applied sciences

MDPI

Article

A Study of Potential Applications of Student Emotion Recognition in Primary and Secondary Classrooms

Yimei Huang, Wei Deng * and Taojie Xu

Hubei Key Laboratory of Digital Education, Central China Normal University, Wuhan 430079, China
* Correspondence: sdengwei@mail.ccnu.edu.cn

Abstract: Emotion recognition is critical to understanding students' emotional states. However, problems such as crowded classroom environments, changing light, and occlusion often affect the accuracy of recognition. This study proposes an emotion recognition algorithm specifically for classroom environments. Firstly, the study adds the self-made MCC module and the Wise-IoU loss function to make object detection in the YOLOv8 model more accurate and efficient. Compared with the native YOL0v8x, it reduces the parameters by 16% and accelerates the inference speed by 20%. Secondly, in order to address the intricacies of the classroom setting and the specific requirements of the emotion recognition task, a multi-channel emotion recognition network (MultiEmoNet) has been developed. This network fuses skeletal, environmental, and facial information, and introduces a central loss function and an attention module AAM to enhance the feature extraction capability. The experimental results show that MultiEmoNet achieves a classification accuracy of 91.4% on a homemade classroom student emotion dataset, which is a 10% improvement over the single-channel classification algorithm. In addition, this study also demonstrates the dynamic changes in students' emotions in the classroom through visual analysis, which helps teachers grasp students' emotional states in real time. This paper validates the potential of multi-channel information-fusion deep learning techniques for classroom teaching analysis and provides new ideas and tools for future improvements to emotion recognition techniques.

Keywords: YOLOv8; object detection; emotion recognition; pedagogical analysis

Citation: Huang, Y.; Deng, W.; Xu, T. A Study of Potential Applications of Student Emotion Recognition in Primary and Secondary Classrooms. *Appl. Sci.* **2024**, *14*, 10875. https://doi.org/10.3390/app142310875

Academic Editor: Douglas O'Shaughnessy

Received: 31 October 2024
Revised: 20 November 2024
Accepted: 21 November 2024
Published: 24 November 2024

1. Introduction

Students' emotional state is not only an important variable in the learning process, but also directly affects their motivation, attention, and engagement [1]. As an important application of artificial intelligence in the field of education, emotion recognition technology is gradually becoming an important tool for enhancing teaching effectiveness in primary and secondary school classrooms, and it also has a profound impact on personalized support for students. Therefore, the study of classroom emotion recognition technology can dynamically monitor and feedback students' emotional states, providing teachers with timely emotional information support. This technology not only has important theoretical value, but also has profound practical significance for improving teaching in practice.

Deep learning techniques have made significant progress in emotion recognition in recent years. For example, convolutional neural networks (CNN) and recurrent neural networks (RNN) have demonstrated significant potential in facial expression and speech emotion recognition [2,3], and these models are able to accurately capture complex emotional features, and improve the accuracy and efficiency of emotion recognition. Compared with traditional emotion assessment methods, deep learning-based emotion recognition technology can provide a more objective and real-time emotion assessment, effectively reducing the influence of human subjective factors. This provides key technical support for teachers to quickly and accurately perceive students' emotional states and then implement personalized teaching.

However, there are challenges to applying deep learning and emotion recognition technology in primary and secondary classrooms. Firstly, the changing positions and postures of students and teachers in classroom environments increase the difficulty of emotion recognition and place higher demands on the robustness of the techniques. Secondly, environmental disturbances, such as classroom background noise [4] and lighting intensity and distribution [5], can weaken the accuracy of emotion recognition. In addition, occlusion problems [6] prevent the full capture of students' facial expressions and body features, which in turn affects the recognition accuracy. In order to effectively address these issues, it is essential to enhance the adaptability and precision of the technique.

Based on this, the main objectives of this study are as follows:

- Improving deep learning-based object detection algorithms to meet the special needs of teaching scenarios for the difficulty of student object detection due to the large number of people in complex classroom environments.
- To develop a multi-channel emotion recognition network aimed at enhancing the accuracy and stability of emotion recognition, addressing the current gap in research attributed to the lack of students' facial expression features within classroom environments.
- This study will provide a visual analysis of student emotions within the classroom, offering comprehensive and accurate data for the continuous monitoring of classroom emotions.

2. Related Work

2.1. Student Object Detection Algorithm

In a classroom setting, accurate student object detection is a prerequisite for classroom emotion recognition. Object detection networks are divided into two main categories: Single-Stage Detectors and Two-Stage Detectors. Single-stage detectors generate target frames and categories simultaneously through a single forward pass, which is fast and suitable for real-time applications. YOLO and SSD are representative algorithms, with the former detecting through a single convolutional network and the latter using multi-scale feature maps to enhance the detection of different target sizes. Chen et al. [7] made student behavior detection in classroom environments more efficient based on the YOLOv4 behavior detection algorithm. A two-stage detector generates candidate frames first, then classifies and locates them with higher accuracy. R-CNN, along with its improved Faster R-CNN, enhances the detection efficiency and accuracy of the regional proposition network (RPN). For example, studies have pointed out the high accuracy of Faster R-CNN in detecting specific student postures [8]. Although these methods offer a preliminary technical foundation for object detection, they still require improvement in their adaptability to complex scenes.

2.2. Classroom Student Emotional Classification

Emotion categorization theory suggests that emotions are adaptive responses of individuals to external stimuli and have evolved over the course of evolution. Ekman [9] first proposed a theory of six basic emotions: happiness, sadness, anger, fear, surprise, and disgust. Plutchik [10] further extended this categorization by proposing eight basic emotions: anger, fear, sadness, disgust, surprise, anticipation, acceptance, and joy. Although these models provide a theoretical framework, there are limitations to these classifications for the identification of emotions in a classroom setting.

Emotional expressions in classroom contexts differ from traditional emotion categorization. Although Ekman's emotion theory is a landmark in the field of emotion recognition, its categorization may not fully encompass the complexity of students' emotions in a particular classroom setting. In order to adapt to classroom scenarios, researchers have developed their own criteria for categorizing emotions based on practical needs. For example, Pekrun et al. [1,11] proposed his Control-Value Theory, which categorizes emotions into four dimensions: positive and negative emotions and activated and inactivated emotions. This framework provides refined theoretical support for analyzing classroom emotions and is particularly suitable for emotion recognition in educational scenarios. Kosti et al. [12]

proposed a classification method based on emotion vocabulary using about 400 emotion words, which were grouped into 26 sentiment categories by cluster analysis (Table 1). This approach not only considers the similarity of emotional words but also pays special attention to the visually separable features of emotions. For this study, we selected six emotion categories related to classroom teaching and learning: Pleasure, Peace, Engagement, Fatigue, Doubt, and Disconnection. The selection of these emotion categories is closer to the reality of teaching and provides a more accurate reference for emotion recognition in teaching scenarios.

Table 1. Classification of 26 distinct emotions [12].

No	Emotion Categories	No	Emotion Categories
1	Affection	14	Excitement
2	Anger	15	Fatigue
3	Annoyance	16	Fear
4	Anticipation	17	Happiness
5	Aversion	18	Pain
6	Confidence	19	Peace
7	Disapproval	20	Pleasure
8	Disconnection	21	Sadness
9	Disquietment	22	Sensitivity
10	Doubt/Confusion	23	Suffering
11	Embarrassment	24	Surprise
12	Engagement	25	Sympathy
13	Esteem	26	Yearning

2.3. Current Status of Emotion Recognition Technology

Emotional recognition determines the emotional state of an individual by analyzing their behavior, mental state, or physiological signals [13]. Mehrabian's study revealed that facial expressions account for 55% of emotional information, intonation for 38%, and speech words for only 7% [14]. This has led to the mainstream of facial expressions in emotion recognition, prompting numerous studies to extract emotions from facial features. For example, Sarvakar et al. [15] used CNN to detect different types of facial expressions. However, there are limitations in emotion recognition for individual facial expressions, and combining other information sources, such as body gestures and background scenes, has emerged as a new research direction.

Researchers have also verified the importance of body posture in emotional expression. Coulson [16] demonstrated the strong association between posture and emotion by analyzing joint movements in static postures. Gavrilescu [17] further combined neural networks and body posture for emotion analysis to improve the accuracy of recognition. Integrating a multi-channel model of facial expression and limb posture allows for a more comprehensive reflection of an individual's emotional state.

Scene contextual information also plays a key role in emotion recognition. The accuracy of emotion recognition significantly improves when facial expressions are consistent with the scene [18]. Kosti et al. [19] proposed a deep learning-based context-aware emotion recognition method that fuses character and scene features to improve the classification accuracy. Lee et al. [20], on the other hand, utilized graph convolutional networks (GCNs) to encode contextual information and enhance the robustness of emotion recognition. Yang et al.'s [21] contextual causal intervention module (CCIM) further improves the performance of the context-aware model by reducing the scene context bias.

In classroom contexts, emotion recognition techniques help teachers adjust teaching strategies based on students' emotional feedback. However, classroom emotion recognition faces challenges such as the scarcity of datasets, occlusion effects, and annotation accuracy. For this reason, researchers have introduced deep learning techniques to improve the accuracy of classroom emotion recognition. Sharma et al. [22] employed a CNN and transfer learning techniques to identify and categorize students' facial emotions in classroom

settings, utilizing the VGG16 model. Chen et al. [23] created a spatio-temporal residual attention network (STRAN)-based model to pull out features of students' facial expressions. They were able to achieve an 80.45% recognition rate on a dataset of students' facial expressions that they made themselves.

These studies have shown that combining multiple information sources can improve the comprehensiveness and accuracy of emotion recognition. However, current classroom emotion recognition still primarily relies on facial expressions [24], and the changing light and occlusion in classroom environments easily obscure facial features, thereby affecting its accuracy and robustness [25].

3. Methodologies

Based on the above review, this study proposes a student object detection algorithm that enhances the YOLOv8 model with the aim of improving the accuracy of student location detection and providing more reliable data support for emotion analysis. Meanwhile, to address the impact of student behaviors such as facial occlusion, head down, or sideways on emotion recognition, this study introduces MultiEmoNet, which aims to overcome the limitations of single facial feature recognition in situations where information is scarce and enhance the robustness of emotion recognition.

3.1. Improved Student Object Detection Algorithm Based on YOLOv8

3.1.1. YOLOv8 Framework

The Ultralytics team developed YOLOv8, a real-time object detection model. YOLOv8 does a better job of finding small and dense targets than the previous version because it has a better feature extraction network, an adaptive anchor mechanism, and a multi-scale feature fusion strategy [26]. However, in complex classroom scenarios, the model frequently results in missed detection and false detection during feature extraction due to factors such as changing light conditions, student postures, and susceptibility to occlusion. This study proposes an improvement scheme based on the MCC module and the WIoU loss function to replace the original C2f module and the CIoU loss function to achieve better feature extraction and loss optimization. Figure 1 displays the YOLOv8x-Improved model.

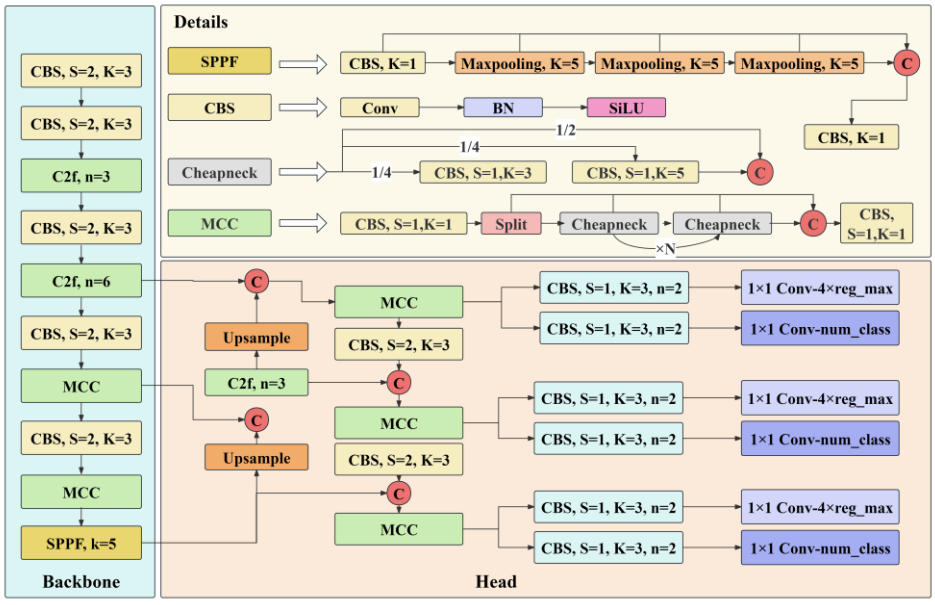


Figure 1. The network architecture of YOLOv8x-Improved.

3.1.2. MCC Module

YOLOv8 enhances the accuracy and efficiency of the model by replacing the C3 module with the C2f module in the CSPLayer to enhance the feature processing capability

and reduce the computational complexity. Practice shows that adjusting the CSPLayer can effectively improve the network performance. When it comes to finding targets, multi-scale features are very important. Adding null convolution helps the growth of multi-scale pyramid models, which greatly enhance the accuracy of localization. Different sensory fields capture detailed and overall information, and after integration, models that integrate multi-scale pyramid structures perform well in object detection. For this reason, this paper proposes the multi-group cheap convolution (MCC) module as an alternative to the C2f module to handle multi-scale features more efficiently and optimize the computational cost.

The MCC module is shown in Figure 2 with the following steps:

1. The input tensor is split into two parts, where x_group occupies half of the channel and x_cheap occupies the other half.
2. The x_group part is rearranged to facilitate processing by group. The term 'group' here refers to the set of convolutional layers with different convolutional kernel scales.
3. Convolutional layers with different kernel sizes (e.g., 3×3, 5×5) are applied to each group of x_group to capture features at different scales.
4. The processed x_group is connected to the unprocessed x_cheap in the channel dimension to combine features at different scales with the original features.
5. Finally, a 1×1 convolutional layer is applied over the fused features in order to integrate all features while keeping the number of channels constant and outputting the final feature tensor.

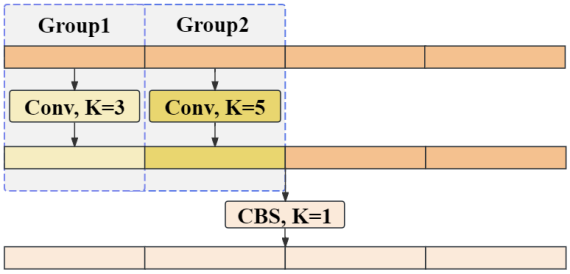


Figure 2. Schematic diagram of the MCC module.

As illustrated in Equation (1), this represents the calculation process for the MCC module.

$$MCC(X) = CBS\left(Concat\left(CW_{k_1*k_1}(x_1), \ldots..CW_{k_n*k_n}(x_n)\right)\right) \quad (1)$$

where $X = [x_1, x_2 \ldots..x_n]$ denotes splitting the input feature X into multiple heads in the channel dimension, $k_i \in \{3, 5, \ldots, k\}$ denotes a monotonic increase in the kernel size by 2, and C denotes the convolution operation, where CBS is a combination of convolution, batch normalization, and a sequence of activation functions, SILU.

To assess the efficiency of the MCC module, this study compares the number of parameters in the YOLOv8 model following the implementation of both the C2f and MCC modules. Table 2 shows the parameter changes after the MCC module. The findings indicate that the MCC module reduces the number of parameters by approximately 25% relative to the C2f module, while also achieving close to a 20% reduction at the model level. This significant decrease enhances both the network efficiency and inference speed. By retaining the first two C2f modules within the Backbone and maintaining one C2f module in the Head section, this study strikes a balance between the model performance and parameter efficiency. This approach ensures that sufficient feature information is captured effectively, allowing for high accuracy in object detection despite reductions in the parameters.

Table 2. Parameters of YOLOv8 model with replaced C2f module.

No	From	N	Params	Module	Arguments
0	−1	1	2320	Conv	[3, 80, 32]
1	−1	1	115,520	Conv	[80, 160, 3, 2]
2	−1	3	436,800	C2f	[160, 160, 3, True]
3	−1	1	461,440	Conv	[160, 320, 3, 2]
4	−1	6	3,281,920	C2f	[320, 320, 6, True]
5	−1	1	1,844,480	Conv	[320, 640, 3, 2]
6	−1	6	9,509,760 (Reduced by 27%)	C2f_MCC	[640, 640, 6, True]
7	−1	1	3,687,680	Conv	[640, 640, 3, 2]
8	−1	3	5,165,760 (Reduced by 26%)	C2f_MCC	[640, 640, 3, True]
9	−1	1	1,025,920	SPPF	[640, 640, 5]
10	−1	1	0	Upsample	[None, 2, 'nearest']
11	[−1, 6]	1	0	Concat	[1]
12	−1	3	5,575,360 (Reduced by 24%)	C2f_MCC	[1280, 640, 3]
13	−1	1	0	Upsample	[None, 2, 'nearest']
14	[−1, 4]	1	0	Concat	[1]
15	−1	3	1,948,800	C2f	[960, 320, 3]
16	−1	1	922,240	Conv	[320, 320, 3, 2]
17	[−1, 12]	1	0	Concat	[1]
18	−1	3	5,370,560 (Reduced by 25%)	C2f_MCC	[960, 640, 3]
19	−1	1	3,687,680	Conv	[640, 640, 3, 2]
20	[−1, 9]	1	0	Concat	[1]
21	−1	3	5,575,360 (Reduced by 24%)	C2f_MCC	[1280, 640, 3]
22	[15, 18, 21]	1	8,795,008	Detect	[80, [320, 640, 640]]

3.1.3. Wise-IoU

The default CIoU loss function utilized in YOLOv8 is optimized for the bounding box overlap, centroid distance, and aspect ratio. However, it exhibits limitations in handling both high- and low-quality samples, particularly when the sample distribution is imbalanced. To mitigate this issue, the present paper proposes substituting CIoU with the WIoUv3 loss function [27]. The WIoU (Wise-IoU) family includes WIoUv1, WIoUv2, and WIoUv3 versions, each tailored to optimize the performance across different scenarios.

WIoUv1 effectively differentiates between high-quality and low-quality samples by incorporating distance attention. This approach mitigates the over-penalization of low-quality samples and enhances the generalization ability. The expression for WIoUv1 is presented in Equations (2) and (3).

$$L_{WIoUv1} = R_{WIoU} L_{IoU} \tag{2}$$

$$R_{WIoU} = \exp\left(\frac{(x - x_{gt})^2 + (y - y_{gt})^2}{\left(W_g^2 + H_g^2\right)^*}\right) \tag{3}$$

This will significantly enhance the L_{IoU} of normal-quality anchor frames when $R_{WIoU} \in [1, \ e]$. It is important to note that L_{IoU} lies between [0, 1] when the anchor frame exhibits substantial overlap with the target frame. This situation markedly reduces the L_{IoU} for high-quality anchor frames, leading to concerns regarding the distance between their centroids. Here, W_g and H_g represent the dimensions of the smallest external frame (Figure 3).

WIoUv2, conversely, enhances the model's emphasis on challenging samples by introducing monotonic focusing coefficients. This approach also addresses the issue of slower convergence during the later stages of training, as illustrated in Equation (4).

$$L_{WIoUv2} = L_{IoU}^{\gamma^*} L_{WIoUv1}, \ \ \gamma > 0 \tag{4}$$

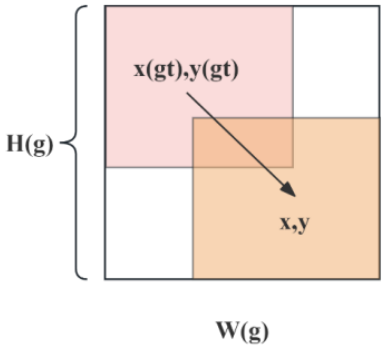


Figure 3. Schematic diagram of WIoU.

Due to the incorporation of the focusing coefficients, the gradient of the WIouv2 backpropagation is also modified, as illustrated in Equation (5).

$$\frac{\partial L_{\text{WioUv2}}}{\partial L_{\text{IoU}}} = L_{\text{IoU}}^{\gamma *} \frac{\partial L_{\text{WioUv1}}}{\partial L_{\text{IoU}}}, \ \gamma > 0 \tag{5}$$

As the reduction in L_{IoU} leads to a corresponding decrease in the gradient gain $r = L_{\text{IoU}}^{\gamma *} \in [0, 1]$, the convergence rate slows down during the later stages of training. To address this issue, we introduce the mean value of L_{IoU} as a normalization factor, as illustrated in Equation (6).

$$L_{WIoUv2} = \left(\frac{L_{\text{IoU}}^{*}}{\overline{L_{\text{IoU}}}} \right)^{\gamma} L_{WIoUv1} \tag{6}$$

where $\overline{L_{IoU}}$ denotes the exponential running average with momentum m. The dynamically updated normalization factor maintains the gradient gain $r = \left(\frac{L_{IoU}^{*}}{\overline{L_{IoU}}} \right)^{\gamma}$ at a generally elevated level, effectively addressing the issue of slow convergence during the later stages of training. The degree of anomaly of the anchor frame is represented by the ratio β of L_{IoU} to $\overline{L_{IoU}}$, as shown in Equation (7).

$$\beta = \frac{L_{\text{IoU}}^{*}}{\overline{L_{\text{IoU}}}} \in [0, +\infty] \tag{7}$$

Lower degrees of abnormality correspond to higher-quality anchor frames; therefore, small gradient gains are assigned to these frames to concentrate the bounding box regression on those of average quality. Concurrently, small gradient gains are also allocated to anchor frames with a high degree of anomaly in order to mitigate the impact of detrimental gradients arising from low-quality samples. The nonmonotonic focusing coefficient was formulated using β and applied to WIoUv1, as illustrated in Equation (8).

$$L_{WIoUv3} = rL_{WIoUv1}, r = \frac{\beta}{\delta \alpha^{\beta - \delta}} \tag{8}$$

where $r = 1$ when $\beta = \delta$. As shown in Figure 4, the highest gradient gain will be obtained when the degree of abnormality of the anchor frame satisfies $\beta = C$ (C is a constant value). Since $\overline{L_{IoU}}$ is dynamic, the quality classification criteria of the anchor frames are also dynamic, which enables WIoUv3 to develop a gradient gain allocation strategy that best fits the current situation during training.

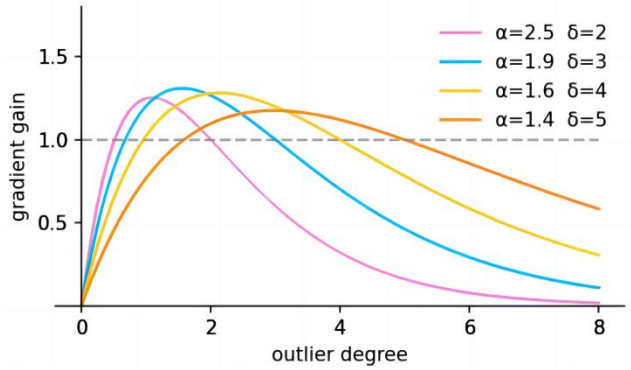


Figure 4. Impact of WIoUv3 hyperparameter selection on gradient gain curve [27].

3.2. Multi-Channel Emotion Recognition Algorithm for the Classroom

In natural classroom environments, the recognition of student emotions encounters several challenges, including occluded faces, inadequate lighting conditions, and variations in posture. Relying solely on facial expression information can significantly impact the accuracy of emotion recognition. To address these issues, this paper proposes a multi-channel fusion-based information processing method aimed at enhancing both the accuracy and robustness of emotion recognition by integrating facial features, the background context, and skeletal data [19,28].

3.2.1. EfficientNetv2 Network

EfficientNetv2 offers several advantages, including a lightweight architecture, rapid processing speed, and high accuracy. These characteristics render it particularly well-suited for efficient image feature extraction tasks. Consequently, this study utilizes EfficientNetv2 to extract image features.

EfficientNet was introduced by Google in 2019, and its core innovation lies in the "composite scaling factor" approach. This method enhances the computational efficiency and accuracy by simultaneously optimizing the network's depth, width, and resolution [29]. While EfficientNetv1 has demonstrated a significant performance on the ImageNet dataset, it suffers from slow inference times and is heavily dependent on large volumes of training data [30].

To address these challenges, the Google team introduced EfficientNetv2 in 2021 (Figure 5). This new architecture is designed to adapt to application scenarios that demand higher speed and efficiency by balancing design elements to enhance the training speed while reducing the number of parameters. EfficientNetv2 presents two significant advancements over its predecessor, EfficientNetv1: First, it employs the Fused-MBConv module instead of the MBConv module for shallow networks. This modification optimizes the balance between the performance and efficiency by eliminating ascending convolutions and enabling shortcut branches under specific conditions. Secondly, EfficientNetv2 introduces a non-uniform scaling strategy for the network. This approach allows differential adjustments based on each stage's contribution to the overall network performance; certain stages are allocated more resources while others are simplified accordingly. Such refinements enhance the overall network efficiency without compromising or even improving its performance.

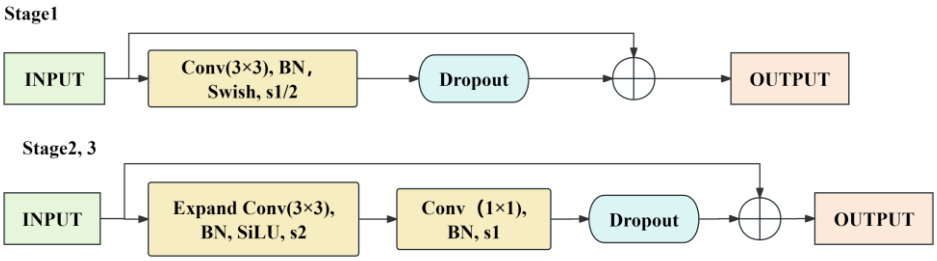


Figure 5. Adjustment module diagram of EfficientNetv2.

3.2.2. Multi-Channel Emotion Recognition Network Architecture

In this study, we aim to develop a Multi-Channel Emotion Recognition Network (MultiEmoNet) utilizing the EfficientNetv2-S model. This model is characterized by its lightweight design and high accuracy, making it suitable for feature extraction [31]. The primary focus of our network will be facial expressions, while background information and skeletal data will serve as supplementary inputs. A detailed description of the MultiEmoNet architecture is provided in the following section.

Face Extraction Module

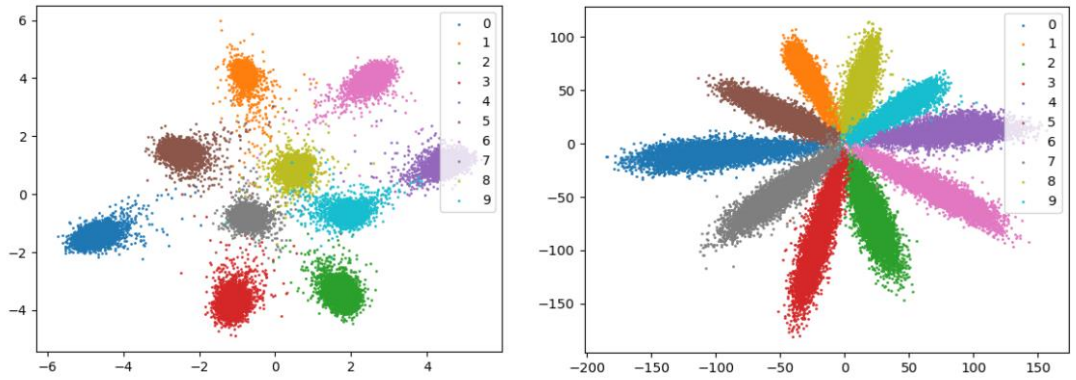
In image classification tasks, the expressions of the same individual exhibit greater similarity than identical expressions across different individuals. This phenomenon complicates the differentiation between various categories of facial expressions. To address this challenge, the present study incorporates a CLF within the feature extraction module to optimize both intra-class and inter-class variations. The implementation of Center Loss enhances the accuracy of facial expression classification by minimizing intra-class variation while indirectly facilitating inter-class separation. The formulation for Center Loss is presented in Equation (9).

$$L_c = \frac{1}{2}\sum_{i=1}^{m}\left\|x_i - c_{yi}\right\|_2^2 \quad (9)$$

where m represents the batch size, x_i denotes the depth feature of sample i, and c_{yi} indicates the feature center of category y_i, to which sample i belongs. The term $\left\|x_i - c_{yi}\right\|_2$ signifies their L2 norm. To enhance the classification performance, center loss is typically employed in conjunction with the cross-entropy loss function, resulting in a joint optimization framework (Equation (10)).

$$L = L_{CE} + \lambda L_c \quad (10)$$

where L_{CE} represents the cross-entropy loss, and λ serves as a balancing parameter to regulate the relative importance of the two losses. Figure 6 illustrates the results of feature visualization on the MNIST dataset, trained using both center loss and cross-entropy loss. The effective application of center loss enhances class separability by optimizing inter-class margins while minimizing intra-class variation.



(**a**) (**b**)

Figure 6. Feature comparison of the MINIST dataset trained with different loss functions. (**a**) Center Loss; (**b**) Cross Entropy Loss.

It has been demonstrated that a pre-trained convolutional neural network utilizing the VGGface2 dataset exhibits a superior performance in face emotion recognition [32]. Consequently, this study employs the pre-trained EfficientNetv2-S model as the backbone network for face feature extraction, while removing its final fully connected layer to facilitate multi-channel information fusion tasks. To further enhance the emotion recognition efficacy, an Auxiliary Attention Module (AAM) is incorporated following the feature ex-

traction network. This module comprises both spatial attention and channel attention components (Figure 7).

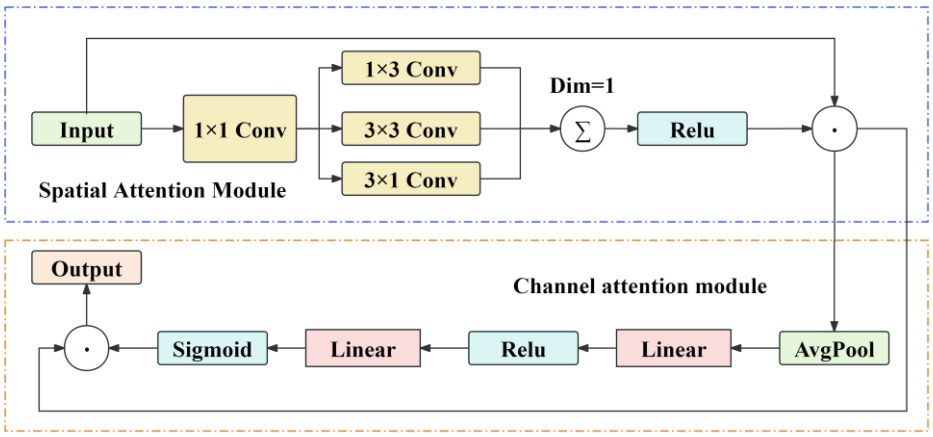


Figure 7. Structure diagram of the AAM module.

The Spatial Attention Module captures spatial features across various orientations through dimensionality-reducing convolutions and a series of specialized convolutions (Conv3 × 3, Conv1 × 3, and Conv3 × 1). The resulting spatial attention maps are then multiplied with the original feature maps to amplify responses at critical spatial locations. The Channel Attention Module compresses spatial dimensions via global average pooling and learns the importance weights for each channel through a sequence of linear transformations. These weights are normalized using a sigmoid function and subsequently multiplied with the original features to enhance the capture of significant channels. By integrating these two modules, AAM effectively identifies essential spatial and channel characteristics. The instantiation of multiple AAM modules enables simultaneous focus on different facial regions, thereby improving the accuracy in emotion recognition. The feature maps generated by one of the AAM modules are utilized to compute the center loss; additionally, areas emphasized by the model are derived from four attention head summation operations. Finally, classification results are produced through a fully connected neural network. The architecture of the face feature extraction network FaceFeatXtractor is illustrated in Figure 8.

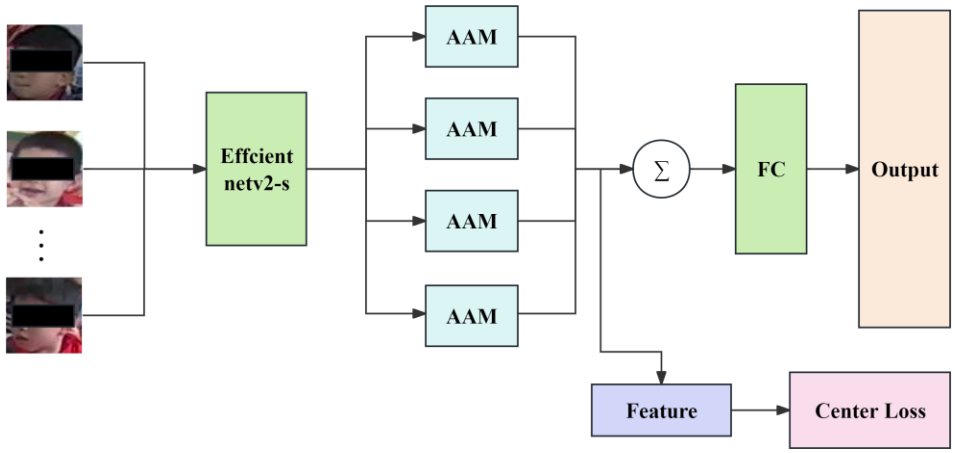


Figure 8. Network architecture of FaceFeatXtractor.

Hybrid Multi-Channel Network

This study introduces the self-developed MultiEmoNet network, which facilitates multi-channel emotion recognition by integrating facial expressions, skeletal information,

and the background context. This approach addresses the limitations associated with single-channel methods in complex classroom environments. While facial expression serves as the primary feature for emotion recognition, it is often vulnerable to occlusion and variations in posture. In contrast, skeletal and background information can provide complementary data when facial cues are obscured, thereby enhancing both the robustness and accuracy of the model.

In this study, we propose a feature fusion approach that not only captures the common features across different modalities, but also accommodates the heterogeneity of information from these modalities, in contrast to traditional data and decision layer fusion methods. To fully leverage multi-type information, we design a network architecture with multiple input channels (Figure 9), where each channel is dedicated to processing specific types of data and performing feature extraction through independent CNNs. Pre-trained weights for single-channel inputs are initialized using EfficientNetv2-S. The facial image channel incorporates AAM, which enhances the representation of key regions associated with facial expressions. The features derived from facial images, skeletal data, and background information are subsequently integrated within a feature fusion unit to create a composite feature vector. This vector undergoes further processing via a fully connected layer (FC) to accomplish the classification task. This architecture aims to improve the classification accuracy by fusing diverse data sources and is particularly well-suited for emotion recognition in complex contexts.

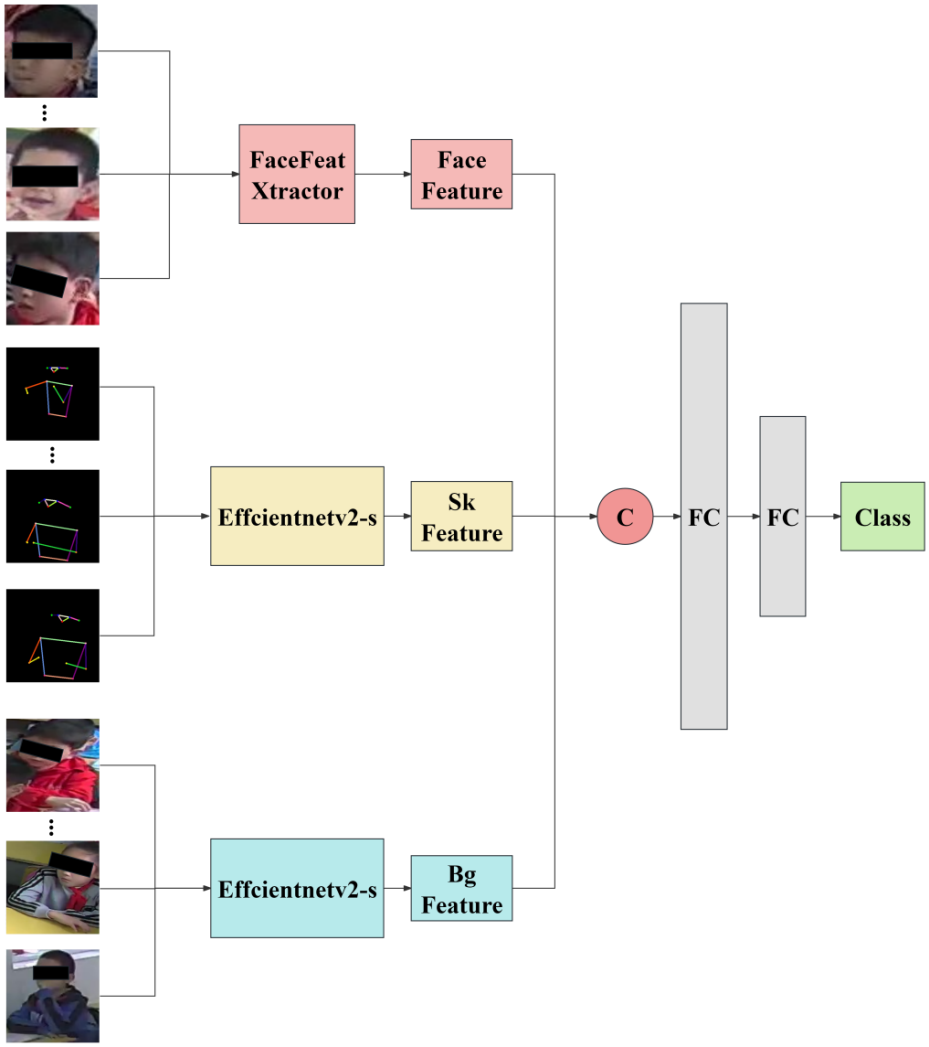


Figure 9. Network architecture of MultiEmoNet.

MultiEmoNet not only enhances the accuracy of emotion recognition, but also improves the model's generalization ability in complex classroom environments. This advancement allows for the more precise processing of multi-source information and facilitates the effective recognition of diverse emotional expressions.

4. Experiments

4.1. Experimental Details

4.1.1. Datasets

Object Detection Dataset

To tackle the challenges posed by classroom environments characterized by high densities of teachers and students, as well as varying camera angles, this study utilized the CrowdHuman dataset for model training [33]. This dataset is recognized for its extensive scale, diversity, and intricate scene annotations, making it particularly suitable for assessing the human detection performance in crowded settings. CrowdHuman comprises approximately 25,000 images that encompass over 40 cities and a variety of activity scenes. Each image contains an average of 22.6 annotated instances—specifically Head BBox, Visible BBox, and Full BBox—thereby providing rich training data for the model (Figure 10).

Figure 10. Annotation examples from the CrowdHuman dataset [33].

Multi-Channel Classroom Emotion Dataset

In order to investigate the issue of emotion recognition within classroom environments, this study collects video data from elementary and middle-school classrooms across various regions and subjects. A total of 200 video files are included, encompassing disciplines such as mathematics, language arts, and others. Each video is recorded at a resolution of 1280 × 720 pixels with a frame rate of 25 frames per second. The duration of each lesson ranges from approximately 40 to 45 min. Informed consent was obtained from all participants and their legal guardians for the classroom experimental data in this study to ensure compliance with ethical standards and the protection of the children's privacy. An example of the collected video data is illustrated in Figure 11.

(**a**) (**b**)

Figure 11. Example of video data collection. (**a**) Front-facing camera video; (**b**) Rear-facing camera video.

In order to accurately label mood changes, this study extracts one frame every 15 s and generates the Student-Full dataset by detecting the students' actual body regions (Visible BBox) using the YOLOv8x-Improved model. A total of 33,481 images were labeled and classified according to the six emotions identified in the previous section (Table 3).

Table 3. Number of annotations in the image emotion dataset.

Emotion Category	Count
Doubt	3857
Engagement	5987
Fatigue	4083
Disconnection	7696
Pleasure	3779
Peace	8078
Total	33,481

In this paper, the YOLOv8-Pose model is used for skeletal feature extraction to generate the Student-Sk dataset. The model uses CSP-darknet as the backbone network and fuses different scale features through PANet, which is able to efficiently detect the skeletal key points. The 17 key points contained in the COCO dataset are used to estimate the human posture (Figure 12). The key point information is normalized and centered to finally generate a skeletal map (Figure 13) for subsequent emotion recognition.

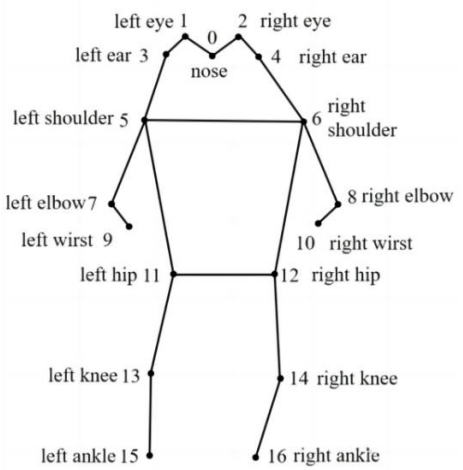


Figure 12. Example of human key points annotation.

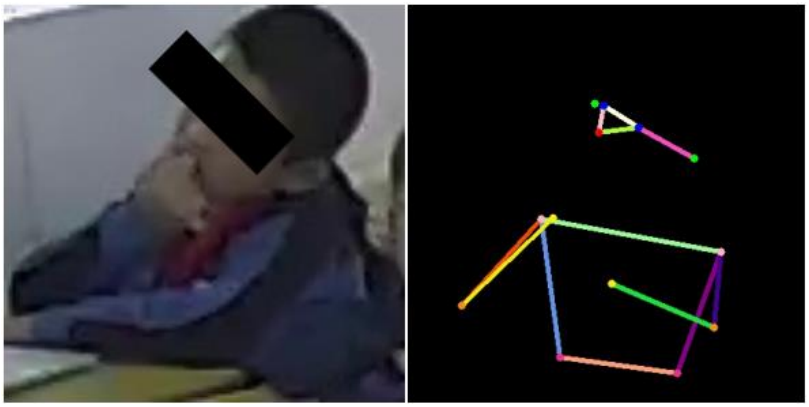


Figure 13. Example of generated skeleton diagram.

In a natural classroom environment, traditional facial detection algorithms often struggle to handle complex scenarios such as occlusion and head-turning. Therefore, in this

paper, we opted for the YOLOv8x-Improved head detection model, which is trained on the CrowdHuman dataset, rather than relying on facial detection methods. The CrowdHuman dataset categorizes human body detection into three distinct categories: head Box, Full BBox, and Visible BBox. The specific detection effect applied to the classroom scene is shown in Figure 14.

Figure 14. Actual detection results using YOLOv8.

In order to enhance the detection accuracy, this study employs a stacking preprocessing method. This involves superimposing the original image twice along both the horizontal and vertical coordinate axes to increase the resolution prior to inputting it into the model. The preprocessed human head image is then fed into the detector for analysis. Subsequently, the coordinates of the detection frame are extracted, yielding the corresponding head image. Finally, a facial information dataset named Student-Head is generated by traversing through the classification folder.

4.1.2. Experimental Setup

The specific hardware and software utilized in the experimental and evaluation processes are detailed in Table 4. The research presented in this paper was conducted on this platform.

Table 4. Experimental configuration.

Experimental Environment	Configuration
Operating system	Windows 10
CPU model	Intel i7-12700KF
Memory	32 GB
GPU model	NVIDIA RTX3080 10 G
Computing platform	CUDA 11.6
Programming language	Python 3.8
Development IDE	PyCharm 2022.2.2
Deep learning framework	Pytorch 11.6

4.1.3. Assessment of Indicators

The object detection performance is primarily assessed based on both the detection accuracy and speed. In this paper, we conduct a comprehensive evaluation of the model's accuracy, efficiency, and robustness using metrics such as Precision, Recall, mean Average Precision (mAP), and Frames Per Second (FPS).

Precision and recall are fundamental metrics for assessing the effectiveness of a model's detection capabilities. They measure, respectively, the accuracy of positive case predictions and the recognition rate of positive cases. These metrics are calculated as outlined in Equations (11) and (12).

$$precision = \frac{TP}{TP + FP} \tag{11}$$

$$Recall = \frac{TP}{TP + FN} \tag{12}$$

Average precision (AP) serves as a metric for evaluating the model performance by computing the area under the Precision–Recall (P-R) curve, as illustrated in Equation (13).

$$AP = \frac{1}{m}\sum_{i=1}^{m}\int_{0}^{1} P_i(r)dr \tag{13}$$

The mAP_0.5 metric is derived from the average AP value when the Intersection over Union (IoU) threshold is set at 0.5, as illustrated in Equation (14).

$$mAP_{\_0.5} = \frac{1}{n}\sum_{i=1}^{n} AP(i) \tag{14}$$

In addition to the detection accuracy, the detection speed serves as a crucial indicator of the performance. The FPS reflect the processing speed of the model; a higher FPS signifies an increased processing speed, making it more suitable for real-time detection tasks.

4.2. Experimental Design

4.2.1. Object Detection Experiment

This experiment utilizes the CrowdHuman dataset for training in object detection. The dataset comprises 15,000 training images and 4370 validation images, all with an image resolution of 640 × 640 pixels. To enhance the model's generalization capability, the training data undergo a series of augmentation processes, including Mosaics, Blur, and Gray transformations.

By incorporating grayscale images into the training set, the model demonstrates enhanced stability in the performance when confronted with missing or inconsistent color information. This capability holds substantial practical significance for deploying the model across diverse environments, particularly in situations characterized by highly variable lighting conditions or unreliable color data. The training hyperparameters utilized in this experiment are detailed in Table 5.

Table 5. Training hyperparameters.

Parameter Name	Value
Image size	640 × 640
Load pretrained weights	True
Training epochs	100
Optimizer	SGD
Initial learning rate	0.01
Final learning rate	0.001
SGD momentum	0.937
Momentum decay	0.0005
Warmup epochs	3.0

4.2.2. Classroom Multi-Channel Emotion Recognition Experiment

For the training of the classroom multi-channel emotion recognition experiment, this study utilized the Student-Full dataset, Student-Head dataset, and Student-Sk dataset as inputs for the background information, facial information, and skeletal information,

respectively. A corresponding file from each dataset with the same name was selected to form a set of input training data. To address missing data issues, an all-black image was employed in place of any missing information images within a channel. In terms of data preprocessing, normalization and random flipping operations were primarily implemented. The division between the training and validation datasets follows an 80:20 ratio. The Adam optimizer was utilized with a learning rate set at 0.001. Based on the results of the experimental iterations, it is observed that the model reaches a more stable state of convergence at 200 rounds. Therefore, a total of 200 epochs is executed in loops, during which the learning rate is halved every three epochs.

4.3. Experimental Results

4.3.1. Performance Evaluation of the CrowdHuman Dataset

This experiment utilizes the CrowdHuman dataset to perform object detection experiments on the YOLOv8x-Improved model. The primary focus is to compare the detection performance of various models, including SSD, Faster-RCNN (ResNet + FPN), YOLOv7x, YOLOv8x, and YOLOv8x-Improved. Table 6 presents a comparative analysis of these models' performances on the CrowdHuman dataset.

Table 6. Comparison of detection results on the CrowdHuman dataset.

Model	Precision	Recall	mAP50	mAP50-95	Model Size	GFLOPS	FPS
SSD	0.661	0.614	0.638	0.482	105.1 M	15.0	43
Faster-RCNN ResNet + FPN	0.684	0.621	0.662	0.524	57.22 M	134.5	18
YOLOv7x	0.842	0.721	0.845	0.581	135 M	188	39
YOLOv8x	0.847	0.769	0.851	0.603	130 M	258.1	40
YOLOv8x-Improved	0.853	0.772	0.862	0.611	109 M	232.2	48

YOLOv8x-Improved significantly enhances key performance metrics. The Precision reaches 0.853, while the Recall stands at 0.772, demonstrating its exceptional capability in recognizing positive samples. In terms of the average precision, YOLOv8x-Improved achieves a score of 0.862 on mAP50 and 0.611 on mAP50-95, indicating its robustness across various IOU thresholds. Furthermore, the model size is reduced to 109 MB, with GFLOPS decreasing to 232.2, resulting in approximately an 18% reduction in weight due to the incorporation of the C2f module MCC. Regarding the processing speed, YOLOv8x-Improved excels with a performance rate of 48 frames per second, surpassing both the original YOLOv8x and other models. This demonstrates significant improvements in speed while maintaining high precision and recall rates.

Combining all performance metrics, the YOLOv8x-Improved model exhibits a more pronounced performance compared to the original YOLOv8x (Figure 15). The enhanced model consistently demonstrates improvements in both precision and recall, ultimately achieving higher steady-state values and enhancing the detection of true-positive samples. In the evaluation of the mean average precision (mAP_0.5), the improved model surpasses its predecessor during the later stages of training while maintaining an exceptional performance on the more rigorous mAP_0.5:0.95 metrics. This indicates substantial advancements in its object detection capabilities.

In terms of training and validation loss, YOLOv8x-Improved employs the WIOU loss function in place of the original CIOU loss (Figure 16). The experimental results indicate that YOLOv8x-Improved exhibits a lower bounding box loss compared to YOLOv8x during both the training and validation phases, thereby significantly enhancing the target localization accuracy. Both models demonstrate a comparable performance regarding category loss, reflecting robust generalization capabilities. The effectiveness of the WIOU loss function further contributes to the improved performance of YOLOv8x-Improved.

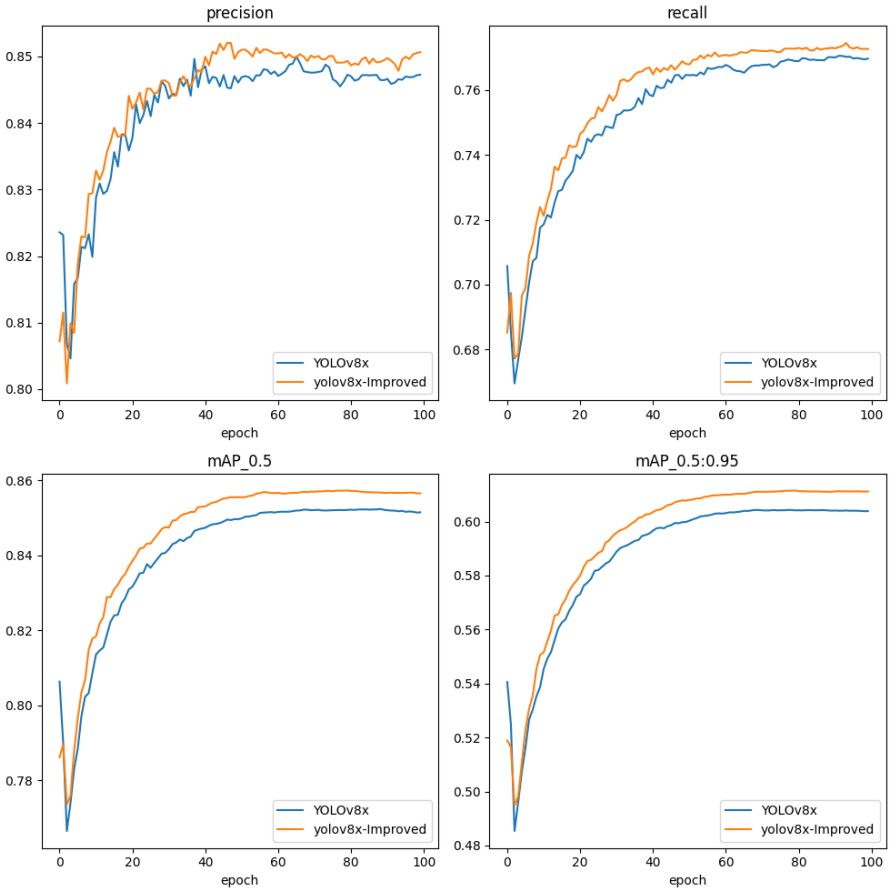


Figure 15. Comparison of training results between YOLOv8x-Improved and YOLOv8x.

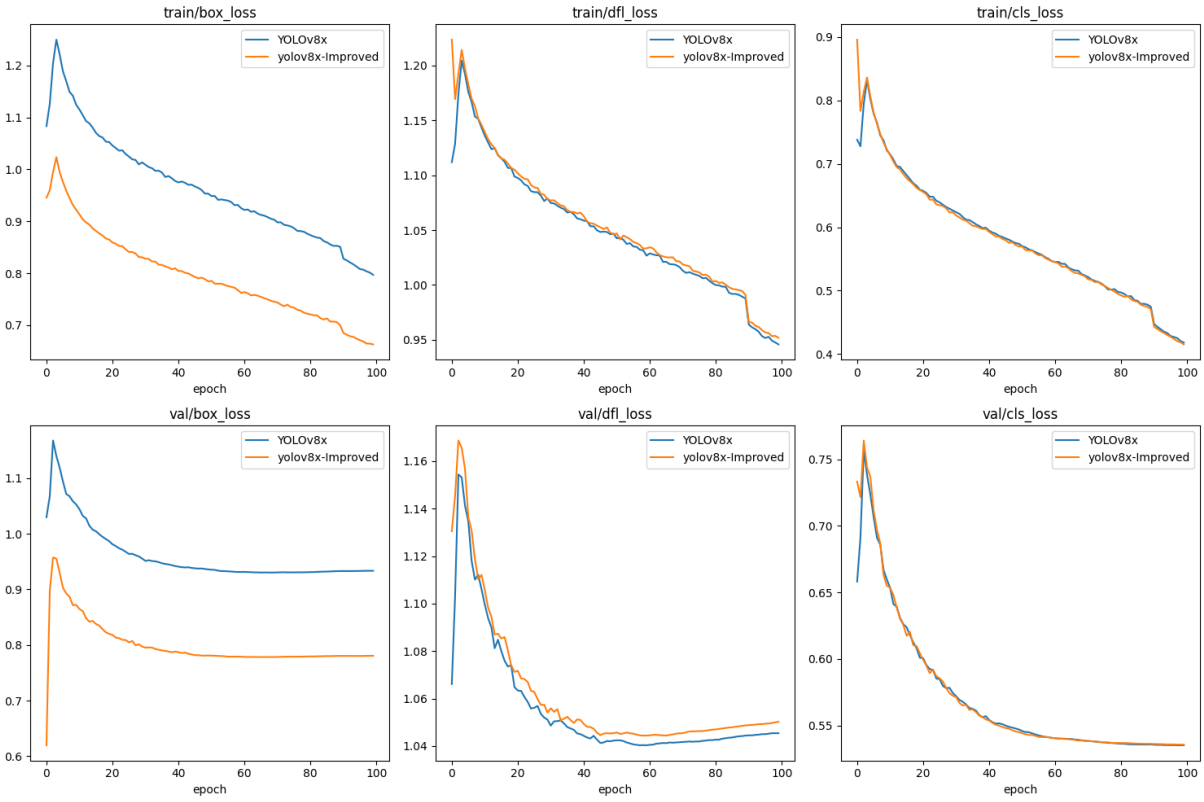


Figure 16. Training and validation loss curves.

4.3.2. Performance Evaluation of a Classroom Student Object Detection Dataset

To further evaluate the inference performance of the emotion recognition algorithm under different hardware configurations to meet the needs of classroom applications for real-time feedback [34], this study tested the inference time and computational efficiency under different GPU configurations (Table 7).

Table 7. Performance comparison of different GPU configurations.

Hardware Configuration	CPU	GPU	Inference Time per Image (ms)	FPS
1	Intel i7-12700KF	NVIDIA RTX 2080	24.39	41
2	Intel i7-12700KF	NVIDIA RTX 2080 Super	22.22	45
3	Intel i7-12700KF	NVIDIA RTX 3080	20.83	48

This study uses a self-made dataset of student classroom object detection to carry out more validation experiments that test how well the YOLOv8x-Improved model works for detection in a real classroom setting. One frame was randomly selected from each of the 100 classroom videos to obtain a total of 100 validation images. Each image contained 30 to 60 students and was accurately labeled using the LabelImg tool (v.1.8.6), as shown in Figure 17.

Figure 17. Annotation example diagram of self-made student classroom object detection dataset.

With the help of the trained model, a validation script was run to quantitatively analyze the detection ability of the model in a real classroom environment. Table 8 summarizes the obtained performance metrics.

Table 8. Comparison of custom classroom object detection dataset results.

Model	Precision	Recall	mAP50	mAP50-95
YOLOv8x-COCO	0.744	0.698	0.77	0.479
YOLOv8x	0.781	0.831	0.851	0.566
YOLOv8x-Improved	0.802	0.852	0.863	0.572

We used the YOLOv8x-COCO model in this study, which we downloaded from the official YOLOv8 website and pre-trained on the COCO dataset. The experiment only evaluates its detection effect on the 'Person' category and compares it with the performance of YOLOv8x and YOLOv8x-Improved in the Visible BBox. The results show that the YOLOv8x-COCO model performs well on general-purpose datasets, but lacks the ability to generalize to specific classroom scenarios. YOLOv8x and YOLOv8x-Improved, specifically trained for crowd detection, demonstrate superior precision and recall. YOLOv8x-Improved, in particular, excels in precision, recall, and mAP50 and mAP50-95 metrics, indicating the model's strong generalization across various scenarios and its exceptional detection performance in natural classroom scenarios following training in the CrowdHuman dataset.

In addition, this paper tests the model through different classroom scenarios (Classroom A and Classroom B). Figure 18 shows the detection results of YOLOv8x-COCO and YOLOv8x-Improved in different classroom scenarios, respectively. YOLOv8x-Improved excels in identifying students at the rear of the classroom and in areas on both sides, while also minimizing missed detection, demonstrating its adaptability in complex classroom environments.

(**a**) (**b**)

Figure 18. Comparison of object detection. (**a**) Classroom A; (**b**) Classroom B.

4.3.3. Results of a Classroom Multichannel Emotion Recognition Experiment

This study conducts experiments on the classroom emotion dataset to verify the effectiveness of the MultiEmoNet network structure. We refer to the background, face, and skeletal information as Bg, Face, and Sk, respectively.

Table 9 shows the experimental results of this study using single-face information. In this study, the FaceFeatXtractor network outperforms other algorithms such as Resnet50, Vgg16, MobileNetv3, and EfficientNetv2-S, achieving an accuracy of 0.721. This paper's proposed face feature extraction network effectively enhances the accuracy of emotion recognition.

Table 9. Comparison of experiments using single-person face information.

Model	Precision
Resnet50	0.689
Vgg16	0.678
MobileNetv3	0.669
EfficientNetv2-S	0.702
FaceFeatXtractor	0.721

Table 10 shows the results of experiments using single-skeletal information in this study. The experiments compare the accuracy of emotion classification using traditional machine learning methods, such as SVM, KNN, and Random Forest, with the use of preprocessed–generated skeletal information maps fed into a convolutional neural network (EfficientNetv2-S). The results demonstrate that the method presented in this paper performs better on skeletal information when compared to traditional machine learning methods.

Table 10. Comparison of experiments using single bone information.

Model	Precision
SVM	0.516
KNN	0.503
Random Forest	0.552
EfficientNetv2-S	0.601

Table 11 shows the experimental results of this study using single background information.

Table 11. Comparison of experiments using single background information.

Model	Precision
Resnet50	0.794
Vgg16	0.774
MobileNetv3	0.771
EfficientNetv2-S	0.813

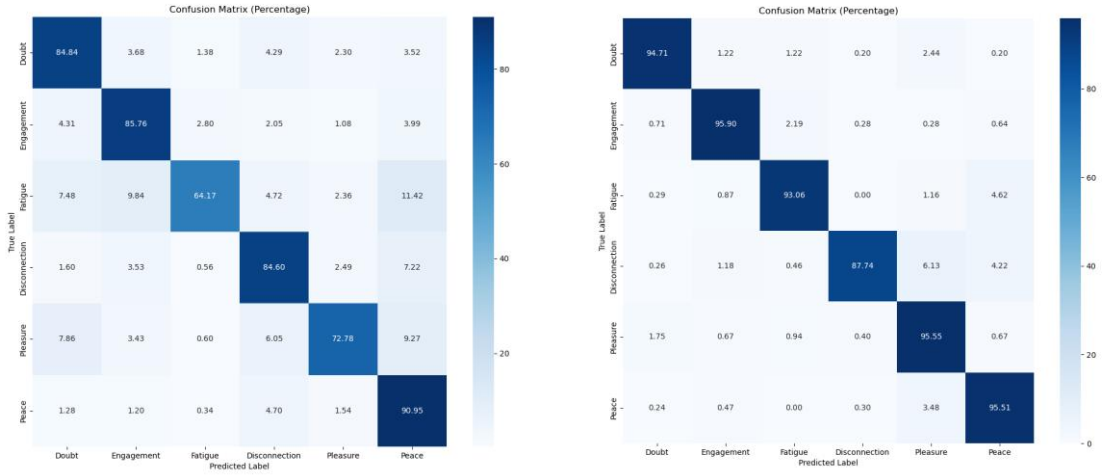
Table 12 MultiEmoNet provides the experimental results for the use of multi-channel information.

Table 12. Comparison of experiments using multi-channel information.

Model	Precision
Face + Sk	0.841
Face + Sk + Bg	0.914

The experimental results show that the MultiEmoNet model combining multi-channel information achieves the highest emotion classification accuracy. The model that uses skeleton information (Sk) and face information (Face) as the inputs performs better than the single-channel information model. This shows that the multi-channel emotion recognition method suggested in this paper works.

In particular, fusing background information (Bg) significantly improves the model's performance. The reason may be that face information is missing in some images, and the background can make up for it; secondly, the resolution of the face images is too low, and the background information contains more emotional features compared to the background information. In this study, we produce confusion matrices using face and skeletal features (Face + Sk) as well as face, skeletal, and background features (Face + Sk + Bg) to verify this (Figure 19).



(**a**) (**b**)

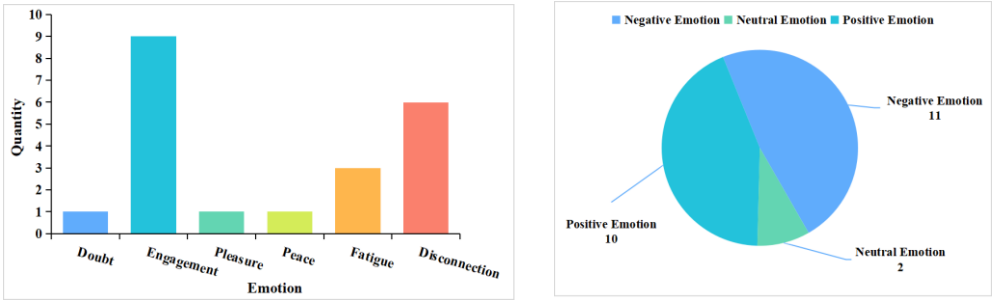
Figure 19. Confusion matrix of MultiEmoNet. (**a**) Face + Sk; (**b**) Face + Sk+ Bg.

The analysis of the confusion matrix reveals that the addition of background information significantly improves the prediction accuracy of the model in all emotion categories, particularly in the 'Fatigue' and 'Pleasure' emotions, increasing the recognition accuracy from 64.17% and 72.78% to 93.06% and 95.55%, respectively. Meanwhile, the addition of contextual information significantly reduces the misclassification rate among different emotion categories, which further proves the key role of contextual information in emotion recognition. The background features provide important cues to the context of emotions, significantly enhance the differentiation ability of the model, and are important for improving the accuracy of emotion recognition systems.

4.4. Subsection Visualization of Student Emotion Recognition in the Classroom

The statistics and visualization of emotion recognition results are important for a deeper understanding of students' emotional states in the classroom. We used an emotion recognition algorithm in this study to generate detailed emotion statistics (Figure 20), which show the specific percentage of different emotion categories and categorize them into positive (Pleasure, Engagement), negative (Fatigue, Disconnection), and neutral (Peace, Doubt) emotions. These data provide teachers with visual feedback on classroom emotions, which helps to assess the classroom climate and adjust teaching strategies in a timely manner. Meanwhile, the mood prediction chart (Figure 21) further visualizes the mood state of each student, allowing teachers to grasp the individual mood distribution in the classroom and identify students who may need extra attention. This visual mood analysis not only provides a global view of the overall climate, but also makes it easier for teachers to further understand the learning progress of different students based on their emotional states after class, thereby optimizing personalized teaching strategies. It is important to note that this technology is intended to support educational and developmental uses only and should not be used for punitive purposes under any circumstances.

On the basis of this, the Emotional Statistics Timing Chart further reveals the trajectory of students' emotional fluctuations in the classroom. The 30 s recording interval can achieve a high recognition accuracy while capturing emotional fluctuations. Therefore, by recording the emotional distribution every 30 s, the chronogram can clearly show the emotional changes of students at different stages, helping teachers to understand students' emotional trends more comprehensively after class so as to make teaching adjustments and support at the appropriate time (Figure 22).



(**a**) (**b**)

Figure 20. (**a**) Detailed emotion statistics; (**b**) Emotion classification statistics.

Figure 21. Emotion prediction results.

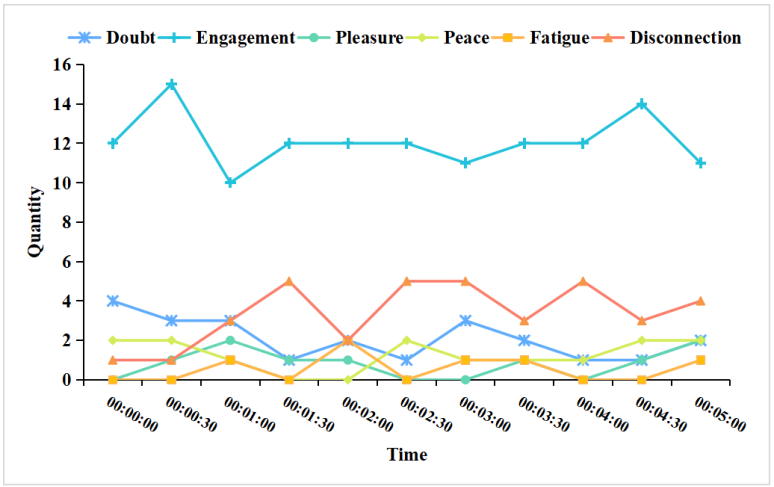


Figure 22. Time series of classroom emotion statistics (five minutes are captured).

Finally, to facilitate the quantitative analysis of classroom teaching, the system will provide an emotion statistics table that systematically summarizes the emotion data in each time period. The table will show the emotion distribution of each time frame in detail, helping teachers track the dynamic changes of students' emotions and provide data support for the timely adjustment of teaching strategies. The table data support downloading, which is convenient for teachers to further analyze students' emotional trends after class and optimize them in combination with specific teaching goals.

5. Conclusions and Future Work

This study provides an in-depth exploration of student emotion recognition and analysis in primary and secondary school classrooms, with the main goal of achieving accurate object detection and emotion recognition in natural classroom environments using deep learning techniques. The main contributions of the study are improving the YOLOv8 model so that it can find emotions more quickly and accurately in complicated classroom settings; creating a multi-channel classroom emotion dataset with information about facial expressions, skeletal structures, and background; and suggesting the MultiEmoNet emotion recognition model. The fusion of multiple information sources enhances the accuracy and robustness of emotion recognition. In addition, this study provides a visual view of classroom student emotion analysis, which provides teachers with data-based teaching feedback and optimization suggestions, enriching the application scenarios of educational technology.

Despite the results, this study still has some limitations. Firstly, although the MultiEmoNet model improves the accuracy of emotion recognition, it also brings higher computational complexity and resource consumption, which may affect its efficiency in real-time applications. Second, the study primarily concentrates on primary and secondary school classroom scenarios, without the extensive verification of its applicability in other educational scenarios and different age groups.

Future research will require further psychological validation, especially with neurodiverse populations, to ensure the robustness and inclusivity of the system across diverse user groups. Future research will focus on model optimization and lightweighting to further enhance the real-time processing capability of the system, especially in resource-limited environments. In addition, the research also plans to add an analysis of misclassification and challenging scenarios to provide directions for the further optimization and performance improvement of the model, thereby enhancing the robustness and adaptability of the system. Future research will also explore the application of emotion recognition technology in more educational scenarios and age groups, aiming to promote the realization of personalized and highly interactive educational environments.

Author Contributions: Conceptualization, Y.H., W.D. and T.X.; methodology, Y.H., W.D. and T.X.; software, Y.H., W.D. and T.X.; validation, W.D. and T.X.; visualization, Y.H.; writing—original draft, Y.H. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by The Humanities and Social Science Research Planning Fund of the Ministry of Education (No. 20YJA880009), and the 2024 "AI + Examination and Evaluation" Special Teaching Reform Project of Central China Normal University (CCNU24GJ17), the Fundamental Research Funds for the Central Universities under Grant CCNU22JC011.

Institutional Review Board Statement: This study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Board of Central China Normal University, Ethic Committee, EC (protocol code CCNU-IRB-202305004a and 24 May 2023).

Informed Consent Statement: Informed consent was obtained from all subjects involved in this study.

Data Availability Statement: The data presented in this study are openly available in the CrowdHuman dataset at https://www.crowdhuman.org/download.html (accessed on 22 May 2024). Relevant data from this study are available from the corresponding author upon reasonable request.

Acknowledgments: The authors would like to thank the Hubei Key Laboratory of Digital Education, Central China Normal University, for providing the research facilities.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Pekrun, R. The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice. *Educ. Psychol. Rev.* **2006**, *18*, 315–341. [CrossRef]
2. Mollahosseini, A.; Hasani, B.; Mahoor, M.H. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Trans. Affect. Comput.* **2017**, *10*, 18–31. [CrossRef]

3. Chernykh, V.; Prikhodko, P. Emotion recognition from speech with recurrent neural networks. *arXiv* **2017**, arXiv:1701.08071.
4. Huang, Y.; Xiao, J.; Tian, K.; Wu, A.; Zhang, G. Research on robustness of emotion recognition under environmental noise conditions. *IEEE Access* **2019**, *7*, 142009–142021. [CrossRef]
5. Liu, Q.; Huang, Z.; Li, Z.; Pointer, M.R.; Zhang, G.; Liu, Z.; Gong, H.; Hou, Z. A field study of the impact of indoor lighting on visual perception and cognitive performance in classroom. *Appl. Sci.* **2020**, *10*, 7436. [CrossRef]
6. Noyes, E.; Davis, J.P.; Petrov, N.; Gray, K.L.; Ritchie, K.L. The effect of face masks and sunglasses on identity and expression recognition with super-recognizers and typical observers. *R. Soc. Open Sci.* **2021**, *8*, 201169. [CrossRef]
7. Chen, H.; Guan, J. Teacher–student behavior recognition in classroom teaching based on improved YOLO-v4 and Internet of Things technology. *Electronics* **2022**, *11*, 3998. [CrossRef]
8. Tang, L.; Gao, C.; Chen, X.; Zhao, Y. Pose detection in complex classroom environment based on improved Faster R-CNN. *IET Image Process.* **2019**, *13*, 451–457. [CrossRef]
9. Ekman, P. An argument for basic emotions. *Cogn. Emot.* **1992**, *6*, 169–200. [CrossRef]
10. Plutchik, R. *Emotions and Life: Perspectives from Psychology, Biology, and Evolution*; American Psychological Association: Washington, DC, USA, 2003.
11. Pekrun, R.; Stephens, E.J. Achievement emotions: A control-value approach. *Soc. Personal. Psychol. Compass* **2010**, *4*, 238–255. [CrossRef]
12. Kosti, R.; Alvarez, J.M.; Recasens, A.; Lapedriza, A. Context based emotion recognition using emotic dataset. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *42*, 2755–2766. [CrossRef]
13. Kumar, A.; Garg, N.; Kaur, G. An emotion recognition based on physiological signals. *Int. J. Innov. Technol. Explor. Eng.* **2019**, *8*, 335–341.
14. Mehrabian, A.; Ferris, S.R. Inference of attitudes from nonverbal communication in two channels. *J. Consult. Psychol.* **1967**, *31*, 248. [CrossRef] [PubMed]
15. Sarvakar, K.; Senkamalavalli, R.; Raghavendra, S.; Kumar, J.S.; Manjunath, R.; Jaiswal, S. Facial emotion recognition using convolutional neural networks. *Mater. Today Proc.* **2023**, *80*, 3560–3564. [CrossRef]
16. Coulson, M. Attributing emotion to static body postures: Recognition accuracy, confusions, and viewpoint dependence. *J. Nonverbal Behav.* **2004**, *28*, 117–139. [CrossRef]
17. Gavrilescu, M. Recognizing emotions from videos by studying facial expressions, body postures and hand gestures. In Proceedings of the 2015 23rd Telecommunications Forum Telfor (TELFOR), Belgrade, Serbia, 24–26 November 2015.
18. Chóliz, M.; Fernández-Abascal, E.G. Recognition of emotional facial expressions: The role of facial and contextual information in the accuracy of recognition. *Psychol. Rep.* **2012**, *110*, 338–350. [CrossRef]
19. Kosti, R.; Alvarez, J.M.; Recasens, A.; Lapedriza, A. Emotion recognition in context. In Proceedings of the IEEE conference on computer vision and pattern recognition, Honolulu, HI, USA, 21–26 July 2017.
20. Lee, J.; Kim, S.; Kim, S.; Park, J.; Sohn, K. Context-aware emotion recognition networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019.
21. Yang, D.; Chen, Z.; Wang, Y.; Wang, S.; Li, M.; Liu, S.; Zhao, X.; Huang, S.; Dong, Z.; Zhai, P. Context de-confounded emotion recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023.
22. Sharma, A.; Mansotra, V. Deep learning-based student emotion recognition from facial expressions in classrooms. *Int. J. Eng. Adv. Technol.* **2019**, *8*, 4691–4699. [CrossRef]
23. Chen, Z.; Liang, M.; Xue, Z.; Yu, W. STRAN: Student expression recognition based on spatio-temporal residual attention network in classroom teaching videos. *Appl. Intell.* **2023**, *53*, 25310–25329. [CrossRef]
24. Noroozi, F.; Corneanu, C.A.; Kamińska, D.; Sapiński, T.; Escalera, S.; Anbarjafari, G. Survey on emotional body gesture recognition. *IEEE Trans. Affect. Comput.* **2018**, *12*, 505–523. [CrossRef]
25. Su, C.; Wang, G. Design and application of learner emotion recognition for classroom. In *Journal of Physics: Conference Series*; IOP Publishing: Bristol, UK, 2020.
26. Terven, J.; Córdova-Esparza, D.-M.; Romero-González, J.-A. A comprehensive review of yolo architectures in computer vision: From yolov1 to yolov8 and yolo-nas. *Mach. Learn. Knowl. Extr.* **2023**, *5*, 1680–1716. [CrossRef]
27. Tong, Z.; Chen, Y.; Xu, Z.; Yu, R. Wise-IoU: Bounding box regression loss with dynamic focusing mechanism. *arXiv* **2023**, arXiv:2301.10051.
28. Mittal, T.; Guhan, P.; Bhattacharya, U.; Chandra, R.; Bera, A.; Manocha, D. Emoticon: Context-aware multimodal emotion recognition using frege's principle. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020.
29. Tan, M.; Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In Proceedings of the 36th International Conference on Machine Learning, Long Beach, CA, USA, 10–15 June 2019.
30. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 84–90. [CrossRef]
31. Wen, Y.; Zhang, K.; Li, Z.; Qiao, Y. A discriminative feature learning approach for deep face recognition. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, the Netherlands, October 11–14, 2016, Proceedings, Part VII 14*; Springer: Cham, Switzerland, 2016.

32. Savchenko, A.V.; Savchenko, L.V.; Makarov, I. Classifying emotions and engagement in online learning based on a single facial expression recognition neural network. *IEEE Trans. Affect. Comput.* **2022**, *13*, 2132–2143. [CrossRef]
33. Shao, S.; Zhao, Z.; Li, B.; Xiao, T.; Yu, G.; Zhang, X.; Sun, J. Crowdhuman: A benchmark for detecting human in a crowd. *arXiv* **2018**, arXiv:1805.00123.
34. Umirzakova, S.; Whangbo, T.K. Detailed feature extraction network-based fine-grained face segmentation. *Knowl.-Based Syst.* **2022**, *250*, 109036. [CrossRef]