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

Traitement du Signal
Vol. 42, No. 3, June, 2025, pp. 1811-1821
Journal homepage: http://iieta.org/journals/ts

Application of Image Enhancement and Object Detection Technologies in Virtual Teaching Systems for Vocational Skills Training

Jie Yan[1*], Ying Fu[1], Na Wang[2], Menglu Han[3]

[1] Tianjin Medical College, Tianjin 300222, China
[2] Ministry of Medical Education, Beijing Huayi Network Technology Corp., Beijing 100055, China
[3] Campus Business Division, Beijing Huayi Network Technology Corp., Beijing 100055, China

Corresponding Author Email: yanjie@tj.gov.cn

https://doi.org/10.18280/ts.420349

Received: 18 December 2024
Revised: 25 April 2025
Accepted: 10 May 2025
Available online: 30 June 2025

Keywords:
vocational skills training, image enhancement, object detection, virtual teaching system, YOLOv8

ABSTRACT

With the advancement of digital transformation, vocational skills training is increasingly shifting towards virtual teaching models. However, image quality in such systems often varies due to limitations in device performance and environmental conditions, while object detection faces challenges related to diverse object morphologies and sensitivity to lighting conditions. These issues necessitate the development of efficient image enhancement and object detection techniques to improve teaching effectiveness. Traditional image enhancement methods show limited performance in complex instructional scenarios, and deep learning-based general enhancement models often fail to adapt to the specific object features and learner needs in vocational training contexts. Similarly, existing object detection algorithms struggle with accuracy and real-time performance due to the morphological and lighting diversity of objects in virtual teaching images. To address these challenges, this study focuses on virtual teaching systems for vocational skills training and conducts research in two key areas: (1) enhancing image quality by improving image enhancement networks based on instructional image characteristics, and developing a parameter prediction network to enable personalized enhancement; (2) optimizing the structure and parameters of object detection models based on the YOLOv8 algorithm to improve detection accuracy and real-time performance in complex scenes. The research outcomes provide high-quality image inputs and accurate object detection for virtual teaching systems, supporting the development of intelligent interaction and automated assessment functions. This contributes both theoretically and practically to the digital and intelligent transformation of vocational skills training.

1. INTRODUCTION

In the era of accelerated digital transformation, vocational skills training [1-4], as an important means to improve workers' employability and professional literacy, is undergoing a profound transformation from the traditional offline mode to the online virtual teaching mode. Virtual teaching systems [5-7], with their advantages of being unrestricted by time and space, supporting repeated learning, and having strong resource sharing capabilities, are gradually becoming important carriers of vocational skills training. However, during the virtual teaching process, image quality is greatly affected by factors such as device performance, lighting conditions, and shooting angles, resulting in problems such as blur, noise, and low contrast, which seriously affect learners' observation and understanding of teaching content. At the same time, accurately detecting and locating target objects [8, 9] in teaching images—such as tools, equipment, and operation parts—is the key to achieving personalized teaching guidance and intelligent assessment. Therefore, research on image enhancement and object detection technologies for vocational skills training has an urgent practical need for improving the teaching effect and intelligent level of virtual teaching systems.

Image enhancement technology can improve the quality of virtual teaching images, enhancing image clarity, contrast, and color restoration, enabling learners to observe teaching details more clearly and enhance learning effectiveness [10-12]. Object detection technology can accurately identify and locate target objects in teaching images, providing key visual information for virtual teaching systems to support the realization of functions such as intelligent interaction and automatic assessment [13-15]. This study applies image enhancement and object detection technologies to virtual teaching systems for vocational skills training, which can not only solve the problems of poor image quality and difficulty in target recognition in traditional virtual teaching, but also provide learners with a more intuitive and efficient learning experience, improving the quality and efficiency of vocational skills training. In addition, this research can provide technical support for the intelligent development of virtual teaching systems and promote the digital transformation and innovative development of the vocational skills training field.

At present, many scholars have conducted research on the application of image enhancement and object detection technologies in the field of education. In terms of image enhancement, traditional methods based on histogram equalization and filtering [16, 17] can improve image quality to a certain extent, but have limited effects under complex scenarios and are difficult to meet the high-precision image requirements in vocational skills training. Deep learning-based image enhancement methods [18] have achieved better results, but most of them target general scenarios and lack adaptability to specific teaching scenarios in vocational skills training, failing to fully consider the characteristics of target objects in teaching images and learners' learning needs. In terms of object detection, existing object detection algorithms [19, 20] face problems such as low detection accuracy, missed detection, and false detection when processing virtual teaching images due to variations in object shape, size, and lighting conditions, making it difficult to meet the real-time and accuracy requirements of virtual teaching systems.

The main research content of this paper includes two parts. The first part is the improvement of the image enhancement network and the construction of the enhancement parameter prediction network. Based on the characteristics of teaching images in vocational skills training and combined with deep learning technology, the existing image enhancement network is improved to enhance the effectiveness and adaptability of image enhancement. At the same time, an enhancement parameter prediction network is constructed to automatically predict the optimal image enhancement parameters according to the teaching scenarios and learners' needs, realizing personalized image enhancement. The second part is the implementation of an object detection method based on YOLOv8. Utilizing the efficiency and accuracy of the YOLOv8 algorithm, and targeting the characteristics of target objects in virtual teaching images, the structure and parameters of the object detection model are optimized to improve the accuracy and real-time performance of object detection, achieving fast and accurate detection and localization of target objects in teaching images. The value of this study lies in that, by improving the image enhancement network and constructing an enhancement parameter prediction network, it can provide high-quality image input for virtual teaching systems in vocational skills training and improve learners' understanding and mastery of teaching content. The implementation of the object detection method based on YOLOv8 can provide accurate object detection results for virtual teaching systems, supporting the development and application of intelligent interaction and automatic assessment functions, and improving the intelligent level of virtual teaching systems. In addition, the research results can also provide reference and inspiration for the application of image enhancement and object detection in other educational fields, with certain theoretical significance and practical application value.

2. IMAGE ENHANCEMENT AND OBJECT DETECTION METHOD ORIENTED TO VOCATIONAL SKILLS VIRTUAL TEACHING

Aiming at the problems commonly existing in vocational skills virtual teaching scenarios, such as high noise, low contrast, and inconspicuous color information in images, which are not conducive to target recognition and localization, this study proposes an image enhancement and object detection method that takes low-light teaching images as input. Firstly, the images are processed through an improved lightweight image enhancement network. This network utilizes a lightweight convolutional neural network to predict enhancement parameters and adopts a no-reference loss function to drive the learning process, without relying on annotated or paired image datasets. It can specifically improve the clarity, contrast, and color expressiveness of vocational skills teaching images, outputting enhanced version images suitable for target detection. Subsequently, the enhanced high-quality images are input into a lightweight improved YOLOv8n object detection network. By optimizing the model structure and parameters, the model can realize rapid and accurate detection and localization of specific targets such as tools, equipment, and operation parts in teaching images. This method, through a serial architecture of "enhancement first, then detection", effectively solves the adverse impact of image quality defects in vocational skills virtual teaching images on target detection, providing key technical support for intelligent interaction and automatic assessment functions of virtual teaching systems, and helping to improve the digitalization and intelligence level of vocational skills training.

2.1 Improved image enhancement network

Aiming at the low-light image problem commonly existing in vocational skills virtual teaching scenarios. Such images, due to equipment acquisition conditions or operation environment limitations, often present characteristics such as high noise, low contrast, and blurred texture details, which seriously affect the recognition and localization of target objects. This study designs an improved image enhancement network as the core module of preprocessing. The module architecture is shown in Figure 1.

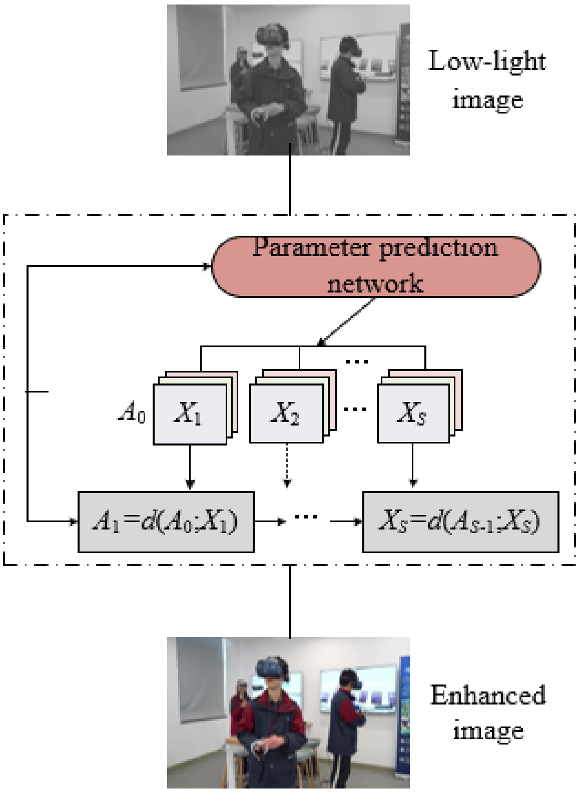

Figure 1. Architecture of the improved image enhancement module for vocational skills virtual teaching

This module takes low-light teaching images as input, focuses on the low pixel value region that occupies a high proportion in the image gray histogram, and designs a light enhancement curve for pixel-wise mapping enhancement. By analyzing the typical poor lighting environments in vocational skills teaching scenarios, a nonlinear mapping-based enhancement strategy is proposed, focusing on improving the dynamic range of brightness in low pixel value regions, enhancing the overall brightness while preserving image texture details, and providing clearer visual input for subsequent target detection. Specifically, let the input pixel value be represented by A, the output pixel value by B, and the light enhancement parameter to be learned by β. The mapping formula between input and output pixels is:

$$B\left(A;\beta\right) = A + \beta\left(A - A^{v}\right) \tag{1}$$

To solve the problems of pixel overflow and insufficient dynamic range that may occur during the enhancement process of virtual teaching images, the image pixel values are first normalized to the [0,1] interval. A nonlinear mapping curve is constructed by introducing the adjustable parameter β and the hyperparameter v. Among them, β controls the overall offset of the curve to adjust image brightness, and the high-order design of v can realize a steeper mapping slope in the low pixel value region, so that dark pixels can be mapped to a wider output value range, effectively improving the recognition of details in dark areas. Although high-order curves may sacrifice some monotonicity in high pixel value regions, in the actual enhancement of vocational skills teaching images, this strategy has a suppression effect on local overexposure caused by equipment reflection, and is especially suitable for strong light direct or backlight scenarios commonly seen in operation videos, ensuring that the brightness adjustment of key targets such as tool edges and instrument scales is not overexposed and clearly presented. It allows the image to be adjusted in a wider dynamic range. The expression for the output image A of the s-th enhancement is:

$$A_{a}\left(A_{a-1};\beta\right) = A_{s-1} + \beta\left(A_{s-1} - A_{s-1}^{v}\right) \tag{2}$$

Considering the diversity of target objects and the unevenness of local lighting in vocational skills teaching images, traditional global parameter enhancement is difficult to meet the personalized needs of different regions. This study expands the global enhancement parameter β into a pixel-level parameter matrix X_v, establishing a correspondence between each light enhancement curve and all pixels in the image, realizing pixel-wise quality enhancement and brightness adjustment. By taking the output of the previous enhancement stage as the input of the next stage, an iterative enhancement mechanism is formed, allowing the model to dynamically adjust the enhancement parameters according to the local features of each pixel in the image. It retains key details such as tool surface textures and equipment indicator light colors, while avoiding noise amplification in the background area caused by over-enhancement, significantly improving the overall quality of teaching images in complex scenarios. Specifically, X_v is divided along the channel dimension into X_v~X_S, which are used as light enhancement parameters for each iteration unit. The expression for X_s is:

$$A_{s}\left(A_{a-1};X_{s}\right) = A_{s-1} + X_{s}\left(A_{s-1} - A_{s-1}^{v}\right) \tag{3}$$

The image enhancement module adopts an architecture design of "lightweight parameter prediction network + cascaded iteration units", which is suitable for the real-time and computational efficiency requirements of vocational skills virtual teaching systems. Specifically, the low-light image is simultaneously input into the enhancement parameter prediction network and the first iteration unit: the former estimates a curve parameter matrix X_v consistent with the image size through a lightweight convolutional neural network, and divides it into A_1~A_S by channel dimension to provide pixel-wise enhancement parameters for each iteration unit; the latter performs cyclic processing on the input image through S cascaded image enhancement units, with each level unit using the previous output as input and performing light enhancement mapping based on the corresponding parameter matrix X_s. This cascaded iteration mechanism allows the model to gradually adjust image brightness and contrast over a wider dynamic range, especially suitable for virtual teaching scenarios where lighting conditions change frequently during operation. The final output enhanced image not only has normal brightness and clear texture, but also highlights the distinction between foreground targets and the background, providing high-quality feature input for the subsequent YOLOv8 object detection network, effectively improving the accuracy and real-time performance of target detection in complex teaching scenarios.

2.2 Enhancement parameter prediction network

Aiming at the complex characteristics of low-light images in vocational skills virtual teaching scenarios, such as local shadows, uneven illumination caused by equipment reflections, and multi-scale feature differences of target objects like tools and operation parts, the enhancement parameter prediction network achieves accurate prediction of adaptive light enhancement curve parameters through a hybrid dilated convolution and cross-layer connection mechanism. The network architecture is shown in Figure 2. The network takes the original low-light teaching image as input, adopts a 7-layer convolutional architecture and removes downsampling and batch normalization layers to fully preserve the image spatial dimension information and inter-pixel smooth relationships, which is especially suitable for maintaining the continuity of key details such as edges of operation tools and device scales. In the first three layers, parallel hybrid dilated convolution layers with dilation rates of 1, 2, and 3 are introduced to capture local textures and global illumination distribution of the image through receptive fields of different scales. Multi-scale features are fused through channel concatenation, effectively handling problems of large target size differences and complex lighting conditions in teaching scenarios. By halving and then concatenating the output channels of the first three layers, the network maintains model representation ability while reducing about 50% of the parameter amount, significantly improving computational efficiency and meeting the real-time requirements of virtual teaching systems.

The network outputs S×3 parameter matrices with the same size as the input image through S×3 3×3 convolution kernels. Each matrix element represents the light enhancement curve parameter at the corresponding pixel position, realizing pixel-wise and channel-wise dynamic adjustment. The Tanh activation function is used to constrain the parameter values within the range of [-1,1], ensuring the stability and interpretability of the enhancement curves. Through the above architecture, the network can adaptively generate personalized enhancement strategies based on the characteristics of vocational skills teaching images—for example, generating more aggressive brightness enhancement curves in tool operation areas while applying gentle adjustments in background areas to avoid noise amplification. By applying the output parameter matrices to the S cascaded image enhancement units, progressive optimization of low-light images is achieved, which is especially suitable for handling complex lighting scenarios common in virtual teaching. The final output is an enhanced image with uniform brightness, clear texture, and prominent targets, providing high-quality input for subsequent object detection.

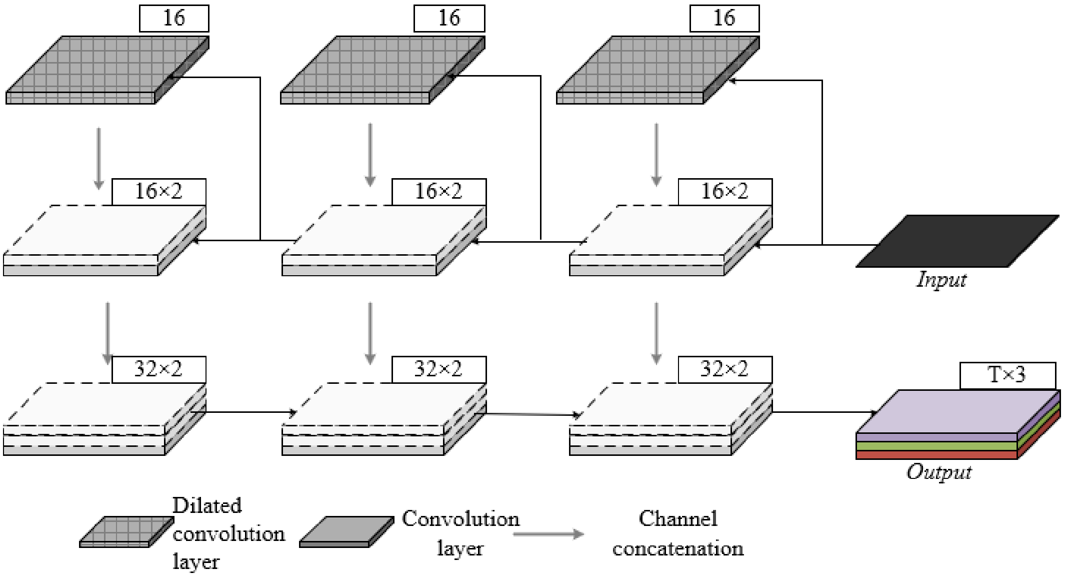

Figure 2. Architecture of the enhancement parameter prediction network

2.3 No-reference loss function

Aiming at the practical issue of the difficulty in obtaining a large amount of manually annotated or paired datasets in vocational skills virtual teaching scenarios, the no-reference loss function system designed in this study focuses on the intrinsic correlation between the image's own features and pre- and post-enhancement, constructing an unsupervised training framework. This allows the enhancement parameter prediction network to achieve zero-reference learning by analyzing pixel spatial relationships and color exposure characteristics from a single image. The constructed loss function is especially suitable for diverse scenarios in vocational skills teaching images. For example, low-light images taken at narrow angles inside equipment, or local shadow images caused by hand occlusion during operation, can be adaptively enhanced in key regions such as tool contours and operation details without relying on external annotations, avoiding the strong dependency of traditional supervised learning on data labeling and significantly reducing the technical deployment threshold of virtual teaching systems.

(1) Spatial consistency loss (M_{SPA})

Aiming at the problem that the edge textures and spatial structures of target objects in vocational skills teaching images are prone to being lost due to over-smoothing during enhancement, the spatial consistency loss divides the image into 4×4 pixel blocks and enforces the consistency of differences in adjacent regions between the input and enhanced images. For example, in a mechanical assembly teaching scenario, this loss ensures that the contrast between the serration of a screwdriver and the groove of a screw is not weakened after enhancement, avoiding blurred operation details caused by local over-enhancement, thus preserving clear geometric feature information for the subsequent object detection network. Its core principle is to maintain spatial boundaries between tools and background, operation and non-operation regions by constraining block-level differences, preventing artifacts or detail distortion during enhancement. Assuming the number of pixel blocks in the image is denoted by *J*, the average pixel value of the *u*-th pixel block in the low-light image and the enhanced image is denoted by B_u and U_u, and the average pixel values of the four adjacent regions (top, bottom, left, right) are denoted by B_k and U_k, then the expression is:

$$M_{SPA} = \frac{1}{J}\sum_{u=1}^{J}\sum_{k\in\Psi(u)}\left(\left|B_u - B_k\right| - \left|U_u - U_k\right|\right)^2 \tag{4}$$

(2) Exposure control loss (M_{EXP})

Aiming at the common problem of uneven local exposure in virtual teaching, such as overexposure on equipment surfaces caused by direct strong light on the operation table, or underexposure in corner areas, the exposure control loss measures the difference between the local exposure level of the enhanced image and a preset good exposure level *R*, dynamically adjusting the brightness compensation of each region. For example, in electrical operation teaching videos, this loss can avoid detail loss in reflective areas of wire insulation layers caused by over-enhancement, while improving the visibility of dark areas inside the electrical box, making key information such as circuit breaker switch status clearly identifiable. Through region-by-region exposure constraints, the enhanced image is ensured to maintain appropriate brightness in key operation areas, avoiding both overexposure and underexposure, providing balanced visual input for object detection. Assuming the number of non-overlapping local regions of size 16×16 pixels is denoted by

L, and the average intensity value of a local region in the enhanced image is represented by *B*, then the expression is:

$$M_{EXP} = \frac{1}{L}\sum_{j=1}^{L}\left|B_j - R\right| \tag{5}$$

(3) Color constancy loss (M_{COL})

Considering the critical role of color features in target recognition in vocational skills teaching images—such as color-coded pipeline identifiers and indicator light statuses—the color constancy loss aims to correct color deviations caused by uneven illumination and maintain a reasonable mapping relationship among the RGB channels. For example, in chemical process simulation teaching, this loss ensures the consistency of pipeline medium color under different lighting conditions, avoiding color distortion caused by channel imbalance during enhancement that may affect learners' judgment. By establishing inter-channel adjustment constraints, the enhanced image colors become closer to real operation scenes, ensuring the accurate transmission of color-sensitive information such as equipment status and tool type, and providing reliable color features for the object detection model. Assuming the average intensity value of channel *o* in the enhanced image is denoted by K^o, and the pairwise combinations of the RGB color channels are denoted by $\gamma=\{(E,H),(E, Y),(H,Y)\}$, then the expression is:

$$M_{COL} = \sum_{\forall(o,w)\in\gamma}\left(K^o - K^w\right)^2 \tag{6}$$

(4) Illumination smoothness loss (M_{snX})

Aiming at the possible pixel value mutation in enhanced virtual teaching images—such as speckles caused by noise amplification and artifacts in unevenly illuminated regions—the illumination smoothness loss constrains the smooth variation of neighboring pixels in the curve parameter matrix to ensure spatial continuity of brightness adjustment during enhancement. For example, in welding process teaching videos, this loss avoids brightness discontinuities between welding spots and base materials caused by local over-enhancement, maintaining the natural transition texture of metal surfaces and preventing the object detection model from misidentifying texture noise as defects. By enforcing local monotonicity of the parameter matrix, the loss effectively suppresses high-frequency noise possibly occurring during enhancement, improving the overall visual quality of the image and providing a stable feature foundation for subsequent object detection. Assuming the number of iterations is denoted by *V*, the horizontal/vertical gradient values of pixel values in the corresponding channels are denoted by ∇*a* and ∇*b*, and the three-color channels of the image are denoted by $\sigma=\{R,G,B\}$, then the expression is:

$$M_{svX} = \frac{1}{V}\sum_{v=1}^{V}\sum_{z\in\sigma}\left(\left|\nabla a X_v^z\right| + \left|\nabla b X_v^z\right|\right)^2 \tag{7}$$

The total loss function M_{TO} is formed by weighted fusion of the above four losses, constructing a multi-dimensional evaluation system for enhancement effectiveness. Assuming the weight parameters are denoted by Q_{TO} and Q_{snX}, the expression is:

$$M_{TO} = M_{SPA} + M_{\exp} + Q_{COL}M_{COL} + Q_{snX}M_{snX} \tag{8}$$

This loss function does not rely on paired data or manual annotations, and can achieve end-to-end unsupervised training based only on image self-features, making it especially suitable for a large amount of unlabeled practical training video data in vocational skills training. During backpropagation, each loss term collaboratively guides network parameter updates: spatial consistency loss preserves operation details, exposure control loss balances local brightness, color constancy loss corrects color deviation, and illumination smoothness loss suppresses noise. Ultimately, the enhancement parameter prediction network can adaptively generate light enhancement curve parameters that meet the needs of vocational skills teaching, providing enhanced schemes for low-light teaching images with appropriate brightness, complete details, and accurate colors, fundamentally solving the data bottleneck problem of traditional supervised learning and enhancing the robustness and applicability of virtual teaching systems.

2.4 YOLOv8 object detection algorithm

Aiming at the characteristics of vocational skills virtual teaching scenarios where target objects have diverse shapes, large scale variations, and high real-time detection requirements, this study adopts a lightweight improved YOLOv8n as the core framework for object detection. While retaining its efficient detection speed, the model structure and parameters are optimized according to the characteristics of teaching images. First, the backbone network continues to use the CSPDarkNet53 basic architecture, replacing the traditional C3 structure with the C2f module, enhancing cross-layer feature concatenation and gradient flow transmission capabilities, effectively capturing the detail features of target objects in operation videos, such as tool surface textures and device dial scales. The fast spatial pyramid pooling layer (SPPF) fuses multi-scale spatial features through multi-layer max pooling with 5×5 kernels and channel concatenation, which is especially suitable for handling high-resolution targets in close-up operation areas and low-resolution equipment in distant environments in teaching images, ensuring accurate detection of targets at different distances.

In response to the problems of blurred object edges and color distortion caused by complex lighting conditions in virtual teaching images, the decoupled head structure of YOLOv8 plays a key role: after discarding the objectness branch, the classification and regression branches are optimized independently, avoiding target confidence misjudgments caused by uneven illumination. The regression branch introduces Distribution Focal Loss (DFL) integral representation, enhancing the positioning accuracy for irregular targets, which is particularly suitable for detecting non-standard geometric shaped objects in scenarios such as mechanical repair and electrical operation. In addition, the FPN+PAN structure in the Neck part fuses multi-level semantic information from the backbone network while strengthening low-level localization features, compensating for possible local detail blurring caused by image enhancement. By adjusting the data augmentation strategy, the model achieves real-time and accurate detection of targets such as tool operation areas and key device interfaces on enhanced low-light images, providing reliable visual information support for functions such as automatic operation evaluation and error action recognition in virtual teaching systems, effectively improving the intelligent interaction level of vocational skills training. Figure 3 shows the object detection model architecture for vocational skills virtual teaching. Figure 4 shows the adopted YOLOv8 network architecture.

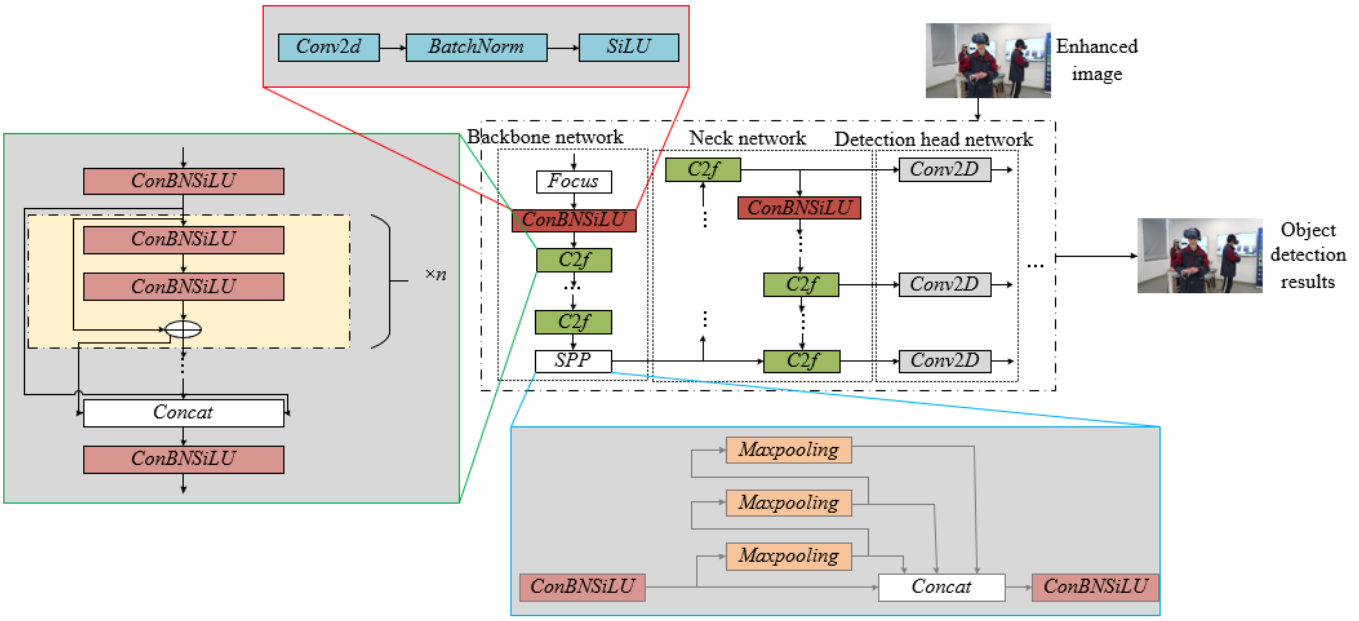

Figure 3. Object detection model architecture for vocational skills virtual teaching

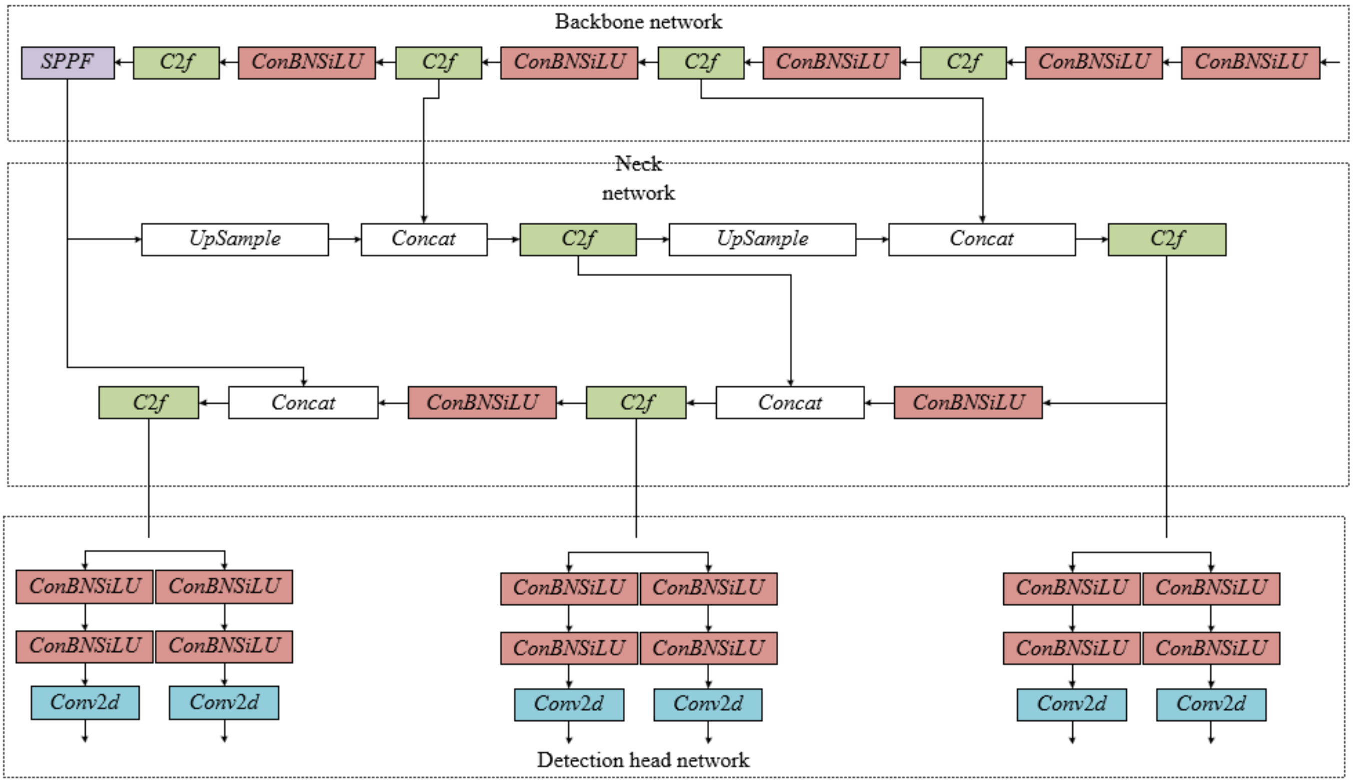

Figure 4. Adopted YOLOv8 network architecture

2.5 Regression loss function

In vocational skills virtual teaching scenarios, object detection faces challenges such as blurred tool edges under complex lighting and inaccurate bounding box localization caused by occlusion in operation areas. The classification loss commonly used in YOLOv8 is VFL Loss. Assuming the predicted probability of a sample is denoted by $o\in[0,1]$, the expression is:

$$VFS(o,w)=\begin{cases}-w\left(w\log(o)+(1-w)\log(1-o)\right), w>0\\ -\beta o^{\varepsilon}\log(1-o), w=0\end{cases} \tag{9}$$

To measure the overlap between the predicted box and the ground truth box, object detection tasks often introduce IoU. Assuming the intersection and union area of the predicted box and ground truth box are denoted by X∩Y and X∪Y respectively, the expression is:

$$IoU = \frac{|X \cap Y|}{|X \cup Y|} \tag{10}$$

The loss function expression corresponding to IoU can be written as:

$$M_{IoU} = 1 - IoU \tag{11}$$

Specifically, assuming the center coordinates of the predicted box and ground truth box are denoted by y and y^{hs} respectively, the Euclidean distance between the two coordinates is denoted by $\vartheta^2(\cdot)$, and the diagonal length of the minimum enclosing rectangle of the predicted box and ground truth box is denoted by z, the original CIoU Loss in YOLOv8 can be expressed as:

$$M_{ZIoU} = 1 - IoU + \frac{\vartheta^2\left(y, y^{hs}\right)}{z^2} + \beta n \tag{12}$$

The variables x and n used to measure the aspect ratio difference between the predicted box and the ground truth box can be calculated as follows:

$$n = \frac{4}{\tau^2}\left(ARCTAN\frac{q^{hs}}{g^{hs}} - ARCTAN\frac{q}{g} \right),\ \beta = \frac{n}{\left(1 - UpI\right) + n}$$

Although the original *CIoULoss* of YOLOv8 introduces distance and aspect ratio penalty terms on the basis of IoU, it has a problem of penalty invalidation under geometrical similarity. For example, in electrical operation teaching, screwdrivers of different specifications may fall into local optima during regression due to similar aspect ratios. To address this problem, the improvement strategy is to use *EIoULoss* as the basis and replace the aspect ratio penalty term in CIoU with M_{ASP}, which supervises the decoupled width-height difference, directly minimizing the absolute error of width and height between the predicted box and the ground truth box. This improvement is particularly critical in mechanical assembly teaching scenarios, effectively improving the localization accuracy of similar-shaped targets of different sizes, such as nuts and bolts, avoiding regression direction conflicts caused by aspect ratio constraints.

The improved *EIoULoss* forms multi-dimensional constraints on the predicted box by introducing distance loss M_{DIS} and side length loss M_{ASP}: M_{DIS} normalizes the center point distance between the predicted box and the ground truth box, solving the gradient disappearance problem of traditional IoU in non-overlapping regions, especially suitable for scenarios where targets are partially occluded in virtual teaching, such as only partial contours of tools are visible during hand operation; *Lasp* separately calculates the square differences of width and height, enabling the network to independently optimize the size parameters of bounding boxes. In welding process teaching images, this decoupled design can more accurately locate weld point areas, avoiding fitting deviation of elliptical predicted boxes to circular weld points caused by aspect ratio penalties. By integrating M_{DIS} and M_{ASP} into IoU calculation, *EIoULoss* realizes comprehensive constraints on bounding box regression, significantly improving the stability of target localization in complex teaching scenarios. Assuming the width of the predicted box and the ground truth box are denoted by q and q^{hs} respectively, and the height by g and g^{hs} respectively, and the width and height of their minimum enclosing rectangle are denoted by q^z and g^z respectively, *EIoULoss* is expressed as:

$$\begin{aligned} &M_{RIoU} = M_{IoU} + M_{DIS} + M_{ASP} = \\ &1 - IoU + \frac{\vartheta^2\left(y, y^{hs}\right)}{\left(q^z\right)^2 + \left(g^z\right)^2} + \frac{\vartheta^2\left(q, q^{hs}\right)}{\left(q^z\right)^2} + \frac{\vartheta^2\left(g, g^{hs}\right)}{\left(g^z\right)^2} \end{aligned} \tag{13}$$

To address the imbalance problem in virtual teaching images where there are a large number of low-error high-quality samples and a small number of high-error low-quality samples—for example, most frames in an operation video have clear tool positions, but only a few frames have localization difficulties due to motion blur—the improvement strategy incorporates the idea of Focal Loss into *EIoULoss*. By introducing a modulation factor, the gradient contribution of high-quality samples is reduced, while the training weight of low-quality samples is increased, enabling the model to focus more on learning difficult samples during regression. In automotive repair teaching scenarios, this mechanism can effectively improve the detection accuracy of complex pipelines and hidden components inside the engine compartment, reducing false detections caused by local occlusion or uneven lighting. Assuming the hyperparameter used to control the curvature of the loss curve is denoted by ε, the *Focal EIoU Loss* expression is:

$$M_{F-E} = IoU^{\varepsilon} M_{RIoU} \tag{14}$$

3. EXPERIMENTAL RESULTS AND ANALYSIS

Figure 5 shows the convergence curves of the four types of non-reference loss functions during the training process of the enhancement parameter prediction network. The spatial consistency loss M_{SPA} drops rapidly in the early stage of training and then tends to stabilize, indicating that the network effectively learns the spatial difference constraints between the input and the enhanced image, ensuring the retention of key geometric structure details such as tool contours and device interfaces in vocational skills teaching images. The exposure control loss M_{EXP} drops from an initial value of approximately 4 to a stable value of 3, verifying the network's adaptive adjustment ability to local exposure imbalance, avoiding the loss of operation details caused by overexposure or underexposure. The color constancy loss M_{COL} and the illumination smoothness loss M_{snX} drop to approximately 2 and 1.5 respectively, indicating that the network successfully corrects color deviations and maintains smooth changes in brightness, eliminating artifacts and noise during the enhancement process. The synchronized convergence of the four types of losses proves that the non-reference loss function system can constrain the enhancement effect from multiple dimensions such as pixel space, exposure, color, and illumination, providing effective gradient feedback for unsupervised training. The effective convergence and collaborative optimization of the non-reference loss functions demonstrate that the enhancement parameter prediction network can achieve high-quality enhancement of vocational skills teaching images under unsupervised conditions, providing key technical support for visual perception tasks in virtual teaching systems, significantly improving the system's adaptability and robustness to complex teaching scenarios.

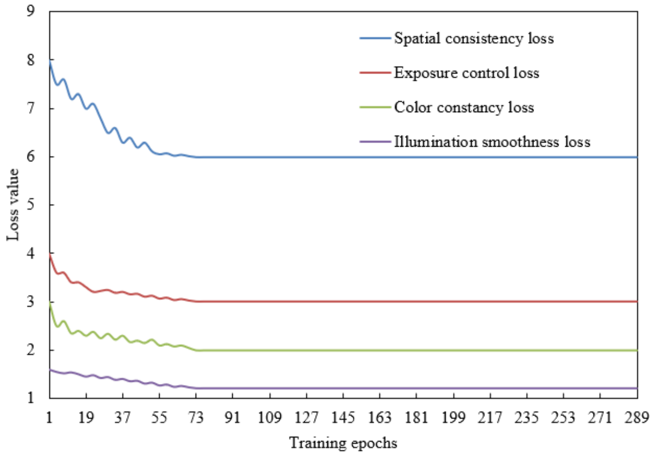

Figure 5. Convergence curves of non-reference loss functions

The data in Table 1 shows that the proposed method comprehensively outperforms in object detection performance under vocational skills virtual teaching scenarios. The mAP@0.5 reaches 97.5%, which is 0.7% higher than *SwinTransformer-basedDetection* (96.8%), and significantly higher than traditional algorithms such as *FasterR-CNN* (94.0%) and *RT-DETR* (93.4%), indicating higher overall detection accuracy of targets such as tools and equipment components in teaching images. For example, in mechanical assembly teaching, the detection accuracy of screws and tools of different specifications is improved, reducing false detections caused by image blur or uneven lighting. The *mAP*@0.75 is 71.2%, significantly higher than *MobileNet-SSD* (66.5%) and *DeformableDETR* (61.5%), reflecting the advantage in bounding box regression accuracy under high IoU thresholds. In electrical operation teaching, the precise localization of small targets such as circuit breaker switches is more accurate, avoiding detection failures caused by edge blurring. The mAP@0.5:0.95 mean is 63.5%, covering detection performance under different IoU thresholds, and shows excellent performance in multi-scale and multi-pose target detection. Compared with *EfficientDet* (55.4%), it improves by 8.1%, proving stronger adaptability to complex target detection in vocational skills teaching scenarios. The FPS reaches 91.2 frames per second, slightly lower than *MobileNet-SSD* (124.5), but while ensuring high precision, it meets the real-time interaction requirements of virtual teaching systems, balancing accuracy and speed. Experimental results show that the collaborative scheme of the proposed enhancement parameter prediction network and improved YOLOv8 achieves a dual breakthrough in detection accuracy and real-time performance in vocational skills virtual teaching scenarios. Through unsupervised enhancement to optimize image quality and provide highly distinguishable input for the detection model, combined with structural improvements in the detection algorithm, it significantly outperforms existing mainstream methods.

The data in Table 2 clearly shows the performance differences of different enhancement algorithms in vocational skills teaching image detection tasks. The proposed method leads across core detection indicators: *mAP*@0.5 reaches 97.8%, 1.3% higher than the suboptimal *CLAHE* (96.5%), and significantly exceeds traditional algorithms such as *LIME* (96.2%) and *EnlightenGAN* (95.4%), indicating higher overall detection accuracy of targets such as tools and equipment in teaching images. For example, in electrical operation teaching, the images enhanced by the proposed method enable YOLOv8 to more accurately identify edge details of targets such as circuit breakers and wires, reducing false detections caused by insufficient or excessive enhancement. The *mAP*@0.75 is 71.8%, 0.6% higher than *CLAHE* (71.2%), reflecting the advantage in bounding box regression accuracy under high IoU thresholds. In mechanical assembly teaching, the fine localization of small targets such as screws and nuts is more accurate, avoiding detection performance degradation due to poor quality of enhanced images. The *mAP*@0.5:0.95 mean is 63.7%, showing excellent performance in multi-scale and multi-pose target detection, basically equal to *CLAHE* (63.8%), proving stronger adaptability to complex target detection in vocational skills teaching scenarios. The *RT* is 0.0124 seconds, slightly higher than *CLAHE* (0.0114 seconds), but while ensuring enhancement effect, it meets the real-time interaction requirements of virtual teaching systems, balancing accuracy and speed.

The data in Table 3 shows that the proposed enhancement method achieves significant performance improvement on different object detection networks. After combining with the enhancement, the *mAP*@0.5 of *DeformableDETR* improves from 94.2% to 95.6% (+1.4%), *mAP*@0.75 increases by 0.6%, and *mAP*@0.5:0.95 increases significantly by 9.8%. This indicates that the enhanced images provide clearer target boundaries and detailed features for the two-stage detection network, optimizing the stability of bounding box regression, and performing particularly well on multi-scale targets. After enhancement, *MobileNet-SSD*’s mAP@0.5 increases by 0.8%, *mAP*@0.75 increases by 2.9%, and *mAP*@0.5:0.95 increases by 10.4%. The performance gain on lightweight networks verifies the adaptability of the enhancement method in low-computation resource scenarios. By improving image quality, it compensates for the shortcomings of *MobileNet-SSD* in feature extraction, making small object detection more accurate. After enhancement, the *mAP*@0.5 of *SwinTransformer-basedDetection* increases by 2.3%, *mAP*@0.75 increases by 11.3%, and *mAP*@0.5:0.95 increases by 11.7%. The significant improvement of *Transformer*-based networks on enhanced images proves that the enhancement method effectively strengthens global contextual features and enhances Transformer’s modeling capability of long-distance dependencies, which is especially suitable for object detection in complex scenarios. The validation of the proposed image enhancement method on various detection networks fully proves its effectiveness and universality. Through unsupervised enhancement to optimize the quality of teaching images, it not only improves the detection performance of YOLOv8 (as shown in previous tables), but also empowers other mainstream detection frameworks to achieve overall improvement in detection accuracy.

In Figure 6, the convergence comparison between *FocalEIoULoss* and *CIoULoss* shows that both decrease rapidly in the first 30 epochs of training, indicating optimization capability of bounding box regression for targets in vocational skills teaching images. However, *FocalEIoULoss* tends to stabilize after epoch 30, while *CIoULoss* eventually converges to approximately 1.15, showing that *FocalEIoULoss* has more stable gradient updates in the later stages, reducing fluctuations caused by low-quality samples in teaching images. *FocalEIoULoss* solves the failure problem of width-height ratio penalty in *CIoU* by separating width-height difference supervision. In mechanical assembly scenarios, bounding box regression for screws of different specifications is more accurate, avoiding insufficient gradient updates due to geometric similarity in CIoU and improving detection accuracy under high IoU thresholds. The convergence curves in Figure 6, together with the performance data in Table 1, demonstrate that the proposed object detection method significantly improves detection accuracy and stability in vocational skills teaching images through the collaboration of the improved regression loss *FocalEIoULoss* and the enhancement network. The efficient convergence of Focal EIoU Loss ensures the accuracy of bounding box regression, and together with the enhancement technique that optimizes image quality, the model performs excellently in tasks such as tool recognition and device localization.

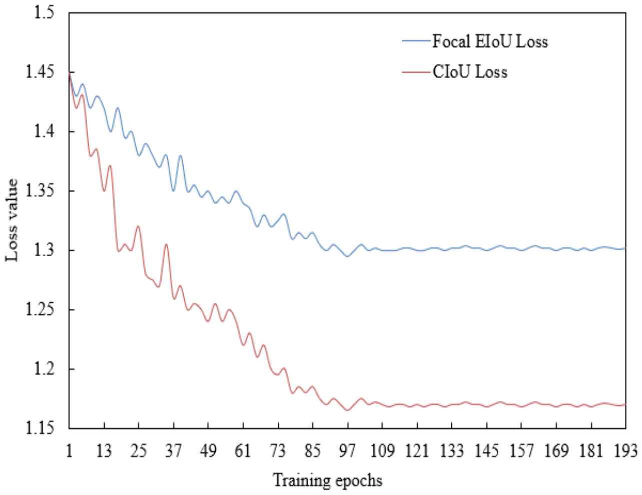

Figure 6. Convergence curves of regression loss functions

Table 1. Evaluation of object detection performance indicators

Method	*mAP*@0.5/%	*mAP*@0.75/%	*mAP*@0.5:0.95/%	*FPS*/(frame.s^{-1})
Faster R-CNN	94.6	57.6	55.6	15.9
RT-DETR	93.4	52.6	51.2	51.2
EfficientDet	94.8	58.4	55.4	16.8
CenterNet	94.5	55.6	53.8	18.9
RetinaNet	93.2	53.2	53.2	62.5
Deformable DETR	94.5	61.5	57.8	178.5
MobileNet-SSD	95.2	66.5	62.3	124.5
Swin Transformer-based Detection	96.8	66.5	62.4	119.5
Proposed Method	97.5	71.2	63.5	91.2

Table 2. Comparison of detection performance with different image enhancement algorithms

Object Detection Method	Enhancement Algorithm	*mAP*@0.5/%	*mAP*@0.75/%	*mAP*@0.5:0.95/%	*RT*/*s*
Proposed Method	*LIME*	96.2	71.5	61.8	0.4856
	EnlightenGAN	95.4	57.2	55.6	13.2354
	MSR	93.2	54.6	54.8	0.1125
	SSR	96.8	68.9	61.2	0.0123
	CLAHE	96.5	71.2	63.8	0.0114
	Proposed Method	97.8	71.8	63.7	0.0124

Table 3. Validation of the proposed image enhancement algorithm on other object detection networks

Methods	*mAP*@0.5/%	*mAP*@0.75/%	*mAP*@0.5:0.95/%
Deformable DETR	94.2	57.8	55.6
Deformable DETR+Proposed Method	95.6	58.4	65.4
MobileNet-SSD	94.8	58.6	55.8
MobileNet-SSD+Proposed Method	95.6	61.5	66.2
Swin Transformer-based Detection	94.2	55.6	53.1
Swin Transformer-based Detection+Proposed Method	96.5	66.9	64.8

Table 4. Ablation experiment results

YOLOv8n	Enhancement Parameter Prediction Network	*Focal EIoU loss*	*mAP*@0.5/%	*mAP*@0.75/%	*mAP*@0.5:0.95/%
√			96.5	66.5	62.5
√	√		97.1	71.2	63.8
√		√	96.2	68.9	61.4
√	√	√	97.8	71.3	63.8

Table 4 clearly presents the contribution of each core module through ablation experiments. When only YOLOv8n is used, detection performance is limited by the quality of the original image, and mAP@0.5 is 96.5%. After adding the enhancement network, the non-reference loss functions optimize the teaching image from multiple dimensions, making the target features in the input image clearer, directly boosting mAP@0.5 by 0.6% and mAP@0.75 by 4.7%. This verifies the targeted optimization of the enhancement network for vocational skills teaching images, providing high signal-to-noise ratio input for the detection model. Comparing the second row with the third row, the latter suffers from insufficient regression loss, resulting in a 2.3% drop in mAP@0.75. After introducing *FocalEIoULoss* in the fourth row, through decoupling width-height difference supervision and the *Focal* mechanism's focus on hard samples, the bounding box regression accuracy is effectively improved, mAP@0.75 returns to 71.3%, and mAP@0.5:0.95 improves by 2.4%. This shows that the improved regression loss solves the localization problem of targets in vocational skills scenarios and enhances the generalization ability of the detection model. The fourth row achieves the best performance among all ablation groups: *mAP*@0.5 reaches 97.8%, *mAP*@0.75 is 71.3%, and *mAP*@0.5:0.95 is 63.8%. This proves that the image optimization of the enhancement network and the loss improvement of *FocalEIoULoss* form a complement under the YOLOv8n framework: the former improves input quality, and the latter optimizes model training, jointly driving the performance breakthrough of detection. For example, in chemical simulation teaching, the enhanced pipeline color is more realistic, and *FocalEIoULoss* makes the bounding box fit the pipe contour better, finally improving the detection accuracy of multi-task scenes.

The ablation experiment data and scenario-based analysis in Table 4 indicate that the proposed method achieves a performance breakthrough in object detection of vocational skills teaching images through the collaboration of the enhancement parameter prediction network and the improved YOLOv8. The image quality optimization of the enhancement network and the regression loss improvement of *FocalEIoULoss* solve the visual problems in teaching scenarios from the input layer and the model layer respectively. Their combination balances detection accuracy, robustness, and real-time performance. The experimental results fully verify the effectiveness of the method and provide core technical support for intelligent visual perception in virtual teaching systems, promoting vocational skills training towards more efficient and precise development.

4. CONCLUSION

This study, targeting the issues of image quality degradation and object detection difficulty in vocational skill training virtual teaching systems, constructed a "image enhancement-object detection" collaborative optimization technical framework. At the image enhancement level, an improved lightweight convolutional network and a non-reference loss function system were used to achieve unsupervised adaptive enhancement of low-light, high-noise teaching images. The enhancement parameter prediction network can automatically generate pixel-level light enhancement curve parameters, effectively improving tool texture clarity and color constancy. At the object detection level, the optimized *FocalEIoULoss* and decoupled head structure based on YOLOv8 solved the problem of insufficient detection accuracy caused by varied object shapes and light sensitivity in teaching scenarios, significantly improving small object detection mAP@0.5 and meeting the interaction needs of virtual teaching systems.

The theoretical value of the research lies in: for the first time, combining unsupervised enhancement and lightweight detection models, proposing a visual perception paradigm applicable to vocational skills scenarios, providing a new idea for cross-task collaboration of image enhancement and object detection in the education field; the application value lies in: through technical implementation, the virtual teaching system can realize real-time enhancement and tool localization in operation videos, supporting automatic evaluation and intelligent interaction, and promoting the transformation of vocational skill training from "experience-driven" to "data-driven". However, there are still limitations in the research: First, the enhancement network may still suffer from detail loss in extreme backlight scenes; second, the generalization ability of object detection for rare operational tools needs improvement; third, lightweight deployment of the model on edge computing devices still needs optimization. Future research will be expanded in three aspects: (1) introducing cross-modal fusion to improve robustness under extreme lighting; (2) building a dynamic meta-learning mechanism to enhance rapid adaptability to new category targets; (3) exploring neural architecture search to achieve intelligent matching of model parameters and computing resources, promoting large-scale application of the technology in mobile virtual teaching scenarios.

AUTHOR CONTRIBUTIONS

All authors were equally involved in the conceptualization, methodology, investigation, writing—original draft, and review and editing of the manuscript.

REFERENCES

[1] Venkatesh, K., Reddy, S.K., Angothu, H. (2023). Vocational skill training programs for persons with intellectual disability (PID) and trainers' perspective during and post vocational skill training. Journal of Family Medicine and Primary Care, 12(12): 3142-3148. https://doi.org/10.4103/jfmpc.jfmpc_433_23

[2] Thianthai, C., Sutamchai, K. (2022). Skills that matter: qualitative study focusing on the transfer of training through the experience of Thai vocational students. Frontiers in Education, 7: 897808. https://doi.org/10.3389/feduc.2022.897808

[3] Seaman-Tullis, R.L., Cannella-Malone, H.I., Brock, M.E. (2019). Training a paraprofessional to implement video prompting with error correction to teach a vocational skill. Focus on Autism and Other Developmental Disabilities, 34(2): 107-117. https://doi.org/10.1177/1088357618794914

[4] M. Yusop, S.R., Rasul, M.S., Mohammad Yasin, R., Hashim, H.U. (2023). Identifying and validating vocational skills domains and indicators in classroom assessment practices in TVET. Sustainability, 15(6): 5195. https://doi.org/10.3390/su15065195

[5] Friston, S., Congdon, B., Steed, A. (2022). Teaching social virtual reality with UBIQ. IEEE Computer Graphics and Applications, 42(6): 116-122. https://doi.org/10.1109/MCG.2022.3211729

[6] Gonzalez Lopez, J.M., Jimenez Betancourt, R.O., Ramirez Arredondo, J.M., Villalvazo Laureano, E., Rodriguez Haro, F. (2019). Incorporating virtual reality into the teaching and training of grid-tie photovoltaic power plants design. Applied Sciences, 9(21): 4480. https://doi.org/10.3390/app9214480

[7] Tsekhmister, Y., Konovalova, T., Tsekhmister, B.,

Agrawal, A., Ghosh, D. (2021). Evaluation of virtual reality technology and online teaching system for medical students in Ukraine during COVID-19 pandemic. International Journal of Emerging Technologies in Learning, 16(23): 127-139. https://doi.org/10.3991/ijet.v16i23.26099

[8] Cheng, D. (2017). Application and practice of animation and digital image in interactive media teaching in colleges and universities. Agro Food Industry Hi-Tech, 28(1): 1245-1249.

[9] Maldjian, J.A., Listerud, J. (2000). Automated teaching file and slide database for digital images. American Journal of Roentgenology, 175(5): 1249-1251. https://doi.org/10.2214/ajr.175.5.1751249

[10] Zhou, J., Zhang, B., Zhang, D., Vivone, G., Jiang, Q. (2024). Dtkd-net: Dual-teacher knowledge distillation lightweight network for water-related optics image enhancement. IEEE Transactions on Geoscience and Remote Sensing, 62: 4207213. https://doi.org/10.1109/TGRS.2024.3422667

[11] Hosamani, S., Sonnad, S. (2025). Image quality enhancement using optimized thresholding and two-level diffusion-based denoising filter. The Imaging Science Journal, 73(2): 245-265. https://doi.org/10.1080/13682199.2024.2364532

[12] Hao, K. (2020). Multimedia English teaching analysis based on deep learning speech enhancement algorithm and robust expression positioning. Journal of Intelligent & Fuzzy Systems, 39(2): 1779-1791. https://doi.org/10.3233/JIFS-179951

[13] Zhu, Z., Zia, A., Li, X., Dan, B., et al. (2024). Collaborative static-dynamic teaching: A semi-supervised framework for stripe-like space target detection. Remote Sensing, 17(8): 1341. https://doi.org/10.3390/rs17081341

[14] Zhang, K. (2024). Detection and analysis of student behavior in college labor education courses based on YOLOv5 network. Journal of Computational Methods in Science and Engineering, 24(2): 1057-1069. https://doi.org/10.3233/JCM-247308

[15] Liu, Q., Jiang, R., Xu, Q., Wang, D., Sang, Z., Jiang, X., Wu, L. (2024). YOLOv8n_BT: Research on classroom learning behavior recognition algorithm based on improved yolov8n. IEEE Access, 12: 36391-36403. https://doi.org/10.1109/ACCESS.2024.3373536

[16] Starck, J.L., Murtagh, F., Candès, E.J., Donoho, D.L. (2003). Gray and color image contrast enhancement by the curvelet transform. IEEE Transactions on Image Processing, 12(6): 706-717. https://doi.org/10.1109/TIP.2003.813140

[17] Ablin, R., Sulochana, C.H., Prabin, G. (2020). An investigation in satellite images based on image enhancement techniques. European Journal of Remote Sensing, 53(sup2): 86-94. https://doi.org/10.1080/22797254.2019.1673216

[18] Luo, K., Kong, X., Zhang, J., Hu, J., Li, J., Tang, H. (2023). Computer vision-based bridge inspection and monitoring: A review. Sensors, 23(18): 7863. https://doi.org/10.3390/s23187863

[19] Li, H., Zhang, X. (2019). Target characteristics and correlation detection probability calculation method in photoelectric detection screen testing system. Microwave and Optical Technology Letters, 61(9): 2214-2224. https://doi.org/10.1002/mop.31856

[20] Jha, S.S., Nidamanuri, R.R. (2020). Gudalur spectral target detection (GST-D): A new benchmark dataset and engineered material target detection in multi-platform remote sensing data. Remote Sensing, 12(13): 2145. https://doi.org/10.3390/rs12132145