Journal of Computer Science and Artificial Intelligence
ISSN: 3078-8242 (Print), ISSN: 3078-2309 (Online) | Vol. 7, No. 2, 2026

Visible Watermark Detection Based on an Improved YOLOv8 Model

Ying Xiao

Southwest Minzu University, Chengdu, Sichuan, China

Abstract: With the widespread dissemination of digital images on the internet, issues related to image copyright protection and content security have become increasingly prominent, making visible watermarks a common method of copyright marking. To address the frequent inaccuracies in visible watermark detection under complex background conditions, this paper proposes an improved deep learning-based visible watermark detection model that achieves high-precision detection and localization of watermark regions. structural optimization is conducted based on the YOLOv8 object detection framework. Considering the characteristics of visible watermarks, such as large scale variations, uneven transparency, and strong coupling with background textures, an attention mechanism module is introduced into the backbone network to enhance feature representation of critical regions. In addition, a watermark feature enhancement module is designed to strengthen multi-scale feature fusion. Furthermore, the improved YOLOv8 model is fused with the RT-DETR model, combining the local feature extraction capability of convolutional neural networks with the global modeling ability of the Transformer architecture to improve detection accuracy and localization robustness. Experimental results demonstrate that the proposed method achieves an mAP50 of 98.9% and an mAP50:95 of 97.6% on the LVW dataset, outperforming multiple mainstream detection models and effectively balancing detection accuracy with efficiency.

Keywords: Visible Watermarks, Deep Learning, Object Detection.

1. Introduction

With the rapid development of internet technology and the comprehensive transformation of the electronic information industry, human society is undergoing a profound digital revolution. The widespread adoption of smart devices and wearable technology has led to the comprehensive digitization of people's lives and work. In this process, the models of information production, dissemination, and consumption have undergone a fundamental shift. In particular, the rise of self-media platforms such as Douyin and Xiaohongshu has significantly lowered the barriers to content creation and dissemination. Vast amounts of multimedia content are flooding into cyberspace at an exponential rate. Among this content, images—with their visual appeal, high information density, and low dissemination costs—have become the primary medium for information exchange and content dissemination.

The ease of image dissemination and the vulnerability of copyright protection have become a prominent contradiction in the digital age, with the protection of creators' legitimate rights and interests becoming increasingly critical. In an open online environment, unauthorized image theft, tampering, and illegal distribution are commonplace. A large number of criminals profit by stealing others' creative works, severely infringing upon the intellectual property rights of original creators. Consequently, digital watermarking technology has emerged as a vital tool in the field of image copyright protection [1]. Since Braudaway [2] and others first introduced visible watermarks for digital image protection, visible watermarking technology has become a hot research topic in the field of image processing. Meng, Chang, and others further extended this technology to the field of video streaming [3], while Kankanhalli and Ramakrishna introduced the Discrete Cosine Transform (DCT) [4]. By calculating the DCT coefficients of specific image regions to dynamically adjust the watermark embedding intensity and combining this with the texture sensitivity characteristics of the human visual system, they significantly improved the visual quality of watermarked images.

In the digital age, the vast quantities of images uploaded by users each day typically contain visible watermarks, posing a significant challenge to traditional, manually assisted watermark detection and removal methods. The scale of image data on the internet has far exceeded the capacity of manual processing, and traditional methods struggle to meet the practical requirements of real-world applications in terms of processing efficiency, real-time performance and restoration quality. It is therefore necessary to develop an efficient, automated and accurate algorithm for detecting visible watermarks.

2. Deep Learning-Based Object Detection Algorithms

Deep learning-based object detection algorithms have overcome the limitations of feature extraction in traditional methods. By utilizing neural networks, deep learning approaches can automatically learn hierarchical features of objects, enabling precise localization and classification. This provides an efficient solution for detecting objects with fixed morphological characteristics, such as visible watermarks in digital images.

These algorithms are primarily divided into two categories: one is two-stage detection algorithms, represented by Faster R-CNN. Their design approach involves first generating candidate target regions using a Region Proposal Network (RPN) [5], followed by feature extraction and classification regression on these candidate regions. These algorithms offer high detection accuracy and can effectively handle scenarios where watermarks are intricately intertwined with background textures, but they incur relatively high

computational costs; The other category consists of single-stage detection algorithms, including the YOLO series [6] and SSD [7]. These algorithms eliminate the candidate region generation step and instead use neural networks to perform dense sampling and prediction across the entire image, achieving an end-to-end detection process. They offer the advantages of fast detection speed and strong real-time performance, making them suitable for scenarios with high demands on detection efficiency. Furthermore, anchor-free detection algorithms that have emerged in recent years (such as CenterNet [8] and FCOS [9]) directly predict the coordinates of key points or bounding boxes of the target, thereby avoiding the cumbersome anchor design steps found in traditional algorithms and further enhancing detection flexibility and generalization capabilities [8, 9]. The primary advantage of these deep learning-based object detection algorithms lies in their ability to automatically learn distinctive features of watermarks, such as texture, shape, and color. This effectively overcomes issues such as background interference, watermark deformation, and scale variations, laying a solid technical foundation for the accurate detection of visible watermarks.

Ultralytics released YOLOv8 in 2023[10], which has become the core reference algorithm for visible watermark detection tasks due to its modular architecture and robust ecosystem. YOLOv8 divides the model into three independent modules—the backbone, neck, and head—facilitating flexible replacement and improvement; it replaces the C3 module in YOLOv5 with a C2f module to enhance feature fusion efficiency; The Neck adopts a PAN structure to optimize feature propagation paths. Its network architecture is shown in Figure 10. YOLOv8 offers an anchor-free detection mode, which eliminates the preprocessing step of anchor clustering and improves the model's generalization ability. Additionally, it employs VFL loss for classification and CIoU loss for regression, enhancing the model's adaptability to small objects and imbalanced samples [10]. Building upon detection, the model can be extended to support tasks such as instance segmentation, keypoint detection, and classification, thereby enabling one-stop multi-task learning. Furthermore, YOLOv8 offers multiple variants ranging from v8n (ultra-lightweight) to v8x (high-performance), allowing adaptation to different deployment scenarios, such as real-time watermark detection on mobile devices or large-scale image processing on servers. Since YOLOv8 features a comprehensive open-source toolchain, it supports rapid hyperparameter tuning and pre-trained weight transfer, thereby reducing the training cost for watermark datasets. It also supports multiple deployment formats, including ONNX and TensorRT, facilitating final engineering implementation. Therefore, this study selected YOLOv8 as the base network and integrated it with RT-DETR (Real-Time Detection Transformer). RT-DETR is an end-to-end Transformer-based detection model that balances detection accuracy with inference speed. Unlike the YOLO series, which relies on an anchor-based detection paradigm with NMS post-processing, this model leverages a hybrid encoding structure and a dynamic query mechanism to effectively capture global image features. By eliminating redundant post-processing operations, it directly predicts both the target class and the bounding box.

3. Visible Watermark Detection Based on the YOLOv8 Model

3.1. Dataset

Datasets form the foundation for training deep learning models and evaluating their performance; their quality directly impacts the models' generalization ability and detection accuracy. For the visible watermark detection task, this paper utilizes the LVW (Large-scale Visible Watermark) dataset [11] as the core training and testing data. This dataset is a commonly used public dataset in the field of visible watermark research, covering various types of visible watermarks in real-world scenarios. It effectively supports the training of models to detect watermarks of different forms and in different scenarios.

The LVW dataset was constructed for large-scale visible watermark detection and removal tasks. It comprises real-world images collected from public online sources, such as news websites, social media platforms, and commercial stock photo libraries, covering diverse scenes including natural landscapes, urban architecture, portraits, and product still lifes, thereby offering exceptional scene diversity. The dataset comprises a total of 60,000 images, containing 80 types of watermarks from companies, organizations, and individuals, featuring various styles such as Chinese text, English text, and logos, with 750 images corresponding to each watermark type. To ensure the generality and usability of the image data, the original watermark-free images were sourced from the publicly available PASCAL VOC 2012 dataset [12]. PASCAL VOC is an image dataset widely used in computer vision research, containing a rich variety of natural scenes and object categories. Utilizing the visible watermarking principle proposed by Tali Dekel [13], each watermark was embedded into the 750 original images with varying sizes, positions, and intensities, as shown in Equation 1:

$$J(p)=\alpha(p)W(p)+(1-\alpha(p))I(p) \quad (1)$$

$p = (x, y)$ denotes the pixel position, α is the intensity with which the watermark is applied to the image (a decimal value in the range [0, 1]), W is the watermark template, I is the original image, and J is the image with the watermark applied.

3.2. Model Structure

The network architecture of YOLOv8 is shown in Figure 1. It primarily consists of three components: the backbone network, the feature fusion network (Neck), and the detection head [10]. These modules work in concert to achieve an end-to-end mapping from the raw image to the object detection results.

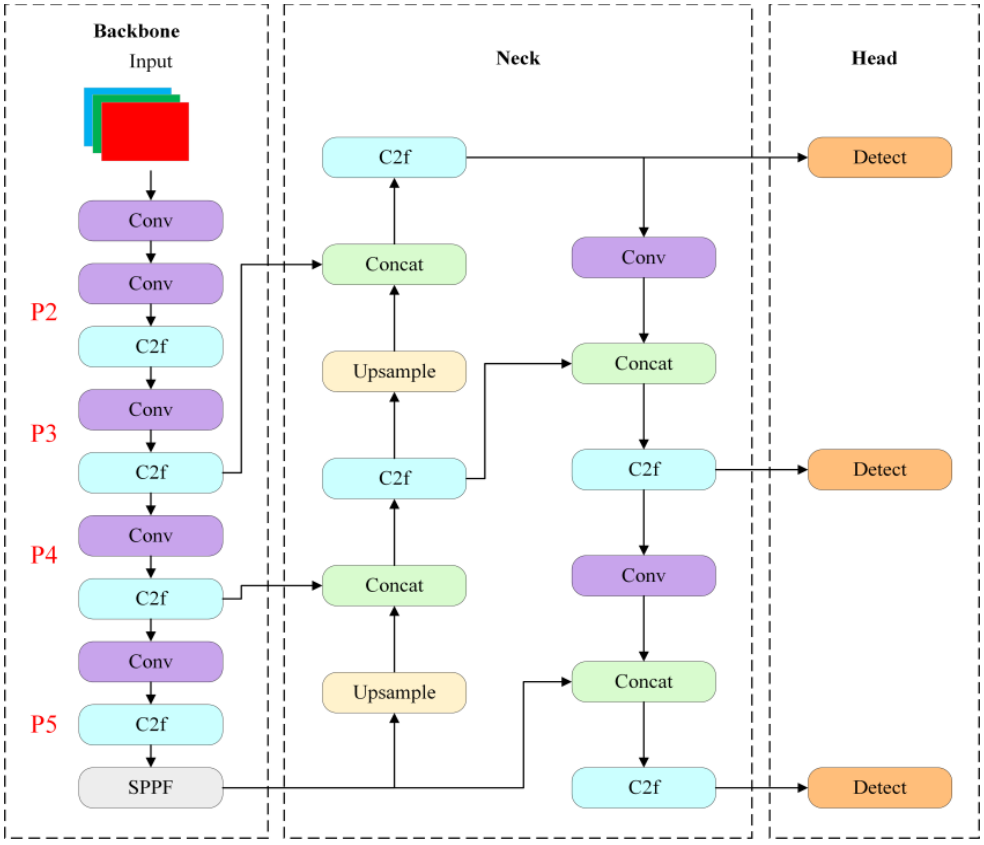

Figure 1. YOLOv8 Network Structure Diagram

In the Backbone section, YOLOv8 employs a hierarchical convolutional structure to extract features from the input image, obtaining feature representations at different resolutions through progressive downsampling. Shallow-layer features primarily capture edge and texture information, whilst deep-layer features focus more on the semantic information of objects. This hierarchical feature extraction approach lays the foundation for subsequent multi-scale object detection, helping the network maintain strong feature representation capabilities in complex scenes. The Neck section is responsible for multi-scale feature fusion, combining feature maps from different stages of the Backbone to enhance feature representation capabilities. The Neck section primarily comprises three major components: the SPPF module (Spatial Pyramid Pooling Fast) for multi-scale pooling operations; the PAA module (Probabilistic Anchor Assignment) for intelligent anchor box allocation; and the PAN module (Path Aggregation Network) for path aggregation of features across different levels [10]. The Head section is responsible for the final object detection and classification tasks, comprising a detection head and a classification head: the detection head contains a series of convolutional and deconvolutional layers to generate detection results; the classification head employs global average pooling to classify each feature map, outputting the probability distribution for each class.

YOLOv8 demonstrates excellent detection performance and high computational efficiency in general object detection tasks, providing a reliable foundational framework for various object detection applications. However, as the network is primarily designed for objects with clear semantic and distinct visual features, there remains room for further optimisation in its feature representation and localisation capabilities when dealing with detection targets such as visible watermarks, which are weakly salient, highly texture-dependent, and exhibit significant scale variations. Therefore, this paper incorporates the characteristics of visible watermarks to introduce corresponding improvements to the original YOLOv8 architecture. The improved YOLOv8 architecture is shown in Figure 2.

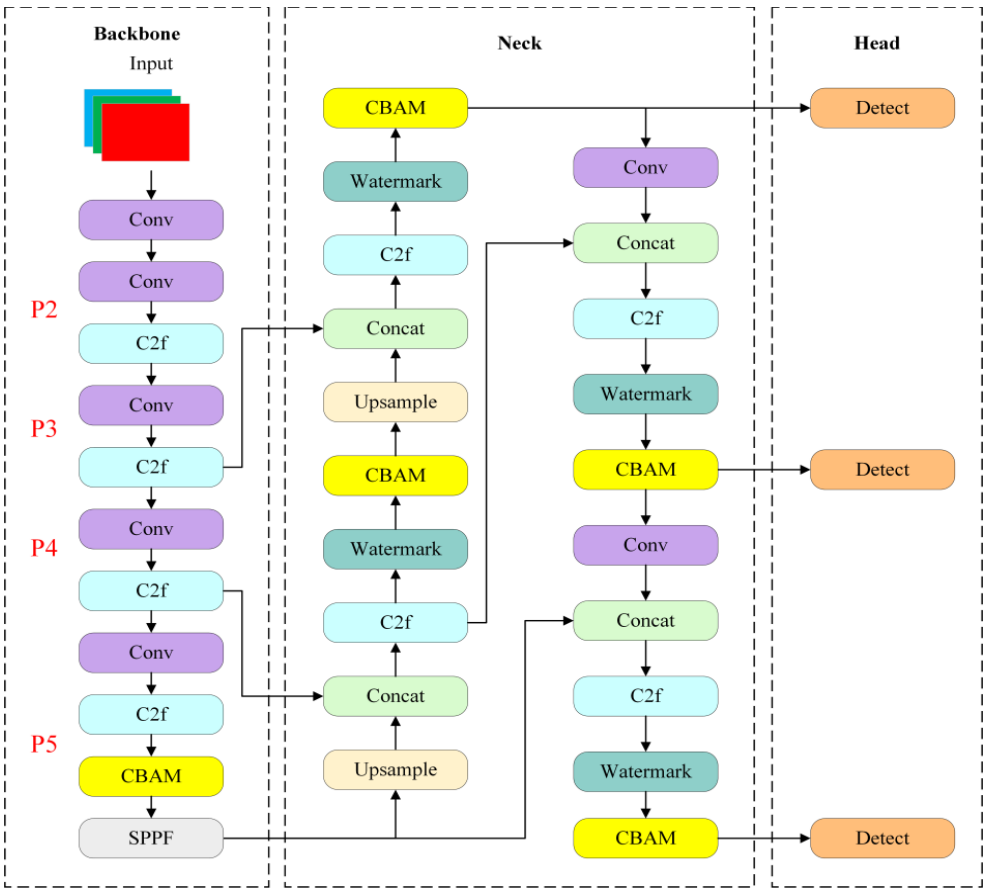

Figure 2. Improved YOLOv8 Network Structure

4. An Improved Method for Visible Watermark Detection Based on YOLOv8

4.1. Addition of a Convolutional Block Attention Module

This paper incorporates a Convolutional Block Attention Module (CBAM) into the original YOLOv8 network. Through a dual-guidance mechanism comprising Channel Attention (CA) and Spatial Attention (SA), during the forward pass, the input features are first processed separately via channel-weighted and spatially-weighted attention. The results of these two attention mechanisms are then fused and combined with the original input features via a residual connection, ultimately producing an enhanced feature representation. This approach strengthens the model's ability to focus on watermark features whilst suppressing interference from redundant background information, thereby enhancing feature representation capabilities.

The spatial attention module primarily focuses on learning the distribution of importance across different spatial locations within the feature map, in order to highlight the positions of areas that may contain watermarks. Given that watermarks typically occupy only localised regions of an image and their spatial distribution is not fixed, spatial attention is introduced to enhance the model's ability to localise watermark regions. In the spatial attention module, the features are first subjected to average pooling and max pooling operations in the channel dimension to obtain two types of spatial statistical features, and the two spatial feature maps are concatenated in the channel dimension; the resulting fused feature map is then passed through a 7×7 convolutional layer, and the convolved feature map is further processed via the Hardswish activation function to generate a spatial attention weight map, the calculation of which can be expressed as:

$$M_s(F)=\text{Hardswish}\left(g\left(\left[\text{AvgPool}_c(F),\ \text{MaxPool}_c(F)\right]\right)\right) \quad (2)$$

In particular, AvgPoolc and MaxPoolc denote pooling operations along the channel dimension, whilst g represents the spatial convolution mapping function. The channel attention module employs a simplified single-layer convolution, with Hardswish replacing the ReLU and Sigmoid activation layers, thereby ensuring greater structural stability.

To further enhance the stability of the module, this paper also performs eigenvalue pruning during the forward propagation of the two sub-modules to prevent extreme eigenvalues from interfering with the learning of attention weights; simultaneously, to avoid potential information bias issues caused by the serial attention structure, this paper employs a method of parallel computation for channel attention and spatial attention, performing a weighted sum of the results from both attention mechanisms with a weighting coefficient set to 0.5; furthermore, on this basis, a residual connection is introduced to superimpose the enhanced features with the original input features. This approach not only ensures the effectiveness of attention enhancement but also preserves the original feature information, alleviating the vanishing gradient problem during model training, thereby improving the stability and convergence speed of the network training. The final output feature is:

$$x_{cbam}=0.5\times(x_{ca}+x_{sa})+x_{identity} \quad (3)$$

x_{ca} represents the output of the channel attention submodule, x_{sa} represents the output of the spatial attention

submodule, and $x_{identity}$ represents the original input features.

4.2. Addition of a Multi-Scale Feature Enhancement Module

It is evident that watermarks in images typically exhibit characteristics such as varying scales, weak edge features, and a high degree of coupling with the background. During the feature extraction and enhancement stages of digital watermarking, single-scale feature processing methods often have limitations, leading to issues such as missed detections and inaccurate localisation. To address this, this paper designs a multi-scale watermark feature enhancement module within the YOLOv8 network. This module uses dilated convolutions to adaptively learn different receptive fields within the feature maps, whilst ensuring that no resolution is lost during the expansion of the receptive fields; by integrating edge-enhanced features with attention-enhanced features, it enables the joint modelling of the watermark's key edge structures and salient regions, thereby enhancing the network's ability to detect and localise watermark targets at different scales. The structure of the multi-scale feature enhancement module is shown in Figure 3.

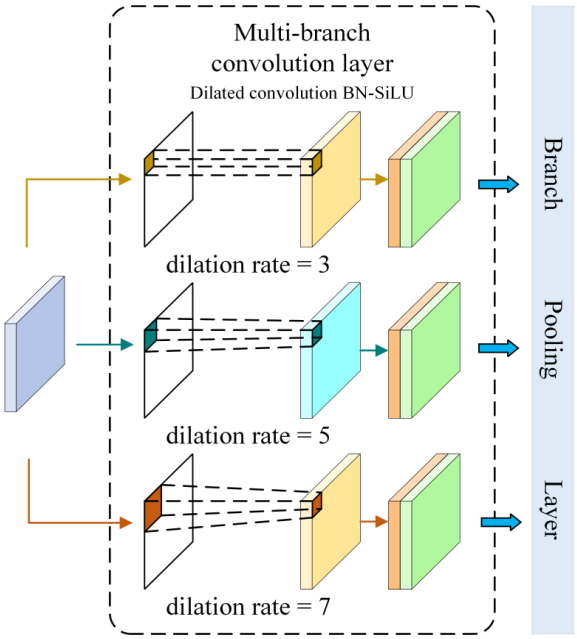

Figure 3. Watermark Enhancement Module Structure Diagram

The multi-scale feature enhancement module is embedded between the base feature extraction layer and the attention processing layer, forming a pipeline. By constructing a multi-branch scale-aware structure, it performs multi-scale feature extraction on the input watermark feature map: the input feature map is fed into branches with dilated convolutions of varying expansion ratios and standard convolutions with different receptive fields, thereby enabling the hierarchical capture of fine-grained edges, local area features, and large-scale watermark features. In standard convolution, the elements of the convolution kernel are adjacent, whereas in dilated convolution, the elements of the convolution kernel are non-adjacent, and the spatial size depends on the dilation rate. The formula for the receptive field of dilated convolution is:

$$r_1 = d \times (k-1) + 1 \tag{4}$$

$$r_n = d \times (k-1) + r_{n-1} \tag{5}$$

k represents the expansion ratio and n represents the stride of the convolution operation.

Once feature extraction is complete across all scale branches, multi-scale feature information is integrated through channel-wise concatenation; batch normalisation is then applied to eliminate differences in feature distributions between the various scale branches; the SiLU activation function is used to enhance the non-linear expressive power of the features; and finally, residual connections are employed to preserve the original feature information.

4.3. A Multi-Model Fusion-Based Visible Watermark Detection Method

Single-object detection models are prone to false negatives or unstable localisation in complex watermarking scenarios. This paper proposes a multi-model fusion watermark detection method based on an improved YOLOv8 and RT-DETR. Compared to CNN-based detection models, RT-DETR utilises the global attention mechanism of the Transformer Encoder to capture long-range dependencies, making it more suitable for detecting weak watermarks that resemble background textures [14]. By employing a prediction box filtering and fusion mechanism based on an IoU threshold, this approach integrates the real-time advantages of YOLOv8 with the global feature capture capabilities of RT-DETR. This not only preserves the inference efficiency of single-model approaches but also addresses the detection shortcomings of single models in complex scenarios, thereby providing a new solution for high-precision real-time object detection.

To minimise the interference of redundant bounding boxes on the fusion results, this paper employs the non-maximum suppression principle to perform an initial screening of the prediction boxes output by individual models: first, the trained modified YOLOv8 model and the RT-DETR model are used separately to detect watermark targets in the input image; if a model outputs only one prediction box, that box is retained directly as the model's valid detection result for the target; if the model outputs multiple prediction boxes, these are sorted by confidence from highest to lowest, and only the box with the highest confidence is retained, thereby avoiding the impact of duplicate detection boxes for the same object on the fusion calculation; as the RT-DETR model employs an end-to-end Transformer architecture, there is no need to use NMS, and the final watermark detection result can be output directly. When both models output valid prediction boxes, the intersection-to-union ratio of the prediction boxes is calculated to determine whether they represent duplicate detections of the same object, thereby enabling result fusion. The calculation formula is:

$$\mathrm{IoU} = \frac{\mathrm{Area}\left(B_y \cap B_r\right)}{\mathrm{Area}\left(B_y \cup B_r\right)} \tag{6}$$

By refers to the detection results from the improved YOLOv8 model, whilst Br refers to those from the RT-DETR model. A threshold is set; if the IoU exceeds this threshold, the prediction box with the highest confidence is retained; otherwise, all detected prediction boxes are retained to expand the coverage of object detection.

5. Experimental Results and Analysis

The configuration remained consistent across all experiments. The CPU used was an AMD Ryzen 9 8945HX with Radeon Graphics, the GPU was an NVIDIA GeForce

RTX 4090 24G, and the development platform was PyCharm 2023 Professional.The training parameter settings are shown in Table 1.

Table 1. Watermark Detection Network Training Parameters

parameter	parameter value
batch	64
max_batches	50200
learning_rate	0.001
momentum	0.9
policy	steps

To validate the effectiveness of the proposed improvement modules in the visible watermark detection task and to analyse the specific impact of different improvement strategies on model performance, systematic ablation experiments were designed and conducted. Using the original YOLOv8 as the baseline model, the experiments sequentially incorporated the CBAM attention module, the multi-scale feature enhancement module, and the RT-DETR fusion strategy to construct four comparison models. These were trained and tested under the same experimental environment and dataset to analyse the impact of each module on detection accuracy and real-time performance. The model configurations are shown in Table 2.

Table 2. Model Configuration for Each Group in the Ablation Experiment

Category	Model configuration
M0 (baseline)	The original YOLOv8 model
M1	YOLOv8+CBAM
M2	YOLOv8+CBAM+Watermarkenhancement
Category	Model configuration
M3	YOLOv8+CBAM+Watermarkenhancement +RT-DETR

The performance metrics of the four model groups on the test set are shown in Table 3.

Table 3. Performance Metrics of Each Model

Category	mAP50	mAP50:95
M0 (baseline)	0.945	0.934
M1	0.957	0.945
M2	0.973	0.962
M3	0.989	0.976

As can be seen from Table 3, the model's detection performance shows an upward trend as each improved module is successively incorporated. Compared with the baseline model M0, the M1 model—which incorporates the CBAM attention module—achieves a 1.27% increase in mAP50 and a 1.18% increase in mAP50:95, thereby validating the positive optimising effect of the CBAM module proposed in this paper on the watermark detection performance of YOLOv8. The core reason lies in the fact that the CBAM module, through the synergistic interaction of channel attention and spatial attention, is able to adaptively enhance the weights of watermark feature channels, suppress interference from redundant background information, and simultaneously focus precisely on the watermark regions within the image. This alleviates the original YOLOv8 model's inadequacy in capturing faint watermark features against complex backgrounds, thereby improving the localisation accuracy and recognition stability of watermark targets.

To comprehensively evaluate the performance of the improved YOLOv8 multi-model fusion detection network proposed in this paper, this section presents comparative experiments. By systematically comparing the proposed method with current mainstream object detection algorithms, we validate its advantages in terms of detection accuracy, localisation capability and overall performance. Several representative object detection models were selected for comparison. To ensure the fairness and comparability of the experimental results, all models were trained and tested using the same dataset split, a uniform number of training iterations, the same input resolution, and consistent hardware environments, with performance evaluated using standardised metrics. A comparison of the F1 score is shown in Figure 4, and a comparison of the mAP50 and mAP50:95 metrics is shown in Figure 5.

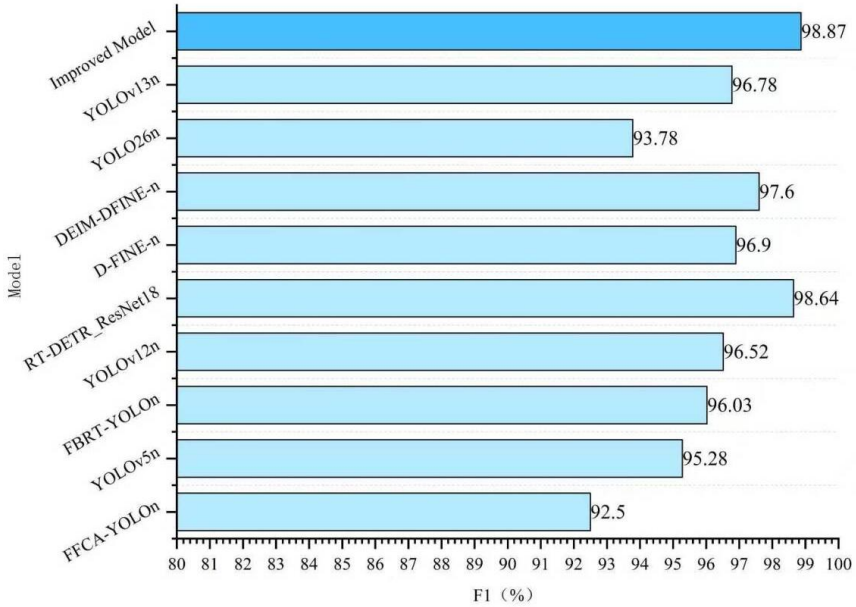

Figure 4. Comparison Chart of F1 Values Across Models

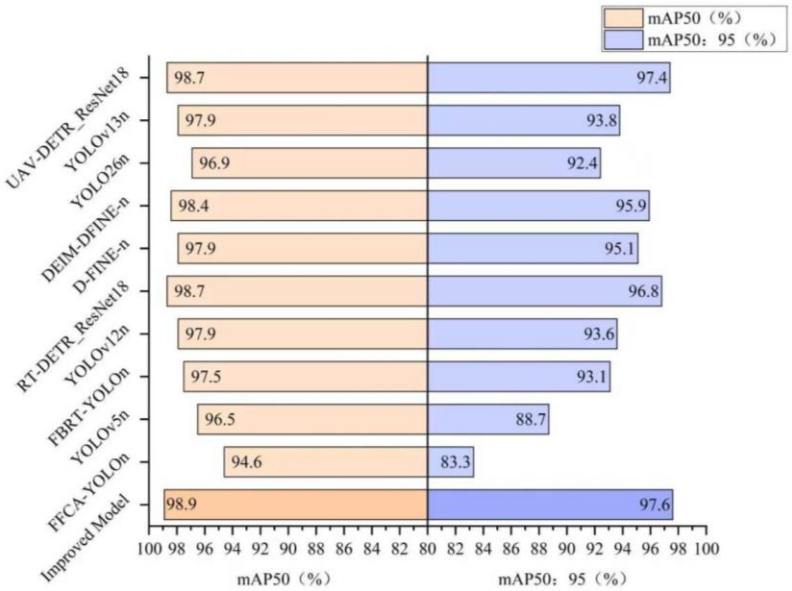

Figure 5. Comparison of mAP50 and mAP50:95 Values Across Models

Analysis of the bar charts above reveals that the fusion model proposed in this paper achieves the best performance on both core accuracy metrics—mAP50 and mAP50:95—with mAP50 reaching 98.9% and mAP50:95 reaching 97.6%. These results are significantly superior to those of the other comparison models, demonstrating the model's exceptional target recognition capability and scale robustness in the task of visible watermark detection.

6. Conclusion

This paper introduces an attention mechanism module into the basic YOLOv8 network architecture to enhance the network's focus on features in key regions; simultaneously, a watermark feature enhancement module is designed to strengthen the fusion capability between shallow and deep features, enabling the model to capture watermark edge and texture information more effectively, thereby improving the detection capability for small-scale and low-contrast watermarks; To address the limitations of convolutional neural networks in global modelling, this paper fuses the improved YOLOv8 model with the RT-DETR model. By integrating local detail features with global semantic information, this fusion achieves a further improvement in detection performance. Furthermore, this chapter validates the effectiveness of the improved modules through ablation experiments and demonstrates, via comparative experiments, that the proposed method is highly competitive in terms of accuracy metrics and overall performance.

References

[1] Zhang, J. R. (2015). The study of digital image watermark algorithm based on transformation domain. Laser Journal, 36(6), 126–129.

[2] Braudaway, G. W., Magerlein, K. A., & Mintzer, F. C. (1996). Protecting publicly available images with a visible image watermark. In Electronic imaging: Science & technology (pp. 126–133). International Society for Optics and Photonics.

[3] Meng, J., & Chang, S. F. (1998). Embedding visible video watermark in the compressed domain. In 1998 International Conference on Image Processing (ICIP 1998) (Vol. 1, pp. 474–477). IEEE.

[4] Kankanhalli, M. S., & Ramakrishnan, K. (1999). Adaptive visible watermarking of images. In 1999 IEEE International Conference on Multimedia Computing and Systems (pp. 568–573). IEEE.

[5] Ren, S. Q., He, K. M., Girshick, R., & Sun, J. (2016). Faster R-CNN: Towards real-time object detection with region proposal networks. arXiv:1506.01497v3.

[6] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. arXiv:1506.02640v5.

[7] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., & Reed, S. (2016). SSD: Single shot multibox detector. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 21–37). IEEE.

[8] Duan, K., Bai, S., Xie, L., Qi, H., Huang, Q., & Tian, Q. (2019). Centernet: Keypoint triplets for object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 6569–6578).

[9] Tian, Z., Shen, C., Chen, H., & He, T. (2019). Fcos: Fully convolutional one-stage object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 9627–9636).

[10] Kukartsev, V. V., Ageev, R. A., Borodulin, A. S., & Boyko, A. A. (2024). Deep learning for object detection in images: Development and evaluation of the YOLOv8 model using Ultralytics and Roboflow libraries. In Proceedings of the 13th Computer Science Online Conference (pp. 629–637). Springer.

[11] Cheng, D., Li, X., Li, W. H., & Wang, H. (2018). Large-scale visible watermark detection and removal with deep convolutional networks. In Chinese Conference on Pattern Recognition and Computer Vision (pp. 31–43). Springer.

[12] Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., & Zisserman, A. (2010). The Pascal Visual Object Classes (VOC) challenge. International Journal of Computer Vision, 88, 303–338. https://doi.org/10.1007/s11263-009-0275-4

[13] Dekel, T., Rubinstein, M., Liu, C., & Freeman, W. T. (2017). On the effectiveness of visible watermarks. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 6864–6872). IEEE. https://doi.org/10.1109/CVPR.2017.726

[14] Zhao, Y., et al. (2024). DETRs beat YOLOs on real-time object detection. In 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 16965–16974). IEEE. https://doi.org/10.1109/CVPR52733.2024.01605