Combining DETR Multi-scale Perception Structure to Improve Target Boundary Extraction Accuracy in Image Data

Yancen Liu
School of Information Science & Engineering, Lanzhou University, Lanzhou 730000, Gansu, China
Corresponding author: Yancen Liu (e-mail: Lyc1128052@163.com)

ABSTRACT Accurate target boundary extraction is essential for high-precision image interpretation in intelligent sensing systems and provides important technical support for electromagnetic imaging, remote sensing, and vision-assisted signal perception applications. Owing to the limited feature representation capability of the original Detection Transformer (DETR) when processing objects at different scales, target boundary extraction accuracy remains insufficient, particularly in scenarios involving fine-grained contour localization. To address this issue, an improved DETR framework based on a multi-scale perception structure is proposed. A multi-scale encoding architecture integrated with a feature pyramid network is employed to capture features at multiple resolutions, enhancing boundary-aware spatial representations through an edge completion mechanism. A scale-aware multi-head attention module is incorporated into the encoder to preserve scale consistency during global feature modeling and reduce semantic drift. Furthermore, a boundary regression head combined with positional embedding jointly exploits spatial and semantic information to improve contour localization, while an IoU-aware loss function optimizes prediction accuracy in overlapping regions. Experimental results demonstrate that the proposed method increases the average IoU from 0.81 to 0.86 and improves target boundary AP from 44.9% to 48.1%. Compared with Deformable-DETR and Elastic-DETR, it also achieves lower boundary error ratios, confirming that the proposed multi-scale perception mechanism effectively enhances high-precision boundary fitting and offers practical value for intelligent visual sensing systems in engineering applications.

INDEX TERMS DETR architecture, multi-scale feature fusion, boundary extraction, intelligent sensing, electromagnetic imaging

I. INTRODUCTION

ACCURATE boundary extraction underpins precise target localization and recognition in image detection tasks.

With the continuous development of deep learning technology, end-to-end detection architecture has made significant progress in reducing artificial prior intervention and improving detection efficiency [1-3]. DETR, as an end-to-end target detection method integrating Transformer structure, realizes unified modeling of various targets in the image by constructing a global attention mechanism, breaking the dependence of traditional detection methods on anchor boxes and NMS (Non-Maximum Suppression), and showing strong structural versatility and modeling capabilities [4-6]. The current target detection system based on DETR still has significant deficiencies in boundary area extraction, which limits its further expansion in fine detection scenarios [7-9].

In practical applications, target size changes, background interference, and boundary ambiguity constitute the main challenges of target boundary extraction. Scale differences lead to the limited expression ability of traditional single-scale feature extraction strategies when dealing with small targets or complex structural areas. The presence of interference information in the background weakens the focusing effect of the attention mechanism on the boundary contour, while edge blur or target occlusion further aggravates the problem of model positioning offset and inaccurate boundary regression [10,11]. The multi-scale perception mechanism applies cross-level semantic fusion and spatial alignment strategies, so that the model can take into account both detail preservation and structural consistency, and improve the response quality of boundary features at different scales. Under the end-to-end detection architecture, since the Transformer relies on fixed position encoding and global feature modeling strategies, it lacks an adaptive processing mechanism for local details and scale changes, ignoring the high-frequency details and local geometric structures of the target boundary area, resulting in a decrease in boundary

prediction accuracy [12,13].

To alleviate the above problems, researchers have proposed a variety of methods based on improving feature expression and enhancing spatial modeling capabilities. The multi-scale noise scheduling strategy selectively enhances the high-confidence features of the boundary area in the feature fusion stage and suppresses the redundant responses of the blurred area, which can achieve directional enhancement of boundary details. In the feature extraction stage, some works apply FPN structures to capture multi-scale semantic information and achieve the fusion of low-level details and high-level semantics through lateral connections, which enhances the response ability to small targets and edge details to a certain extent [14-16]. In the Transformer module, some studies have also tried to embed local attention mechanisms in the encoder or apply multi-scale query schemes in the decoder to improve the structure's recognition accuracy of boundary areas [17,18]. In response to the problem of boundary positioning offset, some works have applied IoU-aware loss or center modulation strategies to enhance the model's ability to regress high-precision bounding boxes [19,20]. In the actual detection process, these methods are prone to apply scale interference during the fusion process, resulting in insufficient feature consistency, and lack of explicit modeling mechanism for scale dependence, and there are still significant errors in boundary area positioning [21,22].

To solve the problem of insufficient scale-aware in DETR in target boundary extraction, an improved detection framework combined with a multi-scale perception structure is therefore designed. This framework constructs a pyramid feature network in the input feature encoding stage to extract semantic representations at different resolutions, and uses a lightweight dynamic convolution module to perform cross-scale feature fusion to improve the expression accuracy of the boundary response area. This proposed framework effectively improves the detection structure's ability to model the boundary areas of targets of different scales, promoting the performance development of end-to-end target detection systems toward fine boundary extraction.

II. RELATED WORK

In the target detection method driven by the Transformer architecture, researchers have continuously promoted the adaptability and detection accuracy of the model in complex scenes through structural innovation and task specialization. Ahmed et al. effectively improved the accuracy and positioning accuracy of remote sensing image target detection by integrating YOLOv7 and DETR algorithms and applying an intelligent selection module. The experiment verified its optimization effect in complex scenes [23].

Researchers gradually explored the deep collaborative path of fusion strategy and feature interaction mechanism to improve the model's adaptability to different data modalities and complex scene conditions [24,25]. Zang et al. proposed a twin network that integrated histogram layers and cross-modal Transformer based on the characteristics of different day and night scenes. By strengthening multi-modal feature interaction during the day, simplifying the structure at night, and focusing on thermal imaging, the pedestrian detection accuracy was effectively improved [26]. Precisely modeling local and global semantic associations under complex spatiotemporal conditions has become an important breakthrough in improving the robustness and generalization ability of target detection [27,28].

In the task of target boundary extraction from image data, multi-scale perception mechanism has been widely used in structural optimization and accuracy improvement as a key means to enhance the model's perception ability and detail recovery ability. Liu et al. achieved fine-grained spatial information capture in the boundary extraction branch by integrating multi-scale dilated convolution and gated feature fusion mechanism, significantly improving the segmentation accuracy of complex building contours [29]. Based on the boundary perception ability of the multi-scale structure enhancement model, multi-level feature fusion is further used to refine the edge information expression [30,31]. The MBR-HRNet (Multi-scale Boundary-Refined High-Resolution Network) model proposed by Yan et al. integrated different levels of features through a multi-scale context fusion module and combined the boundary refinement branch to enhance edge information learning, effectively improving the extraction accuracy and detail integrity of building boundaries in complex scenes [32]. While strengthening the expression of edge details, the application of edge constraint mechanism has become a key strategy to improve boundary discrimination and global consistency [33]. Wang et al. constructed a MEC-Net (Multi-scale Edge Constraint Network) model that integrated multi-scale edge constraints and data enhancement, effectively improving the accuracy of building boundary extraction and overall segmentation performance, and achieved excellent performance on multiple public datasets [34]. The above research has made some progress driven by the multi-scale perception mechanism, but when faced with image scenes with drastic scale changes or high boundary ambiguity, there are still problems such as unstable boundary positioning and insufficient detail recovery.

III. MULTI-SCALE PERCEPTION DETR BOUNDARY ENHANCEMENT MECHANISM

A. FRAMEWORK STRUCTURE AND OVERALL PROCESS

This paper designs an improved framework that integrates multi-scale perception structure to improve the scale adaptability of DETR in target boundary extraction. The overall architecture retains the advantages of Transformer backbone modeling while applying a scale modeling mechanism to enhance the expression ability of boundary features. The structure achieves refined extraction of boundary information while maintaining the end-to-end inference process through the organic integration of multi-scale feature en-

coding, scale-aware attention mechanism, and boundary regression module.

Figure 1 shows the fusion method of the DETR backbone network and the multi-scale perception module. The feature extraction part uses a pyramid style to construct a multi-scale feature layer and outputs image representations with different resolutions and semantic depths. The scale-aware multi-head attention mechanism is applied in the encoding stage to retain the structural consistency between different scales in the global modeling process. The final boundary regression module uses the position embedding information to predict the boundary parameters and combines the IoU-aware loss function to improve the robustness of boundary positioning. The overall design maintains the simplicity of the structure while enhancing the ability to express the scale characteristics of the target boundary.

B. MULTI-SCALE FEATURE ENCODING AND FUSION STRATEGY

This paper applies a multi-scale feature encoding mechanism based on FPN to enhance the modeling ability of the target boundary area. By constructing feature outputs of different levels in the backbone network, a multi-scale feature map with differentiated resolution and receptive field is generated, covering two types of information: local details and global semantics. Assuming that the set of multi-scale features extracted from the backbone network is $\{F_1, F_2, \ldots, F_L\}$, where the l-th layer feature map $F_l \in R^{C_l \times H_l \times W_l}$, and C_l, H_l, and W_l represent the number of channels, height, and width, respectively. To achieve cross-scale semantic alignment, each feature map is mapped to a unified dimension C through a 1×1 convolution, and the spatial resolution is aligned using bilinear interpolation to unify the size of all scale feature maps to (C, H, W). The multi-scale fusion operation adopts a weighted summation strategy, and the response values of all scale feature maps at the spatial position (x, y) are weighted and integrated into the fusion feature $\hat{F}(x, y)$ through Formula (1):

$$\hat{F}(x,y) = \sum_{l=1}^{L} \alpha_l \cdot \mathrm{Upsample}(Q_l * F_l)(x,y) \qquad (1)$$

In Formula (1), α_l is the learnable fusion weight corresponding to the l-th layer feature, satisfying the constraint condition $\sum_{l=1}^{L} \alpha_l = 1$; Q_l represents the convolution weight parameter for channel alignment; $\mathrm{Upsample}(\cdot)$ is the upsampling operation. The fusion output feature $\hat{F} \in R^{C \times H \times W}$ is used in the subsequent Transform encoding stage as the multi-scale boundary feature input for unified representation after scale alignment.

The core of multi-scale feature encoding is to extract the response information of the target boundary at different resolutions to fully cover the hierarchical differences of fine-grained details and semantic structures. High-resolution feature maps help to preserve edge details, while low-resolution features enhance regional perception and context aggregation. The two work together to support precise perception of boundary positions. To verify the response differences of features of different scales in the boundary area, the visualization results of shallow, middle, and deep feature maps based on ResNet are constructed to show the impact of scale changes on the degree of target contour preservation and response intensity, as shown in Figure 2.

Figure 2 shows the average response heat map of the input image and three typical scales, where the input image is annotated with the target area bounding box, and the feature maps of the three scales correspond to the average channel activation distribution of the shallow structure, the middle structure, and the deep structure, respectively. The shallow feature map shows a high response area near the boundary line, reflecting the sensitivity to local texture and contour. The middle feature map still has the ability to recognize the boundary contour, but the response range tends to expand. The deep feature map is concentrated in the target semantic area; the boundary response is blurred; more semantic-level focusing behavior is reflected.

To further verify the coverage ability of features of different scales on boundary information, quantitative indicators are used to statistically analyze the distribution of their response intensity in the boundary area. The feature dimension, theoretical receptive field size, and boundary area response ratio are measured respectively to quantify the boundary expression ability of different feature maps. The shallow feature map has a smaller receptive field, so the response is more concentrated at the edge and the boundary accounts for a higher proportion, while the response of the deep feature map is smoother and distributed within the overall structure of the target. Table 1 gives the comparison results and is used to support the design basis of the importance weights of each feature layer in the scale fusion structure.

TABLE 1. Dimension configuration and boundary response statistics of feature maps of different scales

Scale Level	Feature Dimension	Receptive Field Size (pixels)	Boundary Response Ratio (%)
Level 1 (shallow)	128	33 × 33	78.2
Level 2	256	65 × 65	65.7
Level 3	256	129 × 129	52.3
Level 4	512	257 × 257	39.4
Level 5 (deep)	512	513 × 513	27.1

The boundary response ratio is obtained by counting the average activation intensity of the feature map in the neighborhood of the manually annotated bounding box (set as a boundary band with a width of 15% of the target size) and the average activation intensity of the whole image in the validation set. The receptive field size is calculated based on the hierarchical structure of the backbone network and the convolution kernel superposition relationship theory, reflecting the changing trend of the feature map's coverage of the input image.

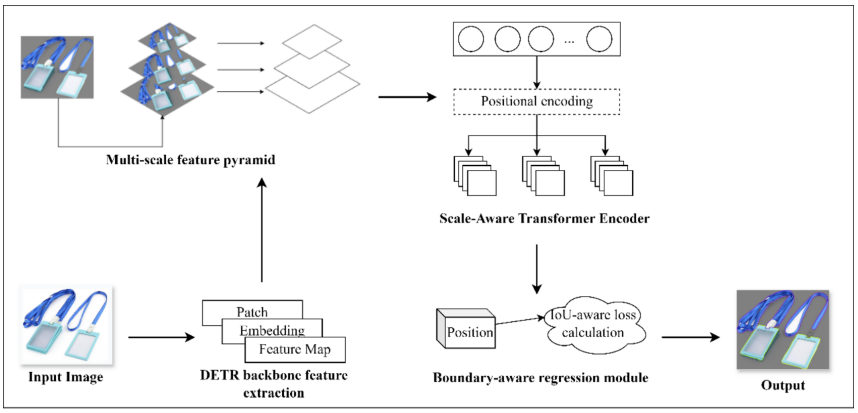

FIGURE 1. Multi-scale perception DETR structure

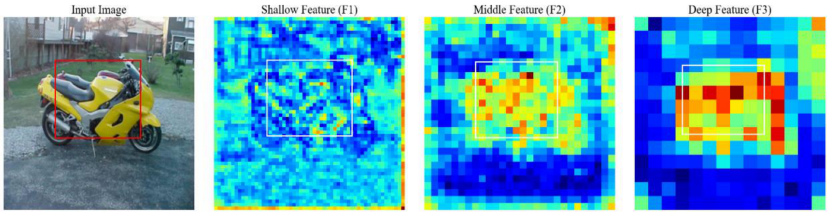

FIGURE 2. Response of feature maps of different scales in the target boundary area

The selection of this width parameter is based on an analysis of the relationship between boundary positioning uncertainty and feature response. Table 2 compares the correlation between the feature map boundary response ratio and the average intersection-union ratio of manually labeled boundary pixels under different boundary band widths.

TABLE 2. Basis for selecting different boundary band widths

Boundary Band Width (% of Target Size)	Boundary Response Ratio (%)	Correlation Coefficient with Ground-Truth Boundary Pixel IoU
5	58.3	0.72
10	69.5	0.85
15	78.2	0.91
20	83.7	0.89
25	86.1	0.85
30	87.9	0.81

As shown in Table 2, when the width is 15%, the boundary response ratio has the highest correlation (0.91) with the intersection-union ratio of the true boundary pixels, indicating that this width best represents the correspondence between the model feature activations and the true boundary. Therefore, 15% was chosen as the standard width for subsequent analysis.

C. SCALE-AWARE TRANSFORMER ENCODING MECHANISM

Based on the boundary expression ability of multi-scale fusion features, to improve the encoder's modeling accuracy of scale information, a scale-aware multi-head attention mechanism is applied into the original DETR encoding structure. The semantic drift discussed in this article refers to the phenomenon that features originating from different resolution levels become semantically misaligned or inconsistent during the multi-scale feature fusion process. This mechanism effectively alleviates the semantic drift and attention ambiguity problems between features of different scales by applying an attention path of explicit scale embedding and scale decoupling while maintaining the global modeling advantage of the self-attention structure. The scale-aware module is embedded in each multi-head attention layer of the Transformer encoder. The adjusted attention mechanism has consistent response to the scale context, ensuring multi-scale alignment modeling of the boundary area.

It is assumed that the fused feature F is expanded into a two-dimensional sequence $\{f_1, f_2, \ldots, f_N\}$, where $f_i \in R^C$ and $N = H \times W$ are the sequence lengths. Each position f_i is attached with a scale embedding vector $s_i \in R^{d_s}$, which encodes the scale attribution information of the corresponding position in the original pyramid. It is constructed through the standard sine-cosine encoding function and is in the form of Formula (2):

$$s_i^{(k)} = \begin{cases} \sin\left(\dfrac{\text{scale}_i}{10000^{2k/d_s}}\right), & \text{if } k \text{ is even,} \\ \cos\left(\dfrac{\text{scale}_i}{10000^{2k/d_s}}\right), & \text{if } k \text{ is odd.} \end{cases} \tag{2}$$

In Formula (2), scale_i represents the original scale number corresponding to position f_i; d_s is the scale embedding dimension; k is the dimension index. The scale embedding and position encoding are superimposed on the input feature,

so that the spatial and scale position information are considered simultaneously when calculating the attention weight.

The scale-decoupled multi-head attention mechanism applies scale-selective projection matrices $W_Q^{(h,l)}$, $W_K^{(h,l)}$, and $W_V^{(h,l)}$ in each head, where h represents the attention head number, and l represents the scale number. After the input features are divided according to the scale source, they are independently projected to generate Query, Key, and Value. Then, the attention is calculated within the scale, and the final output is obtained through inter-scale fusion, as shown in Formula (3):

$$\text{Attention}^{(h)} = \sum_{l=1}^{L} \beta_l^{(h)} \cdot \text{Softmax}\left(\frac{(Q^{(h,l)})(K^{(h,l)})^\top}{\sqrt{d_k}}\right) \times V^{(h,l)} \tag{3}$$

In Formula (3), $Q^{(h,l)} = W_Q^{(h,l)} f_l$, $K^{(h,l)} = W_K^{(h,l)} f_l$, $V^{(h,l)} = W_V^{(h,l)} f_l$, and $\beta_l^{(h)}$ are the fusion weights of the l-th scale in the h-th attention head. Normalization is performed through Softmax to ensure the continuity of attention output across scales. Finally, the outputs of all attention heads are concatenated and linearly transformed to obtain global encoding features with unified scale representation, which are input into the subsequent position-aware regression module.

To verify the effect of scale-aware encoding on the improvement of boundary extraction ability, Table 3 compares the attention focus area indicators under the conditions of scale-free modeling, using only scale embedding, and using a complete scale-aware mechanism. Among them, the focus position shift rate measures the degree of attention shift of the model in the boundary area, and the boundary retention rate reflects the consistency between the attention focus area and the real bounding box. Experimental data shows that the application of scale embedding can significantly reduce the attention shift, and the focus accuracy is further improved after the application of the scale decoupling mechanism, indicating that this mechanism effectively improves the global modeling accuracy of Transformer in multi-scale scenarios.

TABLE 3. Comparison of attention focus performance under different encoding strategies

Encoding Strategy	Focus Shift Rate (%)	Boundary Retention Rate (%)	Boundary Response Gain (dB)
Original DETR Encoder	18.7	62.3	0
With Scale Embedding	12.9	71.4	+2.1
Scale-Decoupled Attention (without Embedding)	10.5	75.2	+3.4
Scale Embedding + Scale-Decoupled Attention	7.8	81.7	+5.2
Multi-scale Embedding + Gated Attention Fuse	6.2	85.1	+6.7

D. REFINED BOUNDARY REGRESSION AND LOSS DESIGN

Based on the global encoding features represented by the unified scale, a boundary-aware regression module is designed to improve the precision and consistency of target boundary prediction. This regression module receives the feature sequence output by the Transformer encoder and outputs the bounding box parameters of each candidate target, including the center coordinates, width, and height information. To fully capture the spatial position information and boundary structure characteristics, a position embedding enhancement strategy is applied in the regression process, and the IoU-aware loss function and the boundary consistency regularization term are combined to construct a joint optimization target, thereby guiding the model to converge to the direction of optimal boundary accuracy during the training phase. It is assumed that the Transformer output is a sequence representation $\{z_1, z_2, \ldots, z_N\}$, where $z_i \in R^d$ is the encoding vector of the i-th position. Each vector is mapped to the bounding box parameters $(\hat{x}_i, \hat{y}_i, \hat{w}_i, \hat{h}_i)$ through a two-layer feedforward network, representing the horizontal and vertical coordinates and width and height of the center of the predicted box, respectively, and all are normalized to the range [0, 1]. To enhance the perception of boundary spatial structure, the position embedding vector p_i and the encoding feature z_i are spliced in the channel dimension, so that the regression function not only relies on semantic information, but also combines the spatial position signal, as shown in Formula (4):

$$[\hat{x}_i, \hat{y}_i, \hat{w}_i, \hat{h}_i] = \phi([z_i \| p_i]) \tag{4}$$

In Formula (4), $\phi(\cdot)$ represents the regression network structure, and $\|$ is the vector splicing operation.

After enhancement, the regression model is more discriminative for the location of the target edge.

This paper applies the IoU-aware loss function as the main component of the optimization target to improve the

boundary consistency of the regression result. The traditional regression loss only minimizes the difference between the predicted box and the true box in the numerical space, ignoring the spatial overlap relationship, while the IoU loss directly optimizes the overlap between the predicted box and the true box, so that the optimization direction is consistent with the detection task. Assuming that the predicted box is $B = (\hat{x}, \hat{y}, \hat{w}, \hat{h})$, and the true box is $B^* = (x^*, y^*, w^*, h^*)$, IoU is defined as the intersection-over-union ratio:

$$\mathrm{IoU}(B, B^*) = \frac{\mathrm{area}(B \cap B^*)}{\mathrm{area}(B \cup B^*)} \tag{5}$$

The loss function is defined as:

$$L_{\mathrm{IoU}} = 1 - \mathrm{IoU}(B, B^*) \tag{6}$$

This loss term reaches its minimum value when the predicted box and the true box overlap, which directly measures the boundary positioning accuracy. This paper constructs a boundary consistency regularization term to further constrain the stability of boundary shape, which is defined as the logarithmic square error of the ratio of the predicted box width and height to the true box width and height, in order to suppress shape drift and scale anomaly. When the target presents an elongated structure or occupies a very small spatial region, enforcing a global width-height ratio constraint introduces an implicit bias toward moderate aspect ratios. Under such conditions, the regularization term may reduce the flexibility of the regression head to follow extreme geometric configurations, resulting in conservative boundary fitting along the long axis. This effect is more pronounced in samples whose aspect ratio deviates persistently from the dataset-level distribution. The loss term is as follows:

$$L_{\mathrm{shape}} = \left(\log \frac{\hat{w}}{w^*}\right)^2 + \left(\log \frac{\hat{h}}{h^*}\right)^2 \tag{7}$$

The final total loss function is a weighted combination of multiple losses:

$$L_{\mathrm{total}} = \lambda_1 \cdot L_{\mathrm{IoU}} + \lambda_2 \cdot L_{\mathrm{shape}} + \lambda_3 \cdot L_{\mathrm{reg}} \tag{8}$$

L_{reg} is the basic coordinate regression loss, and λ_1, λ_2, and λ_3 are loss weight coefficients, which control the influence ratio of each item. This joint optimization strategy constrains boundary regression in terms of spatial overlap, structural consistency, and coordinate accuracy, guiding the model to form a stable and reliable boundary extraction capability during training. After repeated experimental tuning, this loss combination significantly reduces the prediction error fluctuation while maintaining the bounding box convergence speed, and improves the continuity and integrity of the bounding box in the final detection result.

IV. EXPERIMENTAL SETUP AND IMPLEMENTATION PROCESS

A. DATASET AND PREPROCESSING STRATEGY

To ensure that the multi-scale perception DETR structure has stable boundary extraction capabilities under different data environments, this paper designs a systematic data preprocessing strategy and quantifies the impact of each combination on the model training process through multiple sets of experiments. The main datasets used in the experiment are Cityscapes and BDD100K. Cityscapes contains 5,000 high-resolution street scene images, which are widely used in semantic segmentation and boundary detection tasks and have clear edge annotation specifications; BDD100K contains 100,000 driving video frame images of multiple time periods and scenes, with rich target scale changes and complex background information, which is suitable for evaluating the boundary perception ability of the model in real traffic scenes. The Cityscapes dataset is divided into training set (2,975 images), validation set (500 images), and test set (1,525 images) according to the official classification standard, and the BDD100K dataset is divided into training set (70,000 images), validation set (10,000 images), and test set (20,000 images). The comparison summary under different preprocessing methods is shown in Table 4.

TABLE 4. The effect of different preprocessing combinations on model training stability

Preprocessing Combination ID	Image Scaling Strategy	Bounding Box Normalization	Normalization Method	Initial Loss Fluctuation Rate (%)	Convergence Epochs
P1	Fixed-size scaling (800 × 800)	Relative coordinate normalization	Mean-Std normalization	14.7	52
P2	Multi-scale random scaling (640-1024)	Absolute coordinate normalization	Mean-Std normalization	11.3	46
P3	Multi-scale random scaling (640-1024)	Relative coordinate normalization	Mean-Std normalization	9.5	44
P4	Fixed-size scaling (800 × 800)	Absolute coordinate normalization	Division by 255	17.2	58
P5	Multi-scale random scaling (640-1024)	Relative coordinate normalization	Division by 255	13.8	50

B. MODEL CONFIGURATION AND TRAINING DETAILS

To ensure the training efficiency and deployment performance of the multi-scale perception DETR model, this paper conducts systematic experiments on different training configurations, focusing on the impact of factors such as optimizer, learning rate strategy, and batch size on resource consumption and inference efficiency. The performance differences of different configurations in the hardware environment are directly related to the actual application effect of the model. Table 5 shows the detailed performance indicators under multiple sets of training configurations.

TABLE 5. Multi-scale perception DETR training configuration and performance comparison

Configuration ID	Optimizer Type	Learning Rate Schedule	Batch Size	Training Epochs	Memory Usage (GB)
Config-A	AdamW	Cosine Annealing Decay	16	150	11.3
Config-B	AdamW	Linear Decay	16	150	11.1
Config-C	SGD (Stochastic Gradient Descent)	Cosine Annealing Decay	16	150	10.8
Config-D	AdamW	Cosine Annealing Decay	32	150	17.9
Config-E	AdamW	Cosine Annealing Decay	16	200	11.4

Table 5 compares the model's memory usage, single-round training time, and forward inference latency under different optimizer types, learning rate strategies, and batch size combinations. The table shows slight fluctuations in memory usage among different configurations with the same batch size; for example, Config-A differs from Config-B by 0.2 GB. This difference stems from instantaneous fluctuations in GPU memory dynamic management and measurement during training, and its magnitude is within the measurement error range. In multiple repeated measurements under a fixed hardware environment, the standard deviation of memory usage for all configurations was less than 0.15 GB. The data shows that the use of the AdamW optimizer combined with the cosine annealing decay strategy and a medium batch size can help balance training speed and inference efficiency while controlling memory usage. Increasing the batch size causes a simultaneous increase in memory consumption, and the configuration scheme needs to be weighed in combination with actual hardware conditions. The resource demand change trend under different training rounds also provides a parameter basis for the stable training and efficient deployment of the model.

V. RESULTS

A. IMPACT OF MULTI-SCALE PERCEPTION STRUCTURE ON BOUNDARY EXTRACTION ACCURACY

To deeply analyze the actual impact of the multi-scale perception mechanism on the accuracy of target boundary extraction, this paper designs a set of structural ablation experiments, constructs five comparison methods around the bounding box positioning performance, and evaluates them through quantitative and statistical methods. The baseline method is the original DETR, followed by DETR+FPN with FPN, DETR+FPN+SA with scale-aware attention (SA), DETR+Full with full encoding structure but without refined regression module, and finally the complete method proposed in this paper, which adds boundary-aware regression head and IoU-aware loss function to the first three. The relevant results are shown in Figure 3.

As can be seen from Figure 3 (a), the average boundary IoU of the original DETR is 79.6%. Due to its lack of adaptability to different target scales, many bounding boxes have insufficient coverage. After applying FPN, the IoU increases to 82.1%, which is attributed to the enhanced response of multi-scale features to small-scale boundaries. Further adding the scale-aware attention mechanism enables the model to distinguish the semantic and structural differences of each scale during modeling, and the IoU is increased to 83.9%. On this basis, the complete multi-scale perception structure is integrated, and the IoU reaches 85.3%, indicating that scale consistency modeling can effectively improve the fitting ability of the bounding box for complex targets. After further applying the boundary-aware regression head and IoU loss optimization, the proposed method achieves an average IoU of 86%, which is better than all the compared structures, reflecting the optimization effect at the level of boundary positioning details. The vertical axis in Figure 3(b) represents the cumulative probability of boundary positioning error, which is used to reflect the proportion of predicted bounding boxes with errors less than a certain pixel value in all samples. The larger the value, the more concentrated the boundary positioning and the higher the accuracy. The CDF curve shows that the proposed method has the highest cumulative probability in the low error section, and the curve is close to the upper left corner as a whole, indicating that its boundary positioning error distribution is more concentrated and stable.

To clarify the respective contributions of the multi-scale sensing structure and the IoU sensing loss function, a detailed decoupling experiment was designed, and the results are shown in Table 6.

TABLE 6. Analysis of the contribution of different components to the boundary IoU

Experimental Configuration	Mean IoU	Relative Improvement over Baseline
Original DETR (Baseline)	0.81	–
IoU-Aware Loss Only	0.83	+0.02
Multi-Scale Aware Architecture Only (FPN + SA)	0.845	+0.035
Full Method (Multi-Scale Architecture + IoU Loss)	0.86	+0.05

Table 6 shows that applying the IoU sensing loss alone improves IoU by 0.02, while applying the multi-scale sensing structure alone improves it by 0.035. The combined effect of both results in an improvement of 0.05, slightly higher than the simple summation, indicating a synergistic effect between structural improvement and loss optimization. The multi-scale structure contributes the majority of the performance gain.

B. IMPACT OF DIFFERENT SCALE INPUTS ON MODEL BOUNDARY EXTRACTION PERFORMANCE

To deeply analyze the impact of different input scales on target boundary prediction performance, this section of the experiment constructs three input forms based on the

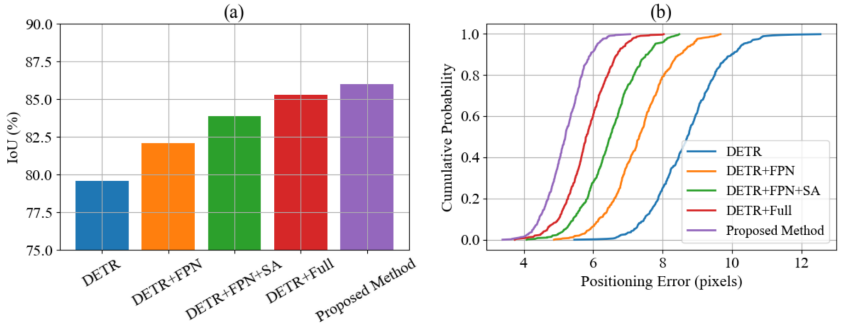

FIGURE 3. The impact of multi-scale structure on bounding box extraction accuracy: (a) Bounding box IoU comparison under different structures; (b) CDF distribution of boundary positioning error of each method

original image: the original image with the original resolution, the image downsampled to half the resolution using average pooling, and the enhanced version of the image after horizontal flipping and contrast enhancement. The model performs boundary regression prediction under these three types of inputs and presents the spatial distribution of the predicted target in the form of a red bounding box in the output results. All input forms keep the target semantics unchanged, aiming to eliminate the interference of content differences and analyze the stability and difference of the model's boundary perception ability from the perspective of input scale and distribution perturbation. The images cover a variety of scenes and target forms, which helps to reveal the universal impact of different scale inputs on the boundary extraction effect, thereby obtaining the experimental results shown in Figure 4.

As can be seen from the results in Figure 4, the predicted boundary under the original image input shows a good fit on the actual contour of the target; the boundary position is relatively close to the edge of the target; the boundary box under the downsampled image generally has edge blur and position offset. The fundamental reason is that the low-resolution input weakens the expression ability of key boundary features, resulting in the lack of sufficient fine-grained structural information support in the feature modeling process of the encoder, thereby affecting the spatial positioning accuracy of the boundary regression module. Under image enhancement input, some target boundary prediction results are closer to the semantic center area in spatial position, which is manifested as boundary reduction or regional offset.

This is because the disturbance of the image contrast and edge gradient distribution caused by the enhancement operation affects the spatial focusing behavior of the model's attention mechanism, causing a slight shift in scale sensitivity at the perceptual level. This shows that the model is more sensitive to the resolution and structural disturbance of the input image when extracting the boundary, and the input scale design has a significant impact on the boundary regression performance.

C. EFFECT OF SCALE-AWARE MECHANISM IN ENCODING MODULE

To analyze the effect of scale-aware attention mechanism on spatial focusing characteristics, this experiment selects three images containing typical vehicle targets as input. Under the premise of keeping the backbone structure consistent, the attention distribution of the model under the application of scale-aware mechanism is shown respectively. The original version uses the standard multi-head attention mechanism, while the improved version incorporates the scale embedding vector and scale decoupling weight learning strategy in the encoding stage to construct a global attention map with consistent scale. In the visualization, the attention heat map is superimposed on the original image to show the significant difference in the focus area, and the rightmost column shows the response distribution of the improved model to the overall target area, as shown in Figure 5.

It can be observed in Figure 5 that when the scale mechanism is not applied, the model's attention is focused on the rear of the vehicle or the local high-texture area, showing a strong position bias. The root cause of this bias phenomenon is that the standard attention lacks the ability to model multi-scale contextual structures, resulting in its reliance on local saliency rather than the overall geometric form of the target in feature selection. After the scale-aware mechanism is applied, the attention hotspot tends to cover the entire target contour, showing a smoother and more coherent focus pattern. This change is attributed to the scale information participating in the attention weight generation process, which enables the model to maintain cross-level feature alignment in the spatial dimension, thereby achieving a unified response to the overall target area. Finally, the scale-aware mechanism effectively promotes the global modeling ability of target structure perception by improving the consistency of attention expression.

D. OPTIMIZATION OF BOUNDARY POSITIONING QUALITY BY REGRESSION HEAD AND LOSS FUNCTION DESIGN

To evaluate the comprehensive impact of the proposed boundary-aware regression structure and IoU-aware loss on the target boundary positioning performance, this paper

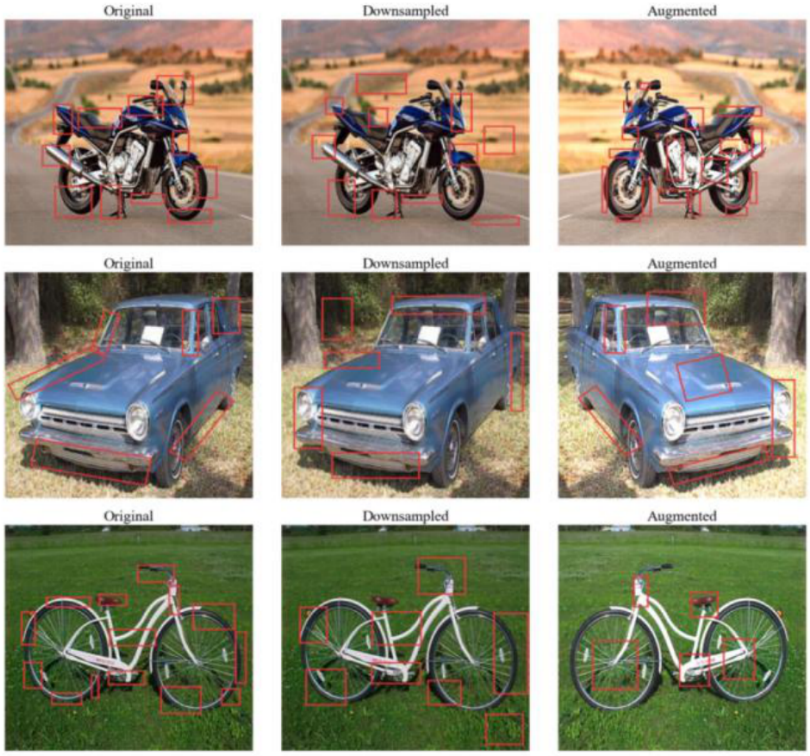

FIGURE 4. Target boundary prediction effect under different scale inputs

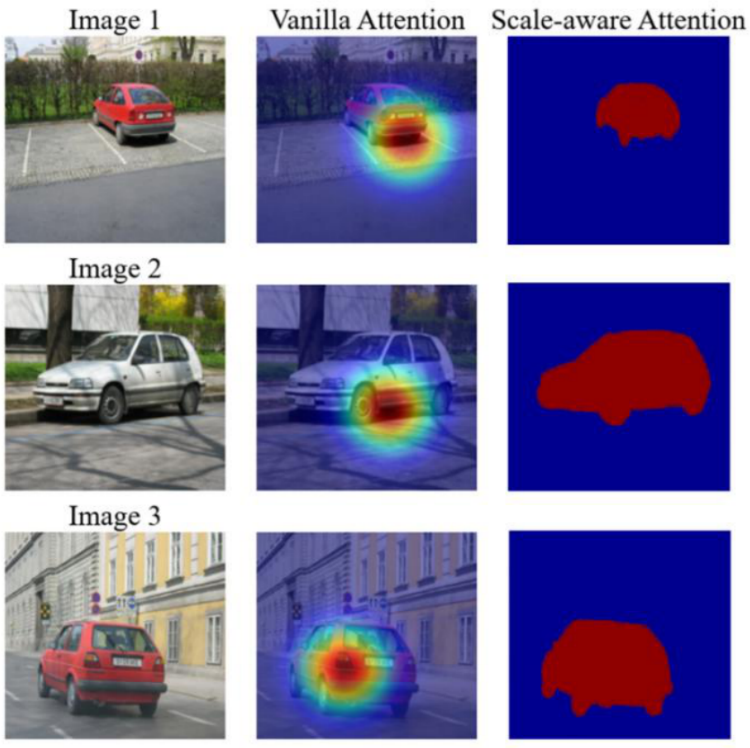

FIGURE 5. Comparison of attention distribution under scale-aware mechanism

constructs four groups of different regression configuration experiments under the unified dataset and model backbone architecture: the basic L1+GIoU loss is used as the control group, and on this basis, the position embedding mechanism

PosEmbed (Positional Embedding), IoU-aware loss function IoU-Loss, and boundary consistency regularization term Reg (Regularization) are applied in turn to observe the contribution of each strategy to the bounding box accuracy and positioning quality. During the experiment, the number of training rounds, data augmentation strategy, and optimizer parameters are kept consistent to control variables and remove interference from other factors. Finally, the bounding box AP value and average IoU performance under different strategies are statistically analyzed. The results are shown in Figure 6.

The data in Figure 6 shows that after the application of position embedding, AP@50 increases from 42.3% to 45.0%; AP@75 increases from 31.0% to 34.5%; the average IoU increases from 0.78 to 0.81. This change is due to the fact that position encoding strengthens the spatial semantic alignment ability in boundary regression, making it easier for the model to learn the actual spatial distribution of the target box. After further adding IoU-aware loss, AP@50 increases to 47.6%; AP@75 reaches 36.2%; IoU increases to 0.85.

This result benefits from the high sensitivity of the loss function to the boundary overlap area, thereby guiding the model to optimize the positioning accuracy rather than simple coordinate regression during gradient return. Finally, after the application of the boundary consistency regularization term, AP@50 increases slightly to 48.1%, and IoU stabilizes to 0.86. To examine the behavior of the shape regularization under extreme geometric conditions, the validation set is partitioned according to ground-truth aspect ratio. For instances with an aspect ratio larger than 5 or smaller than 0.2, the regression performance shows a reduced gain from the regularization term compared with medium-ratio targets, indicating that the constraint mainly benefits shape-stable instances while limiting boundary adaptation in highly elongated cases. This improvement is attributed to the fact that the regularization term suppresses the tendency of the predicted box to deviate from the actual boundary shape, improving the rationality of the overall structure of the boundary box. This series of experiments verifies the synergistic gain effect of regression structure improvement and loss design in improving boundary accuracy.

To examine the different effects of boundary shape constraints on targets with different geometric shapes, the validation set samples were re-divided according to the aspect ratio distribution of the real bounding boxes, classifying the targets into three categories: medium proportion, very small proportion, and high proportion. While maintaining consistency in training configuration and evaluation process, the detection results with and without the inclusion of boundary consistency constraints were compared. The focus was on recording boundary overlap behavior within different aspect ratio ranges, thus characterizing the adaptive behavior of this constraint under extreme geometric conditions. The relevant results are shown in Table 7.

TABLE 7. Effect of boundary shape regularization under different aspect ratio ranges

Aspect Ratio Range	IoU without Reg	IoU with Reg
0.2 – 5.0	0.84	0.86
< 0.2	0.81	0.80
> 5.0	0.82	0.81

As shown in Table 7, when the aspect ratio is between 0.2 and 5.0, the IoU increases from 0.84 to 0.86 after introducing the shape constraint, indicating that this constraint strengthens boundary consistency for geometrically stable targets. When the aspect ratio is less than 0.2, the IoU decreases from 0.81 to 0.80, while when the aspect ratio is greater than 5.0, the IoU decreases from 0.82 to 0.81. This change indicates that the global scaling constraint suppresses the boundary degrees of freedom in extreme cases. This is because the logarithmic scaling penalty introduces a shape bias at the data distribution level, limiting the model's ability to unfold boundaries along the major axis. Overall, the results show that this constraint is more suitable for structurally stable targets, and should be used with caution for targets that are tall and slender or have limited scale.

E. COMPARISON OF BOUNDARY EXTRACTION WITH EXISTING MAINSTREAM METHODS

To comprehensively evaluate the performance of the proposed multi-scale perception structure in the target boundary extraction task, this paper constructs a unified evaluation framework and compares the numerical indicators of 7 representative target detection methods in terms of boundary accuracy. Faster R-CNN (Faster Region-based Convolutional Neural Network), as a representative of the two-stage detection method, has a strong region proposal capability, but there is a boundary regression offset in dense target scenes. YOLOv5 (You Only Look Once version 5) relies on a unified regression structure for end-to-end detection and has high detection efficiency, but its spatial resolution utilization is relatively coarse.

RetinaNet combines FPN structure and Focal Loss to suppress foreground-background imbalance and has certain advantages in multi-scale target detection, but the boundary fitting accuracy is limited by its regression head structure. The original DETR adopts an encoder-decoder end-to-end Transformer architecture, which effectively avoids redundant candidate boxes, but is limited by fixed-scale feature processing and has limited ability to depict boundary details. The Deformable-DETR applies a deformable attention mechanism into the DETR framework and uses dynamic sampling to enhance the adaptability of the boundary feature response area, showing better boundary fault tolerance in occlusion and scale transformation scenarios. Elastic-DETR further combines the elastic feature fusion strategy to improve the scale drift problem in the boundary prediction process through scale invariance constraints, and improves the fitting accuracy of the bounding box to the real

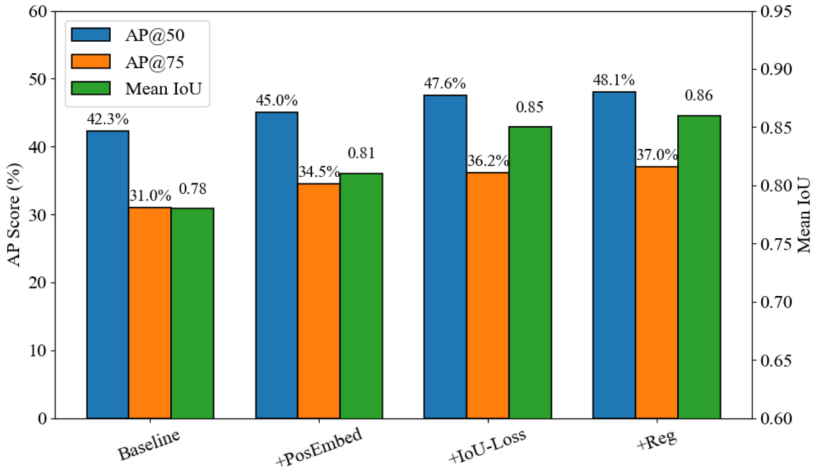

FIGURE 6. Impact of boundary regression configuration on detection accuracy

contour. Based on DETR, this paper applies a multi-scale perception mechanism and synergistically improves the boundary representation capability through pyramid feature enhancement and scale consistency modeling. In addition, to compare with more advanced detectors, DINO-DETR and H-DETR, based on the DETR framework, were introduced as recent representative models. In the experiment, the AP and bounding box IoU performance of each method on the standard validation set are recorded, as shown in Figure 7.

The results in Figure 7 show that the proposed method has significant advantages in boundary extraction, with an AP of 48.1% and a bounding box IoU of 0.86, which are improved compared to the original DETR's 44.9% and 0.81. In contrast, YOLOv5 has an AP of 43.8% and an IoU of 0.791. Although it performs better than Faster R-CNN and RetinaNet, it still has the problem of inaccurate boundary alignment. The reason is that the YOLO series model's compression strategy for feature resolution limits the modeling ability of small-scale boundaries. Faster R-CNN has an IoU of 0.765. Due to the positioning bias of using RoI Pooling, its robustness to boundary prediction is poor. Although the original DETR has good global modeling capabilities, it lacks the ability to distinguish and express boundary features of different scales, resulting in limited improvement in boundary overlap. The Deformable-DETR enhances multi-scale feature perception by applying a deformable sampling mechanism, which increases the AP to 46.2% and the IoU to 0.834; the Elastic-DETR further improves the boundary adaptation capability by using an elastic alignment strategy, with an AP of 47.3% and an IoU of 0.846; DINO-DETR achieved an AP of 47.8% and an IoU of 0.849; H-DETR achieved an AP of 47.5% and an IoU of 0.844. The proposed method enhances the boundary semantic consistency by applying multi-scale feature fusion and scale-aware attention, effectively improving the accuracy and stability of bounding box regression.

F. RELATIONSHIP BETWEEN BOUNDARY ERROR AND SCALE

To further explore the impact of target scale change on the distribution of boundary prediction error, this paper constructs a scale sensitivity analysis experiment based on the overlap error between the true bounding box area and the predicted result of targets of different sizes in the validation set. By extracting the actual size of each target instance and calculating the IoU difference between the corresponding predicted bounding box and the true bounding box, the original DETR method and the multi-scale perception structure proposed in this paper are respectively annotated for error. Further, the scatter plot method shown in Figure 8 is used to map the target scale and prediction error in a bivariate manner, and the overall trend is modeled based on the local weighted regression curve to obtain the error distribution changes of the two methods in each scale segment.

Figure 8 shows that the original DETR method has a significantly higher distribution of boundary errors. The reason is that the fixed-scale feature extraction structure lacks sufficient local detail perception when processing small-size areas, resulting in sparse distribution of feature responses in the boundary area, which in turn affects the encoder's modeling effect on the boundary geometry contour. As the target size increases, the error of the DETR model decreases slightly, but due to the lack of clear modeling of feature scale attribution, there is semantic drift in the cross-scale fusion process, showing insufficient stability of boundary prediction. In contrast, the error of the proposed method in small target areas is lower, which is due to the scale embedding mechanism applied in the encoding stage, strengthening the consistency of boundary responses of features with different resolutions, and allowing the model to retain scale structure information in the global modeling process and reduce the structural uncertainty of the error source.

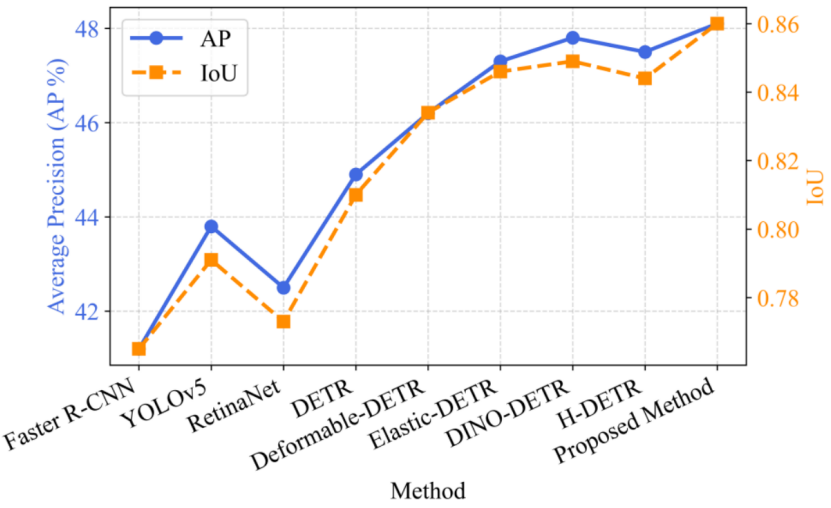

FIGURE 7. Comparison of boundary extraction accuracy of target detection methods

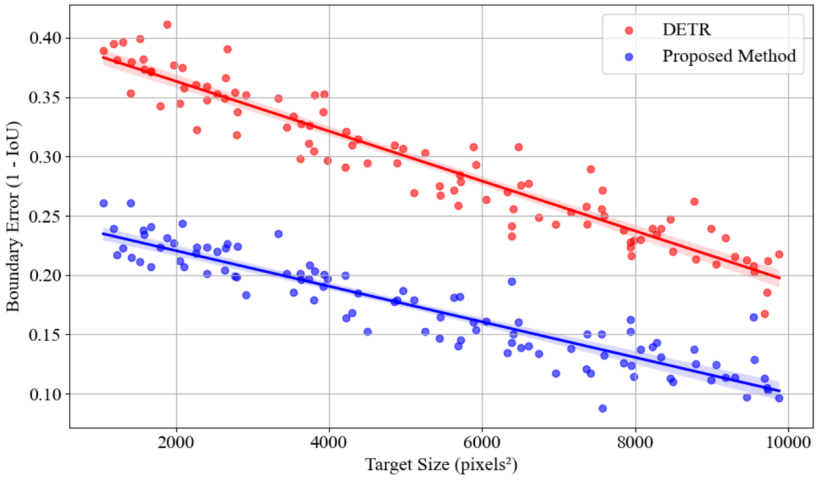

FIGURE 8. Relationship between target scale and boundary prediction error

G. ROBUSTNESS ANALYSIS

To evaluate the adaptability of the proposed scale-aware mechanism to different data preprocessing conditions, the complete model was tested based on the five preprocessing combinations (P1-P5) defined in Section Dataset and Preprocessing Strategy, and its boundary extraction accuracy on the validation set was statistically analyzed. The results are shown in Table 8.

TABLE 8. Boundary extraction performance of the model under different preprocessing conditions

Preprocessing Combination	Average IoU	AP@50 (%)
P1 (Fixed-size)	0.85	47.3
P2 (Multi-scale, Abs)	0.86	47.9
P3 (Multi-scale, Rel)	0.86	48.1
P4 (Fixed-size, Div255)	0.84	46.8
P5 (Multi-scale, Div255)	0.85	47.5

Analyzing the data in Table 8, the P3 combination achieved an IoU of 0.86 and an AP@50 of 48.1%, making it the optimal result. This demonstrates that the preprocessing strategy combining multi-scale random input and relative coordinate normalization is most beneficial for the model to realize its boundary extraction potential. This is because multi-scale training enhances the model's generalization ability to changes in target scale, while relative coordinate normalization provides a stable basis for position regression. The IoUs of the P1 and P4 combinations are 0.85 and 0.84, respectively, indicating that using only fixed-size input limits the model's adaptability to scale information, and different normalization methods have a slight impact on performance. The IoU of the P5 combination is 0.85, with performance between P3 and P1, indicating that after changing the normalization method, the gain from multi-scale training is offset, but the model still maintains better performance than a single fixed scale. These results demonstrate that

the integrated scale-aware encoding mechanism is robust to changes in data distribution introduced by different preprocessing strategies, ensuring the stability of the method's performance.

VI. CONCLUSION

This paper proposes an improved DETR structure that integrates a multi-scale perception mechanism, aiming to improve the accuracy of target boundary extraction in image data. The method utilizes a multi-scale encoding module based on an FPN to extract boundary and semantic features across resolutions.

It constructs a scale-aware multi-head attention mechanism in the Transformer encoder to enhance scale consistency during global modeling. In the decoder, a boundary regression head with position embedding is designed, optimized by an IoU-aware loss function and a boundary consistency regularization term to improve regression accuracy. Experiments show the proposed method achieved an average IoU of 0.86 and an AP value of 48.1%, representing improvements over the original DETR, and also surpassed contemporary DETR variants including Deformable-DETR, Elastic-DETR, DINO-DETR, and H-DETR in boundary extraction accuracy.

Ablation studies confirmed that the multi-scale structure, scale embedding, and joint loss optimization significantly contributed to boundary information modeling, ultimately enhancing the expression and recognition of target boundary structures while maintaining DETR's end-to-end advantages. It is particularly suitable for complex image detection scenarios with blurred boundaries, dense distribution of small targets, or drastic scale changes, showing important application potential in high-precision visual tasks.

AUTHOR CONTRIBUTIONS

Yancen Liu designed, collected and analyzed the data, and drafted the manuscript. Yancen Liu conducted the study, critically revised the manuscript for important intellectual content, and gave final approval of the version to be published. Yancen Liu participated fully in the work, take public responsibility for appropriate portions of the content, and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

CONFLICTS OF INTEREST

The author declares no conflict of interest.

FUNDING

This research received no external funding.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] O. Akcay, A. C. Kınacı, E. O. Avsar, and U. Aydar, "Boundary Extraction Based on Dual Stream Deep Learning Model in High Resolution Remote Sensing Images," Journal of Advanced Research in Natural and Applied Sciences, vol. 7, no. 3, pp. 358-368, 2021, doi: 10.28979/jarnas.911130.

[2] L. Ma, Y. Li, J. Li, J. M. Junior, W. N. Goncalves, and M. A. Chapman, "Boundarynet: Extraction and completion of road boundaries with deep learning using mobile laser scanning point clouds and satellite imagery," IEEE Transactions on Intelligent Transportation Systems, vol. 23, no. 6, pp. 5638-5654, 2022, doi: 10.1109/TITS.2021.3055366.

[3] A. Abdollahi, B. Pradhan, and A. Alamri, "SC-RoadDeepNet: A new shape and connectivity-preserving road extraction deep learning-based network from remote sensing data," IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1-15, 2022, doi: 10.1109/TGRS.2022.3143855.

[4] Z. Wang, Y. L. Li, X. Chen, H. Zhao, and S. Wang, "Uni3detr: Unified 3d detection transformer," in Oh A, Naumann T, Globerson A, Saenko K, Hardt M, Levine S, editors. Advances in Neural Information Processing Systems 36. Proceedings of the 37th Conference on Neural Information Processing Systems, 10-16 December 2023, New Orleans, Louisiana, USA. New York, NY, USA: Curran Associates, Inc., 2023, pp. 39876-39896, doi: 10.48550/arXiv.2310.05699.

[5] L. Dai, H. Liu, H. Tang, Z. Wu, and P. Song, "AO2-DETR: Arbitrary-oriented object detection transformer," IEEE Transactions on Circuits and Systems for Video Technology, vol. 33, no. 5, pp. 2342-2356, 2022, doi: 10.1109/TCSVT.2022.3222906.

[6] M. A. Munir, S. H. Khan, M. H. Khan, M. Ali, and F. S. Khan, "Cal-DETR: Calibrated detection transformer," in Oh A, Naumann T, Globerson A, Saenko K, Hardt M, Levine S, editors. Advances in Neural Information Processing Systems 36. Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS 2023), 10-16 December 2023, New Orleans, Louisiana, USA. New York, NY, USA: Curran Associates, Inc., 2023, pp. 71619-71631, doi: https://doi.org/10.48550/arXiv.2310.05699.

[7] Y. Zeng, Y. Chen, X. Yang, Q. Li, and J. Yan, "ARS-DETR: Aspect ratio-sensitive detection transformer for aerial oriented object detection," IEEE Transactions on Geoscience and Remote Sensing, vol. 62, pp. 1-15, 2024, doi: 10.1109/TGRS.2024.3364713.

[8] X. Song, Z. Hua, and J. Li, "Remote sensing image change detection transformer network based on dual-feature mixed attention," IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1-16, 2022, doi: 10.1109/TGRS.2022.3209972.

[9] Y. Liu, P. Nand, M. A. Hossain, M. Nguyen, and W. Q. Yan, "Sign language recognition from digital videos using feature pyramid network with detection transformer," Multimedia Tools and Applications, vol. 82, no. 14, pp. 21673-21685, 2023, doi: 10.1007/s11042-023-14646-0.

[10] Q. Li, G. Zhong, C. Xie, and R. Hedjam, "Weak edge identification network for ocean front detection," IEEE Geoscience and Remote Sensing Letters, vol. 19, pp. 1-5, 2022, doi: 10.1109/lgrs.2021.3051203.

[11] G. Singh and K. Singh, "Chroma key foreground forgery detection under various attacks in digital video based on frame edge identification," Multimedia Tools and Applications, vol. 81, no. 1, pp. 1419-1446, 2022, doi: 10.1007/s11042-021-11380-3.

[12] M. S. Sivapriya and S. Suresh, "ViT-DexiNet: A vision transformer-based edge detection operator for small object detection in SAR images," International Journal of Remote Sensing, vol. 44, no. 22, pp. 7057-7084, 2023, doi: 10.1080/01431161.2023.2277167.

[13] Y. Xie, Z. Tu, T. Yang, Y. Zhang, and X. Zhou, "EdgeFormer: Local patch-based edge detection transformer on point clouds," Pattern Analysis and Applications, vol. 28, no. 1, pp. 1-13, 2025, doi: 10.1007/s10044-024-01386-6.

[14] Z. Liu and J. Cheng, "Cb-fpn: Object detection feature pyramid network based on context information and bidirectional efficient fusion," Pattern Analysis and Applications, vol. 26, no. 3, pp. 1441-1452, 2023, doi: 10.1007/s10044-023-01173-9.

[15] Q. Liu, J. Bi, J. Zhang, X. Bu, and N. Hanajima, "B-FPN SSD: An SSD algorithm based on a bidirectional feature fusion pyramid," The Visual Computer, vol. 39, no. 12, pp. 6265-6277, 2023, doi: 10.1007/s00371-022-02727-4.

[16] X. Fu, Z. Yuan, T. Yu, and Y. Ge, "DA-FPN: Deformable convolution and feature alignment for object detection," Electronics, vol. 12, no. 6, pp. 1354-1368, 2023, doi: 10.3390/electronics12061354.

[17] H. Wu, J. Liu, T. Jiang, Q. Zou, S. Qi, Z. Cui, et al., "AttentionMGT-DTA: A multi-modal drug-target affinity prediction using graph transformer and

attention mechanism," Neural Networks, vol. 169, pp. 623-636, 2024, doi: 10.1016/j.neunet.2023.11.018.
[18] Y. Ding and M. Jia, "Convolutional transformer: An enhanced attention mechanism architecture for remaining useful life estimation of bearings," IEEE Transactions on Instrumentation and Measurement, vol. 71, pp. 1-10, 2022, doi: 10.1109/tim.2022.3181933.
[19] F. P. An and J. E. Liu, "Medical image segmentation algorithm based on multilayer boundary perception-self attention deep learning model," Multimedia Tools and Applications, vol. 80, no. 10, pp. 15017-15039, 2021, doi: 10.1007/s11042-021-10515-w.
[20] H. Wang, Y. Jin, H. Ke, and X. Zhang, "DDH-YOLOv5: Improved YOLOv5 based on Double IoU-aware Decoupled Head for object detection," Journal of Real-Time Image Processing, vol. 19, no. 6, pp. 1023-1033, 2022, doi: 10.1007/s11554-022-01241-z.
[21] Y. Shin, J. Park, H. Song, S. Yoon, B. S. Lee, and J. G. Lee, "Exploiting representation curvature for boundary detection in time series," in Globerson A, Mackey L, Belgrave D, Fan A, Paquet U, Tomczak J, Zhang C, editors. Advances in Neural Information Processing Systems 37. Proceedings of the 38th Conference on Neural Information Processing Systems, 10-15 December 2024, Vancouver, Canada. New York: Curran Associates, Inc., 2024, pp. 5974-5995, doi: 10.52202/079017-0194.
[22] Y. Ai, R. Song, C. Huang, C. Cui, B. Tian, and L. Chen, "A real-time road boundary detection approach in surface mine based on meta random forest," IEEE Transactions on Intelligent Vehicles, vol. 9, no. 1, pp. 1989-2001, 2024, doi: 10.1109/tiv.2023.3296767.
[23] M. Ahmed, N. El-Sheimy, H. Leung, and A. Moussa, "Enhancing object detection in remote sensing: A hybrid yolov7 and transformer approach with automatic model selection," Remote Sensing, vol. 16, no. 1, pp. 51-67, 2024, doi: 10.3390/rs16010051.
[24] K. M. Elgamily, M. A. Mohamed, A. M. Abou-Taleb, and M. M. Ata, "Enhanced object detection in remote sensing images by applying metaheuristic and hybrid metaheuristic optimizers to YOLOv7 and YOLOv8," Scientific Reports, vol. 15, no. 1, pp. 7226-7256, 2025, doi: 10.1038/s41598-025-89124-8.
[25] H. Nguyen, T. Q. Ngo, H. T. T. Uyen, and M. K. Duong, "Enhanced object recognition from remote sensing images based on hybrid convolution and transformer structure," Earth Science Informatics, vol. 18, no. 2, pp. 228, 2025, doi: 10.1007/s12145-025-01751-x.
[26] Y. Zang, C. Fu, D. Yang, H. Li, C. Ding, and Q. Liu, "Transformer fusion and histogram layer multispectral pedestrian detection network," Signal, Image and Video Processing, vol. 17, no. 7, pp. 3545-3553, 2023, doi: 10.1007/s11760-023-02579-y.
[27] K. Huang, M. Wen, C. Wang, and L. Ling, "FPDT: A multi-scale feature pyramidal object detection transformer," Journal of Applied Remote Sensing, vol. 17, no. 2, pp. 026510-026510, 2023, doi: 10.1117/1.jrs.17.026510.
[28] X. Zhang, Z. Chen, J. Zhang, T. Liu, and D. Tao, "Learning general and specific embedding with transformer for few-shot object detection," International Journal of Computer Vision, vol. 133, no. 2, pp. 968-984, 2025, doi: 10.1007/s11263-024-02199-0.
[29] J. Liu, Y. Xia, J. Feng, and P. Bai, "A Novel Building Extraction Network via Multi-Scale Foreground Modeling and Gated Boundary Refinement," Remote Sensing, vol. 15, no. 24, pp. 5638-5665, 2023, doi: 10.3390/rs15245638.
[30] Q. Yuan, "Building rooftop extraction from high resolution aerial images using multiscale global perceptron with spatial context refinement," Scientific Reports, vol. 15, no. 1, pp. 6499-6513, 2025, doi: 10.1038/s41598-025-91206-6.
[31] H. Zhang, X. Zheng, N. Zheng, and W. Shi, "A multiscale and multipath network with boundary enhancement for building footprint extraction from remotely sensed imagery," IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 15, pp. 8856-8869, 2022, doi: 10.1109/jstars.2022.3214485.
[32] G. Yan, H. Jing, H. Li, H. Guo, and S. He, "Enhancing building segmentation in remote sensing images: Advanced multi-scale boundary refinement with MBR-HRNet," Remote Sensing, vol. 15, no. 15, pp. 3766-3784, 2023, doi: 10.3390/rs15153766.
[33] J. Chang, X. He, P. Li, T. Tian, X. Cheng, M. Qiao, et al., "Multi-scale attention network for building extraction from high-resolution remote sensing images," Sensors, vol. 24, no. 3, pp. 1010-1026, 2024, doi: 10.3390/s24031010.
[34] Z. Wang, Y. Zhou, F. Wang, S. Wang, G. Qin, W. Zou, et al., "A multi-scale edge constraint network for the fine extraction of buildings from remote sensing images," Remote Sensing, vol. 15, no. 4, pp. 927-946, 2023, doi: 10.3390/rs15040927.