Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.079851

ARTICLE

Accurate Real-Time Measurement of Small and Irregular Road Abandoned Objects Using a Lightweight Vision-Based Framework

Ying Tang[1], Chuanyi Ma[2], Feng Guo[1,*] and Wenhao Sun[1]

[1]School of Qilu Transportation, Shandong University, Jinan, China
[2]Shandong Hi-Speed Group Co., Ltd., Jinan, China

*Corresponding Author: Feng Guo. Email: fengg@sdu.edu.cn

Received: 05 February 2026; Accepted: 29 April 2026; Published: 15 June 2026

ABSTRACT: Road Abandoned Objects (RAOs) pose significant threats to traffic safety, particularly due to their small size, irregular shapes, and unpredictable distribution in complex road environments. The primary objective of this study is to develop an accurate and real-time detection framework for RAOs while maintaining low computational cost for practical deployment. To achieve this, we propose RAO-YOLO, a lightweight vision-based detection framework built upon an enhanced YOLO architecture. Specifically, a Mixed Aggregation Network (MANet) is introduced to improve multi-scale feature representation, and a Lightweight Shared Detail-Enhanced Detection (LSDD) head is designed to enhance localization accuracy for small and irregular objects. Furthermore, a Focal-MPDIoU loss function is proposed to address sample imbalance and geometric irregularity during training. Extensive experiments conducted on the RAOD dataset demonstrate that the proposed method achieves superior performance compared to state-of-the-art detectors, achieving a mAP@0.5:0.95 of 56.1% while maintaining real-time inference speed. These results validate the effectiveness of the proposed framework for practical intelligent transportation applications.

KEYWORDS: Road abandoned objects; real-time object detection; road safety; intelligent transportation systems

1 Introduction

Traffic incidents pose serious safety and economic challenges worldwide, resulting in both substantial financial losses and travel disruptions [1,2]. Among the various causes, road abandoned objects (RAOs), referring to unexpected static obstacles such as debris, cargo, or fragments left on the roadway, represent a notable safety hazard.

In this study, real-time detection refers to achieving inference speeds that satisfy practical deployment requirements (typically above 30 FPS) while maintaining reliable detection accuracy. However, detecting small and irregular RAOs in real-time remains a challenging task due to their diverse shapes, small scales, and complex roadside environments. To address this issue, current detection practices primarily rely on static roadside sensors. While these systems can identify anomalies by monitoring traffic flow, they suffer from high false alarm rates and poor adaptability to complex road environments. Thus, there is an urgent need for accurate and efficient RAO detection to ensure smooth operations and maintain traffic safety. An illustration of RAOs highlighted with red boxes, is shown in Fig. 1.

Traditional sensor-based and rule-based approaches have been widely used for RAO detection. Sensor-based methods rely on hardware devices such as inductive loops, acoustic sensors, and passive infrared detectors [3], while rule-based approaches depend on handcrafted thresholds, background modeling, and

heuristic decision rules to identify anomalies. These systems identify anomalies by monitoring traffic flow patterns; however, on the one hand, their fixed-location nature restricts spatial coverage, and on the other hand, their adaptability to complex and dynamic traffic conditions is limited [4]. As a result, detection accuracy can be significantly degraded under varying lighting, weather, and occlusion conditions. Recent advances in computer vision provide solutions for overcoming these limitations. With traffic cameras now widely deployed, modern object detection algorithms indicate strong capabilities in recognizing multi-scale targets under real-time constraints, offering a promising foundation for RAO detection on highways [5]. These limitations highlight the need for more adaptive vision-based detection approaches capable of handling complex roadside scenarios.

Figure 1: Illustration of road abandoned objects (RAOs). Red bounding boxes highlight the RAO targets in the scene.

Early monocular vision-based RAO detection methods relied on edge extraction and background modeling. For instance, the Gaussian Mixture Model (GMM) separates foreground objects by modeling pixel intensities [6], while the Active Contour Model (ACM) detects boundaries via energy minimization [7]. Wavelet filtering and Bayesian decision models have also shown promise in controlled settings [8]. However, these approaches are highly sensitive to lighting, weather, and occlusions, limiting their real-world robustness. To improve geometric awareness, stereo vision techniques were introduced. Using calibrated camera pairs, methods like V-disparity [9] and adaptive baseline stereo [10] estimate object height and distance from disparity maps. Some systems reconstruct 3D point clouds for obstacle clustering and classification [11,12]. Although more robust to illumination changes, stereo setups require precise calibration, lack scalability, and often need manual tuning—hindering large-scale roadside deployment.

Recent advances in AI, particularly convolutional neural networks (CNNs), offer a more adaptable solution. CNNs have achieved remarkable success in object detection across diverse scenarios. Models like Faster R-CNN [13], SSD [14], and YOLO [15,16] enable accurate, real-time object detection and are well-suited for video surveillance. Yet, generic detectors struggle with small, irregular, or context-dependent RAOs. Their practical use remains limited by the scarcity of large-scale RAO-specific datasets and challenges in seamless real-time integration. In particular, standard detection architectures often suffer from insufficient feature representation for small targets and limited localization accuracy for irregularly shaped objects.

Despite recent progress in object detection, effectively detecting small and irregular RAOs under strict real-time constraints remains challenging, we propose an enhanced detection framework, termed RAO-YOLO, built upon YOLOv11m and incorporating targeted improvements. Specifically, our main contributions are as follows:

- We adopt a Mixed Aggregation Network (MANet) to replace the original YOLO feature extraction module. By aggregating multi-scale and multi-branch features, this module significantly enhances the network's representation ability for complex roadside objects.
- We propose a lightweight and efficient detection head, termed LSDD (Lightweight Shared Detail-Enhanced Detection), which improves both classification and localization accuracy through shared detail-enhancing layers and parallel prediction branches.
- We introduce the loss function Focal-MPDIoU, which integrates a Focal IoU mapping and corner distance penalty to better fit the characteristics of small and irregularly shaped RAOs, leading to improved training stability and performance.

2 Related Work

This section reviews existing research relevant to RAO detection. We first present representative datasets and benchmarks, as they provide the foundation for method development and performance evaluation in this domain. We then summarize traditional vision-based methods and AI-based approaches, highlighting their strengths and limitations in complex traffic environments.

2.1 Datasets and Benchmarks

Although general-purpose datasets (e.g., Cityscapes [17], KITTI [18]) support road scene understanding, they lack annotations for abandoned objects, limiting their applicability to RAO detection.

To fill this gap, specialized datasets have emerged. LostAndFound [19] provides 2102 annotated frames of small obstacles from urban roads but suffers from limited scale and diversity. CAOS [20], based on BDD100K [21], targets two anomaly categories, while RoadObstacle21 [22] offers 321 high-resolution, pixel-annotated images of road-surface obstacles, yet both lack scene variability for robust highway generalization. The recently released RAOD dataset [23] addresses these shortcomings with a large-scale, diverse collection of real-world traffic videos featuring various RAOs under diverse conditions, establishing a realistic and challenging benchmark for practical RAO detection in both urban and highway settings.

2.2 Traditional Abandoned Object Detection Methods

Early approaches to abandoned object detection relied on handcrafted features and heuristic rules. Motion-based methods (e.g., frame differencing, optical flow [8,24]) identify static anomalies against moving traffic but fail under camera motion or when RAOs appear in sparse traffic. Background subtraction techniques including GMM, median filtering, and codebook models are efficient but highly vulnerable to shadows, reflections, rain, and illumination changes, leading to excessive false alarms.

Background subtraction methods, including Gaussian Mixture Models (GMM), temporal median filtering, and codebook models, formed the backbone of many early systems [25]. By learning a scene's background over time, they flag foreground anomalies. Although computationally efficient and suitable for real-time use, their accuracy is easily compromised by shadows, reflections, rain, or low-light conditions, leading to false positives or missed detections. Post-processing often employs appearance-based features like color histograms, edge maps, HOG [26], and LBP [27] to refine results. Yet these features lack robustness to intra-class variation and cannot generalize to unseen object categories, limiting their effectiveness in open-set, real-world traffic scenarios.

Moreover, most traditional methods depend on fixed thresholds and static rules, making them inflexible to varying lighting, camera angles, road layouts, or seasonal changes. While efficient and easy to deploy, their reliance on handcrafted heuristics and sensitivity to environmental dynamics severely constrain generalization across diverse highway and urban settings.

2.3 AI-Based Detection Methods

Deep learning has significantly advanced RAO detection by enabling automatic learning of discriminative features. However, RAO detection differs from general object detection due to the small scale, irregular shapes, and ambiguous appearance of targets, which pose additional challenges for existing models. CNN-based detectors like Faster R-CNN [13], SSD [14], and YOLO [15] are widely used for their ability to learn hierarchical representations from annotated data, showing strong generalization—especially when trained on large-scale datasets such as SYNTHIA [28]. However, these models are primarily designed for regular and medium-to-large objects, and often struggle to preserve discriminative features for small and irregular RAOs, leading to missed detections or inaccurate localization. Bakirci [29] comprehensively evaluate YOLOv8 variants in aerial traffic monitoring, offering insights into their robustness under varying illumination, density, and occlusion. Some works further integrate semantic segmentation to better distinguish RAOs from background clutter.

Transformer-based models like DETR and its variants [30,31] capture global context, improving performance in occluded or low-contrast scenes. Streamlined architectures such as D-FINE [32] and RT-DETR [33] enable high-precision, real-time detection, their high computational cost and latency make them less suitable for real-time RAO detection in roadside deployment scenarios.

To mitigate sensor limitations under adverse conditions, recent studies explore multimodal fusion (RGB, LiDAR, radar) via feature- or decision-level strategies [34], enhancing robustness for small or low-contrast RAOs. However, these approaches face challenges in cost, complexity, latency, and computational overhead, hindering real-time highway deployment. Lightweight designs (e.g., MelNet [35], TinyDet [36]) and compression techniques (e.g., pruning, quantization) aim to address these issues. Despite progress, achieving both high accuracy and efficiency in dynamic, real-world traffic environments remains an open challenge.

Despite the progress of existing methods, they remain insufficient for RAO detection due to several key limitations, including inadequate feature representation for small-scale objects that often leads to missed detections, poor modeling of irregular object geometry resulting in inaccurate localization, and the inherent difficulty in balancing detection accuracy with computational efficiency under real-time constraints. To address these challenges, we propose RAO-YOLO, which enhances multi-scale feature representation, improves localization for irregular objects, and maintains real-time performance.

3 Methodology

To address the aforementioned challenges in RAO detection, including insufficient feature representation for small-scale objects, inaccurate localization of irregular targets, and the difficulty of maintaining real-time performance, we propose an improved detection framework termed RAO-YOLO. Built upon the baseline YOLO architecture, the proposed method introduces several targeted enhancements, where each component is explicitly designed to tackle a specific limitation of existing approaches. (1) Mixed Aggregation Network (MANet) module that enhances multi-scale feature aggregation and strengthens the representation of small objects; (2) Lightweight Shared Detail-Enhanced Detection (LSDD) head that improves localization accuracy for irregular targets while maintaining low computational overhead; (3) Focal-MPDIoU loss function designed to address sample imbalance and irregular object shapes during training.

3.1 Overview of Model Architecture

RAO-YOLO builds on YOLOv11 to tackle RAO detection in complex highway scenes. While preserving the one-stage paradigm, it modifies both backbone and detection head (Fig. 2). The backbone retains original

C3K2 modules in shallow layers for efficient low-level feature extraction, while deeper stages replace C3K2 with MANet which is a module that fuses multi-scale features and attention mechanisms [37,38] to better capture discriminative cues from small, irregular, or occluded RAOs. This design is particularly beneficial for RAO detection because abandoned objects in road scenes often occupy only a small number of pixels, exhibit irregular boundaries, and appear in cluttered backgrounds. Multi-scale aggregation helps preserve weak responses from small targets across different feature levels, while the attention mechanism suppresses redundant background interference and highlights informative regions. As a result, MANet improves both local detail sensitivity and global context awareness, which are essential for reliable detection in complex highway environments.

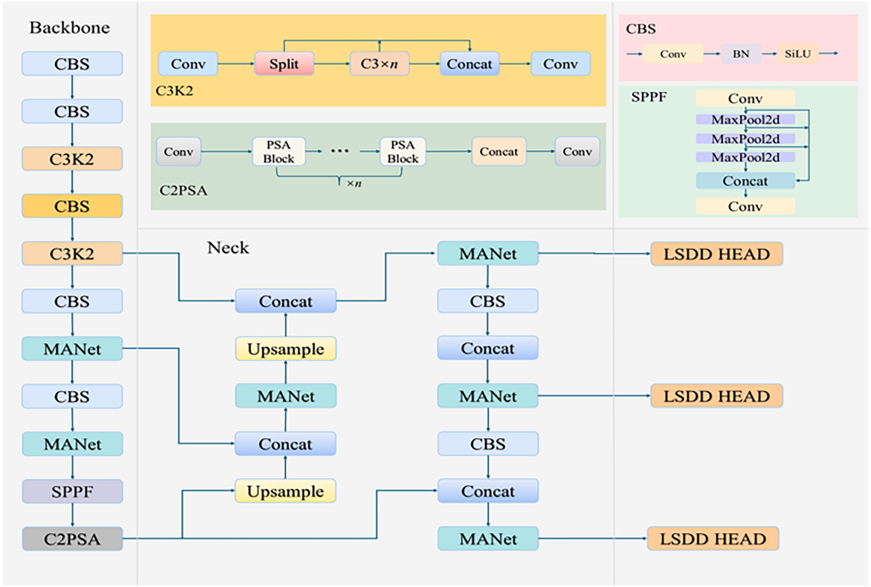


Figure 2: RAO-YOLO model architecture. The proposed framework integrates MANet in the backbone and LSDD in the detection head to enhance feature representation and localization accuracy for small and irregular RAOs.

The detection head is replaced with the lightweight LSDD head, which uses shared convolutions and detail enhancement to reduce complexity while improving fine-grained localization and classification especially effective for RAOs of varying scales and orientations. Combined with a tailored Focal-MPDIoU loss, RAO-YOLO achieves high accuracy and fast inference, making it suitable for real-time highway monitoring in intelligent transportation systems.

3.2 MANet Module

In our proposed framework, we adapt the MANet module (Fig. 3), originally introduced in Hyper-YOLO [39], to address RAO detection challenges. While designed for general object detection, we integrate MANet into the deeper stages of the YOLOv11 backbone and tailor it to better capture small, irregular, and partially occluded roadside objects. The module retains its multi-branch structure (standard convolution, depthwise separable convolution, and bottleneck enhancement), but is re-parameterized to balance accuracy and efficiency for high-resolution highway imagery. Its lightweight attention mechanism further enhances fine-grained feature extraction in cluttered scenes. Compared with the conventional C3K2 module, the adapted MANet provides richer multi-scale representations with lower computational cost, improving robustness to scale variation and complex environments.

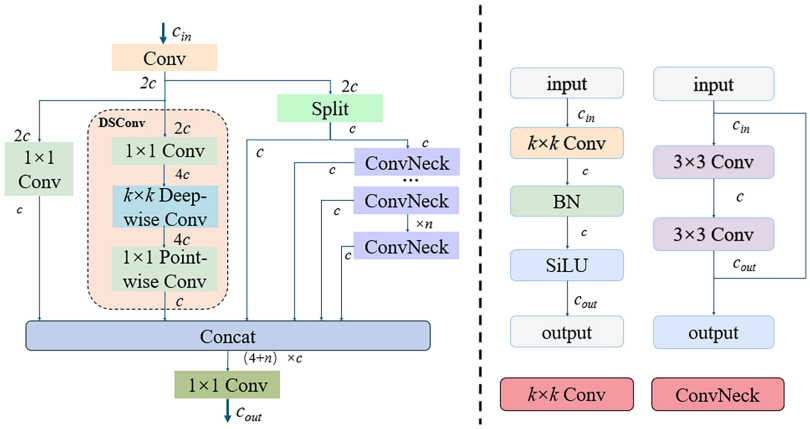


Figure 3: MANet model structure.

These architectural enhancements enable MANet to effectively capture both fine-grained local textures and high-level semantic cues with minimal redundancy. To explain the internal operation of MANet more intuitively, the module can be divided into: initial channel expansion, branch-wise feature transformation, and final feature fusion. First, the input feature is projected to a higher-dimensional intermediate representation to increase feature diversity. Second, the expanded feature is processed by different branches, where each branch focuses on complementary information. The core principle of the MANet module can be expressed as:

$$F_{in} \xrightarrow{Conv_{first}} F_{mid} \in \mathrm{R}^{B\times 2C\times H\times W} \tag{1}$$

After the input feature F_{in} is processed by the first convolutional layer, the number of output F_{mid} channels becomes twice the original (2C).

$$F_{mid} \xrightarrow{Branch_1} F_1 \in \mathrm{R}^{B\times C\times H\times W} \tag{2}$$

$$F_{mid} \xrightarrow{Branch_2} F_2 \in \mathrm{R}^{B\times C\times H\times W} \tag{3}$$

$$F_{mid} \xrightarrow{chunk} F_3, F_4 \in \mathrm{R}^{B\times C\times H\times W} \tag{4}$$

This decomposition allows different branches to focus on complementary information (e.g., compact features, enhanced textures, and preserved representations). Then the transformed feature F is divided into four different branches:

F_1 further compresses the channel through 1 × 1 convolution, F_2 is extracted through a complex convolutional sequence that includes depthwise convolutions, as shown in Eq. (3), F_3 and F_4 features are split directly into two parts along the channel dimension while preserving the original information. The feature enhancement module uses n Bottleneck modules to enhance the last feature. Finally, output result concatenates all the features and fuses them through 1 × 1 convolution, enabling effective integration of multi-scale and multi-type representations.

$$F_{enhanced} = \sum_{i=1}^{n} Bottleneck(F_4) \tag{5}$$

$$F_{out} = Conv_{final}\left([F_1, F_2, F_3, F_4, F_{enhanced}]\right) \tag{6}$$

$$F_{dw} = DWConv\left(F_{mid}\right) = Conv_{depthwise}\left(F_{mid}\right) e\, Conv_{pointwise}\left(F_{mid}\right) \tag{7}$$

Assuming the input feature map has a dimension of $H \times W \times C$, the MANet module performs feature aggregation while maintaining the spatial resolution $H \times W$ and adaptively enhancing informative channel responses. Compared with directly stacking standard convolutional blocks, this design provides a more effective balance between feature richness and computational cost. We adopt MANet because RAO detection requires sensitivity to subtle local cues while remaining efficient enough for real-time deployment. By incorporating this advanced module from Hyper-YOLO, RAO-YOLO achieves a favorable trade-off between detection accuracy and computational efficiency, significantly enhancing the model's practicality for real-world applications. Furthermore, the modular design of MANet allows for easy integration into various backbone architectures, making it a versatile component for modern object detection frameworks.

3.3 Optimized Detection Head

LSDD is a novel detection head module introduced in RAO-YOLO, specifically tailored for lightweight object detection tasks. The main idea of LSDD is to first align features at different scales and then enhance structural details using shared lightweight operators. This allows the detection head to focus more effectively on edge, texture, and directional information that is important for irregular RAOs. The structure of the proposed LSDD head is illustrated in Fig. 4.

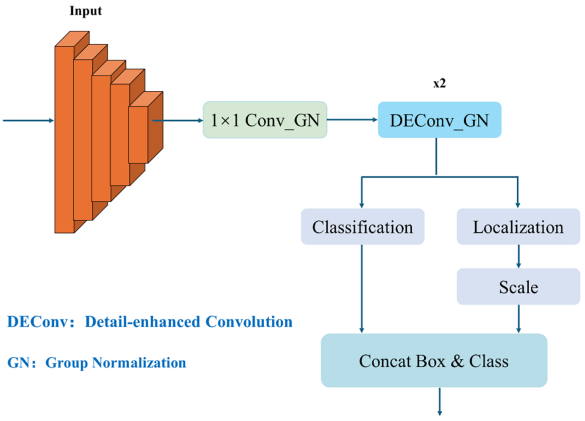


Figure 4: LSDD head structure.

Architecturally, LSDD adopts a multi-scale feature processing strategy, where dedicated 1×1 convolutional layers with group normalization [40] are applied to feature maps at different scales (i.e., P3, P4, and P5). These are followed by a shared detail-enhanced convolutional block, implemented as two consecutive DEConv_GN modules, which are applied across all scales. Group Normalization can better adapt to small-batch training and distributed training, enhancing the stability and robustness of the model.

This parameter-sharing mechanism significantly reduces the overall parameter count and computational burden, while maintaining strong feature extraction capabilities. The central component, DEConv, incorporates five distinct types of differential convolutions: center difference, horizontal difference, vertical difference, diagonal difference, and standard convolution. By capturing edge, texture, and directional features more effectively, these differential convolutions substantially enhance the model's ability to localize and classify objects, especially in challenging scenarios with small or densely packed targets. The calculation formula is as follows:

$$F_i = GN\left(Conv_{1\times1}\left(X_i\right)\right) \tag{8}$$

$$F_i' = DEConv\left(DEConv\left(F_i\right)\right) \tag{9}$$

$$DEConv\left(F\right)=\sum_{k=1}^{5}w_k\cdot D_k\left(F\right) \tag{10}$$

Among them, GN stands for Group Normalization. $D_k\left(F\right)$ represents center, horizontal, vertical, diagonal difference and standard convolution respectively, and w_k is the learnable weight.

From a technical perspective, LSDD leverages parameter sharing and detail enhancement to greatly improve computational efficiency and detection precision. The use of GN ensures robust training stability, even with small batch sizes or distributed training environments. Furthermore, the design of the regression and classification branches where the bounding box regression branch employs a learnable scale module and distribution focal loss (DFL), and the classification branch utilizes a dedicated 1×1 convolution enables the model to better handle multi-scale objects and complex backgrounds. This makes LSDD particularly suitable for real-time RAOs detection, where the detection of small, irregular, or partially occluded objects in complex road environments is critical for traffic safety and intelligent transportation systems. The lightweight and efficient design of LSDD allows for deployment of edge devices and embedded systems, meeting the stringent requirements of real-time RAOs detection in practical applications. The calculation formulas for the regression branch, classification branch, decoding and loss are as follows:

$$B_i=Scale_i\left(Conv_{1\times1}^{bbox}\left(F_i'\right)\right) \tag{11}$$

$$C_i=Conv_{1\times1}^{cls}\left(F_i'\right) \tag{12}$$

$$\hat{b}=\sum_{j=0}^{M-1}p_j\cdot j \tag{13}$$

$$Y=\left[Decode\left(B\right),\sigma\left(C\right)\right] \tag{14}$$

$$L=\lambda_{loc}\cdot L_{loc}+\lambda_{cls}\cdot L_{cls} \tag{15}$$

Among them, B_i and C_i respectively represent the outputs of bounding box regression and category prediction branches, p_j is the probability after softmax, and $M-1$ is the number of buckets divided. After decoding, the final output is Y. σ indicates sigmoid activation, L_{loc} stands for location loss, L_{cls} stands for classification loss, λ_{loc} and λ_{cls} are the corresponding loss weights.

3.4 Improved Loss Function

YOLOv11 originally uses Complete IoU (CIoU) [41] as its regression loss, which improves localization by incorporating center distance and aspect ratio constraints. While effective for regular-shaped RAOs (e.g., cones, barriers), CIoU's rigid aspect ratio term struggles with irregular objects like fragments, spills, or scattered debris leading to localization errors for elongated or flattened targets.

Several IoU variants aim to address this: Shape-IoU enhances shape adaptability but suffers from high computational cost, limiting real-time use; Inner-IoU improves overlap quality for small/occluded objects but underperforms on large or highly irregular RAOs.

In object detection tasks, the training objective typically consists of multiple components, including classification and localization losses. Therefore, the overall optimization objective of the proposed RAO-YOLO detector can be formulated as a composite loss:

$$L=\lambda_{cls}L_{cls}+\lambda_{reg}L_{reg} \tag{16}$$

where L_{cls} represents the classification loss used to predict object categories, and L_{reg} denotes the bounding box regression loss for object localization. The coefficients λ_{cls} and λ_{reg} are weighting parameters used to balance the contributions of classification and localization during training.

To overcome these issues, we propose Focal-MPDIoU, a hybrid loss combining Focal-IoU [42] and Minimum Point Distance IoU (MPDIoU) [43]. MPDIoU eliminates aspect ratio dependence by measuring the minimum point distance between boxes, enabling accurate localization of irregular shapes. Focal-IoU further introduces a piecewise linear weighting scheme (with thresholds α and β) to suppress easy samples and emphasize hard cases such as small, occluded, or scale-varying RAOs. This design improves convergence stability, maintains high localization accuracy across diverse geometries, and supports robust real-time detection. The formulation of Focal-IoU is as follows:

$$IoU_{focal} = \begin{cases} 0, & IoU < \alpha \\ \dfrac{IoU - \alpha}{\beta - \alpha}, & \alpha \le IoU \le \beta \\ 1, & IoU > \beta \end{cases} \tag{17}$$

$$L_{Focal-IoU} = 1 - IoU_{focal} \tag{18}$$

Among them, α and β are hyperparameters defined over the interval (0, 1), and are typically set to 0 and 0.95, respectively. Here, α controls the balance between positive and negative samples, while β adjusts the weighting of hard samples, enabling the model to focus more on difficult instances during training. Meanwhile, MPDIoU improves localization accuracy by introducing a geometric penalty based on the Euclidean distance between the top-left and bottom-right corners of the predicted and ground-truth bounding boxes. Unlike traditional overlap-based metrics, MPDIoU captures subtle misalignments and is more robust to variations in aspect ratio which is an essential property when detecting elongated spills, fragmented RAOs, or non-rectangular roadside objects. The MPDIoU calculation formula is as follows:

$$MPDIoU = IoU - \frac{d_1^2}{w^2 + h^2} - \frac{d_2^2}{w^2 + h^2} \tag{19}$$

$$d_1 = \sqrt{\left(x_1^{pred} - x_1^{gt}\right)^2 + \left(y_1^{pred} - y_1^{gt}\right)^2} \tag{20}$$

$$d_2 = \sqrt{\left(x_2^{pred} - x_2^{gt}\right)^2 + \left(y_2^{pred} - y_2^{gt}\right)^2} \tag{21}$$

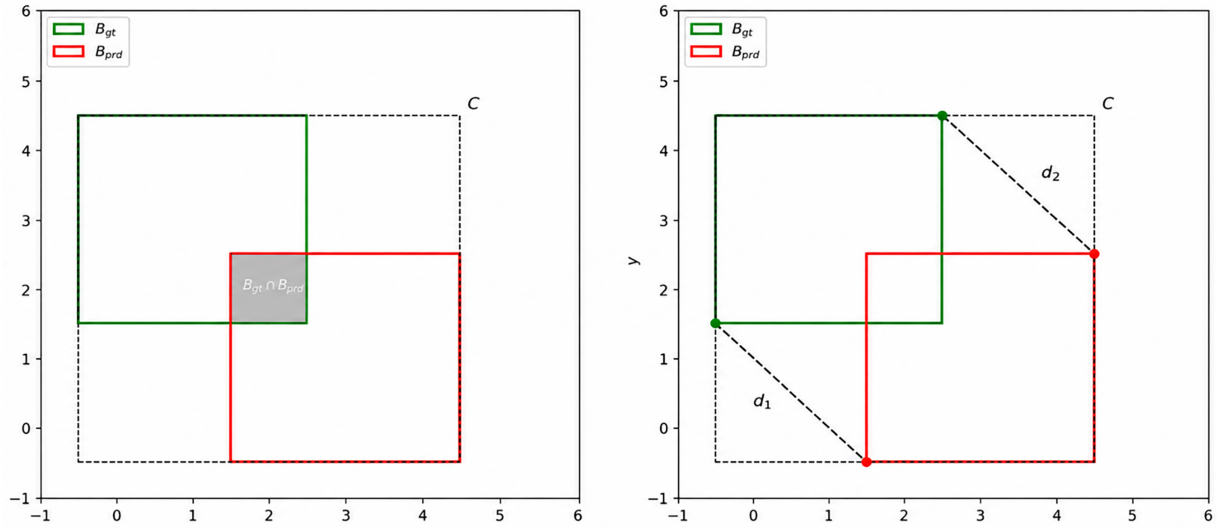
In this formula, w and h are the image width and height. (x_1^{pred}, y_1^{pred}) and (x_2^{pred}, y_2^{pred}) are the predicted box's top-left and bottom-right coordinates, while (x_1^{gt}, y_1^{gt}) and (x_2^{gt}, y_2^{gt}) are those of the ground-truth box. The Euclidean distances between corresponding corners are used to quantify localization error in MPDIoU.

By combining the adaptive sample weighting of Focal-IoU with the geometric sensitivity of MPDIoU, Focal-MPDIoU provides a more discriminative and shape-aware supervision signal, enabling RAO-YOLO to produce tighter, more consistent bounding boxes for challenging objects in real-world highway environments. The Focal-MPDIoU loss function calculation formula as follows:

$$MPDIoU_{focal} = \begin{cases} 0, & MPDIoU < \alpha \\ \frac{MPDIoU - \alpha}{\beta - \alpha}, & \alpha \le MPDIoU \le \beta \\ 1, & MPDIoU > \beta \end{cases} \tag{22}$$

$$L_{Focal-MPDIoU} = 1 - MPDIoU_{focal} \tag{23}$$

Compared to the other three loss functions, Focal-MPDIoU achieves a more balanced computational efficiency, ensuring detection accuracy while meeting real-time requirements. It provides more precise boundary localization for small targets, better handling of overlapping RAOs scenarios, and superior detection performance for partially occluding RAOs through its refined distance-based measurement approach (see Fig. 5).



(**a**) IoU (**b**) Focal-MPDIoU

Figure 5: Comparison of IoU and Focal-MPDIoU penalty. (**a**) IoU overlap: The green box B_{gt} is the ground truth, the red box B_{prd} is the prediction, and the gray area is their intersection. The black dashed box C is the minimum enclosing. (**b**) Focal-MPDIoU penalty: In addition to the overlap, Focal-MPDIoU introduces penalties for the distances between the corresponding corners (d_1, d_2) of the two boxes, as shown by the dashed lines. This helps to penalize misalignment and improve localization accuracy.

4 Experiments and Results

4.1 RAOD Dataset

RAOD [23] is a large-scale, real-world benchmark specifically designed for abandoned object detection in road video surveillance which is a critical safety task for intelligent transportation systems. Addressing the limitations of autonomous-driving-centric datasets, RAOD offers greater diversity and realism through extensive CCTV footage captured across varied times, distances, and highway scenes. It includes 557 video sequences (over 500 with RAOs) and 18,891 pixel-annotated images, making it the largest open-source dataset for this task. Following the official dataset split provided by RAOD, the dataset is divided into 18,915 training images and 2600 testing images, which include both annotated RAO samples and additional background frames without abandoned objects. The data spans more than 70 highway locations and defines 10 common RAO categories, grouped into recognizable items (e.g., plastic bags, boxes, auto parts) and unidentifiable forms (e.g., barrel-, tabular-, or irregular-shaped objects). Representative statistics are shown in Fig. 6.

To simplify the problem and enable models to focus on learning the common features of abandoned objects, all abandoned items are unified into a single “abandoned” class, thereby avoiding potential issues related to class imbalance that fine-grained categorization might introduce. This design allows the model to emphasize shared visual characteristics across categories, improving robustness in detecting small and irregular objects. It facilitates comprehensive evaluation experiments with various baseline models from

diverse fields, including image object detection methods (e.g., YOLO series and DETR-based models) as well as representative approaches for road anomaly detection, which are more suitable for bounding-box-based evaluation in this task. Since the RAO detection task is formulated as a bounding-box-based detection problem, segmentation-based methods are not included in the comparison. Table 1 indicates the compared results between different RAOs datasets.

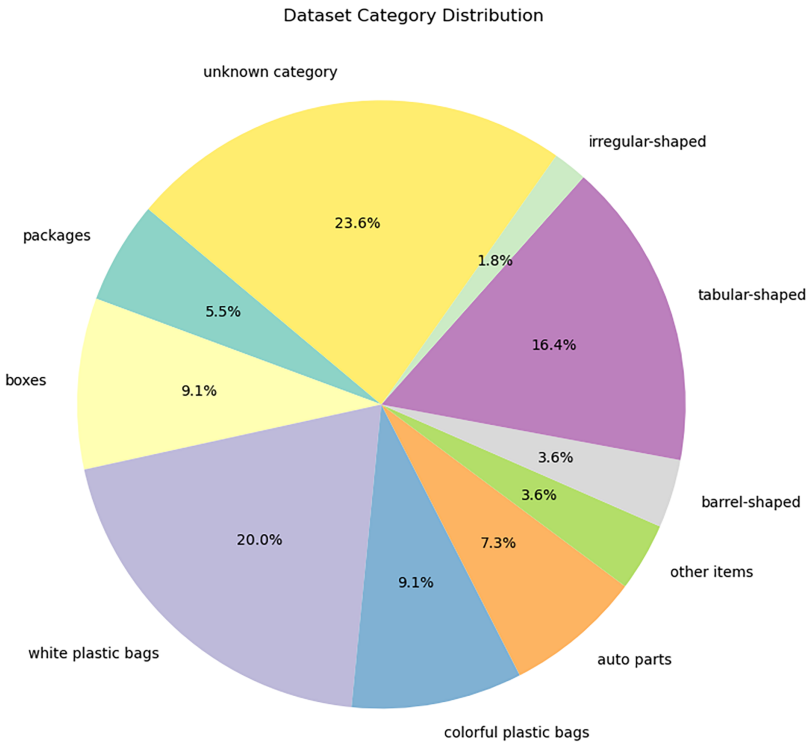


Figure 6: Proportional distribution of object categories in the RAOD dataset.

Table 1: Comparison of publicly available datasets for road abandoned objects in the number of images, positive videos and types.

Dataset	Images	Positives	Types
LostAndFound [19]	2104	112	9
RoadObstacle21 [22]	327	–	–
RAOD [23]	18,891	502	10

4.2 Experimental Detail

All experiments, including those comparing RAO-YOLO with other object detection algorithms, are implemented using PyTorch and conducted on a consistent software and hardware platform to ensure fairness and reproducibility. The hardware setup consists of an NVIDIA RTX 3090 GPU with 24 GB of VRAM, an AMD EPYC 7642 48-core CPU, and 256 GB of system memory. The system operates on the Linux distribution, Ubuntu 22.04. The software stack used for training and evaluation includes Python, the PyTorch deep learning framework, and the NVIDIA CUDA Toolkit, among other supporting libraries. Following the official dataset split, the dataset consists of 18,915 training images and 2600 testing images,

where the training set is used for model learning and the test set is used for final performance evaluation, while data augmentation is applied during training to reduce overfitting. The model was trained using the stochastic gradient descent (SGD) optimizer with an initial learning rate of 0.01 following the default training configuration of the framework, and no warm-up strategy was applied during training.

Regarding the experimental setup and hyperparameter selection, RAO-YOLO and other object detection models primarily rely on default configurations, with the exception of aggregated performance metrics, which are closely tracked throughout the training process. To ensure consistency and minimize the impact of input image resolution on the evaluation outcomes, all experiments were conducted using a standardized input size of 640 × 640 pixels, which is widely adopted in related research. For object detection algorithms, particularly those in the YOLO series, a unified performance metric that is commonly referred to as the *fitness function* is employed during training to guide model evaluation and selection. This function integrates multiple metrics into a single scalar score, with the default formulation defined as:

$$F = 0.1 \times mAP@0.5 + 0.9 \times mAP@0.5{:}\,0.95 \tag{24}$$

In this context, F denotes the fitness score, while mean Average Precision (mAP) quantifies the area under the precision–recall (PR) curve. The key distinction between mAP@0.5 and mAP@0.5:0.95 lies in their respective IoU threshold criteria for bounding box matching. Based on the fitness function formulation, model selection is predominantly guided by performance on the mAP@0.5:0.95 metric. However, in specific tasks such as RAOs detection, researchers often emphasize precision and recall over mAP, due to the critical importance of reducing false positives and false negatives in intelligent transportation applications. The following sections provide a more detailed explanation of the evaluation metrics employed in this study.

4.3 Evaluation Metrics

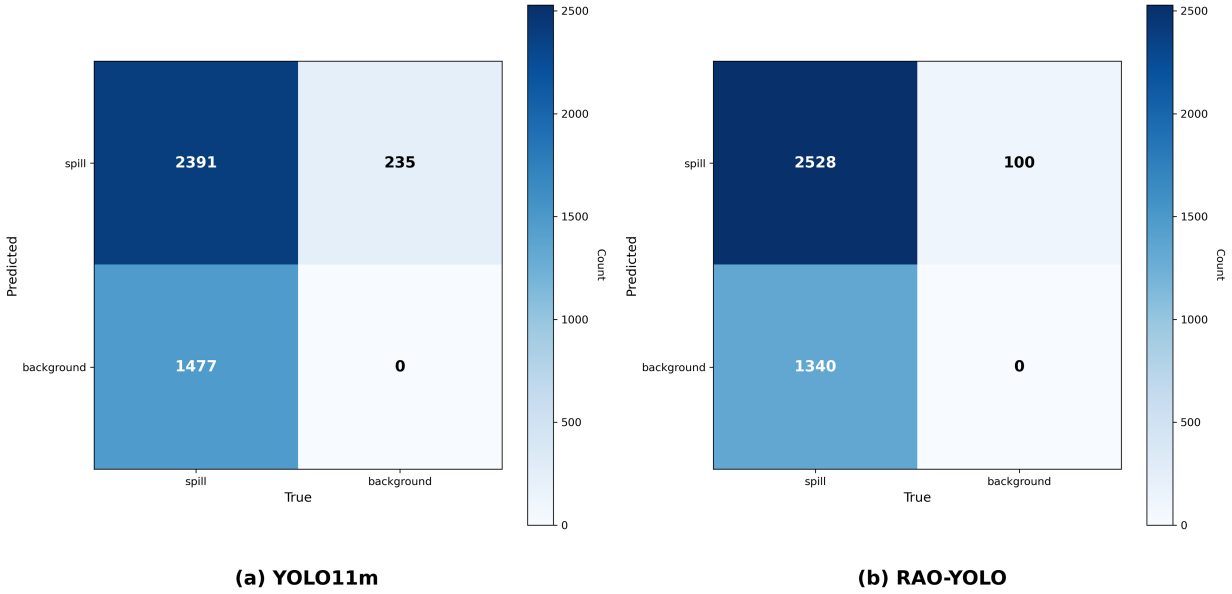
For object detection tasks, a wide range of evaluation metrics [44] are commonly used. In the context of RAOs detection, we highlighted several widely adopted indicators, including precision, recall, and F1 score. Understanding these metrics requires familiarity with the foundational classification outcomes: true positive (TP), false positive (FP), true negative (TN), and false negative (FN). A true positive (TP) occurs when the model correctly identifies the presence of RAOs in an image. FP refers to an incorrect prediction where the model detects RAOs in an image that does not contain one. TN is when the model correctly predicts the absence of RAOs. FN arises when the model fails to detect RAOs that is actually present in the image. Based on these four quantities, the commonly used evaluation metrics are defined as follows:

$$Precision = \frac{TP}{TP + FP} \tag{25}$$

$$Recall = \frac{TP}{TP + FN} \tag{26}$$

Eqs. (25) and (26) present precision (fraction of correct detections among all predictions), recall (fraction of detected RAOs out of all ground-truth instances), and their harmonic mean, the F1 score. However, for RAO detection—where both identification and precise localization are critical—single metrics like recall are insufficient. We therefore adopt mAP@0.5 (at IoU = 0.5) and mAP@0.5:0.95 (averaged over IoU thresholds from 0.5 to 0.95 in 0.05 steps) as primary metrics. The former assesses object recognition capability, while the latter provides a stricter, more comprehensive evaluation of localization accuracy and robustness.

As shown in Fig. 7, the confusion matrix highlights RAO-YOLO's strong performance: it achieves a high true positive count, indicating reliable RAO detection across diverse road conditions, while maintaining low false positives, demonstrating high precision and resistance to over-detection, crucial for minimizing false alarms. Compared to YOLOv11m, RAO-YOLO significantly increases true positives and reduces both false negatives and false positives, confirming its superior overall detection capability. The confusion matrix also reveals the distribution of misclassification cases, providing insight into the typical error patterns of the proposed model.



(a) YOLO11m **(b) RAO-YOLO**

Figure 7: Confusion matrix of different model on the RAOD test set. The matrix illustrates the distribution of true positives, false positives, and false negatives, highlighting the improved detection performance of RAO-YOLO.

While some false negatives still occur, true RAOs missed by the model, the overall pattern suggests that RAO-YOLO prioritizes precision without heavily compromising recall. This trade-off reflects the model's well-calibrated balance between accuracy and efficiency, making it particularly suitable for real-time roadside safety applications. From the confusion matrix, we can find it underscores RAO-YOLO's effectiveness in detecting varied and often subtle RAOs instances, confirming its advantage in both detection reliability and practical deployability compared to conventional models.

4.4 Comparative Experiments

We select YOLOv11m as the baseline due to its strong balance of accuracy, model complexity, and inference speed. While numerous detectors exist, YOLOv11m consistently excels due to architectural enhancements that support both efficiency and precision, crucial for real-time industrial systems. For a fair comparison, we evaluate it against recent models from the YOLOm and YOLOx series, as well as transformer-based RT-DETR, covering different design paradigms (CNN vs. transformer) and accuracy-efficiency trade-offs. All models are tested on the RAOD dataset under identical training and hardware conditions. Since the proposed MANet module operates on multi-scale feature maps without significantly increasing channel dimensions, the overall computational complexity remains comparable to the baseline YOLO architecture while improving feature representation capability.

As shown in Table 2, YOLOx and RT-DETR variants, despite deeper backbones, offer no significant accuracy gains but incur higher GPU memory usage and slower inference, making them less suitable for embedded deployment. Notably, YOLOv11m slightly outperforms YOLOv12m in mAP@0.5 while achieving much higher FPS, confirming its runtime efficiency. More importantly, our proposed RAO-YOLO surpasses all competitors in both mAP@0.5 (Fig. 8) and mAP@0.5:0.95, demonstrating superior detection accuracy and precise localization across varying IoU thresholds. This highlights its robustness in challenging RAO scenarios, where fast and reliable hazard identification is critical. Overall, RAO-YOLO achieves the best balance of accuracy, speed, and efficiency, making it the optimal choice for real-world RAO detection tasks.

Table 2: Comparative experimental results. The best results in each column are highlighted in bold.

Model	Precision	Recall	mAP@50	mAP@50:95	FPS	Size (M)
Faster-RCNN [13]	0.773	0.591	0.702	0.446	15	42
RetinaNet [45]	0.750	0.680	0.692	0.349	22	38
RT-DETR-l [33]	0.901	0.656	0.734	0.454	111.75	63.1
RT-DETR-x	0.865	0.665	0.747	0.46	71.19	129.1
YOLOv9m [46]	0.807	0.632	0.756	0.526	240.27	39.0
YOLOv10m [47]	0.869	0.668	0.778	0.533	286.8	31.9
Hyper-YOLOm [39]	**0.903**	0.685	0.785	0.532	182.11	59.1
YOLOv12m [48]	0.896	0.688	0.791	0.530	218.51	38.9
YOLOv12x	0.865	**0.693**	0.784	0.531	97.89	113.6
YOLOv11m [49]	0.866	0.667	0.792	0.535	261.81	38.6
YOLOv11x	0.872	0.692	0.790	0.531	122.96	109.1
RAO-YOLO	0.880	0.643	**0.808**	**0.561**	204.04	61.4

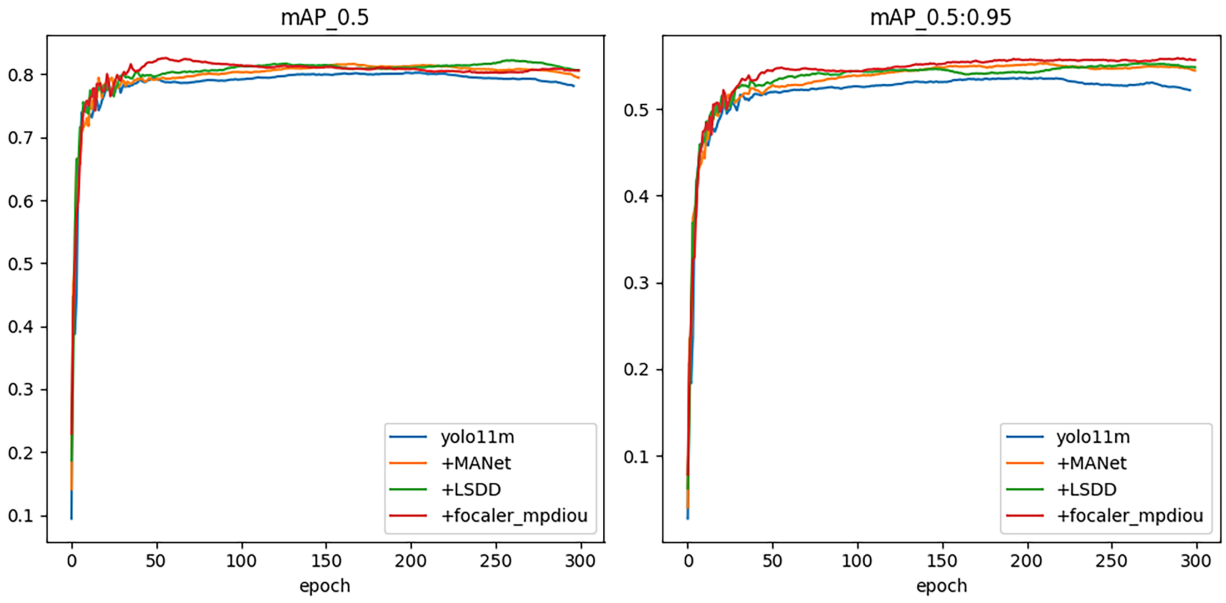


Figure 8: mAP curves of the ablation experiments.

4.5 Ablation Studies

To evaluate the impact of different loss functions on detection performance, we have conducted an ablation study comparing CIoU, Shape-IoU, Inner-IoU, and Focal-MPDIoU within the RAO-YOLO framework. As shown in Table 3, CIoU achieves the highest mAP@0.5 score (0.821), indicating strong performance at loose localization thresholds. However, its mAP@0.5:0.95 score is relatively lower (0.556), suggesting less precise bounding box regression across stricter IoU levels. In contrast, Focal-MPDIoU achieves the highest mAP@0.5:0.95 score (0.561), demonstrating superior localization accuracy under multi-threshold evaluation, which is crucial for detecting small and irregularly shaped RAOs.

Table 3: Comparison of detection performance under different loss functions. The best results in each column are highlighted in bold.

Loss Function	mAP@50	mAP@50:95
CIoU	**0.821**	0.556
Shape-IoU	0.800	0.549
Inner-IoU	0.819	0.551
Focal-MPDIoU	0.808	**0.561**

While Shape-IoU and Inner-IoU show reasonable performance, their lower mAP metrics indicate less robustness in diverse real-world conditions. The Focal-MPDIoU loss, by combining focal weighting with a geometry-adaptive penalty, effectively mitigates the impact of class imbalance and enhances bounding box refinement. Based on this performance gain, particularly in fine-grained localization, we have adopt Focal-MPDIoU as the default loss function in the final RAO-YOLO model. The comparison of detection performance under different loss functions is presented in Table 3.

To assess the contribution of each component in RAO-YOLO, we conduct ablation studies on the RAOD dataset, focusing on three key modules: MANet, LSDD, and Focal-MPDIoU loss (see Table 4). The baseline YOLOv11m achieves 86.6% precision, 66.7% recall, 79.2% mAP@0.5, and 53.5% mAP@0.5:0.95 at 261.81 FPS (38.6 MB). Adding MANet alone boosts performance to 89.1% precision, 68.5% recall, and 81.4% mAP@0.5, while maintaining >200 FPS. Further integrating LSDD raises mAP@0.5 to 82.1% and mAP@0.5:0.95 to 55.6%, albeit with a larger model (61.4 MB) and slightly lower FPS (204.30), enhancing multi-scale RAO detection.

Table 4: Results of the ablation experiments. The best results in each column are highlighted in bold.

Baseline	MANet	LSDD	Focal-MPDIoU	mAP50	mAP50:95	FPS	Size
✓				0.792	0.535	261.81	38.6
✓	✓			0.814	0.552	215.16	55.2
✓	✓	✓		**0.821**	0.556	204.30	61.4
✓ (Ours)	✓	✓	✓	0.808	**0.561**	204.04	61.4

Starting from the YOLOv11m baseline (79.2% mAP@0.5, 53.5% mAP@0.5:0.95), adding MANet improves performance by +2.2% mAP@0.5 and +1.7% mAP@0.5:0.95; further integrating the LSDD head yields an additional +0.7% mAP@0.5 and +0.4% mAP@0.5:0.95; finally, the Focal-MPDIoU loss slightlyincreases mAP@0.5:0.95 by +0.5%, demonstrating its focus on enhancing high-quality localization. As shown in Fig. 9, replacing the original loss with Focal-MPDIoU yields consistently lower training and validation

losses, indicating faster convergence and stable performance, confirming its effectiveness in improving RAO detection accuracy.

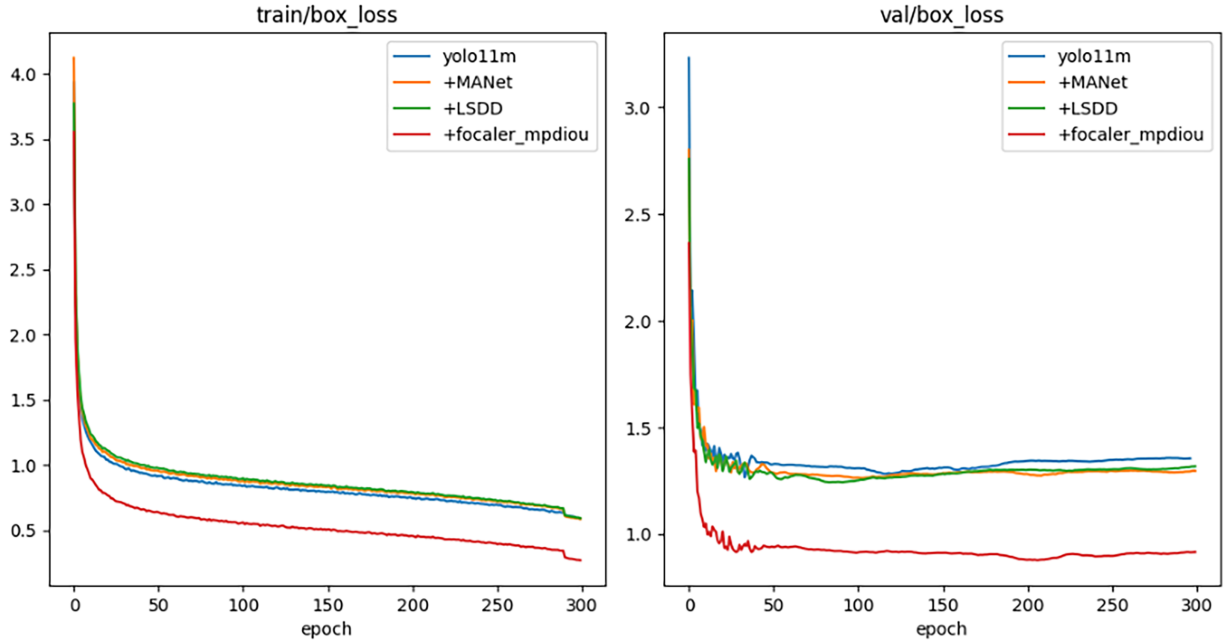


Figure 9: Training and validation loss curves of the ablation experiments.

As shown in Table 4, RAO-YOLO, which includes all three modules (i.e., MANet, LSDD, and the Focal-MPDIoU loss) achieves the best performance in terms of mAP@0.5:0.95 (56.1%) and maintains strong results in all other metrics. This confirms the effectiveness of the combined enhancements in improving bounding box localization and detection reliability, especially for challenging RAO categories under complex background conditions. These results show that RAO-YOLO's architectural and loss function enhancements work synergistically to achieve high accuracy with acceptable computational efficiency, making it well-suited for real-time roadside perception. This is further illustrated in Fig. 8 (mAP curves) and Fig. 10 (PR curves), where our full model consistently outperforms baselines, confirming its superior localization precision.

4.6 Visualization Analysis

To comprehensively evaluate RAO-YOLO, we present heatmaps in Fig. 11 and qualitative results on representative images in Fig. 12, covering diverse real-world conditions including day/night scenes, high-speed traffic, varied weather [50] (e.g., glare, shadows, strong sunlight), and complex road backgrounds (e.g., markings, barriers, curved lanes). Each image includes ground-truth annotations of abandoned objects, which our model accurately localizes with red dotted bounding boxes.

RAO-YOLO shows strong robustness in challenging cases: it reliably detects small, partially occluded, or motion-blurred objects even under low-light conditions. Zoomed-in insets (upper-left corners) highlight its fine-grained localization for small-scale targets which are critical for real-time safety in autonomous driving and roadside monitoring. This visual performance aligns with quantitative gains from our ablation and benchmark studies, driven by the MANet feature extractor, LSDD, and Focal-MPDIoU loss, which together improve boundary precision and detection robustness.

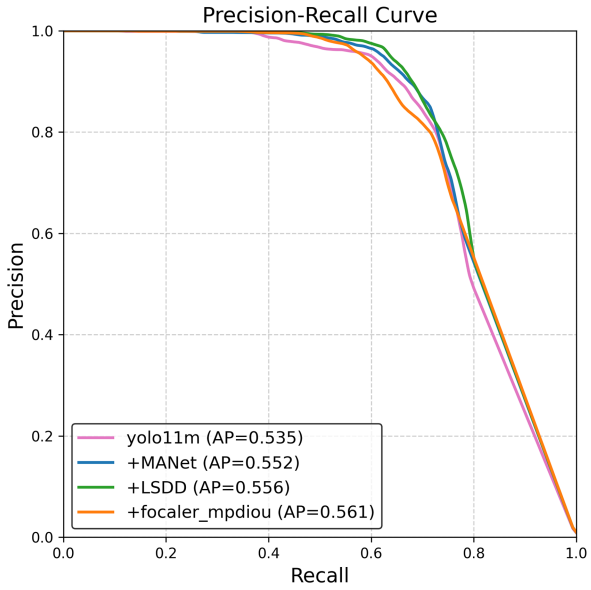


Figure 10: Precision-Recall (PR) curves of the ablation experiments.

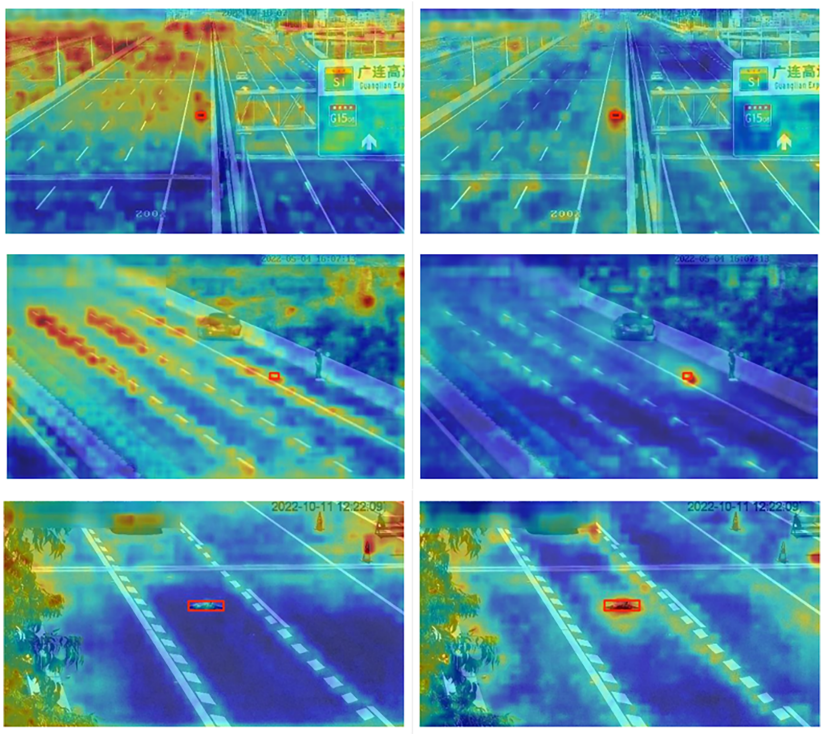


Figure 11: Example heatmap detection results on the RAOD dataset. (**a**) Detection results generated by the YOLOv11m model; (**b**) Detection results generated by the proposed RAO-YOLO model. The heatmaps indicate the response intensity of the models, where warmer colors represent stronger activations. Red bounding boxes highlight the detected RAO targets.

(**a**)

(**b**)

Figure 12: (Continued)

(**c**)

Figure 12: Visual comparison of different models on the RAOD dataset. (**a**) Ground truth, where red dashed bounding boxes indicate the locations of RAO targets to be detected. (**b**) Detection results of YOLOv11m, and (**c**) detection results of the proposed RAO-YOLO. In (**b**,**c**), blue bounding boxes denote the detected RAOs. Additionally, in (**b**), red solid bounding boxes highlight missed or inaccurate detections by YOLOv11m. The top-right corner of each image shows a magnified view of the detected region.

Compared to YOLOv11m, RAO-YOLO produces stronger, more focused activations around true RAOs while suppressing irrelevant background, reflecting enhanced spatial awareness and semantic discrimination. It excels at detecting irregular, elongated, scattered, or occluded RAOs. These improvements stem largely from the Focal-MPDIoU loss, which combines focal weighting with distance-aware IoU to handle class imbalance and geometric variability. This yields tighter bounding boxes, better scale invariance, lower localization error, and greater robustness which are key attributes for deployment in intelligent transportation and edge computing systems where reliability and efficiency are paramount.

5 Conclusion

This paper presents RAO-YOLO, a task-specific object detection framework for addressing the critical safety problem of abandoned objects on highways. By integrating the MANet module for enhanced multi-scale feature representation, the LSDD head for efficient and fine-grained localization, and the Focal-MPDIoU loss for robust training on irregular and small targets, RAO-YOLO effectively overcomes the limitations of conventional monocular vision-based methods.

Experiments on the RAOD dataset demonstrate that RAO-YOLO achieves a mAP@0.5:0.95 of 56.1%, outperforming all evaluated baselines while maintaining high inference speed and a compact model size suitable for edge-oriented deployment. Although real-time performance is validated on a GPU platform, future work will further assess its effectiveness on embedded and edge devices. Qualitative results and

heatmap visualizations indicate that RAO-YOLO produces more accurate and concentrated responses on abandoned objects, while effectively suppressing background interference such as shadows, road markings, and surrounding vehicles. The model also exhibits robust performance under challenging conditions, including low-light environments, motion blur, occlusion, and irregular object shapes, where baseline methods tend to degrade.

Despite these promising results, several limitations remain. The current model is trained solely on RGB images, which may limit robustness under extreme weather and nighttime conditions. Temporal information from consecutive frames is not explicitly exploited, and evaluation is restricted to a single benchmark. In real-world highway monitoring scenarios, abandoned objects are typically observed in continuous video streams, where temporal cues across frames could provide additional contextual information to improve detection stability and reduce prediction fluctuations. Future work will explore hierarchical RAO categorization (e.g., plastic bags, metal fragments, construction materials), enabling more informative hazard classification for road maintenance operations, and explore synthetic data augmentation to generate additional RAO samples. Future work will address these limitations by incorporating multimodal sensor inputs (e.g., LiDAR, radar), integrating temporal reasoning for video-based detection, and extending validation to more diverse datasets and environments. Such efforts will further enhance the robustness, adaptability, and deployment potential of RAO-YOLO in real-world intelligent transportation systems.

Acknowledgement: This work is partially supported by the Natural Science Foundation of, China Postdoctoral Science Foundation and Natural Science Foundation of Shandong Province. All the supports are highly appreciated.

Funding Statement: This work is partially supported by the Natural Science Foundation of China (Grant Number: 52308457), China Postdoctoral Science Foundation (Grant Number: 2024M761811) and Natural Science Foundation of Shandong Province (Grant Number: ZR2023QE220). All the supports are highly appreciated.

Author Contributions: Conceptualization, methodology, writing—original draft preparation, Ying Tang; Resources, investigation, writing—original draft preparation, Chuanyi Ma; Validation, supervision, writing—review and editing, Feng Guo; Software, visualization, Wenhao Sun. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The datasets used in this study are publicly available. The RAOD (Road Abandoned Object Detection) dataset can be accessed at: https://github.com/yajunbaby/A-Benchmark-for-Road-Abandoned-Object-Detection-from-Video-Surveillance.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. NHTSA. Administration, motor vehicle traffic crash data resource page, national highway traffic safety administration. 2022 [cited 2026 Jan 1]. Available from: https://crashstats.nhtsa.dot.gov/#!/.
2. Tan H, Zhao F, Hao H, Liu Z. Cost analysis of road traffic crashes in China. Int J Inj Control Saf Promot. 2020;27(3):385–91. doi:10.1080/17457300.2020.1785507.
3. Zhu L, Yu FR, Wang Y, Ning B, Tang T. Big data analytics in intelligent transportation systems: a survey. IEEE Trans Intell Transport Syst. 2019;20(1):383–98. doi:10.1109/tits.2018.2815678.
4. Sun Z, Bebis G, Miller R. On-road vehicle detection: a review. IEEE Trans Pattern Anal Machine Intell. 2006;28(5):694–711. doi:10.1109/tpami.2006.104.
5. Hoffmann JE, Tosso HG, Santos MMD, Justo JF, Malik AW, Rahman AU. Real-time adaptive object detection and tracking for autonomous vehicles. IEEE Trans Intell Veh. 2021;6(3):450–9. doi:10.1109/tiv.2020.3037928.

6. Stauffer C, Grimson WE. Adaptive background mixture models for real-time tracking. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR); 1999 Jun 23–25; Fort Collins, CO, USA. p. 246–52.
7. Kass M, Witkin A, Terzopoulos D. Snakes: active contour models. Int J Comput Vis. 1988;1(4):321–31. doi:10.1007/BF00133570.
8. Champahom T, Se C, Watcharamaisakul F, Jomnonkwao S, Karoonsoontawong A, Ratanavaraha V. Tree-based approaches to understanding factors influencing crash severity across roadway classes: a Thailand case study. IATSS Res. 2024;48(3):464–76. doi:10.1016/j.iatssr.2024.09.001.
9. Labayrade R, Aubert D, Tarel JP. Real time obstacle detection in stereovision on non flat road geometry through "v-disparity" representation. In: Proceedings of the Intelligent Vehicle Symposium; 2002 Jun 17–21; Versailles, France. p. 646–51.
10. Matthies L, Shafer S. Error modeling in stereo navigation. IEEE J Robot Automat. 1987;3(3):239–48. doi:10.1109/jra.1987.1087097.
11. Chen X, Ma H, Wan J, Li B, Xia T. Multi-view 3D object detection network for autonomous driving. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017 Jul 21–26; Honolulu, HI, USA. p. 1907–15. doi:10.1109/cvpr.2017.691.
12. Li Z, Du Y, Zhu M, Zhou S, Zhang L. A survey of 3D object detection algorithms for intelligent vehicles development. Artif Life Robot. 2022;27(1):115–22. doi:10.1007/s10015-021-00711-0.
13. Ren S, He K, Girshick R, Sun J. Faster R-CNN towards real-time object detection with region proposal networks. In: Advances in neural information processing systems. Montréal, QC, Canada; 2015. p. 91–9.
14. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu CY, et al. SSD: single shot multibox detector. In: Computer vision—ECCV 2016. Berlin/Heidelberg, Germany: Springer; 2016. p. 21–37.
15. Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: unified, real-time object detection. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016 Jun 27–30; Las Vegas, NV, USA. doi:10.1109/cvpr.2016.91.
16. Veres M, Moussa M. Deep learning for intelligent transportation systems: a survey of emerging trends. IEEE Trans Intell Transport Syst. 2020;21(8):3152–68. doi:10.1109/tits.2019.2929020.
17. Cordts M, Omran M, Ramos S, Rehfeld T, Enzweiler M, Benenson R, et al. The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016 Jun 27–30; Las Vegas, NV, USA. p. 3213–23. doi:10.1109/cvpr.2016.350.
18. Geiger A, Lenz P, Urtasun R. Are we ready for autonomous driving? The KITTI vision benchmark suite. In: Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition; 2012 Jun 16–21; Providence, RI, USA. p. 3354–61. doi:10.1109/cvpr.2012.6248074.
19. Pinggera P, Ramos S, Gehrig S, Franke U, Rother C, Mester R. Lost and found: detecting small road hazards for self-driving vehicles. In: Proceedings of the 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); 2016 Oct 9–14; Daejeon, Republic of Korea. p. 1099–106. doi:10.1109/iros.2016.7759186.
20. Yang Z, Radke RJ. Context-aware video anomaly detection in long-term datasets. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2024; 2024 Jun 16–22; Seattle, WA, USA. p. 4002–11. doi:10.1109/CVPRW63382.2024.00404.
21. Yu F, Chen H, Wang X, Xian W, Chen Y, Liu F, et al. BDD100K: a diverse driving dataset for heterogeneous multitask learning. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun 13–19; Seattle, WA, USA. p. 2636–45. doi:10.1109/cvpr42600.2020.00271.
22. Chan R, Lis K, Uhlemeyer S, Blum H, Honari S, Siegwart R, et al. Segmentmeifyoucan: a benchmark for anomaly segmentation. arXiv:2104.14812. 2021.
23. Xu Y, Hu H, Zhu X, Nan Y, Wang K, Liu Z, et al. RAOD: a benchmark for road abandoned object detection from video surveillance. IEEE Access. 2024;12:123985–94. doi:10.1109/access.2024.3407955.
24. Latha YM, Rao BS. A systematic review on background subtraction model for data detection. In: Pervasive computing and social networking. Singapore: Springer; 2022. p. 341–9. doi:10.1007/978-981-16-5640-8_27.

25. Ojala T, Pietikäinen M, Harwood D. A comparative study of texture measures with classification based on featured distributions. Pattern Recognit. 1996;29(1):51–9. doi:10.1016/0031-3203(95)00067-4.
26. Dalal N, Triggs B. Histograms of oriented gradients for human detection. In: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05); 2005 Jun 20–26; San Diego, CA, USA. p. 886–93.
27. Garcia-Garcia B, Bouwmans T, Rosales Silva AJ. Background subtraction in real applications: challenges, current models and future directions. Comput Sci Rev. 2020;35(1):100204. doi:10.1016/j.cosrev.2019.100204.
28. Ros G, Sellart L, Materzynska J, Vazquez D, Lopez AM. The SYNTHIA dataset: a large collection of synthetic images for semantic segmentation of urban scenes. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016 Jun 27–30; Las Vegas, NV, USA. p. 3234–43. doi:10.1109/cvpr.2016.352.
29. Bakirci M. Advanced aerial monitoring and vehicle classification for intelligent transportation systems with YOLOv8 variants. J Netw Comput Appl. 2025;237(B):104134. doi:10.1016/j.jnca.2025.104134.
30. Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S. End-to-end object detection with transformers. In: Computer vision—ECCV 2020. Cham, Switzerland: Springer International Publishing; 2020. p. 213–29. doi:10.1007/978-3-030-58452-8_13.
31. Zhu X, Su W, Lu L, Li B, Wang X, Dai J. Deformable DETR: deformable transformers for end-to-end object detection. arXiv:2010.04159. 2021.
32. Peng Y, Li H, Wu P, Zhang Y, Sun X, Wu F. D-FINE: redefine regression task in DETRs as fine-grained distribution refinement. arXiv:2410.13842. 2024.
33. Zhao Y, Lv W, Xu S, Wei J, Wang G, Dang Q, et al. DETRs beat YOLOs on real-time object detection. arXiv:2304.08069. 2023.
34. Ku J, Mozifian M, Lee J, Harakeh A, Waslander SL. Joint 3D proposal generation and object detection from view aggregation. In: Proceedings of the 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); 2018 Oct 1–5; Madrid, Spain. p. 1–8. doi:10.1109/iros.2018.8594049.
35. Azadvatan Y, Kurt M. Melnet: a real-time deep learning algorithm for object detection. arXiv:2401.17972. 2024.
36. Chen S, Cheng T, Fang J, Zhang Q, Li Y, Liu W, et al. TinyDet: accurate small object detection in lightweight generic detectors. arXiv:2304.03428. 2023.
37. Woo S, Park J, Lee JY, Kweon IS. CBAM: convolutional block attention module. In: Computer vision—ECCV 2018. Cham, Switzerland: Springer International Publishing; 2018. p. 3–19. doi:10.1007/978-3-030-01234-2_1.
38. Wang Q, Wu B, Zhu P, Li P, Zuo W, Hu Q. ECA-net: efficient channel attention for deep convolutional neural networks. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun 13–19; Seattle, WA, USA. p. 11531–9. doi:10.1109/cvpr42600.2020.01155.
39. Feng Y, Huang J, Du S, Ying S, Yong JH, Li Y, et al. Hyper-YOLO: when visual object detection meets hypergraph computation. IEEE Trans Pattern Anal Mach Intell. 2025;47(4):2388–401. doi:10.1109/tpami.2024.3524377.
40. Wu Y, He K. Group normalization. In: Computer vision—ECCV 2018. Cham, Swizerland: Springer International Publishing; 2018. p. 3–19. doi:10.1007/978-3-030-01261-8_1.
41. Zheng Z, Wang P, Liu W, Li J, Ye R, Ren D. Distance-IoU loss: faster and better learning for bounding box regression. Proc AAAI Conf Artif Intell. 2020;34(7):12993–3000. doi:10.1609/aaai.v34i07.6999.
42. Zhang H, Zhang S. Focaler-IoU: more focused intersection over union loss. arXiv:2401.10525. 2024.
43. Ma S, Xu Y. MPDIoU: a loss for efficient and accurate bounding box regression. arXiv:2307.07662. 2023.
44. Padilla R, Netto SL, da Silva EAB. A survey on performance metrics for object-detection algorithms. In: Proceedings of the 2020 International Conference on Systems, Signals and Image Processing (IWSSIP); 2020 Jul 1–3; Niterói, Brazil. doi:10.1109/iwssip48289.2020.9145130.
45. Lin TY, Goyal P, Girshick R, He K, Dollar P. Focal loss for dense object detection. In: Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV); 2017 Oct 22–29; Venice, Italy. p. 2980–8. doi:10.1109/iccv.2017.324.
46. Wang CY, Yeh IH, Liao HYM. YOLOv9: learning what you want to learn using programmable gradient information. arXiv:2402.13616. 2024.

47. Wang A, Chen H, Liu L, Chen K, Lin Z, Han J, et al. YOLOv10: real-time end-to-end object detection. arXiv:2405.14458. 2024.
48. Tian Y, Ye Q, Doermann D. YOLOv12: attention-centric real-time object detectors. arXiv:2502.12524. 2025.
49. Khanam R, Hussain M. YOLOv11: an overview of the key architectural enhancements. arXiv:2410.17725. 2024.
50. Iqra, Giri KJ, Javed M. Small object detection in diverse application landscapes: a survey. Multimed Tools Appl. 2024;83(41):88645–80. doi:10.1007/s11042-024-18866-w.