IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 43, No. 2, April, 2026, pp. 865-878
Journal homepage: http://iieta.org/journals/ts

Multimodal Image Processing and Fine-Grained Learning Behavior Recognition in Higher Education Smart Classrooms

Lipeng Wang

Shijiazhuang University of Applied Technology, Shijiazhuang 050081, China

Corresponding Author Email: 2003100270@sjzpt.edu.cn

https://doi.org/10.18280/ts.430225

Received: 1 November 2025
Revised: 10 February 2026
Accepted: 12 March 2026
Available online: 30 April 2026

Keywords:
multimodal image processing, smart classroom, fine-grained learning behavior recognition, cross-modal feature fusion, spatiotemporal graph convolution, deformable convolution, attention mechanism

ABSTRACT

The intelligent evolution of smart classrooms in higher education has created an urgent need for precise fine-grained learning behavior recognition. Multimodal imaging, leveraging complementary cross-modal information, has emerged as a key technical foundation. However, existing multimodal processing approaches remain limited, hindering deployment in real-world educational environments. To address these challenges, a unified framework for multimodal image processing and fine-grained learning behavior recognition in higher education smart classrooms was proposed, which consists of three core modules. First, a unified multimodal image preprocessing and enhancement module was designed, through which Red, Green, and Blue (RGB), depth, and thermal images were effectively optimized and spatially aligned. Second, a deformable convolution and cross-attention alignment fusion network was constructed, enabling accurate alignment and efficient fusion of heterogeneous cross-modal features. Third, a multi-scale spatiotemporal graph convolutional network was developed to achieve precise recognition of fine-grained learning behaviors by leveraging fused multimodal information. Experiments conducted on a self-constructed SmartClass-MM dataset demonstrated that superior performance was achieved in terms of image enhancement quality, fine-grained behavior recognition accuracy, and real-time efficiency compared with state-of-the-art methods. Significant improvements were observed in peak signal-to-noise ratio and structural similarity index. Furthermore, ablation studies confirmed the independent effectiveness of each core module. The proposed framework provides a novel and efficient solution for multimodal image processing and fine-grained learning behavior recognition in smart classrooms while offering valuable insights for cross-modal feature fusion and spatiotemporal behavior analysis in the image processing field, thereby facilitating the advancement of multimodal technologies in intelligent education.

1. INTRODUCTION

The digital transformation of higher education has driven the evolution of smart classrooms toward greater precision and personalization. Fine-grained learning behavior recognition [1, 2], as a core technology for perceiving classroom dynamics and optimizing instructional processes, has been regarded as a key determinant of the intelligence level of smart classrooms [3]. Fine-grained learning behaviors encompass a wide range of specific in-class actions, which serve as critical indicators for analyzing learners' attention and engagement. Consequently, substantial significance has been attributed to such behaviors for enabling personalized instructional guidance and improving overall teaching quality. Multimodal imaging has been increasingly recognized as an effective approach for fine-grained learning behavior recognition [4, 5]. Red, Green, and Blue (RGB) images [6] are utilized to capture appearance and texture information, depth images [7] are employed to characterize spatial structural features, and thermal images [8] are leveraged to reflect physiological states and attention distribution. Through the complementary integration of these modalities, the inherent limitations of single-modality data can be effectively overcome, thereby enhancing the comprehensiveness and robustness of behavior recognition.

Image processing techniques [9, 10], serving as the foundation of multimodal behavior recognition, play a pivotal role in smart classroom environments, as their performance directly influences the accuracy of subsequent feature fusion and behavior recognition. However, the complexity of smart classroom scenarios introduces several unique challenges to multimodal image processing and fine-grained learning behavior recognition. Variations in classroom lighting conditions, particularly under low-light environments, tend to result in increased image noise and blurred details. In addition, diverse sitting postures, frequent occlusions among learners, and sensor-induced noise further degrade image quality. Due to the heterogeneous nature of multimodal data acquired from different sensors, spatial misalignment across modalities is commonly observed. Meanwhile, the subtle differences among fine-grained learning behaviors make accurate discrimination difficult for conventional methods, while

limited model interpretability hinders the identification of critical factors underlying behavior recognition. These challenges collectively constrain the advancement of intelligent smart classrooms and limit the in-depth application of multimodal image processing technologies in the educational domain. Therefore, the investigation of multimodal image processing and fine-grained learning behavior recognition methods tailored for smart classrooms in higher education is of substantial theoretical importance and practical relevance.

Multimodal image enhancement serves as a fundamental step for improving the performance of behavior recognition. Existing approaches can generally be categorized into traditional methods and deep learning-based methods. Among traditional approaches, the Retinex algorithm [11, 12] enhances images by decomposing illumination and reflectance components; however, under low-light classroom conditions, a balance between noise suppression and edge preservation is difficult to achieve. The Contrast Limited Adaptive Histogram Equalization (CLAHE) algorithm [13, 14] is capable of improving image contrast, yet it is prone to local overexposure and lacks adaptability to the heterogeneous characteristics of multimodal images, thereby limiting its ability to simultaneously optimize RGB, depth, and thermal images. Cross-modal feature fusion [15-17] constitutes a critical stage for integrating multimodal information, with existing strategies broadly classified into early fusion, intermediate fusion, and late fusion. Early fusion methods, such as simple concatenation, fail to account for the heterogeneity of multimodal features, often resulting in feature redundancy and mutual interference. Intermediate fusion approaches driven by conventional attention mechanisms are capable of emphasizing salient features; however, spatial misalignment across modalities is not effectively addressed, and the complementary relationships among multimodal features are insufficiently exploited, thereby limiting the realization of synergistic advantages.

In the domain of fine-grained learning behavior recognition, most existing methods rely on single-modality images or skeletal keypoint representations [18, 19], while neglecting the integration of complementary multimodal information such as texture, depth, and thermal characteristics. Although certain studies have incorporated multimodal data, the temporal dynamics of behaviors have not been comprehensively captured, making it difficult to distinguish highly similar fine-grained actions. Furthermore, the model decision-making process lacks interpretability [20], thereby failing to satisfy the dual requirements of recognition accuracy and interpretability in smart classrooms. Overall, substantial limitations remain in current research, and a comprehensive solution tailored to the complex scenarios of smart classrooms has yet to be established, which constitutes the primary motivation for this study.

A multimodal image processing and fine-grained learning behavior recognition framework tailored for smart classrooms in higher education is developed to address the critical technical bottlenecks in multimodal image enhancement, cross-modal alignment and fusion, and fine-grained behavior recognition. Through this framework, multimodal image quality is optimized, heterogeneous features are efficiently integrated, and fine-grained learning behaviors are accurately recognized, while the requirements for real-time performance and model interpretability in classroom environments are simultaneously satisfied. This framework is expected to provide robust technical support for the intelligent advancement of smart classrooms. The main contributions are summarized below. A unified multimodal image preprocessing and enhancement module is designed to accommodate the complexity of classroom environments. By addressing the inherent limitations of RGB, depth, and thermal image modalities, synchronized optimization and spatial alignment of multimodal images are achieved, thereby establishing a reliable foundation for subsequent cross-modal feature fusion and behavior recognition. A deformable convolution and cross-attention alignment fusion network is constructed. Precise spatial alignment of multimodal features is accomplished through deformable convolution, while complementary relationships among features are effectively exploited via a cross-attention mechanism. A gated adaptive fusion strategy is further introduced to achieve unified feature representation, thereby resolving challenges associated with cross-modal heterogeneous feature alignment and fusion. A multi-scale spatiotemporal graph convolutional network is developed to integrate multimodal features with dynamic spatiotemporal graph information. Behavioral patterns at different temporal scales are captured through multi-scale spatiotemporal convolutions, while critical spatial regions and temporal segments are emphasized using a spatiotemporal attention mechanism. As a result, precise fine-grained learning behavior recognition is achieved, and model interpretability is enhanced. An end-to-end multi-task learning strategy combined with a multi-objective loss function is proposed to jointly optimize the three sub-tasks of image enhancement, feature fusion, and behavior recognition. The issue of class imbalance is alleviated through tailored loss functions, leading to improved overall model performance.

The remainder of this study is organized below. In Section 2, the tasks of multimodal image processing and fine-grained learning behavior recognition in smart classrooms are formally defined, and the core technical challenges are mathematically formulated along with their associated constraints. In Section 3, the proposed framework is described in detail, with particular emphasis on the design principles and technical implementation of each core module, constituting the central part of the study. In Section 4, the effectiveness of the proposed method is quantitatively validated through comparative experiments, ablation studies, visualization analyses, and robustness evaluations. In Section 5, the principal advantages and limitations of the proposed approach are critically discussed, and future research directions are outlined in light of emerging trends in the field. Finally, the conclusions are presented in Section 6, where the key findings and academic contributions are summarized.

2. PROBLEM FORMULATION

In smart classroom environments for higher education, multimodal image data are synchronously acquired using RGB cameras, Azure Kinect depth sensors, and thermal imaging devices. Continuous multimodal image sequences are thus constructed, in which RGB image sequences, depth image sequences, and thermal image sequences are generated through temporally consistent sampling by their respective sensors. Spatial correspondence is maintained across the three modalities, while temporal synchronization is preserved throughout the sequences. The task of fine-grained learning behavior recognition investigated in this study can be formally

defined as a mapping from synchronized trimodal image sequences to fine-grained behavior category labels. Under the condition of multimodal inputs, accurate behavior classification results are required to be produced. In addition, three critical constraints must be satisfied: real-time inference capability, robustness against environmental disturbances such as low illumination and occlusion, and interpretability characterized by transparent and well-defined decision-making logic. These requirements are essential for adapting to dynamic monitoring scenarios in real-world classroom environments.

To address this task, three core technical problems are required to be mathematically formulated and optimized. First, multimodal image enhancement is formulated as an optimization problem in which the feature discrepancy between the enhanced output and high-quality reference images is minimized, while constraints on edge preservation are imposed to achieve a balance between noise suppression and detail preservation. Second, cross-modal feature alignment is modeled through spatial consistency constraints, such that pixel-level spatial displacement errors among multimodal features are minimized, thereby mitigating misalignment caused by sensor disparity and object motion. Third, fine-grained learning behavior recognition is formulated as a nonlinear mapping from spatiotemporal feature representations to behavior category space. Through appropriate objective function design, class imbalance is alleviated, enabling accurate discrimination of highly similar fine-grained behaviors and effective modeling of temporal features.

3. PROPOSED METHOD

3.1 Overall framework overview

A multimodal image processing and fine-grained learning behavior recognition framework for smart classrooms in higher education is developed based on a hierarchically progressive three-layer architecture. These layers are designed to operate in a coordinated and interdependent manner, forming a closed logical loop that closely aligns with the overall pipeline, thereby enabling end-to-end processing from multimodal image input to fine-grained behavior recognition output. The multimodal image preprocessing and enhancement layer serves as the foundational input stage of the framework. RGB, depth, and thermal image sequences are received as inputs, and modality-specific deficiencies are systematically addressed through targeted optimization and spatial alignment. As a result, high-quality and spatially synchronized multimodal image sequences are generated, providing reliable data support for subsequent feature processing. The cross-modal feature alignment and fusion layer operates on the enhanced multimodal images. Through a dedicated network architecture, heterogeneous features are accurately aligned in the spatial domain and efficiently fused. The complementary information across modalities is effectively exploited, enabling the generation of unified cross-modal feature representations that simultaneously encode appearance texture, spatial structure, and physiological state information. The fine-grained learning behavior recognition layer functions as the core output stage of the framework. Spatiotemporal features are extracted from the fused representations, and critical information is selectively emphasized to achieve precise classification of fine-grained learning behaviors and interpretable decision-making. Ultimately, behavior category labels for the corresponding image sequences are output. Through the integration of an end-to-end multi-task learning strategy, the three-layer architecture is jointly optimized, effectively linking the processes of image enhancement, feature fusion, and behavior recognition. As a result, a balanced performance in terms of accuracy, real-time capability, and interpretability is achieved, ensuring adaptability to the complex and dynamic conditions of smart classroom environments.

3.2 Unified multimodal image preprocessing and enhancement module

The unified multimodal image preprocessing and enhancement module is designed to address typical challenges encountered in smart classroom environments, including low-light degradation in RGB images, occlusion-induced missing regions in depth images, and insufficient contrast in thermal image temperature distributions. An integrated framework combining modality-specific fine-grained optimization with global spatial calibration is constructed. While modality-dependent defects are selectively corrected, precise spatial alignment across modalities is simultaneously achieved, thereby providing high-quality data for subsequent cross-modal feature alignment and fusion. The overall workflow is illustrated in Figure 1. To mitigate low-light degradation in RGB images, a Retinex-based decomposition enhancement method incorporating multi-scale weighted guided filtering is developed. According to the illumination imaging model, the original image is decomposed into reflectance and illumination components, which can be expressed as:

$$I(x,y)=R(x,y)\cdot L(x,y) \tag{1}$$

where, $I(x,y)$ denotes the original low-light RGB image, $R(x,y)$ represents the reflectance component encoding intrinsic texture information, and $L(x,y)$ corresponds to the illumination component. Unlike conventional Retinex approaches based on fixed-scale filtering, large-scale guided filtering is utilized to estimate the global illumination distribution, and small-scale guided filtering is employed to capture local illumination details. Adaptive weighting is further introduced based on local pixel neighborhood characteristics, with the weighting function defined as:

$$\omega(x,y)=\lambda_1\cdot|\nabla G(x,y)|+\lambda_2\cdot T(x,y) \tag{2}$$

where, $|\nabla G(x,y)|$ denotes the gradient magnitude within the pixel neighborhood, $T(x,y)$ represents texture complexity, and λ_1 and λ_2 are normalized adaptive coefficients. Regions with pronounced edges and rich textures are assigned higher weights. Finally, the enhanced image is reconstructed using the optimized reflectance component, enabling effective noise suppression while preserving critical details such as human posture and object boundaries.

A multi-stage joint restoration and enhancement strategy is adopted for depth images. Sensor noise is first suppressed using bilateral filtering while spatial structural edges are preserved. Subsequently, small-scale holes are filled through morphological closing operations. Finally, large-area depth missing regions caused by occlusion or sensor failure are restored using a region-filling algorithm, resulting in depth

feature maps with spatial continuity and structural completeness. The optimization of thermal images is achieved through the combination of adaptive histogram equalization and pseudo-color mapping. The dynamic range of the temperature distribution is first expanded using adaptive histogram equalization, thereby enhancing temperature differences in critical regions such as the face and hands. Pseudo-color mapping is then applied to transform grayscale temperature information into high-discriminability color features, while the original temperature values are retained for subsequent attention-based feature extraction.

After modality-specific optimization has been completed, spatial alignment across modalities is achieved through scale normalization and pixel coordinate calibration. RGB, depth, and thermal images are unified into a common spatial resolution and pixel coordinate system, thereby eliminating spatial misalignment caused by sensor placement disparities and temporal acquisition offsets. As a result, precise pixel-level correspondence among the three modalities is ensured, effectively removing spatial discrepancies for subsequent cross-modal feature fusion.

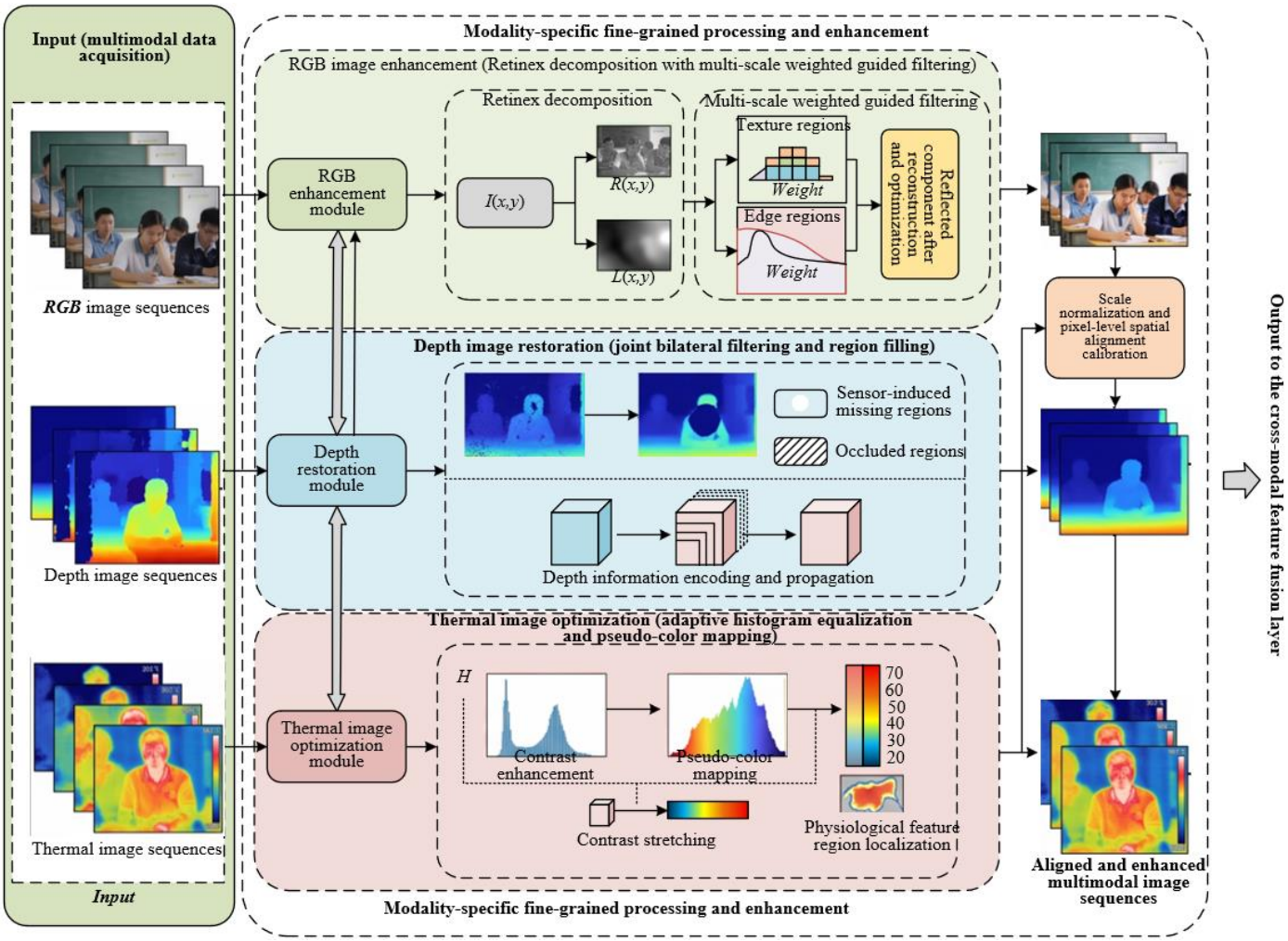

Figure 1. Workflow of the unified multimodal image preprocessing and enhancement module

3.3 Deformable convolution and cross-attention alignment fusion network

To address the challenges of low alignment accuracy and insufficient exploitation of complementary information caused by multimodal feature heterogeneity, a deformable convolution and cross-attention alignment fusion network is developed. RGB features are adopted as the reference baseline, and non-rigid spatial alignment is achieved through deformable convolution. In parallel, a cross-attention mechanism is employed to capture complementary information across modalities. Finally, a gated adaptive fusion strategy is applied to generate unified cross-modal feature representations, thereby providing high-quality feature support for subsequent fine-grained learning behavior recognition. The detailed architecture is illustrated in Figure 2.

The proposed application of deformable convolution extends beyond its conventional use in object detection and is adapted for cross-modal feature alignment. Non-rigid matching is achieved by learning pixel-level two-dimensional spatial offsets. The offset learning process is formulated as:

$$\Delta p_{ij}=W_{offset}\cdot F_{cat}(F_{ref},F_{src}) \tag{3}$$

where, Δp_{ij} denotes the two-dimensional spatial offset of the j-th sampling point in the i-th convolution kernel, W_{offset} represents the learnable offset weights, F_{ref} denotes the RGB reference feature map, and F_{src} represents the source feature map to be aligned (i.e., depth or thermal features). The operator F_{cat} denotes channel-wise feature concatenation. The offset is adaptively learned by jointly considering the local correlations between source and reference features as well as

image gradient information. This enables the sampling points of the features to be aligned to undergo non-rigid deformation in response to variations in human poses and object boundaries within classroom scenes. Consequently, cross-modal spatial misalignment caused by sensor disparity and student movement is effectively eliminated, achieving pixel-level precise alignment.

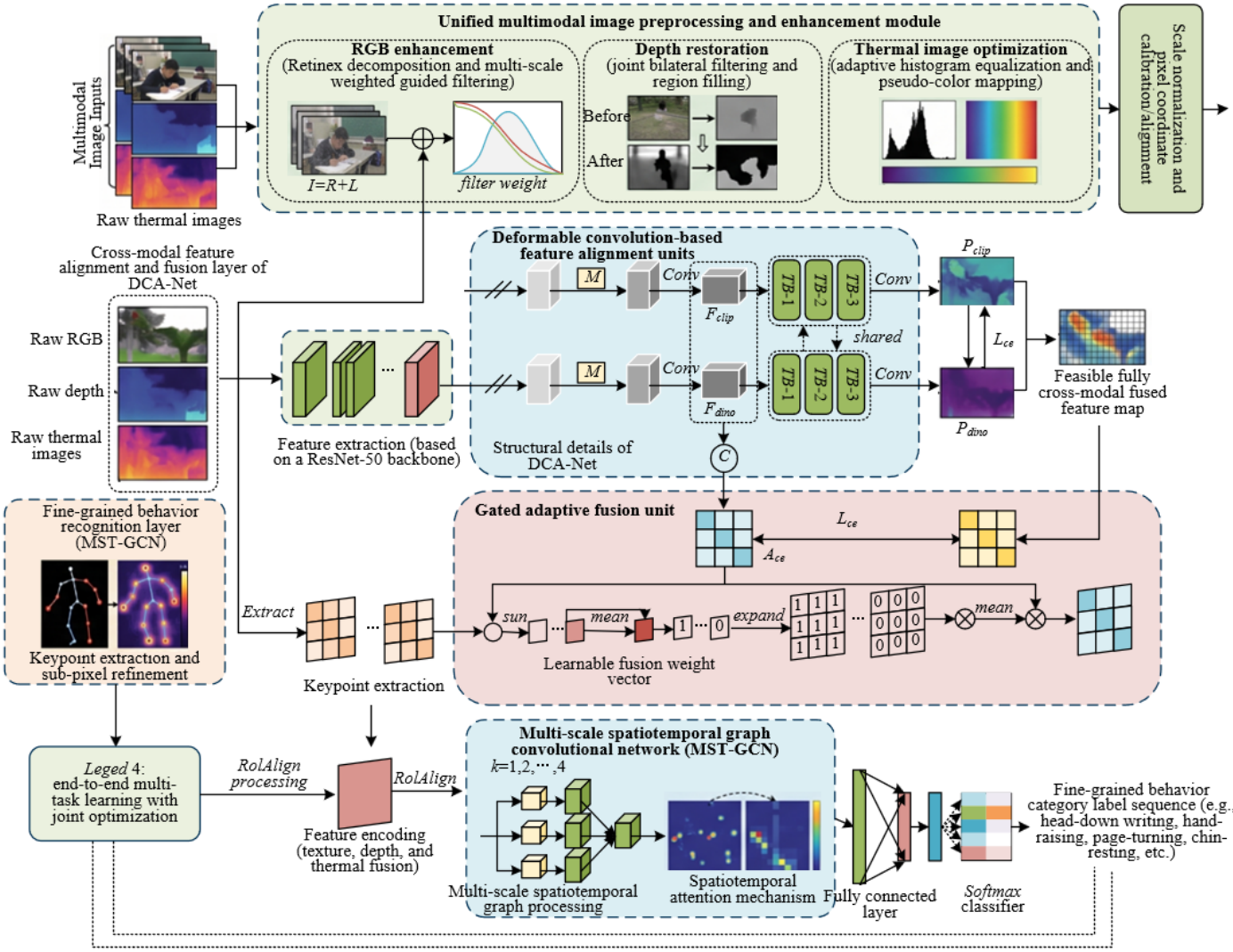

Figure 2. Architecture of the deformable convolution and cross-attention alignment fusion network (DCA-Net)

The aligned multimodal features are further processed through a parallel cross-attention module to exploit complementary information across modalities. This module is designed with two parallel attention branches, enabling feature interaction and weight modulation between depth–RGB and thermal–RGB pairs, respectively. In contrast to conventional single-attention mechanisms, this design allows modality-specific salient information to be selectively emphasized. The attention weight maps are generated based on pairwise feature similarity, which can be formulated as:

$$A_k(x,y)=softmax\left(\frac{F_{RGB}(x,y)\cdot F_k(x,y)^T}{\sqrt{d}}\right) \tag{4}$$

where, $k\in\{D,T\}$, with D and T representing the depth and thermal modalities, $A_k(x,y)$ represents the corresponding attention weight map, and d denotes the feature channel dimension. The softmax function is applied to normalize the weights. In the depth-RGB cross-attention branch, feature similarity between depth and RGB representations is computed to generate attention maps that focus on regions with significant depth variations, thereby enhancing features with weak RGB texture but strong spatial structural cues. In the thermal-RGB branch, attention is directed toward regions associated with physiological characteristics, such as facial temperature variations, resulting in corresponding attention maps. After L2 normalization, the two attention maps are used to modulate the RGB features, yielding depth-modulated and thermal-modulated feature representations. Through this process, complementary information across modalities is effectively extracted.

The gated adaptive fusion unit is responsible for integrating multimodal features and overcoming the limitations of conventional fixed-weight fusion strategies. A learnable weight vector is introduced to dynamically balance the contributions of the three modalities. The fusion process is defined as:

$$F_{fusion}=\alpha\cdot F_{RGB}+\beta\cdot F_{D\text{-}mod}+\gamma\cdot F_{T\text{-}mod} \tag{5}$$

where, F_{fusion} denotes the final cross-modal fused feature map, $F_{D\text{-}mod}$ and $F_{T\text{-}mod}$ represent the depth-modulated and thermal-

modulated features, respectively, and α, β, and γ are learnable fusion weights. These weights are adaptively optimized through a sigmoid activation function and constrained such that $\alpha + \beta + \gamma = 1$. This design enables flexible allocation of modality contributions according to dynamic classroom conditions and feature characteristics. As a result, the fused representation effectively preserves the texture details of RGB images, while incorporating the spatial structural information from depth data and the physiological cues from thermal imaging. Consequently, a comprehensive and discriminative feature representation is obtained, providing robust support for subsequent fine-grained learning behavior recognition.

3.4 Multi-scale spatiotemporal graph convolutional network

To overcome the limitations of conventional behavior recognition methods that rely solely on skeletal keypoints, and to enable fine-grained learning behavior recognition through multimodal information integration, a multi-scale spatiotemporal graph convolutional network is constructed. Through the coordinated integration of multimodal keypoint feature encoding, dynamic spatiotemporal graph construction, multi-scale spatiotemporal convolution, and a spatiotemporal attention mechanism, both spatial structural characteristics and temporal dynamics of behaviors are effectively captured, while model interpretability is simultaneously enhanced. The overall framework is illustrated in Figure 3. A key innovation in the keypoint extraction and feature encoding stage lies in the integration of multimodal information rather than reliance solely on keypoint coordinates. Specifically, a lightweight high-resolution network architecture, combined with depth information, is employed to extract 17 upper-body keypoints of students from the cross-modal fused feature maps. Thermal imaging information is further utilized to refine head-related key regions at the sub-pixel level, thereby ensuring high localization accuracy. Local region feature vectors for each keypoint are subsequently extracted using deformable RoIAlign operations, enabling the encoding of texture, depth, and thermal information. The encoding process is formulated as:

$$f_i = RoIAlign(F_{fusion}, p_i) \cdot W_{enc} + b_{enc} \tag{6}$$

where, f_i denotes the encoded feature vector of the i-th keypoint, F_{fusion} represents the cross-modal fused feature map, p_i denotes the keypoint coordinates, and W_{enc} and b_{enc} correspond to the weights and bias of the encoding layer. This design significantly enhances the discriminative capacity of keypoint features and provides high-quality node representations for subsequent spatiotemporal graph construction.

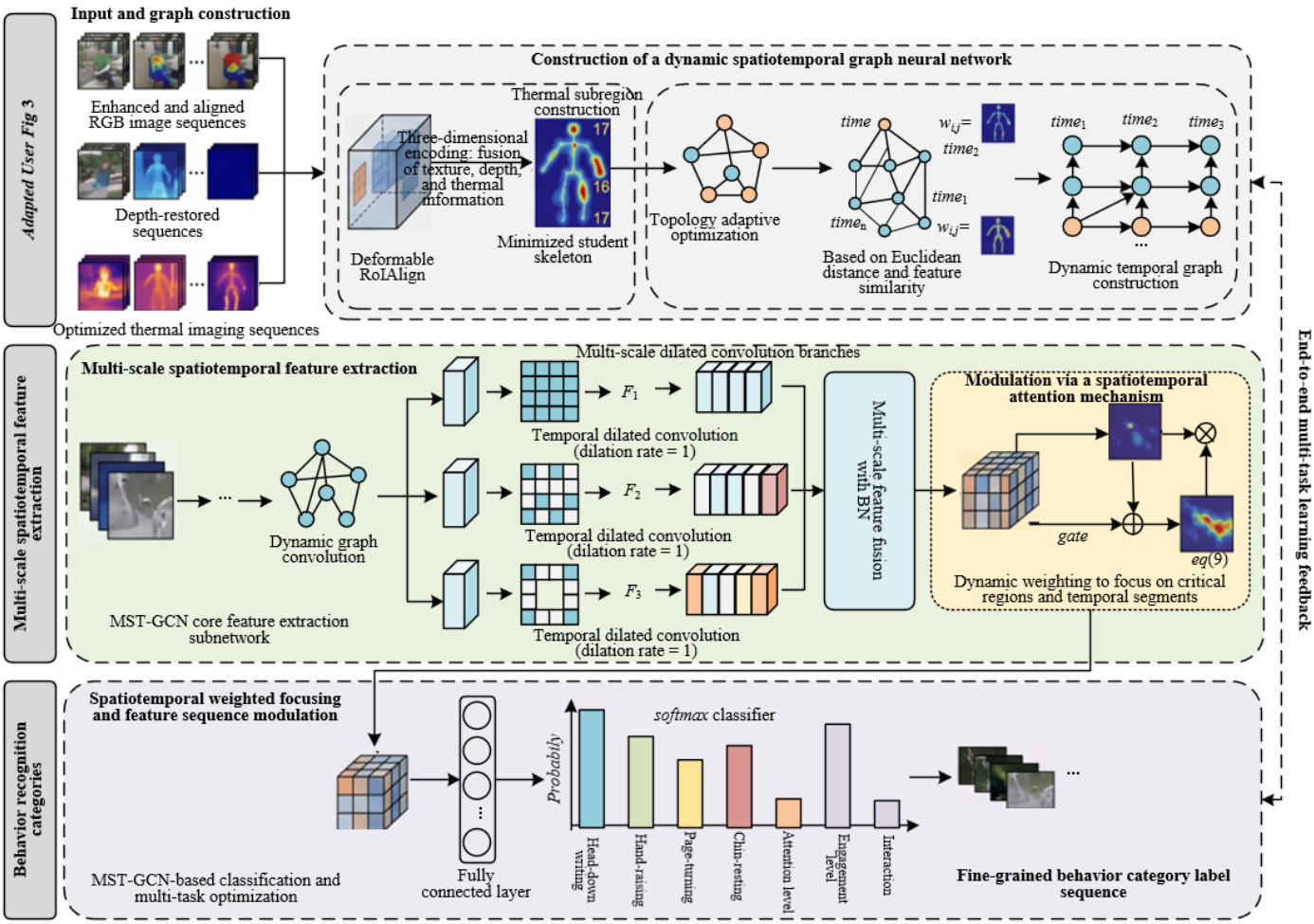

Figure 3. Topology and feature extraction graph of the multi-scale spatiotemporal graph convolutional network (MST-GCN)

The dynamic spatiotemporal graph construction further extends beyond conventional fixed-topology approaches by incorporating both spatial and temporal information for adaptive topology optimization. The graph is decomposed into a spatial graph and a temporal graph. In the spatial graph, encoded keypoints are treated as nodes, and edge weights are

determined by jointly considering Euclidean distance and feature similarity, thereby capturing both spatial proximity and feature consistency. The weight is defined as:

$$w_{ij}=\alpha\cdot exp(-\frac{\|p_i-p_j\|^2}{\sigma^2})+(1-\alpha)\cdot\frac{f_i\cdot f_j^T}{\|f_i\|\|f_j\|} \quad (7)$$

where, w_{ij} denotes the spatial edge weight between node i and node j; α is a balancing coefficient; and σ represents the distance decay parameter. The edge weights are further adaptively adjusted through an attention mechanism, enabling optimization of the spatial graph topology. In the temporal graph, edges are established between corresponding keypoints across adjacent frames, allowing the propagation of temporal dynamic information and facilitating the modeling of continuous behavioral variations. Multi-scale spatiotemporal graph convolution constitutes one of the core innovations of the network. In the spatial domain, graph convolution is employed to aggregate features from neighboring nodes. In the temporal domain, one-dimensional convolutional kernels with dilation rates of 1, 2, and 4 are applied in parallel to process node-wise temporal sequences. This design enables the capture of both rapid behaviors (e.g., hand-raising and page-turning) and slower behaviors (e.g., head-down writing and chin-resting). The multi-scale feature fusion process is expressed as:

$$F_{ST}=BN(ReLU(F_1+F_2+F_4)) \quad (8)$$

where, F_1, F_2, and F_4 denote the output features corresponding to temporal convolutions with dilation rates of 1, 2, and 4, respectively, and BN represents batch normalization. This design effectively enhances the representational capacity of temporal features, enabling comprehensive modeling of behavior patterns at different temporal scales.

A spatiotemporal attention mechanism is further introduced to improve feature discriminability. Through dynamic weighting, attention is focused on keypoint pairs and temporal segments that are most relevant to the current behavior, while suppressing background noise and irrelevant features. The attention weighting process is defined as:

$$F_{att}=softmax(W_{att}\cdot F_{ST})\cdot F_{ST} \quad (9)$$

where, W_{att} denotes the attention weight matrix, and F_{att} represents the attention-enhanced spatiotemporal features. The optimized features are subsequently fed into a fully connected classifier, and category probabilities for fine-grained learning behaviors are obtained through a softmax function, thereby completing the recognition process. Through the integration of multimodal feature fusion, adaptive topology optimization, and multi-scale temporal modeling, the proposed multi-scale spatiotemporal graph convolutional network effectively addresses the limitations of conventional methods, including insufficient temporal feature modeling and limited interpretability, enabling accurate fine-grained learning behavior recognition.

3.5 End-to-end multi-task learning and loss function design

To achieve coordinated optimization across the entire pipeline-including image enhancement, feature fusion, and fine-grained behavior recognition-and to improve overall model performance, an end-to-end multi-task learning strategy is adopted. The enhancement module based on multi-scale weighted guided filtering, the fusion module of the deformable convolution and cross-attention alignment fusion network, and the recognition module of the multi-scale spatiotemporal graph convolutional network are integrated into a unified training framework, enabling joint parameter optimization and mutual adaptation among modules. This design effectively avoids feature inconsistency and performance degradation commonly introduced by stage-wise training. Within this training paradigm, a single backpropagation process is employed to simultaneously propagate gradient information from all three sub-tasks to each module. As a result, the output of the enhancement module is adaptively aligned with the input requirements of the fusion module, while the fused features are further optimized to match the input distribution of the recognition module. A closed-loop optimization mechanism is thereby established, ensuring consistency and synergy across the entire processing pipeline.

To address the distinct optimization challenges associated with each sub-task, a multi-objective loss function is designed. Through the joint contribution of individual loss components, improvements in image enhancement quality, cross-modal fusion effectiveness, and fine-grained recognition accuracy are simultaneously achieved. The total loss is formulated as a weighted sum of the losses from each module:

$$L_{total}=\lambda_1 L_{enh}+\lambda_2 L_{fus}+\lambda_3 L_{rec} \quad (10)$$

where, λ_1, λ_2, and λ_3 denote adaptive weighting coefficients of the modules, and L_{enh}, L_{fus}, and L_{rec} represent the loss functions of the enhancement, fusion, and recognition modules, respectively. The loss function for the enhancement module is composed of a perceptual loss and an edge-preservation loss. The perceptual loss is computed by extracting features using a pre-trained convolutional neural network, enforcing consistency between the enhanced image and a well-illuminated reference image in the feature space. It is defined as:

$$L_{percep}=\|\phi(I_{enh})-\phi(I_{ref})\|_2^2 \quad (11)$$

where, ϕ $(\cdot)$ denotes the feature extraction function, I_{enh} represents the enhanced image, and I_{ref} denotes the reference image. The edge-preservation loss is introduced to constrain the gradient information of the enhanced image, thereby preserving edge details. It is defined as:

$$L_{edge}=\left\|\nabla I_{enh}-\nabla I_{ref}\right\|_1 \quad (12)$$

The weighted sum of the two is obtained as:

$$L_{enh}=\mu L_{percep}+(1-\mu)L_{edge} \quad (13)$$

The fusion module loss consists of a mutual information maximization loss and a feature alignment loss. The mutual information maximization loss is defined as:

$$L_{mi}=-MI(F_{RGB},F_D,F_T) \quad (14)$$

To maximize the complementary information among multimodal features, the feature alignment loss is defined as:

$$L_{align}=\|avg(F_D-F_{RGB})\|_2^2+\|avg(F_T-F_{RGB})\|_2^2 \quad (15)$$

The aligned features are constrained to maintain spatial consistency. For the recognition module, a weighted focal loss is adopted to alleviate the class imbalance problem. The loss function is defined as:

$$L_{w\text{-}focal}=-\sum_{c=1}^{C} w_c y_c \log(p_c) \quad (16)$$

where, w_c denotes the class weight, y_c is the label indicator variable, and p_c represents the predicted category probability. Higher weights are assigned to minority classes, such as hand-raising behaviors.

To ensure coordinated optimization across the three sub-tasks and to prevent dominance by any single loss component, an adaptive weighting optimization strategy is introduced. The coefficients λ_1, λ_2, and λ_3 are not fixed but are dynamically updated during training according to:

$$\lambda_k=\frac{exp(L_k/T)}{\sum_{i=1}^{3} exp(L_i/T)} \quad (17)$$

where, $k\in\{1,2,3\}$ corresponds to the three modules, and T denotes a temperature parameter that controls the sensitivity of weight updates. Through this design, sub-tasks associated with larger loss values are assigned higher weights, allowing current performance bottlenecks to be prioritized during optimization. As a result, a dynamic balance among the losses of different modules is achieved, ensuring the coordinated progression of the three sub-tasks-image enhancement, feature fusion, and behavior recognition-and ultimately improving the overall performance and generalization capability of the model.

4. EXPERIMENTS AND ANALYSIS

4.1 Experimental environment and dataset

A standardized hardware and software environment was adopted to ensure the reproducibility and reliability of the experimental results. The detailed configuration of the experimental setup is summarized in Table 1. At the hardware level, high-performance computing devices were utilized to support both model training and inference. At the software level, the experimental platform was constructed based on widely used deep learning frameworks. Training parameters were carefully tuned through extensive experimentation to achieve a balance between computational efficiency and model performance.

Table 1. Experimental environment configuration

Category	Configuration Details
Hardware	Graphics Processing Unit: NVIDIA RTX 3090 (24 GB); Central Processing Unit: Intel Core i9-12900K; Memory: 64 GB Double Data Rate 5 (DDR5); Storage: 1 TB Solid-State Drive (SSD)
Software	Framework: PyTorch 1.13.1; OpenCV 4.7.0; Python 3.9; Operating System: Ubuntu 20.04 Long-Term Support (LTS)
Training Parameters	Batch size: 32; Initial learning rate: 1×10^{-4}; Number of epochs: 300; Optimizer: AdamW (weight decay: 1×10^{-5}); Temperature coefficient T = 0.1; Early stopping: training is terminated if validation accuracy does not improve for 15 consecutive epochs

Table 2. Details of the SmartClass-MM dataset

Item	Description
Participants	20 undergraduate students (12 male, 8 female; aged 19-24 years)
Behavior Categories	10 classes (hand-raising, page-turning, head-down writing, chin-resting, attentive listening, phone usage, note-taking, drinking water, side conversation, upright sitting)
Total Samples	15,000 samples (1,500 samples per behavior class)
Modalities	Red, Green, and Blue (RGB) images (1920 × 1080), depth images (1920 × 1080), thermal images (1920 × 1080)
Annotation Information	Fine-grained behavior labels, 17 upper-body keypoint coordinates, facial temperature values, and image quality labels
Acquisition Scenarios	Real higher education smart classrooms (three classrooms of varying sizes, including natural lighting, artificial low-light, and mixed lighting conditions)
Data Split	Training set: 12,000 samples (80%); validation set: 1,500 samples (10%); test set: 1,500 samples (10%)

To closely reflect real-world smart classroom scenarios in higher education, a multimodal dataset, termed SmartClass-MM, was constructed. The dataset was synchronously acquired using RGB cameras, Azure Kinect depth sensors, and thermal imaging devices, and it encompassed representative classroom conditions and fine-grained learning behaviors. Detailed information regarding the dataset is provided in Table 2. A total of 20 undergraduate students from different academic years were recruited for data collection, covering 10 categories of fine-grained learning behaviors. In total, 15,000 synchronized trimodal samples were collected, with each sample consisting of RGB, depth, and thermal images along with corresponding annotations. The annotation process was conducted jointly by three researchers in computer vision and two researchers in educational technology. A dual-annotator cross-labeling strategy was adopted, and annotation consistency was evaluated using the Cohen's Kappa coefficient, yielding a value of 0.92, which indicates a high level of annotation reliability. The dataset is specifically designed to capture realistic classroom conditions, including low-light environments and partial occlusions, thereby enabling effective validation of the proposed framework in terms of applicability and generalization capability. Furthermore, it addresses the limitations of existing public datasets, which often lack specificity for smart classroom scenarios.

4.2 Experimental results and analysis

4.2.1 Quantitative results analysis

To validate the performance advantages of the proposed framework, several representative methods from image processing and behavior recognition were selected as baselines. These included a traditional multimodal approach (Retinex + simple concatenation + graph convolutional network), a deep learning-based method (convolutional neural network + conventional attention + graph convolutional network), and a recent multimodal behavior recognition method (multi-modal convolutional neural network + spatiotemporal graph convolutional network). Evaluation metrics encompassed image enhancement quality, feature fusion effectiveness, fine-grained behavior recognition accuracy, and real-time performance. The comparative results are summarized in Table 3.

Table 3. Quantitative comparison of different methods

Method	Peak Signal-to-Noise Ratio (dB)	Structural Similarity	Mutual Information	Edge Preservation Index	Top-1 Accuracy (%)	Top-5 Accuracy (%)	F1 Score (%)	Frames Per Second
Retinex + Simple Concatenation + Graph Convolutional Network	32.15	0.862	0.58	0.79	78.62	89.35	77.94	28.3
Convolutional neural network + Conventional Attention + Graph Convolutional Network	34.87	0.895	0.65	0.83	86.45	94.21	85.87	31.7
Multi-Modal Convolutional Neural Network + Spatiotemporal Graph Convolutional Network	36.47	0.913	0.72	0.86	90.23	96.78	89.76	33.5
Proposed Method	38.62	0.937	0.85	0.92	95.37	98.94	95.12	38.2

It can be observed from Table 3 that superior performance is consistently achieved by the proposed framework across all evaluation metrics, demonstrating its strong overall effectiveness. In terms of image enhancement quality, a peak signal-to-noise ratio of 38.62 dB and a structural similarity index of 0.937 are obtained, representing improvements of 2.15 dB and 0.024, respectively, over the best-performing baseline. These gains can be primarily attributed to the enhancement module based on multi-scale weighted guided filtering, in which a multi-scale weighted guided filtering strategy enables an effective balance between noise suppression and edge preservation under low-light conditions, thereby significantly improving image quality in complex classroom lighting environments. Regarding feature fusion, the proposed method achieves mutual information and edge preservation index values of 0.85 and 0.92, respectively, exceeding those of multi-modal convolutional neural network + spatiotemporal graph convolutional network by 0.13 and 0.06. This improvement is attributed to the deformable convolution and cross-attention alignment fusion network architecture, where precise cross-modal alignment is realized through deformable convolution, and complementary information across modalities is effectively exploited via the parallel cross-attention mechanism, resulting in more informative fused feature representations. For fine-grained learning behavior recognition, Top-1 accuracy, Top-5 accuracy, and F1 score reach 95.37%, 98.94%, and 95.12%, respectively, corresponding to improvements of 5.14%, 2.16%, and 5.36% over the strongest baseline. These gains are primarily enabled by the multi-scale spatiotemporal graph convolutional network, which integrates multimodal keypoint features and captures behavior patterns at different temporal scales through multi-scale spatiotemporal convolution, while focusing on key features via a spatiotemporal attention mechanism. As a result, highly similar fine-grained behaviors can be effectively distinguished. In terms of real-time performance, an inference speed of 38.2 frames per second is achieved, satisfying the requirements of real-time smart classroom monitoring (frames per second $\geq$ 30). This performance surpasses all baseline methods and can be attributed to the use of a lightweight high-resolution network architecture combined with the optimized design of deformable convolution, which reduces computational complexity while maintaining performance.

Further analysis under typical smart classroom conditions, including low illumination and occlusion, demonstrates the robustness of the proposed framework. Under low-light conditions, a peak signal-to-noise ratio improvement of 2.87 dB over multi-modal convolutional neural network + spatiotemporal graph convolutional network is observed, indicating enhanced adaptability to challenging lighting environments. Under occlusion conditions, the reduction in Top-1 accuracy is limited to only 3.21%, which is significantly lower than that of the baseline methods (ranging from 5.87% to 8.34%). This result demonstrates that the alignment and fusion capability of the deformable convolution and cross-attention alignment fusion network, combined with the spatiotemporal feature modeling of the multi-scale spatiotemporal graph convolutional network, effectively mitigates the impact of occlusions, thereby ensuring reliable performance in real-world classroom scenarios.

4.2.2 Ablation study analysis

To evaluate the effectiveness of each core module, ablation experiments were conducted. A baseline model-comprising conventional image enhancement, simple feature fusion, and a standard graph convolutional network-based recognition network without any proposed innovations-was adopted as the reference. The enhancement module based on multi-scale weighted guided filtering, the fusion module of the deformable

convolution and cross-attention alignment fusion network, and the recognition module of the multi-scale spatiotemporal graph convolutional network were then incrementally incorporated to assess their individual and combined contributions. The detailed results are presented in Table 4.

As shown in Table 4, substantial performance improvements are observed with the inclusion of each module, thereby validating the effectiveness of the proposed design. After the integration of the multi-scale weighted guided filtering enhancement module, peak signal-to-noise ratio and structural similarity index are increased by 4.32 dB and 0.051, respectively, while Top-1 accuracy is improved by 8.37%. These improvements indicate that the multi-scale weighted guided filtering module effectively enhances multimodal image quality, providing high-quality input for subsequent feature fusion and behavior recognition. The performance gains are primarily attributed to the multi-scale weighted guided filtering mechanism, which enables targeted restoration of modality-specific degradations. With the further incorporation of the fusion module of the deformable convolution and cross-attention alignment fusion network, additional improvements of 1.74 dB in peak signal-to-noise ratio, 0.018 in structural similarity index, and 6.86% in Top-1 accuracy are achieved. These results demonstrate that the deformable convolution and cross-attention alignment fusion network enables precise cross-modal feature alignment and efficient feature fusion, thereby effectively exploiting complementary information across modalities and overcoming the limitations of conventional fusion methods, such as poor alignment accuracy and insufficient utilization of complementary information. Following the addition of the recognition module of the multi-scale spatiotemporal graph convolutional network, recognition performance is further improved, with Top-1 accuracy increasing by 3.79% and the F1 score by 4.09%. This confirms that the multi-scale spatiotemporal graph convolutional network effectively captures multimodal spatiotemporal features and enhances both the accuracy and interpretability of fine-grained learning behavior recognition. The multi-scale temporal convolution design allows behavior patterns with different temporal dynamics to be comprehensively modeled. It is noteworthy that, although slight fluctuations in inference speed are observed with the addition of each module, the final model maintains a processing speed of 38.2 frames per second, satisfying real-time requirements. This indicates that the proposed framework achieves a favorable balance between performance and computational efficiency, without introducing substantial additional computational overhead.

Table 4. Overall ablation study results

Method	Peak Signal-to-Noise Ratio (dB)	Structural Similarity	Top-1 Accuracy (%)	F1 Score (%)	Frames Per Second
Baseline Model	31.89	0.857	76.35	75.82	39.5
Baseline + Multi-scale Weighted Guided Filtering Enhancement	36.21	0.908	84.72	84.15	37.8
Baseline + Multi-scale Weighted Guided Filtering + Deformable Convolution and Cross-attention Alignment Fusion Network	37.95	0.926	91.58	91.03	35.4
Baseline + Multi-scale Weighted Guided Filtering + Deformable Convolution and Cross-attention Alignment Fusion Network + Multi-scale Spatiotemporal Graph Convolutional Network	38.62	0.937	95.37	95.12	38.2

4.2.3 Robustness analysis

To evaluate the adaptability of the proposed framework under complex interference conditions in smart classroom environments, robustness experiments were conducted. Four representative scenarios were considered, including normal conditions, low-light conditions, severe occlusion, and student movement. The performance of the proposed method was compared with that of the strongest baseline method (multi-modal convolutional neural network + spatiotemporal graph convolutional network). The corresponding results are summarized in Table 5.

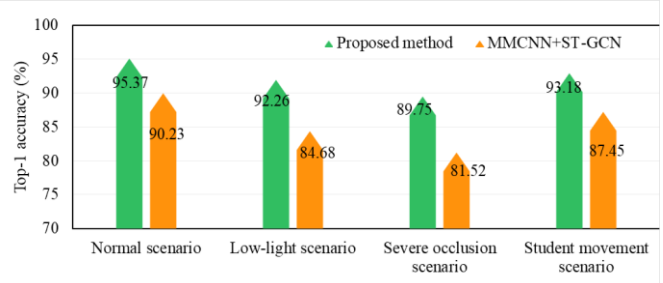

Figure 4. Results of the robustness experiments

As illustrated in Figure 4, high recognition accuracy is consistently maintained by the proposed framework across all interference scenarios, while the performance degradation remains significantly lower than that of the baseline method, indicating strong robustness. Under low-light conditions, a performance degradation of only 3.21% is observed, which is substantially lower than the 5.87% reported for the baseline. This improvement can be attributed to the multi-scale weighted guided filtering enhancement module, by which low-light image quality is effectively improved through noise suppression and edge preservation, thereby providing reliable inputs for subsequent recognition. In the severe occlusion scenario, the performance degradation is limited to 5.99%, compared to 9.65% for the baseline method. This advantage is primarily derived from the fusion module of the deformable convolution and cross-attention alignment fusion network, through which multimodal features are effectively integrated. The depth and thermal information compensates for missing texture information in occluded regions. Furthermore, the multi-scale spatiotemporal graph convolutional network selectively focuses on unoccluded keypoints, thereby enhancing recognition stability. In the student movement scenario, the performance degradation is reduced to 2.29%, slightly lower than that of the baseline method. This

improvement is mainly attributed to the deformable convolution mechanism in the deformable convolution and cross-attention alignment fusion network, which adaptively adjusts feature sampling locations to achieve non-rigid alignment, thereby mitigating cross-modal misalignment caused by student motion. Overall, these results demonstrate that the proposed framework effectively adapts to complex interference conditions in smart classroom environments, exhibiting strong practical applicability.

4.2.4 Real-time performance analysis

To evaluate the engineering feasibility of the proposed framework, a detailed analysis of real-time performance was conducted by decomposing the processing speed of each module. The frame rates (frames per second) of individual modules and the integrated model were compared, and the impact of lightweight design on real-time performance was further examined. The results are summarized in Table 5.

Table 5. Real-time performance analysis

Module	Standalone Inference Speed (Frames Per Second)	Integrated Inference Speed (Frames Per Second)	Lightweight Optimization Strategy	Optimized Inference Speed (Frames Per Second)
Multi-scale Weighted Guided Filtering Enhancement Module	65.3	-	Multi-scale filtering with parallel computation	72.5
Fusion Module of the Deformable Convolution and Cross-attention Alignment Fusion Network	58.7	-	Lightweight deformable convolution design	64.2
Recognition Module of the Multi-scale Spatiotemporal Graph Convolutional Network	42.8	-	Lightweight high-resolution network + temporal convolution optimization	48.6
Full Model (Unoptimized)	-	32.1	-	-
Full Model (Optimized)	-	38.2	All above optimizations + model pruning	38.2

As shown in Table 5, real-time performance is effectively improved through multiple lightweight design strategies. When each module is executed independently, the frames per second of the multi-scale weighted guided filtering, the deformable convolution and cross-attention alignment fusion network, and multi-scale spatiotemporal graph convolutional network are increased by 7.2, 5.5, and 5.8 frames per second, respectively, after optimization, indicating that computational complexity is reduced without compromising performance. For the complete framework, an inference speed of 32.1 frames per second is achieved prior to optimization, which satisfies the basic requirement for real-time processing. After applying lightweight optimization strategies, the frames per second are further improved to 38.2, resulting in enhanced smoothness for real-time monitoring. These improvements can be attributed to several key design choices. The adoption of a lightweight high-resolution network architecture significantly reduces the computational cost of keypoint extraction. The lightweight design of deformable convolution decreases the overhead associated with feature alignment. Parallel computation in multi-scale filtering improves the efficiency of the enhancement module. In addition, model pruning eliminates redundant parameters, further accelerating inference speed. Considering the requirements of real-time monitoring in smart classroom environments, an inference speed of 38.2 frames per second is sufficient to support real-time capture and recognition of learning behaviors, thereby demonstrating strong practical applicability of the proposed framework.

4.3 Ablation study and ablation analysis

To further quantify the contributions of individual components, a fine-grained ablation study was conducted. Four key submodules were examined: (i) the multi-scale weighting scheme in the multi-scale weighted guided filtering enhancement module, (ii) deformable convolution in the fusion module of the deformable convolution and cross-attention alignment fusion network, (iii) multi-dilation temporal convolution in the recognition module of the multi-scale spatiotemporal graph convolutional network, and (iv) the adaptive loss weighting in the end-to-end multi-task learning framework. For each component, a “removal” setting was constructed and compared against the full model. The results are summarized in Table 6.

Table 6. Fine-grained ablation results of submodules

Method	Peak Signal-to-Noise Ratio (dB)	Structural Similarity	Mutual Information	Edge Preservation Index	Top-1 Accuracy (%)	F1 Score (%)
Full Model	38.62	0.937	0.85	0.92	95.37	95.12
Without Multi-Scale Weights of Multi-scale Weighted Guided Filtering	35.79	0.901	0.81	0.89	91.64	91.25
Without Deformable Convolution of the Deformable Convolution and Cross-attention Alignment Fusion Network	37.85	0.928	0.74	0.85	90.87	90.53
Without Multi-Dilation Convolution of the Multi-scale Spatiotemporal Graph Convolutional Network	38.45	0.934	0.84	0.91	92.35	92.01
Without Adaptive Loss Weights	38.17	0.932	0.82	0.90	93.12	92.87

As indicated in Table 6, each proposed submodule contributes substantially to overall performance, thereby validating the rationality and effectiveness of the design. When the multi-scale weighting mechanism in the multi-scale weighted guided filtering module is removed, the peak signal-to-noise ratio decreases by 2.83 dB, the structural similarity index decreases by 0.036, and Top-1 accuracy declines by 3.73%. These results demonstrate that the multi-scale weighting strategy enables adaptive allocation of filtering weights, effectively balancing noise suppression and edge preservation, and significantly improving image enhancement quality under low-light conditions. This component is therefore identified as a critical element of the multi-scale weighted guided filtering module. When deformable convolution in the deformable convolution and cross-attention alignment fusion network is excluded, mutual information and edge preservation index decrease by 0.11 and 0.07, respectively, while Top-1 accuracy drops by 4.50%. These results confirm that deformable convolution enables precise non-rigid alignment of cross-modal features, effectively addressing misalignment caused by sensor disparity and student motion, and significantly improving the quality of fused features. When multi-dilation temporal convolution in the multi-scale spatiotemporal graph convolutional network is removed, Top-1 accuracy decreases by 3.02% and the F1 score decreases by 3.11%. This indicates that multi-dilation convolution is capable of simultaneously capturing both rapid and slow behavior patterns, thereby enhancing the representation of temporal features and improving the discrimination of highly similar fine-grained behaviors. When adaptive loss weighting is removed, a moderate decline is observed across all metrics, with Top-1 accuracy decreasing by 2.25%. This suggests that adaptive weighting effectively balances the contributions of enhancement, fusion, and recognition tasks, ensuring coordinated optimization and preventing any single loss component from dominating the training process.

Overall, the fine-grained ablation results demonstrate that the proposed submodules are not redundant, but instead provide targeted solutions to key technical challenges at different stages of the framework. Through their synergistic interaction, significant improvements in overall performance are achieved, thereby further validating the innovation and effectiveness of the proposed design and providing strong experimental evidence supporting the rationality of the method.

To provide intuitive evidence of robustness and effectiveness under real-world smart classroom conditions, including low illumination and multi-person long-range occlusions, visualization experiments were conducted for multimodal image processing and feature extraction. As illustrated in Figure 5, after multi-scale weighted guided filtering, the illumination distribution of low-light RGB images is significantly equalized, and fine-grained texture details are clearly restored. In addition, missing regions in the original depth images are accurately reconstructed, resulting in well-defined spatial foreground–background separation. During the subsequent cross-modal feature fusion stage, the feature heatmaps generated by the deep cross-attention network are densely and accurately concentrated on key regions associated with student interactions, while background noise from static objects such as desks, chairs, and walls is effectively suppressed.

The final visual semantic recognition results further demonstrate that the proposed multimodal image quality enhancement and fine-grained feature collaborative fusion mechanism effectively mitigate visual distortions introduced by limitations in low-level sensor data acquisition. As a result, accurate recognition of multi-subject learning behaviors is achieved in complex and unconstrained higher education classroom environments, while strong generalization capability is simultaneously maintained.

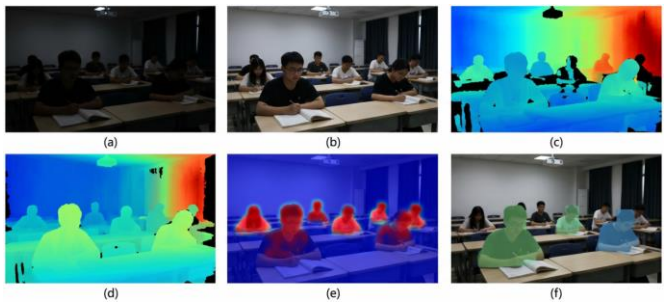

Figure 5. Visual comparison of image processing and recognition results in a long-range multi-student smart classroom scenario
(a) original red, green, and blue (RGB) image; (b) RGB image after enhancement by the multi-scale weighted guided filtering module; (c) original depth image; (d) depth completion and enhancement result; (e) cross-modal fused feature heatmap generated by the deformable convolution and cross-attention alignment fusion network; (f) behavior recognition result produced by the multi-scale spatiotemporal graph convolutional network

5. DISCUSSION

The proposed multimodal image processing and fine-grained learning behavior recognition framework demonstrates substantial advancements across the three core stages-multimodal image enhancement, cross-modal feature fusion, and fine-grained behavior recognition-as evidenced by both quantitative results and visualization analyses. A comprehensive technical framework tailored for smart classrooms in higher education is thus established, exhibiting clear advantages over existing approaches. The multi-scale weighted guided filtering enhancement module extends beyond the limitations of conventional single-modality enhancement and fixed-scale filtering. By explicitly addressing challenges such as low illumination, depth incompleteness, and insufficient contrast in thermal imaging, unified optimization and spatial alignment of RGB, depth, and thermal modalities are achieved. This design effectively balances noise suppression and edge preservation. The deformable convolution and cross-attention alignment fusion network architecture introduces deformable convolution into cross-modal alignment tasks. Through the integration of parallel cross-attention mechanisms and a gated adaptive fusion strategy, limitations associated with conventional fusion methods-such as poor alignment accuracy and feature redundancy-are effectively mitigated. The multi-scale spatiotemporal graph convolutional network integrates multimodal keypoint features and employs dynamic spatiotemporal graph construction alongside multi-scale temporal convolution. This design effectively enhances the accuracy and interpretability of fine-grained behavior recognition. In addition, the end-to-end multi-task learning strategy combined with an adaptive multi-objective loss function enables coordinated optimization across the entire processing pipeline. The coordinated interaction among all modules forms a closed-loop optimization framework. Overall,

the proposed framework is not only well-suited for smart classroom environments but also provides a generalizable paradigm for multimodal image processing and fine-grained behavior recognition in other complex scenarios. The methodological contributions are therefore of considerable academic significance and practical relevance.

Despite the superior overall performance demonstrated in the experiments and the effective adaptation to typical smart classroom scenarios, several limitations remain. In densely populated classroom environments with significant student overlap, inaccuracies in keypoint localization are likely to occur. Consequently, cross-modal feature alignment and fusion may be adversely affected, leading to reduced discrimination accuracy for similar fine-grained behaviors. This limitation is primarily attributed to insufficient feature modeling in overlapping regions. In addition, the temperature calibration accuracy of thermal images is constrained by sensor performance and environmental temperature variations. Although temperature differences are enhanced in the current framework, environmental fluctuations are not explicitly modeled, which may compromise the accuracy of thermal information and subsequently affect recognition accuracy. Furthermore, the degree of model lightweighting remains limited. While real-time requirements are satisfied, computational cost and memory consumption remain relatively high, posing challenges for efficient deployment on resource-constrained platforms such as embedded edge devices. Finally, although the constructed SmartClass-MM dataset reflects realistic classroom conditions, its scale and diversity remain limited, which may restrict generalization across different classroom types and student populations.

In light of these limitations and emerging trends in image processing and intelligent education, future work is expected to be pursued along three primary directions: feature modeling, data optimization, and model deployment. To address the challenge of overlapping student scenarios, Transformer-based architectures will be incorporated to enhance global modeling of cross-modal features. In addition, attention mechanisms specifically designed for overlapping regions will be developed, and keypoint extraction algorithms will be refined to enable accurate localization and separation of overlapping keypoints. To improve the accuracy of thermal information, adaptive temperature calibration methods based on environmental conditions will be investigated, and deep learning–based strategies will be employed to achieve dynamic compensation of temperature errors. To enhance model generalization, the dataset will be expanded in both scale and diversity, and advanced data augmentation techniques will be introduced. Model lightweighting will be further explored through techniques such as model quantization, pruning, and knowledge distillation, enabling compatibility with embedded edge devices. In addition, the integration of federated learning with multimodal techniques will be investigated to facilitate multi-scenario collaborative training while preserving student privacy. Finally, deeper integration with educational management systems will be pursued, allowing behavior recognition outputs to be directly linked with instructional optimization strategies, thereby promoting the development of precise and personalized smart classrooms in higher education.

6. CONCLUSION

To address the critical requirements of multimodal image processing and fine-grained learning behavior recognition in smart classrooms for higher education, a systematic investigation was conducted targeting key technical challenges, including poor adaptability to low-light conditions, cross-modal spatial misalignment, and coarse-grained behavior recognition. A comprehensive multimodal processing and fine-grained behavior recognition framework was developed, enabling end-to-end optimization from multimodal image preprocessing to behavior recognition. This framework provides a robust technical solution for the intelligent advancement of smart classrooms. The core contributions are centered on the design and coordinated integration of three major modules. First, a unified multimodal image preprocessing and enhancement module is introduced, in which a multi-scale weighted guided filtering strategy is employed to achieve precise optimization and spatial alignment of RGB, depth, and thermal images. This module effectively addresses image quality degradation under complex classroom conditions. Second, a deformable convolution and cross-attention alignment fusion network is developed, overcoming the limitations of conventional fusion methods by enabling non-rigid alignment and complementary feature extraction, thereby improving the accuracy and effectiveness of cross-modal feature fusion. Third, a multi-scale spatiotemporal graph convolutional network is constructed, in which multimodal keypoint features are integrated. Through dynamic spatiotemporal modeling and multi-scale temporal feature extraction, accurate recognition and enhanced interpretability of fine-grained learning behaviors are achieved.

Furthermore, an end-to-end multi-task learning strategy combined with an adaptive multi-objective loss function is introduced to enable coordinated optimization across the entire pipeline. This design effectively mitigates class imbalance and ensures model stability and generalization capability. Overall, the proposed framework provides a novel and efficient approach for multimodal image processing and fine-grained learning behavior recognition in smart classrooms. In addition, it extends the application of cross-modal feature fusion and spatiotemporal graph convolution techniques within the field of image processing. The proposed methodology offers valuable insights for interdisciplinary research at the intersection of intelligent education and image processing, with broad practical applicability. It is anticipated that this framework will further promote the development of smart classrooms in higher education toward greater precision and personalization.

REFERENCES

[1] Feng, J., Gou, M. (2023). A progressive region-focused network for fine-grained human behavior recognition. Human-Centric Computing and Information Sciences, 13: 10. https://doi.org/10.22967/HCIS.2023.13.010

[2] Pu, Y., Han, Y., Wang, Y., Feng, J., Deng, C., Huang, G. (2024). Fine-grained recognition with learnable semantic data augmentation. IEEE Transactions on Image Processing, 33: 3130-3144. https://doi.org/10.1109/TIP.2024.3364500

[3] Kwet, M., Prinsloo, P. (2020). The ‘smart’classroom: A new frontier in the age of the smart university. Teaching in Higher Education, 25(4): 510-526. https://doi.org/10.1080/13562517.2020.1734922

[4] Fang, Q., Zhang, Y. (2024). Optimizing remote teaching interaction platforms through multimodal image recognition technology. Traitement du Signal, 41(1): 225-235. https://doi.org/10.18280/ts.410118

[5] Zhou, Y., Wang, J., Zhang, J. (2024). A multimodal image recognition system for student behavior analysis in smart classrooms in universities. Traitement du Signal, 41(6): 3285-3293. https://doi.org/10.18280/ts.410644

[6] Monno, Y., Teranaka, H., Yoshizaki, K., Tanaka, M., Okutomi, M. (2018). Single-sensor RGB-NIR imaging: High-quality system design and prototype implementation. IEEE Sensors Journal, 19(2): 497-507. https://doi.org/10.1109/JSEN.2018.2876774

[7] Le, T.H., Jung, S.W., Won, C.S. (2017). A new depth image quality metric using a pair of color and depth images. Multimedia Tools and Applications, 76(9): 11285-11303. https://doi.org/10.1007/s11042-016-3392-4

[8] Szekely, V., Rencz, M. (2002). Image processing procedures for the thermal measurements. IEEE Transactions on Components and Packaging Technologies, 22(2): 259-265. https://doi.org/10.1109/6144.774742

[9] Qiao, Q. (2022). Image processing technology based on machine learning. IEEE Consumer Electronics Magazine, 13(4): 90-99. https://doi.org/10.1109/MCE.2022.3150659

[10] He, D., Xiong, S. (2021). Image processing design and algorithm research based on cloud computing. Journal of Sensors, 2021(1): 9198884. https://doi.org/10.1155/2021/9198884

[11] Bansal, S., Bansal, R.K., Bhardwaj, R. (2024). A novel low complexity retinex-based algorithm for enhancing low-light images. Multimedia Tools and Applications, 83(10): 29485-29504. https://doi.org/10.1007/s11042-023-16610-4

[12] Majumdar, J., Nandi, M., Nagabhushan, P. (2011). Retinex algorithm with reduced halo artifacts. Defence Science Journal, 61(6): 559-566. https://doi.org/10.14429/dsj.61.753

[13] Alwazzan, M.J., Alattar, A.M. (2025). Modified algorithm to enhance illumination and detail in colour retinal images using the CLAHE technique with a Wiener filter. Biomedical Signal Processing and Control, 110: 108241. https://doi.org/10.1016/j.bspc.2025.108241

[14] Yu, C.Y., Lin, H.Y., Ouyang, Y.C., Yu, T.W. (2013). Modulated AIHT image contrast enhancement algorithm based on contrast-limited adaptive histogram equalization. Applied Mathematics & Information Sciences, 7(2): 449-454. http://doi.org/10.12785/amis/072L10

[15] Gao, Y., Dai, M., Zhang, Q. (2023). Cross-modal and multi-level feature refinement network for RGB-D salient object detection. The Visual Computer, 39(9): 3979-3994. https://doi.org/10.1007/s00371-022-02543-w

[16] Zhao, Y., Li, H. (2026). Feature fusion-based cross-modal proxy hashing retrieval. Applied Sciences, 16(3): 1532. https://doi.org/10.3390/app16031532

[17] Wu, J., Du, J., Hao, F., Hong, J. (2026). Du-CIPT: Dual cross-modal interactive pyramid transformer for RGB-thermal salient object detection and segmentation. Signal Processing: Image Communication, 145: 117551. https://doi.org/10.1016/j.image.2026.117551

[18] Pu, Y., Han, Y., Wang, Y., Feng, J., Deng, C., Huang, G. (2024). Fine-grained recognition with learnable semantic data augmentation. IEEE Transactions on Image Processing, 33: 3130-3144. https://doi.org/10.1109/TIP.2024.3364500

[19] Yu, N., Chen, L., Yi, X., Huang, J. (2025). Attention learning with counterfactual intervention based on feature fusion for fine-grained feature learning. Digital Signal Processing, 163: 105215. https://doi.org/10.1016/j.dsp.2025.105215

[20] Ma, L., Yang, H., Jin, G. (2025). A spatio-temporal feature representation of multimodal surveillance images for behavioral recognition. International Arab Journal of Information Technology, 22(4): 832-843. https://doi.org/10.34028/iajit/22/4/15