SODSR: A Three-Stage Small Object Detection via Super-Resolution Using Optimizing Combination

Xiaoyong Mei, *Member, IEEE*, Kejin Zhang, Changqin Huang, *Member, IEEE*, Xiao Chen, Ming Li, *Member, IEEE*, Zhao Li, Weiping Ding, *Senior Member, IEEE*, and Xindong Wu, *Fellow, IEEE*

***Abstract*—Face detection is a fundamental task in computer vision, yet remains challenging in educational settings due to the presence of objects of various sizes. Subpar detection can significantly impede subsequent tasks' performance. To address this, we present a novel framework, Small Object Detection Super Resolution (SODSR), inspired by super resolution (SR) techniques for feature-level images. SODSR comprises three stages: (1) Constructing a 3D model evaluation matrix to select optimal model combinations based on detection accuracy and image quality metrics. (2) Employing Double-thread FDN in the first stage to preprocess images, enhancing feature resolution for potential facial objects. (3) Leveraging Multi-head HyperNet in the second stage to augment face feature detection and improve accuracy. Finally, in the third stage, we introduce AFPGAN, a facial prior feature enhancement network, coupled with StyleGAN2 for texture and contour detail enhancement. Experimental results demonstrate that SODSR outperforms existing small object detection (SOD) models in both accuracy and visual fidelity.**

***Index Terms*—Combination optimization, face detail restoration, GAN inversion, small object detection, super resolution.**

Received 27 April 2024; accepted 25 July 2024. This work was supported in part by the National Key Research and Development Program of China under Grant 2022YFC3303600, and in part by the National Natural Science Foundation of China under Grant 61877020, Grant 62037001, Grant 62172370, and Grant 62120106008. The work of Ming Li was supported in part by the Jinhua Science and Technology Plan under Grant 2023-3-003a, and in part by the Zhejiang Provincial Natural Science Foundation under Grant LY22F020004. *(Corresponding author: Xiaoyong Mei.)*

Xiaoyong Mei, Kejin Zhang, Changqin Huang, and Xiao Chen are with the Zhejiang Key Laboratory of Intelligent Education Technology and Application, Zhejiang Normal University, Jinhua 321004, China (e-mail: cdmxy@zjnu.edu.cn; kejinz@zjnu.edu.cn; cqhuang@zju.edu.cn; chenxiao@zjnu.edu.cn).

Ming Li is with the Zhejiang Institute of Optoelectronics, Jinhua 321004, China, and also with the Zhejiang Key Laboratory of Intelligent Education Technology and Application, Zhejiang Normal University, Jinhua 321004, China (e-mail: mingli@zjnu.edu.cn.).

Zhao Li is with Hangzhou Yugu Technology Company, Ltd, Hangzhou 311113, China, and also with Zhejiang Lab, Hangzhou 311121, China (e-mail: lzjoey@gmail.com).

Weiping Ding is with the School of Artificial Intelligence and Computer Science, Nantong University, Nantong 226019, China (e-mail: dwp9988@163.com).

Xindong Wu is with the Key Laboratory of Knowledge Engineering With Big Data (The Ministry of Education of China), Hefei University of Technology, Hefei 230009, China, and also with the School of Computer Science and Information Technology, Hefei University of Technology, Hefei 230009, China (e-mail: xwu@hfut.edu.cn).

Recommended for acceptance by D. Tao.

Digital Object Identifier 10.1109/TETCI.2024.3452749

I. Introduction

THE profound merging of artificial intelligence and education is a notable attribute of educational informatization in the current era. As one of the application technologies, object detection can detect and locate students. However, the performance of the general object detection directly applied to the educational scenes is not ideal, because in the education scene, although the indoor object detection of teachers and students has the advantages of small environmental change and relatively fixed location, there are also problems such as occlusion, difference in light amount and varying size of the object. At the same time, due to the problems of image compression loss and block artifacts in image acquisition and storage in different classrooms, and the image sources are quite different. This is mainly caused by the insufficient generalization of the object detection model when small objects are detected. Due to the small number of pixels, the features of small objects are easily mixed with the background during detection, and the detector misses the extraction of features. In addition, the feature information of small objects is incomplete in the shallow layer of the model and does not exist in the deep layer of the model. Therefore, using the object detection model to detect small objects will result in missing or wrong detection. Therefore, how to improve the accuracy of SOD and its generalization ability of model combination in the real scene is a problem that needs to be solved. In addition, considering the particularity of educational scenes, object detection usually involves other downstream tasks, such as expression recognition. In order to ensure the sustainability of downstream tasks, face restoration is included in the research scope of this paper.

In recent times, object detection, a high-level computer vision task, has been increasingly implemented in educational teaching scenes. It can extract teaching information in educational scenes, such as estimating blackboard font and discerning teacher-student classroom behavior, assisting teaching management personnel to better understand classroom information quickly and efficiently. In some researches, object detection has been used to identify and count students in class or identify student behaviors and emotional changes during the learning process. Yi et al. [1] and Liu et al. [2] designed effective student number statistics systems using object detection and Zhang et al. [3] conducted real-time emotion detection of online students based on YOLOv5. However, these methods only focus on the detection of student faces which are regular in sizes, and it is easy to miss

and false detection when the small size face of students appears in the detection screen. This is because small faces often have low resolution and lack obvious texture and shape information. On the one hand, when the neural network is shallow, it may not have enough ability to extract sufficient feature information of the object, resulting in detection errors. On the other hand, when the network is deep, too much pooling may smooth out the details of the object or even remove the object entirely. Some researches have taken SOD into consideration, using multiple-scale learning, data augmentation, and other methods to improve the accuracy of SOD. For example, Li et al. [4] designed a data augmentation method different from traditional methods, expanded the volume of the dataset to improve the ability of model generalization, and improved the detection of student behavior in the classroom by addressing insufficient SOD accuracy in complex classroom environments. However, when the detection image contains multi-scale objects and some sizes are small, most of these methods simply detect objects with size changes, and do not alleviate the problem of visual information loss caused by the natural low-resolution properties of small objects. Therefore, designing a suitable neural network structure that can effectively extract features and distinguish targets in strong interference is still a challenging problem. Obtaining sufficient characteristic information is crucial for accurate detection and reducing the occurrence of false positives.

Due to the limited amount of information provided by small objects, detecting small faces in educational scenes remains a challenge. The detector based on DNN has proposed some effective methods for small objects. As a low-level computer vision task, the high resolution (HR) image reconstructed by SR has more original scene features, which naturally promotes the accuracy of SOD detection, and can be well applied to SOD. At present, a few researches have applied the combination of the two to the field of education. For example, Rothoft et al. [5] proposed a method to detect students' attention based on the low resolution (LR) of face image and focus classification problems, combined with SR and the distribution of the focus points in two dimensions. It is worth noting that although the combination of SR and SOD increases the amount of pixels of the detected object, but the obvious efficiency of the whole image via SR is not high, the increased amount of pixels is limited in visual discrimination, and the loss of texture or shape features seriously hinders human visual perception.

In order to solve the above problems, we screen the SR and SOD models by combining and optimizing the evaluation matrix, and accurately locate the potential areas by suppressing the background and filtering the prominence of the entire image, effectively reducing false positives and improving the performance of SOD. Then, SR is used to highlight the resolution and detail features of face images, suppress irrelevant details, and improve the visibility and discriminability of small object faces. We combine the advantages of SR and SOD and learn from them by combining optimization of SR and SOD seamlessly integrated architecture SODSR. This method makes the architecture combine the advantages of SR and SOD at the same time, thus improving the feature extraction and object detection performance. Specifically, we first construct a 3D model capability evaluation matrix, combining the image quality evaluation metrics, detection accuracy and parameter number of a single SR and SOD model to optimize the basic model required for the first two stages of the architecture. Then, to optimize the basic models, we propose the double-thread feature distillation network (double-thread FDN) to enhance the detailed feature in the images and design a hypernet with multiple heads (multi-head Hyper) to improve SOD accuracy by increasing the number of prediction heads. In the third stage, in order to solve the problem of missing facial semantic and structural information, the face SR network AFPGAN based on GAN inversion is used to promote the recovery and enhancement of face prior SR tasks by using pre-trained network weights.

The contributions of this paper are as follows:

- We propose a three-stage small object detection via super resolution architecture, SODSR for reconstructing faces images in educational scenes to ensure data security for downstream visual tasks. The architecture is extensible and can be optimizeed and combined at all stages using capability evaluation matrix.
- We construct a model capability evaluation matrix for combining super resolution and small target detection models. The evaluation matrix utilizes natural image quantization metrics NIQE, hyperIQA and AP and overall parameter count to balance performance and speed.
- We optimize the first two stages of the SODSR, and design the double-thread FDN and the multi-head HyperNet. The former utilizes multiple information distillation and the attention mechanism to refine and enhance features strengthening the generalization of the network. On the other hand, the latter modifies the detection head and neck part to alleviate the problem of missing detection caused by large size changes of the object, which improves the detection ability of the network.
- Considering the reference-free nature of real-world scene images, GAN inversion is adopted in the third stage to restore facial details and rely on prior information from real images to reconstruct high-definition faces which ensure the authenticity of the repaired faces.

The remainder of this paper is organized as follows. In Section II, we discuss the related work. In Section III, we present the proposed method and architecture. The experiment results are given in Section IV. Section V concludes this paper.

II. Related Work

A. Small Object Detection

The purpose of SOD is to accurately detect small objects (32×32pixels) with limited visual information in an image. Developed as a deep learning-based technique, compared to the traditional DPM detection algorithm [6], SOD R-CNN [7] achieves significant improvement over R-CNN. However, detecting small objects proves challenging considering their low resolution and the complex scenes in real scenarios. Researchers have improved the accuracy of SOD by enhancing sample quantity and feature scale. To address the impact of insufficient sample size, Kisantal et al. [8] oversampled small object images

by copying and pasting them multiple times to enhance the image. ZOPH et al. [9] learned an ideal data augmentation strategy using an automated method that selects the optimal method from a set of techniques. Although data augmentation improves the model's ability to generalize, it is computationally costly and unsuitable strategies can harm model effectiveness.

Small objects hold information within the shallow features of the image, but that information gradually loses as the network layers deepen. To optimize the utilization of shallow features, researchers have utilized multiscale learning methods to fuse shallow and deep features. The most well-known of these is FPN [10], which has progressed by the top-down fusion, simple bidirectional fusion, and complex bidirectional fusion. FPN is a widely accepted approach, acting as the baseline top-down fusion model for major models such as Faster R-CNN [11], YOLOv3 [12], and RetinaNet [13]. Simple bidirectional fusion (PANet) was initially proposed by Lui et al. [14], which integrates a bottom-up fusion path with FPN, proving the benefits of bidirectional fusion. Complex bidirectional fusion builds on PANet, Tan et al. [15] proposed BiFPN to simplify the PANet structure by removing unidirectional nodes and adding skip connections to incorporate more features. To address the feature inconsistencies between different layers of pyramid features, Liu et al. [16] proposed ASFF to readjust the resolutions of each level of the pyramid, and allow for completing integration and better relational understanding between features. Overall, combining shallow feature information with deep semantic information is regularly effective in improving detection performance, but improving the performance of SOD beyond a certain point via multiscale learning is challenging due to various computational and noise factors. Additionally, small objects detected at long distances frequently have low resolution, ultimately impairing overall visual perception.

In addition, it should be noted that the purpose of common SOD is mostly to detect quantity or potential abnormalities, and there is usually no subsequent visual task. The SOD in the educational scene is the person, the detection purpose is targeted, and according to the corresponding educational task, the object to be detected usually needs better image quality, for example, the accuracy rate of emotion recognition needs to ensure that the facial features are rich enough. Therefore, the application of common SOD in the educational scenes may cause the acquired data to be unusable, and the analysis results can not meet the requirements.

B. Super Resolution

CNN-based SR has been concerned ever since deep learning techniques were introduced to the field of SR [17]. SR mainly adopts three types of network construction, such as residual-based, attention-based, and GAN-based algorithms. Residual-based algorithms can minimize complexity and learning difficulty of network by learning the residual image between input and output images. Kim et al. [18] pioneered the use of global residuals in modeling, utilizing 20 convolutional layers. Lim et al. [19] improved the residual block and solved the gradient explosion by stacking residual blocks. Additionally, Kim et al. [20] used a recursive structure to reduce parameters count, while Hui [21] and Liu et al. [22] used information multiple distillation to construct a lightweight network. Residual-based algorithms are a popular feature of lightweight networks, which speed up learning and convergence of network.

Attention mechanisms have demonstrated improvement in SR performance. Mei et al. proposed NLSN [23] and CSNLN [24] methods, which respectively utilized a non-local sparse attention network and fused information to improve computation efficiency. Dai et al. [25] incorporated second-order feature statistics to adaptively adjust high-level features and enhance their discriminative ability. To address correlations of inter-layer, Niu et al. [26] introduced the global attention network (GAN) which assigns different weight values to each layer. However, the scope of improvement using attention mechanisms is restricted by the number of channels and resolution.

The first algorithm to use GAN to produce high-resolution images that are natural and authentic was SRGAN [27]. Though its performance was remarkable, it struggled in generating high-resolution images due to unstable training and pattern collapse. To address this problem, Wang et al. [28] improved the network structure and introduced residual structures to achieve better visual quality. The Real-ESRGAN [29], developed later, trained on purely synthetic data to reduce the noise caused by degradation. Compared to other SR algorithms, GAN-based algorithms leverage their powerful generation ability to achieve better visual reconstruction results.

C. Combination of SOD and SR

In recent years, improving SOD performance via deep learning-based SR guidance has become a research hotspot, and the combination of the SOD and SR has greatly promoted the development and application of SOD in various fields, such as, remote sensing images, aerospace and transportation, Cao et al. [30] used SR to train more robust detectors to transfer images from low-resolution spaces to high-resolution spaces, and avoid the degradation of detection performance caused by low-resolution small objects. Wang et al. [31] constructed a benchmark dataset RSSOD and proposed a multi-class cyclic SR network with residual feature aggregation to assist SOD and improve its performance. In pedestrian detection, Pang et al. [32] designed a framework for small pedestrian detection, and utilized SR to explore the relationship between large and small pedestrians to improve the performance of small pedestrian detection. The combination of SOD and SR has been successful in various fields, however, its application in the education scene is limited. One reason for this is that current combination models mainly focus on detecting the quantity of objects, which in educational scenes is primarily used for people counting, and providing little assistance for other educational visual tasks. Therefore, we expabd the application scenes of combination models to the reconstruction of the detected object and visual presentation.

In summary, we introduce the combination of SR and SOD in educational scenes, emphasizing accuracy in face detection and clear visual effect. To realize this objective, we propose

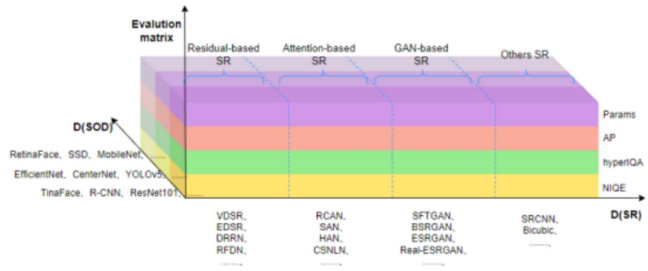


Fig. 1. Capability evaluation matrix diagram of model combination.

a three-stage combination architecture. The specifics of this methodology are detailed in the third section.

III. Methods

In this section, we will introduce the overall process of SODSR construction and the network of each stage. The architecture of SODSR is divided into two parts. In the first part, the basic model of SODSR in the first two stages is optimally selected by using 3D capability evaluation matrix. In the second part, the basic network model of each stage is optimized to improve the performance of face detection and image enhancement. In the first stage, we first use SR network to pre-process the input image to improve the pixel amount and feature richness of the detection object. In the second stage, we use the SOD network to detect the pre-processed images and crop the detected faces. In the third stage, we use the face SR network to restore the texture of the cropped faces, and finally obtain high-quality faces. The above is the overall description of SODSR, the specific description are detailed in the following sections.

A. Model Combination Strategy

In order to optimize and match the most suitable combination model, we propose a model combination strategy, whose 3D capability evaluation matrix is shown in Fig. 1. The whole matching process needs to select and combine optimization from SR sets $D(SR) = \{SR_1, SR_2, \ldots, SR_n\}, n \in N^+$ and SOD sets $D(SOD) = \{SOD_1, SOD_2, \ldots, SOD_m\}, m \in N^+$, and finding the optimal balance point based on a set of evaluation metrics $D(E) = \{E_1, E_2, \ldots, E_z\}, z \in N^+$.

A 3D capability evaluation matrix $M_(ce)$ is created to assess various SR and SOD models using multiple evaluation metrics. This matrix allows for an easier comparison and selection of models based on their respective metric effects, providing guidance on choosing the most effective model. Fig. 1 depicts the x and y-axes as $D(SR)$ and $D(SOD)$, respectively. The x-axis is divided into four categories based on different model strategies, which are residual-based SR, attention-based SR, GAN-based SR and other. Each type of network model contains at least four SR algorithms. According to the number of network layers, the Y-axis is divided into three categories, which are high-performance networks with large number of parameters, high-efficiency networks with small number of parameters, and balanced networks between them. Each type of network model includes at least three SOD algorithms. The z-axis corresponds to the selected evaluation metric, which contain NIQE, hyper-IQA, AP and pre-trained parameter quantity for both SR and SOD.

In order to verify the effect of the model combination in a specific educational scene, we choose a classroom with a dense population of students as the research object, with the purpose of finding and clearly showing the faces of students, and the steps are as follows: First step, selecting three SOD algorithms from $D(SOD)$, according to the detected AP value and the pre-training weight of the SR, select the most appropriate SR with its combination which yields three combination models. In the second step, for the three combination models obtained, we refer to the overall pre-training parameters and AP values and select the combination model that can balance them. Using these steps, we select RFDN (SR) [22] and YOLOv5 (SOD) [3] as the basic models for the first two stages. In addition, to address the problem of restoring facial texture to generate realistic images, we use the GAN Inversion method to restore facial details.

B. Small Object Detection Super-Resolution

Our goal is to detect student faces which sizes vary due to the distance from the camera in the classroom and restore their textures. To do this, we propose a three-stage small object detection via super-resolution architecture called SODSR, which is illustrated in Fig. 2. The architecture consists of three models, of which the first model is used to improve the pixel amount and object size of the image, the second model is used to detect the face of student, and the third model is used to repair the face. The specific process is as follows. Given a LR image of student faces, the double-thread FDN is used in the first stage to enrich the feature of the image and generate an intermediate SR image named SR_{medi}. After SR_{medi} is sent into multi-head HyperNet of the second stage, features are extracted and fused using bankbone and neck, and the four prediction heads in the head are used to detect and locate the faces of the students. The optimal bounding boxes are screened by non-maximum suppression and the detected images SR_{ded} are obtained. To obtain clearly distinguishable student faces, the AFPGAN model based on GAN Inversion is used in the third stage, and a pre-trained facial reconstruction model is used to degrade and reconstruct the image P_{face} processed by SR_{ded}, and a recognizable face SR_{face} is obtained. The proposed architecture is detailed later in this paper.

C. Double-Thread FDN

Larger displays of faces can provide models to obtain richer facial features. However, when the proportion of face pixels in an image is small, fewer features can be extracted, negatively impacting the success of face visual tasks. To address this, we pre-process the image by utilizing SR to increase its size and pixel count, thereby reconstructing some texture details and enriching feature representation.

According to the model combination strategy, RFDN is selected as the basic model, and on this basis, the model of the first

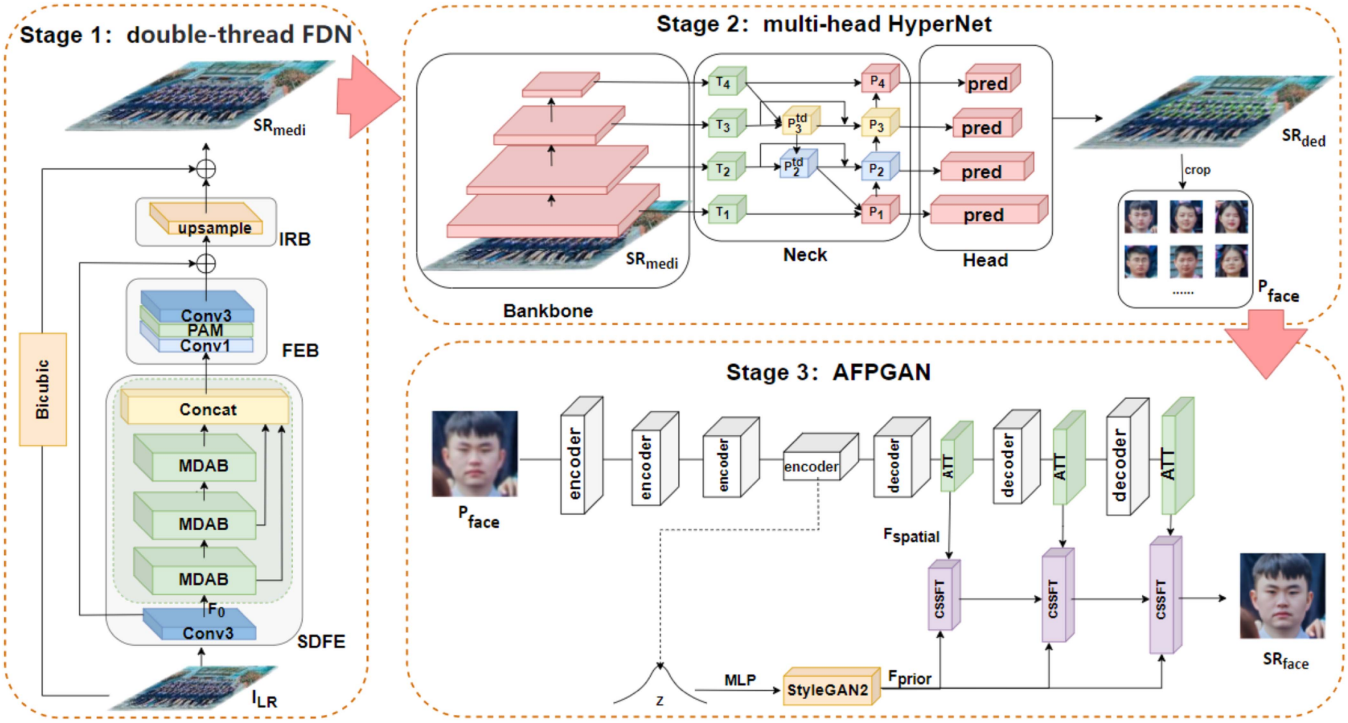


Fig. 2. The architecture of SODSR. Stage 1 is a SR model which used to improve image resolution and target size. Stage 2 is a SOD model which used to detect faces. Stage 3 is also a SR model which used to restore texture details of the detected faces. SDFE refers to deep and shallow layer feature extraction, MDAB reconstruct rich refining features. FEB enhanced detection of image features. PAM is used to enhance pixel-level features. IRB stands for the image reconstruction block. ATT refers to the attention mechanism. CSSFT refers to the channel segmentation space feature transformation layer.

stage of SODSR is designed, which is called double-thread FDN, as shown in Fig. 2. To further improve the image reconstruction capabilities of the RFDN and to enrich the semantic feature of images, we have implemented the following reconstructs: firstly, we introduce self-attention and pixel attention mechanisms. The former captures long-range dependencies within features, while the latter enhances pixel-level responses of features. secondly, added an additional branch to preserve the shallow information of the input LR image, which enriches the features while minimizing loss. Double-thread FDN comprises three main network components which contains Shallow and Deep Feature Extraction (SDFE), Feature Enhancement Block (FEB), and Image Reconstruction Block (IRB). To initiate its execution process, we input the LR image into double-thread FDN, which subsequently undergoes the following networks.

SDFE: Through the extraction of shallow features from the original image and the stacking of refined features of various depths, a fusion of shallow and deep features is attained. This process mitigates the information loss caused by an increase in convolutional layers during forward propagation while also preserving the shallow feature information of the original image. SDFE realizes two functions which contain shallow feature extraction and deep feature refinement, the former is completed using 3×3 convolution as follows:

$$F_0 = f_{conv3}(I_{LR}) \tag{1}$$

where f_{conv3} denotes the 3×3 convolution, and F_0 denotes the extracted shallow features.

Deep feature refinement is completed by three cascaded Multi-Residual Attention Block (MDAB). MDAB consists of two parts, which are Progressive Refinement Block (PRB) and Channel-Spatial Attention Block (CSAB), where the former learns to reconstruct rich hierarchical features and the latter assigns weight values to dual-level features. Fig. 3 depicts the network.

PRB: The $F_0 \in R^{h\times w\times c_{64}}$ with 64 channels is initially fed to the PRB, which is bifurcated into two components called Distillation Components (DC) and Cascade Components (CC) respective. DC compresses and reduces the number of channels to half their original value using a 1x1 convolution for efficient compression while removing redundant parameters. Meanwhile, CC gradually maps deeper features using a residual structure while retaining effective information through the Leaky ReLu activation function to avoid the complete rejection of smooth features. The final layer in CC uses a 3×3 convolution. The formula is as follows:

$$C_1 = H_{CC_1}(F_0) \tag{2}$$

$$C_n = f_{conv3}(H_{CC_{n-1}}(\ldots H_{CC_2}(C_1))), n \in [2,3] \tag{3}$$

$$D_1 = H_{DC_1}(F_0), D_n = H_{DC_n}(C_{n-1}), n \in [2,3] \tag{4}$$

where H_{CC_1} and H_{CC_n} denote the first and n-th deep feature extraction block, C_{n-1} and C_n denote the input and output features of the H_{CC_n}, H_{DC_1} and H_{DC_n} denote the first and n-th feature distillation block, and D_{n-1} and D_n denote the input and output features of the H_{DC_n}.

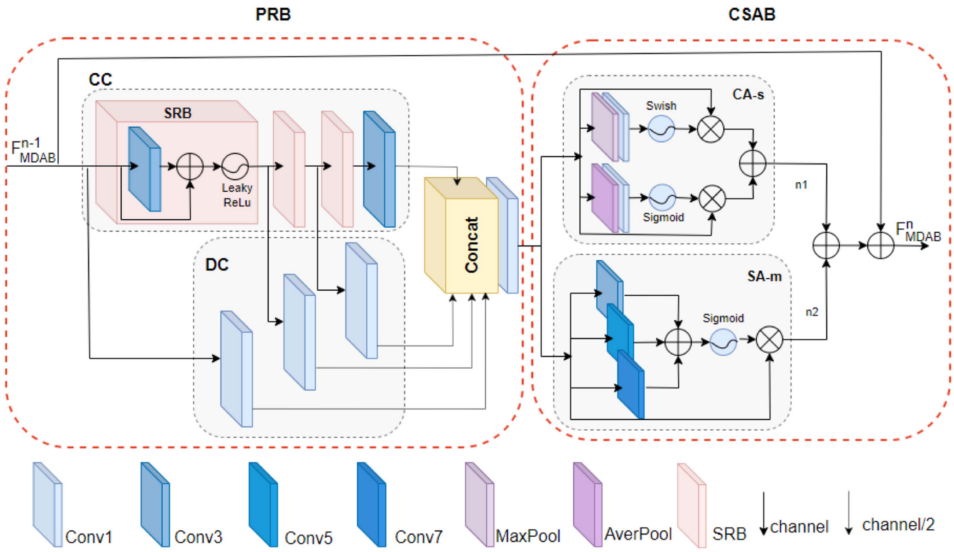


Fig. 3. The network of MDAB.

Following incremental feature computation, the feature map from the last layer of CC and a fixed number of n-th distilled features of DC are aggregated to fuse multi-channel feature fusion. This deepens the representation abilities of the features at different levels and retains a part of the original featue of image. Additionally, a 1×1 convolution is performed on all intermediate features to fuse and reduce their dimensionality, which reduces redundancy and noise. The formula is as follows:

$$F_{PRB} = f_{conv1}([D_1, \ldots, D_n, C_n]) \tag{5}$$

where F_{PRB} represents the fused feature, $[\cdot]$ denotes the aggregation on the channel and f_{conv1} denotes the 1×1 convolution.

CSAB: By feeding F_{PRB} into CSAB, parallel channels and spatial attention are used to obtain the weight value of F_{PRB} effectively and highlight the important features in the two domains, where channel attention concerns correlations between channels in the feature maps, while spatial attention highlights significant areas in the feature maps of PRB.

Channel Attention (CA-s). Each channel of the feature map is treated as a feature detector where each channel is assigned different weight values. Channels with significant features are assigned more weight while less attention is given to others. This attention mechanism improves the ability of the network to extract features through the fused feature F_{PRB}, which is able to focus on channels with more importance and suppress irrelevant features.

The $F_{PRB} \in R^{h\times w\times c_{128}}$ with 128 channels is fed into the CA-s, with average and max pooling in parallel to aggregate spatial information of the feature map. h_{Avg} and h_{Max} are normalized channel attention weights by corresponding nonlinear activation functions. Then the attention weights are multiplied by F_{PRB} to generate the corresponding channel-weighted feature map, and the two feature maps are fused to produce the channel-weighted feature map F_{CA-s}. The formula is as follows:

$$h_{Avg} = \delta_{sigmoid}(f_{conv1}(f_{AvgPool}(F_{PRB}))) \tag{6}$$

$$h_{Max} = \delta_{swish}(f_{conv1}(f_{MaxPool}(F_{PRB}))) \tag{7}$$

$$F_{CA-s} = h_{Avg} * F_{PRB} + h_{Max} * F_{PRB} \tag{8}$$

where $f_{AvgPool}$ and $f_{MaxPool}$ denote average pooling and max pooling, respectively. $\delta_{sigmoid}$ and δ_{swish} are sigmoid and swish of activation functions respectively. * is the weighted operation where the attention weights act on F_{PRB}, and F_{CA-s} is the channel-weighted feature map.

Spatial Attention (SA-m). Key facial regions such as the eyebrows, eyes, mouth and nose contain rich texture information of recognition tasks in facial image recognition tasks. Leveraging the spatial attention mechanism to assign a weight for different facial regions in the feature map makes the network focus more on the recognition of facial features and enhances the feature extraction ability of the network. According to the proportion of the face in the input image, if the receptive field is too small, only local features can be recognized, whereas too much irrelevant information is obtained if it is too big. As a result, the combination of multi-scale feature extraction and spatial attention is more effective in extracting weighted features more robustly.

The $F_{PRB} \in R^{h\times w\times c_{128}}$ is fed into the SA-m where it enters different convolutional layers with kernel sizes of 3x3, 5x5 and 7x7 to extract features with different receptive field sizes. To avoid boundary information loss, zero padding is added to the edges of feature map. The multi-scale spatial features extracted from the three branches are then element-wise summed to produce the fused feature map. It is then subjected to the sigmoid activation function from which a spatial attention weight map is generated within the range of [0,1]. The spatial attention map is subsequently multiplied by F_{PRB} to generate the spatially

weighted feature map F_{SA-m}. The formula is as follows:

$$h_{weight} = \delta_{sigmoid}(f_{conv3}(F_{PRB}) + f_{conv5}(F_{PRB}) + f_{conv7}(F_{PRB})) \quad (9)$$

$$F_{SA-m} = h_{weight} * F_{PRB} \quad (10)$$

where f_{conv5} and f_{conv7} denotes the 5×5 and 7×7 convolution respectively, h_{weight} is the spatial attention weights, and F_{SA-m} is the spatial-weighted feature map.

After completing the attention mapping on the different paths, the F_{CA-s} and the F_{SA-m} are aggregated using adaptive weight n1 and n2 to form a dual-layer attention structure F_{CSAB} in the current MDAB. The F_{CSAB} attends to both essential regions in the spatial dimension and important features in the channel dimension. To acquire the output, an identity connection is used to combine the input to the current MDAB and F_{CSAB} generated by this layer. Three deep features F^1_{MDAB}, F^2_{MDAB} and F^3_{MDAB} generated by the three MDAB are connected in the channel dimension to obtain the feature output F_{SDFE} of SDFE. The formula is as follows:

$$F_{CSAB} = F_{CA-s} * n1 + F_{SA-m} * n2 \quad (11)$$

$$F^m_{MDAB} = F^m_{CSAB} + F^{m-1}_{MDAB}, m \in [1, 3] \quad (12)$$

$$F_{SDFE} = [F^1_{MDAB}, F^2_{MDAB}, F^3_{MDAB}] \quad (13)$$

where n1 and n2 denote the randomly generated adaptive weights respectively, F^{m-1}_{MDAB} and F^m_{MDAB} denote the input and output of the m-th MDAB respectively, and in the first MDAB the F^0_{MDAB} is F_0. F_{SDFE} denotes the output of the SDFE.

FEB: To improve the generation of the reconstructed feature maps, the Pixel Attention Mechanism (PAM) is employed to focus on the global features. The feature F_{DFEB} undergoes nonlinear mapping prior to the utilization of PAM to generate a three-dimensional map of attention coefficients for the global feature pixels. Subsequently, a 3x3 convolution is applied to smooth the features, generating the reconstructed feature F_{PAM}. To retain the shallow feature information from the original image, both F_0 and F_{PAM} are aggregated, resulting in the attention-enhanced reconstructed feature F_{FEB}. The formula is as follows:

$$F_{PAM} = f_{conv3}(f_{PAM}(f_{conv1}(F_{SDFE}))) \quad (14)$$

$$F_{FEB} = F_{PAM} + F_0 \quad (15)$$

where f_{PAM} denotes the PAM, which supports 1x1 convolution, sigmoid activation function, and an identity connection. F_{PAM} denotes the smoothed pixel feature and F_{FEB} denotes the reconstructed feature.

IRB: LR images are subjected to bicubic interpolation, and are combined with the up-sampled F_{FEB} to obtain an SR image I_{SR}. The formula is as follows:

$$SR_{medi} = f_{up}(F_{FEB}) + H_{Bicubic}(I_{LR}) \quad (16)$$

where $H_{Bicubic}$ denotes the bicubic interpolation, SR_{medi} denotes the SR image of the double-thread FDN.

In the double-thread FDN network, the LR image $I_{LR} \in R^{h \times w}$ is fed into the SR model, which upscales it s times to generate the SR image $SR_{medi} \in R^{sh \times sw}$, which enhances the resolution of the image while increasing the size of the target object, thus facilitating SOD in the ensuing stage.

D. Multi-Head HyperNet

Objects with a large proportion area and high resolution are generally easier to detect compared to smaller objects, which is due to the fact that small objects tend to have low resolutions as well as weak feature representation, resulting in fewer latent features when features are being extracted. In the first stage of our architecture, we pre-proces the detected images using SR, which is beneficial as it not only increases the size of smaller objects but also preserves the finer details, thus minimizing any loss of shallow semantic information. Through model combination, we choose YOLOv5 as the base model for the second stage, but it has poor performance on SOD, and it is prone to miss or false detection. This is due to the feature pyramid structure in the algorithm is not excellent enough. Therefore, we optimize the original CSP (Cross Stage Partial) with BiFPN, which improves the quality and diversity of features, and adding a prediction head, which increases the output capability of the model and covers more target scales. The optimized model is a multi-head hyper feature learning network Multi-head HyperNet, and its specific network is shown in Fig. 2.

Bankbone: We adopt the CSPDarknet53 network to extract features of various scales and depths, which is essential for object detection. To preserve the characteristics of small objects, we preserve the shallow features when the channel number is 128, providing precise location information to prevent the significant loss of small object features that can occur with an increase in network depth. At the bottleneck of the CSPDarknet53, we introduce the Transformer encoder to capture both global and contextual information. The SR_{medi} from the previous stage is fed into the FEM to acquire four features T_1, T_2, T_3 and T_4 with different dimensional sizes. The formula is as follows:

$$T_i = H_{Bankbone}(SR_{medi}|\theta_i) \in R^{h \times w \times c}, i \in [1, 4] \quad (17)$$

where $H_{Bankbone}$ denotes the feature extraction of the CSP-Darknet53, and θ_i denotes the parameters of each layer.

Neck: To increase the receptive field size and enhance the representation capability of the network, we introduce the BiFPN structure for feature fusion in both top-down and bottom-up for the extracted features T_1, T_2, T_3 and T_4.

The top-down feature fusion pathway sequentially fuses the features T_1, T_2, T_3 and T_4. For example, to fuse T_4, we first adjust the channel number to 512 using 1x1 convolution. We then adjust the feature size using the nearest interpolation to obtain T'_4, which is the same size as T_3. We concatenate the two features and input them into learning module C3 to perform residual learning for feature fusion and expand the receptive field. The result is the output P^{td}_3. The formula is as follows:

$$P^{td}_i = C_3([f_{nearest}(P^{td}_{i+1}), T_i]), i \in [2, 3] \quad (18)$$

where $f_{nearest}$ denotes the nearest interpolation with an upsampling factor of 2, C_3 denotes the feature fusion module, and P^{td}_i

denotes the top-down intermediate temporary feature in the i-th layer.

In the bottom-up feature fusion pathway, we use skip connections to feed T_1, T_2, T_3 and T_4 into the pathway, retaining the shallow feature information. For T_2, to obtain the P_2, downsample the obtained P_1 and aggregate with T_2, and further fuse the features to obtain the feature output P_2. The relationship between each layer is shown as follows:

$$\begin{cases} P_1 = g_{conv}([T_1, f_{nearest}(P_2^{td})]), \\ P_2 = g_{conv}([f_{conv3}(P_1), T_2, P_2^{td}]), \\ P_3 = g_{conv}([f_{conv3}(P_2), T_3, P_3^{td}]), \\ P_4 = g_{conv}([f_{conv3}(P_3), T_4]), \end{cases} \tag{19}$$

where g_{conv} denotes the convolution group, P_i denotes the output of the i-th layer.

Head: The Head contains four prediction heads to detect objects of different sizes, alleviate the problem of missing detection caused by large size changes of the object, and increase the accuracy of SOD. P_i is transmitted into the Head to make predictions, and its feature information is converted into category and coordinate information. After multiple predictions of an object, a large number of candidate bounding boxes is generated. Each candidate bounding box includes not only the location information, but also the accuracy score of the candidate box, which can help the model select the best bounding box.

We filter all the candidate bounding boxes into non-maximum suppression. Assuming that there are M detection objects, N candidate bounding boxes are predicted, and the candidate bounding boxes are sorted according to the score to find the candidate bounding box $B^{\max}$ with the largest score. Calculate the IoU value between the candidate bounding boxes $B^{\max}$ and other N-1 candidate bounding boxes B^i respectively to check the overlap degree between bounding boxes. The formula is as follows:

$$IoU = \frac{B^{\max} \cap B^i}{B^{\max} \cup B^i}, i \in [1, N-1] \tag{20}$$

Set the threshold to 0.5, the IoU value is compared with the threshold value. When the IoU value is greater than the threshold, which indicates that the candidate bounding box B^i is coincided with $B^{\max}$. The score value of B^i is set to 0 and candidate bounding box of B^i is removed. Loop N-1 times until all different predictions that overlap with the candidate bounding boxes $B^{\max}$ and the IoU value is less than the threshold.

$$B^i.\text{score} = \begin{cases} 0, \text{IoU} > 0.5 \\ B^i.\text{score}, \text{IoU} <= 0.5 \end{cases} \tag{21}$$

After the above operation, the optimal bounding boxes of M objects are screened out, and the detected image SR_{ded} is obtained.

E. AFPGAN

Despite the fact that the small object image has undergone one round of SR, the restoration of facial details is insufficient. To improve visual perception, inspired by GFPGAN [33], we design an attention-based facial prior feature enhancement network AFPGAN. Fig. 2 illustrates the network, which comprises three

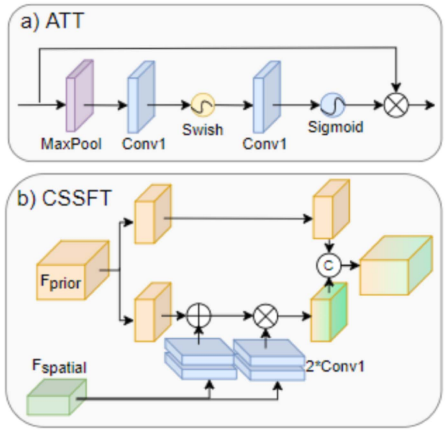


Fig. 4. The network of ATT and CSSFT.

parts,which are the Spatial Feature Generate Block (SFGB), the Priori Generate Block (PGB), and the Fine-processing Fuse Block (FFB). The SFGB compresses the input facial feature in pixel space to obtain the most important latent features and reconstructs the image space features of varying resolutions through interpolation, and then uses the attention mechanism to enhance the features. The PGB uses a pre-trained facial reconstruction model StyleGAN2 to acquire prior information such as facial geometry and texture by utilizing the generated feature weights. The FFB achieves more realistic and reasonable facial images using spatial feature modulation, with further details is as follows.

SFGB: This feature has been compressed to retain the most important facial features and remove irrelevant texture information. The SFGB utilizes a U-Net network, where the encoder use 3 down-samples to scale the feature size for an enhanced receptive field. The bottleneck of SFGM extracts potential features with 512 channels, which are compressed to preserve primary facial features while discarding inconsequential texture information. In the decoder, SFGB generates three resolution sizes of spatial features, which are weighted through the attention mechanism ATT, as shown in Fig. 4(a). The enhanced features enable to identify the positions of facial feature and the spatial relationships among them. The formula is as follows:

$$F_{latent} = B_{enc}(P_{face}) \tag{22}$$

$$F^i_{spatial} = f_{ATT}(B^i_{dec}(F_{latent}|\theta)), i \in [1, 3] \tag{23}$$

where P_{face} denotes the student face after cutting processing, B_{enc} denotes the encoder, B^i_{dec} denotes the decoder, and f_{ATT} denotes a similar SE block attention operation. F_{latent} denotes latent features, $F^i_{spatial}$ denotes i spatial features generated by the decoder, and θ denotes parameters of the layer.

PGB: Subsequently, the latent representation z is obtained as feature F_{latent} in the SFGB. However, due to its normal distribution, the z space lacks the ability to effectively represent semantic attributes. Thus, in the PGB, we map it to F_{latent} code w using the MLP structure. We then feed w to the P_rGAN to produce intermediate features F^i_{GAN} with different resolution sizes. These features incorporate numerous generative facial

priors, resulting in rich attribute, texture details, and other high-frequency feature required to restore low-quality facial images. The formula is as follows:

$$w = MLP(F_{latent}) \tag{24}$$

$$F^i_{GAN} = P_rGAN(w|\theta), i \in [1,3] \tag{25}$$

where MLP denotes the MLP structure, P_rGAN denotes the pre-trained facial reconstruction network, and F^i_{GAN} denotes the generated prior features.

We choose StyleGAN2 as the pre-trained facial GAN network, but once a more appropriate pre-trained facial GAN model is identified, it can be substituted for the current StyleGAN2.

FFB: We utilize the CSSFT, as shown as in Fig. 4(b), which is based on the Spectral Feature Transformation (SFT) to fuse the $F_{spatial}$ and F_{GAN}, resulting in high-quality facial feature map with the spatial characteristics from the input image preserved.

The F_{GAN} is partitioned into two unequal halves,which are one half remains unchanged while the other half undergoes a spatial affine transformation operation using $F_{spatial}$. Two neural networks are utilized to derive the scaling parameter matrices α^i and translation parameter matrices β^i. The two channels construct feature F_{out}, which is subsequently unified for each resolution size to reconstruct SR_{face}. The formula is as follows:

$$F^{i-0}_{GAN}, F^{i-1}_{GAN} = f_{spilt}(F^i_{GAN}) \tag{26}$$

$$\alpha^i, \beta^i = f_{conv}([F^i_{spatial}, F^{i-1}_{GAN}]) \tag{27}$$

$$F^i_{out} = [Identity(F^{i-1}_{GAN}), F^{i-1}_{GAN} \odot \alpha^i + \beta^i], i \in [1,3] \tag{28}$$

$$SR_{face} = Fuse(F^1_{out}, F^2_{out}, F^3_{out}) \tag{29}$$

where f_{spilt} denotes channel separation, F^{i-0}_{GAN} and F^{i-1}_{GAN} denote the two features obtained by dividing the prior features into channels, α^i and β^i denote the affine transformation parameter respectively, $\odot$ denotes the dot product, $Fuse$ denotes the fusion of multiple features, and SR_{face} denotes the student face image after the third stage of SR.

The above describes the specific module of SODSR. Using architecture, we achieve reliable facial detection and clear facial images of students, which can provide substantial task support. Steps are summarized as Algorithm 1.

Algorithm 1: Model Combination and SODSR.

Input:SOD sets $D(SOD)$, SR sets $D(SR)$, Evaluation metrics sets $D(E)$, Image and I_{LR}
Output:restored face sets

Stage 0: Determine combination model

for each SOD ∈ D(SOD) **do**
  for each SR ∈ D(SR) **do**
    Param(SR), NIQE, Image-S ← SR(Image)
    $AP_{l,m,s}$ ← SOD(Image-S)
  end for
  Y+R,T+R,R+R ← Compare $AP_{l,m,s}$, Param(SR) and NIQE
end for
Y+R ← Compare(Y+R,T+R,R+R) the basic combination model

Stage 1: Perform SR on images

input a real education scenario image I_{LR}
F_{SDFE} ← Feature Extraction(I_{LR})
F_{FEB} ← Attention(F_{SDFE})
SR_{medi} ← Double-Path(F_{FEB}, I_{LR})

Stage 2: Detecting faces in images

T_1, T_2, T_3, T_4 ← Bankbone(SR_{medi})
P_1, P_2, P_3, P_4 ← BiFPN(T_1, T_2, T_3, T_4)
SR_{ded} ← Head(P_1, P_2, P_3, P_4)

Stage 3: Restoring faces in images

Crop SR_{ded} and generate face sets D(face)
for each P_{face} ← D(face) **do**
  $F_{latent}, F^i_{spatial}$ ← U-Net(P_{face})
  F^i_{GAN} ← Pretraining model(F_{latent})
  SR_{face} ← CSSFT($F^i_{GAN}, F^i_{spatial}$)
end for
Return restored faces stes

IV. Experiments and Results

This section presents a comprehensive description of our experiment, which comprises four parts, namely, (1) experimental settings, containing the dataset and evaluation metrics; (2) model combination analysis; (3) presentation and the analysis of the experiment results with an emphasis on both the quantitative and qualitative performance evaluation of SOD and reconstructed facial images; and (4) ablation experiments.

A. Dataset and Evaluation Metrics

Dataset: DIV2K dataset. It contains 1000 images. Of which 800 images are used for training, 100 images for validation, and the rest for testing. This experiment uses 800 high-low resolution image pairs for training.

300W dataset: It contains two sub-datasets which are indoor and outdoor. SOD uses outdoor datasets for testing.

Set5 dataset: It contains 5 images which are used in this experiment for performance testing.

Set14 dataset: It contains 14 images which are used in this experiment for performance testing.

B100 dataset: It contains 100 images which are used in this experiment for performance testing.

Urban100 dataset: It consists of 100 images of challenging urban scenes with various frequency bands, which are utilized in this experiment for performance evaluation.

Widerface dataset: It contains 32,203 images, of which 12,881 images are utilized for training, 3,220 images for validation, and the remaining images for testing purposes. 393,703 facial annotations are marked in this dataset.

Evaluation Metrics: Various evaluation metrics, namely accuracy (AP), precision, recall and F1-score, are utilized to assess the performance of SOD in comparison to other published object detection methods. For SR, a range of reference metrics are utilized for evaluation. Full-reference evaluation metrics such as Peak Signal-to-Noise Ratio (PSNR) and Structural SIMilarity (SSIM) are employed to measure the distortion and detect the similarity between images before and after SR, respectively. Additionally, no-reference evaluation metrics such as Natural Image Quality Evaluator (NIQE), Blind/Referenceless Image Spatial QUality Evaluator (BRISQUE) and a deep learning-based image quality evaluator hyperIQA are employed to predict the image quality scores.

B. Model Combination Analysis

Although theoretically any SR algorithm can be applied for the magnification and enhancement of LR images in the first stage, the parameter quantity and performance of our architecture must be taken into account. Thus, it is necessary to filter various existing SR algorithms and pair them with benchmark SOD algorithms to assess the performance and overall parameter quantity of the architecture to obtain the most suitable combination model. For this purpose, we prefer several mainstream SR and SOD algorithms. Four mainstream SR algorithms from three categories are optimized and selected, namely, residual-based VDSR, EDSR, RFDN and DRRN [34], attention-based RCAN [35], CSNLN, HAN and NLSN, GAN-based SFT-GAN [36], ESRGAN, BSRGAN [37] and Real-ESRGAN. Furthermore, three typical SOD algorithms are optimized and combined, namely, YOLOv5 with CSPNet backbone, TinaFace [38] with ResNet backbone, and RetinaFace [39] with MobileNet backbone.

The purpose is to repair low resolution images and enlarge the small target size within them. We evaluate the performance of 42 SR+SOD tasks by transferring typical algorithms to each model combination. Before the model is combined, the WiderFace dataset is subjected to pre-processing operations, wherein a down-sampling factor of 2 is used to obtain small object images of size 512×384. For the task of model combination, small objects of 512× 384in size undergo SR to improve image quality for SOD. In the first stage of SR, the purpose is to enhance LR images and enlarge the size of small objects. NIQE and hyperIQA are employed to measure the image quality after SR. With regard to the SR-SOD model combination, we optimize three SOD algorithms with different backbone networks and combine them with the SR algorithms using AP values for accuracy evaluation to measure the performance of the model combination.

We perform an ablation analysis on each combination mode to evaluate the efficacy of various model combination configurations. Apart from the standard metrics designated for each stage, we employ the overall parameter quantity, which is the total of the parameters in both the SR and SOD algorithms as a comprehensive evaluation metric.

We construct two benchmark datasets for each type of SOD. Table I lists the evaluation metrics that compare the image quality and accuracy of SOD with and without the assistance of SR. For instance, L1 displays the results without using SR in YOLOv5, and L2 indicates the results after employing SRCNN. L15–L16 and L29–L30 present the benchmark data of the model combination for TinaFace and RetinaFace, respectively. It is evident from the benchmark data that the use of SRCNN in SOD moderately improves the AP value, which demonstrates the efficiency of SR pre-processing for improving SOD performance.

We perform ablation experiments on the SR algorithms for each type of SOD to determine the optimal combination model. There are three sets of SR algorithms in YOLOv5, of which EDSR (L4) and RFDN (L5) demonstrate superior performance in the residual-based SR with AP values for the easy, medium and hard subsets of 94.33, 92.21, 80.98 (L4) and 94.37, 92.30, 80.96 (L5). Compared to the benchmark data (L1), we note a considerable increase in the AP values for the easy, medium and hard subsets, especially the hard subset, which has the most significant improvement by 11.46% and 11.44% in the hard subsets. In the attention-based SR, HAN (L9) and NLSN (L10) achieve the highest AP values for the easy, medium and hard subsets in SOD with 94.25, 92.17, 80.42 (L9) and 94.23, 92.15, 80.47 (L10). In the GAN-based SR, BSRGAN (L12) and Real-ESRGAN (L14) achieve the highest AP values for the easy, medium and hard subsets in SOD of 94.15, 91.97, 80.09 (L12) and 93.15, 91.50, 79.53 (L14). Taking into account the performance of SOD and the parameter quantity of SR, we compare the top two performing combination models of the three sets of algorithms and optimize that RFDN is the best SR for YOLOv5, which is the combination of YOLOv5+RFDN (L5). We adopt the same way to TinaFace and RetinaFace by selecting TinaFace+RFDN (L19) and RetinaFace+RFDN (L33) as the optimal combination, respectively.

Finally, we select the optimal model combination from the top three contenders. The evaluation metrics comprise the total of the pre-training parameters for the combination model and the accuracy of SOD. After comparing the total of the pre-training parameters of YOLOv5+RFDN (L5), TinaFace+RFDN (L19) and RetinaFace+RFDN (L33), we observe that although RetinaFace+RFDN (L33) requires the least pre-training parameters (0.96M), its AP values for the easy, medium and hard subsets of SOD is the lowest at 90.45, 90.85 and 78.22. YOLOv5+RFDN (L5) requires the second least pre-training parameters (7.596M), with AP values for the easy, medium and hard subsets of 94.37, 92.30 and 80.96. TinaFace+RFDN (L19) has the highest AP values for the easy, medium and hard subsets of 96.47, 95.46 and 90.77, but it requires the most pre-training parameters (38.501M). Considering the AP value and parameter quantity comprehensively, YOLOv5+RFDN (L5) is optimized to choose

TABLE I
MODEL COMBINATION EXPERIMENT

Pre-training SOD	Model Combination	NIQE	hyperIQA	$AP_l^{IoU=0.5}$	$AP_m^{IoU=0.5}$	$AP_s^{IoU=0.5}$	Params(SR)	Params(SBD)	Total Params
YOLOv5	1 None+YOLOv5	6.81	38.21	93.71	90.60	69.52		7.075M	7.075M
	2 SRCNN+YOLOv5	6.25	37.52	93.72	90.54	69.92	8k	7.075M	7.082M
	3 (ResNet)VDSR+YOLOv5	4.51	45.93	94.28	92.19	79.63	694k	7.075M	7.752M
	4 (ResNet)EDSR+YOLOv5	4.49	49.35	94.33	92.31	80.98	4.5M	7.075M	11.575M
	5 (ResNet)RFDN+YOLOv5	4.30	47.79	94.37	92.30	80.96	534k	7.075M	7.596M
	6 (ResNet)DRRN+YOLOv5	4.32	49.49	94.19	92.12	80.38	307k	7.075M	7.374M
	7 (Attention)RCAN+YOLOv5	5.73	48.37	93.99	92.13	82.41	15M	7.075M	22.075M
	8 (Attention)CSNLN+YOLOv5	4.32	49.55	94.22	92.15	80.45	3M	7.075M	10.075M
	9 (Attention)HAN+YOLOv5	4.30	49.08	94.25	92.17	80.42	15.92M	7.075M	22.995M
	10 (Attention)NLSN+YOLOv5	4.32	47.00	94.23	92.15	80.47	41.80M	7.075M	48.875M
	11 (GAN)SFTGAN+YOLOv5	2.47	62.80	93.68	90.46	73.83	1.61M	7.075M	8.685M
	12 (GAN)BSRGAN+YOLOv5	3.81	72.73	94.15	91.97	80.09	16.6M	7.075M	23.675M
	13 (GAN)ESRGAN+YOLOv5	4.24	30.10	91.66	87.11	59.76	16.7M	7.075M	23.775M
	14 (GAN)Real-ESRGAN+YOLOv5	4.01	75.68	93.85	91.50	79.53	16.7M	7.075M	23.775M
TinaFace	15 None+TinaFace	6.81	38.21	95.68	93.39	78.24		37.98M	37.98M
	16 SRCNN+TinaFace	6.25	37.52	95.73	93.43	78.25	8k	37.98M	37.987M
	17 (ResNet)VDSR+TinaFace	4.51	45.93	96.55	95.44	89.63	694k	37.98M	38.657M
	18 (ResNet)EDSR+TinaFace	4.49	49.35	96.53	95.52	90.96	4.5M	37.98M	42.48M
	19 (ResNet)RFDN+TinaFace	4.30	47.79	96.47	95.46	90.77	534k	37.98M	38.501M
	20 (ResNet)DRRN+TinaFace	4.32	49.49	96.51	95.46	90.45	307k	37.98M	38.279M
	21 (Attention)RCAN+TinaFace	5.73	48.37	95.78	94.42	87.17	15M	37.98M	52.98M
	22 (Attention)CSNLN+TinaFace	4.32	49.55	96.50	95.45	90.51	3M	37.98M	40.98M
	23 (Attention)HAN+TinaFace	4.30	49.08	96.53	95.45	90.53	15.92M	37.98M	53.90M
	24 (Attention)NLSN+TinaFace	4.32	47.00	96.51	95.47	90.54	41.80M	37.98M	79.78M
	25 (GAN)SFTGAN+TinaFace	2.47	62.80	95.05	92.68	79.90	1.61M	37.98M	39.59M
	26 (GAN)BSRGAN+TinaFace	3.81	72.73	95.66	94.41	88.96	16.6M	37.98M	54.58M
	27 (GAN)ESRGAN+TinaFace	4.24	30.10	94.00	90.57	68.94	16.7M	37.98M	54.68M
	28 (GAN)Real-ESRGAN+TinaFace	4.01	75.68	94.96	93.43	88.00	16.7M	37.98M	54.68M
RetinaFace	29 None+RetinaFace	6.81	38.21	89.71	85.62	62.83		0.44M	0.44M
	30 SRCNN+RetinaFace	6.25	37.52	89.79	85.75	63.22	8k	0.44M	0.44M
	31 (ResNet)VDSR+RetinaFace	4.51	45.93	90.80	88.14	70.98	694k	0.44M	1.11M
	32 (ResNet)EDSR+RetinaFace	4.49	49.35	90.85	88.16	72.22	4.5M	0.44M	4.94M
	33 (ResNet)RFDN+RetinaFace	4.30	47.79	90.45	90.85	78.22	534k	0.44M	0.96M
	34 (ResNet)DRRN+RetinaFace	4.32	49.49	90.85	88.01	71.63	307k	0.44M	0.73M
	35 (Attention)RCAN+RetinaFace	5.73	48.37	87.17	90.43	87.59	15M	0.44M	15.44M
	36 (Attention)CSNLN+RetinaFace	4.32	49.55	90.51	90.86	88.05	3M	0.44M	3.44M
	37 (Attention)HAN+RetinaFace	4.30	49.08	90.53	90.91	88.11	15.92M	0.44M	16.34M
	38 (Attention)NLSN+RetinaFace	4.32	47.00	90.54	90.89	88.10	41.80M	0.44M	42.24M
	39 (GAN)SFTGAN+RetinaFace	2.47	62.80	79.90	89.93	86.06	1.61M	0.44M	2.05M
	40 (GAN)BSRGAN+RetinaFace	3.81	72.73	88.96	90.81	88.19	16.6M	0.44M	17.04M
	41 (GAN)ESRGAN+RetinaFace	4.24	30.10	68.94	86.89	81.89	16.7M	0.44M	17.14M
	42 (GAN)Real-ESRGAN+RetinaFace	4.01	75.68	88.00	90.60	87.66	16.7M	0.44M	17.14M

*The optimal results for all data in the indicator are highlighted in red.

as the most suitable combination model. Therefore, our proposed architecture is based on the model combination YOLOv5 and RFDN.

C. Experiment Results and Analysis

Analysis of small object detection results: In this part, we analyze the experimental results of SOD. We compare our algorithm with mainstream first-order and second-order methods on subsets with various levels of accuracy, as shown in Table II, and all tests are conducted on the Widerface dataset. Table II shows that our algorithm attains the highest detection accuracy on the easy, medium and hard subsets, with an accuracy of 94.3%, 93.5% and 88.6%. Although the accuracy of YOLOv5s on the easy subset is close to ours, the accuracy of our algorithm is 0.9% and 5.5% higher than YOLOv5s on the medium and hard subsets. In addition, compared with the classical SOD algorithm Tiny Face, our algorithm achieves AP values which are increased by 1.8%, 2.5%, and 8.0% on the easy, medium and hard subsets respectively. The comparison between our algorithm with other

TABLE II
TESTING ON THE WIDERFACE DATASET

Method	Easy	Medium	Hard
ACF	65.9	54.1	27.3
Faceness	71.3	53.4	34.5
Two Stage CNN	68.1	61.4	32.3
Tiny Face	92.5	91.0	80.6
SSH	93.1	92.1	84.5
Blaze Face	88.5	85.5	73.1
yolov5-blazeface	90.4	88.7	78.0
Yolov5s	**94.3**	92.6	83.1
Ours	**94.3**	**93.5**	**88.6**

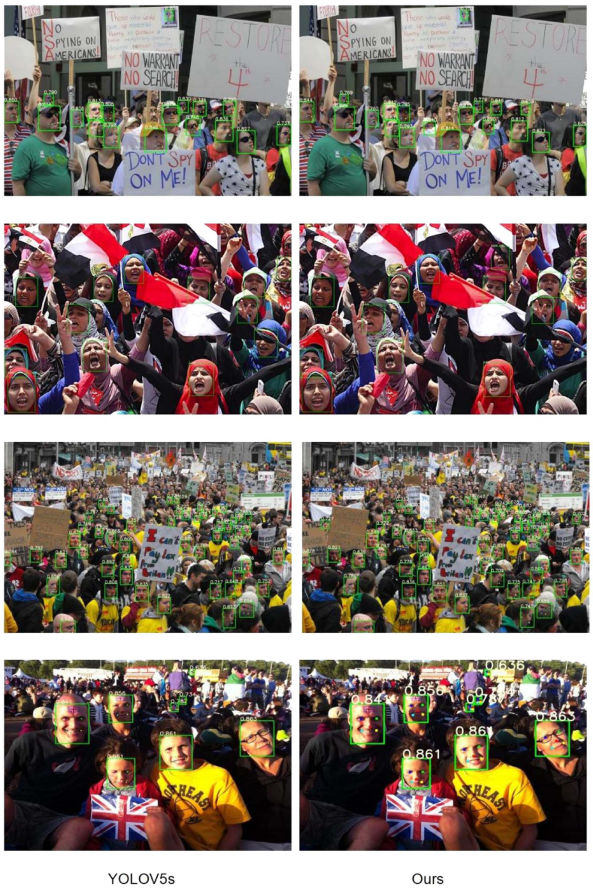


Fig. 5. Testing on the 300 W dataset.

algorithms shows that our solution considerably improves the detection accuracy in all subsets, particularly with the most significant improvement on the hard subset which is a still a challenging direction for SOD tasks. In addition, in practical applications, it is beneficial for detecting small facial features in educational scenes.

In addition, to visually demonstrate the validity of our algorithm, qualitative analysis is performed using an outdoor sub-dataset in 300W dataset. Yolov5s and our algorithm are used to test the 300W dataset, and the test results are shown in Fig. 5. In Fig. 5, the left side images are the visual result of detection by Yolov5s, and the right side images are the visual result of detection by our algorithm. The number of detection results from top to bottom on the left side is 18,13,49 and 8 respectively, while the number of detection results from top to bottom on the right side is 19,14,56 and 8 respectively. The results show that our algorithm performs better in detecting SOD and can detect more small objects.

To verify the practicality of our proposed method in educational scenes, we test it on images captured in real-world educational scenes, so we choose graduation images and teaching images. Fig. 6 illustrates the visual output of our method, with the left and right columns showing the original images and the detected results, respectively. The number of detected faces in the images are 47 and 29, respectively. The aforementioned results indicate that our method can accurately detect faces of

Fig. 6. The testing results in the educational scenes.

varying sizes. Nevertheless, it is essential to note that if a part of a face is obscured or angled away from the camera, the detection of our facial detection algorithm may be incomplete.

Results of restoring small facial objects: Subsequent tasks, such as expression recognition, may cause certain obstacles due to small faces in the educational scenes, which usually have lower resolutions. To mitigate the harm caused by these situations, we reconstruct the detected faces of the students, which helps to restore image details and improve the resolution of the images.

We select images of graduation scenes, crop the face images of the students and conduct SR restoration. To better compare the before and after SR reconstruction, we concentrate on displaying the cropped face images from the original images in Fig. 7(a) and the results of the SR reconstruction in Fig. 7(b). A visual comparison shows that the resolution of a restored image is higher than the original image and the restoration of its details and textures is excellent. We plan to objectively validate the remarkable restoration results of our proposed method by employing different SR methods to restore and compare the cropped facial images. Moreover, we use no-reference evaluation metrics to quantitatively assess the effectiveness of the results.

To demonstrate the superiority of our approach in facial restoration, multiple SR algorithms, such as Bicubic, SRCNN, ESRGAN, Real-ESRGAN, BSRGAN and PFSR [40], are employed to reconstruct the faces of students. As shown in Fig. 7(c), our proposed method successfully restores facial details, such as teeth and eye, whit its powerful facial prior, while also achieving high fidelity restoration for lighting and hair. Another GAN-based algorithm, such as BSRGAN, has unsatisfactory results. Although it recognizes facial features, it does not restore facial details and produces a large number of textures which affects the authenticity of small target faces. Although PFSR increases the resolution of facial images, the generated facial textures are unrealistic and exhibits minor differences compared to other SR algorithms but the effect on restoring facial details is not very significant.

A no-reference quality evaluation is conducted using NIQE, BRISQUE and HyperIQA as shown in Fig. 7(c), and the evaluation results are shown in Table III. NIQE measures the deviation between the measured image and the natural images in multiple

(a) croped face images before restored

(b) croped face images after restored

(c) Comparison of differnt methods

Fig. 7. Cropped faces before (a) and after restoration (b), as well as the visualization comparison (c).

TABLE III
NR-IQA OF IMAGES

Methods	NIQE↓			BRISQUE↓			hyperIQA↑		
	Pic1	Pic2	Pic3	Pic1	Pic2	Pic3	Pic1	Pic2	Pic3
Bicubic	14.87	14.64	14.95	74.59	76.38	76.68	30.4	29.71	30.25
SRCNN	12.91	12.11	14.21	71.79	72.49	73.99	28.36	30.12	29.62
ESRGAN	12.39	14.1	14.43	38.6	54.05	48.24	29.11	29.46	31.53
Real-ESRGAN	14.44	12.36	11.35	54.62	48.7	43.55	29.33	28.17	25.11
BSRGAN	7.94	7.7	6.86	21.07	27.38	25.18	37.7	40.65	27.52
PFSR	5.75	5.35	5.45	28.47	9.02	16.14	27.97	29.09	32.49
ours	**4.01**	**3.87**	**5.12**	**5.52**	**1.57**	**3.53**	**79.63**	**81.03**	**79.04**

TABLE IV
THE EXPERIMENT IN NETWORK REPLACEMENT

Model Combination	NIQE	hyperIQA	$AP_l^{IoU=0.5}$	$AP_m^{IoU=0.5}$	$AP_s^{IoU=0.5}$
RFDN+Yolov5s	4.30	47.79	94.37	92.30	80.96
RFDN+Yolov8n	4.30	47.79	94.63	92.16	76.25
CAMixerSR+Yolov5s	4.07	55.11	94.77	92.82	81.92
CAMixerSR+Yolov8n	4.07	55.11	94.84	92.35	76.51

dimensions, where the larger the difference, the greater the difference between the measured image and the natural image. BRISQUE predicts image evaluation scores by comparing the pixel intensity of a distorted image against natural images. The lower the BRISQUE score, the better the quality. HyperIQA evaluates the local regions of the image for both local and global human visual perceptions. A higher HyperIQA value indicates better quality. The results show that our proposed method performs significantly better in the three metrics, delivering images that closely match human visual perception and better aesthetics effect compared to other SR algorithms. However, we observe that some SR algorithms do not improve the restoration quality of an image compared with Bicubic, as shown in Fig. 7(c). We speculate that this it is due to the noise present in low-quality images, which is not eliminated during the SR process, resulting in excessive artifacts and ringing appearing in the HR image.

The effectiveness of three-stage architecture replacement: In this section, we analyze the replaceable of the three-stage architecture. The architecture we currently use, especially the first two stages, are selected according to the 3D capability evaluation matrix. However, it should be noted that the selected RFDN and Yolov5s are not the latest models. Therefore, in order to verify the feasibility of architecture replacement and whether the latest models can replace the stage models, We choose the latest lightweight SR networks and SOD networks to replace stage models and measure their performance metrics. In the SR network, we choose the lightweight SR network CAMixerSR [41], and in the SOD network, we choose Yolov8-n [42]. The experimental results are shown in the Table IV.

The Table IV contains four groups of combined models, the first group of which represents the measuring model and its evaluation value. As it can be observed from the Table IV, the performance metrics of CAMixerSR in the first stage is better than that of the benchmark model, the performance metrics of its non-reference image is better than that of the basic SR model, and the performance metrics of SOD is also better than that of the benchmark SOD model, which indicates the effectiveness of CAMixerSR, which can replace the measuring first-stage model. However, in the second stage model, although the performance metrics of the latest Yolov8n is higher than that of Yolov5s

TABLE V
ABLATION EXPERIMENTS OF COMPONENTS

MDAB	PAM	Bicubic	Set5	Set14	BSD100	Urban100
✓	×	×	34.67/0.9317	28.91/0.8203	31.21/0.9057	27.49/0.8642
✓	✓	×	34.76/0.9324	28.99/0.8207	31.26/0.9060	27.55/0.8645
✓	×	✓	34.68/0.9319	28.94/0.8202	31.26/0.9059	27.58/0.8641
✓	✓	✓	**34.85/0.9329**	28.98/**0.8213**	**31.31/0.9064**	**27.64/0.8655**

TABLE VI
ABLATION EXPERIMENTS OF CA-S AND SA-M

CA-s	SA-m	Set5	Set14	BSD100	Urban100
Basic Block		32.18/0.8952	28.54/0.7819	27.54/0.7360	25.52/0.7858
✓	×	34.65/0.9312	28.94/0.8196	31.28/0.9058	27.57/0.8636
×	✓	34.73/0.9323	28.99/0.8209	31.26/0.9063	27.58/0.8651
✓	✓	**34.85/0.9329**	28.98/**0.8213**	**31.31/0.9064**	**27.64/0.8655**

in the facial detection metrics of large size, while it is lower than that of Yolov5s in the facial detection of SOD. In view of this phenomenon, we suspect that this is because the research of Yolov8 does not specifically target SOD. In addition, due to the number of parameters of Yolov8n is lower than that of Yolov5s, for comparative fairness, we should choose Yolov8s as our benchmark model. However, the relevant codes and models of Yolov8s have not been released at present. Therefore, it is not possible to upgrade the performance metrics of Yolov8n to Yolov8s.

Furthermore, qualitative and quantitative experiments demonstrate that Super-Resolution (SR) technology consistently enhances object detection performance. In our initial model combination analysis, we compared SOD performance with and without SR using the results presented in Table I. The findings consistently show that integrating SR improves SOD performance compared to baseline results without SR. Furthermore, Table IV highlights that modern SR techniques exhibit stronger enhancements for SOD compared to classical methods. These results across different combinations underscore the efficacy of SR in enhancing SOD. However, we observed varying contributions from different SR technologies. Hence, practical implementation of SR to enhance SOD requires careful selection and optimization to maximize effectiveness.

D. Ablation Experiments

We perform comprehensive ablation experiments to investigate the functionality of each module. We apply identical parameter settings in all of the experiments, except for the specified modifications made to certain module, which are detailed in each subsequent section.

The effectiveness of MDAB: To assess the effectiveness of the double-thread FDN, we conduct two ablation experiments. The first experiment evaluates the effectiveness of each module of the model, whereas the second experiment validates the efficacy of the attention mechanism within the core module of the MDAB.

We implement ablation experiments to assess the effectiveness of each module in our model. Specifically, we need to verity the efficacy of the MDAB, the PAM and the Bicubic Up-sampling. To evaluate the effectiveness of MDAB, we set it as our benchmark model, remove PAM and Bicubic, and train it for 1000 epochs with the same number of parameters. Then we test it on Set5, Set14, B100, and Urban100 datasets for upscaling 4, the value of PSNR and SSIM metrics are generated, as shown in Table V. We validate the effectiveness of PAM by adding it to the benchmark model and observing the improvements in the PSNR and SSIM values. A similar validation is performed for Bicubic, which shows less significant improvement than PAM. Finally, combining PAM and Bicubic into the benchmark model leads to a substantial improvement of PSNR, increasing by 0.18, 0.07, 0.1 and 0.15in each dataset. Our proposed method demonstrates that both PAM and Bicubic improve the effect of SR.

The effectiveness of Attecntion: To verify the effectiveness of CA-s and SA-m in MDAB. In this experiment, we employ CA-s and SA-m for training without changing other MDAB structures, using the same training parameters and datasets as the previous ablation experiment. The test data obtained using both CA-s and SA-m are taken as the benchmark data, and test data using CA-s or SA-m alone are taken as the comparison data, as shown in Table VI. The results indicate that the baseline data is significantly higher than the other two sets of data, which demonstrates the effectiveness of using both CA-s and SA-m. In addition, we add an ablation experiment to compare the original attention layer in the RFDN with the attention structure we designed. The experimental results are shown in line 2 of Table VI. The experimental results prove the effectiveness of the attention structure we designed.

The effectiveness of SOD: In this part, ablation experiments are carried out on the structure of the SOD network. The structures involved in the experiments include the number of predicted heads and the structure of neck. In addition, the enhanced effect of super resolution on SOD is validated. Since we use the same evaluation metrics, the results of both the ablation experiments are presented in the same table.

Firstly, we verity the effectiveness of multiple prediction heads using the HyperNet with three prediction heads as our benchmark algorithm. We assess the performance of the model using a diverse set of metrics, which contain mean average precision (mAP), F1-score, precision and recall, in addition to testing the AP values of the easy, medium and hard subsets under varying IoU thresholds, as shown in column 2 of Table VII. Subsequently, we employ multi-head HyperNet with identical parameters and the results of the verification are shown in column 5 of Table VII. A comparison between these two column data reveals that our proposed method outperforms the benchmark model in most cases. Notably, ours improves by 1.59% in mAP @.5:.95 and 16.9% in precision when compared to the benchmark data. However, as precision and recall are interrelated, we also evaluate our model in terms of F1-score, which exhibits a 3.74% improvement over the benchmark model. The AP values with different intersection over union (IoU) are also vastly improved, especially on the hard subset, which shows a significantly steeper growth rate than that of the easy and medium subsets. Under the IoU of 0.5, 0.6, 0.75 and 0.9, the AP increases by a considerable margin of 1.89%, 2.73%, 3.06%, and 0.65%, respectively. Yet,

TABLE VII
ABLATION EXPERIMENTS OF SR AND SOD

Metric/Methods	With three heads			With four heads		
	Original	SR	FPN	Original	SR	FPN
mAP@.5	71.81	74.85	71.93	72.10	73.72	73.51
mAP@.5:.95	71.18	74.77	71.58	72.77	75.27	73.04
Precision	73.90	76.40	68.10	90.80	98.90	84.52
Recall	58.59	60.34	52.62	55.91	64.48	54.46
F1-score	65.36	67.43	59.36	69.20	81.00	66.24
$AP_l^{IoU=0.5}$	93.99	94.44	93.56	94.32	94.37	93.29
$AP_m^{IoU=0.5}$	92.14	93.54	90.74	92.21	93.53	89.65
$AP_s^{IoU=0.5}$	82.42	88.84	75.01	84.31	88.57	68.27
$AP_l^{IoU=0.6}$	91.52	91.99	90.71	91.59	91.72	91.06
$AP_m^{IoU=0.6}$	88.72	90.56	86.96	88.69	90.59	86.44
$AP_s^{IoU=0.6}$	73.16	82.51	65.43	75.89	82.28	61.13
$AP_l^{IoU=0.75}$	75.27	75.51	72.95	75.37	73.23	74.38
$AP_m^{IoU=0.75}$	69.63	73.15	66.86	70.08	72.75	69.68
$AP_s^{IoU=0.75}$	44.96	55.32	40.03	48.02	55.24	39.94
$AP_l^{IoU=0.9}$	11.40	10.16	8.55	11.58	8.91	9.31
$AP_m^{IoU=0.9}$	8.52	9.14	6.51	9.19	9.06	7.26
$AP_s^{IoU=0.9}$	4.21	5.39	3.08	4.86	5.63	3.53

it is worth noting that under the same IoU, the easy and medium subsets achieve only a marginal improvement of -0.33%, 0.07%, 0.1%, 0.18%, and 0.07%, -0.03%, 0.45%, 0.67%, respectively. In summary, the above experiments successfully demonstrate the effectiveness of multiple prediction heads.

Secondly, the validity of different neck structures for the SOD is tested. The FPN is used in the Yolov5, and in this paper we adopt the BIFPN. Using our measuring model (column 6) as the benchmark algorithm, BIFPN in the benchmark algorithm was replaced by FPN, and the experiment is carried out with different number of prediction heads. The experimental results are shown in columns 4 and 7 of Table VII. Through the comparison of column 6 and column 7 of Table VII, it is find that the performance detection using BIFPN is better than that using FPN.

Finally, we verify the effectiveness of SR for SOD. We feed the SR image to the SOD algorithm and the comparative metrics data before and after SR for the benchmark algorithm and the multi-head HyperNet algorithm are presented in column 3 and 6 of Table VII, respectively. A comparison of the data reveals an increase in the performance metrics for testing the image after SR, suggesting that SR can improve the accuracy of SOD.

The effectiveness of ATT: To verify the effectiveness of the attention structure ATT in the AFPGAN, we perform two ablation experiments, which one based on the CSSFT control experiment and another based on the SFT control experiment. In the CSSFT control experiment, we keep the other network unchanged, and only add or delete the ATT. We use the model that without the ATT as the benchmark model, while the test data from this model serves as the baseline data. The model with the ATT is used as the control model, and its test data is used as the control data. The testing process incorporates PSNR, NIQE, BRISQUE and HyperIQA. The same approach is taken for the control experiment based on SFT, and the experimental results are shown in Table VIII. By comparison, it can be seen that adding the ATT structure to both the CSSFT and SFT control

TABLE VIII
ABLATION EXPERIMENTS OF ATT

Methods	PSNR↑	NIQE↓	BRISQUE↓	hyperIQA↑
SFT	24.66	4.45	10.76	**78.02**
SFT+ATT	**24.68**	**4.22**	**8.03**	76.40
CSSFT	24.90	4.52	4.11	**78.91**
CSSFT+ATT	**24.99**	**4.33**	**1.97**	78.33

experiments improves their overall performance. In the CSSFT control experiment, the ATT structure increases the PSNR value by 0.09, at the same time, the NIQE and BRISQUE values decrease by 0.19 and 2.14, respectively. Similarly, in the SFT control experiment, the ATT structure increases PSNR while reducing the NIQE and BRISQUE values by 0.23 and 2.72, respectively. Consequently, our objective metric comparison supports the efficacy of the attention structure ATT and its benefits for image quality restoration.

V. CONCLUSION

In response to the problems of poor detection performance, the false and missing detection of small student faces in educational scenes, which is due to a different scales facial imaging area, low resolution, and insufficient feature, we propose a three-stage architecture SODSR, which effectively detects student faces with different scales, restores facial textures and enhances its detailed features. In the first stage, we construct the double-thread FDN (SR) to pre-process the LR images, and introduce the attention mechanism to enhance features while designing a dual-path network to increase the utilization rate of shallow information, so as to improve the image resolution. In the second stage, the multi-head HyperNet (SOD) is used to detect small student faces caused by being far away from the camera. Four prediction heads are used to alleviate the impacts caused by the changes of multi-dimensional features, which increase the robustness of the model and improve detection accuracy. In the third stage, we construct the AFPGAN (SR) based on GAN Inversion to restore the texture of small student faces, which employ the StyleGAN2 network to extract prior features and modulate them with the enhanced spatial features to fuse features. Extensive experiments on datasets prove that our proposed SODSR has a high object detection capability and face restoration capability, and achieves promising performance of the method is effective in recognizing the small faces of students in educational scenes.

For future research directions, there are several avenues for optimizing our proposed method. Firstly, efforts will focus on reducing missed and false detection rates, particularly for non-frontal and occluded faces. Secondly, we aim to develop a more lightweight network architecture for SODSR, aiming to streamline network parameters and enhance detection speed. Additionally, integrating visualization tasks such as facial expression recognition into this framework will broaden its applicability. Furthermore, the growing interest in domain adaptation within object detection, as evidenced in recent studies [43], [44], could motivate exploring its integration into educational SOD scenarios. This exploration will investigate the potential of using

SR images as a source domain to improve SOD performance in educational contexts.

REFERENCES

[1] K. Yi, S. Yan, L. Liu, J. Zhu, W. Liang, and J. Xu, "CCSS: An effective object detection system for classroom crowd statistics," in *Proc. IEEE Annu. Comput., Softw., Appl. Conf.*, 2022, pp. 111–116.

[2] M. Liu, X. Zhang, and Y. Han, "Intelligent counting system for classroom numbers based on video surveillance," in *Proc. IEEE Int. Conf. Power, Intell. Comput. Syst.*, 2020, pp. 242–245.

[3] Z. Zhang, P. Lin, S. Ma, and T. Xu, "An improved YOLOv5s algorithm for emotion detection," in *Proc. Int. Conf. Pattern Recognit. Artif. Intell.*, 2022, pp. 1002–1006.

[4] Z. Li, J. Xiong, and H. Chen, "Based on improved YOLO_v3 for college students' classroom behavior recognition," in *Proc. Int. Conf. Artif. Intell. Comput. Inf. Technol.*, 2022, pp. 1–4.

[5] V. Rothoft, J. Si, F. Jiang, and R. Shen, "Monitor pupils' attention by image super-resolution and anomaly detection," in *Proc. Int. Conf. Comput. Syst., Electron. Control*, 2017, pp. 843–847.

[6] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.

[7] C. Chen, M.-Y. Liu, O. Tuzel, and J. Xiao, "R-CNN for small object detection," in *Proc. Asian Conf. Comput. Vis.*, 2017, pp. 214–230.

[8] M. Kisantal, Z. Wojna, J. Murawski, J. Naruniec, and K. Cho, "Augmentation for small object detection," 2019, *arXiv:1902.07296*.

[9] B. Zoph, E. D. Cubuk, G. Ghiasi, T.-Y. Lin, J. Shlens, and Q. V. Le, "Learning data augmentation strategies for object detection," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 566–583.

[10] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 936–944.

[11] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[12] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.

[13] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2999–3007.

[14] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8759–8768.

[15] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: Scalable and efficient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10778–10787.

[16] S. Liu, D. Huang, and Y. Wang, "Learning spatial fusion for single-shot object detection," 2019, *arXiv:1911.09516*.

[17] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 2016.

[18] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1646–1654.

[19] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit Workshops*, 2017, pp. 1132–1140.

[20] J. Kim, J. K. Lee, and K. M. Lee, "Deeply-recursive convolutional network for image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1637–1645.

[21] Z. Hui, X. Gao, Y. Yang, and X. Wang, "Lightweight image super-resolution with information multi-distillation network," in *Proc. ACM Int. Conf. Multimedia*, 2019, pp. 2024–2032.

[22] J. Liu, J. Tang, and G. Wu, "Residual feature distillation network for lightweight image super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 41–55.

[23] Y. Mei, Y. Fan, and Y. Zhou, "Image super-resolution with non-local sparse attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 3517–3526.

[24] Y. Mei, Y. Fan, Y. Zhou, L. Huang, T. S. Huang, and H. Shi, "Image super-resolution with cross-scale non-local attention and exhaustive self-exemplars mining," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 5689–5698.

[25] T. Dai, J. Cai, Y. Zhang, S.-T. Xia, and L. Zhang, "Second-order attention network for single image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 11057–11066.

[26] B. Niu et al., "Single image super-resolution via a holistic attention network," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 191–207.

[27] C. Ledig et al., "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 105–114.

[28] N. C. Rakotonirina and A. Rasoanaivo, "ESRGAN+: Further improving enhanced super-resolution generative adversarial network," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2020, pp. 3637–3641.

[29] X. Wang, L. Xie, C. Dong, and Y. Shan, "Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 1905–1914.

[30] L. Cao, R. Ji, C. Wang, and J. Li, "Towards domain adaptive vehicle detection in satellite image by supervised super-resolution transfer," in *Proc. AAAI Conf. Artif. Intell.*, 2016, pp. 1138–1144.

[31] Y. Wang et al., "Remote sensing image super-resolution and object detection: Benchmark and State of the Art," *Expert Syst. Appl.*, vol. 197, 2022, Art. no. 116793.

[32] Y. Pang, J. Cao, J. Wang, and J. Han, "JCS-Net: Joint classification and super-resolution network for small-scale pedestrian detection in surveillance images," *IEEE Trans. Inf. Forensics Secur.*, vol. 14, no. 12, pp. 3322–3331, Dec. 2019.

[33] X. Wang, Y. Li, H. Zhang, and Y. Shan, "Towards real-world blind face restoration with generative facial prior," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 9164–9174.

[34] Y. Tai, J. Yang, and X. Liu, "Image super-resolution via deep recursive residual network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2790–2798.

[35] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 294–310.

[36] X. Wang, K. Yu, C. Dong, and C. C. Loy, "Recovering realistic texture in image super-resolution by deep spatial feature transform," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 606–615.

[37] D. Akiyama and T. Goto, "Improving image quality using noise removal based on learning method for surveillance camera images," in *Proc. IEEE Glob. Conf. Life Sci. Technol.*, 2022, pp. 325–326.

[38] Y. Zhu, H. Cai, S. Zhang, C. Wang, and Y. Xiong, "TinaFace: Strong but simple baseline for face detection," 2020, *arXiv:2011.13183*.

[39] J. Deng, J. Guo, Y. Zhou, J. Yu, I. Kotsia, and S. Zafeiriou, "RetinaFace: Single-stage dense face localisation in the wild," 2019, *arXiv:1905.00641*.

[40] D. Kim, M. Kim, G. Kwon, and D.-S. Kim, "Progressive face super-resolution via attention to facial landmark," in *Proc. Brit. Mach. Vis. Conf.*, 2019, pp. 161.1–161.12.

[41] Y. Wang, S. Zhao, Y. Liu, J. Li, and L. Zhang, "CAMixerSR: Only details need more "attention"," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 25837–25846.

[42] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics YOLO," Jan. 2023. [Online]. Available: https://github.com/ultralytics/ultralytics

[43] J. Pang, W. Liu, B. Zhang, X. Yang, B. Liu, and D. Tao, "MCNet: Magnitude consistency network for domain adaptive object detection under inclement environments," *Pattern Recognit.*, vol. 145, 2024, Art. no. 109947.

[44] W. Li, X. Liu, and Y. Yuan, "SIGMA++: Improved semantic-complete graph matching for domain adaptive object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 7, pp. 9022–9040, Jul. 2023.

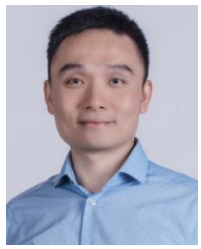
Xiaoyong Mei (Member, IEEE) received the Ph.D. degrees in computer software and theory from Sun Yat-sen University, Guangzhou, China, in 2011. He completed the Postdoctoral Fellowship position with the School of Information Technology in Education, South China Normal University, Guangzhou, China, in 2017. He is currently a Shuang Long Scholar Distinguished Professor with the Key Laboratory of Intelligent Education Technology and Application of Zhejiang Province, Zhejiang Normal University, Jinhua, China. He has authored or coauthored more than 20 scientific papers in international journals and conferences. His research interests include image processing and analysis, video analysis and dynamic scene processing, information extraction, and visual applications and systems.

Kejin Zhang received the bachelor's degree in intelligent educational technology from Jiangxi Agricultural University, Nanchang, China, in 2021. She is currently working toward the master's degree in educational technology with Zhejiang Normal University, Jinhua, China. Her research interests include super resolution and object detection.

Changqin Huang (Member, IEEE) received the Ph.D. degree in computer science and technology from Zhe jiang University, Hangzhou, China, in 2005. He completed two Postdoctoral Fellowship positions with the ECNU TCL Joint Workstation on Educational Technology and the Sun Yat-sen University, Guangzhou, China, on computer software and theory. He completed visiting research with the University of California Irvine, Irvine, CA, USA, in 2011, and La Trobe University, Melbourne, VIC, Australia, in 2018. He is currently a Distinguished Professor with Zhejiang Normal University, Jinhua, China, and also the Director of the Key Laboratory of Intelligent Education Technology and Application of Zhejiang Province, China. He has authored or coauthored several papers in prestigious journals in Computer Science and Educational Technology, including IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, *Communications on Applied Electronics*, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON CYBERNETICS, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS, CHB, BJET, and acted as PI for many projects related to AI and Its Applications in Education. Dr. Huang is a Guangdong Specially-Appointed Professor (Pearl River Scholar). His research interests include Big Data in education to machine learning and intelligent education. He is an Associate Editor for IEEE TRANSACTIONS ON LEARNING TECHNOLOGIES.

Xiao Chen received the bachelor's degree in computer science and technology from the Zhejiang University Of Finance & Economics, Hangzhou, China, in 2022. He is currently working toward the master's degree in Intelligent educational technology with Zhejiang Normal University, Jinhua, China. His research interests include human pose estimation and object detection.

Ming Li (Member, IEEE) received the Ph.D. degree from the Department of Computer Science and IT, La Trobe University, Melbourne, VIC, Australia, in 2017. He is currently a Shuang Long Scholar Distinguished Professor with the Zhejiang Key Laboratory of Intelligent Education Technology and Application, Zhejiang Normal University, Jinhua, China. He completed two Postdoctoral Fellowship positions with the Department of Mathematics and Statistics, La Trobe University, and the Department of Information Technology in Education, South China Normal University, Guangzhou, China, respectively. He has authored or coauthored 80 papers in top-tier journals and conferences, including IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, *Artificial Intelligence*, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, NeurIPS, ICML, IJCAI. His research interests include graph neural networks, graph representation learning, graph data mining, learning theory for neural networks. He, as a leading Guest Editor, organized a special issue *Deep Neural Networks for Graphs: Theory, Models, Algorithms and Applications* in IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, and a special session on *Recent Advances in Deep Learning for Graphs* in LOD2022. He is a Member of IEEE Task Force on Learning For Graphs, an Associate Editor for *Neural Networks*, *Applied Intelligence*, *Alexandria Engineering Journal*, *Network: Computation in Neural Systems*, *Soft Computing*, *Neural Processing Letters*.



Zhao Li received the Ph.D. degree in computer science from the University of Vermont, Burlington, VT, USA, in 2012. He is currently with Link2Do Technology Ltd, and an adjunct Professor with Zhejiang University, Hangzhou, China. He has authored or coauthored several papers in prestigious conferences and journals including WWW, AAAI, ICDE, VLDB, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING. His research interests include multi-agent reinforcement learning, Big Data driven security, and large-scale graph computing. He was the recipient of the Excellent Graduate Award from the University of Vermont, in 2012

Weiping Ding (Senior Member, IEEE) received the Ph.D. degree in computer science from the Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2013. From 2014 to 2015, he was a Postdoctoral Researcher with the Brain Research Center, National Chiao Tung University, Hsinchu, Taiwan, China. In 2016, he was a Visiting Scholar with the National University of Singapore, Singapore. From 2017 to 2018, he was a Visiting Professor with the University of Technology Sydney, Ultimo, NSW, Australia. He ranked within the top 2% Ranking of Scientists in the World by Stanford University, Stanford, CA, USA, during 2020–2023. He has authored or coauthored more than 300 articles, including more than 120 IEEE Transactions papers. His nineteen authored/coauthored papers have been selected as ESI Highly Cited Papers. He has coauthored four books. He has holds 32 approved invention patents, including two U.S. patents and one Australian patent. His main research interests include deep neural networks, granular data mining, and multimodal machine learning. He is an Associate Editor/Area Editor/Editorial Board member of IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, IEEE TRANSACTIONS ON FUZZY SYSTEMS, IEEE/CAA JOURNAL OF AUTOMATICA SINICA, IEEE TRANSACTIONS ON EMERGING TOPICS IN COMPUTATIONAL INTELLIGENCE, IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, IEEE TRANSACTIONS ON INTELLIGENT VEHICLES, IEEE TRANSACTIONS ON ARTIFICIAL INTELLIGENCE, *Information Fusion*, *Information Sciences*, *Neurocomputing*, *Applied Soft Computing*, *Engineering Applications of Artificial Intelligence*, *Swarm and Evolutionary Computation*. He was/is the Leading Guest Editor of Special Issues in several prestigious journals, including IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION, IEEE TRANSACTIONS ON FUZZY SYSTEMS, *Information Fusion*, *Information Sciences*. He is currently the Co-Editor-in-Chief of both *Journal of Artificial Intelligence* and *Systems and Journal of Artificial Intelligence Advances*.

Xindong Wu (Fellow, IEEE) received the bachelor's and master's degrees in computer science from the Hefei University of Technology, Hefei, China, and the Ph.D. degree in artificial intelligence from the University of Edinburgh, Edinburgh, U.K., in 1993. He is currently the Director and a Professor of the Key Laboratory of Knowledge Engineering with Big Data (the Ministry of Education of China), Hefei University of Technology. His research interests include Big Data analytics, data mining and knowledge engineering. He is a Foreign Member of the Russian Academy of Engineering, and a Fellow of the AAAS (American Association for the Advancement of Science). Dr. Wu is the Steering Committee Chair of the IEEE International Conference on Data Mining (ICDM) and the Editor in-Chief of *Knowledge and Information Systems* (KAIS, by Springer). He was the Editor-in-Chief of the IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING between 2005 and 2008 and Co-Editor-in-Chief of the *ACM Transactions on Knowledge Discovery from Data Engineering* between 2017 and 2020. He was the Program Committee Chair/Co-Chair for ICDM 2003 (the 3 rd IEEE International Conference on Data Mining), KDD 2007 (the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining), CIKM 2010 (the 19th ACM Conference on Information and Knowledge Management), and ICBK 2017 (the 8th IEEE International Conference on Big Knowledge).