Check for updates

CAAI Transactions on Intelligence Technology

WILEY

The Institution of Engineering and Technology

ORIGINAL RESEARCH OPEN ACCESS

Dynamic Facial Expression Recognition of Learners via Adaptive Global Attention and Differential Temporal Transformer

Wei Liu[1] | Lujia Li[1] | Chun Yan[2] | Yulin Zhang[2] | Xiaochun Cheng[3] | Xinyan Zhao[1] | Mingshi Liu[2]

[1]College of Computer Science and Engineering, Shandong University of Science and Technology, Qingdao, China | [2]College of Mathematics and Systems Science, Shandong University of Science and Technology, Qingdao, China | [3]Computer Science Department, Swansea University, Swansea, UK

Correspondence: Xiaochun Cheng (Xiaochun.Cheng@Swansea.ac.uk)

Received: 12 October 2025 | **Revised:** 30 January 2026 | **Accepted:** 8 February 2026

Keywords: face analysis | facial expression recognition | spatial-temporal feature | transformer

ABSTRACT

Analysing learners' facial expressions during learning and exploring their learning processes and emotional changes are of great significance for assisting teachers' teaching and promoting smart education. In complex learning environments, static facial expression recognition fails to capture the dynamic changes of learners' expressions losing the continuous features in the learning process, and its recognition effect is easily interfered with by factors such as occlusion and lighting variations during learning. To address the above issues, a network model based on adaptive global attention and temporal difference is proposed to recognise learners' dynamic expression sequences. Firstly, we have designed an Adaptive Global Attention (AGA) block, which adaptively models inter-channel relationships to dynamically enhance key channels that are highly correlated with learners' states while suppressing redundant information, thereby improving the model's feature representation capability under noisy environments. Secondly, we have designed a Differential Temporal Transformer (DTFormer) to extract differential information between consecutive frames, increasing the model's sensitivity to learners' facial expression dynamics and improving recognition performance. The two components complement each other in terms of spatial feature enhancement and temporal dynamic modelling effectively improving the model's overall capability for representing learners' dynamic facial expressions. Experiments were conducted on public datasets DFEW, FERV39k and the learner E-learning emotional state data set DAiSEE, and comparisons were made with classical methods using objective indicators. The results demonstrate that the proposed method outperforms the comparison methods in multiple performance indicators, thereby verifying its effectiveness.

1 | Introduction

With the continuous advancement of artificial intelligence technology, educational methods are undergoing profound changes, and the concept of smart education has been widely recognised. As an important part of the smart education field, expression recognition technology realises accurate perception of learners' emotions by analysing facial expressions during learning, which is an indispensable component in smart classrooms. Emotion is a key factor in the learning process. Learners' emotional states are usually reflected through their expressions, and variations in emotional states exert an impact on their learning motivation and learning outcomes [1]. Therefore, using expression recognition technology to analyse learners' facial expressions in the classroom and explore their learning processes and emotional changes can help teachers better understand learners' classroom states, thereby improving teaching efficiency and quality.

Over the past few years, learners' facial expression recognition has emerged as a critical technique for evaluating learning

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2026 The Author(s). *CAAI Transactions on Intelligence Technology* published by John Wiley & Sons Ltd on behalf of The Institution of Engineering and Technology and Chongqing University of Technology.

states, and numerous scholars have conducted extensive research on expression recognition algorithms. At present, most expression recognition tasks for learners' learning states rely on feature extraction from single-frame images of learners' faces [2]. However, in real-world classroom learning scenarios, learner facial expression recognition faces challenges from specific facial occlusion behaviours, such as resting the chin on the hand and wearing glasses as well as significant pose variations caused by activities such as note-taking and problem-solving. Learners' learning states are a process rather than an instant. Therefore, although static image-based expression recognition methods are efficient, they lack motion information of learners' expressions and cannot well reflect learners' states. To better adapt to complex learning scenarios, we choose to perform dynamic expression recognition based on image sequences of learners' learning expressions, which can better reflect the changes in learners' expressions during learning.

However, current dynamic image sequence-based expression recognition also faces challenges. Traditional convolutional neural networks (CNNs) [3] struggle to capture long-term dependencies and dynamic evolution patterns in consecutive frames [4]. Moreover, indiscriminate extraction of spatial and channel-wise features during facial expression recognition tends to introduce background noise leading to inefficient resource utilisation and insufficient discriminative features, which may cause key facial regions to be overlooked and ultimately degrade recognition performance. In addition, undifferentiated and uninhibited extraction of expression features leads to resource waste and insufficient features, affecting the final recognition effect. Unlike generic dynamic facial expression recognition tasks, learners' expressions in real learning scenarios exhibit distinctive characteristics, such as subtle emotional variations and short expression durations. Moreover, from a spatial perspective, key facial regions often yield weak responses while redundant features dominate the representation; from a temporal perspective, local variations (e.g., head movements) coexist with long-term emotional states. These characteristics render traditional facial expression recognition methods based on static features or single-frame modelling inadequate for accurately characterising learners' true learning states.

To address the aforementioned challenges of facial expression recognition in learning scenarios, this paper proposes a dynamic facial expression recognition method for learners based on Adaptive Global Attention (AGA) and Differential Temporal Transformer (DTFormer). The proposed approach focuses on two complementary dimensions—selective spatial–channel modelling and temporal differential perception—to achieve precise extraction and effective modelling of discriminative facial expression features in real-world learning environments. The main contributions of this work are summarised as follows:

1. We have proposed an Adaptive Global Attention (AGA) mechanism with a gated fusion strategy. To address the weak responses of key facial features and the presence of various interferences in learning scenarios, the AGA module is designed with two complementary attention branches: adaptive channel attention and global convolutional attention. Specifically, the adaptive channel attention branch captures local feature variations by modelling fine-grained inter-channel dependencies and employs a gated fusion mechanism to adaptively integrate global and local channel information, enabling dynamic allocation of channel-wise feature weights and deep modelling of spatial–channel correlations. The global convolutional attention branch extracts global contextual semantic information to ensure the stability of global feature representations. In addition, a temporal global pooling operation is introduced to jointly model channel-wise and temporal information enhancing the sensitivity of channel responses to inter-frame variations. As a result, the proposed AGA module selectively emphasises salient features while suppressing noncritical information under complex conditions such as pose variations and facial occlusions, thereby effectively enhancing the discriminative power of features.
2. We have constructed a Differential Temporal Transformer (DTFormer). Given that learners' facial expressions exhibit intertwined short-term rapid changes and long-term emotional states over time, a bidirectional temporal enhancement module is introduced before the Transformer encoder. This module explicitly models the temporal dynamics of facial expressions through forward and backward frame difference operations between adjacent frames, whereas depthwise separable convolutions are incorporated to enhance local temporal modelling capability. Temporal feature reshaping is thus completed prior to the multi-head self-attention mechanism. The proposed DTFormer not only preserves the Transformer's strength in capturing long-range temporal dependencies, but also significantly improves the model's sensitivity to short-term dynamic variations, leading to enhanced temporal feature extraction and effective collaborative modelling of short-term facial changes and long-term emotional states in learning scenarios.
3. To assess the validity of the proposed method in the task of dynamic facial expression recognition for learners, we utilise the DAiSEE dataset, which contains affective state data collected from learners in online learning environments. In addition, the large-scale dynamic facial expression datasets DFEW and FERV39k contain samples with extreme illumination conditions, self-occlusions and significant pose variations, which effectively simulate the influences of lighting variations, facial occlusions and head movements in real classroom learning scenarios. Therefore, we have conducted experiments on the DFEW and FERV39k datasets using Weighted Average Recall (WAR) and Unweighted Average Recall (UAR) as evaluation metrics to validate the effectiveness of the proposed method. Compared with other network models, it demonstrates excellent recognition performance.

2 | Related Work

2.1 | Learner Expression Recognition

Different from traditional expression recognition, the classroom environment for facial expression recognition is more complex. Furthermore, some facial expressions classified in traditional

24682322, 0, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/cit2.70115 by Swansea University, Wiley Online Library on [17/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

facial expression classification are not easy to occur in a classroom setting. Therefore, for facial expression recognition in classroom scenarios, it is necessary to redefine facial expression categories according to the characteristics of the application scenario, while excluding those facial expression categories that are irrelevant to the classroom [5, 6]. Nowadays, there is no unified concept for the classification of learners' classroom emotions. Savchenko et al. analysed learners' learning behaviours in online classrooms and divided learners' expressions into anger, contempt, disgust, fear, happiness, neutrality, sadness and surprize [7]. Gupta et al. collected learners' online learning videos, constructed the DAiSEE dataset and divided it into four emotional states: boredom, confusion, engagement and frustration [8].

With the in-depth research and wide application of deep learning methods, learner classroom expression recognition has made significant progress in accuracy and speed. In their study, Tang et al. introduced a classroom evaluation system utilising CNN-based object recognition and further designed an emotion recognition model by eliminating fully connected layers and combining depthwise separable convolutions with other components [9]. A deep attention network was utilised by Yu et al. to construct both a learner expression recognition system and a teaching evaluation algorithm [10]. Firstly, multipath facial images were generated through preprocessing methods such as cropping and occlusion and input into multipath deep attention networks respectively; then, different weights were assigned to the multipath networks through the self-attention mechanism, and constraint loss functions were used to limit weight assignment; finally, expression classification was performed based on the learnt global features to realise learner classroom expression recognition under occlusion. By combining the varying states of multiple facial tissues, Li et al. evaluated learners' states, they also developed a video-based classroom evaluation technique, leveraging facial features—including head angle, eyebrows, eyes and lips—to interpret learners' expressions and evaluate classroom effectiveness [11]. Krishnnan et al. come up with a new algorithm framework using SSIM to identify key frames in videos to improve recognition rate and combined facial expressions with drowsiness detection to perceive learners' learning states [12]. The SFER-MDFAE model, proposed by Shou et al., achieved a significant improvement in the accuracy of learner facial expression recognition within smart classroom scenarios by leveraging multiscale feature aggregation and fine-grained regional attention mechanisms [13]. Chen et al. proposed an attention fusion-based occluded facial expression recognition model (AFNet), which employs a multibranch spatial attention network to extract local facial features, automatically perceive occluded facial regions and address the occlusion issue in online learning [14]. Additionally, Chen et al. derived learners' emotional states from classroom expressions and cognitive feedback from behaviours subsequently integrating these two aspects to generate a comprehensive evaluation of classroom states [15]. Wang et al. used vision-language models to analyse 5000 images depicting confused, distracted, happy, neutral and tired expressions through zero shot prompt analysis, achieving moderate performance in academic facial expression recognition [16].

The existing learner expression recognition methods mentioned above are mostly static expression recognition methods based on images, but static expression recognition methods are difficult to adapt to complex expression recognition scenarios. Therefore, we choose to perform dynamic expression recognition on learners' expressions. In addition, since most current learner expression datasets are private, we select the abovementioned public dataset DAiSEE, which classifies learners' expressions into four categories: boredom, confusion, engagement and frustration, and perform preprocessing such as frame processing, face detection and alignment on the learners' online learning videos in this data set to confirm the effectiveness of the proposed method in learner expression recognition.

2.2 | Dynamic Expression Recognition

Currently, static image-based facial expression analysis remains the mainstream in the field of expression analysis. Due to the similarity between different facial expression emotions, Lee et al. proposed a divide-conquer learning strategy to precisely classify plenty of facial expression data, thereby more effectively assisting in extracting expression features [17]. Luo et al. proposed a facial expression recognition network integrated with a multiscale fusion attention mechanism aiming to enhance the representation capability of critical facial features [18]. Song et al. proposed HFE-Net, which achieves collaborative modelling of local and global facial expression information through parallel fusion of convolutional features with a multi-head self-attention mechanism, thereby enhancing the model's ability to perceive overall facial expression structures [19].

With advances in deep learning, particularly the use of temporal neural networks and 3D convolutional models, research has moved from static to dynamic facial expression recognition aiming to more effectively capture the evolution of facial expressions. In comparison with static facial expression recognition (SFER), dynamic facial expression recognition (DFER) methods take into account the spatial features of single-frame images as well as the dependencies and temporal information among frames in video sequences [20, 21]. Early studies mostly used CNN models to extract spatial features from each frame of images, then used RNN models to analyse the temporal relationships between image frames and later proposed 3D-CNN-based methods for 3D data modelling and joint learning of spatial-temporal features. Liu et al. proposed a dynamic expression recognition framework based on a hybrid attention mechanism, aiming to balance the relationship between local changes and global feature representation of facial expressions [22]. An et al. proposed channel-adaptive model parameters for the initialisation problem of CNN and LSTM [23], which can effectively reduce gradient explosion during training and further improve the accuracy of expression classification and recognition. In recent years, with the successful application of Transformers in the field of computer vision [24], temporal modelling methods based on self-attention mechanisms have gradually emerged as an important research direction in dynamic facial expression recognition (DFER), and researchers have begun

24682322, 0, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/cit2.70115 by Swansea University, Wiley Online Library on [17/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

utilising the structural merits of such models for extracting spatial-temporal information from video sequences [25]. Former-DFER, designed by Zhao et al., utilises a Transformer architecture combining CS-Former and T-Former modules to separately learn spatial and temporal characteristics in dynamic facial expression recognition [26]. Ma et al. developed a Spatial-Temporal Transformer aimed at capturing key intra-frame details, exploring frame-to-frame relationships and fusing spatial and temporal information in a unified manner [27]. Wang et al. designed a CNN-Transformer model, integrating Spa-CNN for spatial processing and T-Former for temporal modelling, to capture global temporal relations at consistent spatial positions across frames while maintaining feature resolution [21, 28]. Beyond architectural innovations, several studies have explored dynamic facial expression recognition from the perspectives of robustness and representation consistency, aiming to alleviate the impact of noisy frames and improve cross-domain generalisation in real-world scenarios. Li et al. introduced NR-DFERNet to suppress the influence of noise frames in video sequences [29]. Wei et al. proposed a super image-based spatiotemporal convolutional model and a two-stream LSTM model for facial expressions to capture local spatial-temporal features and learn global temporal clues of emotional changes [20]. Han et al. proposed the RTT model, which utilises an IR50 network pretrained on large-scale face datasets to extract static facial expression features and further incorporates a Transformer-based Temporal Feature Enhancement Module (TFEM) to strengthen the perception of complex dynamic expression variations [30]. Lu et al. proposed the Multi-Snippet Spatiotemporal Learning (MSSL) framework, which integrates (2 + 1)D multi-snippet spatiotemporal modelling, the BTMSE module, and a Temporal Transformer module to hierarchically capture subtle expression dynamics and long-term temporal dependencies [31]. Inspired by the reference encoding paradigm in neuroscience in neuroscience, Stettler et al. proposed the multi-domain neural reference encoding (MD-NRE) model, which represents facial expressions as offset vectors relative to domain-specific neutral references. This design enables effective cross-domain facial expression recognition in few-shot scenarios, significantly improving data efficiency and generalisation ability [32]. Focussing on the complementary roles of global and local information, Jiang et al. adopted a dual-path modelling strategy to separately learn structural features and high-frequency texture details and achieved collaborative optimisation through multi-scale feature fusion, thereby mitigating structural bias and excessive information smoothing commonly observed in single-network architectures for complex visual tasks [33]. Shi et al. proposed a Dual-Spike Self-Attention (DSSA) mechanism and designed the SpikingResformer architecture, which integrates the hierarchical structure of ResNet with spike-based self-attention. This approach preserves multilevel local feature extraction while introducing global attention modelling, effectively mitigating the limitations of pure attention-based architectures in local representation learning [34].

Despite the notable progress achieved by the aforementioned methods in improving dynamic facial expression recognition, several limitations remain. On the one hand, most existing channel- or spatial-attention mechanisms rely on coarse-grained operations such as global average pooling or max pooling. In particular, traditional attention modules such as SE and CBAM generate channel weights through a single-path global pooling strategy. In learning scenarios where learners' facial expression variations are subtle and discriminative regions are highly localised, such designs tend to overlook fine-grained spatial information within channels leading to weakened responses of critical local expression features and the loss of essential spatial positional cues [35, 36]. On the other hand, some approaches primarily focus on global temporal modelling and assume equal-quality temporal features before entering the temporal modelling stage resulting in insufficient attention to short-term temporal patterns and rapid local variations. Consequently, these methods struggle to effectively capture subtle dynamic expression changes [37, 38]. In addition, spatiotemporal features across different temporal scales often lack effective collaborative modelling mechanisms, which further limits the robustness and generalisation ability of models in complex real-world scenarios.

To address the above limitations, this paper proposes a learner dynamic facial expression recognition method based on fine-grained feature modelling and multiscale temporal representation, which integrates Adaptive Global Attention and Differential Temporal Transformer. An improved fine-grained channel attention mechanism is introduced, which incorporates a dual-branch attention structure and a gated fusion strategy to model channel-wise relationships between global contextual semantics and local fine-grained features. Learnable gating weights are employed to achieve adaptive fusion, enabling the model to maintain global perceptual capability while significantly enhancing responses to discriminative channel features. Meanwhile, a bidirectional temporal enhancement module is combined with a Temporal Transformer to collaboratively model short-term temporal patterns and long-term emotional states, thereby improving the model's sensitivity to facial expression variations and enabling effective recognition of complex dynamic expressions in real-world learning scenarios.

3 | Methodology

The overall architecture of the proposed learner dynamic expression recognition method based on Adaptive Global Attention and Differential Temporal Transformer is shown in Figure 1. The model adopts ResNet18 as the spatial feature extraction backbone, which consists of four residual stages (Block1–Block4). To mitigate the interference of noise and redundant information commonly present in classroom environments, Adaptive Global Attention (AGA) modules are inserted into the second and third residual stages of ResNet18. These modules enhance channel responses that are highly relevant to facial expression discrimination while suppressing redundant features, thereby providing reliable and discriminative representations for subsequent temporal modelling. Furthermore, a Differential Temporal Transformer (DTFormer) is introduced to perform differential temporal modelling on the feature sequences refined by AGA, enabling the model to focus on both short-term evolution and long-term dependencies of facial expression features and to capture the temporal progression of learners' emotional states. The collaborative and complementary design of AGA and DTFormer allows the proposed model to effectively address the challenges of subtle expression

24682322, 0, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/cit2.70115 by Swansea University, Wiley Online Library on [17/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

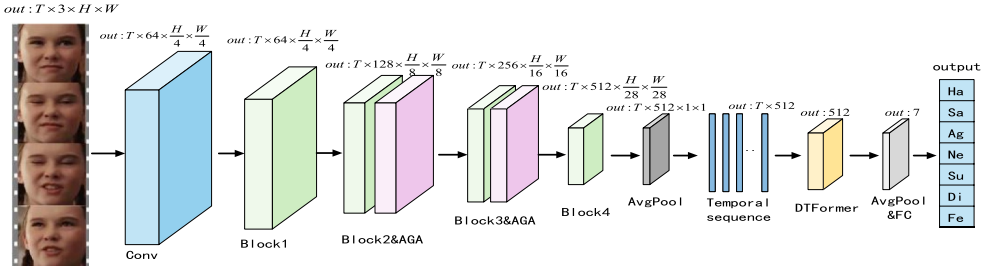

FIGURE 1 | Overall architecture of the proposed learner dynamic facial expression recognition network derived from Adaptive Global Attention (AGA) and Differential Temporal Transformer (DTFormer). The AGA module dynamically fuses global and local information to emphasise salient features, whereas the DTFormer captures subtle temporal changes through inter-frame difference operations.

variations and temporal instability in real-world classroom scenarios.

Firstly, the expression video sequence is preprocessed to obtain an image sequence with a fixed number of frames; the ResNet18 backbone network serves to extract features from the input image sequence, and then the adaptive global attention module is employed to obtain global and local features. The feature map is flattened in the time dimension to obtain a new temporal feature sequence, and the differential temporal Transformer module is used to model the temporal dependencies between frames to obtain temporal features; finally, the fully connected network generates the final classification output yielding the learner's dynamic facial expression recognition results.

3.1 | Backbone Network for Image Feature Extraction

The image feature extraction network uses the ResNet18 network as the backbone to extract image features and obtain basic global features. The network takes a frame set composed of T RGB face images with height H and width W as input, which is obtained by random dynamic sampling of segments from the original video. To be more specific, all frames in the training video samples are first evenly divided into U segments, and then V frames are randomly chosen from each segment. For the test video samples, all frames are initially divided into U segments, but then V frames are sequentially selected at the middle position of each segment [39]. Thus, frames are generated as input for training or testing.

3.2 | Adaptive Global Attention Mechanism (AGA)

In traditional learner classroom facial expression recognition tasks, most methods rely on convolutional neural networks (CNNs) to extract facial features and then adopt global pooling or single-branch channel attention mechanisms for global feature modelling. However, in real-world scenarios, learners' facial expression variations are often subtle. Conventional attention mechanisms based on a single global pooling pathway are therefore insufficient to simultaneously capture global emotional states and fine-grained expression changes, which limits the discriminative capability of the model.

To address this issue, we have proposed a dual-branch adaptive channel attention mechanism that differs from conventional single-branch channel attention mechanisms as illustrated in Figure 2. Specifically, this mechanism consists of two complementary attention branches: an Adaptive Channel Attention (ACA) module and a Global Convolutional Attention (GCA) module [40]. The ACA branch models fine-grained inter-channel dependencies to capture subtle local feature variations, whereas the GCA branch focuses on extracting global contextual semantic information. By jointly modelling local and global channel relationships, the proposed mechanism enhances sensitivity to fine-grained features while preserving global perceptual awareness, thereby improving recognition accuracy. Furthermore, considering that most existing channel attention mechanisms lack explicit temporal modelling capability, we introduce global convolution and temporal global pooling operations to integrate channel-wise features with temporal dynamics. This design enables dynamic adjustment of channel responses across frames, endowing the model with cross-frame channel selection capability and facilitating the capture of stable and discriminative facial expression evolution patterns.

3.2.1 | Adaptive Channel Attention (ACA)

This module obtains channel correlation features and channel compression expressions through local convolution and fully connected operations, and adaptively adjusts attention weights using the attention degree between channels and the gated fusion mechanism, finally obtaining channel attention to achieve accurate enhancement of emotion-related channels.

Specifically, to capture feature information in the channel dimension, we initially perform Global Average Pooling (GAP) and Global Max Pooling (GMP) operations on the network feature $X \in \mathbb{R}^{C \times H \times W}$ with input dimension $C \times H \times W$ to obtain global and local feature representations at the channel level as shown in Equations (1) and (2):

$$f_{\text{avg}} = GAP(X_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} X_c(i,j) \in \mathbb{R}^{T \times C \times 1 \times 1}, \quad (1)$$

$$f_{\max} = GMP(X_c) = \max_{1 \le i \le H, 1 \le j \le W} X_c(i,j) \in \mathbb{R}^{T \times C \times 1 \times 1}. \quad (2)$$

where f_{avg} and $f_{\max}$ are the global channel feature representation and local significant feature representation respectively,

24682322, 0, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/cit2.70115 by Swansea University, Wiley Online Library on [17/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

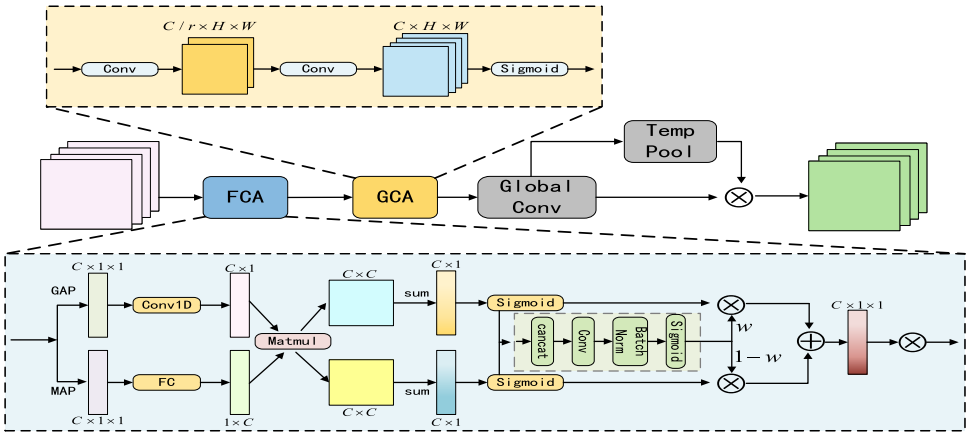

FIGURE 2 | The proposed Adaptive Global Attention (AGA) module diagram. The module is composed of adaptive channel attention and global convolutional attention and employs gated fusion to model cross-channel and temporal dependencies, dynamically emphasising salient features and suppressing irrelevant information.

and $X_c(i,j)$ represents the feature response of the cth channel at spatial position (i,j).

For the global average feature f_{avg} extracted by GAP, local modelling is performed using a 1D convolution operation to obtain $l_1 \in \mathbb{R}^{C\times 1}$; for the local significant feature $f_{\max}$ extracted by GMP, global modelling is performed through a fully connected operation to obtain $l_2 \in \mathbb{R}^{C\times 1}$. For the purpose of promoting effective interplay between global and local information, the obtained global and local information are combined, and a cross-similarity matrix is constructed by matmul product of their transposes to capture their correlation as shown in Equations (3) and (4):

$$A = l_1 \cdot l_2^T \in \mathbb{R}^{T\times C\times C}, \tag{3}$$

$$B = l_2 \cdot l_1^T \in \mathbb{R}^{T\times C\times C}. \tag{4}$$

Subsequently, to enable effective collaboration between different attention branches, an adaptive gated fusion mechanism is designed to dynamically regulate the integration of fine-grained features and global semantic features at the channel level. Since the channel correlation matrix contains rich inter-channel dependency information, row-wise and column-wise statistical information is first extracted from the channel correlation matrices A, B, respectively, to generate global and local weight vectors a, b. These vectors serve as prior weights that quantify the importance of global contextual semantics and local fine-grained responses.

Building upon this, a learnable gating parameter vector is introduced, which is normalised through a Sigmoid activation function to produce channel-wise gating parameters constrained within the range [0,1] as formulated in Equation (5):

$$w = \sigma(Conv1D([a,b])) \in [0,1]. \tag{5}$$

here, $w \in \mathbb{R}^{1\times C\times 1\times 1}$ denotes the learnable gating parameters vector, and $\sigma(\cdot)$ represents the Sigmoid activation function. This design ensures the continuity and differentiability of the fusion weights facilitating stable network training. Subsequently, the generated gating coefficients are used to perform adaptive weighted fusion of the fine-grained features and global features. The computation process is formulated as Equation (6):

$$f_{att} = w \times a + (1-w) \times b \in \mathbb{R}^{T\times C\times 1\times 1}. \tag{6}$$

here, $\otimes$ denotes element-wise multiplication across channels, and f_{att} represents the channel attention vector obtained after adaptive fusion. To further constrain the magnitude of channel responses and enhance numerical stability, the fused channel attention vector is normalised again using a Sigmoid activation function. It is then applied to the original input feature map to achieve channel-wise feature recalibration as formulated in Equation (7):

$$X^c = X \otimes \sigma(f_{att}) \in \mathbb{R}^{T\times C\times H\times W}. \tag{7}$$

Through the above process, the model can adaptively emphasise more discriminative feature representations across different samples and channels. It selectively highlights informative features while suppressing noncritical ones, enabling more precise weighting of facial expression-related features. As a result, the proposed mechanism effectively enhances both the accuracy and robustness of facial expression recognition in complex classroom scenarios.

3.2.2 | Global Convolutional Attention (GCA)

To boost the model's spatial representation capability for key facial regions of learners, a global convolutional attention module is adopted to model dependencies between various spatial positions in facial images. It can not only suppress relatively unimportant channels but also retain feature position information through global convolution, which is crucial for improving the performance of learner expression recognition.

For the feature map X^c adjusted through the channel attention described earlier, global average pooling and global maximum pooling operations are first executed in the channel dimension. This results in two distinct spatial feature descriptions z_{avg} and $z_{\max}$ as presented in Equations (8) and (9), which not only

24682322, 0, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/cit2.70115 by Swansea University, Wiley Online Library on [17/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

realise the aggregation of feature maps in the channel dimension but also emphasise spatial position-related information.

$$z_{avg} = GAP(X^c) = \frac{1}{C}\sum_{i=1}^{C} X^c(i) \in \mathbb{R}^{T\times 1\times H\times W}, \tag{8}$$

$$z_{\max} = GMP(X^c) = \max_{1\le i\le C} X^c(i) \in \mathbb{R}^{T\times 1\times H\times W}. \tag{9}$$

Then, the above two spatial feature descriptions are concatenated in the channel dimension, and a convolution layer is used for feature fusion and dimensionality reduction to further extract feature relationships in the spatial dimension as shown in Equation (10):

$$z_{att} = Conv\left(\left[z_{avg}, z_{\max}\right]\right) \in \mathbb{R}^{T\times 1\times H\times W}. \tag{10}$$

The result processed by the convolution layer is transmitted through a Sigmoid activation function to obtain spatial attention weights, which are multiplied by the input feature map to obtain the output feature map X^s as shown in Equation (11):

$$X^s = X^c \otimes \sigma(z_{att}) \in \mathbb{R}^{T\times C\times H\times W}. \tag{11}$$

Finally, to achieve dynamic adjustment of channel response, global convolution is introduced to fuse channel and temporal information as shown in Equations (12) and (13):

$$z_{gc} = \sigma(Conv_{1\times k}(X) + Conv_{k\times 1}(X)) \in \mathbb{R}^{T\times C\times 1\times 1}, \tag{12}$$

$$X^t = z_{gc} \otimes GAP_t\left(z_{gc}\right) \in \mathbb{R}^{T\times C\times H\times W}. \tag{13}$$

where z_{gc} is the feature representation obtained after global convolution, and GAP_t represents the global pooling operation in the temporal dimension. To retain the original input features and improve expressive ability, a residual connection is ultimately incorporated to obtain the final attention output feature map $M \in \mathbb{R}^{T\times C\times H\times W}$. Although learning key regions, global or local information is not lost, thereby further improving the accuracy of expression recognition.

3.3 | Differential Temporal Transformer (DTFormer)

In previous studies, most dynamic facial expression recognition (DFER) methods relied on optical flow-based handcrafted features. Although such methods can capture inter-frame motion information, they are typically loosely coupled with the backbone network and lack end-to-end adaptive learning capabilities. Some approaches introduced temporal positional encoding or temporal attention mechanisms within Transformers [26]; however, they still mainly perform global modelling, which is insufficient for capturing instantaneous expression variations.

Although the multi-head self-attention mechanism in Transformers enables direct global temporal modelling and long-term dependency extraction, it tends to overlook short-term temporal information. In classroom learning scenarios, learners' emotional states are often reflected in rapid facial muscle movements or instantaneous expressions over short time periods, which possess strong short-term temporal characteristics and are critical for emotion discrimination. To fully leverage short-term temporal features and enhance sensitivity to expression dynamics, we have introduced a differential temporal enhancement module prior to the Transformer encoder. Unlike existing differential methods, such as LG-TDL or five-frame differencing techniques [41, 42], which either regard temporal differences as auxiliary constraints without directly participating in temporal modelling or operate independently of global temporal modelling, these approaches often fail to capture the holistic dynamics of facial expressions and introduce additional parameter overhead. For dynamic expression recognition tasks, relying solely on either short-term temporal changes or auxiliary constraints is insufficient to simultaneously capture both short-term and long-term temporal dependencies.

To this end, we have proposed the Differential Temporal Transformer (DTFormer), which leverages forward and backward temporal differencing as a temporal feature enhancement module for Transformer-based temporal modelling. This module captures bidirectional changes between adjacent frames and employs a learnable gating mechanism to adaptively regulate the differential information at a fine-grained scale. The reshaped temporal features are then fed into the multi-head self-attention mechanism, guiding the model to naturally assign higher weights to regions with significant motion. Consequently, DTFormer not only retains the advantages of Transformers in global temporal modelling but also enhances the extraction of short-term temporal features, enabling effective collaborative modelling of short-term facial dynamics and long-term temporal dependencies in learner expressions.

As shown in Figure 3, the module comprises differential temporal enhancement module and M temporal encoders, realising a differential temporal Transformer module with short-term temporal modelling capability.

First, a convolution layer is used for local context modelling, and a global average pooling operation is performed to generate frame-level image representations x_t, so that the representations of all frames form a sequence feature $X' = \{x'_1, x'_2, ..., x'_T\} \in \mathbb{R}^{T\times C}$, that is, T spatial feature vectors are obtained. As shown in Equations (14) and (15), the feature representation of the feature vector at any time t is denoted as X'_t, and forward and backward frame shift operations are performed on it to obtain forward shifted features and backward shifted features:

$$\vec{X}'_t = Roll\left(X'_t, +s\right), \tag{14}$$

$$\overleftarrow{X}'_t = Roll\left(X'_t, -s\right). \tag{15}$$

where $Roll(\cdot)$ represents the cyclic shift operation of time frame positions, s is the displacement step size, defaulting to 1 and $\vec{X}'_t$, $\overleftarrow{X}'_t$ represent forward displacement by one frame and backward displacement by one frame, respectively, and the module preserves the sequence length without introducing additional parameters, thereby avoiding potential interference from padding on the temporal structure. Subsequently,

24682322, 0, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/cit2.70115 by Swansea University, Wiley Online Library on [17/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

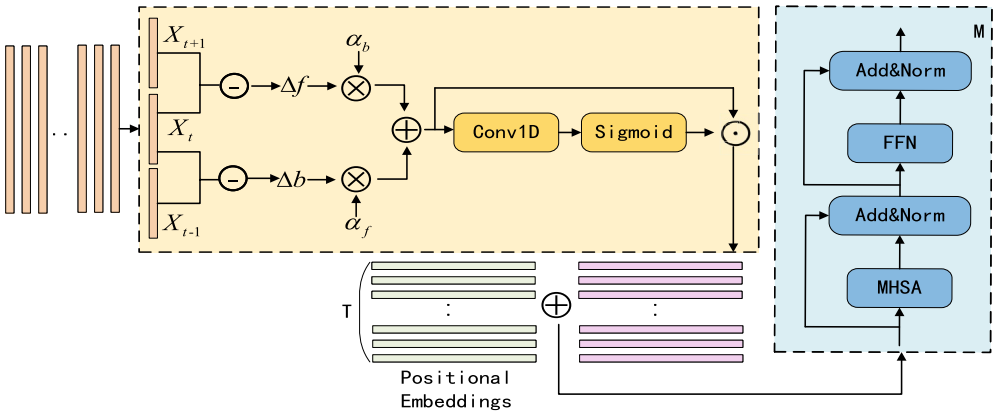

FIGURE 3 | The proposed Differential Temporal Transformer (DTFormer) module diagram. The module utilises inter-frame difference operations combined with depthwise separable convolutions to capture facial expression dynamics, improving temporal feature sensitivity and model robustness.

bidirectional differential branches are constructed to capture dynamic changes as formulated in Equations (16) and (17):

$$\Delta f = \vec{X}_t' - X_t', t = 1, ..., T-1, \tag{16}$$

$$\Delta b = X_t' - \overleftarrow{X}_t', t = 2, ..., T. \tag{17}$$

where Δf, Δb represent forward difference and backward difference, respectively, and the forward difference represents the feature change from the current frame to the next frame, whereas the backward difference captures the feature changes from the previous frame to the current frame. This approach not only preserves the variations between adjacent frames but also helps suppress background bias, thereby better capturing the dynamic patterns of inter-frame changes. To further enhance the model's ability to adaptively model dynamic changes in both directions, learnable weighting parameters α_f, α_b are introduced to perform adaptive fusion of the forward and backward differential features as shown in Equation (18):

$$\Delta = \alpha_f \cdot \Delta f + \alpha_b \cdot \Delta b \in \mathbb{R}^{T \times C}. \tag{18}$$

where α_f, α_b are learnable parameters, initialised to 0.5 and used to balance the contribution of differential features in different directions to the representation at the current time step. To further model local temporal consistency and suppress abrupt changes caused by noise, a 1D depthwise separable convolution is applied to the fused differential features, and the Δ is processed by 1D convolution in the temporal dimension as shown in Equation (19):

$$\Delta' = Conv1D_{dw}(\Delta) \in \mathbb{R}^{T \times C}. \tag{19}$$

where the convolution kernel size is set to 3 and symmetric padding is applied to preserve the temporal dimension, allowing each channel to independently learn its local temporal variation patterns and enhancing temporal representation capability.

Considering that differential features may contain irrelevant dynamics caused by head movements or background changes, a gating mechanism is further introduced to selectively regulate the differential information. The Sigmoid activation function is used as the gating function to generate a gating weight matrix G as shown in Equation (20):

$$G = \sigma(\Delta') \in \mathbb{R}^{T \times C}. \tag{20}$$

This gating tensor adaptively adjusts the retention degree of dynamic features at each moment within the interval [0,1], thereby realising feature-level temporal information control. Finally, as shown in Equation (21), the original input and temporal differential enhanced features are fused with residuals through a gating mechanism to obtain the module output.

$$X_{t'}' = X_t' + G \odot \Delta \in \mathbb{R}^{T \times C}. \tag{21}$$

here $\odot$ denotes element-wise multiplication. Through this residual fusion approach, the model can selectively preserve regions with key dynamic changes based on the magnitude of temporal variations in the input, effectively suppressing irrelevant or noisy information. The above differential method is applied before the Transformer as a short-term dynamic enhancement layer, which enhances the Transformer's sensitivity to short-term dynamic facial expressions.

Subsequently, the temporal features $X_{t'}'$ obtained from the Differential Temporal Enhancement module are combined with learnable temporal positional encodings and used as the input to the temporal encoder as shown in Equation (22):

$$Z = X_{t'}' + e_{pos}. \tag{22}$$

where e_{pos} represents a learnable position embedding for encoding temporal positions, $t' \in \{0, 1, ..., T\}$. The obtained $Z_{t'}$ is input into the temporal encoder to obtain the attention output as shown in Equations (23) and (24):

$$D_{t'}^{(l,k)} = Softmax\left(\frac{Q_{t'}^{(l,k)} K_{t'}^{(l,k)^T}}{\sqrt{C'}}\right), \tag{23}$$

$$S_{t'}^{(l,k)} = B^{(l,k)} V^{(l,k)}. \tag{24}$$

where $C' = C/k^s$ head is the channel dimension of each attention head, k^s is the total number of attention heads and l is the lth encoder. The outputs of all attention heads are

24682322, 0, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/cit2.70115 by Swansea University, Wiley Online Library on [17/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

concatenated and the dimension is restored through linear mapping, and the output $Z_{t'}^{l}$ is obtained by combining residual connections. After normalisation, it is input into the feedforward neural network for nonlinear transformation, and the representation of this layer's spatial encoder is output as shown in Equation (25):

$$Z_{t'}^{L} = MLP\left(LN\left(Z_{t'}^{l}\right)\right) + Z_{t'}^{l}. \tag{25}$$

where $MLP(\cdot)$ is a feedforward neural network composed of two fully connected layers and a GELU function. The Transformer output is obtained after M encoders. Finally, the cross-entropy loss function is used as a supervised training method to acquire the ultimate classification outcome.

4 | Experiments

For the purpose of evaluating the effectiveness of the proposed method, experimental parameters and datasets are first given, ablation experiments are performed, and comparisons are made with other advanced dynamic facial expression recognition methods.

4.1 | Experimental Datasets

To assess the performance of the proposed learner dynamic facial expression recognition model, experiments were conducted on two large-sized public dynamic facial expression datasets: DFEW [43] and FERV39k [44]. Both datasets were collected from a large number of movie clips across multiple countries and contain various real-world scenarios, effectively simulating diverse and highly challenging conditions, including extreme lighting, self-occlusion and unpredictable head movements as shown in Figure 4. They can appropriately replicate factors such as lighting, occlusion and head motion in real classroom scenarios. In addition, both datasets are annotated with seven basic expression categories: happiness, sadness, neutral, anger, surprize, disgust and fear.

In addition, the DAiSEE dataset [8] is used to validate the effectiveness of the method in recognising learner states in learning scenarios. Designed specifically for affective computing in educational settings, the DAiSEE dataset supports multilevel engagement analysis. It is a multi-label video classification dataset composed of 9068 video clips captured from 112 users, including video recordings of subjects in e-learning environments with four types of labels: engagement, frustration, confusion and boredom. Each emotion is divided into four levels: very low, low, high and very high [45, 46].

To utilise this dataset in the present study, we followed the processing method proposed by Zhu [47]. The engagement category was assigned the lowest priority, as it is considered a baseline in educational scenarios and is only labelled when other more meaningful emotional signals are absent. Accordingly, the multi-label DAiSEE dataset was converted into a single-label task following these rules: for each video sample, the emotion with the highest intensity level was selected as the dominant label. In cases where multiple dimensions share the same intensity, a priority ranking based on educational relevance from high to low is applied: frustration, confusion, boredom and engagement. To preserve the characteristics of real classroom scenarios while mitigating class imbalance, all frustration and confusion samples from the original dataset were retained, whereas boredom and engagement samples were randomly undersampled. In the final experimental subset, the proportion of boredom: engagement: confusion: frustration is approximately 3:1:1:1. The processed dataset information is summarised in Table 1, which alleviates the original distribution bias of high engagement and low negative emotions, thereby enhancing the model's ability to discriminate features of minority classes.

FIGURE 4 | Partial images of DFEW, FERV39k and DAiSEE datasets.

TABLE 1 | Sample counts by emotion category for the DAiSEE dataset: Before and after processing (unit: Samples).

	Number of training samples		Number of validation samples		Number of test samples	
Emotion category	**Before processing**	**After processing**	**Before processing**	**After processing**	**Before processing**	**After processing**
Boredom	849	259	264	91	470	139
Engagement	4163	582	1301	204	1040	312
Confusion	275	275	90	90	106	106
Frustration	194	194	68	68	104	104

24682322, 0, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/cit2.70115 by Swansea University, Wiley Online Library on [17/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

4.2 | Data Preprocessing

In real classroom scenarios, learner dynamic facial expression recognition is significantly affected by complex lighting conditions, partial occlusions and variations in head pose. To ensure that experimental results accurately reflect the model's performance in practical applications, a systematic preprocessing and data augmentation strategy adopted. All operations were performed on video frame groups as the fundamental unit, ensuring spatial transformations are consistent within the same sequence and preserving the temporal continuity required for effective temporal modelling.

4.2.1 | Video Processing and Geometric Alignment

To eliminate geometric differences caused by head rotation, pitch or spatial shifts and to ensure structural consistency of input faces, a standardised processing pipeline was applied for different datasets. For publicly available datasets with already aligned faces (DFEW and FERV39k), the officially provided aligned data were used directly, and all face images were resized to 112× 112 using bilinear interpolation as model input. For the DAiSEE dataset, which consists of raw classroom videos, frames were first extracted at fixed intervals using FFmpeg. Then, MTCNN was applied to detect faces in each frame, and the face region was cropped and aligned based on bounding boxes and facial keypoints. To ensure completeness and clarity of the face region, cropped images were first resized to 224× 224 and then scaled to 112× 112 as the final model input.

4.2.2 | Data Augmentation for Complex Environments

To simulate uncontrollable factors in real classroom scenarios and avoid overfitting to idealised distributions, a frame group-based augmentation strategy implemented. By maintaining consistent spatial transformations across frames within the same sequence, the temporal continuity of facial expression evolution was effectively preserved. Random gamma correction and colour jittering were applied to dynamically adjust image brightness, contrast and saturation, simulating lighting changes during different periods of a class. Multiscale cropping with fixed offset sampling and horizontal flipping were also employed to simulate visual deformations caused by head pitch and involuntary forward or backward leaning, enhancing the model's ability to capture facial features under nonfrontal viewing angles.

4.3 | Training Strategy

All experiments for every dataset and model were conducted using the PyTorch deep learning framework on a single NVIDIA GeForce RTX 4060 GPU (16 GB VRAM). Data loading was performed with 4 parallel threads (num_workers = 4), and the batch size was set to 40 to balance computational efficiency and stable gradient updates.

For DFEW and FERV39k, the models were optimised using the stochastic gradient descent (SGD) algorithm with an initial learning rate of 1×10^{-3}, a momentum of 0.9 and a weight decay of $1e-4$ to enhance regularisation and prevent overfitting in complex video backgrounds. The learning rate was scheduled using the StepLR policy, decaying by 50% every 7 epochs, with a total training duration of 100 epochs. For DAiSEE, considering the extreme class imbalance, a weighted random sampling mechanism implemented during training. Instead of conventional linear decay, a cosine annealing with warm restarts strategy employed for learning rate scheduling with an initial period of $T_0 = 10$ and a period multiplication factor of $T_{mult} = 2$. For DFEW and FERV39k, we followed previous work and adopted the unweighted average recall (UAR, i.e., the accuracy of each class divided by the number of classes, with no consideration of the number of instances in each class) and weighted average recall (WAR, i.e., accuracy) as metrics [21, 28]. For the DAiSEE dataset, accuracy serves as the evaluation metric following previous research methods.

4.4 | Ablation Studies

To validate the effectiveness of each module in the proposed method, we conducted ablation experiments on the ResNet18 backbone by progressively introducing different temporal modelling and attention mechanisms. Model performance was evaluated across DFER datasets using UAR and WAR as metrics with results summarised in Table 2.

Comparing ResNet18 + TFormer with ResNet18 + DTFormer, it can be observed that incorporating differential temporal modelling improves WAR from 67.42% to 68.13%, indicating that differential temporal features can more effectively capture dynamic information during facial expression changes positively contributing to emotion classification. Further introducing channel attention mechanisms, although SE and CBAM provide stable performance improvements under the ResNet18 + DTFormer framework, the use of the AGA module combined with the differential temporal Transformer achieves even more significant gains in both UAR and WAR. Moreover, when AGA and DTFormer are directly combined without the gated fusion mechanism, the performance improvement is limited, suggesting that simply stacking attention and differential temporal modelling is insufficient to fully exploit their complementary advantages. The complete model achieves the best performance with WAR reaching 69.30%, demonstrating that the AGA module and

TABLE 2 | Ablation study results on the DFEW dataset (w/o gate = without gated fusion strategy).

Setting	Metrics (%)	
	UAR	WAR
Resnet18 + TFormer	53.56	67.42
Resnet18 + DTFormer	54.22	68.13
Resnet18 + CBAM + DTForme	55.14	68.65
Resnet18 + SE + DTForme	55.35	68.79
Resnet18 + AGA + TForme	54.98	68.51
Resnet18 + AGA + DTFormer (w/o gate)	54.67	68.44
Resnet18 + AGA + DTFormer	55.63	69.30

24682322, 0, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/cit2.70115 by Swansea University, Wiley Online Library on [17/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

differential temporal Transformer can form an effective synergy under a proper fusion mechanism. Their complementarity enhances the model's robustness in capturing dynamic facial expression features in complex, nonideal classroom scenarios.

4.5 | Comparative Experiments

TTo confirms the validity of the proposed method in classroom expression recognition, and comparative experiments are conducted not only on DFEW and FERV39k datasets with other models but also on the DAiSEE dataset. The comparative analysis of the proposed method and other methods on the three datasets is shown in Tables 3–5. For previously published studies, we directly cited the reported performance on the respective datasets. For models that did not provide results on the DAiSEE dataset, we re-implemented them based on their official source code under the same preprocessing and evaluation metrics.

TABLE 3 | Comparison of different methods on the FERV39k dataset.

	Metrics (%)	
Method	**UAR**	**WAR**
C3D [44]	22.68	31.69
I3D-RGB [48]	30.17	38.78
VGG13 + LSTM [49]	32.41	43.37
Resnet18 + LSTM [50]	30.92	42.59
Former-DFER [25, 26]	37.20	46.85
NR-DFERNet [25, 29]	33.99	45.97
NSNP-DFER [51]	37.80	46.40
MSSL [31]	39.27	48.01
TG-DFER [52]	41.50	51.67
AGDTNet (ours)	35.25	47.08

4.5.1 | FERV39k

The recognition results of different methods on the FERV39k dataset are summarised in Table 3. Methods based on 3D convolution or static features (e.g., C3D, I3D-RGB) exhibit relatively limited performance in complex real-world scenarios, whereas methods incorporating temporal modelling mechanisms (e.g., LSTM or Transformer architectures) achieve more consistent performance gains in both UAR and WAR. In terms of weighted average recall (WAR), the proposed method outperforms most baseline approaches, for example, it surpasses ResNet18 + LSTM by 4.49% and NR-DFERNet by 1.11% but still trails behind some weakly supervised or self-supervised methods (e.g., TG-DFER). This indicates that there is still room for improvement in the model's class-level balanced recognition especially for classes with fewer samples.

The confusion matrix of our method on FERV39k is illustrated in Figure 5 where diagonal elements represent the proportion of correctly classified samples for each emotion category, and off-diagonal elements reflect inter-class confusion. It can be observed that emotions with more pronounced expressions, such as happiness and sadness, are recognised effectively, whereas noticeable confusion occurs among surprize, disgust and fear mainly because these emotions exhibit subtle facial changes and similar expressions, which are easily influenced by individual differences. These results further indicate that, although the proposed method demonstrates stable overall

TABLE 5 | Comparison of different methods on the DAiSEE dataset.

Method	**Accuracy (%)**
I3D [54]	51.03
ResNet + LSTM [55]	58.24
Former-DFER [26]	59.83
LSTPNet [56]	61.37
AGDTNet (ours)	62.56

TABLE 4 | Comparison of different methods on the DFEW dataset.

	Accuracy of each emotion (%)							Metrics (%)	
Method	**Happy**	**Sad**	**Neutral**	**Angry**	**Surprise**	**Disgust**	**Fear**	**UAR**	**WAR**
C3D [44]	75.17	39.49	55.11	62.49	45.00	1.38	20.51	42.74	53.54
I3D-RGB [48]	78.61	44.19	56.69	55.87	45.88	2.07	20.51	43.40	54.27
VGG11 + LSTM [49]	76.89	37.65	58.04	60.70	43.70	0.00	19.73	42.39	53.70
Resnet18 + LSTM [50]	78.00	40.65	53.77	56.83	45.00	4.14	21.62	42.86	53.08
Former-DFER [26]	84.05	62.57	67.52	70.03	56.43	3.45	31.78	53.69	65.70
NR-DFERNet [29]	88.47	64.84	70.03	75.09	61.60	0.00	19.43	54.21	68.19
LOGO-former [38]	85.39	66.52	68.94	71.33	54.59	0.00	32.71	54.21	66.98
Dual-STI [39]	87.65	68.32	67.55	71.70	57.93	8.97	36.14	56.89	68.29
NSNP-DFER [51]	87.93	65.08	62.48	71.89	62.46	0.00	40.56	56.05	68.02
MSSL [31]	—	—	—	—	—	—	—	59.86	68.58
ST_RDGCN [53]	89.57	69.92	62.17	79.31	62.24	20.69	30.39	59.18	69.37
TG-DFER [52]	90.92	74.92	71.41	73.36	59.43	13.10	38.03	60.17	71.62
AGDTNet (ours)	88.96	68.07	67.79	76.27	65.99	0.00	22.35	55.63	69.30

24682322, 0, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/cit2.70115 by Swansea University, Wiley Online Library on [17/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

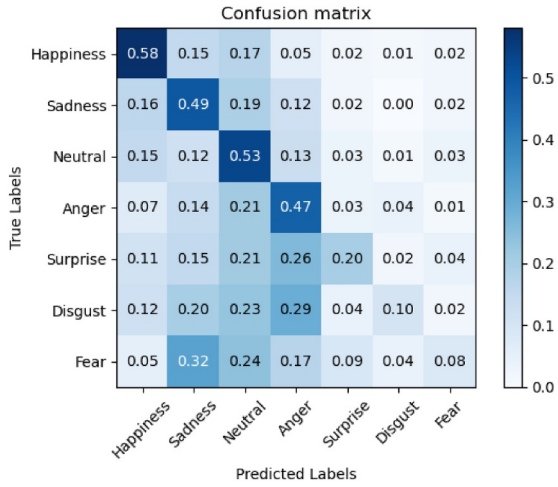

FIGURE 5 | Confusion matrix on the FERV39k dataset.

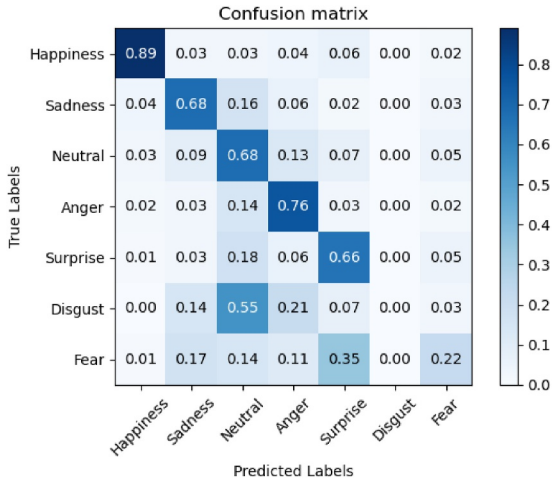

FIGURE 6 | Confusion matrix on the DFEW dataset.

recognition performance, there remains room for improvement in fine-grained discrimination of low-intensity emotions.

4.5.2 | DFEW

The recognition results of different methods on the DFEW dataset are presented in Table 4. The proposed method achieves a Weighted Average Recall (WAR) of 69.30%, outperforming ResNet18 + LSTM by 17.15%, Former-DFER by 4.53% and LOGO-Former by 2.04% among others. For emotions with more pronounced facial changes, such as happiness and anger, most methods achieve relatively high recognition accuracy. In contrast, for categories like Neutral and Fear, which involve subtler facial muscle movements, recognition performance is generally lower across methods. Although the model maintains strong overall recognition capability under class-imbalanced conditions, the accuracy for subtle emotions such as Disgust and Fear still requires improvement.

The confusion matrix of our method on the DFEW dataset is illustrated in Figure 6. The accuracy for the Happy class reaches 89%, and for Angry it exceeds 70%. However, confusion remains between categories such as Disgust and Fear with Disgust often misclassified as Neutral. This is primarily because these emotions exhibit subtle facial changes in real-world scenarios, and the facial action patterns across different emotions share certain similarities, increasing the difficulty of discrimination.

4.5.3 | DAiSEE

The recognition results on the DAiSEE dataset are shown in Table 5. The proposed AGDT method achieves an overall accuracy of 62.56%, outperforming comparative methods such as I3D and Former-DFER, which demonstrates the effectiveness of our approach for dynamic facial expression recognition in real learning scenarios. As shown in the confusion matrix in Figure 7, there are differences in recognition accuracy across categories. The Boredom, Engagement and Frustration classes achieve relatively high accuracies of 65%, 56% and 76%, respectively, whereas the Confusion class has a slightly lower

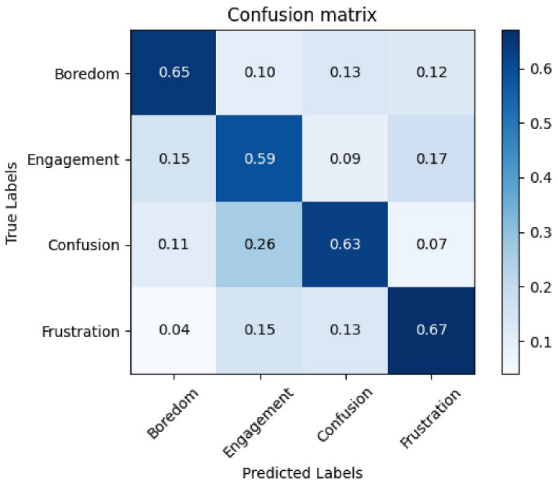

FIGURE 7 | Confusion matrix on the DAiSEE dataset.

accuracy of 63%. This is mainly because Confusion exhibits partial facial similarity with Boredom or Engagement leading to misclassifications in some samples.

To address this, our method redefines the labels, randomly samples and balances the sample numbers resulting in improved recognition performance. For categories with subtle expression differences that are easily confused, some misclassifications still occur. This indicates that, although the proposed method effectively captures dynamic facial expression features in real learning scenarios and improves overall performance, its recognition capability for subtle emotional categories still has room for improvement providing guidance for future model enhancements.

4.5.4 | Complexity Analysis

The proposed method was compared with representative baseline approaches in terms of model complexity and inference efficiency as summarised in Table 6. Model complexity was evaluated using parameter count (Params), floating-point operations (FLOPs) and inference time. FLOPs were calculated

24682322, 0, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/cit2.70115 by Swansea University, Wiley Online Library on [17/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 6 | Comparison of model complexity and inference efficiency among representative methods.

Method	Params (M)	FLOPs (G)	Inference time (ms)
C3D [44]	78	39	—
Former-DFER [26]	18	9	8
NR-DFERNet [29]	—	6	—
MAE-DFER [57]	85	50	—
AGDTNet (ours)	19	6	11

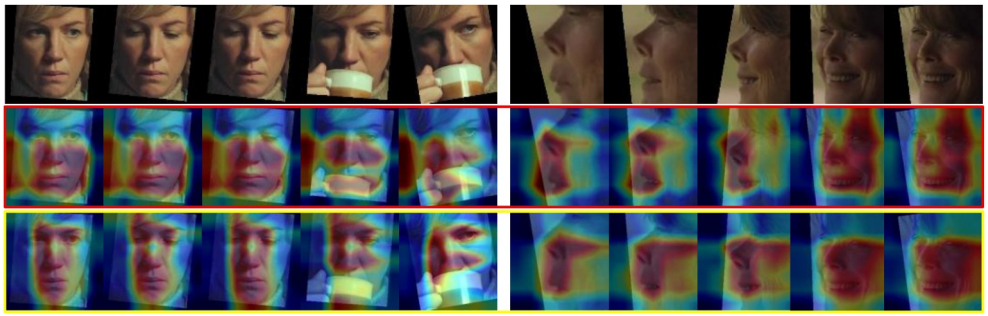

FIGURE 8 | Attention visualisation. The feature maps within the yellow boxes correspond to the outputs of the complete model, whereas the red boxes show the feature maps of the model without the gated fusion mechanism.

per video clip, and inference time was measured on a single-GPU setup.

The results show that the proposed model has 19.10 M parameters and a computational complexity of 5.98 GFLOPs placing it at a moderate scale. Compared with some baselines that rely heavily on large-scale temporal modelling, the proposed model achieves strong recognition performance while maintaining relatively low computational cost. In terms of inference efficiency, under a batch size of 1, the average GPU inference time is 11.71 ms making it suitable for offline analysis and near real-time emotion recognition in classroom scenarios. These results indicate that the proposed method achieves a reasonable trade-off between recognition performance and computational efficiency.

4.6 | Visual Analysis

To further validate the model's ability to focus on key discriminative regions in dynamic facial expression recognition under real classroom scenarios, visual analyses were conducted on samples with occlusion and pose variations. As illustrated in Figure 8, the first row displays the original video frames, the second row shows visualisation results of the model without the gated fusion mechanism and the third row presents results of the complete model.

Without the gated fusion mechanism, the model's attention responses are relatively scattered with some activations spreading to background regions or nondiscriminative facial contours and exhibiting unstable spatial variations across adjacent frames. In contrast, the complete model concentrates consistently on subtle facial regions such as the eyes, eyebrows, mouth and other facial muscles effectively highlighting key expression changes. This indicates that the model can suppress redundant and noisy features, enhancing selective modelling of critical facial regions and improving the discriminability and robustness of dynamic expression recognition. However, in cases of weak expression intensity or severe occlusion, attention responses may still show some dispersion, suggesting that future work should explore more robust feature modelling strategies.

5 | Conclusion

This paper proposes a learner expression recognition model based on adaptive global attention and differential temporal Transformer. The model introduces adaptive global attention to selectively emphasise informative features and suppress noncritical features and adds a differential temporal Transformer module to enhance sensitivity to expression changes, thereby improving recognition performance. However, in real classroom scenarios, this task still faces significant challenges. When learners' expressions are weak, accompanied by large head pose variations or substantial occlusions, local facial cues can be easily diminished. Under extreme interference conditions, the model's attention may still become somewhat dispersed affecting prediction stability. Furthermore, due to the lack of explicit annotations for demographic attributes such as skin colour and ethnicity in existing public classroom datasets, this study could not perform grouped evaluations to assess potential individual biases. Future work will consider incorporating more diverse datasets with fairness annotations to further improve the model's robustness and fairness across different populations and real-world scenarios. Additionally, given the limited size of publicly available learner expression video datasets, imbalanced class distributions and the constraints of single-modality information, future research could explore multi-modal inputs (e.g., audio, text) or self-supervised learning

24682322, 0, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/cit2.70115 by Swansea University, Wiley Online Library on [17/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

frameworks to mitigate the dependency on large-scale labelled data and further enhance recognition performance.

Acknowledgements

Authors are funded by UKRI (Grant EP/W020408/1) and Grant RS718 through Swansea University. This work was supported by the Humanities and Social Science Fund of Ministry of Education of China (23YJAZH084).

Funding

The study was supported by UKRI (Grant EP/W020408/1); the Humanities and Social Science Fund of Ministry of Education of China (23YJAZH084).

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

Data available on request from the authors.

References

1. J. Li, E. Xue, C. Li, and Y. He, "Investigating Latent Interactions Between Students' Affective Cognition and Learning Performance: Meta-Analysis of Affective and Cognitive Factors," *Behavioral Sciences* 13, no. 7 (2023): 555, https://doi.org/10.3390/bs13070555.

2. Q. Yuan, "Research on Classroom Emotion Recognition Algorithm Based on Visual Emotion Classification," *Computational Intelligence and Neuroscience* 2022, no. 1 (2022): 6453499, https://doi.org/10.1155/2022/6453499.

3. L. Peng, S. Baoye, and X. Lin, "Bearing Fault Diagnosis in Multiple Working Conditions Based on One Dimensional Convolutional Neural Network," *Journal of Shandong University of Science and Technology* 42, no. 5 (2023): 88–96, https://doi.org/10.16452/j.cnki.sdkjzk.2023.05.010.

4. Q. Xiaohe, L. Zhiwei, L. Da, D. Xiao, H. Kaixun, and Z. Kai, "Fault Diagnosis Method of Tansfer Learning Based Multi-Scale Graph Convolutional Networks," *Journal of Shandong University of Science and Technology* 44, no. 5 (2025): 119–129, https://doi.org/10.16452/j.cnki.sdkjzk.2025.05.012.

5. J. Wang, "Research on Classroom Emotion Recognition Algorithm Based on Visual Emotion Classification," (In Chinese) *Educational Theory and Research* 2, no. 4 (2024): 35–36, https://doi.org/10.61369/etr.6415.

6. B. Fang, X. Li, G. Han, and J. He, "Facial Expression Recognition in Educational Research From the Perspective of Machine Learning: A Systematic Review," *IEEE Access* 11 (2023): 112060–112074, https://doi.org/10.1109/ACCESS.2023.3322454.

7. A. V. Savchenko, L. V. Savchenko, and I. Makarov, "Classifying Emotions and Engagement in Online Learning Based on a Single Facial Expression Recognition Neural Network," *IEEE Transactions on Affective Computing* 13, no. 4 (2022): 2132–2143, https://doi.org/10.1109/TAFFC.2022.3188390.

8. A. Gupta, A. D'Cunha, K. Awasthi, and V. Balasubramanian, "Daisee: Towards User Engagement Recognition in the Wild," preprint arXiv:1609.01885 (2016), https://doi.org/10.48550/arXiv.1609.01885.

9. J. Tang, X. Zhou, and J. Zheng, "Design of Intelligent Classroom Facial Recognition Based on Deep Learning," *Journal of Physics: Conference Series* 1168, no. 2 (2019): 022043, https://doi.org/10.1088/1742-6596/1168/2/022043.

10. W. Yu, M. Liang, and X. Wang, "Learner Expression Recognition and Intelligent Teaching Evaluation Algorithm Based on Deep Attention Network," (In Chinese) *Journal of Computer Applications* 42, no. 3 (2022): 743–749, https://doi.org/10.11772/j.issn.1001-9081.2021040846.

11. H. Li, L. Yang, Z. Weijia, and S. Peixuan, "Analysis of Teaching Effect Based on Facial Expression in Classroom Environment," *Modern Distance Education Research* 30, no. 4 (2017), https://doi.org/10.1145/3383923.3383949.

12. N. Krishnnan, S. Ahmed, T. Ganta, and G. Jeyakumar, "A Video Analytics Based Solution for Detecting the Attention Level of the Students in Class Rooms," in *2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, (2020), 498–501.

13. Z. Shou, Y. Huang, D. Li, et al., "A Student Facial Expression Recognition Model Based on Multi-Scale and Deep Fine-Grained Feature Attention Enhancement," *Sensors* 24, no. 20 (2024): 6748, https://doi.org/10.3390/s24206748.

14. Y. Chen, K. Li, F. Tian, G. Wei, and M. Seberi, "Lightweight Expression Recognition Combined Attention Fusion Network With Hybrid Knowledge Distillation for Occluded e-Learner Facial Images," *Neurocomputing* 628 (2025): 129656, https://doi.org/10.1016/j.neucom.2025.129656.

15. Z. Chen, M. Liang, Z. Xue, and W. Yu, "STRAN: Student Expression Recognition Based on Spatio-Temporal Residual Attention Network in Classroom Teaching Videos," *Applied Intelligence* 53, no. 21 (2023): 25310–25329, https://doi.org/10.1007/s10489-023-04858-0.

16. D. Wang, C. Yang, and G. Chen, "Using Vision Language Models to Detect Students' Academic Emotion Through Facial Expressions," preprint arXiv:2506.10334 (2025), https://doi.org/10.48550/arXiv.2506.10334.

17. D. H. Lee and J. H. Yoo, "CNN Learning Strategy for Recognizing Facial Expressions," *IEEE Access* 11 (2023): 70865–70872, https://doi.org/10.1109/ACCESS.2023.3294099.

18. S. Luo, M. Li, and M. Chen, "Multi-Scale Integrated Attention Mechanism for Facial Expression Recognition Network," *Computer Engineering and Applications* 59, no. 1 (2023): 199–206, https://doi.org/10.3778/j.issn.1002-8331.2203-0170.

19. D. Song and C. Liu, "A Facial Expression Recognition Network Using Hybrid Feature Extraction," *PLoS One* 20, no. 1 (2025): e0312359, https://doi.org/10.1371/journal.pone.0312359.

20. J. Wei, G. Hu, X. Yang, A. T. Luu, and Y. Dong, "Learning Facial Expression and Body Gesture Visual Information for Video Emotion Recognition," *Expert Systems with Applications* 237 (2024): 121419, https://doi.org/10.1016/j.eswa.2023.121419.

21. L. Wang, "Research on Facial Expression Recognition Based on Deep Learning," in *2019 3rd International Conference on Electronic Information Technology and Computer Engineering (EITCE)* (IEEE, 2019), 1144–1147, https://doi.org/10.1109/EITCE47263.2019.9095140.

22. X. Liu, X. Gong, and H. Zhao, "Dynamic Face Expression Recognition Based on Mixed Attention Mechanism (In Chinese)," supplement, *Journal of Computer Applications* 43, no. S1 (2023): 1–7, https://doi.org/10.11772/j.issn.1001-9081.2022101472.

23. F. An and Z. Liu, "Facial Expression Recognition Algorithm Based on Parameter Adaptive Initialization of CNN and LSTM," *Visual Computer* 36, no. 3 (2020): 483–498, https://doi.org/10.1007/s00371-019-01635-4.

24. X. Yan, L. Xianglan, P. Xuguang, L. Fang, and Z. Haiyan, "Research on Person Re-Identification Method Based on CNN-Transformer and Attention Pyramid," *Journal of Shandong University of Science and Technology* 44, no. 1 (2025): 110–118, https://doi.org/10.16452/j.cnki.sdkjzk.2025.01.010.

25. S. Yan, Y. Wang, X. Mai, et al., "Empower Smart Cities With Sampling-Wise Dynamic Facial Expression Recognition via Frame-

24682322, 0, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/cit2.70115 by Swansea University, Wiley Online Library on [17/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Sequence Contrastive Learning," *Computer Communications* 216 (2024): 130–139, https://doi.org/10.1016/j.comcom.2023.12.032.

26. Z. Zhao and Q. Liu, "Former-DFER: Dynamic Facial Expression Recognition Transformer," in *Proceedings of the 29th ACM International Conference on Multimedia* (Association for Computing Machinery, 2021), 1553–1561.

27. F. Ma, B. Sun, and S. Li, "Spatio-Temporal Transformer for Dynamic Facial Expression Recognition in the Wild," preprint arXiv: 2205.04749 (2022), https://doi.org/10.48550/arXiv.2205.04749.

28. L. Wang, X. Kang, F. Ding, S. Nakagawa, and F. Ren, "A Joint Local Spatial and Global Temporal CNN-Transformer for Dynamic Facial Expression Recognition," *Applied Soft Computing* 161 (2024): 111680, https://doi.org/10.1016/j.asoc.2024.111680.

29. H. Li, M. Sui, and Z. Zhu, "Nr-dfernet: Noise-Robust Network for Dynamic Facial Expression Recognition," preprint arXiv:2206.04975 (2022), https://doi.org/10.48550/arXiv.2206.04975.

30. T. Han, S. Dou, W. Zhang, and R. Liu, "Enhanced Dynamic Temporal Feature Extraction With Static Expression Insights for Dynamic Facial Expression Recognition," *Digital Signal Processing* 168 (2025): 105470, https://doi.org/10.1016/j.dsp.2025.105470.

31. Y. Lü, F. Zhang, Z. Ma, B. Zheng, and Z. Nan, "Dynamic Facial Expression Recognition in the Wild via Multi-Snippet Spatiotemporal Learning," *Neurocomputing* 636 (2025): 130020, https://doi.org/10.1016/j.neucom.2025.130020.

32. M. Stettler, A. Lappe, and M. A. Giese, "Facial Expression Recognition Based on Multi-Domain Norm-Referenced Encoding," *Neural Networks* 197 (2025): 108384, https://doi.org/10.1016/j.neunet.2025.108384.

33. K. Jiang, Z. Wang, P. Yi, T. Lu, J. Jiang, and Z. Xiong, "Dual-Path Deep Fusion Network for Face Image Hallucination," *IEEE Transactions on Neural Networks and Learning Systems* 33, no. 1 (2020): 378–391, https://doi.org/10.1109/TNNLS.2020.3027849.

34. X. Shi, Z. Hao, and Z. Yu, "Spikingresformer: Bridging Resnet and Vision Transformer in Spiking Neural Networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2024), 5610–5619.

35. X. Liu, H. Xu, and M. Wang, "Sparse Spatial-Temporal Emotion Graph Convolutional Network for Video Emotion Recognition," *Computational Intelligence and Neuroscience* 2022, no. 1 (2022): 3518879, https://doi.org/10.1155/2022/3518879.

36. R. Singh, S. Saurav, T. Kumar, R. Saini, A. Vohra, and S. Singh, "Facial Expression Recognition in Videos Using Hybrid CNN & ConvLSTM," *International Journal of Information Technology* 15, no. 4 (2023): 1819–1830, https://doi.org/10.1007/s41870-023-01183-0.

37. L. Wang, X. Kang, F. Ding, S. Nakagawa, and F. Ren, "MSSTNet: A Multi-Scale Spatio-Temporal CNN-Transformer Network for Dynamic Facial Expression Recognition," in *ICASSP 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2024), 3015–3019.

38. F. Ma, B. Sun, and S. Li, "Logo-Former: Local-Global Spatio-Temporal Transformer for Dynamic Facial Expression Recognition," in *ICASSP 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2023), 1–5.

39. M. Li, X. Zhang, C. Fan, T. Liao, and G. Xiao, "Dual-STI: Dual-Path Spatial-Temporal Interaction Learning for Dynamic Facial Expression Recognition," *Information Sciences* 678 (2024): 120953, https://doi.org/10.1016/j.ins.2024.120953.

40. S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional Block Attention Module," in *Proceedings of the European Conference on Computer Vision (ECCV)* (Springer International Publishing, 2018), 3–19.

41. Y. Xiao, Q. Yuan, K. Jiang, et al., "Local-Global Temporal Difference Learning for Satellite Video Super-Resolution," *IEEE Transactions on Circuits and Systems for Video Technology* 34, no. 4 (2023): 2789–2802, https://doi.org/10.1109/tcsvt.2023.3312321.

42. C. Huang, Q. Zeng, F. Xiong, and J. Xu, "Space Dynamic Target Tracking Method Based on Five-Frame Difference and Deepsort," *Scientific Reports* 14, no. 1 (2024): 6020, https://doi.org/10.1038/s41598-024-56623-z.

43. X. Jiang, Y. Zong, W. Zheng, et al., "Dfew: A Large-Scale Database for Recognizing Dynamic Facial Expressions in the Wild," in *Proceedings of the 28th ACM International Conference on Multimedia*, (2020), 2881–2889.

44. Y. Wang, Y. Sun, Y. Huang, et al., "Ferv39k: A Large-Scale Multi-Scene Dataset for Facial Expression Recognition in Videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2022), 20922–20931.

45. Y. Ma, Y. Wei, Y. Shi, X. Li, Y. Tian, and Z. Zhao, "Online Learning Engagement Recognition Using Bidirectional Long-Term Recurrent Convolutional Networks," *Sustainability* 15, no. 1 (2023): 198, https://doi.org/10.3390/su15010198.

46. M. M. Santoni, T. Basaruddin, and K. Junus, "Convolutional Neural Network Model Based Students' Engagement Detection in Imbalanced DAiSEE Dataset," *International Journal of Advanced Computer Science and Applications* 14, no. 3 (2023), https://doi.org/10.14569/IJACSA.2023.0140371.

47. D. Zhu, *Student Online Learning Dynamic Facial Expression Recognition Based on an Improved 3D Residual Neural Network(In Chinese)*. Master's thesis (Jiangxi University of Finance and Economics, 2025).

48. J. Carreira and A. Zisserman, "Quo Vadis, Action Recognition? a New Model and the Kinetics Dataset," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2017), 6299–6308.

49. K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," preprint arXiv:1409.1556 (2014).

50. K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2016), 770–778.

51. Z. Han, X. Meichen, P. Hong, L. Zhicai, and G. Jun, "NSNP-DFER: A Nonlinear Spiking Neural P Network for Dynamic Facial Expression Recognition," *Computers & Electrical Engineering* 115 (2024): 109125, https://doi.org/10.1016/j.compeleceng.2024.109125.

52. G. Jung, H. Kong, S. W. Lee, "Text-Guided Weakly Supervised Framework for Dynamic Facial Expression Recognition," *Pattern Recognition* (2025): 112910: SSRN 5282338, https://doi.org/10.48550/arXiv.2511.10958.

53. C. Huang, F. Jiang, Z. Han, et al., "Modeling Fine-Grained Relations in Dynamic Space-Time Graphs for Video-Based Facial Expression Recognition," *IEEE Transactions on Affective Computing* 16, no. 3 (2025): 1675–1692, https://doi.org/10.1109/taffc.2025.3530973.

54. H. Zhang, X. Xiao, T. Huang, S. Liu, Y. Xia, and J. Li, "An Novel End-to-End Network for Automatic Student Engagement Recognition," in *2019 IEEE 9th International Conference on Electronics Information and Emergency Communication (ICEIEC)* (IEEE, 2019), 342–345.

55. A. Abedi and S. S. Khan, "Improving State-of-the-Art in Detecting Student Engagement With Resnet and Tcn Hybrid Network," in *2021 18th Conference on Robots and Vision (CRV)* (IEEE, 2021), 151–157.

56. C. Lu, Y. Jiang, K. Fu, Q. Zhao, and H. Yang, "Lstpnet: Long Short-Term Perception Network for Dynamic Facial Expression Recognition in the Wild," *Image and Vision Computing* 142 (2024): 104915, https://doi.org/10.1016/j.imavis.2024.104915.

57. L. Sun, Z. Lian, B. Liu, and J. Tao, "Mae-Dfer: Efficient Masked Autoencoder for Self-Supervised Dynamic Facial Expression Recognition," in *Proceedings of the 31st ACM International Conference on Multimedia*, (2023), 6110–6121.

24682322, 0, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/cit2.70115 by Swansea University, Wiley Online Library on [17/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License