Steel Surface Defect Detection Algorithm Based on Dual Attention Mechanism Fusion with YOLOv12

Yajie Ouyang[a], Neng Wen[b], Miaoling Miao[c]

Jiangxi University of Finance and Economics, Nanchang, 330013, China

[a]2202303937@stu.jxufe.edu.cn, [b]3225339794@qq.com, [c]ndml111@icloud.com

Abstract. To address the strong reliance on traditional manual vision and the frequent missed detection of subtle defects against complex backgrounds in steel surface defect detection, this paper proposes an improved detection algorithm—CA-ViT-YOLOv12—based on dual attention mechanism fusion. Built upon the YOLOv12 framework, the algorithm innovatively integrates a Coordinate Attention (CA) module and a Vision Transformer (ViT) module. The CA module dynamically captures critical features across spatial and channel dimensions, enhancing precise localization of minute flaws; the ViT module leverages self-attention to establish global semantic dependencies, effectively overcoming the limited receptive field bottleneck of conventional convolutional neural networks and improving robustness under multi-texture interference and illumination variation. Experiments on a dataset comprising 5,000 real-world steel surface images from diverse scenarios show that the CA-ViT-YOLOv12 fusion model achieves a mean Average Precision (mAP@0.5) of 0.809 across all categories—significantly outperforming baseline methods—and provides reliable foundational algorithmic support for automated, high-precision quality inspection in steel production lines.

Keywords: Steel defect detection, YOLOv12, Computer vision, Coordinate attention

1. Introduction

In today's globalized and highly advanced industrial manufacturing systems, steel—serving as a foundational and strategic core material—continues to play an irreplaceable supporting role. Surface quality of steel is not only a key indicator of metallurgical process capability but also directly affects the safety performance and service life of products across numerous downstream applications, including machinery, automotive, aerospace, and shipbuilding. During complex production processes such as rolling, cooling, and transportation, steel surfaces inevitably develop various defects—including cracks, scratches, rust spots, pits, and mill scale—due to equipment aging, fluctuations in process parameters, and non-ideal production environments. Historically, surface defect inspection in industrial settings has heavily relied on manual visual inspection or traditional measuring tools. However, this conventional inspection paradigm suffers from inconsistencies arising from individual differences in inspector experience, visual fatigue caused by prolonged high-intensity work, and variable lighting conditions in workshop environments—both subjective and objective factors that lead to highly inconsistent inspection standards, low efficiency, and frequent false positives and missed detections. As noted in relevant industrial literature and industry pain points, the low efficiency and poor stability of traditional manual inspection can no longer meet the stringent quality control requirements of modern high-speed, fully automated production lines. Accordingly, metal surface quality inspection urgently needs to shift from experience-driven to data-driven approaches to fully overcome the subjectivity inherent in manual inspection.

With the deepening of Industry 4.0 concepts and continuous advancement of intelligent manufacturing technologies, computer vision- and deep learning–based intelligent recognition of steel surface defects has rapidly emerged as a key research focus for both academia and industry worldwide. At the algorithmic evolution level, convolutional neural networks (CNNs), leveraging their strong local feature extraction capability, have long dominated industrial vision inspection. Researchers globally have achieved high recognition accuracy for typical steel surface defects—including cracks, pitting, and scratches—by introducing lightweight network architectures and multi-scale feature fusion mechanisms. Notably, the YOLO series of algorithms reformulates object

detection as a single-stage regression problem, significantly boosting inference speed. Taking the YOLOv12 framework as an example, its backbone adopts the Residual Efficient Layer Aggregation Network (R-ELAN), which integrates deep convolution with residual connections to effectively reduce computational load while preserving high-level spatial perception capability; its Neck incorporates the FlashAttention regional attention mechanism, further enhancing feature aggregation efficiency in cluttered scenes. The YOLOv12 architecture is illustrated in Figure 1.

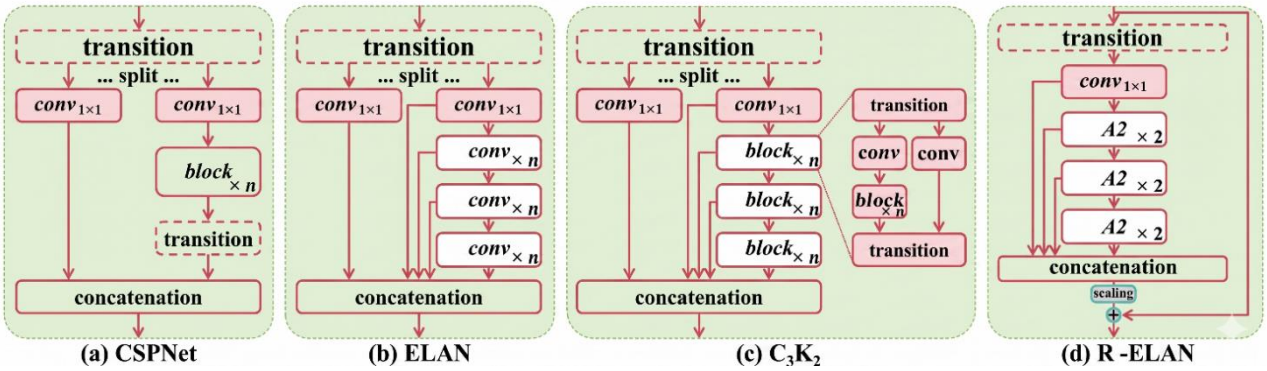

Figure 1. YOLOv12 Architecture

Although CNN-based YOLO series algorithms show great potential in industrial quality inspection, they still exhibit significant performance bottlenecks when confronting the extremely complex real-world steel production environment. First, traditional visual detection has long suffered from the "low accuracy, limited speed, and missing details" problem. Second, steel plate surfaces in industrial settings often feature severe background interference such as glare and mill scale—visual features of which closely resemble genuine defects like "rust spots" and "scratches", easily causing model confusion and misjudgment. More critically, due to their inherent local receptive field limitation, conventional convolutional networks can only focus on local textures and struggle to establish semantic dependencies among features at a global scale. When confronted with densely distributed multi-defects or long, slender cracks spanning large areas, CNN models relying solely on local features often fail to achieve comprehensive and precise detection. To overcome this bottleneck, cutting-edge research is increasingly incorporating attention mechanisms and Transformer architectures to enable global feature modeling and high-precision small-object detection in complex scenarios.

Addressing the core pain points in industrial quality inspection and the limitations of existing algorithms, this paper proposes an improved defect detection algorithm CA-ViT-YOLOv12 based on dual attention mechanism fusion. Building upon the YOLOv12 architecture, which balances high accuracy and low latency, our approach innovatively integrates the Coordinate Attention (CA) mechanism and Vision Transformer (ViT) deeply into the feature extraction and processing network. Specifically, the CA module performs 1D global pooling on the feature tensor across both spatial and channel dimensions, dynamically and precisely capturing key feature regions along vertical and horizontal directions—preserving fine-grained positional information while enhancing feature representation for minute flaws. Meanwhile, the ViT module leverages self-attention to directly model global image context, successfully breaking through the receptive-field limitation of conventional CNNs, thereby enabling the model to maintain exceptional recognition stability under complex backgrounds, drastic illumination changes, and multiple texture interferences.

The main contributions of this paper are as follows:

1. CA-ViT-YOLOv12 employs the CA module to dynamically capture key features across spatial and channel dimensions, enhancing the model's precise localization capability for minute flaws.

2. CA-ViT-YOLOv12 employs the ViT module to construct global semantic dependencies via self-attention, effectively overcoming the receptive-field limitation of conventional CNNs and improving model robustness under multi-texture interference and illumination variation.

3. A novel dual-attention fusion network architecture is proposed, integrating the CA module with the ViT module to achieve complementary local fine-grained features and global contextual information, significantly improving detection performance in complex scenarios.
4. Experiments on real-world datasets validate the superiority of the proposed method.

2. Related Research

Steel surface defect detection is a critical task in industrial quality inspection. Historically, it has relied heavily on manual visual inspection or traditional measuring tools. These approaches are significantly affected by subjective factors such as inspector experience variation and visual fatigue, and suffer from low detection efficiency and high false-negative/positive rates—making them inadequate for modern high-throughput automated production lines. Early automated detection methods predominantly employed conventional image-processing–based machine vision algorithms (e.g., edge detection, morphological operations), yet these exhibit poor robustness under complex textures and varying illumination. Metal surface quality inspection urgently needs to shift from experience-driven, handcrafted-feature–based approaches to data-driven ones to fully overcome the subjectivity inherent in manual inspection. Moreover, conventional image-processing methods—including edge detection, threshold segmentation, and texture analysis perform adequately only in simple backgrounds; they lack robustness, yield high miss-detection rates, and suffer from low efficiency when confronted with complex textures, illumination variations, background interference, or minute defects.

With the rapid advancement of deep learning, convolutional neural network (CNN)–based object detection algorithms have gradually become mainstream. Early studies primarily adopted two-stage detectors for instance, Faster R-CNN [1], which uses a Region Proposal Network (RPN) followed by classification and regression to localize and identify defects—achieving good accuracy on steel defect datasets (e.g., NEU-DET), but suffering from poor real-time performance, thus failing to meet the high-throughput requirements of production lines. Wang et al. [2] instead adopted an improved Cascade R-CNN to address challenges posed by high reflectivity on metal surfaces and strong similarity between defects and background. However, due to their large number of parameters and high computational complexity, two-stage algorithms often fail to satisfy the real-time online inspection demands of high-throughput steel production lines.

To balance detection accuracy and inference speed, one-stage regression algorithms represented by YOLO[3] have gradually become the core technical approach for industrial quality inspection. YOLO directly partitions an image into a grid and outputs bounding boxes and class probabilities, enabling end-to-end fast detection. To address multi-scale, small-object, and complex-background challenges in steel surface defect detection, researchers have made extensive improvements to YOLO . Li et al.[4] proposed STE-YOLO, optimizing feature extraction and fusion modules based on YOLOv8 to enhance detection capability for typical defects (e.g., cracks, scratches, rust spots). Zhou et al.[5] introduced a multi-path attention mechanism, proposing MPA-YOLO to further improve YOLOv8's accuracy in steel defect detection. Zhang et al.[6] proposed LAM-YOLOv10n based on YOLOv10, improving small-defect detection under complex backgrounds via latent-space attention and multi-scale fusion modules.

Introducing attention mechanisms is a key direction for enhancing model capability in detecting minute defects and suppressing interference. Wang et al.[7] integrated coordinate attention (CA) into YOLOv5, embedding positional information along horizontal and vertical axes in the channel dimension, effectively improving feature extraction for strip steel surface defects; they further optimized detection performance using a decoupled head. CASI-Net, proposed by Li et al.[8], similarly leverages the CA module to embed positional information, improving precise localization of defect regions.

To overcome the limited local receptive field of traditional CNNs, Transformer architectures and self-attention mechanisms have been progressively introduced into steel defect detection. Fan et al.[9]

were among the first to apply Vision Transformer (ViT) to steel plate surface defect detection, modeling global dependencies via self-attention. Liu et al.[10] proposed a Transformer-based framework for global contextual modeling of steel defects in complex scenes. Wu et al.[11] designed DSAT, which employs dynamic sparse attention windows and hierarchical feature fusion, achieving breakthroughs in minute-defect recognition while controlling computational cost. Li et al.[12] proposed a sparse global attention Transformer that balances global modeling and computational efficiency via sliding-window sparse self-attention.

As research advances, hybrid architectures have become a current research hotspot, simultaneously leveraging CNNs' strength in local feature extraction and Transformers' capability in global semantic modeling. MSFT-YOLO, proposed by Guo et al. [13], integrates Transformer modules and BiFPN into YOLOv5 to enhance multi-scale feature fusion. Wu et al. [14] proposed SH-DETR, a DETR-based model that employs Transformer self-attention encoders to handle extremely complex industrial backgrounds; additionally, Pan et al. [15] introduced DEENet (a dual-encoder model), which fuses CNNs' edge-enhancement capability with Transformers' global contextual awareness, achieving exceptionally high accuracy in real-world industrial scenarios.

Although existing methods have achieved significant progress in accuracy and speed, they still suffer from the following limitations: First, they exhibit high miss rates for extremely small and densely distributed defects, and lack robustness under strong illumination, glare, and multi-texture interference; furthermore, their fusion of global contextual modeling and local fine-grained feature extraction remains inefficient, and most models incur high parameter counts or computational complexity, making direct deployment on edge devices challenging. As the latest version of the YOLO series, YOLOv12 introduces mechanisms such as the R-ELAN backbone and FlashAttention, yielding baseline performance improvements—but its adaptability to complex industrial steel surface scenarios still requires further optimization.

In summary, steel surface defect detection is evolving from single-CNN approaches toward attention-enhanced, Transformer-integrated, and hybrid architectures. The CA-ViT-YOLOv12 algorithm proposed in this paper innovatively fuses Coordinate Attention (CA) with ViT modules, aiming to improve detection accuracy and robustness for minute defects in complex scenes through complementary local position-sensitive feature enhancement and global semantic dependency modeling.

3. Model

To achieve high-accuracy and high-robustness defect detection on complex industrial steel surfaces, the proposed CA-ViT-YOLOv12 model comprises four key modules in its overall architecture: a backbone feature extraction network, a Coordinate Attention mechanism module, a Vision Transformer (ViT) global modeling module, and a multi-scale feature fusion and detection head. The model's structure is illustrated in Fig. 2

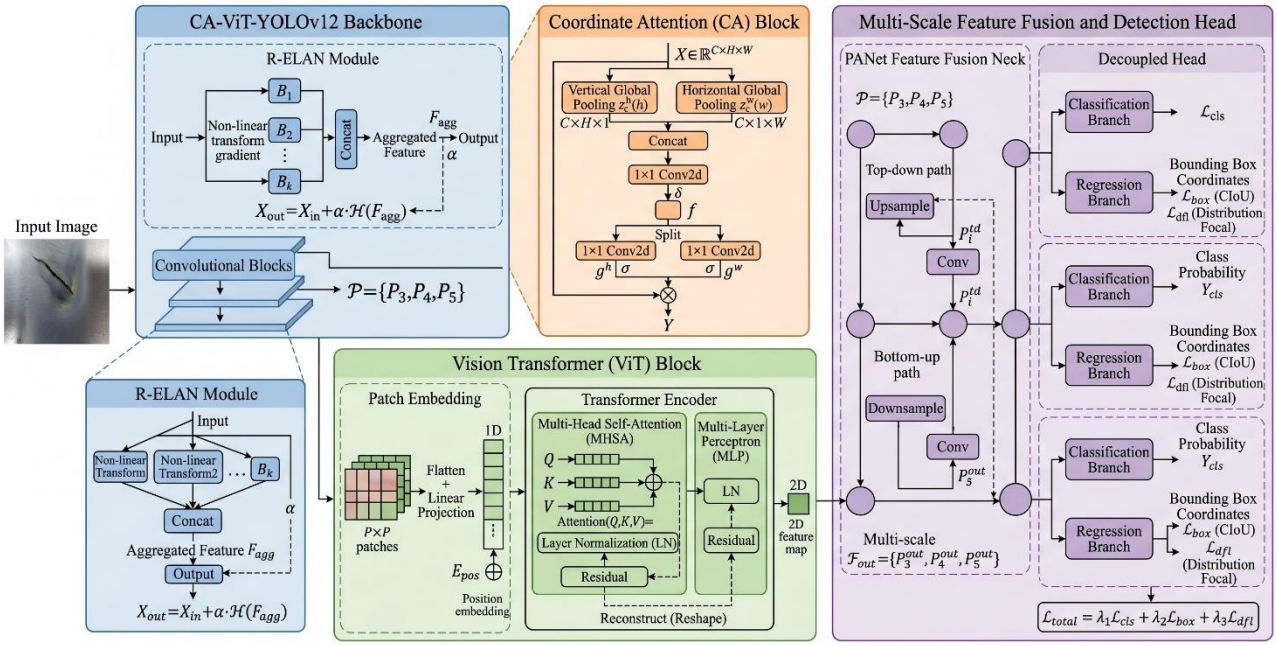

Figure. 2. Algorithm architecture diagram

3.1. Backbone Feature Extraction Network Module

The backbone network's primary task is to extract multi-scale foundational features from input steel surface images. This model adopts YOLOv12's core Residual Efficient Layer Aggregation Network (R-ELAN) as the foundational architecture for feature extraction.

In steel surface defect detection tasks, the wide variety of defects (e.g., cracks, scratches, pits) and their extremely large scale variations demand a backbone network capable of powerful multi-scale feature extraction—while preserving high-resolution spatial information and controlling computational complexity. To this end, this algorithm adopts YOLOv12's core backbone architecture: the Residual Efficient Layer Aggregation Network (R-ELAN).

Traditional deep convolutional networks often suffer from gradient vanishing and low feature reuse when stacking layers progressively. R-ELAN addresses these issues by introducing carefully designed residual connections and gradient path aggregation, enabling efficient fusion of deeper convolutional layers.

Mathematically, let the input feature tensor to the R-ELAN module be X_{in} . The network first maps and splits feature channels via a transition layer, generating multiple parallel gradient computation branches (Blocks). Denoting the nonlinear transformation functions of these branches as $B_1, B_2, \dots, B_k$, the multi-branch aggregated feature F_{agg} can be expressed as:

$$F_{agg} = \text{Concat}\big(B_1(X_{in}), B_2(X_{in}), \dots, B_k(X_{in})\big) \tag{1}$$

To enhance feature reuse and prevent gradient vanishing in deep networks, R-ELAN introduces a residual connection architecture with a scaling factor after feature aggregation. Its final output feature X_{out} can be formalized as:

$$X_{\text{out}} = X_{\text{in}} + \alpha \cdot \mathcal{H}\big(F_{\text{agg}}\big) \tag{2}$$

Here, $\mathcal{H}(\cdot)$ denotes the channel-mapped convolution operation applied to the aggregated features, and α is a residual scaling factor dynamically learned during training. This design greatly enhances the network's ability to represent subtle defect features—without compromising the underlying physical texture characteristics of raw steel images (e.g., the original rolling texture of steel plates).

Moreover, to maintain strong spatial perception capability with fewer parameters for industrial deployment, the model extensively employs a novel 7×7 depthwise separable convolution block within the backbone network.

Depthwise separable convolution decomposes conventional standard convolution into two sequential steps: depthwise convolution and pointwise convolution. Suppose the input feature map has spatial dimensions $D_F \times D_F$, input channels M , output channels N , and kernel size $D_K \times D_K$ (where $D_K = 7$).

The computational cost C_{std} of conventional standard convolution is:

$$C_{std} = D_K \cdot D_K \cdot M \cdot N \cdot D_F \cdot D_F \tag{3}$$

Whereas the computational cost C_{dws} of depthwise separable convolution is:

$$C_{dws} = D_K \cdot D_K \cdot M \cdot D_F \cdot D_F + M \cdot N \cdot D_F \cdot D_F \tag{4}$$

Thus, the computational compression ratio R between the two is:

$$R = \frac{C_{dws}}{C_{std}} = \frac{1}{N} + \frac{1}{D_K^2} \tag{5}$$

When employing 7×7 large convolutional kernels, $1/D_K^2 \approx 0.0204$. The formula shows that this design expands the receptive field to 7×7 while reducing computational cost by nearly 50× compared to a standard convolution of the same size.

The large 7×7 -sized receptive field enables the network to effectively capture macro-level contextual information on steel plate surfaces (e.g., large-scale rust scale, uneven illumination gradients), thereby effectively suppressing macro-level background noise; meanwhile, R-ELAN's residual aggregation mechanism ensures feature flow at micro-level resolution, allowing the model to retain exceptional sensitivity to fine defects such as "cold cracks" or "micro-scratches." This synergy provides a highly expressive multi-scale feature foundation for the subsequent dual attention modules (CA and ViT).

3.2. Coordinate Attention Mechanism

In steel surface defect detection, defects such as scratches and fine cracks typically exhibit pronounced spatial directionality, random distribution, and low contrast. Conventional channel attention mechanisms compress spatial information into a 1D vector via 2D global average pooling—modeling inter-channel dependencies but discarding spatial location information critical for defect localization. To overcome this limitation, our model introduces the Coordinate Attention (CA) mechanism. Innovatively integrating channel attention with positional information, CA performs 1D feature aggregation along two orthogonal spatial directions, preserving fine-grained positional cues while successfully capturing long-range dependencies. The CA module's computation consists primarily of two stages: coordinate information embedding and coordinate attention generation.

Assume the input feature tensor is $X \in \mathbb{R}^{C \times H \times W}$, where C , H , and W denote the number of channels, height, and width, respectively. To encourage the network to capture features with precise positional encoding, the CA module avoids conventional 2D global pooling and instead applies two 1D global pooling layers—one along the horizontal (X) direction and the other along the vertical (Y) direction—for feature encoding.

Specifically, for the c -th channel, the pooling output $z_c^h(h)$ along the height (vertical) dimension of size $H \times 1$ can be expressed as:

$$z_c^h(h) = \frac{1}{W} \sum_{0 \le j < W} x_c(h, j) \tag{6}$$

Similarly, the pooling output $z_c^w(w)$ along the width (horizontal) dimension of size $1 \times W$ can be expressed as:

$$z_c^W(w) = \frac{1}{H} \sum_{0 \le i < H} x_c(i, w) \tag{7}$$

These two transformations generate feature maps of sizes $C \times H \times 1$ and $C \times 1 \times W$, respectively. This aggregation operation along two independent dimensions enables the network to capture long-range feature dependencies along one dimension while preserving precise spatial location information along the other—making it highly sensitive to elongated scratch or crack features on steel surfaces.

After embedding coordinate information, the extracted direction-aware features must be fused to generate an attention weight matrix. First, the two feature maps z^h and z^w from the independent dimensions are concatenated along the spatial dimension, then fed into a shared 1×1 convolutional layer (Conv2d) for channel-wise compression and dimensionality reduction, producing the intermediate feature map f :

$$f = \delta\Big(\mathrm{Conv2}d([z^h, z^w])\Big) \tag{8}$$

Here, [∵] denotes concatenation along the spatial dimension; δ is a nonlinear activation function; the resulting feature map is $f \in \mathbb{R}^{C/r \times 1 \times (H+W)}$, where r is the channel compression ratio.

Next, the fused feature map f is split (Split) along the spatial dimension into two independent tensors $f^h \in \mathbb{R}^{C/r \times H \times 1}$ and $f^w \in \mathbb{R}^{C/r \times 1 \times W}$, corresponding to vertical and horizontal directions. Then, two separate 1×1 convolutional layers restore the number of channels to the original input dimension C, followed by Sigmoid activation, yielding attention weight maps $g^h \in \mathbb{R}^{C \times H \times 1}$ and $g^w \in \mathbb{R}^{C \times 1 \times W}$ for the height and width dimensions, respectively:

$$g^h = \sigma\Big(\mathrm{Conv2d}_h(f^h)\Big) \tag{9}$$

$$g^w = \sigma\big(\mathrm{Conv2d}_w(f^w)\big) \tag{10}$$

Finally, the generated attention weight maps are used to re-weight (Re-weight) the original input feature tensor X by coordinate. The value $y_c(i,j)$ at position (i, j), channel c of the final output feature tensor $Y \in \mathbb{R}^{C \times H \times W}$ is computed as follows:

$$y_c(i,j) = x_c(i,j) \times g_c^h(i) \times g_c^w(j) \tag{11}$$

The above attention maps are then applied precisely to the input feature maps via element-wise multiplication, thereby accurately highlighting and enhancing representations of the defect objects of interest in the image. The feature stream enhanced by the CA module exhibits strong directional awareness and positional sensitivity in the spatial domain. This design effectively guides the model to focus on defect-coordinate activations amid complex steel plate background textures, significantly improving its capability to localize dense or minute defects.

3.3. Vision Transformer Global Modeling Module

In traditional deep convolutional neural networks (CNNs), the size of the receptive field directly determines the model's ability to capture contextual information. However, surface defects on steel exhibit significant scale variation and morphological diversity—for instance, long scratches spanning an entire steel plate or large-scale rust patches. Relying solely on the local receptive field of convolutional operations often fails to establish global semantic associations across regions, leading to misclassification of locally similar, normal rolling textures as defects. To overcome this bottleneck, our algorithm introduces a Vision Transformer (ViT) module into the backbone network, leveraging its powerful self-attention mechanism to directly model global contextual relationships in the image and achieve deep understanding of macro-level semantics on steel surfaces. The ViT module's computation pipeline consists of three core steps: patch embedding, positional encoding, and multi-head self-attention encoding.

Unlike CNNs, which apply convolution pixel-by-pixel across layers, the ViT module treats the feature map as sequential data. Suppose the local feature map tensor input to the ViT module is $X \in$

$\mathbb{R}^{C\times H\times W}$; it is first partitioned spatially into N non-overlapping image patches of fixed size $P\times P$. Thus, the length of the image patch sequence is $N=\frac{H\times W}{P^2}$.

Next, these one-dimensional image patches are flattened and mapped via a learnable linear projection matrix $E\in\mathbb{R}^{(P^2\cdot C)\times D}$ into a hidden space of dimension D . To preserve the original spatial topology of the image during serialization, a learnable one-dimensional positional encoding $E_{\text{pos}}\in\mathbb{R}^{N\times D}$ is added to each image patch vector. The resulting input sequence Z_0 can be expressed as:

$$Z_0=\left[x_p^1E;x_p^2E;\ldots;x_p^NE\right]+E_{\text{pos}} \tag{12}$$

where x_p^i denotes the i -th flattened image patch. Through this operation, originally local visual features are transformed into a globally aware sequence with positional sensitivity.

The constructed sequence data then enters a multi-layer Transformer encoder for feature interaction. At the core of the encoder lies the multi-head self-attention (MHSA) mechanism, enabling the model to dynamically aggregate information from all other image patches globally when processing the current patch.

For a feature vector in the sequence, MHSA first maps it via three distinct learnable linear transformation matrices W_Q,W_K,W_V into a query vector matrix (Query, Q), a key vector matrix (Key, K), and a value vector matrix (Value, V):

$$Q=Z_{l-1}W_Q,\ K=Z_{l-1}W_K,\ V=Z_{l-1}W_V \tag{13}$$

Self-attention weights are computed via dot-product. Specifically, the similarity between different image patches is calculated using the dot product of Q and K , scaled by the factor $\sqrt{d_k}$ (where d_k is the dimensionality of the key vectors) to prevent gradient explosion; finally, the result is normalized via the Softmax function to obtain the attention weight matrix, which is then multiplied by the value matrix V :

$$\text{Attention}(Q,K,V)=\text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{14}$$

To capture feature representations across distinct subspaces, MHSA performs h such self-attention operations (i.e., multiple heads) in parallel, then concatenates the outputs of all heads and applies a linear projection:

To capture feature representations across distinct subspaces, MHSA performs h such self-attention operations (i.e., multiple heads) in parallel, then concatenates the outputs of all heads and applies a linear projection:

$$\text{MHSA}(Z_{l-1})=\text{Concat}(\text{ head }_1,\text{ head }_2,\ldots,\text{ head }_h)W^0 \tag{15}$$

Here, $\text{head}_i=\text{Attention}(Q_i,K_i,V_i),W^O$ denotes the output projection matrix.

The full Transformer encoder module consists of alternating stacked MHSA and multilayer perceptron (MLP) layers, with layer normalization applied before each sublayer and residual connections preserved to ensure stable gradient flow. The forward pass of the l -th encoder layer is given by:

$$Z_l'=\text{MHSA}\big(\text{LN}(Z_{l-1})\big)+Z_{l-1} \tag{16}$$

$$Z_l=MLP\big(LN(Z_l')\big)+Z_l' \tag{17}$$

After processing through multiple Transformer encoder layers, the output sequence Z_L has fully integrated global contextual information. Finally, the system reshapes the 1D sequence back into 2D spatial dimensions via a shape reconstruction operation and feeds it into the subsequent YOLOv12 network.

The self-attention mechanism in the ViT module enables the model to bridge spatial distances—linking a localized minor defect with the overall lighting distribution or macroscopic background of the steel plate. For instance, in complex regions where water stains overlap with surface rust scale, ViT accurately identifies the semantic nature of the region as "background interference" rather than "actual defect" by computing similarities among global feature patches (i.e., interactions between Q and K), thereby significantly enhancing the model's discriminative capability and robustness in complex, multi-textured environments.

The ViT module's decoding process is illustrated in Figure 3

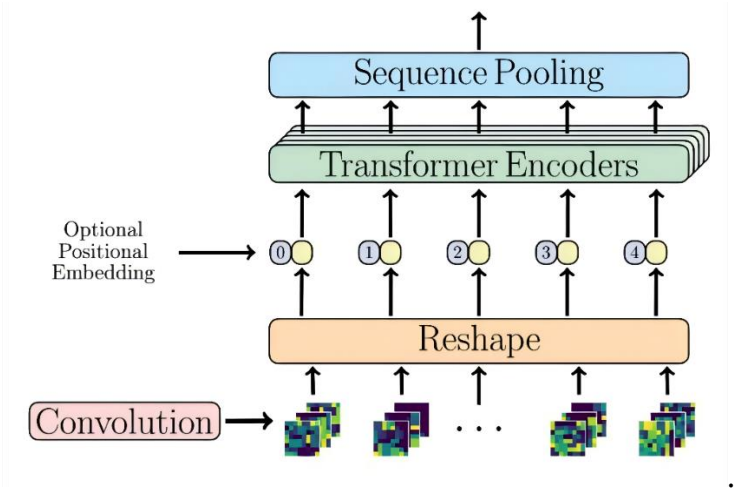

Figure 3. Schematic of ViT module decoding

3.4. Multi-scale Feature Fusion and Detection Head

After basic feature extraction by the backbone network and local-global feature reconstruction via the CA and ViT modules, the network has acquired feature maps rich in semantic meaning and precise localization information. However, steel surface defects vary widely in scale—from micron-level pitting and fine cracks to macroscopic patches occupying most of the field of view. To effectively address multi-scale object detection and convert high-dimensional features into final predictions, this algorithm designs an efficient multi-scale feature fusion mechanism in the network's neck and adopts an optimized decoupled prediction head in the head.

The conventional Feature Pyramid Network (FPN) contains only a top-down path, making it prone to losing high-resolution localization information from shallow layers in deep networks. This model adopts PANet's bidirectional fusion concept and integrates YOLOv12's simplification strategy for multi-scale detection paths to construct a more efficient feature fusion neck.

Assume the three feature maps of different scales (large, medium, small) output from the backbone network and attention module are denoted as $\mathcal{P} = \{P_3, P_4, P_5\}$, where P_3 has the highest resolution and is used for detecting small objects; P_5 carries the deepest semantics and is used for detecting large objects.

First, via the top-down path, high-level semantic information from deeper layers is propagated to shallower layers, generating intermediate features P_i^{td} .

$$P_i^{td} = \text{Conv}(P_i) + \text{Upsample}\left(P_{i+1}^{td}\right),\ i \in \{3,4\} \tag{18}$$

Next, via the bottom-up path, high-resolution spatial localization information from shallower layers is fed back to deeper layers, generating the final output features P_i^{out} for detection:

$$P_i^{\text{out}} = \text{Conv}\left(P_i^{td}\right) + \text{Downsample}\left(P_{i-1}^{\text{out}}\right),\ i \in \{4,5\} \tag{19}$$

After bidirectional cross-layer fusion, the output feature set $\mathcal{F}_{\text{out}} = \{P_3^{\text{out}}, P_4^{\text{out}}, P_5^{\text{out}}\}$ achieves perfect unification of high-level semantics and low-level geometric details, effectively enhancing the model's adaptability to steel defect detection across large scale variations.

In object detection tasks, class classification focuses more on extracting texture and semantic features of defects, whereas bounding box regression emphasizes geometric edges and spatial coordinates of defects. To avoid mutual interference between these two tasks during feature extraction, the model adopts a Decoupled Head architecture.

For each fused feature map P_i^{out} at every level, the Decoupled Head splits it into two parallel convolutional branches, respectively outputting class prediction probabilities Y_{cls} and bounding box coordinate predictions Y_{reg} :

$$Y_{cls} = \text{Conv}_{cls}(P_i^{out}) \tag{20}$$

To better balance localization and classification objectives, the model's total loss function $\mathcal{L}_{\text{total}}$ comprises three weighted components: classification loss $(\mathcal{L}_{\text{cls}})$, bounding box regression loss $(\mathcal{L}_{\text{box}})$, and distribution focal loss $(\mathcal{L}_{dfl})$:

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{cls}} + \lambda_2 \mathcal{L}_{\text{box}} + \lambda_3 \mathcal{L}_{\text{dfl}} \tag{21}$$

Among these, the classification loss $\mathcal{L}_{cls}$ employs binary cross-entropy loss to compute the error between predicted classes and ground-truth labels. The bounding box regression loss $\mathcal{L}_{\text{box}}$ adopts the CloU loss. Scratches on steel surfaces often exhibit extreme aspect ratios; CloU not only considers the overlapping area and center-point distance between predicted and ground-truth boxes, but also introduces an aspect-ratio penalty term αv . Distribution focal loss transforms box positions into continuous distributions, prompting the network to rapidly focus on probability distributions near defect edges—further improving regression accuracy for ambiguous boundaries (e.g., the gradual transition zone at rust edges).

The integration of a multi-scale feature pyramid with a decoupled head design enables the CA-ViT-YOLOv12 model to handle complex scenarios in steel industry settings with exceptional precision. Whether it is a "long scratch" spanning the full width of a steel plate or a "tiny pit" occupying only a few pixels, the system delivers predictions with high confidence and high coordinate accuracy. Coupled with an optimized multi-task loss function, the model maintains high recall while significantly reducing false positives caused by background texture.

4. Experiments and Results Analysis

To comprehensively evaluate the effectiveness of the proposed algorithm, we designed extensive experiments and compared it against multiple mainstream methods. This chapter details the datasets used, evaluation metrics, baseline methods for comparison, and provides both qualitative and quantitative analyses of the experimental results.

4.1. Dataset

The primary dataset used in this study originates from proprietary data collected in real production environments of automotive manufacturing and steel processing enterprises, comprising 5000 surface images of steel components captured under diverse production conditions and lighting environments. Additionally, to test and validate the model's generalization capability and reliability across different data sources, publicly available steel component surface defect datasets (e.g., the Northeastern University NEU-DET dataset) were incorporated to augment the sample pool. After meticulous annotation, the dataset covers six highly representative steel surface defect types: Crazing, Inclusion, Patches, Pitted Surface, Rolled-in Scale, and Scratches. All images were annotated with bounding-box labels by domain experts with industrial quality inspection experience, meeting the requirements of high-precision object detection tasks.

To ensure objectivity and fairness in model evaluation, this experiment follows a standard machine learning workflow and systematically partitions the above-constructed hybrid dataset. The dataset is rigorously split into training, validation, and test sets, ensuring consistent class distribution across all subsets—thus avoiding model evaluation bias arising from imbalanced data distribution.

In real-world industrial visual inspection scenarios, imaging conditions often deviate from the ideal due to minor camera position shifts, workshop lighting flicker, and variations in surface reflectivity across different steel batches. Meanwhile, certain severe defects (e.g., large-scale mill scale) occur frequently, whereas fine cracks appear rarely—resulting in a pronounced class imbalance. To address these issues and enhance model robustness, this experiment employs a comprehensive data augmentation strategy prior to model training. Specifically:

Color-space perturbation: Primarily addresses uneven illumination. Random adjustments to hue, saturation, and brightness are applied in the HSV color space.

Geometric-space transformation: Primarily addresses variations in workpiece orientation and scale. Images undergo random rotation (angle range $\pm 15^{\circ}$), random translation (translation ratio relative to image size ± 0.3), and random scaling (scale ratio relative to original size $= 0.5$).

Deformation simulation: Random shear transformation (Shear) is applied to images, with shear angle range set to $\pm 5^{\circ}$.

4.2. Evaluation Metrics and Baseline Methods

This study adopts Precision (P), Recall (R), and mean Average Precision (mAP) across all classes as core quantitative evaluation metrics.

To comprehensively and objectively assess the overall performance of the proposed CA-ViT-YOLOv12 algorithm—and to scientifically validate the practical contribution of its core innovative modules—this experiment selects multiple widely representative object detection algorithms as comparative baseline (Baseline) methods, covering classic two-stage networks, mainstream one-stage networks, and the original base architecture of our algorithm:

Faster R-CNN: A classic two-stage object detection algorithm that generates region proposals via a Region Proposal Network (RPN), followed by classification and bounding-box regression.

YOLOv8: A benchmark one-stage detection architecture; YOLOv8 exhibits high maturity in feature extraction and multi-scale fusion (PAN-FPN), and is widely deployed in current industrial visual quality inspection tasks.

YOLOv12: By introducing an attention-centric architectural design, it significantly improves detection accuracy while maintaining real-time inference speed—achieving a better balance between speed and accuracy.

4.3. Experimental Results

To verify the effectiveness of the proposed algorithm, we conduct a comprehensive comparative analysis between our method and multiple mainstream approaches on the dataset. Experimental results are shown in Table 1.

Table 1. Experimental results

Method	Precision (%)	Recall (%)	mAP@0.5 (%)	mAP@0.5:0.95(%)
Faster R-CNN	76.4	68.2	70.1	41.5
YOLOv8	79.8	71.5	73.8	45.6
YOLOv12	82.1	74.3	76.5	48.2

As shown intuitively in Table 1, the proposed CA-ViT-YOLOv12 fusion model achieves state-of-the-art performance across all core quantitative metrics. On the key metric of mean Average Precision at IoU=0.5 (mAP@0.5), CA-ViT-YOLOv12 reaches 80.9%. Although traditional two-stage networks employ a Region Proposal Network (RPN), their lack of attention guidance for tiny targets makes them highly prone to anchor box redundancy and misclassification against complex steel surface backgrounds. Our algorithm outperforms them by over 10 percentage points in accuracy, and its single-stage architecture inherently better satisfies real-time requirements in industrial production lines. While YOLOv8 exhibits strong feature pyramid fusion capability, its purely convolutional

architecture faces limitations when handling large-area background noise. Our algorithm's 7.1% accuracy advantage fully demonstrates the generational superiority of the "attention + Transformer" dual-driven paradigm for low-contrast industrial images. The original YOLOv12, leveraging its advanced architecture design, already achieves 76.5% mAP@0.5 and 82.1% precision. After integrating the dual-attention mechanism, CA-ViT-YOLOv12's Recall improves significantly from 74.3% to 78.5%, and mAP@0.5:0.95 rises from 48.2% to 52.4%.

The substantial improvements in Recall and strict-threshold precision (mAP@0.5:0.95) strongly validate the effectiveness of our novel modules: First, the Coordinate Attention (CA) mechanism endows the network with directional perception capability in the spatial dimension, enabling it to sensitively detect ultra-low-contrast, hairline-thin "cracks" and "micro-pits", thereby greatly reducing false negatives (i.e., improving Recall); Second, the Vision Transformer (ViT) module, via global self-attention, successfully establishes contextual semantic associations between defect features and macro-scale steel surface textures. When confronted with extremely complex background interferences—such as water stains, glare, or large-area rust—ViT effectively suppresses false positives (FPs), preventing normal background regions from being misclassified as defects, thus elevating overall detection precision (Precision) to 85.2%.

5. Conclusion

To address key challenges in steel industry surface defect detection—such as frequent missed detections of tiny defects, severe interference from complex backgrounds, and the lack of deep decision support in traditional detection systems—this paper proposes and implements a steel surface defect detection method based on YOLOv12 fused with a dual-attention mechanism. Specifically, this work innovatively integrates the Coordinate Attention (CA) mechanism and a Vision Transformer (ViT) module into the YOLOv12 base architecture. Experiments show that the CA module's capability to capture fine-grained features complements the ViT module's ability to model global semantic context. The fused model achieves a mean Average Precision (mAP@0.5) of 0.809 across all classes on a real industrial dataset. Compared with classic two-stage and one-stage baseline models, this algorithm demonstrates overwhelming performance advantages in detecting low-contrast defects (e.g., minute scratches) and resisting complex background interference (e.g., large-scale mill scale), significantly reducing both missed detections and false alarms in industrial settings. Although the proposed CA-ViT-YOLOv12 algorithm and its collaborative system excel across multiple metrics, further exploration remains necessary to meet increasingly sophisticated intelligent manufacturing demands. Currently, the system relies primarily on 2D visual images. Future work plans to incorporate 3D laser profilometer or infrared thermal imaging data to enable multimodal physical feature fusion; simultaneously, research will explore unsupervised or semi-supervised continual learning mechanisms, allowing the model to adaptively iterate using the massive volume of daily unlabeled production-line data, further enhance the model's generalization ability in extremely rare defect cases.

References

[1] Girshick R. Fast r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2015: 1440-1448.

[2] Wang Y, Wang X, Hao R, et al. Metal surface defect detection method based on improved cascade r-cnn[J]. Journal of Computing and Information Science in Engineering, 2024, 24(4): 041002.

[3] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 779-788.

[4] Li D, Wang E, Li Z, et al. STE-YOLO: A surface defect detection algorithm for steel strips[J]. Electronics, 2024, 14(1): 54.

[5] Zhou Y, Zhao Z. MPA-YOLO: Steel surface defect detection based on improved YOLOv8 framework[J]. Pattern Recognition, 2025, 168: 111897.

[6] Zhang L, Wang Z, Ma Y, et al. Steel surface defect detection algorithm based on improved YOLOv10[J]. Scientific Reports, 2025, 15(1): 32827.

[7] Wang B, Wang M, Yang J, et al. YOLOv5-CD: Strip steel surface defect detection method based on coordinate attention and a decoupled head[J]. Measurement: Sensors, 2023, 30: 100909.

[8] Li Z, Wu C, Han Q, et al. CASI-Net: A novel and effect steel surface defect classification method based on coordinate attention and self-interaction mechanism[J]. Mathematics, 2022, 10(6): 963.

[9] Fan J, Ling X, Liang J. Detection of surface defects of steel plate based on vit[C]//Journal of physics: conference series. IOP Publishing, 2021, 2002(1): 012039.

[10] Liu G, Chen Y, Ye J, et al. A transformer neural network based framework for steel defect detection under complex scenarios[J]. Advances in Engineering Software, 2025, 202: 103872.

[11] Wu S, Yang H, Liao L, et al. DSAT: a dynamic sparse attention transformer for steel surface defect detection with hierarchical feature fusion[J]. Scientific Reports, 2025, 15(1): 29198.

[12] Li Y, Han Z, Wang W, et al. Steel surface defect detection based on sparse global attention transformer[J]. Pattern Analysis and Applications, 2024, 27(4): 152.

[13] Guo Z, Wang C, Yang G, et al. Msft-yolo: Improved yolov5 based on transformer for detecting defects of steel surface[J]. Sensors, 2022, 22(9): 3467.

[14] Wu S, Yang H, Liao L, et al. SH-DETR: Enhancing steel surface defect detection and classification with an improved transformer architecture[J]. PLoS One, 2025, 20 (11): e0334048.

[15] Pan W, Zhong R, Huang J, et al. DEENet: an edge-enhanced CNN–Transformer dual-encoder model for steel surface defect detection[J]. Scientific Reports, 2026.