Multi-Modal Intelligent Student Monitoring Model Based on Attention and Lightweight Modules

Xiaowei Feng*[a]
Shenzhen Middle School, Shenzhen, China 518001
* Corresponding author: Xiaoweifeng2025@163.com

ABSTRACT

Campus gender-based violence often occurs in hidden spaces, where supervision struggles with privacy and behavior recognition, and victims often remain silent. This paper proposes a multi-modal intelligent model with two algorithms. POSTER V2-CGNet enhances expression recognition by adopting Coordinate Attention to capture spatial patterns and a gate-based module to filter cross-modal noise. YOLOv8s-Pose-SRFA improves pose estimation via a Slim-Neck network that reduces complexity with skip connections and an RFA head that boosts keypoint accuracy in occluded scenes. Experimental verification based on RAF-DB and AffectNet datasets shows that POSTER V2-CGNet achieves an accuracy of 91.35% on the RAF-DB dataset, 1.24% higher than the base model POSTER V2, and 66.25% on the complex AffectNet-7 scenario, balancing accuracy and efficiency; the mAP50 index of YOLOv8s-Pose-SRFA reaches 86.4%, surpassing mainstream models such as HRNet, with significantly reduced parameters and computational complexity, and requires approximately 84.8% fewer FLOPs than OpenPose.

Keywords: POSTER V2-CGNet; YOLOv8s-Pose; Gate-based Spatial-Channel; RFAConv

1. INTRODUCTION

In recent years, incidents of campus gender-based violence and sexual assault have frequently been exposed, attracting widespread social attention. However, most sexual assaults occur in places such as hallways and offices. These places lack effective supervision and often become "gray areas" where teachers abuse power and students find it difficult to seek help. At the same time, victims of campus sexual assault often face long-term psychological distress, such as Post-Traumatic Stress Disorder (PTSD), depression, and self-harm behaviors. Due to being under the suppression of power, victimized students often fall into a silence mechanism or develop Stockholm syndrome, making it difficult for them to take the initiative to seek help and even coexist with perpetrators for a long time. This psychological "secondary harm" is often more concealed and persistent.

Existing supervision mechanisms have a contradiction between protecting privacy and recognizing behaviors. On the one hand, surveillance coverage is incomplete; on the other hand, installing too many surveillance devices is likely to infringe on students' privacy rights. Therefore, there is an urgent need for a new type of system that balances protectiveness, non-intrusiveness, and intelligent recognition to accurately analyze and promptly warn of risky interactive behaviors in key spaces. Artificial intelligence technologies, especially the emerging Large Language Models (LLMs), multi-modal recognition systems, and emotion recognition algorithms in recent years, have provided new solutions for the prevention and psychological intervention of campus gender-based violence. Through AI technology to assist in monitoring risky behaviors such as speech, body language, and emotions in counseling scenarios, it can provide students with a silent safety protection mechanism.

This project focuses on building an AI-based intelligent recognition and intervention system for campus seminar rooms, aiming to effectively identify inappropriate teacher behaviors and protect students' psychological safety from two dimensions: in-event early warning and post-event psychological intervention, thereby delving into the structural roots of

campus sexual assault from a technical perspective. This paper constructs a real-time multi-modal perception system, including: body recognition: using deep learning to track human poses and detect inappropriate behaviors such as physical proximity and touching; expression recognition: analyzing changes in students' facial expressions (emotions such as fear, tension, and disgust).

2. RELATED WORK

2.1 Sexual Harassment Detection Based on Conversation Text

Early work on textual harassment detection primarily employs classical machine learning with feature engineering. Edupuganti [1] adopted Random Forest (RF) for chat classification, citing its ability to handle high-dimensional inputs, evaluate variable importance, tolerate construction errors, mitigate class imbalance, and offer lower computational cost than deep learning. Empirical comparison with Logistic Regression confirmed RF's superiority. Potha and Maragoudakis [2] took a temporal perspective, using time-series models with Singular Value Decomposition (SVD) for dimensionality reduction and denoising. The sorted eigenvalues indicate semantic strength—multiple "peak" words in a conversation signal higher harassment likelihood.

Rizwan et al. [3] designed a real-time system combining NLP and Bayesian classification. User inputs are monitored live, stored as CSV files, and checked per user to accommodate individual definitions of harassment; detected cases trigger pop-up alerts for user action. However, single-level detection suffers from accuracy loss due to subjective interpretation.

On social media, particularly Twitter, scalability and filtering become critical. Saeidi et al. [4] benchmarked ten algorithms (Logistic Regression, Gaussian Naive Bayes, Decision Tree, RF, Linear/Gaussian/Polynomial SVM, MLP, AdaBoost) on the SIMAH dataset, using TF-IDF and Word2Vec as features. Results showed that Linear and Gaussian SVM with TF-IDF yielded the most accurate predictions, while Decision Tree and RF also performed well. Extending this, Saeidi et al. [5] applied Graph Convolutional Networks (GCNs) to the same task and demonstrated that GCNs maintain strong performance even with reduced dataset sizes, outperforming classical methods in data-scarce scenarios.

Similarly, Alawneh et al. [6] built a sentiment-based detection system using TF-IDF and machine learning, enhancing cross-population generalisability. James and Osubor [7] took a rule-based approach, proposing a dictionary of aggressive/harassment-related terms and defining logic rules for pattern matching, offering a straightforward yet interpretable alternative.

Most textual approaches rely on bag-of-words or embedding representations with supervised classifiers; ensemble methods (RF) and kernel methods (SVM) are prevalent; recent work leans into graph neural networks for improved sample efficiency. A persistent challenge is the subjective nature of harassment, which drives interest in user-adaptive or multi-level systems.

2.2 Sexual Harassment Detection Based on Computer Vision

Visual harassment detection faces the dual hurdles of limited labelled datasets and the complexity of spatiotemporal action recognition. Islam et al. [8] constructed the Sexual Harassment Video (SHV) dataset from YouTube and proposed a CNN-LSTM architecture: CNN (with pre-trained backbones) extracts spatial features, while LSTM captures temporal dynamics, followed by three dense layers. Experiments on benchmark action datasets (Hockey Fight, Movie Violence) achieved state-of-the-art accuracy at the time, demonstrating cross-domain robustness.

Shahria et al. [9] also collected social-media videos and designed a CNN with multiple convolution kernels (32, 64, 128, 256 layers), each with zero-padding and max-pooling. They fused standard and separable convolutions for speed, and used three fully-connected layers for final feature reduction—a pragmatic trade-off between accuracy and inference time.

Aggarwal, Agrawal, and Sureka [10] tackled the problem via metadata mining rather than pixel-level analysis. Using YouTube API downloads and manual annotation, they defined discriminative features (linguistic cues in titles/descriptions, popularity, duration, category) and employed one-class classifiers to identify vulgar, abusive, or bullying videos. Their evaluation exceeded 80% accuracy, proving metadata-only approaches viable for large-scale screening. Gangwar et al. [11] focused on Child Sexual Abuse (CSA) material detection, decomposing the problem into two simpler sub-tasks and constructing large-scale labelled datasets for deep neural network training—addressing the data scarcity issue in a critical subdomain.

Deep spatiotemporal networks (CNN-LSTM) dominate visual detection, but they require substantial labelled data. Metadata-based methods offer a lighter, scalable alternative. The lack of publicly available, large-scale sexual harassment video datasets remains a major bottleneck; efforts like SHV and CSA datasets are initial steps, yet generalisation across diverse real-world scenarios is still unverified.

Overall, text-based methods can effectively identify linguistic features in harassing behaviors, yet fail to fully capture non-verbal cues such as facial expressions, body postures, interpersonal distances and physical interactions. Computer-vision-based approaches can supply complementary behavioral information, whereas their practical performance is constrained by insufficient scenario-specific data, object occlusion, high computational cost and uncertain generalization capability in real-world settings. Few existing studies simultaneously consider recognition accuracy, deployment efficiency and privacy protection within a unified framework. To address these research gaps, this study integrates lightweight facial expression recognition and human pose estimation components to improve the accuracy-efficiency trade-off, and lays a technical foundation for future privacy-aware multimodal risk-perception systems.

3. MODELING ALGORITHMS

3.1 Facial Expression Recognition Algorithm Based on POSTER V2-CGNet

In recent years, various deep learning models have emerged in the field of Facial Expression Recognition (FER). Among them, POSTER V2 has achieved breakthrough progress in balancing accuracy and efficiency with its streamlined two-stream network, window cross-attention design, and lightweight Transformer architecture. Compared with traditional single-stream models, POSTER V2 separates image appearance features and key point geometric features through a two-stream design, and uses the sparsity and geometric invariance of key points to guide image features to focus on expression-sensitive areas, effectively alleviating the inter-class confusion problem caused by global feature redundancy in single-stream models. Although POSTER V2 has shown high accuracy and computational efficiency in complex facial expression recognition tasks, its limitations in the single-modal feature optimization stage restrict the model's robustness in real-world scenarios. Therefore, this chapter improves its shortcomings to improve the accuracy and robustness of expression recognition.

In the POSTER V2 framework, IR50 serves as the Image Backbone, responsible for extracting multi-scale global expression features from original input images. Its basic module adopts an improved Bottleneck_IR_SE structure, which realizes adaptive calibration of channel dimensions by embedding a lightweight Squeeze-and-Excitation (SE) attention module to enhance the network's expression recognition ability. This module includes two key stages: Global Information Embedding (Squeeze) and Adaptive Recalibration (Excitation), and the specific process is shown in Figure 1.

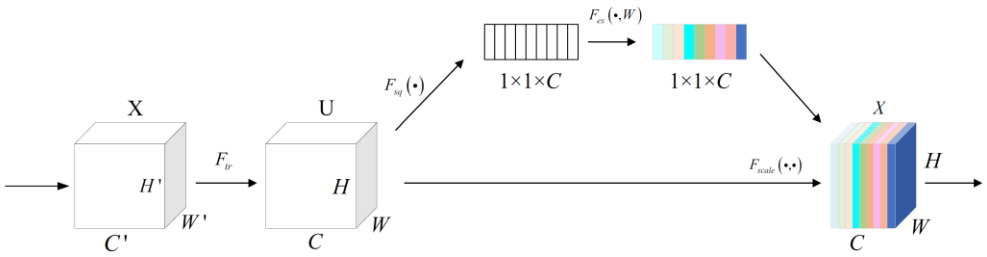


Figure 1. Squeeze-and-Excitation attention mechanism.

The SE module enhances the model's ability to perceive key semantic channels by modeling the dependencies between channels. However, in complex facial expression recognition tasks, the global channel compression strategy of the SE module has obvious limitations. On the one hand, the global average pooling operation compresses the spatial dimension into a single scalar, resulting in the complete loss of local spatial structure information, making it difficult to distinguish regions with significant spatial distribution differences but similar channel features. On the other hand, the SE module only focuses on the global dependencies of the channel dimension and does not consider the anisotropic distribution

characteristics of expression features, making it difficult to capture the spatial distribution characteristics of expression changes.

To address the above shortcomings, this study adopts the Coordinate Attention (CA) module to replace the SE module and construct a position-sensitive channel enhancement strategy. By combining the position information of the feature space with channel attention, it enhances the model's ability to perceive the target spatial direction, thereby enabling more accurate identification and positioning of multi-scale objects in images.

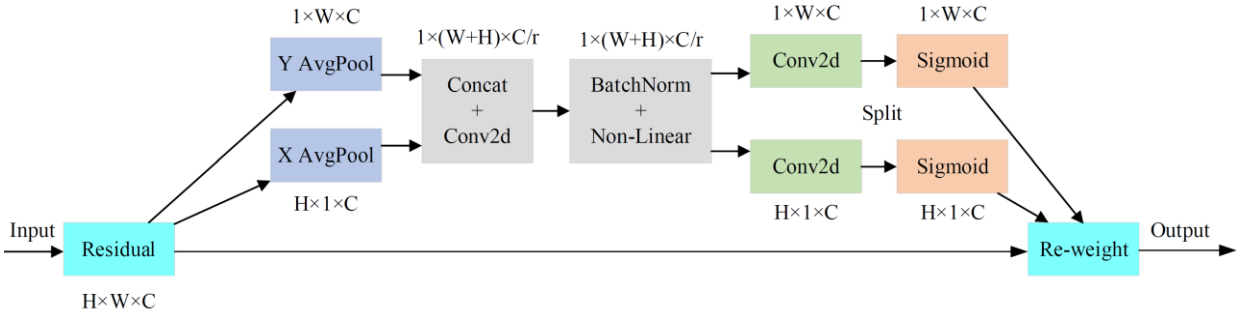


Figure 2. Coordinate Attention mechanism.

The CA module realizes fine-grained position perception by decoupling spatial dimensions, and the implementation process includes two steps: Coordinate Information Embedding (CIE) and Coordinate Attention Generation (CAG). Its specific process is shown in Figure 2. By performing attention modeling in both horizontal and vertical directions, the CA module effectively performs refined processing on feature maps, helping to capture spatial position information, thereby enhancing the model's ability to locate and recognize target features, and improving the ability to distinguish intra-class differences and robustness in complex scenarios.

In the POSTER V2 architecture, image features and key point features directly perform cross-modal interaction without dynamic screening of single-modal features, resulting in irrelevant features and effective information participating equally in attention calculation. To solve this problem, this section designs a lightweight Gate-based Spatial-Channel (G-SC) module before cross-modal interaction to dynamically filter spatial and channel features respectively. The G-SC consists of a Spatial Gate (SG) and a Channel Gate (CG) working in parallel.

The Spatial Gate adopts depthwise separable convolution. First, it extracts local expression texture features through 3×3 depthwise convolution, then uses 1×1 pointwise convolution to perform non-linear transformation on these features to generate intermediate response features. After Sigmoid normalization, this feature restores to the original resolution through Bilinear Upsample to generate a spatial weight map for enhancing the response of key expression regions. The Channel Gate counts the global response at the channel level through global average pooling and models the dependencies between channels using a Multi-Layer Perceptron to generate a channel weight vector. Finally, the weights of the two branches are fused with the original feature map through residual connection to obtain the final output feature map. After introducing G-SC, the cross-modal interaction process of the model is optimized as follows: image features and key point features pass through independent gating units respectively to generate purified features before information interaction. The structure of the POSTER V2-CGNet model is shown in Figure 3.

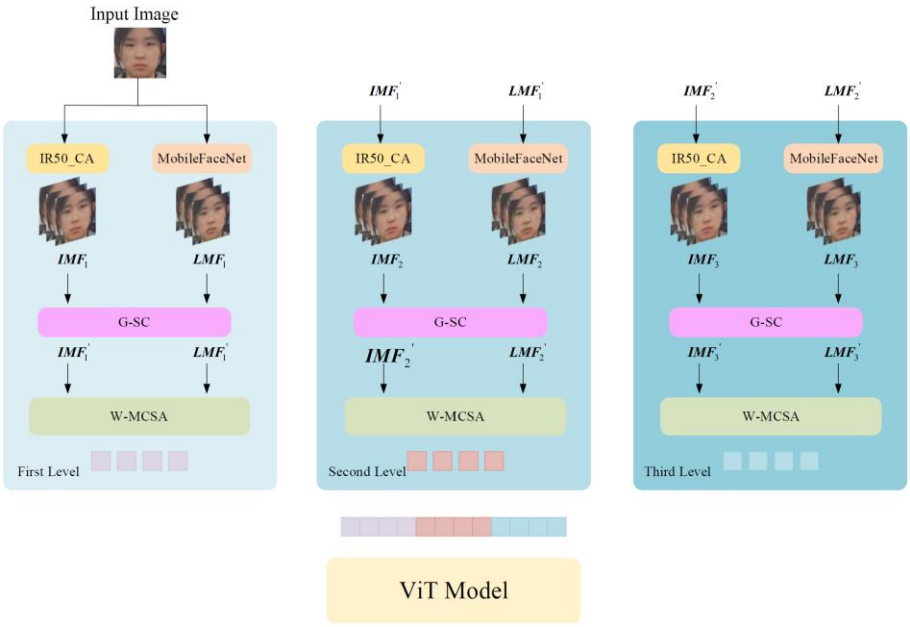


Figure 3. Architecture of POSTER V2-CGNet.

3.2 Human Pose Estimation Algorithm Based on YOLOv8s-Pose-SRFA

Compared with traditional two-stage pose estimation algorithms, YOLOv8-Pose has achieved dual breakthroughs in efficiency and accuracy through end-to-end architectural innovation. Traditional methods usually need to first locate the human body frame through a target detector, then perform key point regression on the cropped region. This cascaded structure not only doubles the video memory usage but also increases redundant calculations due to the separation of the detection and pose estimation models, resulting in repeated feature extraction. YOLOv8-Pose combines detection and pose estimation into a single-stage framework. By sharing backbone network features, it introduces depthwise separable convolution and Cross Stage Partial (CSP) lightweight modules, significantly reducing the model's parameters and computational complexity. At the same time, it adopts a Multi-Task Decoupled Head to separate detection and key point regression tasks, and dynamically integrates multi-level semantic information with a Path Aggregation Network. On the premise of maintaining the positioning ability of high-resolution feature space, it effectively alleviates the key point misdetection problem in scenarios of limb crossing and occlusion. This chapter takes YOLOv8s-Pose as the baseline model and improves its insufficient cross-level feature fusion efficiency and limited local key point positioning accuracy to enhance performance in real-time monitoring scenarios.

As a key module connecting the backbone network and the detection head, the Neck structure's multi-scale feature fusion ability directly determines the upper limit of model performance. Aiming at the feature mismatch and computational redundancy problems in the Feature Pyramid Network of YOLOv8s-Pose, this study proposes a Slim-Neck (Slimmed Feature Fusion-Neck) lightweight feature fusion network, whose design concept is reflected in two aspects: feature fusion and computational optimization. The Neck structure of YOLOv8s-Pose only realizes the fusion and interaction of features at different levels through a single linear upsampling operation and skip connection, without channel alignment of features, which is likely to lead to channel mismatch between features, resulting in information loss and reduced feature fusion accuracy. On the other hand, this simple operation lacks deep reuse of low-level high-resolution features, leading to the gradual attenuation of detail information in the process of feature transmission. In addition, when constructing the feature pyramid, although the C2f module improves the model's feature expression ability by doubling the number of channels step by step, it is difficult to offset the huge parameters and computational complexity it brings.

In terms of feature fusion, Slim-Neck realizes a multi-level cascaded feature fusion strategy, breaking through the limitation of single-path fusion of the original model. It adopts a bidirectional cross-layer skip connection structure to transmit low-level high-resolution features processed by pointwise convolution to each decoding stage, realizing bidirectional fusion of detail and semantic features. In addition, Slim-Neck introduces pointwise convolution before each upsampling and feature concatenation to compress and adjust the channels of feature maps, unify the feature dimensions

of different levels, thereby alleviating information distortion and interference caused by scale differences, and improving the information interaction ability of multi-level features.

In terms of lightweight, aiming at the problem of exponential expansion of the number of channels in the feature pyramid of the original model, Slim-Neck realizes the non-linear mapping of high-level features to low-dimensional space through pointwise convolution, which not only reduces the number of channels of feature maps, lowers parameters and computational load, but also promotes the fusion and interaction between features at different levels. In addition, Slim-Neck adopts a unified C2f module in all stages, and the features at all levels maintain 256 channels, retaining depthwise separable convolution inside to reduce parameters. At the same time, Slim-Neck reconstructs the downsampling method of the original model into cascaded 3×3 convolution, compressing the feature map step by step, which can effectively alleviate the information loss problem caused by direct downsampling and compression of the feature map in the original model. This structure reduces the computational complexity of the feature pyramid from $O(N^2)$ to $O(N)$ while maintaining multi-scale perception ability, improving the model's computational efficiency. The structure of Slim-Neck is shown in Figure 4.

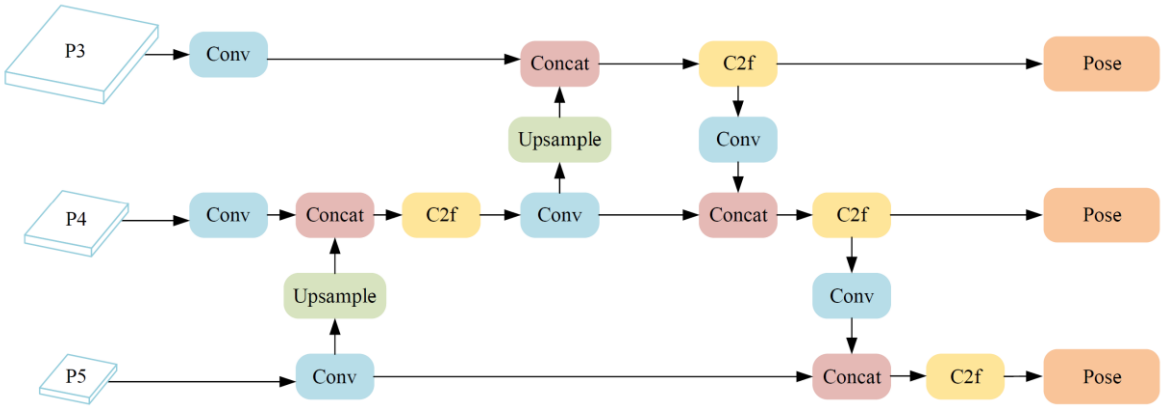


Figure 4. Structure of Slim-Neck module.

We propose a novel Receptive-Field Attention (RFA) mechanism and design the RFAConv convolution operation based on it, aiming to optimize the feature extraction process by dynamically generating non-shared parameters at the receptive field level. The structure of RFAConv is shown in Figure 5.

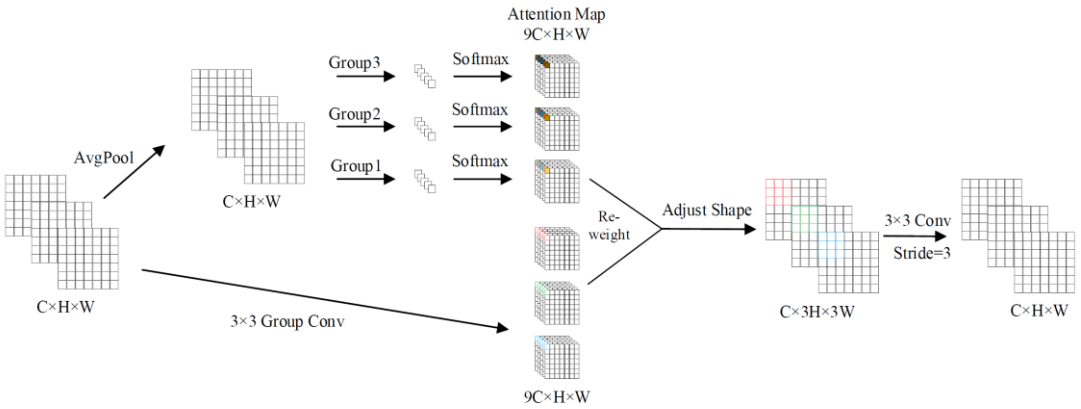


Figure 5. Architecture of RFAConv module.

An improved RFAHead detection head is proposed based on RFAConv. Its core innovation is to enhance the expression ability of local features through the dynamic receptive field attention mechanism and realize the parallel processing of detection and key point tasks by combining the decoupled prediction head. On the basis of sharing backbone features, RFAHead optimizes detection confidence and key point heatmaps through independent branches respectively, thereby realizing efficient collaboration of multi-task.

The structural design of RFAHead revolves around two characteristics of pose estimation tasks: first, the positioning accuracy of key points is highly dependent on the fine-grained expression of local features; second, the spatial relationship between key points needs to be modeled through contextual information. The feature enhancement module of the detection head adopts a two-level RFAConv stacking structure. The first-level RFAConv generates receptive field attention weights through average pooling and grouped convolution, captures the local spatial correlation of joint regions, and dynamically enhances the feature response in target regions; the second-level RFAConv further captures the spatial dependence between key points on the basis of refined features. This hierarchical design gradually focuses on the semantic regions related to the target pose through two weight adjustments. In the feature reconstruction stage, tensor deformation operation is used to restore the weighted features to high-resolution space, and geometric prior is introduced to help improve the sub-pixel level prediction accuracy of key point coordinates.

4. EXPERIMENTATION AND ANALYSIS

4.1 Experimental setup

This study conducts model training and validation based on the RAF-DB and AffectNet datasets. RAF-DB contains 15,339 diverse images collected from the Internet, covering seven basic emotional categories (Happy, Sad, Surprise, Fear, Disgust, Anger, and Neutral). The data covers different ages, races, lighting conditions, and head poses, and is completed by 315 annotators through multiple rounds of quality verification to ensure high-consistency annotation. To improve the model's robustness, this study strictly cleans the original data, removing blurred or severely occluded samples. Finally, 12,271 training images and 3,068 test images are used, and about 2,454 validation images are divided from the training set at a ratio of 4:1 to optimize the model's generalization performance. AffectNet is currently the largest facial expression recognition benchmark database, containing about 1.2 million web images, but there is a serious class imbalance with the maximum class difference reaching 20:1. The data is captured from search engines through emotional keywords, covering real-world scenarios with cross-race, age, and complex lighting conditions, and is refined annotated by a professional annotation team combined with semi-automatic tools. This study mainly adopts its 7-class basic expression version, with 280,401 training images and 3,500 test images.

All experiments are conducted based on the PyTorch deep learning framework, with experimental equipment being a single NVIDIA RTX3090 24GB, and all experiments use unaligned original data. This paper adopts the Adam optimizer with a hidden dimension of 64 and an output dimension of 32. The learning rate is set to 1e-5, and the batch size is set to 32. This experiment uses early stopping to train the model to prevent overfitting, with the early stopping parameter set to 10. Specifically, after each training epoch on the training set, the model's performance is validated on the validation set. When the task loss function of the validation set does not decrease for 10 consecutive times, the early stopping mechanism is triggered to terminate training. The parameters from the previous iteration result are used as the final parameters of the model, and the model is tested on the test set to obtain the final experimental results.

4.2 Experimental results

This experiment verifies the comprehensive performance advantages of POSTER V2-CGNet by comparing it with mainstream algorithms. The experimental results are shown in Table 1. On the RAF-DB natural scene dataset, the POSTER V2-CGNet model achieves an accuracy of 91.35%, which is 0.28% higher than the second-best model LFNSB, 1.24% and 1.74% higher than the base models POSTER V2 and POSTER respectively, and 4.23% and 1.52% higher than the early representative methods MixAugment and EAC, showing significant performance advantages. In cross-dataset validation, the model achieves an accuracy of 66.25% in the complex AffectNet-7 scenario, 1.18% higher than POSTER V2, and close to the single-modal method LFNSB, but with lower computational complexity, indicating that the model achieves a better balance between efficiency and accuracy. On the AffectNet-8 dataset with sparse annotated samples and class imbalance, POSTER V2-CGNet achieves an accuracy of 62.39%, surpassing POSTER V2 and FerNeXt, and 1.34% higher than the POSTER method, verifying the model's strong adaptability to differences in data distribution. Although LFNSB has a slightly higher accuracy of 62.68% than POSTER V2-CGNet on AffectNet-8, its parameters and computational complexity are higher. Through the collaborative design of coordinate attention and lightweight gating units, POSTER V2-CGNet achieves significant model compression and computational acceleration while ensuring accuracy, and has higher practical deployment value.

Table 1. Experimental results.

Model	RAF-DB(%)	AffectNet-7(%)
MixAugment	87.12	65.47
EAC	89.83	64.89
Ada-DF	90.04	65.16
FerNeXt	88.43	64.28
POSTER	89.61	64.73
LFNSB	91.07	66.32
POSTER V2	90.11	65.07
Ours	91.35	66.25

The visualization effect of expression recognition by POSTER V2-CGNet is shown in Figure 6. It can be seen from the figure that the expression recognition effect is stable, the facial detection bounding box is accurate without offset or recognition errors, and the classification results are correct without misclassification.

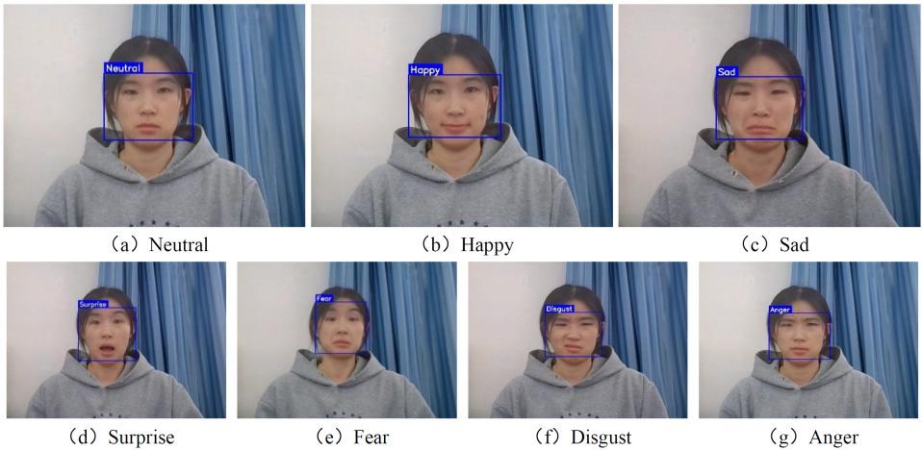


Figure 6. Visualization of expression recognition results using improved POSTER V2.

This section compares YOLOv8s-Pose-SRFA with other pose estimation algorithms to analyze its performance in terms of accuracy and computational efficiency, so as to comprehensively evaluate the model's performance. The experimental results are shown in Table 2. YOLOv8s-Pose-SRFA surpasses the second-best model HRNet in the mAP50 index, and its parameters are only one-third of HRNet, revealing the limitations of the two-stage model in the computational paradigm. It relies on multi-scale features and complex post-processing operations, resulting in high computational complexity and poor real-time performance; while YOLOv8s-Pose-SRFA improves accuracy and computational efficiency through a single-stage architecture and lightweight attention mechanism. Compared with OpenPose, the computational load of YOLOv8s-Pose-SRFA is only one-sixth of that of OpenPose, and the accuracy is increased by 1.5%. In addition, the parameters of YOLOv8s-Pose-SRFA are only 26% of that of Hourglass. It requires approximately 84.8% fewer FLOPs than OpenPose and achieves 49% higher computational efficiency than Hourglass.

Table 2. Experimental results of pose estimation.

Model	mAP50(%)	mAP50-95 (%)	FLOPs(G)
HRNet	86.2	58.5	57.6
OpenPose	84.9	57.1	153.2
Hourglass	81.8	54.3	45.8

YOLOv8s-Pose	85.3	59.5	30.4
Ours	86.4	61.2	23.3

5. CONCLUSION

This paper focuses on the prevention of gender-based violence and sexual assault in concealed campus scenarios such as seminar rooms. Aiming at the contradiction between privacy protection and behavior recognition in existing supervision mechanisms, it proposes a multi-modal intelligent student monitoring model that balances protectiveness, non-intrusiveness, and intelligence. In terms of technical innovation, the POSTER V2-CGNet algorithm is proposed, which replaces the traditional channel attention module with a coordinate attention module, adds a lightweight spatial-channel gating unit, and optimizes cross-modal feature interaction; the YOLOv8s-Pose-SRFA algorithm is designed, which adopts a Slim-Neck lightweight feature fusion network and a receptive field attention mechanism to improve the accuracy and efficiency of pose detection. Experiments based on RAF-DB and AffectNet datasets show that POSTER V2-CGNet achieves an accuracy of 91.35% on the RAF-DB dataset, and the mAP50 index of YOLOv8s-Pose-SRFA reaches 86.4%, both superior to mainstream comparative models. This study realizes in-event early warning of risky behaviors in concealed campus scenarios and support for post-event psychological intervention, providing an effective solution for the technical governance of campus sexual assault problems.

ACKNOWLEDGMENTS

The author would like to thank Professor Yi Yang of Tsinghua University for valuable discussions and methodological feedback during the development of this research.

REFERENCES

[1] EDUPUGANTI V. Harassment detection on twitter using conversations [D], Ohio: Wright State University, 2017.
[2] POTHA N, MARAGOUDAKIS M. Cyberbullying detection using time series modeling [C]. In: Proceedings of the 2014 IEEE International Conference on Data Mining Workshop. Shenzhen: IEEE, 373-382, 2017.
[3] RIZWAN K, BABAR S, ARISHA A. HarX: Real-time harassment detection tool using machine learning [C]. In: Proceedings of the 2021 International Conference of Modern Trends in Information and Communication Technology Industry. Sana'a: IEEE, 1-6, 2021.
[4] SAEIDI, MOZHGAN, MILLIOS, BERTON. Categorizing online harassment on Twitter [C]. Machine Learning and Knowledge Discovery in Databases. Communications in Computer and Information Science, Cham: Springer, 1168: 283-297, 2019.
[5] SAEIDI, MOZHGAN, EVANGELOS MILIOS. Graph Convolutional Networks for Categorizing Online Harassment on Twitter [C]. In: Proceedings of the 2021 20th IEEE International Conference on Machine Learning and Applications. Pasadena: IEEE, 946-951, 2021.
[6] ALAWNEH, EMAD, JAFAR, Al FAYOUMI. Sentiment analysis-based sexual harassment detection using machine learning techniques [C]. In: Proceedings of the 2021 International Symposium on Electronics and Smart Devices. Bandung: IEEE, 1-6, 2021.
[7] JAMES, I. I., and V. I. OSUBOR. Hostile social media harassment: A machine learning framework for filtering anti-female jokes [J]. Nigerian Journal of Technology, 2022, 41(2): 311-317.
[8] ISLAM M S, HASAN M M, AKBAR, ARAFAT, MURAD. A deep spatio-temporal network for vision-based sexual harassment detection [C]. In: Proceedings of the Emerging Technology in Computing, Communication and Electronics, Dhaka: IEEE, 2021: 1-6.
[9] SHAHRIA M T, PROGGA F T, AHMED S, ARISHA, A. Application of neural networks for detection of sexual harassment in workspace [C]. In: Proceedings of the 2021 International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies, Bhilai: IEEE, 1-4, 2021.

[10] AGGARWAL N, AGRAWAL S. Mining YouTube metadata for detecting privacy invading harassment and misdemeanor videos [C]. In: Proceedings of the 2014 12th Annual International Conference on Privacy, Security and Trust, Toronto: IEEE, 84-93, 2014.
[11] GANGWAR A, GONZALEZ-CASTRO V, ALEGRE E, FIDALGO. AttM-CNN: Attention and metric learning based CNN for pornography, age and child sexual abuse detection in images [J]. Neurocomputing, 2021, 445: 81-104.