*2025 International Conference on Intelligent Computing*
*July 26-29, Ningbo, China*
https://www.ic-icc.cn/2025/index.php

# Low-Light Gaze Estimation for Fine-grained Intelligent Classroom Behavior Monitoring

Jin Wang[1][0000-0002-5113-6296], Meng Chen[1][0009-0005-4666-6988] and Dandan Wang[1,*][0009-0003-2026-533X]

[1] School of Artificial Intelligence and Computer Science, Nantong University, Jiangsu Province, China
wang.dd@ntu.edu.cn

**Abstract.** Computer Vision (CV) technology is crucial for intelligent classroom behavior monitoring. Current CV methods can only measure coarse-grained metrics like attendance rate and head-up rate, while gaze estimation enables fine-grained monitoring of each student. However, existing gaze estimation algorithms struggle in low-light classroom environments. To address this, we propose the LLSGE-Net framework, which integrates low-light image enhancement with gaze estimation. This multi-stage enhancement and calibration process significantly improves image quality. Our method utilizes Local-Global Context Fusion (ALGCF) for better eye and face feature integration, and a feature enhancement technique combining 1D convolution and group normalization. The Enhanced Local Spatial and Global Channel Attention (ELSCA) improves the localization of regions of interest, while the Deep Feature Extraction Network (DFENet) refines high-level features. Extensive experiments demonstrate the superiority of our approach in real-world low-light classroom scenarios for student attention detection and behavior monitoring.

**Keywords:** Gaze Estimation, Classroom Behavior Monitoring, Local-Global Context Fusion, Attention Mechanism, Deep Feature Extraction Network.

## 1 Introduction

Intelligent classroom behavior analysis [1] is of great significance in the education field [2] , providing multi-dimensional data related to students' learning status [3], emotional responses [4] , attention levels [5] , and more. This data is essential for improving teaching quality, promoting students' overall development, and optimizing teaching strategies. Current classroom behavior monitoring systems focus on coarse-grained metrics such as seat occupancy and head-up rate [6]. However, the future of classroom behavior analysis lies in fine-grained monitoring, enabling detailed tracking of each student's behavior.

Fine-grained classroom behavior monitoring can be divided into explicit and implicit monitoring. Explicit monitoring uses devices like eye trackers, smartwatches, and smart glasses, which are costly, suitable only for small-scale scenarios, and may interfere with students' behavior [7]. In contrast, implicit monitoring generally involves cameras to analyze students' behavior without their awareness, minimizing the impact on their actions [8]. Gaze estimation is a key technology for implicit monitoring, as it allows for the detection and analysis of students' gaze direction, providing valuable insights into their attention and engagement.

However, existing gaze estimation methods are primarily designed for ideal lighting conditions, such as in Human-Computer Interaction (HCI) [9], Virtual Reality (VR) [10], and Driver Attention Monitoring [11], making them unsuitable for classroom scenarios. In classrooms, lighting conditions are often poor, with only the teacher's area being well-lit, while students' positions typically suffer from low-light environments. Gaze estimation in such low-light conditions presents a significant challenge.

To address this, we propose the LLSGE-Net framework, depicted in **Fig. 1**, which integrates image enhancement with gaze estimation in an end-to-end system. The enhancement module improves low-light image quality through a multi-stage process, while the gaze estimation module, based on an improved ResNet18 architecture, applies the novel ELSCA attention mechanism to focus on key eye regions. Additionally, the Deep Feature Extraction Network (DFENet) refines the feature extraction process. These innovations enhance gaze estimation accuracy, particularly in low-light classroom scenarios, and represent a significant advancement in implicit monitoring for classroom behavior analysis.

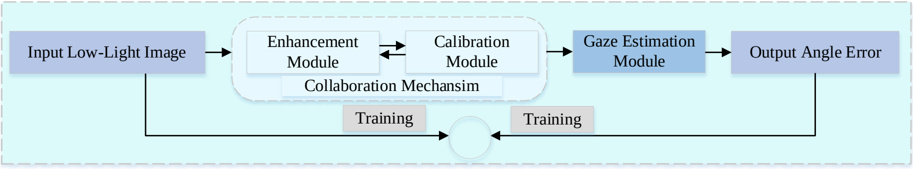


**Fig. 1.** The proposed end-to-end  LLSGE-Net model framework.

The main contributions of this paper are as follows:

We propose a synergistic image enhancement module that addresses the challenges posed by low-light conditions. The module enhances eye details by leveraging multi-scale feature fusion and global context information, while preserving the overall image quality. A subsequent calibration module refines the image further, removing artifacts and making it closer to real-world classroom conditions.

We introduce a novel gaze estimation model that integrates a deep attention mechanism with feature extraction, specifically designed to work effectively in low-light classroom environments. This model significantly improves the accuracy and robustness of gaze estimation by focusing on key eye regions, even under challenging lighting conditions.

We propose a new research direction by integrating low-light image enhancement with gaze estimation, which provides a solution for real-world classroom behavior monitoring. This novel approach enhances the reliability and effectiveness of gaze estimation, offering a significant step forward in implicit monitoring systems for human-computer interaction in practical, low-light scenarios.

# 2 Related Work

## 2.1 Classic Gaze Estimation

With the rapid development of deep learning, many appearance-based gaze estimation methods have been proposed, which typically map facial or eye images to gaze points. Some studies have focused on enhancing network architectures to improve estimation accuracy. Cheng et al. introduced dual-view interactive convolution blocks and transformers to better integrate feature representations [12], while Nagpure et al. leveraged neural architecture search to design a compact and efficient model [13]. Wang et al. optimized facial feature extraction using capsule networks [14], and Jin et al. developed a personalized calibration pipeline to address the eyeball's rolling state [15].

Other research has aimed at improving cross-domain performance to enhance model generalization. Cai et al. explored uncertainty minimization techniques for domain adaptation [16], Zhang et al. proposed a latent gaze pattern-based estimation approach [17], and Liu et al. introduced model variants combined with intra-group attention mechanisms for better adaptability [18]. Hisadome et al. further proposed a multi-view gaze estimation task and a cross-view feature fusion strategy to address variability across head poses [19].

In addition to architecture and generalization improvements, there has been growing attention on enhancing the practicality and robustness of gaze estimation systems. Zhang et al. proposed a real-time, single-stage, multi-person gaze estimation method [20], while Balim et al. developed a model capable of predicting 3D gaze direction directly from raw camera frames [21].

Despite these advances, the accuracy of gaze estimation remains heavily dependent on precisely capturing features from the eye region. Traditional attention mechanisms, such as SE [22], CBAM [23], ECA [24], and CA [25], have contributed to performance improvements but still face challenges in modeling global context and long-range dependencies. SE enhances channel-wise information via global pooling but neglects spatial details; ECA refines channel attention yet struggles with capturing complex dependencies; CBAM incorporates spatial attention but remains limited in modeling long-range spatial relationships; CA embeds spatial information into channel attention, but its generalization ability is weak and suffers from the reduction in channel dimensions. MSCA [26], although utilizing multi-scale convolutional kernels to improve receptive fields, introduces high computational complexity and inefficient feature fusion, making it less suitable for real-time applications.

Furthermore, achieving high-precision gaze estimation under low-light conditions remains an open and significant challenge, as such environments introduce additional noise and variability that further complicate accurate feature extraction and prediction.

## 2.2 Gaze Estimation in Low-Light Scenarios

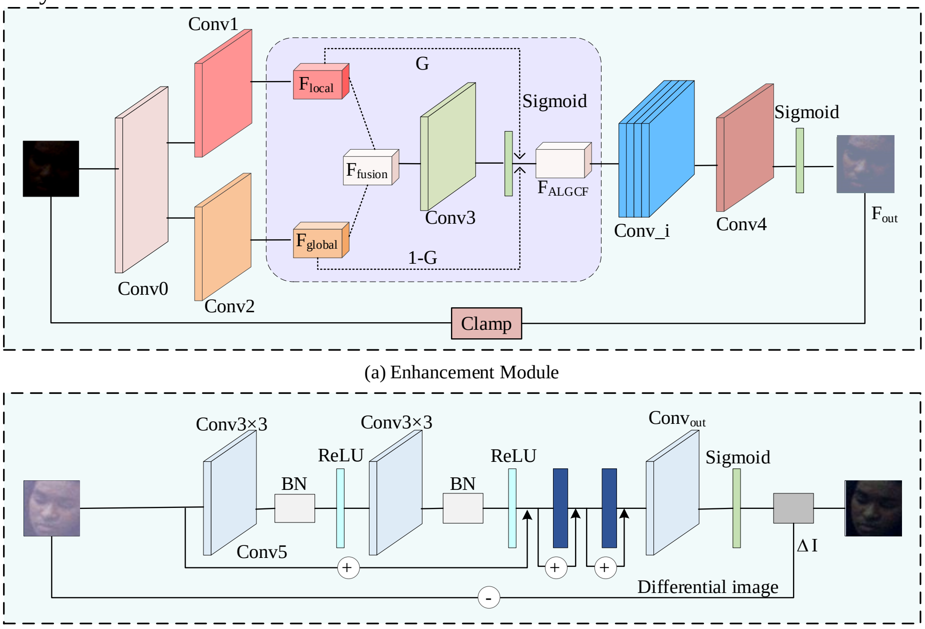
Gaze estimation in low-light conditions remains a significant challenge. The primary approaches are using infrared cameras or low-light RGB images for gaze regression tasks. While infrared cameras are effective, they are expensive and raise privacy concerns. For low-light RGB gaze regression, X. Zhang et al. [27] trained models by adjusting illumination levels, and Tobias Fischer et al. [28] used grayscale conversion and histogram equalization for diverse lighting conditions. However, these deep networks still struggle to effectively capture eye details and lack robustness.

## 3 Proposed Method

We propose the LLSGE-Net model, a deep learning framework for gaze estimation in low-light environments. It integrates an image enhancement module with a gaze estimation module.

### 3.1 Low-light Image Enhancement Model

Low-light image enhancement aims to reveal hidden information and improve image quality. Inspired by the Self-Calibrated Illumination (SCI) method [29], we introduce the ALGCF module in the enhancement network to address the dynamic nature of eye features in gaze estimation tasks. The image enhancement module (**Fig. 2** (a)) boosts eye feature brightness and contrast by fusing multi-scale features and extracting global context information. The calibration module (**Fig. 2** (b)) refines the image through multiple layers of convolution, batch normalization, and activation functions, enhancing high-level feature expressiveness and robustness. Dark blue indicates two of the operations and a residual link. This process enhances and calibrates the model synergistically.



(a) Enhancement Module

(b) Calibration Module

**Fig. 2.** Overview of the low-light image enhancement framework.

The enhancement module extracts multi-scale features from the input image and integrates them with global contextual information to ultimately achieve detail enhancement and lighting balance.

The input image passes through the initial convolutional layer, extracting the fundamental features $F_0$:

$$F_0 = ReLU(Conv0(I)) \tag{1}$$

Where $Conv0$ is the initial convolutional layer that extracts features from the input image through convolutional kernels.

The ALGCF module is then applied to fuse local and global information. The local feature extractor captures detailed information from the eye region:

$$F_{local} = Conv1(F_0) \tag{2}$$

Where $Conv1$ is the convolutional layer of the local feature extractor. By using 3×3 depthwise separable convolution, local details of each channel are independently extracted, focusing on capturing eye-specific details from the initial feature map.

The global context information extractor captures overall facial illumination and structural information:

$$F_{global} = In(Conv2(AAP(F_0)), size = F_{local}) \tag{3}$$

Where the adaptive average pooling layer $AAP$ extracts illumination and structural information from the global context. $Conv2$ uses a 1×1 convolution kernel to generate a global feature map.

Local and global features are then combined to generate fusion weights. Adaptive feature fusion is performed using these fusion weights:

$$\begin{cases} F_{fusion} = Concat(F_{local}, F_{global}) \\ G = \sigma(Conv3(F_{fusion})) \\ F_{ALGCF} = G \times F_{local} + (1 - G) \times F_{global} \end{cases} \tag{4}$$

Where the $Concat$ operation combines local and global features into a fused feature map. $Conv3$ is the convolution layer of the gating mechanism, where $\sigma$ represents the $Sigmoid$ function. The fused feature $F_{ALGCF}$ adaptively combines local and global features according to the weights $G$ and $1-G$, ensuring global illumination consistency while preserving local eye details in the facial feature map.

The fused feature map is further processed through multiple convolutional layers to extract more features:

$$F_{i+1} = F_i + ReLU(BN(Conv_i(F_i))) \tag{5}$$

Where $Conv_i$ represents the convolution layer of each convolutional block, $BN$ performs feature normalization. Through residual connections (+), each convolutional block extracts features while preserving the integrity of the input features.

The output convolution layer converts the fused feature map into an enhanced image:

$$F_{output} = \sigma(Conv4(F_{i+1})) \tag{6}$$

Where $Conv4$ is the output convolution layer, converting the feature map into the final enhanced image. The activation function normalizes the pixel values to the range (0,1).

Finally, the enhanced feature map is added to the input image to obtain the final enhanced image:

$$I_{enhanced} = Clamp(F_{output} + I, 0,1) \tag{7}$$

Where the $Clamp$ function ensures that the pixel values of the enhanced image are within a reasonable range.

The enhanced image $I_{enhanced}$ passes through the initial convolution layer to obtain the calibrated feature map:

$$F_{calib(0)} = ReLU(BN(Conv5(I_{enhanced}))) \tag{8}$$

Where $Conv5$ is the input convolution layer of the calibration network, extracting the calibrated feature map.

The calibrated feature map $F_{calib(0)}$ undergoes multiple convolutional blocks for illumination and detail adjustment:

$$F_{calib(i)} = F_{calib(i\text{-}1)} + ReLU(BN(Conv_i(F_{calib(i\text{-}1)}))) \tag{9}$$

Where $F_{calib(i\text{-}1)}$ represents the output feature map of layer $(i-1)$ . $F_{calib(i)}$ is the output feature map of the convolutional block of layer $i$ , combining the original input features $F_{calib(i\text{-}1)}$ with the new features obtained through convolution, $BN$ , and $ReLU$ processing.

The final calibrated feature map passes through the output convolution layer to generate the differential image:

$$\Delta I = \sigma(Conv_{out}(F_{calib(i)})) \tag{10}$$

Where $Conv_{out}$ is the output convolution layer, generating the final difference image through convolution.

Finally, the differential image $\Delta I$ is subtracted from the enhanced input image to obtain the final calibrated image $I_{calibrated}$ :

$$I_{calibrated} = I_{enhanced} - \Delta I \tag{11}$$

## 3.2 Synergistic Mechanism Model

At each stage, the input image is first enhanced by the enhancement module to obtain the enhanced image $I_{enhanced}$ . Then, the calibration module calibrates the enhanced image to obtain the calibrated feature map $\Delta I$ . The calibrated feature map is added to the input image to generate a new input image for the next stage of processing.

$$\begin{cases} I_{enhanced}^{(i)} = EnhanceNetwork(I_{input}^{(i)}) \\ \Delta I^{(i)} = CalibrateNetwork(I_{enhanced}^{(i)}) \\ I_{input}^{(i+1)} = I + \Delta I^{(i)} \end{cases} \tag{12}$$

Where $I_{input}^{(i)}$ is the input image of the $i-th$ stage, $I_{enhanced}^{(i)}$ is the enhanced image of the $i-th$ stage, and $I$ is the input image.

As shown in **Fig. 2**, both the input and output in **Fig. 2** (a) and **Fig. 2** (b) undergo a single round of enhancement and calibration. Throughout the process, the model is enhanced and calibrated several times. This final image integrates multi-stage feature extraction, enhancement, and calibration, resulting in the optimal output. Saving this final enhanced image ensures the higher quality input for subsequent gaze estimation. Experimental results validate that after three iterations, the resulting image can be regarded as the final enhanced output.

### 3.3 Gaze Estimation Model

Traditional gaze estimation methods struggle under low-light conditions due to significant image quality degradation, which impairs accurate eye feature extraction. To address this challenge, we propose a gaze estimation model utilizing processed low-light images. The model is based on an improved ResNet18 (**Fig. 3**). ELSCA attention mechanism modules are integrated in the second, third, and fourth stages to enhance recognition and processing of key eye regions under low-light conditions. Additionally, a DFENet deep feature extraction module is incorporated in later stages to further refine high-level features. This end-to-end integrated approach enables the LLSGE-Net model to deliver exceptional performance in gaze estimation tasks under low-light conditions.

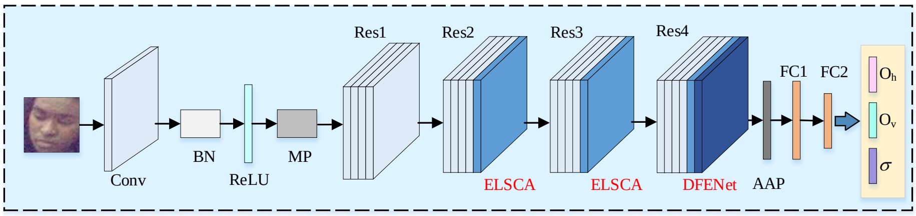


**Fig. 3.** Overview of the gaze estimation framework.

The accuracy of gaze estimation depends on capturing eye region features accurately. The proposed ELSCA attention mechanism, inspired by CA [25], improves both spatial and channel attention. By integrating feature correlation analysis, it effectively enhances eye features in low-light environments, improving the model's ability to handle long-range dependencies. The ELSCA mechanism is shown in **Fig. 4**.

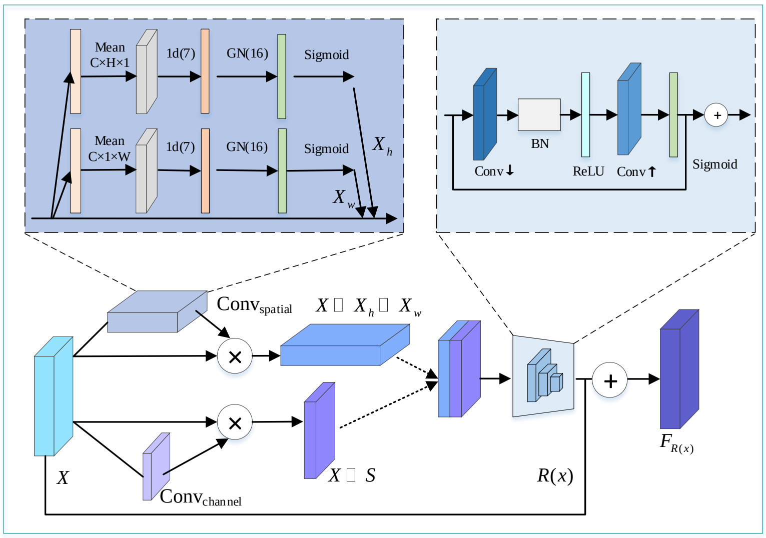


**Fig. 4.** The ELSCA attention mechanism network architecture.

Instead of global spatial pooling, strip pooling is used in the spatial dimension to extract feature vectors in horizontal and vertical directions. This maintains elongated kernel shapes to capture long-range dependencies.

The input feature map is $X \in \mathbb{R}^{B\times C\times H\times W}$, where $B$ is the batch size, $C$ is the number of channels, and $H$ and $X_w$ are the height and width of the feature map, respectively.

The width dimension is averaged to result in a feature map of size $B\times C\times H\times 1$, denoted as $\overline{X}_h$. Then, a $Conv1d$, $GN$, and $Sigmoid$ activation function are applied to obtain a feature map $X_h$ of $B\times C\times H\times 1$.

$$\begin{cases} \overline{X}_h = \dfrac{1}{W}\sum_{i=1}^{W} X[:,:,:,i], \quad \overline{X}_h \in \mathbb{R}^{B\times C\times H\times 1} \\ X_h = \sigma(GN(Conv1d(\overline{X}_h))), \quad X_h \in \mathbb{R}^{B\times C\times H\times 1} \end{cases} \tag{13}$$

Where $\overline{X}_h \in \mathbb{R}^{B\times C\times H\times 1}$, $X_h \in \mathbb{R}^{B\times C\times H\times 1}$. Here, $X[:,:,:,i]$ selects a sub-tensor of shape $(B,C,H)$, choosing only the $i$-th element along the $W$ dimension. $GN$ represents group normalization, with each group containing 16 channels. $Conv1d$ denotes the 1D convolution operation with a kernel size of 7, capturing long-range dependencies.

Then, the height dimension is averaged to result in a feature map of size

$B \times C \times 1 \times W$, denoted as $\overline{X}_w$. Next, a $Conv1d$, $GN$, and $Sigmoid$ activation function are applied to obtain a feature map $X_w$ of size $B \times C \times 1 \times W$.

$$\begin{cases} \overline{X}_w = \dfrac{1}{H} \sum_{j=1}^{H} X[:,:,j,:], \quad \overline{X}_w \in \mathbb{R}^{B \times C \times 1 \times W} \\ X_w = \sigma(GN(Conv1d(\overline{X}_w))), \quad X_w \in \mathbb{R}^{B \times C \times 1 \times W} \end{cases} \tag{14}$$

Where $\overline{X}_w \in \mathbb{R}^{B \times C \times 1 \times W}$, $X_w \in \mathbb{R}^{B \times C \times 1 \times W}$. Here, $X[:,:,j,:]$ selects a sub-tensor of shape $(B, C, W)$, choosing only the $j$-th element along the $H$ dimension.

The final input feature map $X$ is multiplied by the two attention feature maps $X_h$ and $X_w$ to obtain the spatial attention feature map $Y$:

$$Y = X \odot X_h \odot X_w, \quad Y \in \mathbb{R}^{B \times C \times H \times W} \tag{15}$$

Where $Y \in \mathbb{R}^{B \times C \times H \times W}$, and $\odot$ represents element-wise multiplication.

To further enhance the feature relationships between channels, we propose a channel attention mechanism. Adaptive average pooling reduces the feature values of each channel of the input feature map to a single value, resulting in a feature map of size $A \in \mathbb{R}^{B \times C \times 1 \times 1}$. Next, the average pooling result $A$ is processed, flattening the pooled feature map into a vector of size $\mathbb{R}^{B \times C}$.

The flattened feature map is processed by an $MLP$ network, which includes two fully connected layers and a $ReLU$ activation function. The processed result $A'$ undergoes $Sigmoid$ activation to obtain a weight map of size $S \in \mathbb{R}^{B \times C}$. The weight map is then reshaped to match the shape of the input feature map, resulting in a weight map of size $S \in \mathbb{R}^{B \times C \times 1 \times 1}$. The final output is obtained by element-wise multiplication of the input feature map $X$ and the weight map $S$, producing the channel attention map $Z$.

$$\begin{cases} A = AdaptiveAvgPool2d(X), \quad A \in \mathbb{R}^{B \times C \times 1 \times 1} \\ A' = FC2(\mathrm{Re}LU(FC1(A))), \quad A' \in \mathbb{R}^{B \times C} \\ S = \sigma(A'), \quad S \in \mathbb{R}^{B \times C \times 1 \times 1} \\ Z = X \odot S, \quad Z \in \mathbb{R}^{B \times C \times H \times W} \end{cases} \tag{16}$$

Where $A \in \mathbb{R}^{B \times C \times 1 \times 1}$, $A' \in \mathbb{R}^{B \times C}$, and $Z \in \mathbb{R}^{B \times C \times H \times W}$.

Channel attention evaluates the contribution of each channel to gaze estimation by enhancing important channel features, thereby optimizing the overall feature quality.

Next, feature correlation analysis is proposed. This integrates spatial and channel attention and optimizes the interactions between features. The spatial and channel attention-weighted features are concatenated along the channel dimension, resulting in feature $F_{\text{concat}}$:

$$F_{\text{concat}} = [Y, Z] \tag{17}$$

Where $Y$ and $Z$ are the feature maps with spatial and channel weights applied, respectively.

Then, feature correlation analysis is performed on the merged features:

$$R(x) = \sigma(Conv_{\uparrow}(ReLU(BN(Conv_{\downarrow}(F_{concat}))))) \tag{18}$$

Where $Conv_{\uparrow}$ is a dimension-reducing convolution layer that helps the module focus on capturing the most critical feature correlation information. $Conv_{\downarrow}$ is an expanding convolution layer that generates the final feature adjustment weights.

Through feature correlation analysis, the model can dynamically evaluate and adjust the relationships between features, highlighting important features while suppressing irrelevant features.

Combining the above components, the enhanced feature $F_{R(x)}$ can be obtained through the following steps:

$$F_{R(x)} = X \,□\, R(x) + W \tag{19}$$

Where □ represents element-wise multiplication. The feature correlation analysis module generates weights $R(x)$, which are multiplied element-wise with the original input $X$.

In the later stages of LLSGE-Net, DFENet is introduced to refine high-level features. The DFENet module consists of several convolution blocks. The output channels follow this sequence: 64×2、128×3、256×3、512×3 and 1024. Max pooling is applied after the 2nd, 5th, and 8th convolution blocks to reduce feature map dimensions. Each convolution block includes batch normalization and activation. Finally, an adaptive average pooling layer adjusts feature maps to 1×1×1024. The DFENet structure is shown in **Fig. 5**.

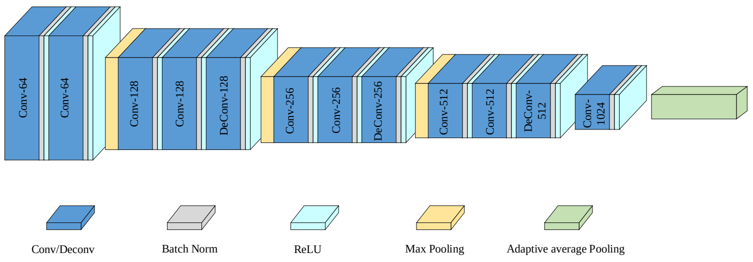


**Fig. 5.** The DFENet module network architecture.

### 3.4 Residual Connection Model

In our improved ResNet18 model, attention mechanisms are applied after the 2nd, 3rd, and 4th residual modules. Additionally, a DFENet is introduced at the end of the residual blocks to integrate and extract high-level semantic features. Finally, $fc$ maps the feature vector to a three-dimensional output vector. The first two elements are transformed by the hyperbolic tangent function to obtain the horizontal angle $O_h$ and vertical angle $O_v$ of the gaze. The third element is transformed by the $Sigmoid$ function and multiplied by $\pi$ to obtain the gaze prediction uncertainty $\sigma$ .

The loss function for this study, denoted as $MSELoss$ (mean squared error). A smaller $MSELoss$ indicates a smaller error. Therefore, The loss function of gaze estimation is:

$$\mathrm{L}_{gaze} = \frac{1}{n}\sum\left\|\xi_{gt} - \xi_{pred}\right\|_2^2 \tag{20}$$

Where the true value of the gaze is $\xi_{gt}$ , and the predicted value of the gaze estimation is $\xi_{pred}$ .

# 4 Experimental Analysis

## 4.1 Datasets

To evaluate the performance of the proposed network model, we trained and tested it on the widely-used Gaze360 dataset [30], a standard in gaze estimation tasks. We used the data preprocessing method proposed by Zhang et al. [31] to ensure consistency and effectiveness.

To simulate various low-light conditions, we applied four preprocessing strategies to the Gaze360 dataset: dark scenes (Dark_comp), extremely dark scenes (Dark_super), low illumination simulated by gamma correction (Dark_gamma), and dark scenes with unknown light source positions (Dark_light). These modifications allowed us to construct a variant of Gaze360 that reflects different real-world low-light scenarios.

Specifically, for Dark_comp and Dark_super, we adjusted brightness and contrast to enhance the realism of nighttime environments. Brightness was modified by compressing the dynamic range of dark and bright regions, while contrast adjustments focused on narrowing the distribution of pixel intensities to strengthen the distinction between illuminated and dark areas.

For Dark_gamma, we simulated low-light environments through gamma correction, selecting a gamma value of 0.7. Lower gamma values darken images, and a value of 0.7 was chosen to realistically replicate low-light visual conditions while maintaining sufficient detail for gaze estimation.

To further explore the challenges of gaze estimation under complex illumination settings, we developed the Dark_light subset, simulating environments such as university lecture halls, planetariums, and nighttime driving scenarios. These scenes feature strong directional lighting and extensive dark regions, significantly increasing the difficulty of gaze estimation.

To generate the Dark_light data, we applied a two-step processing pipeline: first, the overall image brightness was reduced to deepen dark regions; second, localized illumination patterns were introduced at random positions using Gaussian-blurred illumination masks, simulating point light sources such as streetlights. This approach enhanced the realism and spatial complexity of the scenes, creating more challenging and naturalistic low-light conditions for evaluation.

Using these methods, we successfully built a new dataset that captures a variety of low-light conditions, with sample results shown in **Fig. 6**.

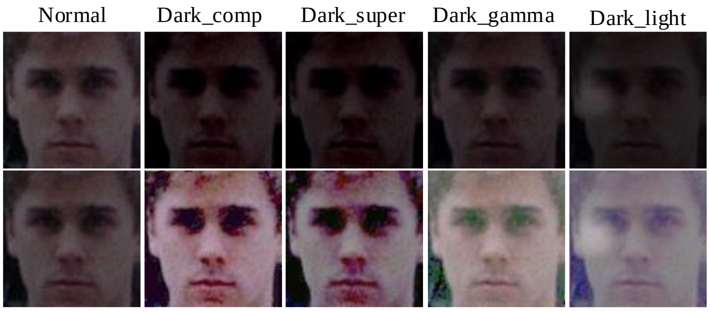


**Fig. 6.** Low-light Image Processing Dataset.

As shown in **Fig. 6**, the first row displays the visual effects after dataset processing, while the second row shows the results after enhancement by the low-light image enhancement model.

We use the main evaluation index angle error of gaze estimation to compare the performance with other gaze estimation models. Assuming the actual gaze direction is $\mathbf{g} \in \mathbb{R}^3$ and the estimated gaze direction is $\mathbf{g} \in \mathbb{R}^3$, the angular error can be calculated as:

$$\mathrm{L}_{\text{angular}} = \frac{\mathbf{g} \cdot \hat{\mathbf{g}}}{\|\mathbf{g}\| \|\mathbf{g}\|} \tag{21}$$

### 4.2 Experimental Results

**Analysis of Comparative study.** The comparative experiments results are presented in **Table 1**.The proposed method achieves lower angular errors than other advanced methods under all low-light conditions, demonstrating superior performance.

To visually demonstrate the superiority of our approach, we performed a visualization analysis of the gaze estimation results using the image with the test set sequence number 186811.jpg from the Gaze360 dataset. The visualization results are shown in **Fig. 7**.

**Table 1.** Performance evaluation on the Gaze360 dataset under four designed low-light envronments, the results are angular error (°).

| Module | Dark_comp | Dark_super | Dark_gamma | Dark_light |
|---|---|---|---|---|
| FullFace [32] | 16.82 | 17.21 | 15.91 | 16.51 |
| RT-Gene [28] | 14.30 | 15.42 | 13.07 | 12.57 |
| Dilated-Net [33] | 14.67 | 15.23 | 13.40 | 13.64 |
| Gaze360 [30] | 12.99 | 13.92 | 12.46 | 11.94 |
| CA-Net [34] | 12.55 | 13.46 | 12.25 | 11.78 |
| GazeTR [35] | 12.36 | 12.67 | 11.05 | 11.55 |
| L2CS-Net [36] | 12.26 | 13.87 | 12.24 | 11.86 |
| **LLSGE-Net (ours)** | **11.75** | **12.17** | **10.43** | **11.27** |

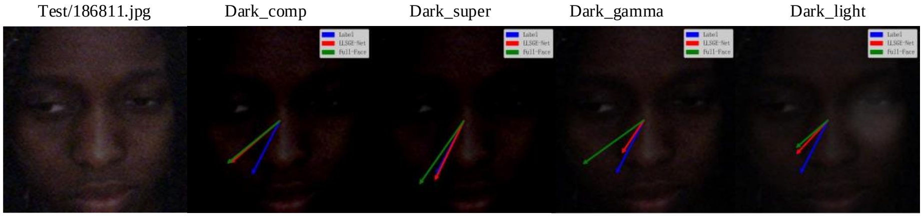


**Fig. 7.** Visualization of the comparison results between LLSGE-Net and FullFace.

As shown in **Fig. 7**, the blue arrows represent the ground truth gaze directions (Label), the red arrows indicate the predictions from the proposed LLSGE-Net model, and the green arrows correspond to the predictions from the classical FullFace [32]. It can be observed that the predictions of LLSGE-Net are generally closer to the ground truth, with noticeably smaller angular deviations compared to those of FullFace. This demonstrates that the proposed model achieves higher prediction accuracy and stability across the four low-light conditions.

The detailed visualization results are shown in **Table 2**.

**Table 2.** Comparison of Experimental Results between LLSGE-Net and FullFace.

| Module | Dark_comp | Dark_super | Dark_gamma | Dark_light |
|---|---|---|---|---|
| FullFace | 16.32° | 15.76° | 22.49° | 16.39° |
| **LLSGE-Net** | **14.40°** | **4.71°** | **13.43°** | **12.59°** |

**Analysis of Ablation Study.** We used ResNet18 as the baseline model (B). First, we added the low-light image enhancement module (L). Next, the attention module (E) was added to the baseline and L modules. Finally, we added the deep feature extraction module (D) to refine high-level features further. The results of the ablation study are shown in **Table 3**. The best results are highlighted in bold.

**Table 3.** Comparison of Experimental Results between LLSGE-Net and FullFace.

| Module | Dark_comp | Dark_super | Dark_gamma | Dark_light |
|---|---|---|---|---|
| B | 13.50 | 14.58 | 12.78 | 12.35 |
| B+L | 12.64 | 13.29 | 12.14 | 11.92 |
| B+L+E | 11.80 | 12.52 | 10.58 | 11.50 |
| **B+L+E+D** | **11.75** | **12.17** | **10.43** | **11.27** |

From the experimental results, it can be seen that using only ResNet18 as the baseline network (B) yields average performance.

**Analysis of ELSCA Attention Mechanism Effectiveness.** We selected SE [22], CBAM [23], ECA [24], CA [25] and MSCA [26] for comparison. The baseline for this comparison is (B+L+D) from the ablation study, which includes all modules except the ELSCA attention mechanism. The evaluation results of different attention mechanisms are shown in **Table 4**, with the best results highlighted in bold.

**Table 4.** Evaluation results of different attention mechanisms, the results are angular error (°).

| Attention | Dark_comp | Dark_super | Dark_gamma | Dark_light |
|---|---|---|---|---|
| SE [22] | 12.34 | 12.56 | 11.85 | 11.9 |
| CBAM [23] | 12.13 | 13.00 | 11.83 | 11.84 |
| ECA [24] | 11.98 | 12.61 | 11.32 | 11.89 |
| CA [25] | 11.95 | 12.87 | 11.21 | 11.79 |
| MSCA [26] | 12.28 | 13.2 | 11.07 | 11.75 |
| **ELSCA** | **11.75** | **12.17** | **10.43** | **11.27** |

From the experiments, it can be observed that the ELSCA attention mechanism shows significant advantages over other traditional attentions.

**Analysis of Low-light Image Enhancement Model Effectiveness.** We selected FullFace [32], RT-Gene [28], Dilated-Net [33], and Gaze360 [30] for validation. The baseline for these experiments involves first processing through the low-light image enhancement module (L) and then through each respective network model. Experiments were conducted under four low-light conditions. The results are shown in **Table 5**.

The experimental data shows that under four different lighting conditions, all models performed better after processing with the low-light image enhancement module (L).

**Analysis of Gaze Estimation Model Effectiveness.** This paper focuses on gaze estimation in low-light scenarios. To validate the effectiveness of the gaze estimation module in all-day scenarios, additional experiments were conducted on the Gaze360 dataset. The results are shown in **Table 6**, with the best performance highlighted in **bold**.

**Table 5.** The results of different models' evaluation with and without the low-light image enhancement module (L), the results are angular error (°).

| Models | Dark_comp | Dark_super | Dark_gamma | Dark_light |
|---|---|---|---|---|
| FullFace | 16.82 | 17.21 | 15.91 | 16.51 |
| L+FullFace | **15.02** | **16.87** | **14.07** | **15.58** |
| RT-Gene | 14.30 | 15.42 | 13.07 | 12.57 |
| L+RT-Gene | **13.98** | **15.02** | **12.71** | **12.05** |
| Dilated-Net | 14.67 | 15.23 | 13.40 | 13.64 |
| L+Dilated-Net | **14.35** | **14.39** | **12.87** | **13.01** |
| Gaze360 | 12.99 | 13.92 | 12.46 | 11.94 |
| L+Gaze360 | **12.54** | **13.02** | **11.96** | **11.38** |

**Table 6.** Performance evaluation of our gaze estimation method (ours) compared with advanced methods under the original Gaze360 dataset. The results are average angular error (°).

| Module | Gaze360 |
|---|---|
| FullFace [32] | 14.94 |
| RT-Gene [28] | 12.73 |
| Dilated-Net [33] | 13.73 |
| Gaze360 [30] | 11.47 |
| CA-Net [34] | 11.20 |
| GazeTR [35] | 10.85 |
| L2CS-Net [36] | 10.41 |
| **(Ours)** | **10.23** |

Based on the data in **Table 6**, our gaze estimation model (Ours) demonstrates excellent performance under the original Gaze360 dataset.

## 5 Conclusion

This paper proposes the LLSGE-Net framework, which combines low-light image enhancement with gaze estimation to address challenges in low-light classroom environments. It shifts gaze estimation from coarse-grained to fine-grained student behavior analysis. Unlike existing methods focused on near-field applications, LLSGE-Net enhances eye detail visibility and image quality through a synergistic enhancement and calibration module, offering a novel solution for classroom attention detection. The framework integrates multi-scale feature fusion, global information extraction, and the Enhanced Local Spatial and Global Channel Attention (ELSCA) mechanism to improve accuracy and robustness. Extensive experiments on the Gaze360 dataset demonstrate LLSGE-Net's superiority in both low-light and well-lit conditions, making it a valuable tool for intelligent classroom behavior monitoring and human-computer interaction.

**Acknowledgments.** This work was supported in part by the National Natural Science Foundation of China under Grant 62002179.

**Disclosure of Interests.** The authors have no competing interests to declare that are relevant to the content of this article.

## References

1. Zhang M, Zhao Q, Li J, et al. Research on Hidden Mind-Wandering Detection Algorithm for Online Classroom Based on Temporal Analysis of Eye Gaze Direction[C]//International Conference on Intelligent Computing. Singapore: Springer Nature Singapore, 2024: 188-196.
2. Feng Y. Information Security Protection for Online Education Based on Blockchain[C]//International Conference on Intelligent Computing. Singapore: Springer Nature Singapore, 2023: 466-474.
3. Huang C J, Luo Y C, Chen C H, et al. A Learning Assistance Tool for Enhancing ICT Application Ability of Elementary and Secondary School Students[C]//Advanced Intelligent Computing Theories and Applications. With Aspects of Theoretical and Methodological Issues: 4th International Conference on Intelligent Computing, ICIC 2008 Shanghai, China, September 15-18, 2008 Proceedings 4. Springer Berlin Heidelberg, 2008: 661-668.
4. Qiang Z, Nagai Y, Fukuda H, et al. Impression Evaluation of Chat Robot with Bodily Emotional Expression Incorporating Large-Scale Language Models[C]//International Conference on Intelligent Computing. Singapore: Springer Nature Singapore, 2024: 439-449.
5. Wu H, Li X, Yang D, et al. DocBAN: An Efficient Biaffine Attention Network for Document-Level Named Entity Recognition[C]//International Conference on Intelligent Computing. Singapore: Springer Nature Singapore, 2024: 65-76.
6. Song P H, Li Y, Xin G Y, et al. Evaluating Effect of Classroom Interior Layouts on Crowd Evacuation Efficiency Using Improved Evacuation Model[C]//International Conference on Intelligent Computing. Singapore: Springer Nature Singapore, 2024: 171-180.
7. Wang T, Zheng C. Face Swapping via Reverse Contrastive Learning and Explicit Identity-Attribute Disentanglement[C]//International Conference on Intelligent Computing. Singapore: Springer Nature Singapore, 2024: 216-227.
8. Li H, Jia K, Jin Z, et al. Weakly Supervised Object Localization Based on Implicit Spatial Constraints[C]//International Conference on Intelligent Computing. Singapore: Springer Nature Singapore, 2024: 416-429.
9. Shi Y, Li X, Ouyang D, et al. Multimodal Usability of Human-Computer Interaction Based on Task-Human-Computer-Centered Design[C]//Intelligent Computing Theories and Application: 13th International Conference, ICIC 2017, Liverpool, UK, August 7-10, 2017, Proceedings, Part I 13. Springer International Publishing, 2017: 459-468.
10. Mazzoleni S, Battini E, Buongiorno D, et al. Design and Development of a Robotic Platform Based on Virtual Reality Scenarios and Wearable Sensors for Upper Limb Rehabilitation and Visuomotor Coordination[C]//Intelligent Computing Methodologies: 15th International Conference, ICIC 2019, Nanchang, China, August 3–6, 2019, Proceedings, Part III 15. Springer International Publishing, 2019: 704-715.
11. Xi Y, Guo J, Ma M. YOLO-BS: A Better Object Detection Model for Real-Time Driver Behavior Detection[C]//International Conference on Intelligent Computing. Singapore: Springer Nature Singapore, 2024: 78-89.
12. Y. Cheng and F. Lu, “DVGaze: Dual-view gaze estimation,” in Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 20632–20641, 2023.

13. V. Nagpure and K. Okuma, “Searching efficient neural architecture with multi-resolution fusion transformer for appearance-based gaze estimation,” in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 890–899, 2023.
14. H. Wang, J. O. Oh, H. J. Chang, J. H. Na, M. Tae, Z. Zhang, and S.-I. Choi, “GazeCaps: Gaze estimation with self-attention-routed capsules,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2669–2677, 2023.
15. S. Jin, J. Dai, and T. Nguyen, “Kappa angle regression with ocular counter-rolling awareness for gaze estimation,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2659–2668, 2023.
16. X. Cai, J. Zeng, S. Shan, and X. Chen, “Source-free adaptive gaze estimation by uncertainty reduction,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 22035–22045, 2023.
17. H. Zhang, S. Wu, W. Chen, Z. Gao, and Z. Wan, “Self-calibrating gaze estimation with optical axes projection for head-mounted eye tracking,” IEEE Transactions on Industrial Informatics, vol. 20, no. 2, pp. 1397–1407, 2023.
18. R. Liu, Y. Liu, H. Wang, and F. Lu, “PNP-GA+: Plug-and-play domain adaptation for gaze estimation using model variants,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024.
19. Y. Hisadome, T. Wu, J. Qin, and Y. Sugano, “Rotation-constrained cross-view feature fusion for multi-view appearance-based gaze estimation,” in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 5985–5994, 2024.
20. M. Zhang, Y. Liu, and F. Lu, “GazeOnce: Real-time multi-person gaze estimation,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4197–4206, 2022.
21. H. Balim, S. Park, X. Wang, X. Zhang, and O. Hilliges, “EFE: End-to-end frame-to-gaze estimation,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2688–2697, 2023.
22. J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141, 2018.
23. 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), pp. 3–19, 2018.
24. Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, “ECA-Net: Efficient channel attention for deep convolutional neural networks,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11534–11542, 2020.
25. Q. Hou, D. Zhou, and J. Feng, “Coordinate attention for efficient mobile network design,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13713–13722, 2021.
26. M.-H. Guo, C.-Z. Lu, Q. Hou, Z. Liu, M.-M. Cheng, and S.-M. Hu, “SegNext: Rethinking convolutional attention design for semantic segmentation,” Advances in Neural Information Processing Systems, vol. 35, pp. 1140–1156, 2022.
27. X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, “MPIIGaze: Real-world dataset and deep appearance-based gaze estimation,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 1, pp. 162–175, 2017.
28. T. Fischer, H. J. Chang, and Y. Demiris, “RT-GENE: Real-time eye gaze estimation in natural environments,” in Proceedings of the European Conference on Computer Vision (ECCV), pp. 334–352, 2018.
29. L. Ma, T. Ma, R. Liu, X. Fan, and Z. Luo, “Toward fast, flexible, and robust low-light image enhancement,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5637–5646, 2022.

30. P. Kellnhofer, A. Recasens, S. Stent, W. Matusik, and A. Torralba, “Gaze360: Physically unconstrained gaze estimation in the wild,” in Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6912–6921, 2019.
31. Y. Cheng, H. Wang, Y. Bao, and F. Lu, “Appearance-based gaze estimation with deep learning: A review and benchmark,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024.
32. X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, “It’s written all over your face: Full-face appearance-based gaze estimation,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 51–60, 2017.
33. Z. Chen and B. E. Shi, “Appearance-based gaze estimation using dilated convolutions,” in Asian Conference on Computer Vision, pp. 309–324, Springer, 2018.
34. Y. Cheng, S. Huang, F. Wang, C. Qian, and F. Lu, “A coarse-to-fine adaptive network for appearance-based gaze estimation,” in Proceedings of the AAAI Conference on Artificial Intelligence, pp. 10623–10630, 2020.
35. Y. Cheng and F. Lu, “Gaze estimation using transformer,” in 2022 26th International Conference on Pattern Recognition (ICPR), pp. 3341–3347, IEEE, 2022.
36. Abdelrahman, T. Hempel, A. Khalifa, A. Al-Hamadi, and L. Dinges, “L2CS-Net: Fine-grained gaze estimation in unconstrained environments,” in 2023 8th International Conference on Frontiers of Signal Processing (ICFSP), pp. 98–102, IEEE, 2023.