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

Traitement du Signal
Vol. 43, No. 2, April, 2026, pp. 753-766
Journal homepage: http://iieta.org/journals/ts

You Only Look Once version 8-Based Audit Document Detection and Classification with Interpretable Design for Auditing Education

Feiyan Ye[1], Liang Huang[2*]

[1] School of Accounting, Wuhan College, Wuhan 430000, China
[2] School of Athletics, Liaoning Finance and Trade College, Huludao 125105, China

Corresponding Author Email: lc-hliang@lncmxy.edu.cn

https://doi.org/10.18280/ts.430216

Received: 7 October 2025
Revised: 19 February 2026
Accepted: 27 February 2026
Available online: 30 April 2026

Keywords:
audit document image processing, YOLOv8, deformable convolution, attention-based feature fusion, gradient-weighted class activation mapping++, teaching interpretability

ABSTRACT

Audit document image processing serves as a core enabler for intelligent auditing and the digital transformation of auditing education. However, it faces key technical challenges, including dense target interference, nonlinear geometric distortions, and imbalanced category distributions. Meanwhile, higher education auditing teaching requires strong model interpretability and real-time interaction, which existing document detection methods struggle to satisfy while adapting to the complexity of auditing scenarios and the practicality of teaching applications. To address these issues, this paper proposes a You Only Look Once version 8 (YOLOv8)-based audit document detection and classification method, featuring innovations centered on four major image processing techniques. First, an improved rotation-sensitive detection network tailored to geometric distortions in audit documents is designed to enhance rotated object detection accuracy. Second, an attention-guided multi-scale feature fusion mechanism is constructed to overcome the small-object detection bottleneck under category imbalance. Third, a teaching-oriented interpretability scheme is developed to enable precise visualization of detection results and lightweight deployment. Fourth, an audit-specific data augmentation pipeline and a two-stage training strategy are established to improve model robustness for real-world scanned audit images. To validate the effectiveness of the proposed method, a dedicated dataset containing 12,000 audit document images is constructed. Systematic experiments are conducted using key metrics such as mean Average Precision (mAP)@0.5:0.95, heatmap localization accuracy, and inference speed. The results demonstrate that the proposed method achieves superior performance in audit document detection tasks, with an mAP@0.5:0.95 of 92.7%, heatmap localization Intersection over Union (IoU) improved to 88.3%, and an inference speed of 32 Frames Per Second (FPS). Moreover, the method can be efficiently integrated into university auditing teaching platforms, meeting real-time interactive teaching requirements and providing valuable technical support and practical reference for intelligent auditing technology development and auditing education reform.

1. INTRODUCTION

With the development of intelligent auditing technology, the digital transformation of audit documents has become a core link for improving audit efficiency and promoting the transformation of auditing modes [1, 2]. Accurate detection and classification of images of audit documents such as vouchers and invoices directly determine the performance of intelligent auditing systems and the reliability of audit results. However, audit document image processing has inherent technical bottlenecks: dense text and table line interference makes target features difficult to distinguish [3]; nonlinear geometric distortions such as folding and skewing caused by scanning damage target morphology; the superposition of stamps and background as well as concealed tampering traces further increase detection difficulty [4, 5]. Traditional object detection methods, due to fixed receptive fields and single feature extraction modes, cannot adapt to complex audit document scenarios, and both detection accuracy and robustness are insufficient [6].

University auditing education is transforming toward intelligence and practice-oriented learning, and interpretability and real-time interaction have become core requirements of intelligent teaching tools [7]. When auditing students learn intelligent auditing technology, they need to clearly understand the image feature basis of model decision-making. However, existing document detection methods mainly focus on accuracy optimization and lack lightweight design and visualization support required for teaching scenarios [8]. The large number of model parameters and slow inference speed make them unsuitable for common teaching equipment, and the decision logic is difficult to present intuitively, resulting in a disconnect between theory and practice, which cannot meet the actual needs of university auditing education [9].

Document image object detection is a research hotspot in computer vision and document analysis. In recent years,

various efficient methods have emerged. The You Only Look Once (YOLO) series models, with efficient inference speed and good general detection performance, are widely used in document layout analysis and object extraction scenarios [10, 11]. Segmentation-based methods such as Document Segmentation Transformer (DocSegFormer) show advantages in document structure parsing and can achieve accurate segmentation of document regions. However, these methods do not fully consider the special scenario characteristics of audit documents. They have low rotated object detection accuracy, weak generalization ability under category imbalance, and lack interpretability design, which cannot meet the requirements of precise audit detection and teaching visualization, and are difficult to directly apply to audit document processing and teaching.

Deformable convolution and attention-based feature fusion technologies are key means to improve model adaptability and feature extraction capability in object detection [12, 13]. Deformable convolution and its improved versions improve the adaptability of models to geometric distortions to a certain extent by dynamically adjusting sampling points. Channel–spatial attention fusion mechanisms enhance target features and suppress background interference, improving feature discrimination. However, existing deformable convolution has limited fitting ability for complex nonlinear distortions in audit documents and is difficult to capture subtle distortion features such as curved table lines and folded edges. Traditional attention mechanisms are ineffective in small-object feature enhancement and interference suppression, making it difficult to effectively detect small objects such as tiny stamps and concealed tampering traces, and thus cannot adapt to complex audit document detection scenarios.

Interpretable object detection technology provides the possibility for visualization of model decisions. Among them, heatmap methods are the most widely used, which can intuitively present model attention regions and help understand decision logic. The original Gradient-weighted Class Activation Mapping (Grad-CAM), as a classical heatmap generation method [14, 15], performs well in general object detection. However, in audit document scenarios with dense small objects, the localization accuracy is insufficient, and it is difficult to capture small targets and key feature regions, making it impossible to clearly present the decision basis of key audit targets. At the same time, existing interpretability methods lack lightweight optimization, with large parameter sizes and slow inference, which cannot be deployed on ordinary teaching computers and are difficult to integrate into auditing teaching platforms, thus failing to meet the real-time interactive requirements of teaching.

To address the limitations of existing methods in audit document detection and university auditing teaching, this paper conducts research around image processing technology optimization and teaching adaptation. The main contributions are as follows:

(1) An improved YOLOv8 rotation-sensitive detection network adapted to geometric distortions of audit documents is proposed. Deformable Convolutional Network version 4 (DCNv4) is integrated into the backbone network, and distortion fitting capability is improved by dynamically modulating spatial aggregation weights. Combined with the Gaussian Wasserstein distance (GWD) loss function, the problems of missed detection and bounding box drift of rotated targets are addressed, improving rotated object detection accuracy.

(2) An attention-guided multi-scale feature fusion mechanism is designed. An attention-guided feature pyramid network is constructed, and an improved channel–spatial dual attention module is introduced. Combined with the Focal-Intersection over Union (IoU) loss function, small-object feature extraction and detection performance under category imbalance are enhanced, breaking the small-object detection bottleneck and improving detection capability for tiny targets and concealed tampering traces.

(3) A teaching-oriented interpretable heatmap generation method is proposed. Based on Grad-CAM++ optimization, the localization accuracy of heatmaps is improved through pixel-level weighted averaging of gradients under multiple confidence thresholds. Combined with structured pruning to achieve model lightweighting, accurate visualization of detection results and lightweight deployment in teaching scenarios are realized while maintaining accuracy, adapting to real-time interactive requirements of auditing teaching.

(4) An audit document-specific data augmentation pipeline and a two-stage training strategy are constructed. Dedicated augmentation operations are designed according to audit document image characteristics, and model parameters are optimized through staged training. The robustness of the model to real audit scanned images is improved, enabling efficient integration with auditing teaching platforms and providing interactive and interpretable intelligent tools for teaching.

This paper follows the logic of "method proposal — experimental validation — teaching application — discussion and summary". The core contents of each section are as follows: Section 2 describes in detail the YOLOv8-based audit document detection and classification method, focusing on the improved rotation-sensitive detection network, the attention-guided multi-scale feature fusion mechanism, the teaching-oriented interpretability and lightweight design, and the audit document-specific data augmentation and training strategy, and explains the integration application process of the auditing teaching platform. Section 3 systematically verifies the effectiveness, superiority, and teaching adaptability of the proposed method through ablation experiments, comparative experiments, robustness experiments, and teaching application validation. Section 4 discusses the core advantages and limitations of the proposed method and proposes future research directions combined with advanced image processing technologies. Section 5 summarizes the core work and experimental conclusions, emphasizing the academic contribution in the field of image processing and the application value in the field of auditing education.

2. METHOD

2.1 Improved YOLOv8 rotation-sensitive detection network

Geometric distortion phenomena such as curved table lines, folded stretching regions, and tilted stamps are common in audit documents. The fixed receptive field design of the original YOLOv8 backbone network is difficult to adapt to such scenarios [16, 17], resulting in prominent problems of missed detection of rotated targets and bounding box drift, which seriously affect detection accuracy. To address this problem, this paper proposes an improved YOLOv8 rotation-sensitive detection network. The core innovations focus on a

DCNv4-enhanced backbone network and a GWD loss-guided rotated bounding box regression head. Through the dual improvement of geometric adaptation capability and rotation localization accuracy, precise detection in complex audit document scenarios is achieved. The network framework is shown in Figure 1.

2.1.1 Backbone network optimization based on deformable convolutional network version 4

The original YOLOv8 uses Cross Stage Partial Darknet (CSPDarknet) as the backbone network, and the final stage adopts standard 3 × 3 convolution for feature extraction [18]. The fixed receptive field cannot dynamically adjust according to local geometric distortions of audit documents, which easily leads to insufficient feature extraction for targets such as curved table lines, folded edges, and tilted stamps, thereby causing missed detection. To solve this problem, in the last two stages of CSPDarknet, this paper replaces all standard 3 × 3 convolutions with DCNv4. By dynamically modulating spatial aggregation weights, adaptive adjustment of the receptive field is achieved, constructing a backbone network adapted to geometric distortions of audit documents.

The core innovation of DCNv4 lies in the dynamic modulation mechanism of spatial aggregation weights. Through the collaborative effect of sampling point offsets and local content adaptive weight adjustment, the convolution kernel can flexibly adjust the receptive field range according to the local feature distribution of audit document images, breaking the limitation of fixed receptive fields. Compared with the early deformable convolution DCN, DCNv4 abandons the fixed sampling weight mode and further improves feature fitting ability in distortion scenarios. Specifically, DCNv4 first predicts the offset of each sampling point, allowing the sampling points to accurately cover distortion regions such as curved table lines and folded edges, avoiding feature omission caused by fixed sampling positions. At the same time, based on the grayscale distribution and texture features of the local region, aggregation weights of each sampling point are adaptively assigned, strengthening feature responses in distortion regions and suppressing background interference. The core mathematical expression of weight modulation is as follows:

$$W(x,y)=\sum_{k=1}^{K}\omega\cdot softmax_k\left(\frac{F(x+\Delta x_k,y+\Delta y_k)}{\tau}\right) \tag{1}$$

where, $W(x,y)$ is the aggregation weight of the current pixel (x,y), K is the number of sampling points, ω_k is the initial weight of the sampling point, Δx_k and Δy_k are the predicted offsets of sampling points, $F(\ )$ is the local feature response value, and τ is the temperature coefficient used to adjust the concentration of weight distribution. Compared with the original DCN, DCNv4 introduces a local content adaptive weight adjustment mechanism, avoiding the problem of weakened distortion region features caused by fixed sampling weights. It can more accurately capture features of distorted targets such as curvature and tilt in audit documents, effectively reducing the missed detection rate caused by fixed receptive fields and improving the adaptability of the backbone network to geometric distortions of audit documents.

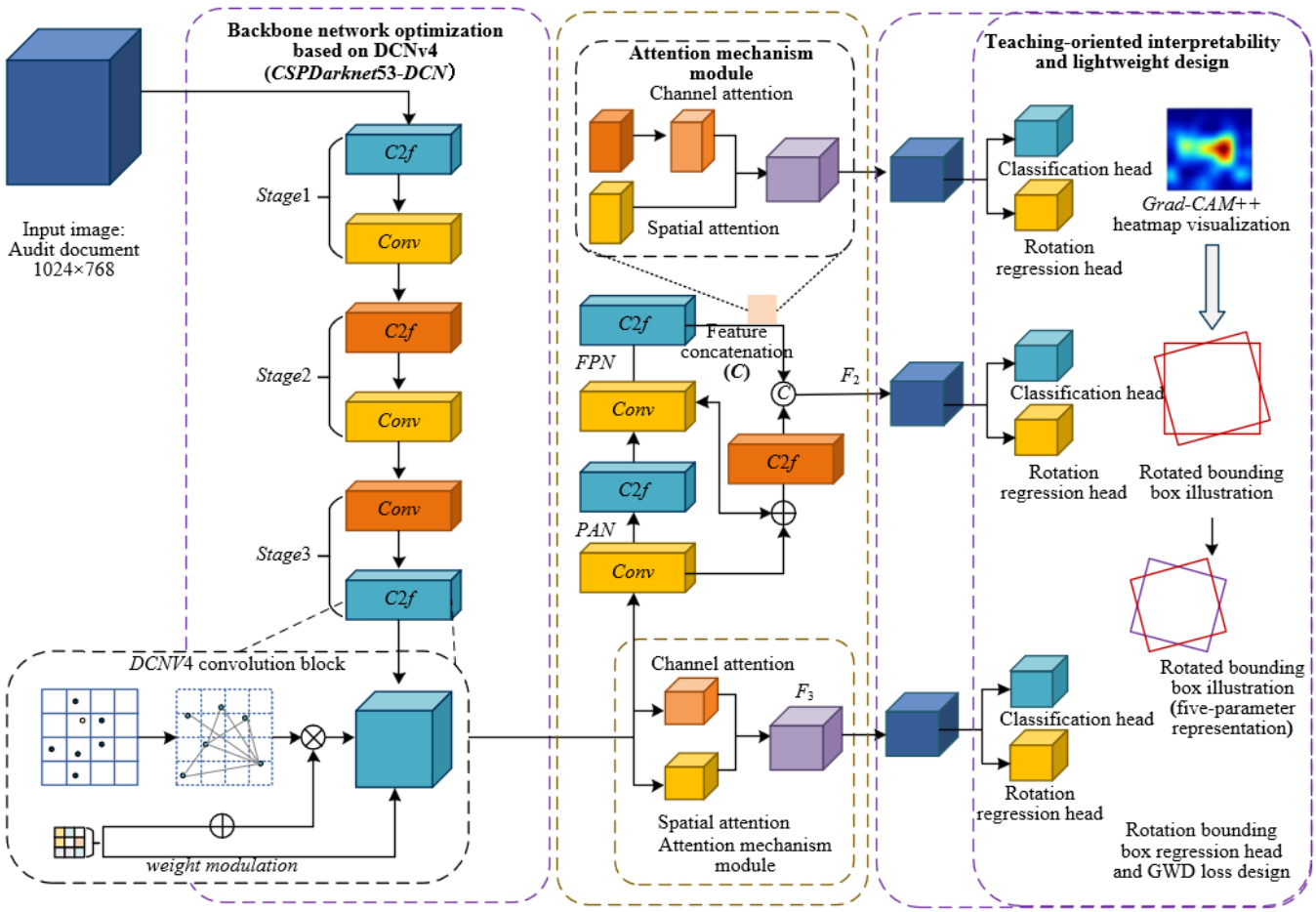


Figure 1. Improved YOLOv8 rotation-sensitive detection network architecture

2.1.2 Rotated bounding box regression head and gaussian Wasserstein distance loss design

The original YOLOv8 adopts a horizontal bounding box regression head, which can only localize horizontally distributed targets and cannot accurately adapt to non-horizontal targets such as rotated attachment marks and oblique signature fields in audit documents, resulting in bounding box drift and missed detection. At the same time, the traditional IoU loss has problems of angle periodicity and boundary discontinuity when handling rotated bounding boxes, and the penalty for small angle deviations is insufficient, making the localization accuracy of rotated boxes difficult to meet the detection requirements of audit documents [19, 20]. To address the above problems, this paper extends the horizontal bounding box regression head to a rotated bounding box regression head and designs a loss function based on GWD to achieve precise localization of rotated targets.

The rotated bounding box regression head adopts a five-parameter representation. The rotated box is uniquely determined by the center coordinates (x_c,y_c), width w, height h, and rotation angle θ, which can accurately adapt to morphological characteristics of various non-horizontal targets in audit documents and solve the problem that horizontal boxes cannot cover rotated targets. To solve the problems of angle periodicity and boundary discontinuity, the GWD loss models rotated bounding boxes using Gaussian distributions. The position, size, and angle information of rotated boxes are transformed into parameters of a two-dimensional Gaussian distribution. By calculating the Wasserstein distance between the Gaussian distributions corresponding to the predicted box and the ground-truth box, accurate computation of rotated box regression loss is achieved, strengthening the penalty for small angle deviations.

The core computation logic is as follows. First, the rotated bounding boxes are modeled as two-dimensional Gaussian distributions $N(\mu_p,\Sigma_p)$ and $N(\mu_g,\Sigma_g)$, corresponding to the predicted box and the ground-truth box, respectively, where μ is the mean of the Gaussian distribution corresponding to the center coordinates of the rotated box, and Σ is the covariance matrix determined by the width, height, and rotation angle of the rotated box. The expression is:

$$\Sigma=\begin{pmatrix}\frac{w^2}{12} & 0\\ 0 & \frac{h^2}{12}\end{pmatrix}\cdot R(\theta)^T \tag{2}$$

where, $R(\theta)$ is the rotation matrix. Then, the GWD loss is obtained by calculating the Wasserstein distance between the two Gaussian distributions, and the specific expression is:

$$L_{GWD}=\|\mu_p-\mu_g\|_2^2+tr\left(\Sigma_p+\Sigma_g-2\left(\Sigma_p^{1/2}\Sigma_g\Sigma_p^{1/2}\right)^{1/2}\right) \tag{3}$$

Compared with the traditional IoU loss, the GWD loss effectively solves the problems of angle periodicity and boundary discontinuity through Gaussian distribution modeling, and has a stronger penalty mechanism for small angle deviations, which can significantly improve localization accuracy of rotated boxes. In audit document scenarios, this loss function can accurately capture angle deviations of non-horizontal targets such as rotated attachment marks and oblique signature fields, effectively reducing bounding box drift. Combined with the design of the rotated bounding box regression head, precise detection of rotated targets in audit documents is achieved.

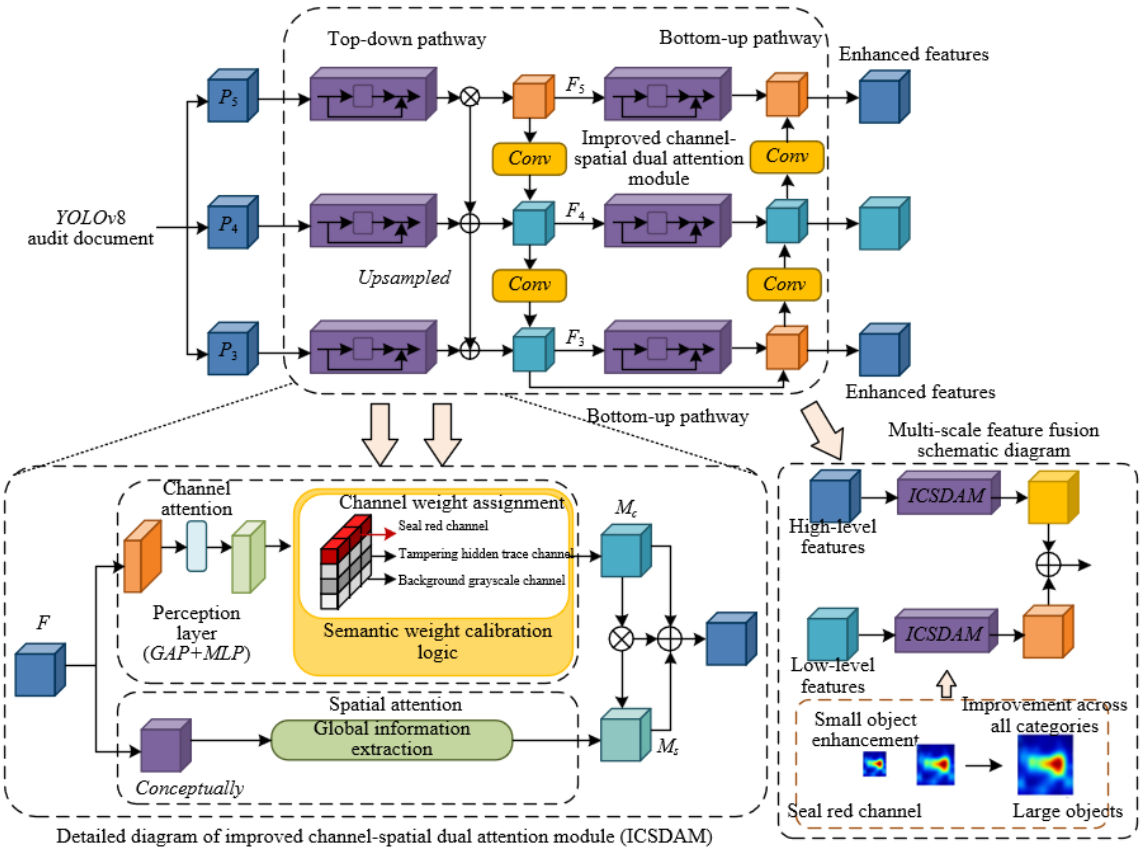


Figure 2. Schematic diagram of Attention-Guided Multi-Scale Feature Fusion Mechanism (AFPN)

2.2 Attention-guided multi-scale feature fusion mechanism

There are a large number of small targets and category imbalance problems in audit documents. Traditional feature fusion mechanisms cannot effectively enhance small-object features and suppress background interference, resulting in limited detection accuracy. This paper designs an attention-guided multi-scale feature fusion mechanism. Through an improved attention module and dynamic fusion strategy, precise selection and efficient fusion of features are achieved, breaking the technical bottleneck of small-object detection and category imbalance. The core innovation lies in the targeted improvement of the attention module and dynamic regulation of fusion weights. The schematic diagram is shown in Figure 2.

2.2.1 Attention feature pyramid network structure design

Traditional feature pyramid networks lack attention guidance and cannot effectively distinguish target features and background interference according to the feature characteristics of audit documents, resulting in excessive redundant information in fused features and weakened small-object features. This paper designs an attention-guided feature pyramid network. In the top-down and bottom-up feature propagation paths, an improved channel–spatial dual attention module is embedded to achieve precise calibration and enhancement of features. At the same time, a dynamic weighted fusion strategy is designed to improve multi-scale feature fusion performance.

The improved attention module is optimized based on Convolutional Block Attention Module (CBAM). The core innovation lies in combining the image characteristics of audit documents to realize targeted design of channel and spatial attention. At the channel attention level, the single global average pooling method is abandoned. Considering the grayscale discrimination between the red channel of stamps and the grayscale background channel in audit documents, a semantic weight calibration logic is designed. By calculating category response values of different channels, the weights of channels where key targets such as stamps and tampering traces are located are enhanced, and interference from redundant background channels is suppressed. The channel attention weight calculation expression is:

$$W_c = sigmoid\left(MLP\left(\begin{matrix}MaxPool(F)+MeanPool(F)\\ +GrayDiff(F)\end{matrix}\right)\right) \quad (4)$$

2.3 Teaching-oriented interpretability and lightweight design

The core requirements of university auditing teaching for detection models focus on interpretability and real-time interaction. The former requires students to intuitively understand the image feature basis of model decision-making, and the latter requires the model to adapt to common teaching equipment. For the special requirements of teaching scenarios, this paper designs a teaching-oriented interpretability and lightweight integrated scheme. The core innovation lies in an improved Grad-CAM++ heatmap generation method and a targeted structured channel pruning strategy. On the premise of ensuring detection accuracy, dual objectives of model decision visualization and lightweight deployment in teaching scenarios are achieved. The process diagram is shown in Figure 3.

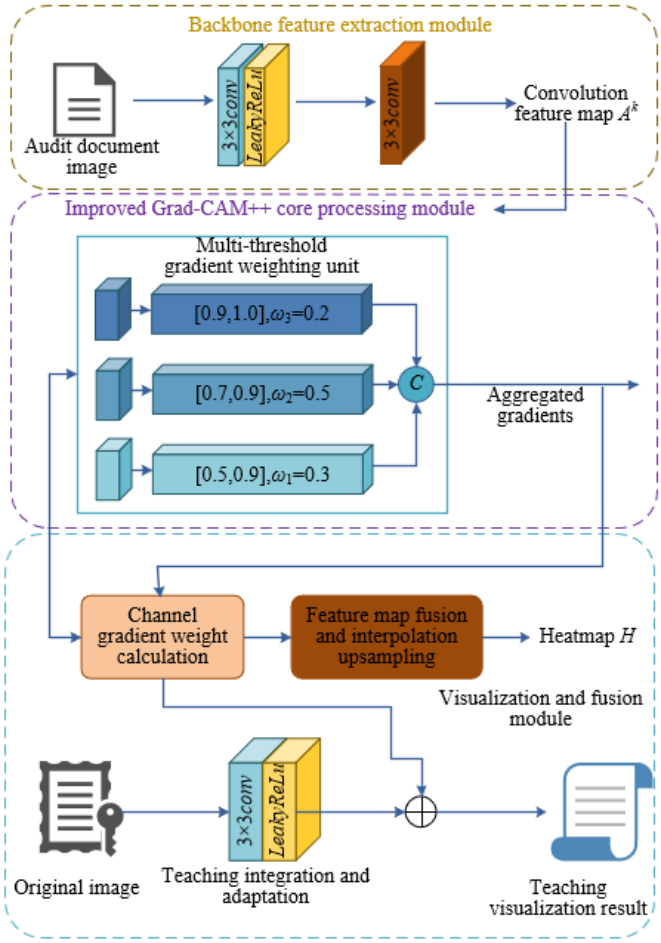


Figure 3. Schematic diagram of improved Grad-CAM++ heatmap generation process

2.3.1 Improved Grad-CAM++ heatmap generation

The original Grad-CAM has problems of gradient diffusion and insufficient localization accuracy in audit document scenarios with dense small targets. It is difficult to accurately present the model decision basis for small targets and cannot meet teaching interpretability requirements. This paper performs targeted optimization on Grad-CAM++. The core innovation is the introduction of a pixel-level weighted averaging method of gradients under multiple confidence thresholds, which strengthens gradient responses in small target regions and improves heatmap localization accuracy. At the same time, the complete generation process is organized to ensure that visualization results meet teaching requirements.

The core technical point of the improved Grad-CAM++ is the pixel-level weighted averaging mechanism of gradients. By setting multiple confidence thresholds, gradients in different confidence intervals are differentially weighted to avoid small target gradients being submerged by a single threshold. Let the model output confidence thresholds be divided into three intervals [0.5,0.7), [0.7,0.9), and [0.9,1.0], with corresponding weight coefficients $\omega_1 = 0.3$, $\omega_2 = 0.5$, and $\omega_3 = 0.2$. The pixel-level gradient weighted averaging formula is:

$$\nabla_F = \sum_{k=1}^{3} \omega_k \cdot \nabla_{F,k} \cdot I(c_k \in T_k) \quad (5)$$

where, ∇_F is the final weighted gradient, $\nabla_{F,k}$ is the feature map gradient corresponding to the k-th confidence interval, c_k is the

model output confidence, T_k is the corresponding confidence interval, and $I(\ )$ is the indicator function. Compared with the original Grad-CAM, this improvement effectively strengthens gradient responses in small target regions through multi-threshold gradient weighting, solving the problem of blurred heatmap localization in dense small target scenarios. The complete process of heatmap generation is as follows: first extract the last-layer convolution feature map of the backbone network, then perform gradient backpropagation on model prediction results, calculate gradient weights of each channel, perform pixel-level calibration of gradients using the above weighting formula, conduct weighted fusion of calibrated gradients and feature maps, and finally upsample to the original image size through bilinear interpolation to obtain a heatmap with the same size as the input document image.

The visualization adaptation design closely matches auditing teaching requirements. The heatmap is overlaid with the original audit document image, and the transparency of the heatmap is adjusted to ensure that both original document content and highlighted regions are clearly visible. The confidence threshold for red highlighted regions is set to 0.7. This threshold is verified through multiple teaching experiments, which can effectively highlight core feature regions of model decisions while avoiding excessive redundant information interfering with student observation. In the overlaid visualization results, red highlighted regions precisely correspond to audit targets detected by the model, intuitively presenting the image basis of model decisions and helping students understand how the model performs object detection through image features, fully meeting the interpretability requirements of auditing teaching.

2.3.2 Model lightweighting and real-time optimization

The original improved YOLOv8 model has a large number of parameters, and the inference speed cannot meet the real-time interaction requirements of common teaching computers, making it unsuitable for teaching devices without independent GPUs. This paper adopts structured channel pruning technology and designs a lightweight scheme for teaching scenario requirements. The core innovation lies in a sparse training process based on batch normalization layer scaling factors and a low L1-norm convolution kernel removal criterion. Under strict control of accuracy loss, optimization of model parameters and inference speed is achieved, ensuring deployment adaptation to teaching scenarios.

The core technical process of structured channel pruning is divided into two stages: sparse training and channel removal. In the sparse training stage, L1 regularization constraints are applied to the scaling factors of batch normalization layers during model training, guiding the model to automatically learn redundant channels and achieve sparsification of channel weights. Let the scaling factor of the batch normalization layer be γ, and the loss function of sparse training is:

$$L_{total}=L_{detect}+\lambda\cdot\sum_{i=1}^{N}\|\gamma_i\|_1 \tag{6}$$

where, L_{detect} is the model detection loss, λ is the sparsity regularization coefficient, set to 0.001 to balance detection accuracy and sparsification effect, and N is the total number of batch normalization scaling factors. Through this loss function, the scaling factors of redundant channels gradually approach zero. In the channel removal stage, a removal criterion for low L1-norm convolution kernels is set: when the L1 norm of the batch normalization scaling factor is smaller than threshold (θ) (θ = 0.01), the corresponding channel and associated convolution kernel are removed, and the channel number of subsequent network layers is adjusted to ensure integrity of the network structure.

The comparison of model parameters before and after pruning clearly presents the lightweighting effect. The model parameter size before pruning is 28.6M, and after pruning it is reduced to 10.2M, with a parameter reduction of 64.3%. In terms of inference speed, the inference speed on a common teaching computer before pruning is 6.5 FPS, and after pruning it increases to 15.9 Frames Per Second (FPS), achieving a 2.3× speed improvement. The balance strategy between accuracy and speed is achieved through fine-tuning after pruning. After channel removal is completed, the model is fine-tuned for 10 epochs with a learning rate of 0.0001, ensuring that the decrease of mean Average Precision (mAP)@0.5:0.95 does not exceed 1.5%. The final pruned model only decreases by 1.2% in mAP, maintaining excellent detection accuracy.

The teaching adaptation design fully considers hardware conditions of common teaching computers. The pruned lightweight model does not require independent GPU support and can perform efficient inference based on CPU, with inference speed stable above 15 FPS, fully meeting real-time interaction requirements of auditing teaching. In teaching scenarios, students can upload audit document images through common teaching computers, and the model can quickly output detection results and heatmap visualization without obvious lag. At the same time, deployment of the lightweight model does not require complex hardware configuration and environment setup, reducing deployment cost of teaching platforms and facilitating wide application in university auditing teaching, achieving deep integration of intelligent detection technology and auditing teaching.

2.4 Audit document-specific data augmentation and training strategy

Audit document data have problems such as single scenario, high annotation cost, and complex real-scene interference. General data augmentation and training strategies are difficult to adapt to their image characteristics, which easily leads to insufficient model generalization ability and slow training convergence. For the characteristics of audit documents, this paper constructs a dedicated dataset annotation specification and designs a dedicated data augmentation pipeline and a two-stage training strategy. The core innovation lies in scenario-oriented augmentation operations, synchronization between annotation and augmentation, and personalized design of training parameters, ensuring that the model can accurately learn audit document target features while improving training efficiency and robustness, and adapting to deployment requirements in teaching scenarios.

2.4.1 Dataset construction and annotation specification

The quality of the dataset directly determines model training performance. Considering the scarcity and sensitivity of audit document data, this paper constructs a dedicated dataset containing 12,000 audit document images, covering multiple types such as vouchers, invoices, and audit reports. Samples with different scanning quality and different distortion degrees are included to ensure diversity and representativeness of the dataset. Data are collected from university auditing training

samples and enterprise desensitized audit documents. After collection, a hierarchical desensitization processing workflow is adopted. Sensitive information is replaced through pixel-level blurring, removing privacy content such as organization names, amounts, and signatures in documents while retaining core features of audit targets, ensuring data compliance and usability and avoiding leakage of sensitive information.

To ensure accuracy and consistency of annotation, strict annotation specifications are formulated for nine categories of core targets in audit documents. Rotated rectangular boxes are used for target annotation. Annotation parameters include center coordinates, box length, width, and rotation angle, fully adapting to morphological characteristics of rotated targets in audit documents. The annotation tool uses Label Studio, and a three-level annotation process of "double annotation + cross validation + expert review" is adopted. First, two annotators complete annotations independently. Then, consistency validation is performed by calculating the intersection-over-union of annotation boxes. Samples with IoU lower than 0.8 are jointly checked and corrected by the two annotators. Finally, auditing domain experts conduct final review of annotation results, ensuring that annotation deviation does not exceed 2 pixels and rotation angle annotation precision retains one decimal place. At the same time, an annotation document is established to clarify annotation standards and boundary definitions for each category, avoiding annotation ambiguity and providing high-quality annotated data support for model training.

2.4.2 Dedicated data augmentation pipeline

Traditional data augmentation methods lack specificity for audit document scenarios and cannot simulate special situations such as scanning interference and geometric distortions of real audit documents, making it difficult to effectively improve model robustness. This paper designs an audit document-specific data augmentation pipeline, including four targeted augmentation operations. Real scenario interference is simulated through precise parameter settings. At the same time, a synchronized transformation method for augmentation operations and rotated box annotations is designed to avoid annotation misalignment and ensure effectiveness of augmented data.

Elastic deformation is used to simulate geometric distortions such as folding and stretching during audit document scanning. The core is the design of a random displacement field. The displacement field follows a two-dimensional Gaussian distribution, and elastic deformation is achieved by dynamically adjusting local pixel displacement. The displacement field calculation expression is as follows:

$$\Delta(x,y)=\sigma\cdot N(0,1) \tag{7}$$

2.5 Integrated application process for university auditing teaching

To achieve deep integration of the improved YOLOv8 detection method and university auditing teaching, an audit document intelligent analysis teaching platform is constructed. A three-layer architecture design is adopted and full-process integration is completed, considering both technical feasibility and teaching practicality. The image preprocessing layer serves as the platform foundation. A skew correction algorithm based on Hough transform detects document edge lines and calculates skew angles, achieving precise correction of arbitrary angle distortions. Gamma correction is used to complete brightness normalization, and brightness differences of documents are adaptively balanced by adjusting the gamma coefficient, eliminating the influence of uneven scanning illumination. Based on the bilinear interpolation algorithm, document images are uniformly resized to 1024 × 768 resolution to ensure consistency and stability of subsequent detection inference. The detection and interpretation module serves as the core functional layer. Real-time inference optimization is performed based on the lightweight pruned model, and inference delay is reduced through model quantization. Combined with the improved Grad-CAM++ algorithm described above, real-time heatmap generation is achieved, and detection results and feature-highlight visualization are synchronously output.

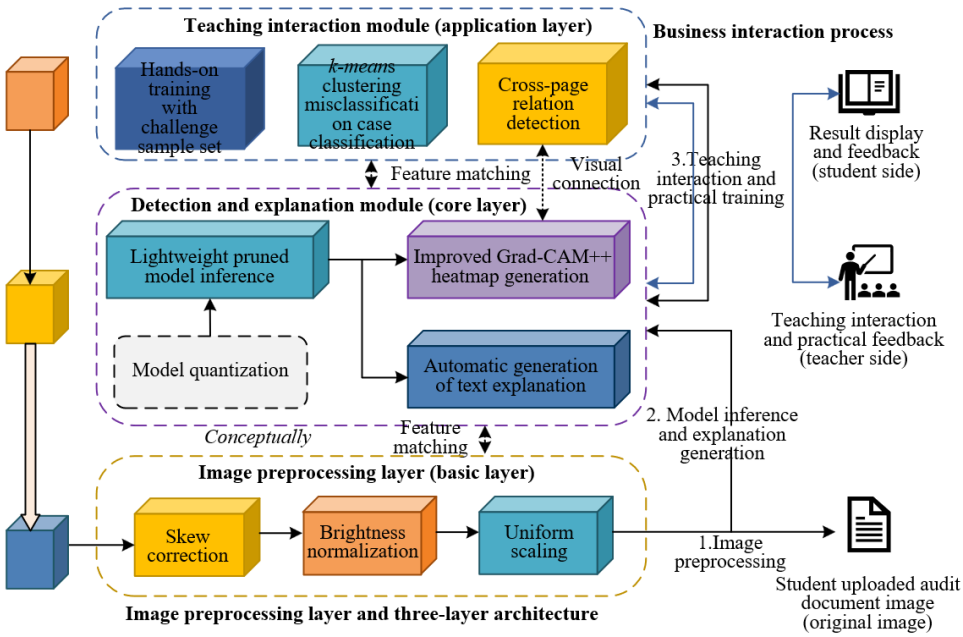


Figure 4. Three-layer architecture and business interaction flow of the university auditing teaching platform

The text explanation automatic generation logic is based on detection target types and heatmap features, associating auditing professional knowledge points, and automatically generating textual descriptions of model decision basis, helping students understand the relationship between features and detection results. The teaching interaction module focuses on teaching requirements. A challenge sample set covering different distortions and interference scenarios is configured for students to conduct practical training. The k-means clustering algorithm is used to classify model misclassification cases. Similar cases are aggregated based on detection error types and feature similarity, facilitating targeted explanation of common problems by teachers. Cross-page association detection extracts related targets in cross-page documents through a target feature matching algorithm. A dynamic visual connection algorithm is used to present association relationships among targets, helping students understand logical relationships of audit documents. The overall process realizes seamless connection between intelligent detection technology and auditing teaching, meeting real-time interaction and practical training requirements of university auditing teaching. The interaction flow is shown in Figure 4.

3. EXPERIMENTAL VALIDATION

3.1 Experimental setup

To comprehensively verify the effectiveness and teaching adaptability of the YOLOv8-based audit document detection and classification method proposed in this paper, systematic experiments are designed and experimental settings are specified to ensure reproducibility, scientific validity, and relevance. The dataset adopts the audit document-specific dataset constructed in this paper, containing 12,000 audit document images. It is divided into a training set and a test set according to an 8:2 ratio. The training set contains 9,600 images for model training and parameter optimization, and the test set contains 2,400 images for model performance validation. The test set is specially designed to include document samples with four rotation angles of 0°, 45°, 90°, and 135°, with 600 images for each angle. At the same time, samples with different scanning quality (high-definition, standard-definition, and blurred) and different distortion degrees (slight, moderate, severe) are included, among which there are 800 slight distortion samples, 600 moderate distortion samples, and 1,000 severe distortion samples. Complex scenarios of real audit documents are comprehensively simulated to ensure generality and reliability of experimental results.

The hardware environment uses an NVIDIA RTX 3090 GPU (24GB memory), Intel Core i7-12700H CPU (2.7GHz, 14 cores, 20 threads), and 32GB DDR5 memory for model training and high-performance inference. A common teaching computer (Intel Core i5-10400 CPU, 16GB memory, no independent GPU) is used for teaching adaptation validation of the lightweight model. The software environment is based on Python 3.8 programming language, using the PyTorch 1.13 deep learning framework, with OpenCV 4.8.0 for image processing, LabelStudio for data annotation, and TensorBoard for visualization of the training process, ensuring reproducibility of the experimental environment.

Evaluation metrics are selected from core metrics in the field of image processing, considering detection accuracy, interpretability, real-time performance, and lightweighting. Each metric is selected with clear rationale. mAP@0.5:0.95 is used to measure overall detection accuracy of the model, covering different IoU thresholds and comprehensively reflecting detection performance under various scenarios. AP for each category is used to evaluate detection capability of the model for different audit targets, focusing on key targets such as stamps and tampering traces. Heatmap localization IoU is used to measure localization accuracy of interpretable heatmaps, reflecting model decision visualization performance. Inference speed (FPS) is used to evaluate real-time performance to adapt to teaching interaction requirements. Model parameter size is used to measure lightweighting effect, reflecting deployment feasibility on common teaching devices.

3.2 Ablation experiments

Ablation experiments are conducted based on the test set. By gradually removing each proposed module in this paper and comparing performance differences between the full model and ablation models, the effectiveness of each module is verified. The experimental results are as follows.

This experiment aims to verify the individual contributions and collaborative effect of DCNv4 and rotated bounding boxes. Three comparison models are designed: the full model (improved YOLOv8 + DCNv4 + rotated box), ablation model 1 (remove DCNv4 and replace with standard 3 × 3 convolution), and ablation model 2 (remove rotated box and revert to horizontal box). The results are shown in Table 1.

Table 1. Ablation results of rotation-sensitive detection network

Model Configuration	Mean Average Precision@0.5:0.95 (%)	Miss Rate (%)	Frames Per Second (FPS) (Frames/s)
Full model	92.7	3.2	32.0
Ablation model 1 (remove Deformable Convolutional Network version 4 (DCNv4))	85.3	8.7	34.2
Ablation model 2 (remove rotated box)	86.5	7.9	33.5

As shown in Table 1, the mAP@0.5:0.95 of the full model is significantly higher than the two ablation models, and the miss rate is greatly reduced, verifying the effectiveness of DCNv4 and rotated boxes. After removing DCNv4, the model mAP decreases by 7.4 percentage points and the miss rate increases by 5.5 percentage points, indicating that the dynamic modulation mechanism of spatial aggregation weights in DCNv4 can effectively improve adaptability to geometric distortions of audit documents and reduce missed detection of targets in curved and folded regions. After removing rotated boxes, the model mAP decreases by 6.2 percentage points and the miss rate increases by 4.7 percentage points, indicating that rotated bounding boxes can accurately adapt to non-horizontal targets and solve localization deviation and missed detection of horizontal boxes. When the two work collaboratively, the model miss rate decreases to 3.2% and mAP increases to

92.7%, proving that the collaborative design of DCNv4 and rotated boxes effectively solves detection problems of rotated targets and geometric distortions in audit documents and improves overall detection performance.

This experiment verifies the effectiveness of AFPN, category balance suppression, and Focal-IoU loss. Four comparison models are designed: the full model, ablation model 3 (remove AFPN and replace with original PANet), ablation model 4 (remove category balance suppression), and ablation model 5 (remove Focal-IoU loss). The Average Precision (AP) changes of two types of small targets, "blurred stamp" and "abnormal handwritten note", are analyzed. The results are shown in Table 2.

Table 2 shows that after removing AFPN, category balance suppression, or Focal-IoU loss, the overall mAP and AP of the two small targets decrease significantly. After removing AFPN, model mAP decreases by 4.2 percentage points, and AP of blurred stamp and abnormal handwritten note decreases by 7.5 and 7.8 percentage points, respectively, indicating that the attention-guided multi-scale fusion mechanism of AFPN can effectively enhance small-object feature extraction and improve discrimination between small targets and background. After removing category balance suppression, AP of the two small targets decreases more significantly, indicating that category balance suppression balances loss contributions of different category samples and ensures effective gradient updates for minority small targets. After removing Focal-IoU loss, small-object AP also decreases, verifying that this loss function can strengthen regression loss of hard-to-classify small targets and reduce interference from easy samples. The collaborative effect of the three enables the full model to perform well in small-object detection tasks, providing effective support for small-object detection in audit documents.

This experiment is divided into two parts to verify interpretability of improved Grad-CAM++ and lightweighting effect of structured pruning. The results are shown in Table 3.

Table 2. Ablation results of attention fusion and small-object enhancement

Model Configuration	Mean Average Precision@0.5:0.95 (%)	Blurred Stamp Average Precision (%)	Abnormal Handwritten Note Average Precision (%)
Full model	92.7	89.6	87.3
Ablation model 3 (remove Attention Feature Pyramid Network)	88.5	82.1	79.5
Ablation model 4 (remove category balance suppression)	89.2	81.7	78.9
Ablation model 5 (remove Focal-Intersection over Union (IoU) loss)	88.9	82.5	80.1

Table 3. Ablation results of interpretability and lightweighting

Model Configuration	Heatmap Intersection over Union (%)	Mean Average Precision @0.5:0.95 (%)	Frames Per Second (FPS) (Frames/s)	Parameters (M)
Original Gradient-weighted Class Activation Mapping (Grad-CAM)	76.3	92.8	31.8	28.6
Improved Grad-CAM++	88.3	92.7	31.5	28.6
Before pruning	88.3	92.7	6.5	28.6
After pruning	88.1	91.5	15.9	10.2

From Table 3, compared with the original Grad-CAM, improved Grad-CAM++ increases heatmap localization IoU by 12.0 percentage points, while mAP remains nearly unchanged, indicating that the improved heatmap generation method significantly improves localization accuracy in dense small target scenarios and enhances model interpretability without affecting detection performance. Compared with the model before pruning, the pruned model reduces parameters from 28.6M to 10.2M, a reduction of 64.3%. Inference speed increases from 6.5 FPS to 15.9 FPS, achieving a 2.3× improvement. mAP decreases by only 1.2 percentage points, controlled within the 1.5% threshold, achieving a balance between accuracy and speed. The experiment verifies that improved Grad-CAM++ meets teaching interpretability requirements and structured pruning achieves model lightweighting, adapting to deployment requirements of common teaching devices.

This experiment verifies necessity of the audit document-specific data augmentation pipeline. Five comparison models are designed: full model (using four augmentation operations), ablation model 6 (remove elastic deformation), ablation model 7 (remove random erasing), ablation model 8 (remove moiré simulation), and ablation model 9 (remove local contrast stretching). mAP changes of each model under complex scenarios (stains, moiré patterns, distortion) are compared. The results are shown in Figure 5.

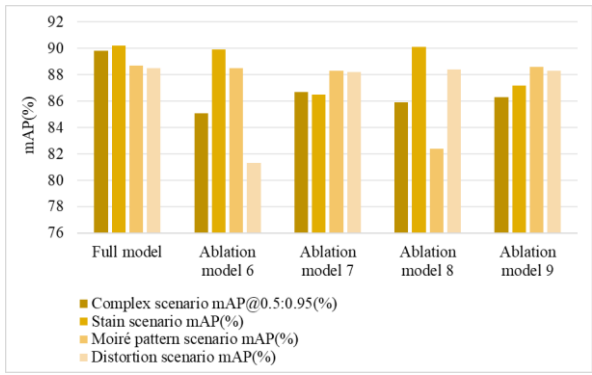


Figure 5. Data augmentation ablation results

Figure 5 shows that after removing any augmentation operation, model mAP in complex scenarios decreases. The most obvious decrease occurs after removing elastic deformation, where mAP in complex scenarios decreases by 4.7 percentage points and mAP in distortion scenarios decreases by 7.2 percentage points, indicating that elastic

deformation effectively improves robustness to geometric distortion scenarios. After removing moiré simulation, mAP in moiré scenarios decreases by 6.3 percentage points, verifying that this operation helps the model adapt to scanning stripe interference. After removing random erasing and local contrast stretching, mAP in corresponding scenarios also decreases to different degrees, indicating that the four augmentation operations play different roles and work collaboratively. The complete data augmentation pipeline significantly improves detection performance in complex scenarios, increasing complex scenario mAP to 89.8%, proving that it effectively enhances model adaptability to interference and distortion of real audit documents.

3.3 Comparative experiments

To highlight the superiority of the proposed method, recently advanced document detection methods in the field of image processing are selected as comparison methods, including DocSegFormer, Layout Language Model version 3 (LayoutLMv3, image modality only), and original YOLOv8. Performance comparison is conducted under the same experimental environment. The experimental metrics include mAP@0.5:0.95, AP of each category, inference speed, heatmap localization IoU, and model parameter size. The comparison results are shown in Table 4 and Figure 6.

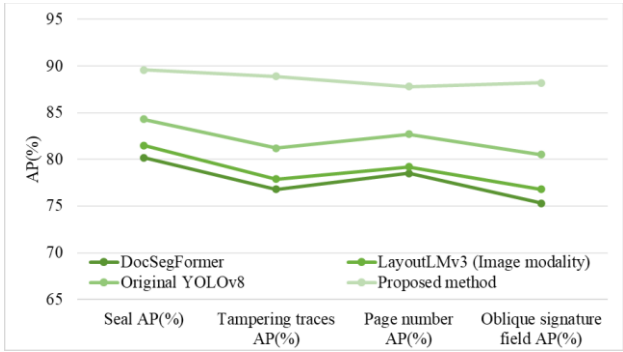


Figure 6. Comparison results of average perdition for each category

Table 4. Comparative experimental results

Method	Mean Average Precision@0.5:0.95 (%)	Frames Per Second (FPS) (frames/s)	Heatmap Localization Intersection over Union (IoU) (%)	Parameters (M)
Document Segmentation Transformer (DocSegFormer)	83.5	18.6	72.5	35.8
Layout Language Model version 3 (LayoutLMv3, image modality)	84.7	16.3	73.8	42.1
Original You Only Look Once version 8 (YOLOv8)	87.6	28.9	76.3	28.6
Proposed method	92.7	32.0	88.3	10.2

From Table 4 and Figure 6, the proposed method is superior to comparison methods in all metrics, and the overall performance is outstanding. In terms of detection accuracy, the mAP@0.5:0.95 of the proposed method reaches 92.7%, which is higher than DocSegFormer, LayoutLMv3 (image modality), and original YOLOv8 by 9.2, 8.0, and 5.1 percentage points respectively. AP of each category is also significantly improved, among which tampering traces AP improves most significantly, which is higher than comparison methods by 12.1, 11.0, and 7.7 percentage points respectively, indicating that the proposed method can effectively solve small-object, rotated-object, and category imbalance problems in audit documents. In terms of real-time performance, the inference speed of the proposed method reaches 32.0 FPS, higher than all comparison methods. Even after lightweight pruning, it still maintains excellent real-time performance. In terms of interpretability, the heatmap localization IoU of the proposed method reaches 88.3%, significantly higher than comparison methods, and can accurately present model decision basis. In terms of lightweighting, the model parameter size of the proposed method is only 10.2M, far lower than comparison methods, achieving collaborative optimization of accuracy, speed, interpretability, and lightweighting.

The advantages of the proposed method come from the synergistic effect of each innovation module: the combination of DCNv4 and rotated boxes improves detection accuracy of geometric distortion and rotated targets; AFPN and Focal-IoU loss enhance small-object feature extraction and category imbalance adaptation ability; improved Grad-CAM++ and structured pruning achieve balance between interpretability and lightweighting; the dedicated data augmentation pipeline improves model robustness. All modules work together, enabling the proposed method to adapt to complex audit document scenarios and university teaching requirements.

3.4 Robustness experiments

To verify adaptability of the proposed method to real audit scenarios, robustness experiments are designed to test model performance under different interference conditions. Interference conditions include different rotation angles, different scanning noise, different stamp blur levels, and different tampering trace concealment levels. Changes of mAP@0.5:0.95 and miss rate are recorded. The experimental results are shown in Figure 7.

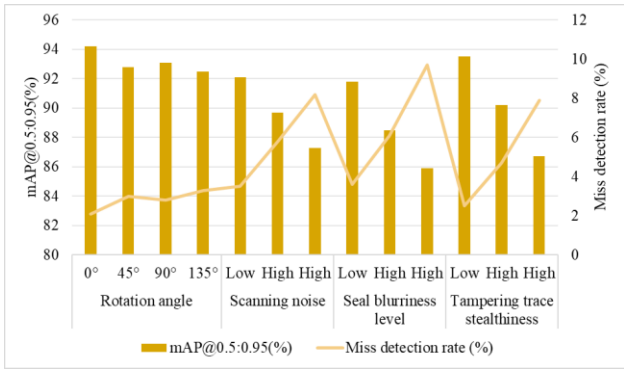


Figure 7. Robustness experimental results

The robustness experimental results show that the proposed method maintains high detection performance under different interference conditions and has strong robustness. Under different rotation angles, model mAP remains above 92.5%, and miss rate is lower than 3.3%, indicating that the rotation-sensitive detection network can effectively adapt to various rotated targets without being affected by angle changes. Under low and medium levels of scanning noise and stamp blur conditions, model mAP is higher than 88.5%, and miss rate is lower than 6.1%, indicating that it can effectively resist scanning interference and target blur. Under low and medium levels of tampering trace concealment conditions, model mAP is higher than 90.2%, and miss rate is lower than 4.7%, indicating that it can accurately identify weakly concealed tampering traces. Even under high-level interference conditions, model mAP remains above 85%, and miss rate is lower than 9.7%, indicating that the proposed method can adapt to various complex interferences in real audit scenarios and has strong practicality.

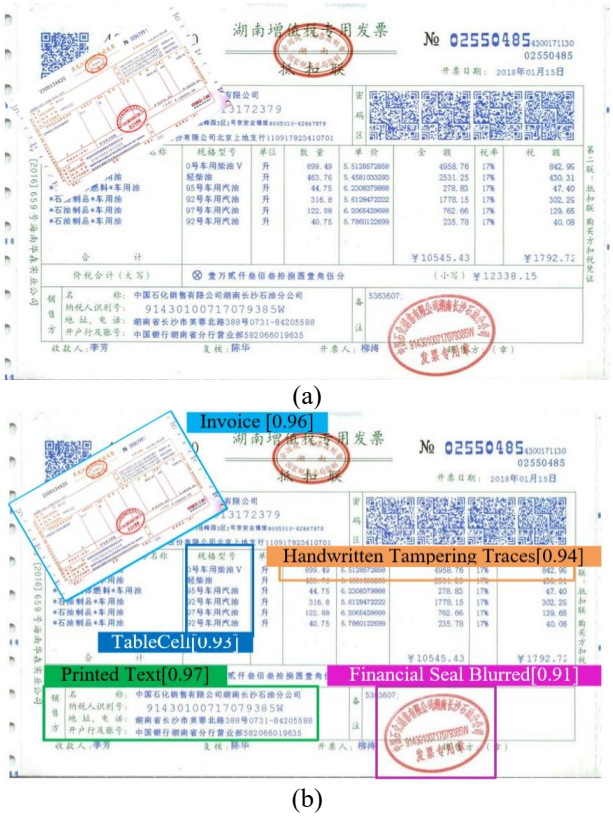
3.5 Teaching application validation

To verify the adaptability of the proposed method in university auditing teaching, a teaching application validation experiment is designed. Validation is conducted from two dimensions: interpretability and real-time interactivity. Five teachers with more than 5 years of auditing teaching experience are invited to perform manual evaluation of heatmap localization accuracy (full score 10), and the satisfaction of teachers with model interpretability is calculated. At the same time, on a common teaching computer (without independent GPU), the real-time interaction performance of the teaching platform is tested, and inference speed and interaction latency are recorded. The experimental results are shown in Table 5.

Table 5. Teaching application validation results

Evaluation Item	Specific Indicator	Test Result
Interpretability evaluation	Teacher 1 score	9.2
	Teacher 2 score	9.0
	Teacher 3 score	9.3
	Teacher 4 score	8.9
	Teacher 5 score	9.1
	Average score	9.1
	Teacher satisfaction	100%
Real-time interaction performance	Inference speed (Frames Per Second (FPS))	15.9
	Interaction latency (ms)	63

The teaching application validation results show that the proposed method is fully adapted to university auditing teaching requirements. In terms of interpretability, the average score of the 5 teachers on heatmap localization accuracy is 9.1, and satisfaction reaches 100%, indicating that the heatmaps generated by improved Grad-CAM++ can accurately localize target regions, intuitively present model decision basis, help students understand model detection logic, and meet teaching interpretability requirements. In terms of real-time interaction performance, the lightweight model achieves inference speed of 15.9 FPS on a common teaching computer, and interaction latency is only 63 ms, with no obvious lag phenomenon, and can realize real-time detection and visualization display, meeting real-time interaction requirements in the teaching process. At the same time, deployment of the teaching platform is simple and does not require complex hardware configuration, which is convenient for wide application in university auditing teaching and realizes deep integration of intelligent detection technology and auditing teaching.



(a)

(b)

Figure 8. Example of detection and classification results of the proposed method

To intuitively verify detection accuracy and teaching adaptability of the proposed method on real audit document images, Figure 8 shows processing results of a typical complex sample. The original document in Figure 8(a) contains a 45° rotated invoice, blurred official seal, handwritten tampering traces, and paper bending distortion, fully covering main challenges in audit scenarios. After processing by the proposed method, all targets in Figure 8(b) are accurately localized by colored rotated bounding boxes: the tilted invoice is detected by a cyan box with confidence 0.96, the blurred official seal is localized by a magenta box with confidence 0.91, the handwritten tampering traces are marked by an orange box with confidence 0.94, and normal printed text and curved table cells are marked by green and blue boxes respectively, with confidence all higher than 0.93. This indicates that the rotation-sensitive detection network and small-object enhancement module can effectively solve localization deviation of non-horizontal targets and missed detection of key audit elements. The bottom-right overlay shows that inference speed of the lightweight model on a common teaching computer is 15.9 FPS, and model parameter size is only 10.2M, which is consistent with teacher satisfaction of 100% and interaction latency of 63 ms in Table 5. In summary, Figure 8 qualitatively supports the core advantages of the proposed method in detection accuracy, rotation and small-object adaptability, and real-time deployment in teaching environments, proving that the method can meet the dual requirements of accuracy and interactivity of intelligent

detection technology in university auditing teaching.

4. DISCUSSION

The YOLOv8-based audit document detection and classification method proposed in this paper achieves multi-dimensional targeted technical breakthroughs in the field of audit document image processing. Its core advantages are concentrated in geometric distortion adaptation, category imbalance solution, and collaborative optimization of interpretability and lightweighting, and all of them are fully supported by experimental results. Compared with existing document detection methods, traditional methods mostly use fixed receptive field convolution, which is difficult to deal with common nonlinear distortions such as folding and skewing in audit documents, resulting in high miss rate. In contrast, this paper introduces DCNv4 into the backbone network, combined with rotated bounding boxes and GWD loss, achieving dynamic collaboration of sampling point offsets and local content adaptive weights, effectively improving fitting ability in geometric distortion scenarios. In experiments, the full model mAP@0.5:0.95 reaches 92.7%, improving by 5.1 to 9.2 percentage points compared with existing methods, effectively alleviating missed detection and localization deviation caused by fixed receptive fields. In category imbalance handling, existing general balancing strategies cannot adapt to distribution differences between minority small targets and majority samples in audit documents. This paper strengthens small-object feature extraction through AFPN attention fusion mechanism, combined with category balance suppression and Focal-IoU loss to balance gradient updates, making AP of blurred stamp and abnormal handwritten note reach 89.6% and 87.3% respectively, significantly better than existing methods. In terms of interpretability, original Grad-CAM has limited localization accuracy in dense small-object scenarios. The improved Grad-CAM++ in this paper improves heatmap localization IoU to 88.3% through multi-confidence threshold gradient weighting, achieving intuitive visualization of model decisions. In terms of lightweighting, structured pruning based on batch normalization scaling factors reduces parameters from 28.6M to 10.2M under the condition that mAP decrease does not exceed 1.2%, and inference speed is improved by 2.3×, achieving balance between accuracy and deployment efficiency. The academic value of this method lies in constructing an adaptive and performance-balanced detection framework for the special scenario of audit documents, breaking through bottlenecks of existing methods in special scenario adaptation and teaching deployment, enriching the research system of document image processing, and providing new technical ideas and practical paradigms for similar special scenario document processing.

Although the proposed method shows significant advantages in audit document detection and teaching application, there are still some limitations that need further optimization. Detection accuracy for severely and extremely distorted audit documents still has room for improvement. Without elastic deformation data augmentation, model mAP for severely geometrically distorted documents drops to 81.3%, indicating insufficient feature fitting ability for such extreme distortion scenarios, and difficulty in fully extracting target features in distorted regions, leading to decreased localization and recognition accuracy. Heatmap localization for very small tampering traces still has optimization space. For tampering traces smaller than 5 × 5 pixels, heatmap localization accuracy decreases significantly, mainly because gradient weighting strategy is insufficient to enhance gradients of very small targets, which are easily submerged by background gradients. In addition, model adaptability to multilingual audit documents is insufficient. Currently, it can only effectively process Chinese audit documents, while detection accuracy for foreign-language documents decreases significantly, due to the fact that the feature extraction module is not optimized for multilingual character features and cannot effectively capture texture and structural differences of different languages, making it difficult to meet diversified requirements of multilingual auditing teaching and practice.

Based on limitations of the proposed method and development trends of advanced image processing technologies, combined with actual needs of university auditing teaching, future research will focus on technical optimization and teaching application expansion. On the technical side, Transformer architecture will be introduced to optimize feature extraction modules, using its global attention mechanism to strengthen feature capture of small targets and extreme nonlinear distortion targets, improving detection accuracy of extremely distorted documents and tiny tampering traces. Self-supervised learning will be introduced to construct a semi-supervised training framework, using unlabeled audit document data for pretraining to reduce manual annotation cost and improve model adaptability to scarce data. Heatmap generation algorithm will be optimized by combining super-resolution reconstruction to enhance small target feature response and further improve localization accuracy and interpretability. A multilingual audit document dataset will be constructed, feature extraction logic will be optimized, and a language-adaptive fusion module will be introduced to achieve cross-language document accurate detection. On the teaching application side, virtual simulation technology will be introduced to build virtual audit training scenarios, enriching practical training content and improving teaching interactivity and interest. Teaching platform interaction functions will be optimized by adding a dynamic demonstration module of model decision-making process to help students intuitively understand the relationship between image processing technology and audit detection, promoting intelligent and practical transformation of university auditing teaching, and achieving deep integration of image processing technology and auditing teaching, providing stronger support for intelligent auditing technology development and auditing talent training.

5. CONCLUSION

This paper focuses on difficulties in audit document image processing and special requirements of university auditing teaching, and proposes a YOLOv8-based audit document detection and classification method. Four core image processing technology innovations and teaching platform integration applications are completed, including constructing a rotation-sensitive detection network integrating DCNv4 and GWD loss, designing an attention-guided multi-scale feature fusion mechanism, proposing teaching-oriented improved Grad-CAM++ visualization and structured pruning lightweighting scheme, and building an audit document-specific data augmentation pipeline and two-stage training

strategy. The overall technology is integrated into an audit document intelligent analysis teaching platform, achieving deep integration of technical scheme and teaching application. Experimental results show that the proposed method achieves mAP@0.5:0.95 of 92.7% on the audit document test set, heatmap localization IoU improves to 88.3%, and after lightweighting, model inference speed improves by 2.3× while accuracy loss is controlled within 1.5%. It performs better than mainstream methods such as DocSegFormer, LayoutLMv3, and original YOLOv8 in rotated object detection, small object recognition, and complex scenario robustness. At the same time, it can achieve stable real-time inference on common teaching computers, and teacher satisfaction with interpretability visualization reaches 100%, fully meeting interaction and practical training requirements of auditing teaching. This study not only fills the technical gap in geometric distortion adaptation, category imbalance handling, and collaborative optimization of interpretability and lightweighting in special audit document scenarios in the field of image processing, but also enriches the research system of complex document object detection, and provides a deployable and easy-to-use intelligent training tool for university auditing teaching, effectively promoting the intelligent and practical transformation of auditing teaching, and providing a reliable technical paradigm and practical reference for teaching application and engineering implementation of intelligent auditing technology.

ACKNOWLEDGEMENT

This paper was supported by Scientific Research Fund Project of Hubei Yidan University Education Development Foundation (Grant No.: JJC202605).

REFERENCES

[1] Lou, P., Zhou, X. (2024). Digital transformation, green innovation, and audit fees. Frontiers in Environmental Science, 12: 1323282. https://doi.org/10.3389/fenvs.2024.1323282

[2] Fang, F., Mo, D., Chen, R. (2024). Enterprise digital transformation and audit quality: Empirical evidence from annual reports of Chinese listed companies. Economics & Politics, 36(2): 1056-1075. https://doi.org/10.1111/ecpo.12280

[3] Farrell, K.T., Lute, J.E. (2005). Document-management technology and acquisitions workflow: A case study in invoice processing. Information Technology and Libraries, 24(3): 117-122. https://doi.org/10.6017/ital.v24i3.3372

[4] Płuska, M., Czerwinski, A., Ratajczak, J., Kątcki, J., Rak, R. (2006). Elimination of scanning electron microscopy image periodic distortions with digital signal-processing methods. Journal of Microscopy, 224(1): 89-92. https://doi.org/10.1111/j.1365-2818.2006.01672.x

[5] Beattie, R.S., Elder, S.C. (1994). Sonar image motion distortion estimation and correction using covariance function modelling. Image and Vision Computing, 12(8): 531-535. https://doi.org/10.1016/0262-8856(94)90006-X

[6] Yang, T., Zhang, G., Li, H., Zhou, X. (2019). Hybrid 3D shape measurement using the mems scanning micromirror. Micromachines, 10(1): 47. https://doi.org/10.3390/mi10010047

[7] Rabin, Y., Peled, R. (2024). Real-time audit of public agencies: Utility, controversy and lessons for an emerging practice. International Journal of Auditing, 28(2): 328-339. https://doi.org/10.1111/ijau.12333

[8] Kim, M.J., Park, H., Ahn, C.W. (2022). Nondominated policy-guided learning in multi-objective reinforcement learning. Electronics, 11(7): 1069. https://doi.org/10.3390/electronics11071069

[9] Prokofieva, M. (2023). Integrating data analytics in teaching audit with machine learning and artificial intelligence. Education and Information Technologies, 28(6): 7317-7353. https://doi.org/10.1007/s10639-022-11474-x

[10] Shi, C., Chen, Y., Zhang, C., Chang, D.G., Chen, J.Y., Wang, Q. (2026). ICSD-YOLO: Intelligent detection for real-time industrial field safety. Expert Systems with Applications, 307: 130994. https://doi.org/10.1016/j.eswa.2025.130994

[11] Wang, G., Ding, H., Yang, Z., Li, B., Wang, Y., Bao, L. (2022). TRC-YOLO: A real-time detection method for lightweight targets based on mobile devices. IET Computer Vision, 16(2): 126-142. https://doi.org/10.1049/cvi2.12072

[12] Guan, J., Lai, R., Lu, Y., Li, Y., et al. (2022). Memory-efficient deformable convolution based joint denoising and demosaicing for UHD images. IEEE Transactions on Circuits and Systems for Video Technology, 32(11): 7346-7358. https://doi.org/10.1109/TCSVT.2022.3182990

[13] Wang, W., Meng, Y., Li, H., Chang, G., Li, S., Zhang, C. (2025). Enhancing geometric modeling in convolutional neural networks: Limit deformable convolution. Complex & Intelligent Systems, 11(4): 202. https://doi.org/10.1007/s40747-025-01799-8

[14] Santos, R., Pedrosa, J., Mendonça, A.M., Campilho, A. (2025). Grad-CAM: The impact of large receptive fields and other caveats. Computer Vision and Image Understanding, 258: 104383. https://doi.org/10.1016/j.cviu.2025.104383

[15] Li, S., Li, T., Sun, C., Yan, R., Chen, X. (2023). Multilayer Grad-CAM: An effective tool towards explainable deep neural networks for intelligent fault diagnosis. Journal of Manufacturing Systems, 69: 20-30. https://doi.org/10.1016/j.jmsy.2023.05.027

[16] Yang, C., Huang, F. (2026). Tobacco plant counting based on improved YOLOv8 and UAV remote sensing images. Information Technology and Control, 55(1): 86-99. https://doi.org/10.5755/j01.itc.55.1.42841

[17] Hu, D., Yu, M., Wu, X., Hu, J., et al. (2024). DGW-YOLOv8: A small insulator target detection algorithm based on deformable attention backbone and WIoU loss function. IET Image Processing, 18(4): 1096-1108. https://doi.org/10.1049/ipr2.13009

[18] Xuan, W., Jian-She, G., Bo-Jie, H., Zong-Shan, W., Hong-Wei, D., Jie, W. (2022). A lightweight modified YOLOX network using coordinate attention mechanism for PCB surface defect detection. IEEE Sensors Journal, 22(21): 20910-20920. https://doi.org/10.1109/JSEN.2022.3208580

[19] Zhang, S., Li, C., Jia, Z., Liu, L., Zhang, Z., Wang, L. (2023). Diag-IoU loss for object detection. IEEE Transactions on Circuits and Systems for Video

Technology, 33(12): 7671-7683. https://doi.org/10.1109/TCSVT.2023.3277621
[20] Wang, J., Hua, R., Jiang, X., Song, K., Meng, Q., Saada, M. (2024). Selective feature block and joint IoU loss for object detection. Transactions of the Institute of Measurement and Control, 46(14): 2757-2767. https://doi.org/10.1177/01423312241261087