IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 26 May 2025, accepted 2 July 2025, date of publication 15 July 2025, date of current version 23 July 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3589159

RESEARCH ARTICLE

Improved RT-DETR Framework for Railway Obstacle Detection

PENG LI[1], YANHUI PENG[1,2], SU-MEI WANG[2,3], (Member, IEEE), AND CHENG ZHONG[1]

[1]School of Railway Tracks and Transportation, Wuyi University, Jiangmen 529020, China
[2]National Rail Transit Electrification and Automation Engineering Technology Research Center (Hong Kong Branch), Hong Kong
[3]Department of Civil and Environmental Engineering, The Hong Kong Polytechnic University, Hong Kong

Corresponding author: Su-Mei Wang (may.sm.wang@polyu.edu.hk)

This work was supported in part by Wuyi University's Hong Kong and Macao Joint Research and Development Fund under Grant 2019WGALH15 and Grant 2019WGALH17, and in part by the Innovation and Technology Commission of Hong Kong SAR Government to Hong Kong Branch of Chinese National Rail Transit Electrification and Automation Engineering Technology Research Center under Grant K-BBY1.

ABSTRACT Obstacle intrusion detection in railway systems is a critical technology for ensuring the operational safety of trains. However, existing algorithms face challenges related to insufficient multiscale object detection, high model redundancy, and poor real-time performance. Building upon the RT-DETR framework, this study proposes a Multiscale Separable Deformable (MSD) module that integrates depthwise convolution with deformable convolution to enhance feature extraction capabilities while reducing computational load. Additionally, a Deformable Agent Attention (DAA) mechanism is designed to optimize attention weights through sparse queries, effectively improving detection accuracy for small targets and enhancing inference speed in complex scenarios. Experimental results demonstrate that the improved model achieves 87.9% mean average precision (mAP) on a railway dataset, with a detection speed of 90 frames per second (FPS). The proposed model achieves a +1.7% mAP improvement and 13.9% faster inference speed compared to RT-DETR, while simultaneously reducing model parameters by 24.6%. As a result, the proposed model is highly effective for multiple obstacle intrusion detection in complex real-world scenarios.

INDEX TERMS Convolutional neural network (CNN), deep learning, obstacle intrusion detection, railway traffic, transformer.

I. INTRODUCTION

High-speed railways and subways are widely constructed globally, which serve as the primary modes of transportation for both passengers and freight due to their cost-effectiveness, speed, and safety. However, rail accidents remain a significant concern, mostly due to obstacles intruding onto the tracks. According to the International Union of Railways (UIC), 90% of railway accidents are attributed to third-party intrusions into rail lines [1].

Accidents caused by obstacle intrusions often occur, involving people, animals, trees, kites, and other objects. These accidents have resulted in significant losses, including service disruptions, delays, damage to trains and infrastructure, increased maintenance costs, and potential safety risks to passengers and staff. These risks are particularly pronounced in open railway environments, such as level crossings and suburban sections [2], [3], as illustrated in Figure 1. When trains encounter hazardous obstacles during operation, they must execute emergency braking to avoid collisions. However, due to their high speeds, early identification of obstacles and effective protective measures are imperative. These incidents also lead to reputational damage and financial impacts due to compensation claims and operational inefficiencies. Considering the limited reaction time and visual field of train drivers, avoiding collision with obstacles based solely on manual observation is challenging. These intrusions pose serious risks to train safety and require urgent attention. Reliance on manual observation by train drivers is susceptible to environmental and human factors,

The associate editor coordinating the review of this manuscript and approving it for publication was Wei Xu.

resulting in delayed response times. Consequently, achieving high-precision detection of multiscale intrusive obstacles under real-time conditions to enable timely detection and warning remains a significant challenge. Recent developments in artificial intelligence and autonomous driving technologies have driven remarkable progress in machine vision for real-time railway obstacle detection. Deep learning and image processing techniques can rapidly identify and analyze potential hazards in rail transit systems, thereby enhancing operational safety. With the development of railway obstacle intrusion detection system, the system typically consists of multiple interconnected layers, each playing a critical role in ensuring the safe operation of trains. At the hardware layer, visual sensors and LiDAR devices capture real-time data about the railway environment. The algorithm layer, which forms the core focus of our research, processes this sensor data using advanced object detection and tracking models to identify potential hazards along the tracks. Above this sits the decision layer, responsible for risk assessment and generating appropriate responses, while the execution layer interfaces directly with train control systems to implement necessary safety measures. Our work specifically targets the algorithm layer, recognizing its pivotal role in enabling real-time, accurate detection of obstacle on railway tracks. By leveraging onboard camera systems, we aim to enhance the system's ability to identify small and obstructed objects that pose significant safety risks. This research ultimately seeks to improve railway safety by strengthening the algorithmic foundation for obstacle detection, thereby reducing the likelihood of accidents caused by obstacle intrusion.

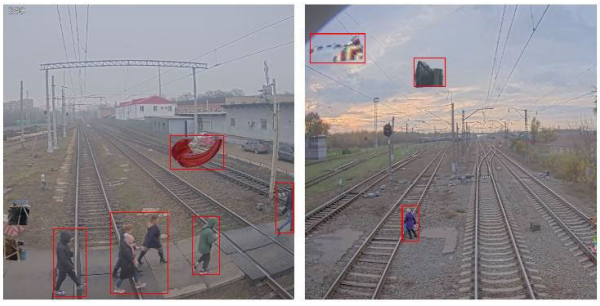

FIGURE 1. Potential obstacles in railway environments (Red boxes indicate annotated obstacle regions).

The traditional manual inspection methods, characterized by low efficiency, high cost, and subjectivity, have gradually been replaced. sensors, along with image processing techniques and AI-based technology, have been used to develop obstacle intrusion detection systems (OIDS) for railway applications [4], [5], [6]. High-resolution visual sensors, such as cameras, LiDAR, and millimeter-wave radar, are typically installed on locomotives, level crossings, and stations. These sensors collaborate with obstacle intrusion detection (OID) algorithms to detect obstacles that are currently or potentially entering the rail area, thereby issuing a danger warning to operating trains. The algorithms used in OID can be categorized into conventional image processing techniques and AI-based deep learning (DL) methods. Traditional image processing relies on local operations to extract shallow, low-level features from rail and obstacle images. In contrast, deep learning (DL)-based methods, such as convolutional neural networks (CNNs), can automatically extract features with high abstraction, accuracy, and robustness, which makes them widely utilized in OID. Given the constraints of computational resources and the high-performance demands of railway applications, current research has focused on designing efficient and accurate obstacle detection and segmentation frameworks that efficiently abstract and quantify image content for improved identification.

Existing methods for railway obstacle detection [7], [8] still face limitations in comprehensively identifying both ground and aerial obstacles along rail lines, particularly small-scale objects. Furthermore, airborne obstacle (e.g., balloons, plastic bags) often evade timely detection despite drivers' continuous vigilance. These floating objects can obstruct the driver's vision and potentially damage power transmission lines. Thus, real-time identification of such objects is crucial for accident prevention. Railway OID algorithms must prioritize real-time performance, as early obstacle detection enables prompt driver intervention to mitigate accidents. While ground-based railway inspection systems suffer from limited coverage, onboard detection systems can dynamically monitor the train's operational environment, effectively compensating for the shortcomings of fixed monitoring infrastructure. Therefore, alongside improving detection accuracy, model complexity must be optimized to facilitate deployment on embedded devices.

In summary, railway intrusion detection algorithms must simultaneously satisfy real-time requirements, achieve accurate multiscale object detection in complex environments, and maintain balanced model complexity. Current algorithms have computational redundancy and struggle to deliver optimal performance in real-world deployments under increasingly intricate railway scenarios.

To address these challenges, this paper proposes an improved real-time high-precision detection model based on RT-DETR [9], specifically designed for OID in complex railway environments. This model enhances multiscale object detection performance while optimizing real-time efficiency, thereby reducing the frequency of train emergency braking and reducing accident risks. Compared with existing models, its compact architecture is specially tailored for deployment on embedded devices in train scenarios. Moreover, the proposed modules exhibit cross-domain applicability, as these components can be adapted for use in unmanned aerial vehicle (UAV) inspection and autonomous driving scenarios, providing reusable building blocks. The contributions of this work mainly lie in two aspects:

1. The Multiscale Separable Deformable (MSD) Module is proposed by integrating depthwise convolution (DWC) with deformable convolution network v4 (DCNv4). This module significantly reduces computational costs while expanding the receptive field and improving multiscale object detection

capabilities, thus enhancing both real-time performance on embedded devices and detection accuracy in railway environments.

2. The Deformable Agent Attention (DAA) Mechanism is proposed by employing deformable sampling points to focus on critical regions, thus extracting more essential features. By utilizing a sparse set of agents as queries, it reduces redundancy in attention weight computation, thus significantly accelerating inference speed.

II. RELATED WORK

Machine learning-based railway obstacle detection methods primarily fall into two categories: background modeling-based approaches and classifier-based approaches. The former is suitable for detecting moving obstacles intrusion but exhibits limited adaptability in complex scenarios, while the latter can process single-frame images yet suffers from restricted accuracy in multiscale and small object detection. In recent years, the rapid advancement of deep learning in image processing has demonstrated the significant advantages of Convolutional Neural Network (CNN) and Transformer-based models in railway obstacle detection. These models automatically extract features and perform classification, substantially enhancing feature extraction efficiency and robustness, making them particularly suitable for obstacle intrusion detection in railway scenarios. This technological evolution has crystallized into two distinct architectural frameworks: the accuracy-oriented two-stage detectors and efficiency-focused single-stage implementations.

Two-stage neural network approaches [10] first generate region proposals from input images, followed by classification and regression on these candidate regions. He et al. [11] enhanced small object recognition by improving Swin Transformer [12] and PAFPN [13], proposing a modified Mask R-CNN. Separately, Li et al. [14] further improved object detection in complex traffic environments through cross-layer fusion of backbone network features. Despite these architectural innovations, the inherent multi-stage processing mechanism imposes substantial computational demands, fundamentally limiting their real-time applicability in railway monitoring systems. Conversely, Single-stage neural network approaches [15] directly predict object categories and locations from input images, offering high computational efficiency and streamlined architectures. This significantly improves real-time detection performance, making them more suitable for deployment on embedded devices in rail transit systems [16], [17], [18]. Nevertheless, their simultaneous handling of anchor generation, classification, and regression—coupled with reliance on predefined anchor box designs—results in lower detection accuracy, particularly for small objects. To address this, Tian [19] proposed a variable-zoom multiscale enhancement method, leveraging detection results as prior knowledge to identify more small targets. Meng et al. [20] improved local feature representation by integrating SAM and SENet, enhancing attention to small objects. While these models advance detection performance, limitations persist in global context modeling and long-range dependency capture—precisely the challenges that Transformer architectures are positioned to address.

The widespread adoption of Transformers in computer vision is largely due to their inherent ability to model global relationships through self-attention mechanisms. This capability is exemplified by DETR [21], which revolutionized object detection with its end-to-end set prediction approach, effectively capturing long-range dependencies. However, DETR's reliance on global feature processing results in significant computational overhead, posing challenges for real-time deployment. Consequently, there has been increasing research attention on improving DETR.

To accelerate model convergence, several DETR variants—including Deformable-DETR [22] and Conditional-DETR [23]—have introduced innovative attention mechanisms. In parallel, DINO [24] addresses both positive and negative noise object queries, which accelerates model convergence and enhances the accuracy of detecting small targets. MS-DETR [25] improves the training efficiency of DETR by explicitly supervising the candidate generation process through a mix of one-to-one and one-to-many supervision. Additionally, RT-DETR addresses this limitation through its decoupled multiscale interaction architecture, which is the first to separate multiscale feature interaction from cross-scale fusion operations to enhance computational efficiency. Wang et al. [26] introduced cascaded group attention to propose CGA-IFI for intra-scale feature interaction. They further designed a dilated reparam block (DRB-CFFM) to enhance cross-scale feature interactions, thereby improving high-resolution remote sensing object detection through advanced image feature fusion. Yu et al. [27] enhanced railway turnout defect detection performance by modifying the RT-DETR backbone: (1) replacing the multi-head self-attention mechanism with the Hilo attention mechanism, and (2) substituting the original cross-scale feature fusion module (CCFF) with an optimized fusion structure.

Despite these advancements, hybrid CNN-Transformer detectors continue to struggle with balancing model compactness, detection accuracy, and inference speed. To address these challenges in complex railway scenarios, this study introduces an improved RT-DETR framework that facilitates real-time, high-precision multiscale obstacle intrusion detection while maintaining architectural simplicity.

III. METHODOLOGY AND IMPROVEMENTS

A. IMPROVED NETWORK ARCHITECTURE

As illustrated in Figure 2, we implement algorithmic improvements based on the RT-DETR model to achieve efficient and accurate detection of railway obstacles. The input images from the dataset are first processed through the MSD_Block (Multiscale Separable Deformable Block), which enables enhanced multiscale feature extraction

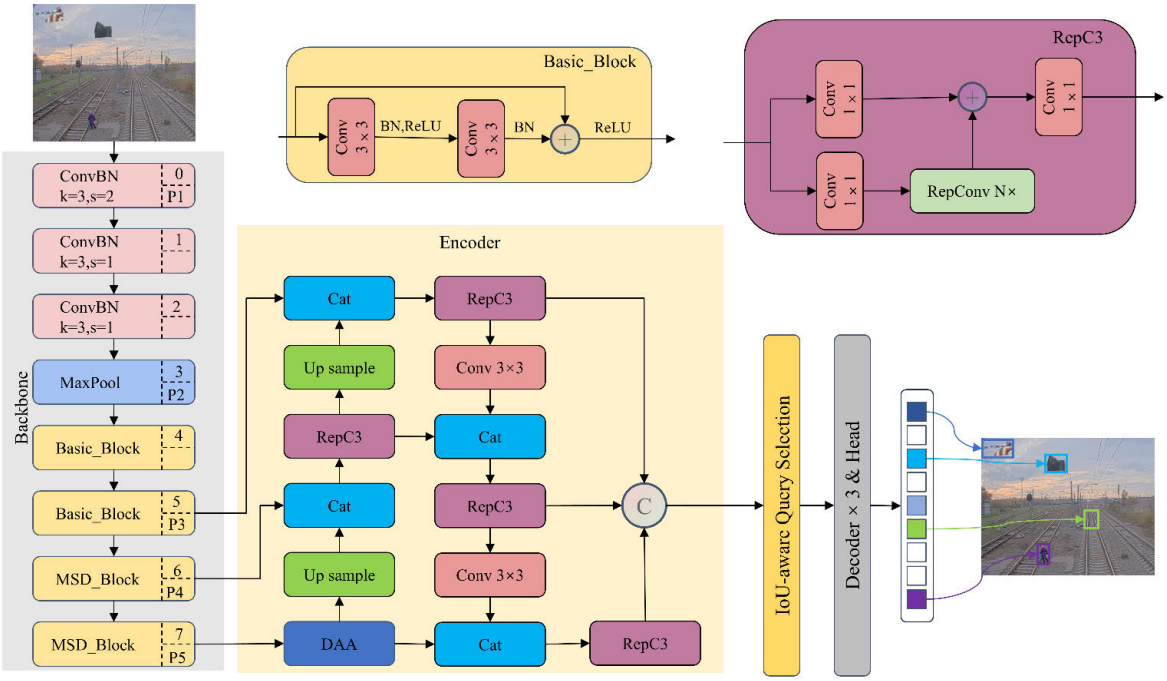

FIGURE 2. Improved model architecture.

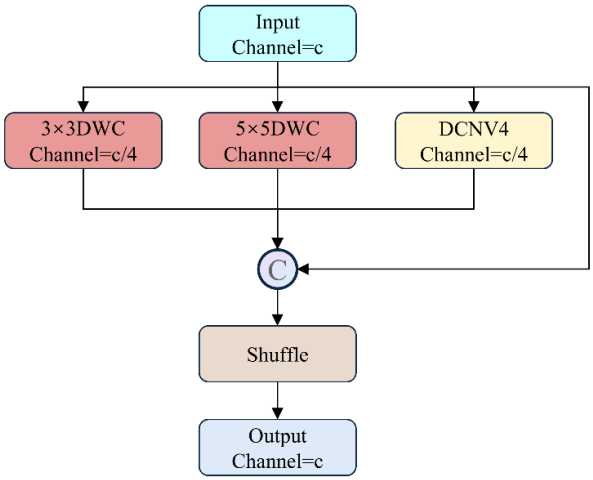

FIGURE 3. MSD module.

by synergizing depthwise convolution (DWC) with feature redundancy reduction operations, thereby significantly reducing computational overhead. Subsequently, the low-resolution feature maps are fed into the DAA module (Deformable Agent Attention) to acquire richer and more accurate semantic information. By leveraging a sparse set of agents as queries, this module eliminates redundancy in attention weight computation, accelerating the model's inference speed while maintaining detection precision.

B. MSDBLOCK

For computationally constrained embedded devices, model lightweighting is critical, while the complexity of railway environments imposes stricter demands on real-time performance. The original model employs stacked ResNet blocks [28], which not only fail to meet lightweight and real-time requirements due to their bulky architecture, but also suffer from limited receptive fields. The conventional convolutional stacking used in these blocks exhibits restricted capability in multiscale feature extraction. To address these issues, we propose the MSD module, which enables more comprehensive feature extraction while significantly reducing computational overhead.

As illustrated in Figure 3, the MSD module integrates depthwise convolution (DWC) and deformable convolution network v4 (DCNv4) [29]. Based on the effective receptive field theory [30], enlarging convolutional kernel sizes proves more effective in expanding the model's feature receptive field. Consequently, we employ kernels of varying dimensions to enhance multiscale object detection, while DWC reduces computational costs compared to standard convolutions without compromising feature extraction capability. To address feature redundancy in preliminary extraction stages [31], we preserve 25% of the input channels without being processed, thereby maintaining accuracy while minimizing computational overhead.

For complex images with deformable or non-rigid variations, fixed convolutional kernels struggle to adapt to content changes. In contrast, DCNv4 dynamically adjusts kernel shapes and sampling positions to adaptively focus on critical feature regions, facilitating the extraction of intricate local patterns. Following feature concatenation from four branches, a shuffle operation [32] is applied to enable cross-channel information fusion, thereby enhancing

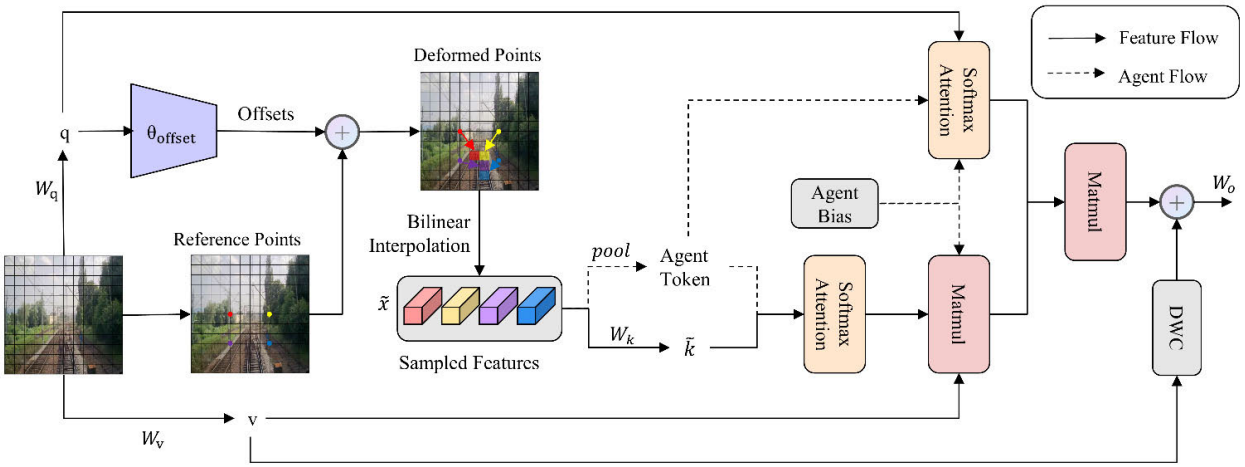

FIGURE 4. Deformable agent attention.

the model's representational capacity. The operation is formulated as:

$$Output = Shuffle(cat(DWC_3\,(x_0)\,, DWC_5\,(x_1)\,, DCN\,(x_2)\,, x_3) \quad (1)$$

where DWC_3, DWC_5 denote depthwise convolutions with kernel sizes of 3×3 and 5×5, respectively.

C. DEFORMABLE AGENT ATTENTION MECHANISM

The standard Transformer suffers from high computational complexity and excessive reliance on computational resources, making it challenging to deploy in complex railway environments. To address this limitation, we propose a Deformable Agent Attention (DAA) mechanism by synergizing the strengths of Agent Attention [33] and Deformable Attention [34]. As illustrated in Figure 4, the proposed DAA extends the traditional attention triplet (Q, K, V) by introducing an additional set of agent vectors A, forming a novel quadruple attention mechanism (Q, A, K, V). In this framework, the agent vectors A first act as proxies for the query vectors Q to aggregate information from K and V, and then broadcast the refined information back to Q. Since the number of agent vectors can be designed to be significantly smaller than that of query vectors, agent attention achieves global information modeling at reduced computational cost. The DAA mechanism seamlessly integrates the powerful global modeling capability of Softmax attention with the computational efficiency of linear attention, inheriting their advantages while maintaining low computational complexity and high model expressiveness. To further enhance the module's ability to focus on task-relevant regions and capture discriminative features, the DAA incorporates deformable points from Deformable Attention. Specifically, a deformable attention block with an offset network generates adaptive offsets based on query features, dynamically determining the positions of agents A and keys K. This design enables agents A and keys K to concentrate on critical regions with heightened flexibility and efficiency, thereby enhancing the original self-attention module and extracting more informative features.

Figure 4 Given an input feature map $x \in R^{H\times W\times C}$, we first downsample it with a factor r to obtain a uniform grid size $H_G = H/r$, $W_G = W/r$, and reference points $p \in R^{H_G\times W_G\times 2}$ are generated from this grid for subsequent deformable convolution operations. The offset generation involves linearly projecting the input features to produce queries $q = xW_q$, which are fed into a lightweight network $\theta_{offset}(\cdot)$ to generate offsets $\Delta p = \theta_{offset}(\mathrm{q})$. Deformed features $\tilde{x}$ are then sampled via spatial interpolation, followed by linear projections and pooling to derive keys and agent tokens, as are formalized in equation (2) and (3):

$$q = xW_q, \tilde{k} = \tilde{x}w_k, AgentToken = pool(\tilde{x}w_a) \quad (2)$$

$$\Delta p = \theta_{offset}(q), \tilde{x} = \varphi(x; p + \Delta p) \quad (3)$$

where $\tilde{k}$ and Agent Token represent the deformed key and agent embeddings. The sampling function $\varphi(\cdot;\cdot)$ employs bilinear interpolation, defined in equation (4):

$$\varphi\left(z;\left(p_x, p_y\right)\right) = \sum\nolimits_{r_y, r_y} g\left(p_x, r_x\right) g\left(p_y, r_y\right) z\left[r_y, r_x, :\right] \quad (4)$$

where g(a, b) denotes the interpolation weighting function along the x and y-axes, and (r_x, r_y) indexes all spatial positions in $z \in R^{H\times W\times C}$.

With q and Agent Token, the Deformable Agent Attention (DAA) mechanism is formulated as:

$$O^A = \sigma\left(qA^T\right)\underbrace{\sigma\left(A\tilde{k}^T\right)v}_{V_A} \quad (5)$$

where $A \in R^{n\times c}$ denotes the agent tokens from equation (2), $\sigma(\cdot)$ is the softmax function, and $v = xW_v$ is the linearly projected value. Specifically, agents A first aggregate global information via $\sigma(AK^T)V$ to produce V_A. Subsequently, A serves as keys to broadcast agent-refined features back to

TABLE 1. Data distribution in TAD dataset.

Class	Number	Quantity		
		Train set	Val set	Test set
Crossing	N1	1499	194	292
Turnout	N2	2593	356	506
Semaphore	N3	3561	453	642
Person	N4	1495	256	267
Nest	N5	4482	602	914
Plastic bag	N6	1502	209	318
Floater	N7	6169	843	1268
Balloon	N8	2521	348	518
Total		23822	3261	4725

all queries through $\sigma(qA^T)$, avoiding quadratic QK^T computation. By setting the number of agents, the mechanism achieves linear complexity while preserving global modeling capacity.

The final output of DAA integrates agent-guided features and local diversity preservation:

$$W_o = \sigma\left(qA^T + B_2\right)\sigma\left(A\tilde{k}^T + B_1\right)v + DWC(v) \quad (6)$$

where B_1, B_2 are learnable agent biases enhancing positional awareness, and the depthwise convolution (DWC) term maintains feature diversity.

D. REPLACING AIFI WITH DDA MODULE

In the AIFI module, each query q interacts with a large number of keys k, leading to excessive computational costs, slow convergence, and increased risk of overfitting. To address these limitations, we replace the AIFI module in RT-DETR with the proposed Deformable Agent Attention (DAA) module. The DAA mechanism focuses candidate agents and keys on critical regions, thereby avoiding the high computational complexity caused by applying uniform global attention across the entire image in standard multi-head self-attention (MHSA). The DAA module concentrates attention on pivotal regions and captures more informative features. By utilizing a small set of agents A as query proxies, it eliminates direct interactions between individual queries and keys, mitigating redundancy in attention weight computation while achieving high model expressiveness with low computational complexity. The operational process is formulated as:

$$F_5 = DAA(P_5) \quad (7)$$

where DAA denotes the Deformable Agent Attention mechanism, and P_5 represents the higher-level feature layer from the backbone network (see Figure 2), which contains richer semantic information.

IV. EXPERIMENTAL RESULTS

To evaluate the detection capability and improvement effectiveness of the proposed algorithm, we conducted comparative and ablation experiments on the railway dataset. All experiments were performed under identical software and hardware configurations for consistent comparison and analysis. The implementation utilized the PyTorch 2.0.1 deep learning framework on Ubuntu 22.04, with a workstation equipped with an Intel Core i9-12900K CPU and an NVIDIA GeForce RTX 3090 GPU (24GB VRAM). Input images were resized to 640 × 640 pixels before being fed into the model. During training, the batch size was set to 16 for 120 epochs, using a stochastic gradient descent (SGD) optimizer with momentum. The learning rate, momentum, and weight decay were configured as 0.0001, 0.9, and 0.0001, respectively.

A. DATA PREPARATION

Pretrained models on common public datasets (e.g., MS COCO, PASCAL VOC) often fail to demonstrate comparable performance in railway environments. To address this limitation, we constructed a railway-specific dataset comprising training, validation, and test sets, synthesized from three sources: Infra Dataset [35], Rail Dataset [36], and RailFOD23 [37].

To achieve real-time obstacle detection ahead of trains and ensure consistency with actual application scenarios, we manually removed non-driver perspective images from these datasets, retaining only those captured from advantageous positions on the train, simulating the viewing angle of onboard cameras. This selection enhances the model's generalization capability in railway scenarios ahead of the train. We utilized the LabelMe annotation tool to rectify the annotation information of the integrated dataset, correcting inconsistencies in the original datasets (e.g., unlabeled pedestrians, railroad switches, and level crossings). We also established annotation granularity requirements, such as using minimum bounding rectangles for large objects and covering the entire body for small objects, to prevent missed detections due to occlusion. We made every effort to ensure that there were no omissions or errors in the annotations of the integrated dataset. Cross-validation was conducted by two annotators to ensure data accuracy. The optimized dataset was named the Train Assisted Dataset (TAD). Figure 5 illustrates

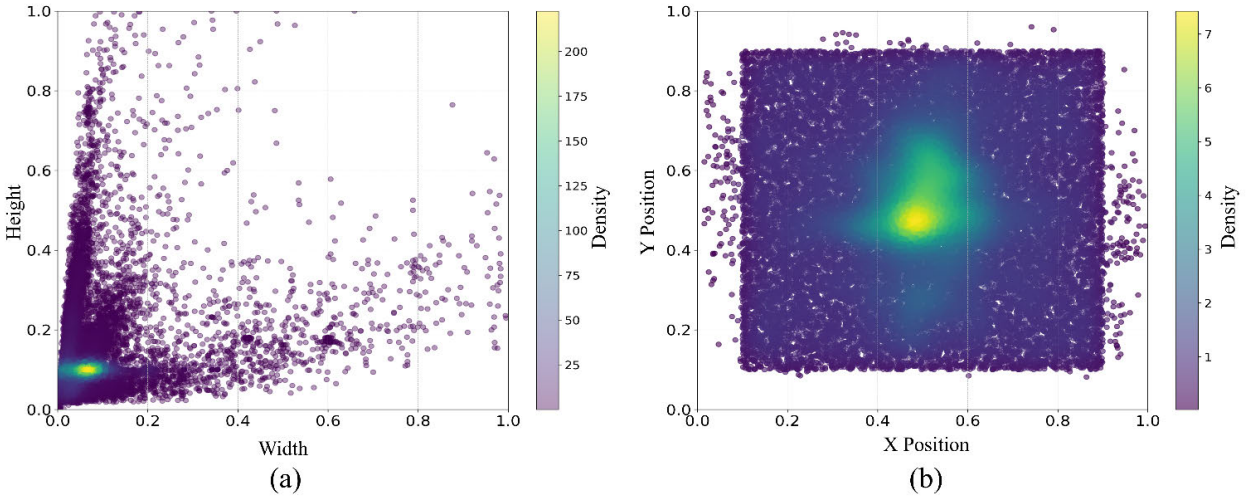

FIGURE 5. Distributions of object dimensions and locations in TAD dataset: (a) Height-width distribution; (b) Center-point spatial distribution. Heatmap colors indicate density levels (yellow: highest;green, blue, and purple: descending order).

the height-width distribution (Fig. 5(a)) and center-point location distribution (Fig. 5(b)) of the objects in the dataset, revealing that most obstacles in railway environments are small targets, posing significant challenges for high-precision real-time detection. The detection scenarios cover objects near the tracks under both daytime and nighttime conditions. These objects are categorized into eight classes: Crossing, Turnout, Semaphore, Person, Nest, Plastic bag, Floater, and Balloon. Detecting Crossings, Turnout, and Semaphore aims to alert drivers to exercise heightened caution in these areas, while Nest, Plastic bag, Floater, and Balloon threaten driver visibility and critical infrastructure such as pantographs and power transmission lines. Detailed category descriptions are provided in Table 1. The dataset contains 11,957 images, which are randomly split into training, validation, and test sets at ratios of 75%, 10%, and 15%, respectively. In terms of data augmentation, we maintained consistency with the data augmentation strategies of the original RT-DETR model, which helps to fairly validate the effectiveness of the improved model.

B. EVALUATION METRICS

To evaluate the performance of the proposed model under identical experimental conditions, this study compares detection models using the following metrics: mean Average Precision (mAP), F1-score, Giga Floating-point Operations per Second (GFLOPs), Frames Per Second (FPS), and model parameters (Params). The F1-score, defined as the harmonic mean of precision and recall, serves as a comprehensive metric to balance the trade-off between these two indicators. A higher F1-score indicates better equilibrium between precision and recall. The calculation for the F1 score is detailed in Equation (8).

$$F1 = 2\frac{Precision \cdot Recall}{Precision + Recall} \tag{8}$$

The Average Precision (AP) measures the average precision at various recall levels per category. The mAP reflects the model's average detection accuracy across all obstacle categories, with higher values denoting superior multi-class recognition performance, as illustrated in the following equations, Equations (9) and (10).

$$AP = \int_0^1 P(r)dr \tag{9}$$

$$mAP = \frac{\sum_{j=1}^{S} AP(j)}{S} \tag{10}$$

where S denotes the total category count and the denominator is the summation of AP values across all categories.

FPS quantifies real-time inference speed, where higher values correspond to faster processing capabilities, as shown in Equation (11).

$$FPS = \frac{1}{t_{avg}} \tag{11}$$

where t_{avg} represents the per-image inference time.

Moreover, GFLOPs and Params measure computational complexity and model size, respectively. Lower values in these metrics indicate reduced algorithmic complexity, enhancing deployability on resource-constrained embedded devices.

C. MODEL COMPARISON EXPERIMENTS

In comparative experiments with state-of-the-art detection algorithms, we benchmarked Improved RT-DETR against advanced models including YOLOv11 [38] and YOLOv12 [39]. The experimental results are summarized in Table 2. Improved RT-DETR achieves 87.9% mean average precision (mAP) at a detection speed of 90 FPS. Due to their bulky backbone networks, RT-DETR-34 and RT-DETR-50 exhibit slower inference speeds, higher computational complexity, excessive parameter counts, and larger model sizes,

TABLE 2. Results of model comparison experiments.

Model	mAP50(%)	FPS	FLOPs(G)	Params(M)	Size(MB)
RT-DETR-R18	86.2	79	57.2	19.89	38.6
RT-DETR-R34	86.9	69.5	87.5	30	57.9
RT-DETR-R50	87.1	44.7	125.7	41.95	82.1
YOLOv11n	81.7	440	6.3	2.58	5.2
YOLOv11m	87.1	134	67.7	20.04	38.7
YOLOv12m	87.3	104	67.1	20.11	38.9
Improved RT-DETR	87.9	90	50.2	14.99	29.3

TABLE 3. Results of ablation experiments.

Model	mAP50(%)	FPS	Parameter（M）	GFLOPs
RT-DETR	86.2	79	19.89	57.2
RT-DETR_MSD	87.3	82	14.78	49.8
RT-DETR_DAA	87.4	89.5	20.1	57.5
Improved RT-DETR	87.9	90	14.99	50.2

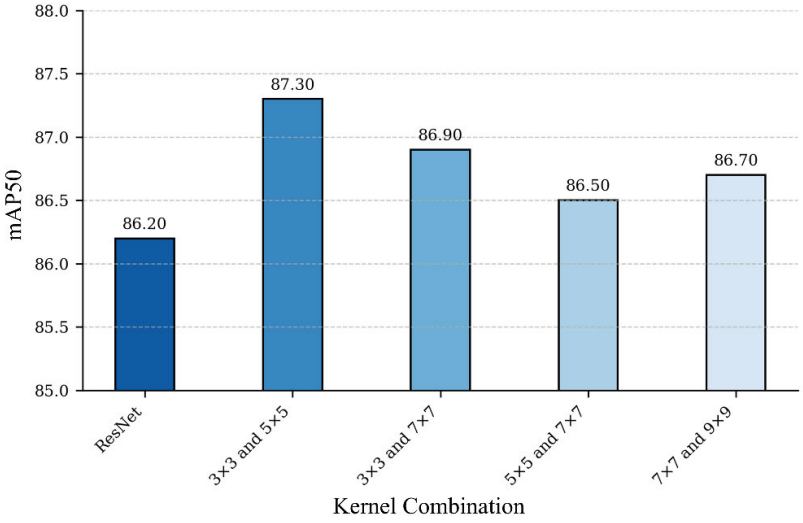

FIGURE 6. Results of multi-scale convolutional combinations.

rendering them unsuitable for real-time deployment on embedded onboard devices. While YOLOv11n is deployable on embedded platforms, its insufficient detection accuracy (81.7% mAP) makes it impractical for railway environments requiring high precision. Although YOLOv11m has a detection accuracy 0.8% lower and YOLOv12m has a detection accuracy 0.6% lower than that of our improved model, respectively. Moreover, they require 17.5 GFLOPs and 16.9 GFLOPs more in computations, 5.05M and 5.12M additional parameters, and 9.4 MB and 9.6 MB larger model sizes compared to our proposed method. These results conclusively validate that our optimized model is more amenable to deployment on embedded systems such as onboard devices.

D. ABLATION STUDY

To validate the effectiveness of the proposed modules, we conduct ablation experiments using RT-DETR as the baseline model. By modifying the backbone network, we achieve multiscale feature extraction while significantly reducing computational complexity and parameter count. Additionally, the original attention in the AIFI module is replaced with our proposed Deformable Agent Attention (DAA) to enhance inference speed and accuracy. Experimental results are summarized in Table 3.

The RT-DETR_MSD variant, which integrates the MSD module into the backbone, demonstrates a 1.1% improvement in mean average precision (mAP) while reducing parameters by 5.11M and computational load by 7.4 GFLOPs, confirming its suitability for edge device deployment. For the RT-DETR_DAA variant, replacing the standard transformer attention with DAA increases parameters by 0.21M and computation by 0.3 GFLOPs, yet achieves a 7.1 FPS acceleration and a 1.2% mAP gain. This improvement stems from the DAA mechanism's ability to focus dynamically on critical regions, accelerating both training convergence and inference

TABLE 4. Comparative results of different attention mechanisms.

Attention mechanisms	mAP50 (%)	mAP50:95 (%)	Params (M)	FLOPs (G)	FPS	F1
AIFI	86.2	64.2	19.89	57.2	79	85.46
Cascade Group Attention[40]	86.4	64.2	19.72	57.3	91.8	85.04
Deformable Attention[34]	86.9	64.7	21.17	57.4	88.8	85.35
Agent Attention[33]	87	65	19.96	57.4	89	85.89
TBSN [41]	86.4	64.3	21.76	58.9	90	85.82
Deformable Agent Attention(our)	87.4	65.5	20.1	57.5	89.5	86.45

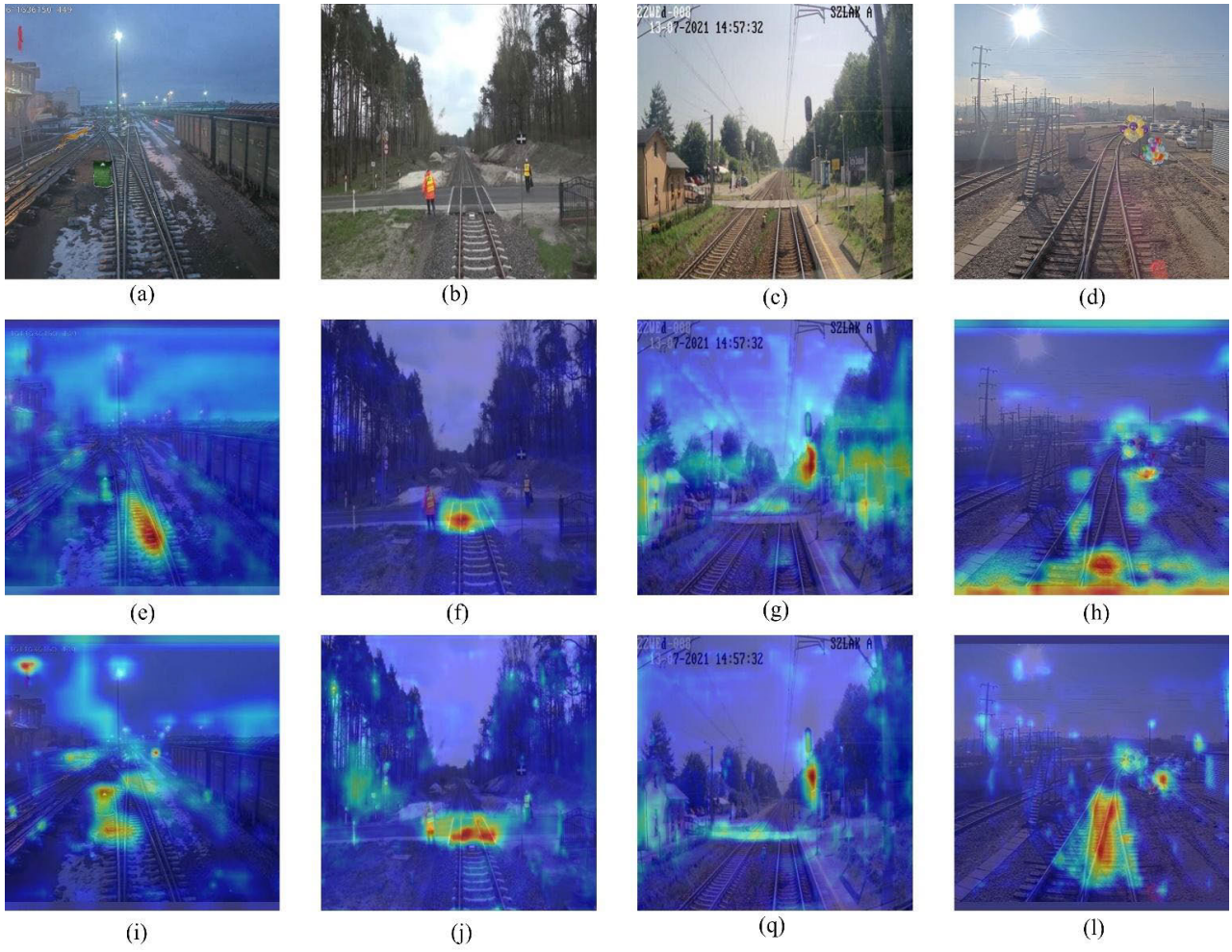

FIGURE 7. Grad-CAM visualization of feature activation maps before and after integrating the DAA module. (a–d) Original images. (e–h) Feature heatmaps of the baseline RT-DETR. (i–l) Feature heatmaps of the proposed model with DAA.

efficiency. The final optimized RT-DETR model outperforms the baseline with a 1.7% mAP50 increase, an 11 FPS speedup, 12.2% fewer GFLOPs, and a 4.9M parameter reduction. The ablation study results, as comprehensively analyzed in Table 3, confirm the effectiveness of each proposed module.

E. RESULTS OF MULTI-SCALE CONVOLUTIONAL COMBINATIONS

To validate the effectiveness of multi-scale feature extraction in enhancing detection accuracy for railway environments, we conducted comparative experiments on the proposed MSD module with different convolutional kernel configurations. As shown in Figure 6, models incorporating multi-scale convolutional kernels achieved superior detection accuracy compared to those using single-scale kernels. Specifically, the parallel architecture combining 3×3 and 5×5 convolutional kernels yielded the highest precision. Notably, all parallel configurations outperformed the original ResNet-based backbone network in detection accuracy. All convolutional operations employed lightweight depthwise

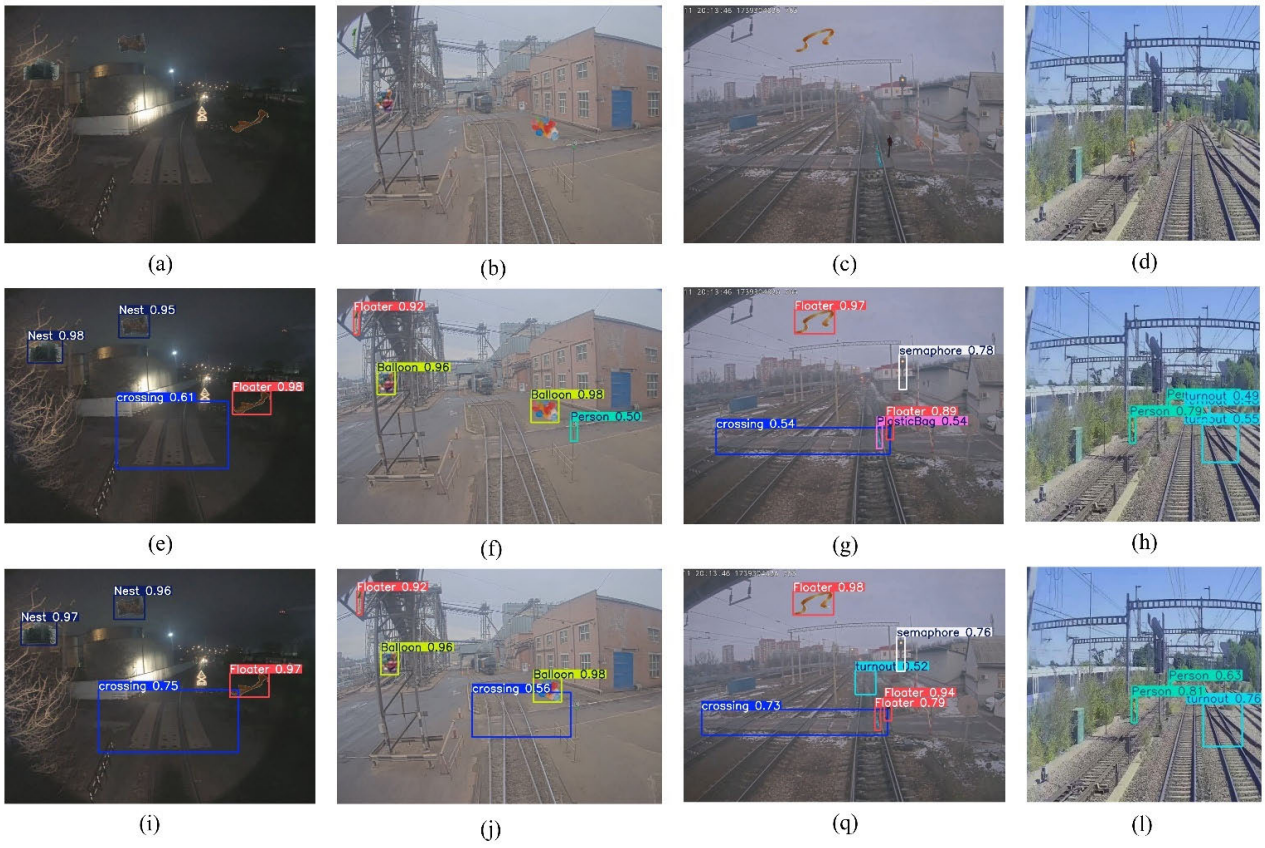

FIGURE 8. Partial detection results of two models. (a-d) Ground truth bounding boxes, (e-h) Predicted images from RT-DETR model, (i-l) Predicted images From improved RTDETR.

convolution (DWC), demonstrating that parallel multi-scale convolutions effectively capture multi-scale features while maintaining computational efficiency. These results also confirm the existence of feature redundancy in image feature extraction stages. As illustrated in Figure 3, selective feature extraction can be implemented without compromising final detection accuracy.

F. COMPARATIVE ANALYSIS OF ATTENTION MECHANISMS

To validate the superiority of our proposed Deformable Agent Attention (DAA) in obstacle intrusion detection under complex environments, we conducted comparative experiments against state-of-the-art attention mechanisms. All attention variants were evaluated by replacing the AIFI module in the RT-DETR model on the TAD dataset, with the original AIFI module serving as the baseline. To minimize testing variance, the reported frames per second (FPS) were averaged over five independent trials. The experimental results are presented in Table 4.

From Table 4, it can be observed that the proposed Deformable Agent Attention achieves significant improvements in mAP, F1-score, and FPS with only marginal increases in model size and parameter count compared to the baseline model. Specifically, mAP50 and mAP50:95 are enhanced by 1.2% and 1.3%, respectively, while the FPS improves by 10.5 frames per second. Although the Cascade Group Attention module accelerates detection speed, this comes at the cost of reduced accuracy. Both Deformable Attention and Agent Attention exhibit improvements in accuracy and inference speed, yet their performance remains suboptimal. The Transformer-based Blind-Spot Network (TBSN) module introduces substantial computational and parametric overhead for a mere 0.2% accuracy gain, rendering it impractical for onboard device deployment. Furthermore, our proposed attention mechanism achieves the highest F1-score (86.45) among these modules, indicating an optimal balance between precision and recall.

G. VISUALIZATION EXPERIMENTS AND ANALYSIS

To validate the attention capability of the proposed Deformable Agent Attention (DAA) mechanism on critical features, we employ Grad-CAM [42] to visualize the regions of interest in both the original RT-DETR and our modified model with DAA. Figure 7(a–d) displays the original input images, while (e–h) and (i–l) present the feature heatmaps of the baseline RT-DETR and our DAA-enhanced model, respectively. Deeper color intensity indicates higher model attention to the corresponding regions. As shown in the first and second columns, our model exhibits superior perception in complex railway scenarios, such as low-light conditions

and small targets, achieving more accurate detection of floating objects and pedestrians. This demonstrates enhanced global contextual modeling. Notably, in the third and fourth columns, our model reduces excessive focus on distracting elements by prioritizing essential regions, thereby strengthening target discriminability. These visual results confirm that our approach maintains robust global modeling and stable anti-interference capabilities even in intricate railway environments.

Figure 8 shows the detection results of the improved model in different complex railway environments. Fig. 8(a)-(l) demonstrate a comparative analysis of detection results between RT-DETR and the proposed improved model across diverse railway environments. Subfigures 8(a)-(d) display original input images, while 8(e)-(h) present the detection outcomes of RT-DETR, and 8(i)-(l) showcase the results from the enhanced model. These visual comparisons conclusively demonstrate that our improved model surpasses RT-DETR in detection accuracy, with significant reductions in missed detections and false positives. Specifically, Fig. 8(e) reveals that RT-DETR produces low-confidence detections under nighttime conditions, a limitation also observed in Figs. 8(g) and (h). Fig. 8(f) further illustrates RT-DETR's failure to detect a level crossing and its erroneous classification of a roadside sign as a pedestrian. In contrast, the proposed model achieves accurate detection across all scenarios, as evidenced in Figs. 8(i)-(l).

V. CONCLUSION

This study proposes a hybrid model combining convolutional neural networks (CNNs) and Transformers to enhance object detection performance in railway operations. By introducing two key modules, e.g., MSD and DAA, the model achieves superior real-time detection accuracy in complex railway environments. Specifically, the MSD module strengthens the model's capability to detect multiscale targets, enabling robust recognition of objects with varying sizes. The DAA module optimizes attention weight allocation to reduce computational redundancy, thereby accelerating inference speed. Experimental results on the railway dataset TAD demonstrate that the improved model attains 87.9% mean average precision (mAP) at 90 frames per second (FPS) on an NVIDIA GeForce RTX 3090 GPU, indicating high detection accuracy alongside real-time efficiency. To simplify the initial research phase, this study implemented real-time detection through full-image processing. However, this approach exhibits limitations in spatial granularity due to the absence of Region of Interest (ROI) partitioning. The lack of ROI segmentation leads to computational redundancy in non-critical regions (e.g., vegetation, distant infrastructure), thereby increasing false positive rates from background artifacts. To address this limitation, future work will incorporate semantic-guided dynamic ROI partitioning. This approach integrates a lightweight semantic segmentation module to prioritize high-resolution analysis for high-risk areas (e.g., near-track regions) while applying downsampling to peripheral regions, thereby minimizing computational overhead. Additionally, we will extend the model's applicability to diverse railway scenarios and enhance its capability to detect additional obstacle categories, thereby improving generalization performance. These enhancements will facilitate broader deployment across various railway environments while improving operational safety.

ACKNOWLEDGMENT

The authors gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

[1] Y. Qu, B. Wang, R. Qiu, and C. Zheng, "The research and realization of contactless collision early warning assistant system," in *Proc. Int. Conf. Electr. Inf. Technol. Rail Transp.*, Jan. 2016, pp. 147–154.

[2] P. Guha, S. Kool, R. Pipalwa, A. Acharjee, and A. Paul, "A prototype model of an unmanned automated railway level crossing traffic control system using ultrasonic and infrared sensors," *Int. J. Wireless Mobile Comput.*, vol. 26, no. 4, pp. 342–353, 2024.

[3] M. Zhang, A. J. Khattak, J. Liu, and D. Clarke, "A comparative study of rail-pedestrian trespassing crash injury severity between highway-rail grade crossings and non-crossings," *Accident Anal. Prevention*, vol. 117, pp. 427–438, Aug. 2018.

[4] M. Ghorbanalivakili, J. Kang, G. Sohn, D. Beach, and V. Marin, "TPE-net: Track point extraction and association network for rail path proposal generation," in *Proc. IEEE 19th Int. Conf. Autom. Sci. Eng. (CASE)*, Auckland, New Zealand, Aug. 2023, pp. 1–7.

[5] E. H. Assaf, C. von Einem, C. Cadena, R. Siegwart, and F. Tschopp, "High-precision low-cost gimballing platform for long-range railway obstacle detection," *Sensors*, vol. 22, no. 2, p. 474, Jan. 2022.

[6] D. He, Z. Zou, Y. Chen, B. Liu, and J. Miao, "Rail transit obstacle detection based on improved CNN," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–14, 2021.

[7] D. Ristić-Durrant, M. Franke, and K. Michels, "A review of vision-based on-board obstacle detection and distance estimation in railways," *Sensors*, vol. 21, no. 10, p. 3452, May 2021.

[8] S. Athira, "Image processing based real time obstacle detection and alert system for trains," in *Proc. 3rd Int. Conf. Electron., Commun. Aerosp. Technol. (ICECA)*, Coimbatore, India, Jun. 2019, pp. 740–745.

[9] Y. Zhao, W. Lv, S. Xu, J. Wei, G. Wang, Q. Dang, Y. Liu, and J. Chen, "DETRs beat YOLOs on real-time object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 16965–16974.

[10] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[11] D. He, Y. Qiu, J. Miao, Z. Zou, K. Li, C. Ren, and G. Shen, "Improved mask R-CNN for obstacle detection of rail transit," *Measurement*, vol. 190, Feb. 2022, Art. no. 110728.

[12] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.

[13] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8759–8768.

[14] C.-J. Li, Z. Qu, S.-Y. Wang, and L. Liu, "A method of cross-layer fusion multi-object detection and recognition based on improved faster R-CNN model in complex traffic environment," *Pattern Recognit. Lett.*, vol. 145, pp. 127–134, May 2021.

[15] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[16] Z. Jin, Z. Hu, H. Wang, and P. Li, "Personnel intrusion detection in railway perimeter with improved YOLOv7," in *Proc. Int. Symp. Artif. Intell. Robot.*, Jan. 2024, pp. 238–249.

[17] Y. Gao, Y. Qin, Z. Cao, L. Lian, J. Bai, X. Ge, and H. Yu, "A lightweight pedestrian intrusion detection algorithm based on on-board video," in *Proc. Int. Conf. Electr. Inf. Technol. Rail Transp.*, Jan. 2024, pp. 63–72.
[18] Z.-C. Feng, J. Yang, F. Li, Z.-C. Chen, Z. Kang, and L.-M. Jia, "An efficient foreign object recognition model in rail transit based on real-time railway region extraction and object detection," *J. Electr. Eng. Technol.*, vol. 19, no. 6, pp. 3723–3734, Feb. 2024.
[19] R. Tian, H. Shi, B. Guo, and L. Zhu, "Multi-scale object detection for high-speed railway clearance intrusion," *Appl. Intell.*, vol. 52, no. 4, pp. 3511–3526, Mar. 2022.
[20] C. Meng, Z. Wang, L. Shi, Y. Gao, Y. Tao, and L. Wei, "SDRC-YOLO: A novel foreign object intrusion detection algorithm in railway scenarios," *Electronics*, vol. 12, no. 5, p. 1256, Mar. 2023.
[21] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 213–229.
[22] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," 2020, *arXiv:2010.04159*.
[23] D. Meng, X. Chen, Z. Fan, G. Zeng, H. Li, Y. Yuan, L. Sun, and J. Wang, "Conditional DETR for fast training convergence," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 3631–3640.
[24] H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, L. M. Ni, and H.-Y. Shum, "DINO: DETR with improved DeNoising anchor boxes for end-to-end object detection," 2022, *arXiv:2203.03605*.
[25] C. Zhao, Y. Sun, W. Wang, Q. Chen, E. Ding, Y. Yang, and J. Wang, "MS-DETR: Efficient DETR training with mixed supervision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 17027–17036.
[26] A. Wang, Y. Xu, H. Wang, Z. Wu, and Z. Wei, "CDE-DETR: A real-time end-to-end high-resolution remote sensing object detection method based on RT-DETR," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2024, pp. 8090–8094.
[27] C. Yu and X. Chen, "Railway rutting defects detection based on improved RT-DETR," *J. Real-Time Image Process.*, vol. 21, no. 4, p. 146, Aug. 2024.
[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[29] Y. Xiong, Z. Li, Y. Chen, F. Wang, X. Zhu, J. Luo, W. Wang, T. Lu, H. Li, Y. Qiao, L. Lu, J. Zhou, and J. Dai, "Efficient deformable ConvNets: Rethinking dynamic and sparse operator for vision applications," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 5652–5661.
[30] X. Zhu, D. Cheng, Z. Zhang, S. Lin, and J. Dai, "An empirical study of spatial attention mechanisms in deep networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6687–6696.
[31] J. Chen, S.-H. Kao, H. He, W. Zhuo, S. Wen, C.-H. Lee, and S.-H.-G. Chan, "Run, don't walk: Chasing higher FLOPS for faster neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 12021–12031.
[32] H. Li, J. Li, H. Wei, Z. Liu, Z. Zhan, and Q. Ren, "Slim-neck by GSConv: A lightweight-design for real-time detector architectures," 2022, *arXiv:2206.02424*.
[33] D. Han, T. Ye, Y. Han, Z. Xia, S. Pan, P. Wan, S. Song, and G. Huang, "Agent attention: On the integration of softmax and linear attention," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2024, pp. 124–140.
[34] Z. Xia, X. Pan, S. Song, L. E. Li, and G. Huang, "Vision transformer with deformable attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 4784–4793.
[35] M. Ćwiek. *Infra Dataset*. Accessed: Dec. 15, 2024. [Online]. Available: https://universe.roboflow.com/mateusz-cwiek/infra
[36] Rail-HP8IJ_dataset. *Rail Dataset*. Accessed: Dec. 15, 2024. [Online]. Available: https://universe.roboflow.com/rail-psseq/rail-hp8ij
[37] Z. Chen, J. Yang, Z. Feng, and H. Zhu, "RailFOD23: A dataset for foreign object detection on railroad transmission lines," *Sci. Data*, vol. 11, no. 1, p. 72, Jan. 2024.
[38] R. Khanam and M. Hussain, "YOLOv11: An overview of the key architectural enhancements," 2024, *arXiv:2410.17725*.
[39] Y. Tian, Q. Ye, and D. Doermann, "YOLOv12: Attention-centric real-time object detectors," 2025, *arXiv:2502.12524*.
[40] X. Liu, H. Peng, N. Zheng, Y. Yang, H. Hu, and Y. Yuan, "EfficientViT: Memory efficient vision transformer with cascaded group attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 14420–14430.
[41] J. Li, Z. Zhang, and W. Zuo, "Rethinking transformer-based blind-spot network for self-supervised image denoising," 2024, *arXiv:2404.07846*.
[42] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

PENG LI received the Ph.D. degree in vehicle engineering from Central South University, in 2017. He is currently a Lecturer with the School of Railway Tracks and Transportation, Wuyi University. His main research interests include machine vision and damage identification.

YANHUI PENG received the bachelor's degree in communication engineering from Wuyi University, Jiangmen, China, in 2022, where he is currently pursuing the master's degree. His research interests include computer vision and deep learning in railway engineering.

SU-MEI WANG (Member, IEEE) received the B.S. degree in civil engineering from Hebei University, Baoding, China, in 2013, and the Ph.D. degree in bridge and tunnel engineering from Zhejiang University, Hangzhou, China, in 2018. She is currently a Research Assistant Professor with The Hong Kong Polytechnic University, Hong Kong. Her research interests include structural health monitoring, high-speed railway and maglev structural dynamics, interaction of train-rail-bridge systems, vision-based deep learning, and monitoring and control in rail engineering.

CHENG ZHONG received the bachelor's degree in engineering from Nanchang Hangkong University, in 2015. He is currently pursuing the master's degree in mechanical engineering with Wuyi University, Jiangmen, Guangdong. During the postgraduate study, his research direction is machine vision. His main research interests include object detection and semantic segmentation of rail transit scenes.