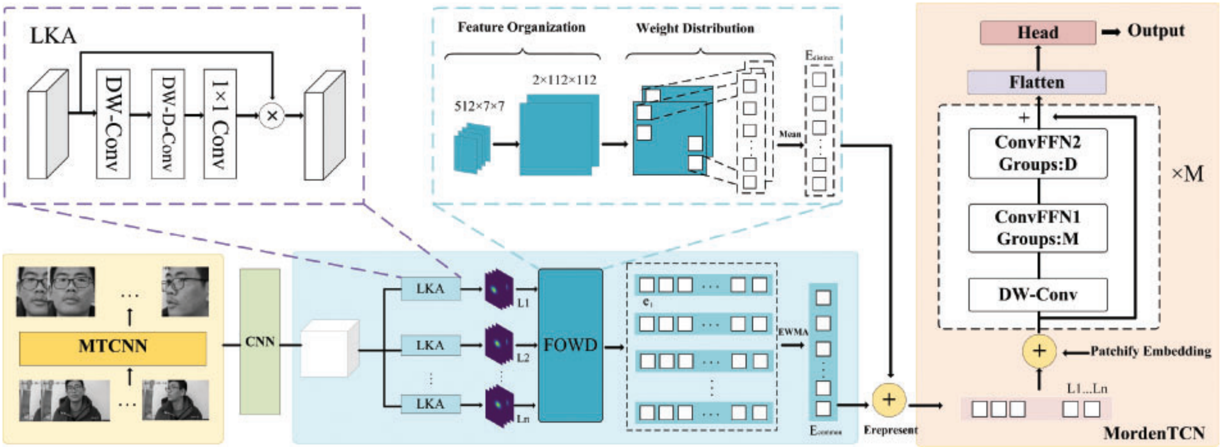

Figure 2: The overall architecture of our RefEIP model. The RefEIP model takes a sequence of L raw video frames as input and outputs an ordinal class that indicates the student's engagement level in the video. MTCNN extracts facial images from the video dataset, and CNN is used to extract features from each image. We apply the LKA mechanism to capture key local regions and use the FOWD module to consolidate feature maps into a single channel, enhancing efficiency by reducing the weight of redundant features. Extracted features, categorized into E_{common} and $E_{distinct}$, are fed into the ModernTCN module to model temporal information from consecutive frames

The main contributions of this paper are as follows:

1: We propose a new end-to-end engagement recognition framework called RefEIP. Compared to existing methods, RefEIP can integrate facial spatiotemporal information, which is beneficial for capturing detailed engagement states and enhancing the accuracy of engagement prediction.

2: We introduce the redistributing facial features method, which focuses on effective features. This method captures local patches to eliminate white erosion present in padding convolutions, protecting representations from the influence of white features and optimizing facial features to enhance representation quality.

3: To better evaluate the effectiveness of the proposed methods, we collected the NEVD. To our knowledge, existing publicly available engagement evaluation datasets mainly focus on the visible light domain. The NEVD is the first near-infrared video dataset for evaluating engagement, validating the effectiveness of our methods.

2 Related Work

In recent years, computer vision has played a significant role in automated measurement [14], particularly in student engagement detection [15,16]. Researchers increasingly employ computer vision and deep learning techniques to enhance the accuracy and efficiency of detection. For touch detection problems based on computer vision, researchers have explored two main methods: feature-based models and end-to-end models. Feature-based models first extract specific visual features from images or videos and then use these features to infer student engagement. On the other hand, end-to-end models learn directly from raw data and output engagement predictions without the need for manual feature extraction. Research on these methods aims to improve the accuracy and efficiency of engagement detection, providing more automated solutions for the education sector.

In feature-based engagement measurement methods, multimodal manual features are commonly extracted from single-frame videos, such as head posture, eye gaze angle, and facial expressions [17–19], and then input into a classifier to detect the engagement level in the image. Whitehill et al. first utilized the Cognitive and Emotional Response Tracker (CERT) to detect students' facial expressions during learning activities and then relied on students' head pose and basic facial movements to detect their level of engagement in individual images [18]. Chun et al. [20] utilized Haar-like feature detection and the Adaptive Boosting (AdaBoost) algorithm to learn facial features, employing optical flow techniques and template-matching methods to track facial features. Wu et al. [21] proposed a feature-based student engagement detection method in the EmotiW dataset [22], which involves extracting facial and upper body features from videos and utilizing long short-term memory (LSTM) and gated recurrent units (GRUs) combined for feature classification to detect student engagement. Wang et al. [23] introduced a novel convolutional neural network (CNN) architecture that extracts various facial expression features using facial landmarks and applied them to the DAiSEE [13] to predict engagement levels. However, during the feature extraction process, valuable information may be lost, potentially leading to the absence of key features and affecting the robustness of the model. Some existing methods use ensemble models to integrate different features, but this approach is effective only when the base learners perform well and diversity is sufficient.

The former method often focuses on the spatial information of single-frame images, while engagement is often a spatiotemporal emotional behaviour. End-to-end methods utilize video data to assess students' engagement, which is advantageous compared to methods based on single-frame images because they can capture spatiotemporal features of engagement behaviour. Compared to frame-based methods, this approach needs fewer annotations because each video segment only needs one label. Nevertheless, this method may face greater challenges in dealing with classification problems because annotations are less precise. Furthermore, most datasets used for engagement assessment are either small in scale or not publicly available, leading to difficulties in comparing different algorithms and constraining progress in the field [24]. To address this issue, Gupta et al. proposed the DAiSEE, which is also the largest publicly available dataset for engagement classification and evaluated various end-to-end convolutional video classification techniques, such as InceptionNet, three-dimensional convolutional neural network (C3D), and long-term recurrent convolutional networks [25–27]. The accuracy rates were 46.4%, 56.1%, and 57.9%, respectively. Geng et al. [28] proposed an automatic engagement recognition method based on a C3D, which introduced focal loss. This method achieved an accuracy of 56.2% on the DAiSEE. Zhang et al. [29] proposed an improved inflated 3D convolutional network (I3D) model; specifically, the I3D [26] model merged highly inattentive with inattentive and attentive with highly attentive, achieving an accuracy of 98.82% in binary engagement classification on the DAiSEE. Modelling based on image recognition and video classification is transitioning from CNNs to transformers. Wu et al. [8] introduced CavT to construct a video transformer for predicting engagement intensity using a second-order representation sampling method to address sample scarcity. Huang et al. [30] presented a deep engagement recognition network (DERN), which initially captures multimodal features from faces using the OpenFace library [31] and combines them for classification using temporal convolution, bidirectional long short-term memory [32], and attention mechanisms, achieving 60% detection accuracy on the DAiSEE. However, this network has a large parameter size and is not suitable for practical online applications. Liang et al. [33] proposed online engagement detection based on spatiotemporal attention mechanisms, which consisted of pre-trained ResNet-18 and LSTM, with Shuffle Attention and Global Attention incorporated into spatial and temporal modules, achieving an accuracy of 60.2% on the DAiSEE.

Although these methods demonstrate excellent performance on individual datasets, they are confined to visible light datasets, making their robustness difficult to ascertain. Furthermore, compared to other video datasets, the number and scale of datasets used for engagement prediction remain limited.

3 Methods

Fig. 2 illustrates the structure of the RefEIP hybrid neural network architecture used to classify student engagement in videos. This approach is end-to-end, meaning that features do not need to be manually extracted from videos or frames; instead, features are dynamically learned during network training. The network takes the raw frame sequence of the video as input, with the model's input being a tensor, where L, C, H, and W correspond to the number of frames, channels, frame height, and frame width, respectively, and outputs ordinal categories corresponding to the student's engagement level in the video.

3.1 Multitask Convolutional Neural Network

First, we preprocess the faces in the input video. Facial features are the most expressive and reflect the key areas of individual emotional states. These consecutive frames often contain a large amount of background information, which not only increases the computational burden on the network but also may affect the accurate extraction of facial features. To extract the facial parts in the video frames, we employ a multitask convolutional neural network (MTCNN) for face detection. The MTCNN consists of three cascaded lightweight CNN models: the proposal network (P-Net), the refine network (R-Net), and the output network (O-Net). The P-Net is responsible for extracting features from the image pyramid and determining candidate bounding boxes. The R-Net is used to filter the candidate bounding boxes generated by the P-Net, while the O-Net further refines the candidate boxes and locates facial key points. The entire MTCNN model progressively extracts and optimizes candidate boxes through the cascade of these three networks. We can extract regions containing only the facial parts from the original image, obtaining high-quality facial images of size C × 112 × 112 for subsequent steps.

3.2 Redistributing Facial Features Module

3.2.1 Large Kernel Attention Mechanism

As shown in Fig. 2, for a given facial image, we utilize convolutional neural networks (CNNs) to extract feature mappings. Subsequently, we employ multiple spatial attention mechanisms to automatically capture local patches. However, without proper guidance, ensuring comprehensive recognition of facial regions cannot be guaranteed, especially for faces with significant pose variations or strong occlusions, leading to a decline in attention recognition performance. Traditional attention mechanisms treat images as one-dimensional sequences, ignoring their two-dimensional structure and only achieving spatial adaptability while neglecting channel adaptability. To mitigate these issues, we employed large kernel attention (LKA) [11], which combines the advantages of convolution and self-attention mechanisms, allowing for better adaptation to local contextual information and long-term dependencies [34]. As shown in Fig. 3, LKA convolution is divided into three parts: spatial local convolution (depthwise convolution), spatial remote convolution (depthwise dilated convolution), and channel convolution (1 × 1 convolution). A $b \times b$ convolution is replaced with a $(b/d) \times (b/d)$

dilated convolution, where d denotes the dilation rate, followed by another convolution of size $(2d - 1) \times (2d - 1)$ to further process the captured long-range dependency information. Finally, a 1×1 convolution is applied for the ultimate feature integration. With the above decomposition, we can capture long-range relationships at a relatively small computational cost and parameter count. The module can be written as follows:

$$Attention = Conv_{1\times 1}\left(DW\text{-}D\text{-}Conv\left(DW\text{-}Conv\left(DW\text{-}Conv\left(F\right)\right)\right)\right) \tag{1}$$

$$Output = Attention \otimes F \tag{2}$$

where F represents the input features $F \in \mathbb{R}^{C\times H\times W}$, $Attention \in \mathbb{R}^{C\times H\times W}$ represents the attention map, where different values denote the importance of different features, and $\otimes$ is the elementwise product. LKA integrates convolution with a self-attention mechanism, considering local contextual information, large receptive fields, linear complexity, and dynamic processes. Additionally, LKA adopts adaptiveness in both spatial and channel dimensions. In deep neural networks, each channel typically corresponds to different features or objects, and adaptiveness in the channel dimension plays a crucial role in visual tasks [35,36].

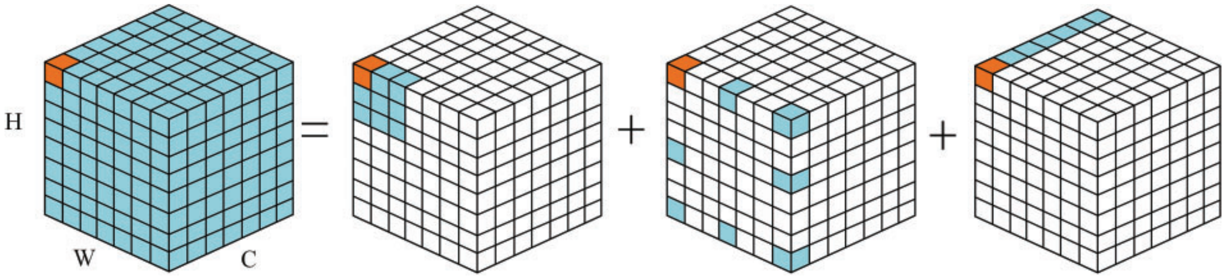

Figure 3: The decomposition diagram of the large nuclear convolution. Standard convolution operations can be broken down into three parts: deep convolution (DW-Conv), deep extended convolution (DW-D-Conv), and point convolution (1 × 1 Conv). In the diagram, the blue grid represents the location of the convolution kernel, while the yellow grid represents the centre point. The diagram demonstrates how a 13 × 13 convolution is broken down into a 5 × 5 depth-wise convolution, a 5 × 5 dilated depth-wise convolution with a dilation rate of 3, and a pointwise convolution

3.2.2 Feature Organization and Weight Distribution Module

Due to the nature of convolution, padding ensures the size of feature maps, alleviating information loss and improving network performance. However, excessive padding introduces too much irrelevant information to the image, and conventional pooling layers cannot mitigate this impact. To address this issue, we propose the FOWD module. In this section, we will discuss the construction and functionality of this module.

The module consists of two parts: feature organization (FO) and weight distribution (WD). The FO module serves as an auxiliary amplification function, amplifying the feature before distributing its weights through the WD module's convolution. The greater the padding in the previous convolution process is, the lower the weight. Padding during the convolution process leads to the formation of white features. By observing the padding locations, we can infer that edges and corners are the main areas where white features aggregate. The white pixels are evenly distributed in each channel, which is not conducive to precise information processing. Slightly expanding the white range may damage critical information. Therefore, we need a method to aggregate white pixels to the same extent while maintaining their relative positional distribution, concentrating the most eroded pixels at the periphery of the feature map and reconstructing the previously extracted feature map. The specific implementation process is shown in Fig. 2. Additionally, we weaken the weight of edge-whitening information by using unpadded blocks to counteract the adverse effects of padding operations. We chose the method mentioned in the effective subpixel convolutional layer to reconstruct the feature map to meet this requirement [37]. The specific implementation is shown in Fig. 4. In this process, only the absolute position of the feature points is changed, but the relative position remains unchanged. The channel is scaled down to 4, which corresponds to the original feature point position and the point position in the feature cluster.

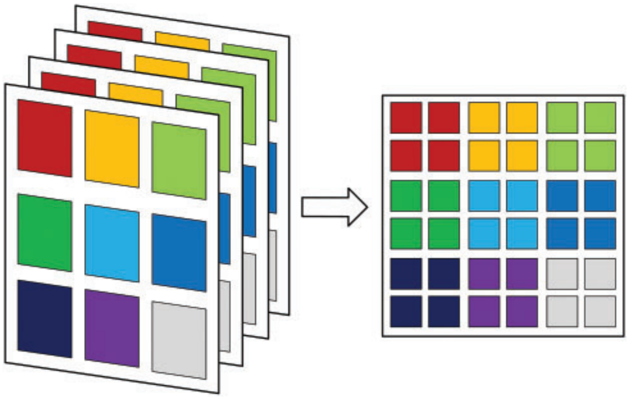

Figure 4: Example of our feature map rearrangement method

To counteract the impact of whitening features on the network, we designed a dedicated convolutional layer aimed at mitigating the influence of peripheral pixels. As illustrated in Fig. 4, the most severely eroded points are concentrated at the edges of the rearranged feature maps. We ingeniously utilized perceptual bias, as depicted in Fig. 5, this diagram shows a refill-free convolution over a pixel image. The convolution kernel has a size of 3 × 3 and a step size of 1. The depth of the colour on the pixel indicates the perceived frequency, and the darker the colour is the greater the perceived frequency. The convolutional kernel has a lower perceptual frequency for edge pixels than for internal pixels. Smaller convolutional kernels tend to extract information more concentratedly, which is disadvantageous to the network. Given the large volume of channel information entering the space, we opt for convolutional kernels several times larger than conventional kernels, enabling them to effectively ignore more information, which is more advantageous to the network. By adjusting the size and stride of the convolutional kernel, we can filter out white features while retaining crucial parts.

3.2.3 Facial Affinity Enhancement

Convolutional neural networks tend to assign higher weights to features that are beneficial for achieving evaluation metrics, which may lead to overly narrow learning of the network. However, considering changes in facial features is essential in the task of attention detection, especially considering the variation in facial features. Yang et al. argued that the face is a combination of expressive and neutral components [38]. When facing the task of facial attention detection, the relative changes between adjacent frames are small. Therefore, we divide facial features into two major categories:

common features and distinct features.

$$E_{represent} = E_{common} + E_{distinct} \tag{3}$$

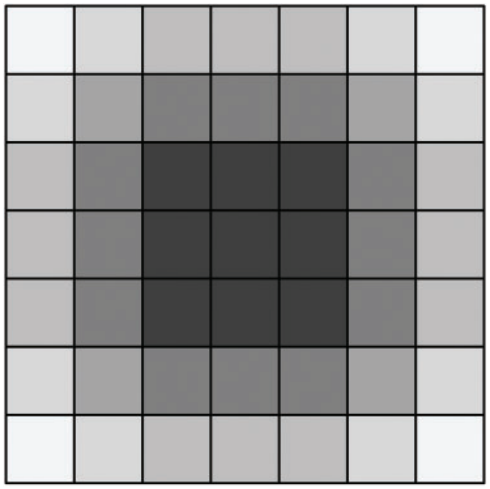

Figure 5: A refill-free convolution over a 7 × 7 pixel image

For similar faces, learning common features can help simplify the process of feature extraction. As such, the network does not have to learn the full features of each face from scratch but only needs to distinguish the differences between individuals based on common features. Therefore, we can extract the average characteristics of said as the initial common representation and use the exponential moving average exponentially weighted moving average (EWMA) to continually update it to adapt to the different concentrations.

$$E_{batch}, E_{init} = \frac{\sum_{i=1}^{N} e_i}{N} \tag{4}$$

$$E_{common} = \lambda E_{batch} + (1 - \lambda) E'_{common} \tag{5}$$

The features of the current batch are denoted as E_{batch} batches, while the previous average features are labelled as E'_{common}. The coefficient λ represents the rate of weight reduction, which is a constant between 0 and 1 used to adjust smoothness.

3.3 Timing Feature Extraction Module

The feature vectors extracted from consecutive frames are considered multidimensional inputs for the continuous time steps of the ModernTCN [6] module, simulating temporal information in the video. In our proposed architecture, both temporal and spatial modelling are jointly trained on the frame sequence data of the video. We employ two stacked ModernTCN blocks to model the temporal changes, which output the final engagement level. ModernTCN, as a pure convolutional structure, maintains the efficiency advantages of convolutional models while exhibiting excellent performance in time series analysis tasks. Inspired by some architectural designs in transformers [39], the modern convolutional block incorporates some design elements from transformers, thus having a structure similar to that of transformer blocks and utilizing large kernels to increase receptive fields.

Specifically, ModernTCN combines the design of modern convolution with the traditional TCN [40], enabling not only increased global awareness of temporal sequence data but also capturing correlations across variables. Before the backbone, we embed the channel features at each pixel into a D-dimensional vector, mixing the information across channels through an embedding layer. Simply embedding variables into fixed-dimensional vectors may fail to capture the complex dependencies between variables and may even lead to the loss of individual characteristics of variables. Moreover, this embedding method may not effectively preserve the dimensional information between variables, thereby limiting the study of intervariable dependencies. To address this issue, we adopt a method called patchwise variable-agnostic embedding.

After appropriately padding a time series $x_{in} \in R^{K\times L}$ consisting of K variables (each of length L), it is further divided into N patches of size P, spaced apart by a step size of S to ensure nonoverlapping. Then, these patches are embedded into vectors.

$$x_{emb} = Embbdding\left(X_{in}\right) \tag{6}$$

$x_{emb} \in \mathbb{R}^{K\times D\times N}$ is the embedded input. Then, these patches are embedded into D-dimensional embedding vectors using an equivalent fully convolutional approach. After reshaping the shape to $x_{in} \in R^{K\times 1\times L}$, the padded input is mapped to a vector with D output channels through a 1D convolutional layer. Notably, embedding is carried out independently for each univariate time series to capture information from additional variable dimensions.

The Depthwise Separable Convolution (DWConv) in the ModernTCN block is responsible for learning the temporal information between tokens based on each feature, independently learning the temporal dependencies of each univariate time series, similar to the self-attention module in Transformer [41]. Moreover, the Convolutional Feed-Forward Network (ConvFFN), similar to the Feed-Forward Network (FFN) module in Transformer, independently learns new feature representations for each token, supplementing mixed information across feature and variable dimensions. This design separates the mixture of temporal and feature information, making the task easier to learn and reducing computational complexity. The processed features are input into DWConv, which enhances its nonlinear modelling capability by employing large kernels. For information across variable dimensions, ModernTCN decomposes a single ConvFFN into ConvFFN1 and ConvFFN2. ConvFFN1 learns new feature representations for each variable, while ConvFFN2 focuses on capturing cross-variable dependencies for each feature.

The embedded $x_{emb} \in \mathbb{R}^{K\times D\times N}$ is then input into the backbone network to capture the dependencies between time and variables and learn a representative Z with information, with a dimensionality of $Z \in \mathbb{R}^{K\times D\times N}$.

$$Z = Backbone\left(X_{emb}\right) \tag{7}$$

$Backbone(\ )$ is two stacked ModernTCN blocks, each organized in a residual formula.

$$Z_{i+1} = Block\left(Z_i\right) + Z_i \tag{8}$$

In this architecture, $Z_i \in \mathbb{R}^{K\times D\times N}$ and $i \in \{1, \ldots, K\}$ represent the inputs for the i-th block. After passing through the backbone, the output is denoted as $Z_i \in \mathbb{R}^{K\times D\times N}$. The flattened layer then reshapes the final representation to $Z_i \in \mathbb{R}^{1\times(K\times D\times N)}$. Subsequently, a projection layer with SoftMax activation maps the final representation to the ultimate classification result. The overall structure of the ModernTCN block is illustrated in Fig. 2.

4 Experimental Validation

In this section, we conduct experiments to investigate the performance of our method, comparing it to the most advanced methods.

4.1 Datasets

The experiment utilized the self-collected Near Infrared Engagement Dataset (NEVD) and the publicly available Database for Affect in Situations of Elicitation (DAiSEE) [13] to train and validate our method.

To validate the effectiveness of the proposed network, Hikvision network cameras were used to record near infrared (NIR) engagement videos of 25 university students (18 males and 7 females) aged between 18 and 22 years. These students are enrolled in a four-year undergraduate program in Computer Science. The recorded videos capture their engagement during various stages of their coursework, which includes foundational and advanced courses such as programming, data structures, algorithms, and software engineering. Each student was recorded in four videos, each lasting four minutes, covering four different levels of engagement. For low-engagement scenarios, behaviours such as yawning, nodding off, and looking down were simulated to mimic common distractions during studying, while high-engagement scenarios featured students displaying behaviours such as forward gaze and attentiveness. Fig. 6 shows examples of videos from the NEVD. The camera parameters were set to a resolution of 1920 × 1080 pixels at a frame rate of 30 frames/second. The engagement processing method for the NEVD dataset was consistent with the categorization used in the DAiSEE. To further augment the dataset, different starting frames were selected, and individual videos were resampled at equal intervals. In the end, we obtained 2088 videos, each lasting 10 s, for this experiment, with the sample counts shown in Table 1. The engagement processing method for the NEVD dataset was aligned with that for the DAiSEE to ensure experimental consistency and comparability.

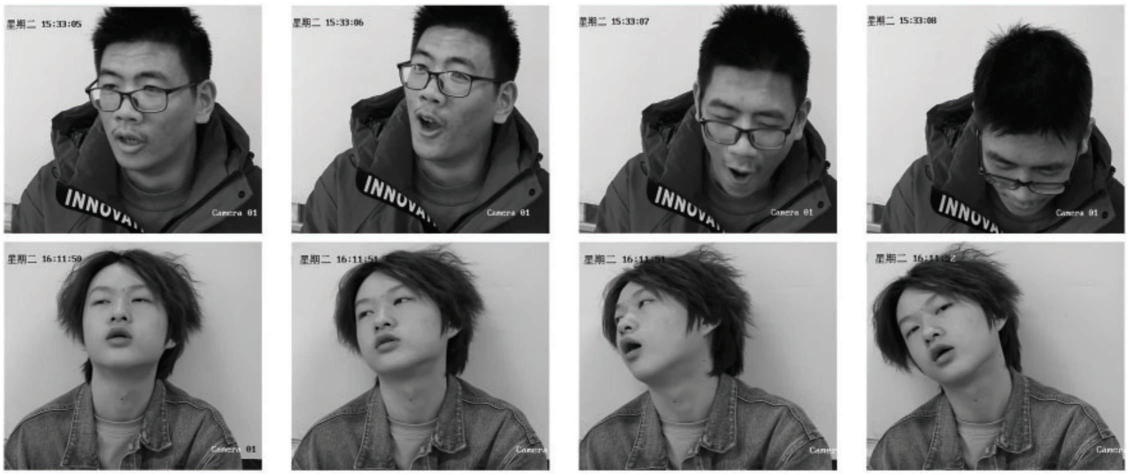

Figure 6: (Continued)

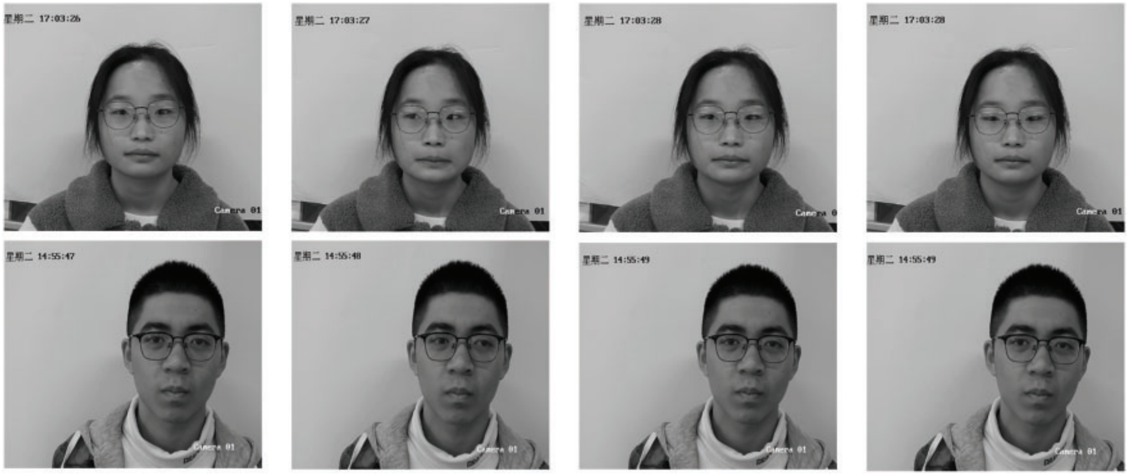

Figure 6: Examples of videos from the NEVD. The engagement intensity levels are shown from top to bottom: [0 (Very low)–3 (Very high)]. The frames are displayed sequentially from left to right

Table 1: Samples in train and validation set on NEVD

Leval	Train	Validation
0	444	111
1	416	105
2	396	100
3	412	104
Total	1668	420

DAiSEE is one of the largest publicly available datasets in the field of engagement detection. This dataset includes 9068 videos from 112 students participating in online courses aimed at analysing their emotional states in natural environments, such as boredom, confusion, engagement, and frustration. Each video lasts up to 10 s and is categorized into four levels of engagement intensity: 0 (very low), 1 (low), 2 (high), and 3 (very high), with a frame rate of 30 fps and a resolution of 640 × 480 pixels. We combined the training and validation sets to train the model and tested the classification results on the test videos. Table 2 below shows the number of samples for each category, indicating a significant imbalance in the dataset. Fig. 7 shows examples of videos from the DAiSEE.

4.2 *Experimental Evaluation Metrics*

Gupta et al. [13] used Top-1 accuracy as the evaluation metric for classification models in their research on the DAiSEE. To maintain consistency in the evaluation, this paper also employs the same metric. Top-1 accuracy selects the category with the highest probability value from the model's predictions and compares it with the true category. If they match, the prediction is considered correct; otherwise, it is considered incorrect. The specific calculation method is shown in the following formula:

$$P = \frac{TP}{TP + FP} \tag{9}$$

In the formula: *P* represents Precision, *TP* represents True Positives, and *FP* represents False Positives.

Table 2: Samples in train, validation, and test sets on DAiSEE

Leval	Train	Validation	Test
0	34	23	4
1	213	143	84
2	2617	813	882
3	2494	450	814
Total	5258	1429	1784

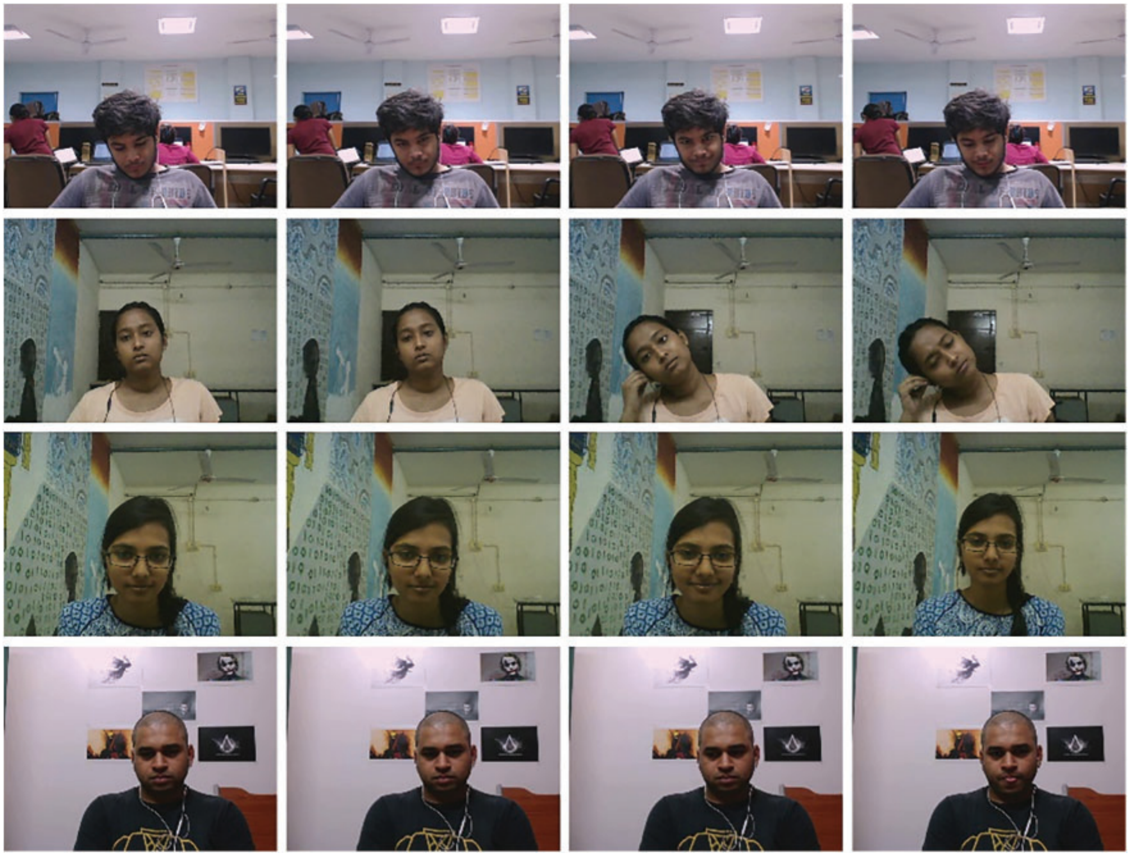

Figure 7: Examples of videos from the DAiSEE. The engagement intensity levels are shown from top to bottom: [0 (Very low)–3 (Very high)]. The frames are displayed sequentially from left to right

4.3 Experimental Details

Our method employs the Stochastic Gradient Descent (SGD) optimizer with L2 loss for training. The initial learning rate is set to 10-4, and the momentum is 0.9. The default training process consists of 100 epochs with early stopping applied as necessary. The resolution of the facial images is adjusted to (112, 112). Videos are downsampled temporally and spatially to obtain a 16 × C × 112 × 112 (L × C × H × W) tensor as the input to the architecture, with C adjusted according to the different datasets. The backbone network is part of ResNet-18 [10], initialized with pre-trained weights from ImageNet. We use this part of the network to extract whitened features with a size of 512 × 7 × 7. According to the rearrangement principle in Section 3.2.2, due to the number of channels not being a multiple of 4, only two-channel feature maps can be obtained. As the number of feature map channels decreases, the spatial area of the feature map increases to 112 × 112 pixels. No padding is added during the convolution process, making the choice of convolution kernel size crucial. We set the convolution kernel size to 32 × 32 × 1 pixels, which is larger than the size of the feature clusters, with a stride of 8. The two channels output from the WD block are averaged to represent each expression. For facial affinity enhancement, the λ gain hyperparameter is set to a default of 0.2, as recommended by the literature [38], with the gain hyperparameter depending on the affinity within the dataset and updated through gradients. Through the network output, we can effectively obtain high-quality representations for each feature. The extracted feature vectors are fed into ModernTCN, which consists of 2 ModernTCN blocks. According to the definition of ModernTCN [12] for classification tasks, during the patch embedding process, the patch size and stride are set to $P = 1$ and $S = 1$, respectively. All the deep learning networks were implemented in PyTorch, and the experiments were conducted on an NVIDIA A100 40 GB Graphic Processing Unit (GPU).

4.4 Comparison with State-of-the-Art Methods

Our method was tested on the NEVD, and Table 3 presents the results compared with those of other works. We implemented methods based on several GitHub public repositories, including I3D, C3D, S3D (Spatial-Temporal Separation Network), ResNet + LSTM, ResNet + TCN, and Shifted Window Transformer Base (Swin-B) [26,27,42–44]. Table 3 provides the experimental results of different methods for predicting engagement intensity on the NEVD. It can be observed from the table that our proposed method outperforms traditional end-to-end learning methods such as I3D, C3D, and S3D, achieving the highest accuracy of 90.8%. Compared to the state-of-the-art end-to-end Swin-B algorithm, our method also shows a 1.6% improvement. Traditional methods often face the challenges of high computational cost and high memory usage, while our method exhibits superior performance in these aspects. We achieve higher accuracy using fewer frames, and compared to the state-of-the-art Swin-B, our method demonstrates lower computational cost and memory usage.

Table 3: Performance comparisons on NEVD. "Acc" represents the accuracy of the predictions. "Frames" refers to the individual images in a video sequence. The magnitudes for floating point operations (FLOPs) and parameters are measured in giga (10e9) and mega (10e6), respectively

Method	Acc	Frames	FLOPS	Param
I3D	83.4	64	37.2	14.4
C3D	85.7	64	38.6	63.2
S3D	87.7	64	24.4	10.3
ResNet-LSTM	87.8	32	9.4	11.2
Swin-B	89.2	32	147.5	88.1
ResNET-TCN	89.5	16	6.3	12.9
RefEIP	**90.8**	16	19.8	46.2

To validate the effectiveness of our method, we conducted the same experiments on the DAiSEE. Table 4 presents the results of different methods for engagement intensity prediction on the DAiSEE. The data in the table show that our proposed method outperforms most end-to-end and feature-based methods. We implemented methods based on several approaches, including Deep Facial Spatiotemporal Network (DFSTN) [7], Deep Emotion Recognition Network (DERN) [30], Shuffle Attention-Global Attention-Residual Long Short-Term Memory (SA-GA-Res-LSTM) [6], and Squeeze-and-Excitation ResNet with Temporal Convolution Network (SE-ResNET-TCN) [33]. Compared to the state-of-the-art end-to-end SE-GA-RES-LSTM algorithm, our method achieves a 1% increase in accuracy. Our algorithm demonstrates greater flexibility by employing fewer attention modules compared to the complex attention mechanism utilized by the aforementioned algorithm, which allows us to better capture local patterns and extract effective features. Notably, our method slightly lags behind the fusion feature-based algorithm mentioned in [6]. We speculate that this lag is due to the larger backbone network used in the model in [6], which integrates multiple features such as facial features, head posture, and blinking. This complexity leads to increased inference time and resource consumption. The fourth column in the table displays the number of frames in the video sequences. Despite using fewer frames, we still achieved excellent results.

Table 4: Comparison results on the DAiSEE dataset

Method	Author	Year	Frames	Acc
I3D [29]	Zhang	2019	64	52.4
C3D [28]	Geng	2019	64	57.6
S3D [42]	Xie	2018	64	59.7
DFSTN	Liao	2021	20	58.8
DERN	Huang	2019	64	60.0
Swin-B	Liu	2022	32	60.2
SA-GA-Res-LSTM	Liang	2023	16	60.2
SE-ResNET-TCN	Lu	2022	10	61.4
RefEIP	Ours	2024	16	61.2

Fig. 8 shows the confusion matrices of engagement levels for different methods on the NEVD and DAiSEE databases. The actual classes are along the vertical axis and the predicted classes are along the horizontal axis. The numbers in the confusion matrix represent the proportion of samples predicted as a particular class out of the total number of actual samples for that class. The first row shows that our method effectively distinguishes between different levels of engagement. However, in the DAiSEE dataset, the distribution of engagement labels is extremely imbalanced (1:8:73:67), with the ratio of the largest to the smallest sample count reaching 67. Even the most advanced end-to-end methods currently available struggle to correctly classify minority classes (low engagement levels, i.e., engagement levels 0 and 1). By examining the confusion matrix, we can see that some samples in Class 1 are correctly classified, further demonstrating that the introduction of the redistributing facial features module effectively improves the adaptation to local contextual information and reduces the whitening features introduced by padding. This improvement significantly enhances the model's performance in facial feature extraction and engagement recognition tasks, particularly in the classification accuracy of low-engagement-level samples.

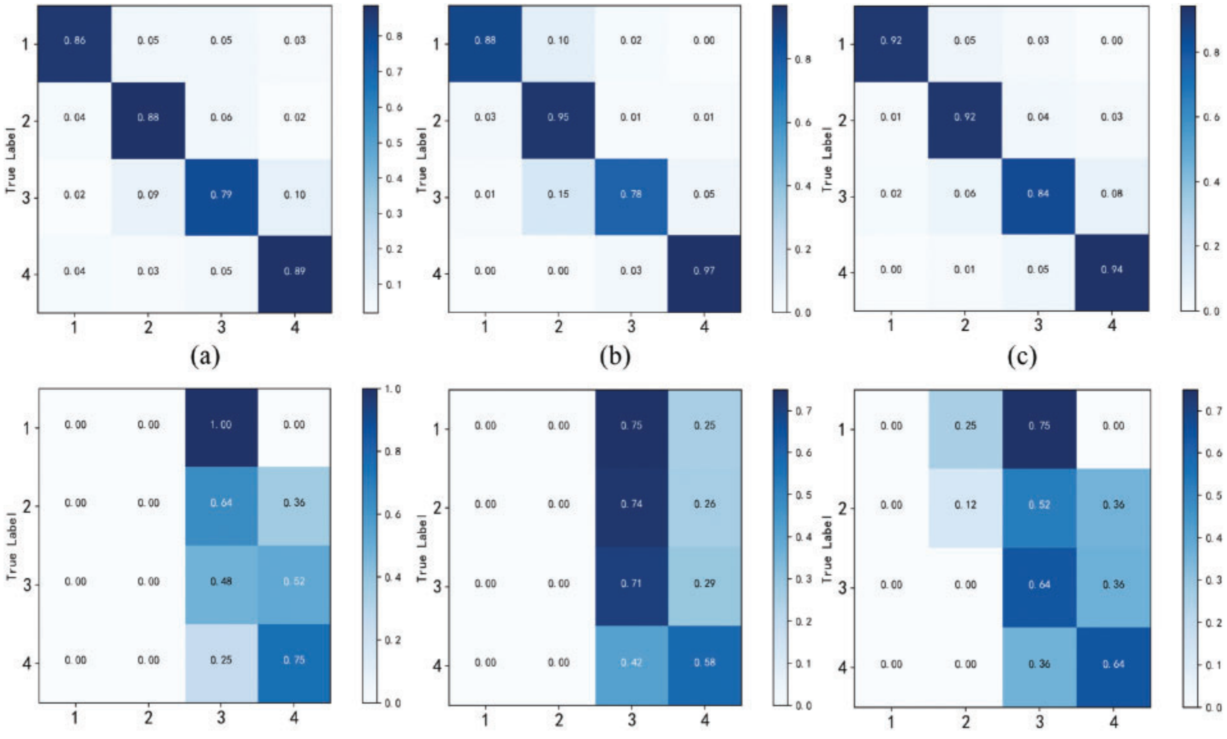

Figure 8: Confusion matrices of engagement levels for different methods on the NEVD and DAiSEE databases. (a) Results of the C3D method on the NEVD. (b) Results of the Swin-B method on the NEVD. (c) Results of the RefEIP method on the NEVD. (d) Results of the C3D method on the DAiSEE. (e) Results of the Swin-B method on the DAiSEE. (f) Results of the RefEIP method on the DAiSEE

4.5 Ablation Study

4.5.1 Face Extraction

Facial extraction is the primary step in handling noise, prompting an in-depth study of the relationship between facial features and engagement. We used the MTCNN to perform facial extraction on the NEVD and DAiSEE databases. According to the results in Table 5, the cases with facial extraction outperformed those without it. This empirical result underscores the critical role of the face in expressing the intensity of emotional connections.

Table 5: Ablation study on face extraction approach with RefEIP on NEVD and DAiSEE databases

	NEVDAcc	DAiSEEAcc
Face extraction	90.8	61.2
No face extraction	84.4	57.4

4.5.2 The Negative Impact of Albinism Characteristics

To ensure the credibility of our research, we conducted an independent experiment to explore the potential impact of white features on the results. We adjusted ResNet-18 [10] to meet the requirements of the experiment. We introduced the FOWD block into the model to replace the global average pooling layer and applied convolutional processing to the output 512 × 7 × 7 feature map. Within the FOWD block, we utilized an adjustable kernel size, k, while keeping the stride constant. Finally, we retained the fully connected layer and made adaptive adjustments. The experimental results are depicted in Fig. 9. Based on the chart, we conclude that as k increases, the gain effect exhibits a specific peak shape, consistent with the trend of changes in the ratio of perceptual frequencies. This finding suggests that increasing k within a certain range can enhance the weighting of specific directional features in the feature map, thereby increasing the sensitivity of the model to these directional features. Experiments on the NEVD clearly indicate that by reducing the weight of the white features at the edges of the feature maps, the representation quality can be effectively improved, thereby slightly enhancing the model performance. However, the introduction of white features does indeed have a certain impact on the representation of images.

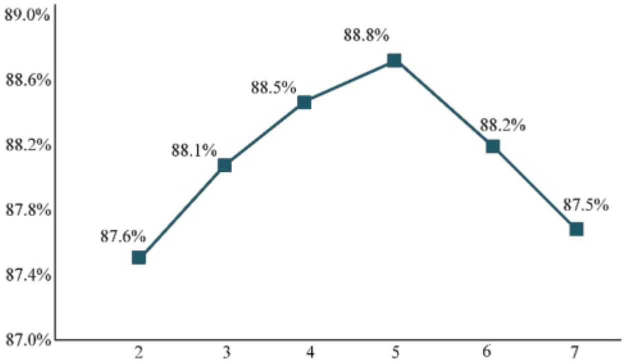

Figure 9: Evaluation of the kernel size, k, of the FOWD block on the NEVD

4.5.3 Time Series Network Comparison

The feature vectors extracted from consecutive frames are used as multidimensional inputs for modelling temporal information in videos. To validate the effectiveness of our proposed hybrid neural network architecture, we conducted experiments on the NEVD and DAiSEE databases by replacing ModernTCN with these methods. Table 6 presents a comparison between the ModernTCN and other recurrent temporal networks. LSTM introduces three gate units that are adept at handling long-term dependencies by learning to control the flow of information. TCNs adopt convolutional methods to capture long-term dependencies in time series and have advantages in modelling long sequences and retaining historical memory. The Continuous Kernel Convolution (CKCNN) parallelizes the processing of arbitrary-length sequences by representing convolutional kernels as continuous functions in a single operation. Patch time series Transformer (PatchTST) excels in handling multivariable time series by leveraging channel independence, particularly in long-term sequence tasks. The table shows that ModernTCN performs better on both datasets. We believe that by incorporating transformer architecture design and increasing the effective receptive field (ERF) through the adoption of large kernels, ModernTCN effectively enhances participation prediction performance. In this experiment, we observed significant differences in the performances of different models on the two datasets. Although ModernTCN performs better overall, a traditional TCN also achieves excellent results on the DAiSEE, while PatchTST performs better on the NEVD. This difference may be attributed to the DAiSEE containing more training data, allowing the model to learn from a more diverse range of contexts, thereby enhancing its generalizability and performance. Additionally, another reason that the TCN achieves better results may be its superior classification performance on short time series.

Table 6: Comparison of ModernTCN with other temporal convolutional networks

Dataset	LSTM [32]	TCN [40]	CKConv [45]	PatchTST [39]	ModernTCN [12]
NEVD	88.3	90.0	89.9	90.1	90.8
DAiSEE	59.5	61.0	60.1	60.4	61.2

4.5.4 Effectiveness of the Proposed Modules

To validate the modules proposed in RefEIP, we conducted an ablation experiment to explore the impact of the LKA, FOWD, and facial affinity enhancement (FAE) modules on the NEVD and DAiSEE databases, as illustrated in Table 7. The baseline strategy (first row) involved feeding the feature maps extracted from ResNet-18 directly into a standard ModernTCN without guidance from any LKA or FOWD components. Compared to the baseline, integrating the LKA module resulted in performance improvements on both datasets (1.5% and 1.1%, respectively). The LKA module, which combines convolution and self-attention mechanisms, adeptly captures local contextual information and long-range dependencies in facial images, addressing challenges such as pose variations and incomplete facial regions, which ultimately enhances facial recognition performance. Furthermore, the addition of the FOWD module led to significant performance enhancements (1.6% and 1.2%, respectively), attributed to its ability to adjust the weight distribution within feature maps based on padding levels, thereby mitigating the introduction of extraneous information and exploring more identifiable feature regions. The proposed facial affinity enhancement (FAE) module simplified the feature extraction process, prompting the network to learn universal features to adapt to different engagement states, resulting in state-of-the-art performances of 90.8% and 61.2%, respectively. This significant improvement indicates that multiple modules complement each other, enabling comprehensive learning and capturing useful representation features. We believe that the synergistic effect of these modules has led the model to achieve optimal performance on the NEVD and DAiSEE databases.

Table 7: Results of ablation experiments on the NEVD and DAiSEE databases

LKA	FOWD	FAE	NEVD	DAiSEE
			87.4	58.7
✓			88.9	59.8
✓	✓		90.5	61.0
✓	✓	✓	90.8	61.2

4.6 Attention Visualization

To illustrate the effectiveness of our approach more intuitively, we adopted the Gradient-weighted Class Activation Mapping (GradCAM) [46] method for visualizing heatmap images of image frames. Specifically, we first resized the visualization attention map to the same size as the input image and then generated heatmap images highlighting regions associated with specific classes by backpropagating the model's gradient information to the output feature maps of the convolutional layers.

Fig. 10 shows attention maps for different engagement levels. The following findings were obtained when comparing different columns (training strategies): Compared to the baseline method (I), the addition of the LKA module allows for more comprehensive coverage of facial regions, including key features such as the eyes, nose, and mouth, which are crucial for engagement prediction tasks. Notably, as shown in Fig. 10b,c, when the target object's pose changes significantly, our method can still accurately locate key areas that are helpful for the task. This accuracy is mainly due to LKA's ability to comprehensively consider local context information, flexibly capturing both the local context and long-range dependencies of facial images. Compared to strategy (III), the introduction of the FOWD module further optimizes the entire framework. FOWD reconstructs the previously extracted feature maps by reducing the whitening features and irrelevant information introduced by padding operations. As a result, we can more accurately locate areas of interest, such as the hand occlusion area in Fig. 10a of (III). This module eliminates redundant information, making localization more precise. With the application of the LKA and FOWD modules, the entire framework can focus on more distinctive facial regions. Comparing the visualizations of frame heatmaps with Shuffle Attention and SE attention mechanisms added to the ResNet18 network in methods (IV) and (V) clearly shows that our method (III) better focuses on facial regions that represent engagement. Although the Shuffle Attention and SE (Squeeze-and-Excitation) attention mechanisms can improve the capture of feature areas to some extent, they still have deficiencies in the precise localization of facial features. In contrast, our method not only accurately locates key areas by incorporating the LKA and FOWD modules, even with significant pose changes, but also better focuses on facial features, such as the eyes and mouth, which are closely related to engagement prediction. This finding demonstrates that our method has significant advantages in the extraction and focus on facial features, making it better suited for engagement prediction tasks.

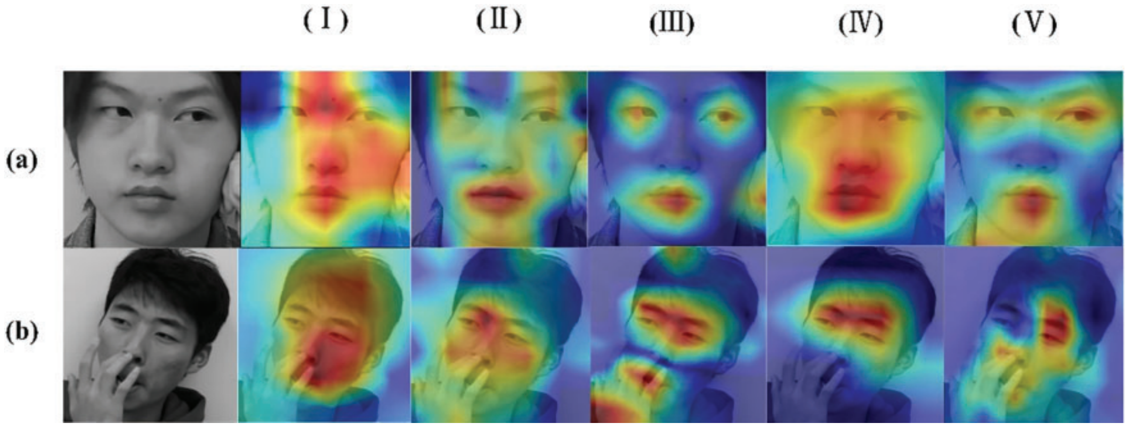

Figure 10: (Continued)

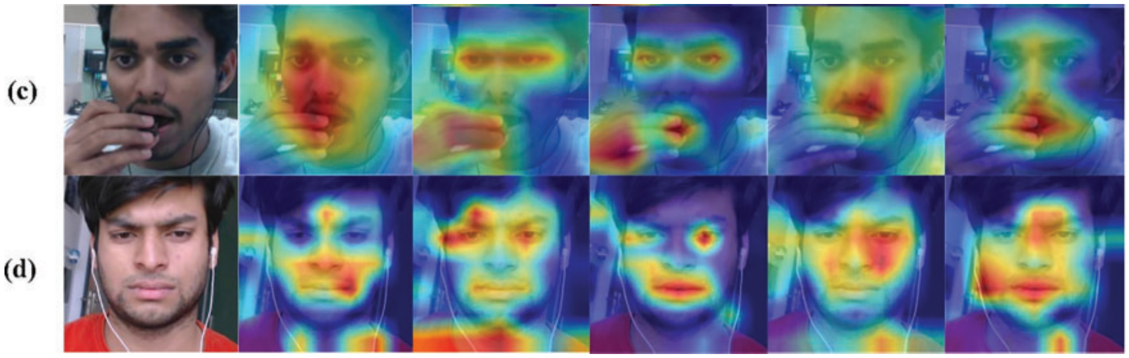

Figure 10: Attention visualization of different engagement levels on some example face images from the NEVD and DAiSEE datasets. (a)–(b) are collected from the NEVD dataset, and (c)–(d) are collected from the DAiSEE dataset. (I)–(V) denote five training strategies. (I) denotes the baseline strategy (II) denotes the baseline strategy with the addition of the LKA module (III) denotes the strategy using both LKA and FOWD modules. (IV) SE-ResNET-TCN, (V) SA-GA-Res-LSTM

5 Conclusion

This paper introduces an end-to-end residual network and temporal convolutional network hybrid neural architecture called RefEIP for engagement prediction. This architecture integrates facial spatiotemporal information aimed at capturing subtle changes in student engagement states to enhance engagement prediction performance. We propose the redistributing facial features module, through which we successfully capture local patches and eliminate whitening erosion present in padding convolutions, optimizing facial features and improving representation quality. Additionally, we collected a near-infrared engagement video dataset (NEVD) to evaluate the effectiveness of our approach. The experimental results demonstrate satisfactory detection accuracy on the NEVD and outstanding performance on the publicly available Database for Affect in Situations of Elicitation (DAiSEE), surpassing current state-of-the-art methods.

Acknowledgement: The authors would like to express their gratitude for the valuable feedback and suggestions provided by all the anonymous reviewers and the editorial team.

Funding Statement: This work was supported by the National Natural Science Foundation of China (No. 62367006) and the Graduate Innovative Fund of Wuhan Institute of Technology (Grant No. CX2023551).

Author Contributions: Conceptualization, Xi Li and Weiwei Zhu; methodology, Weiwei Zhu; software, Qian Li; validation, Xi Li, Weiwei Zhu and Yaozong Zhang; formal analysis, Qian Li; investigation, Changhui Hou; resources, Yaozong Zhang; data curation, Xi Li; writing—original draft preparation, Qian Li; writing—review and editing, Changhui Hou and Weiwei Zhu; visualization, Qian Li; project administration, Xi Li. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data presented in this study are available on request from the corresponding author. Data are not publicly available due to privacy considerations.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] O. Mohamad Nezami, M. Dras, L. Hamey, D. Richards, S. Wan and C. Paris, "Automatic recognition of student engagement using deep learning and facial expression," in *Machine Learning and Knowledge Discovery in Databases*, Cham, Switzerland: Springer International Publishing, 2020, pp. 273–289.

[2] J. Yang, K. Wang, X. Peng, and Y. Qiao, "Deep recurrent multi-instance learning with spatio-temporal features for engagement intensity prediction," in *Proc. 20th ACM Int. Conf. Multimodal Interact.*, Seattle, WA, USA, Oct. 16–20, 2018, pp. 594–598.

[3] P. Guhan, M. Agarwal, N. Awasthi, G. Reeves, D. Manocha and A. Bera, "ABC-Net: Semi-supervised multimodal GAN-based engagement detection using an affective, behavioral and cognitive model," 2020, *arXiv:2011.08690*.

[4] D. Li, H. Liu, W. Chang, P. Xu, and Z. Luo, "Visualization analysis of learning attention based on single-image PnP head pose estimation," in *2017 2nd Int. Conf. Educ., Sports, Arts Manag. Eng. (ICESAME 2017)*, Zhengzhou, China, Apr. 29–30, 2017, pp. 1508–1512.

[5] S. D'Mello, E. Dieterle, and A. Duckworth, "Advanced, analytic, automated (AAA) measurement of engagement during learning," *Educ. Psychol.*, vol. 52, no. 2, pp. 104–123, 2017. doi: 10.1080/00461520.2017.1281747.

[6] Y. Lu, Y. Zhan, Z. Yang, and X. Li, "Student engagement recognition network integrating facial appearance and multi-behavior features," *Comput. Sci. Appl.*, vol. 12, p. 1163, 2022.

[7] J. Liao, Y. Liang, and J. Pan, "Deep facial spatiotemporal network for engagement prediction in online learning," *Appl. Intell.*, vol. 51, no. 10, pp. 6609–6621, 2021. doi: 10.1007/s10489-020-02139-8.

[8] H. Wu *et al.*, "CvT: Introducing convolutions to vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Montreal, QC, Canada, Oct. 11–17, 2021, pp. 22–31.

[9] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, 2016. doi: 10.1109/LSP.2016.2603342.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, NV, USA, Jun. 27–30, 2016, pp. 770–778.

[11] M. -H. Guo, C. -Z. Lu, Z. -N. Liu, M. -M. Cheng, and S. -M. Hu, "Visual attention network," *Comput. Vis. Media*, vol. 9, no. 4, pp. 733–752, 2023. doi: 10.1007/s41095-023-0364-2.

[12] D. Luo and X. Wang, "ModernTCN: A modern pure convolution structure for general time series analysis," in *Twelfth Int. Conf. Learn. Represent.*, Vienna, Austria, Apr. 2024, pp. 25–36.

[13] A. Gupta, A. D'Cunha, K. Awasthi, and V. Balasubramanian, "DAiSEE: Towards user engagement recognition in the wild," 2016, *arXiv:1609.01885*.

[14] S. Mandia, R. Mitharwal, and K. Singh, "Automatic student engagement measurement using machine learning techniques: A literature study of data and methods," *Multimed. Tools Appl.*, vol. 83, no. 16, pp. 49641–49672, 2023. doi: 10.1007/s11042-023-17534-9.

[15] P. Sharma *et al.*, "Student engagement detection using emotion analysis, eye tracking and head movement with machine learning," in *Int. Conf. Tech. Innov. Learn., Teach. Educ.*, Beijing, China, Jun. 25, 2022, pp. 52–68.

[16] N. Alruwais and M. Zakariah, "Student-engagement detection in classroom using machine learning algorithm," *Electronics*, vol. 12, no. 3, p. 731, 2023. doi: 10.3390/electronics12030731.

[17] M. He, J. Zhang, S. Shan, M. Kan, and X. Chen, "Deformable face net for pose invariant face recognition," *Pattern Recognit.*, vol. 100, no. 10, p. 107113, 2020. doi: 10.1016/j.patcog.2019.107113.

[18] J. Whitehill, Z. Serpell, Y. -C. Lin, A. Foster, and J. R. Movellan, "The faces of engagement: Automatic recognition of student engagementfrom facial expressions," *IEEE Trans. Affect. Comput.*, vol. 5, no. 1, pp. 86–98, 2014. doi: 10.1109/TAFFC.2014.2316163.

[19] C. Qi, J. Zhang, H. Jia, Q. Mao, L. Wang and H. Song, "Deep face clustering using residual graph convolutional network," *Knowl. Based Syst.*, vol. 211, no. 2, p. 106561, 2021. doi: 10.1016/j.knosys.2020.106561.

[20] J. Chun and W. Kim, "3D face pose estimation by a robust real time tracking of facial features," *Multimed. Tools Appl.*, vol. 75, no. 23, pp. 15693–15708, 2016. doi: 10.1007/s11042-014-2356-9.

[21] J. Wu, B. Yang, Y. Wang, and G. Hattori, "Advanced multi-instance learning method with multi-features engineering and conservative optimization for engagement intensity prediction," in *Proc. 2020 Int. Conf. Multimodal Interact.*, Darmstadt, Germany, Nov. 9, 2020, pp. 777–783.

[22] A. Kaur, A. Mustafa, L. Mehta, and A. Dhall, "Prediction and localization of student engagement in the wild," in *2018 Dig. Image Comput.: Tech. App. (DICTA)*, Melbourne, Australia, Dec. 3–5, 2018, pp. 1–8.

[23] Y. Wang, A. Kotha, P. -H. Hong, and M. Qiu, "Automated student engagement monitoring and evaluation during learning in the wild," in *2020 7th IEEE Int. Conf. Cyber Secur. Cloud Comput. (CSCloud)/2020 6th IEEE Int. Conf. Edge Comput. Scal. Cloud (EdgeCom)*, Beijing, China, Jul. 6–8, 2020, pp. 270–275.

[24] J. Lee, W. Reade, R. Sukthankar, and G. Toderici, "The 2nd YouTube-8M large-scale video understanding challenge," in *Proc. Eur. Conf. Comput. Vis. (ECCV) Workshops*, Munich, Germany, Sep. 8–14, 2018.

[25] J. Donahue *et al.*, "Long-term recurrent convolutional networks for visual recognition and description," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, MA, USA, Jun. 7–12, 2015, pp. 2625–2634.

[26] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, Las Vegas, NV, USA, Jun. 27–30, 2016, pp. 2818–2826.

[27] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Santiago, Chile, Dec. 7–13, 2015, pp. 4489–4497.

[28] L. Geng, M. Xu, Z. Wei, and X. Zhou, "Learning deep spatiotemporal feature for engagement recognition of online courses," in *2019 IEEE Symp. Series Computat. Intell. (SSCI)*, Athens, Greece, Dec. 6–9, 2019, pp. 442–447.

[29] H. Zhang, X. Xiao, T. Huang, S. Liu, Y. Xia and J. Li, "An novel end-to-end network for automatic student engagement recognition," in *2019 IEEE 9th Int. Conf. Electron. Inf. Emerg. Commun. (ICEIEC)*, Beijing, China, Jun. 21–23, 2019, pp. 342–345.

[30] T. Huang, Y. Mei, H. Zhang, S. Liu, and H. Yang, "Fine-grained engagement recognition in online learning environment," in *2019 IEEE 9th Int. Conf. Electron. Inf. Emerg. Commun. (ICEIEC)*, Harbin, China, Jan. 20–22, 2019, pp. 338–341.

[31] T. Baltrušaitis, P. Robinson, and L. -P. Morency, "OpenFace: An open source facial behavior analysis toolkit," in *2016 IEEE Winter Conf. App. Comput. Vis. (WACV)*, NV, USA, Mar. 7–10, 2016, pp. 1–10.

[32] A. Graves and A. Graves, "Long short-term memory," in *Supervised Sequence Labelling with Recurrent Neural Networks*, 1st ed. Berlin, Germany: Springer, 2012, pp. 37–45.

[33] Y. Liang, Z. Zhou, W. Huang, and Z. Guo, "Attention detection in online education based on spatiotemporal attention mechanism," *Softw. Guide*, vol. 23, no. 1, pp. 150–155, 2024. Accessed: May 1, 2024. https://kns.cnki.net/kcms/detail/42.1671.TP.20230904.1124.html

[34] Z. Qin, P. Zhang, F. Wu, and X. Li, "FcaNet: Frequency channel attention networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Montreal, QC, Canada, Oct. 11–17, 2021, pp. 783–792.

[35] M. -H. Guo *et al.*, "Attention mechanisms in computer vision: A survey," *Computat. Vis. Media*, vol. 8, no. 3, pp. 331–368, 2022.

[36] L. Chen *et al.*, "SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, Honolulu, HI, USA, Jul. 21–26, 2017, pp. 5659–5667.

[37] W. Shi *et al.*, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, Las Vegas, NV, USA, Jun. 27–30, 2016, pp. 1874–1883.

[38] H. Yang, U. Ciftci, and L. Yin, "Facial expression recognition by de-expression residue learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, Salt Lake City, UT, USA, Jun. 18–22, 2018, pp. 2168–2177.

[39] Y. Nie, N. H. Nguyen, P. Sinthong, and J. Kalagnanam, "A time series is worth 64 words: Long-term forecasting with transformers," 2022, *arXiv:2211.14730*.

[40] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," 2018, *arXiv:1803.01271*.

[41] A. Vaswani *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems*, Long Beach, CA, USA, Dec. 4–9, 2017, vol. 30, pp. 3–19.

[42] S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murphy, "Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, Sep. 8–14, 2018, pp. 305–321.

[43] A. Abedi and S. S. Khan, "Improving state-of-the-art in detecting student engagement with Resnet and TCN hybrid network," in *2021 18th Conf. Robots Vis. (CRV)*, Ottawa, ON, Canada, May 26–28, 2021, pp. 151–157.

[44] Z. Liu *et al.*, "Video swin transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recogn.*, New Orleans, LA, USA, Jun. 19–24, 2022, pp. 3202–3211.

[45] D. W. Romero, A. Kuzina, E. J. Bekkers, J. M. Tomczak, and M. Hoogendoorn, "CKConv: Continuous kernel convolution for sequential data," 2021, *arXiv:2102.02611*.

[46] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, Venice, Italy, Oct. 22–29, 2017, pp. 618–626.