algorithms

MDPI

Article

A Lightweight and Efficient Detection Transformer for Highway Abandoned Objects

Biao Zhang, Chishe Wang * and Jie Wang

School of Network and Communication Engineering, Jinling Institute of Technology, Nanjing 211169, China; zhangbiao9992023@163.com (B.Z.); wangjie@jit.edu.cn (J.W.)
* Correspondence: wyxcs@163.com

Abstract: Abandoned objects on highways seriously threaten traffic safety, and their prompt identification and removal are crucial. Existing methods struggle to balance computational cost and detection accuracy due to the significant scale differences of abandoned objects on highways. To address these problems, we propose a Lightweight and Efficient Detection Transformer for highway abandoned objects (LE-DETR). This study first designs a real-time feature extraction module that effectively captures essential information and accelerates information flow. Building on this module, we construct a lightweight backbone network for feature extraction, enhancing parameter utilization. A Triple Fusion (TFusion) module is proposed, integrating high-level semantic information with low-level spatial information to increase detailed information. A Cross-Layer Multi-Scale Interaction (CMI) module is designed, utilizing large-kernel depth-wise convolutions of various sizes to extract features from different receptive fields, enhancing the multi-scale representation of abandoned objects. The LE-DETR model is trained and evaluated using a constructed Highway Abandoned Object Dataset (HAOD). The experimental results indicate that compared to the suboptimal RT-DETR-R18, LE-DETR improves accuracy by 6.5%, reduces the number of parameters by 27.1%, and decreases floating-point operations (FLOPs) by 21.1%. These improvements demonstrate the great potential of LE-DETR for detecting abandoned objects on highways.

Keywords: abandoned object detection; deep learning; lightweight structure; multi-scale

check for updates

Academic Editors: Laura Antonelli and Lucia Maddalena

Received: 11 February 2025
Revised: 25 February 2025
Accepted: 1 March 2025
Published: 2 March 2025

Citation: Zhang, B.; Wang, C.; Wang, J. A Lightweight and Efficient Detection Transformer for Highway Abandoned Objects. *Algorithms* **2025**, *18*, 133. https://doi.org/10.3390/a18030133

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Highway traffic safety has long been one of the main issues in road safety. However, abandoned objects seriously threaten highway safety. Therefore, the rapid identification and removal of abandoned objects on highways are essential for improving road safety. Manual inspection is the primary method for identifying abandoned objects, but it is time-consuming, ineffective, and unable to meet real-time monitoring needs. Moreover, manual inspection is inappropriate in an environment with rapidly moving traffic.

Many traditional techniques for detecting abandoned objects have been presented in recent years. Fu et al. [1] used a Gaussian Mixture Model (GMM) [2] for background modeling and edge statistical methods to eliminate noise and optimize detection using an integrated object-tracking module. However, when the dynamic environment changes heavily, this method does not update frequently enough, delaying the recognition of emerging static objects and reducing overall detection accuracy. Szwoch [3] proposed a method for identifying lost luggage based on the stability of foreground pixel values. Nevertheless, this method is primarily appropriate for interior environments and is only

effective with larger objects, rendering it unsuitable for complex highway scenarios. Zeng et al. [4] proposed a detection system that utilizes three-dimensional image information. The system initially employs a monocular camera video to perform preliminary static foreground segmentation to identify suspected abandoned objects. Subsequently, a stereo camera performs 3D reconstruction to evaluate potential traffic threats. However, the effectiveness of this system in dynamic and complex road environments is significantly restricted by its reliance on complex image processing techniques and precise stereo camera calibration. Lin et al. [5] employed a GMM to build a double learning rate background modeling system that analyzes the differences between various frames to identify static or dissipating objects in the foreground. However, due to its high computational cost, this system is inappropriate for application scenarios requiring a prompt response.

With the rapid advancement of deep learning technologies, significant progress has been made in applying these methods for detecting abandoned objects on highways. Park et al. [6] presented a technique that integrates dual-background modeling with Mask R-CNN [7] to detect abandoned and stolen objects from video surveillance systems. This method effectively detects stationary objects and reduces errors induced by variations in lighting, but it is unsuitable for highway scenarios. Song et al. [8] proposed a method combining MOG2 [9] background modeling and YOLOv9 [10] for abandoned object detection on highways. This approach improves detection accuracy in dynamic environments but needs optimization for handling noise and vehicle occlusion. Qian et al. [11] proposed a novel 20-layer residual network (Nov-ResNet-20) to detect abandoned objects in real-time traffic surveillance videos. This network combines selective search [12,13] and Non-Maximum Suppression (NMS) [14,15] to optimize the object detection process and improve abandoned objects' recognition rate and accuracy. Although the method can effectively identify abandoned objects in traffic videos, the processing time of the images remains lengthy, affecting the method's efficiency in practical applications. Liu et al. [16] proposed a lightweight YOLOv5-MN3 convolutional neural network, significantly reducing the parameters and computational costs by employing a depth-separable convolution instead of a standard convolution. Furthermore, the network integrates an attention mechanism to improve the extraction of features and optimizes the model by employing knowledge distillation techniques. While the model successfully attains a lightweight design, the implementation method is more complicated. Zhou et al. [17] proposed a small object detection method that integrates an adaptive dual-background model with SAO-YOLO, aiming to reduce false positives and false negatives for small and occluded objects. However, this method is not specifically designed for highway scenarios.

Despite these advancements, several research gaps remain. First, existing models struggle to balance computational cost and real-time performance, making them unsuitable for real-time highway monitoring. Second, multi-scale detection capabilities are insufficient for complex highway scenarios, where abandoned objects vary significantly in size and distance from the camera. Finally, there is a lack of lightweight detection models specifically designed for highway environments.

To resolve these issues, we propose LE-DETR, a lightweight end-to-end highway abandoned object detection model based on the Real-Time Detection Transformer with a ResNet18 backbone (RT-DETR-R18) [18]. Compared to existing methods, LE-DETR introduces a novel lightweight design and efficient multi-scale detection capabilities. Specifically, to address the challenge of balancing computational cost and real-time performance in existing methods, we designed a lightweight and efficient Re-parameterized Efficient Layer Aggregation Network (RepELAN) module for real-time feature extraction. This module reduces the use of residual structures and incorporates the RepConv module with several 3×3 convolution combinations to simplify the structure and accelerate information flow

while capturing more effective data. Based on the RepELAN module, we constructed a Re-parameterized Lightweight Network (RLNet) as a feature extraction backbone for real-time object detection tasks. This network significantly optimizes detection efficiency while maintaining or enhancing detection accuracy. To further enhance the model's multi-scale detection capabilities in complex highway environments, we propose an Efficient Multi-scale Hybrid encoder called the EMHencoder to address significant scale variations in abandoned objects. This encoder consists of two main modules: the Triple Fusion (TFusion) module and the Cross-Layer Multi-Scale Interaction (CMI) module. The TFusion module enriches the representation of detail by combining high-level semantic information from deeper layers with low-level spatial information from superficial layers, thereby improving detection performance. The CMI module enhances the capability to detect multi-scale abandoned objects by merging features across different layers and employing large-kernel depth convolutions of varying sizes to extract features from diverse receptive fields. Finally, to validate our model's performance, we constructed a new dataset of abandoned objects on highways and performed a series of detailed experiments.

In summary, the contributions in this paper can be summarized as follows:

(1) We present LE-DETR, an end-to-end abandoned object detection model distinguished by its real-time response capability, efficient processing performance, and lightweight design.
(2) We designed an efficient real-time feature extraction module, RepELAN. The experimental results demonstrate that this module surpasses similar real-time detection modules in terms of parameter count, floating-point operations, inference speed, and detection accuracy. Building on this module, we constructed a new feature extraction backbone network, RLNet, which reduces computational load, enhances parameter utilization, and significantly improves detection performance.
(3) We introduce an encoder with efficient multi-scale feature extraction capabilities named EMHencoder. It includes two core modules: TFusion and CMI. The interaction effect between these modules optimizes handling multi-scale objects in complex scenarios, resulting in more accurate and efficient detection outcomes in practical applications.
(4) We constructed a new highway abandoned object dataset, which includes images of abandoned objects on highways under various angles and lighting conditions.

2. Methodology

2.1. Introduction to the Baseline Model

This paper utilizes RT-DETR-R18 as the baseline model, an advanced real-time end-to-end object detection model. Compared to traditional YOLO models, RT-DETR incorporates the self-attention mechanism of the Transformer [19] architecture, significantly enhancing its ability to capture interrelations within images, especially improving detection accuracy in complex scenarios. Additionally, RT-DETR is optimized for real-time processing, making it more suitable for applications requiring rapid response.

The structure of the RT-DETR-R18 model, as shown in Figure 1, consists of four main components: a feature extraction backbone network, an efficient hybrid encoder, an IoU-aware Query Selection mechanism, and a Transformer decoder with an auxiliary prediction head. The feature extraction backbone network uses an improved ResNet18, which includes ConvBNReLU (CBR), max pooling, and BasicBlock modules. The efficient hybrid encoder processes high-level features from the backbone network using an Attention-based Intra-scale Feature Interaction module (AIFI) and fuses features across scales using a CNN-based Cross-Scale Feature Fusion module (CCFF). This design significantly reduces computational demands while enhancing processing speed and maintaining detection performance. The

IoU-aware Query Selection mechanism selects image features with high classification and IoU scores from the encoder output as initial queries for the decoder. The Transformer decoder, equipped with an auxiliary prediction head, continually refines object queries to generate more accurate bounding boxes and confidence scores.

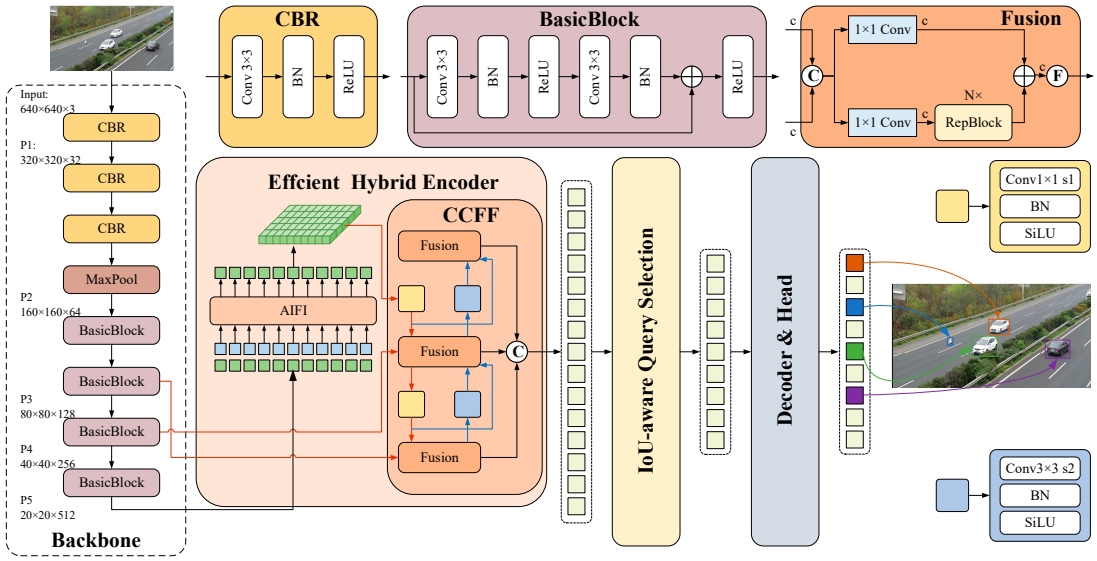


Figure 1. Structure of the RT-DETR-R18 model.

2.2. Proposed Methodology for the Detection of Abandoned Objects on Highways

As illustrated in Figure 2, this study presents LE-DETR, a lightweight end-to-end model for detecting abandoned objects on highways, which includes three main modules: RepELAN, TFusion, and CMI. These modules enhance the model's capacity to detect abandoned objects on highways, effectively handling considerable differences in the scale of objects and reducing the model's high computational costs. In the following sections, we will briefly introduce RepELAN and RLNet in Section 2.2.1, followed by an overview of TFusion in Section 2.2.2 and, finally, an introduction to CMI in Section 2.2.3.

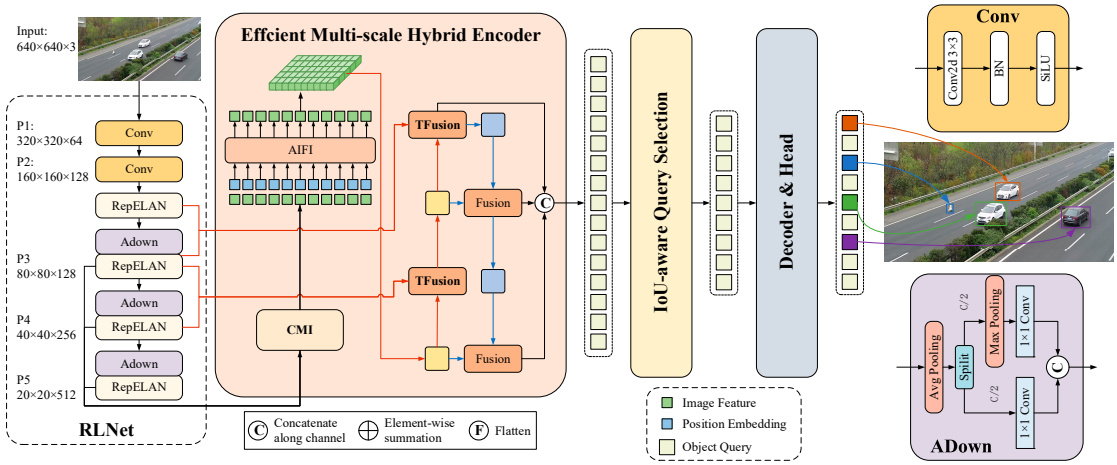


Figure 2. Structure of the LE-DETR model.

2.2.1. Design of RepELAN and RLNet

The task of abandoned object detection poses significant challenges in balancing efficiency and accuracy. Although previous research has explored various design strategies, these strategies are still unable to avoid inherent computational redundancies that result in an extra increase in computational overhead. To address this challenge, we

propose RepELAN, a lightweight cascading feature extraction module. By optimizing computational architecture and memory demands, this module enhances real-time capabilities, significantly reduces computational resource consumption, and greatly improves detection performance, making it more suitable for deployment on resource-constrained embedded devices.

As depicted in Figure 3, we designed RepELAN based on the C2f feature extraction module from YOLOv8 and the backbone network concept from RepVGG [20,21]. To accelerate information flow, we removed the bottleneck layer residual structures from the C2f module and replaced the original n bottleneck layers with $n-1$ 3×3 convolution layers. This modification increased the detection speed by reducing the model's memory occupation and access costs, as illustrated in Figure 4. Additionally, the NVIDIA library has been specifically optimized for 3×3 convolutions. Before the $n-1$ 3×3 convolutions, we introduced RepConv, a training-inference-decoupled convolution, to compensate for potential performance loss from removing residual structures. During training, RepConv enhances the training performance of the model by parallelly performing 3×3 convolutions, 1×1 convolutions, and identity mappings. It is effectively transformed into a single-path 3×3 convolution during the inference stage to optimize inference speed. After applying RepConv and 3×3 convolutions, we added a 1×1 convolution to produce additional feature maps that improve network performance without incurring substantial computational overhead [22]. Finally, the key component of RepELAN during inference consists of a sequence of n 3×3 convolutions, followed by a solitary 1×1 convolution. This design significantly increases the rate of information transmission and decreases the computational burden. In addition, we introduced a scaling factor in the proposed module, allowing for flexible adjustment of the model's size and complexity to suit different application scenarios. In this paper, the scaling factor is set to 0.5.

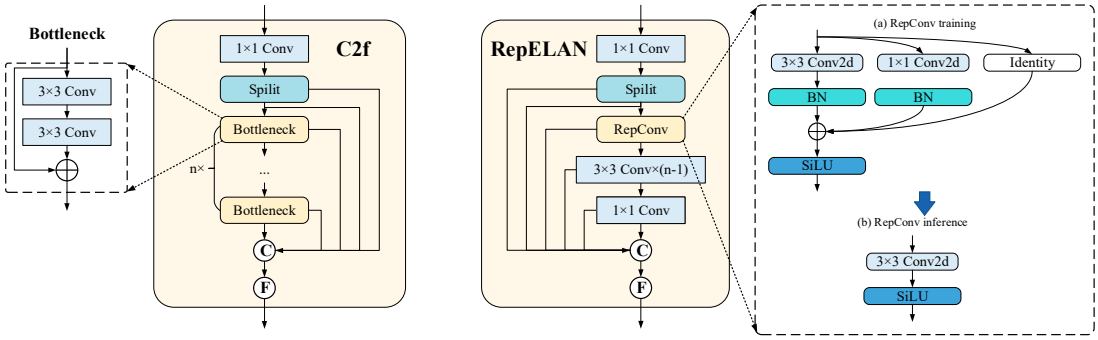


Figure 3. Detailed structural comparison of the C2f and RepELAN modules.

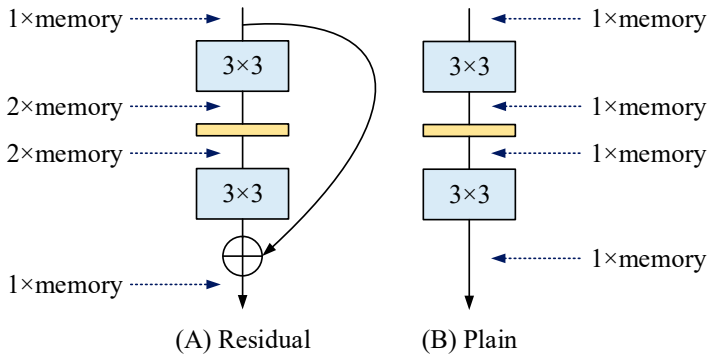


Figure 4. Comparison of the residual structure with the serial structure, where the residual structure requires the input x to be continuously stored in the memory before the addition operation, doubling the memory occupation.

Expanding upon the RepELAN foundation, we constructed the feature extraction backbone network, RLNet, as depicted in the backbone section of Figure 2. Images with dimensions of 640 × 640 × 3 are inputted into this network, generating a feature map with dimensions of 160 × 160 × 128 after two convolution operations and one feature extraction process. Subsequently, the network is divided into three stages, with each stage preceded by an Adown module that performs spatial downsampling while retaining more image information. Every stage has a RepELAN module to process and extract feature representations effectively.

2.2.2. Principle of the Proposed TFusion Module

While large-size feature maps are crucial for object detection tasks, the main approach of the RT-DETR encoder for fusion is upsampling of smaller feature maps and combining them with the features from the previous layer. In this process, the rich detail information contained in larger feature maps is ignored, consequently restricting the ability to detect abandoned objects accurately. To tackle this problem, we designed the TFusion module, which effectively handles features of different sizes (large, medium, and small) and greatly enhances the richness of detail by merging these features [23].

Figure 5 shows the structure of the TFusion module. Initially, the spatial size and number of channels of different feature maps are adjusted to align with the dimensions of the intermediate-size feature maps. Large-size feature maps are processed through a convolution module and downsampled using a combination of max and average pooling. This step helps reduce spatial dimensions and enhances translational invariance. We employed nearest neighbor interpolation to perform upsampling of small-size feature maps. This method effectively preserves local features and prevents the loss of details in small objects by utilizing adjacent pixel values. Lastly, the following equation illustrates the concatenation of feature maps of large, medium, and small sizes along the channel dimension to facilitate subsequent fusion:

$$F_{TF} = Fusion(Concat(F_l, F_m, F_s)) \tag{1}$$

where F_l, F_m, and F_s denote feature maps of large, medium, and small sizes, respectively. $Fusion$ denotes the original Fusion module. The output F_{TF} is obtained by inputting the concatenated result of F_l, F_m, and F_s into the subsequent module. The resolution of F_{TF} is equivalent to that of F_m.

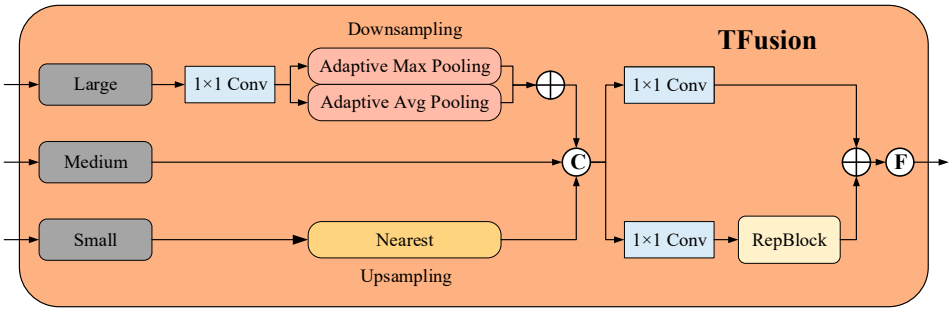


Figure 5. Detailed structure of the TFusion module.

2.2.3. Implementation Details of the Proposed CMI Module

The significant scale variations of abandoned objects complicate model recognition compared to common object detection. Accurately identifying abandoned objects depends on their visual characteristics and contextual information. To tackle the issue of significant scale variations, some researchers enhance the model's robustness to scale changes through explicit data augmentation techniques [24–26]. Moreover, other methods extract

rich scale information via multi-scale feature fusion [27,28] or pyramid architectures [29,30]. Nonetheless, a constraint of these methods is that the receptive fields for objects of varying scales remain unchanged, limiting their ability to provide adequate contextual information for larger objects. To address the challenges presented by scale variations of abandoned objects, we draw inspiration from PKINet and propose the Cross-Layer Multi-Scale Interaction (CMI) module, which is incorporated into the initial part of the encoder. The CMI module is designed to efficiently extract multi-scale features of abandoned objects across different receptive fields while minimizing the contamination of small object features by large convolution kernels and reducing substantial I/O overhead and latency [31].

The structure of the CMI module, as shown in Figure 6, adopts an inception-style design [32–34]. The CMI module merges the feature maps from the backbone network's P3, P4, and P5 layers to generate a fused feature map called x. The map undergoes parallel processing using large-kernel depth-wise (DW) convolutions of sizes 5×5, 7×7, 9×9, and 11×11. This process effectively captures spatial features of different scales and prevents excessively sparse feature representations. The identity map of feature map x is integrated with the results of the DW convolutions of varying sizes using element-wise addition. The above process is followed by adding 1×1 pointwise (PW) convolutions to merge features from various receptive fields and represent channel relationships. To avoid the problem of gradient vanishing, we combine the processed results with the identity map of feature map x using element-wise addition. This approach ensures that multi-scale features are effectively fused. In the end, the AIFI module utilizes a self-attention mechanism to extract and utilize contextual information from the rich multi-scale feature maps, thereby optimizing detection performance, as demonstrated in Formulas (2)–(4).

$$x = concat(P_3', P_4', P_5') \tag{2}$$

$$Feature = Conv_{1\times1}(x + \sum_{k\in\{5,7,9,11\}} DWConv_{k\times k}(x)) \tag{3}$$

$$M = Conv_{3\times3,stride=2}(x + Feature) \tag{4}$$

where x represents the result after processing and merging P_3', P_4', and P_5' and k denotes the list of sizes for depth convolution kernels. *Feature* is the outcome after x undergoes a series of DW convolutions, feature map additions, and PW convolutions. M is the result obtained after adding *Feature* to the feature map and resizing it.

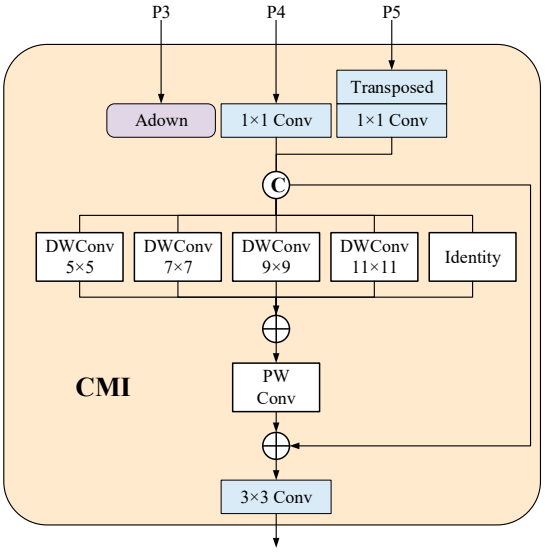


Figure 6. Detailed structure of the CMI module.

When detecting abandoned objects of different sizes, the model exhibits notable accuracy and robustness improvements attributed to the collaborative effect of the TFusion and CMI modules. These improvements demonstrate the EHMencoder's advanced capabilities and effectiveness in practical applications.

3. Experimental Results and Performance Analysis

3.1. Dataset and Implementation Details

To evaluate the effectiveness of the proposed algorithm, we extracted images containing abandoned objects from roadside surveillance videos captured along specific highway segments in Nanjing, Jiangsu Province, China, and manually annotated these images. We used these data to establish the Highway Abandoned Object Dataset (HAOD), with some sample images illustrated in Figure 7. The dataset contains a total of 4840 images of abandoned objects captured in highway daytime scenes, taken from various angles, lighting conditions, and categories, with each image having a resolution of 1920 × 1080. We consolidated all abandoned objects into one category in the dataset. Additionally, to reduce the impact of environmental factors during training and evaluation, we annotated the abandoned objects and labeled additional factors that could potentially influence the detection of abandoned objects, including vehicles, road signs, traffic cones, roadside leaves, shadows, tripods, and birds. These additional annotations enhance the model's capacity to identify abandoned objects within intricate backgrounds, improving its object detection accuracy. The dataset was divided into training, validation, and test sets in an 8:1:1 ratio.



Figure 7. Sample images from the dataset.

The experimental hardware platform for this study was an NVIDIA Tesla P100 GPU, with Ubuntu 16.04.6 LTS as the operating system. Python version 3.8.19 was used, and the LE-DETR model was built using the PyTorch framework. The AdamW optimizer was employed, with an initial learning rate set to 0.0001 and weight decay set to 0.0001. For images with an adaptive size of 640 × 640, the batch size was set to 4. Data augmentation techniques included random color distortion, expansion, cropping, flipping, and resizing, following [35]. Additionally, all mentioned models employed the train-from-scratch strategy and utilized the more persuasive COCO evaluation metrics.

3.2. Verifying That RepELAN Is More Efficient

To demonstrate the RepELAN's fastness and effectiveness, we employed a classic dimensional feature map from a real-time detection model as the input, maintaining a consistent batch size of 8. We conducted detailed testing on the module's floating-point operations, the number of parameters, and inference speed. In addition, we performed the same experiments on other state-of-the-art modules, including C3, ELAN [36], C2f, and RepNCSPELAN [10], to evaluate their performance.

The experimental results in Table 1 indicate that RepELAN has significant advantages in multiple respects. RepELAN necessitates the least floating-point operations among similar feature extraction modules, with a mere 10.7 G. RepELAN outperforms other comparative modules in terms of the number of parameters across all feature map sizes. In particular, the number of parameters in RepELAN is at least 50% lower than that of the ELAN and C2f modules across various feature map sizes. Additionally, RepELAN exhibits superiority in the speed of inference. For example, RepELAN achieves 291.3 frames per second (FPS) at a feature map size of 40 × 40 × 512, approximately 19.3% faster than the second-ranked RepNCSPELAN. In conclusion, RepELAN achieves exceptional inference speeds while maintaining the lowest computational complexity and number of parameters.

Table 1. Comparison between the RepELAN module and other state-of-the-art modules of the same type.

Module	Feature Map Size	Params (M)	FLOPs (G)	FPS
C3	320 × 320 × 64	0.019	15.1 G	90.7
	160 × 160 × 128	0.074	15.1 G	149.8
	80 × 80 × 256	0.296	15.1 G	202.6
	40 × 40 × 512	1.181	15.1 G	241.7
	20 × 20 × 1024	4.722	15.1 G	229.7
ELAN	320 × 320 × 64	0.031	25.2 G	76.7
	160 × 160 × 128	0.123	25.2 G	118.0
	80 × 80 × 256	0.492	25.2 G	153.6
	40 × 40 × 512	1.968	25.2 G	169.9
	20 × 20 × 1024	7.867	25.2 G	147.7
C2f	320 × 320 × 64	0.029	23.5 G	71.3
	160 × 160 × 128	0.115	23.5 G	107.6
	80 × 80 × 256	0.460	23.5 G	147.9
	40 × 40 × 512	1.837	23.5 G	175.6
	20 × 20 × 1024	7.343	23.5 G	155.4
RepNCSPELAN	320 × 320 × 64	0.014	11.5 G	57.4
	160 × 160 × 128	0.057	11.5 G	116.4
	80 × 80 × 256	0.226	11.5 G	185.8
	40 × 40 × 512	0.903	11.5 G	245.0
	20 × 20 × 1024	3.608	11.5 G	252.4
RepELAN (ours)	320 × 320 × 64	0.013	10.7 G	93.7
	160 × 160 × 128	0.053	10.7 G	161.8
	80 × 80 × 256	0.210	10.7 G	225.1
	40 × 40 × 512	0.837	10.7 G	291.3
	20 × 20 × 1024	3.345	10.7 G	298.4

To evaluate the RepELAN module's performance across different models, we replaced the feature extraction modules with RepELAN in various models. We compared the models' FLOPs, the number of parameters, inference speeds, and abandoned object detection accuracy before and after integrating the RepELAN module.

The computational costs of various models are significantly reduced by the RepELAN module, as illustrated in Table 2. In particular, the FLOPs of the YOLOv7 model decreased from 103.3 G to 94.7 G, while the FLOPs of the YOLOv5-N model decreased from 4.2 G to 3.4 G. These results illustrate the RepELAN module's efficacy in decreasing computational costs. The application of RepELAN significantly reduced the number of parameters in different models, including the YOLOv9-C model, which showed a 22.1% reduction. This reduction made the models more lightweight and easier to deploy. The RepELAN module significantly improved the inference speed for all models. The FPS of the YOLOv9-C

model increased by 44.6%, and the YOLOv8-N model saw an 18.8% increase. These improvements further demonstrate the potential of the RepELAN module in enhancing real-time performance. Regarding detection accuracy, models using RepELAN demonstrated stable or improved performance in their AP50 and AP50:95 metrics. For instance, the AP50:95 of the YOLOv5-N model increased from 33.3% to 35.1%, and the AP50 of the YOLOv9-C model rose from 78.7% to 81.2%.

Table 2. Comparison between the original models and those with the feature extraction modules replaced by RepELAN, where the model with RepELAN in its name is the replaced model.

Model	Epochs	Params (M)	FLOPs (G)	FPS (bs = 8)	AP50	AP50:95
YOLOv5-N	125	1.8	4.2	685.2	77.5	33.3
YOLOv5-N-RepELAN	200	1.5	3.4	740.2	77.7	35.1
YOLOv7	200	36.5	103.3	66.5	81.6	39.2
YOLOv7-RepELAN	200	34.8	94.7	67.6	83.5	39.0
YOLOv8-N	200	3.0	8.1	506.4	77.0	37.3
YOLOv8-N-RepELAN	200	2.2	6.1	601.8	77.6	37.2
YOLOv9-C	200	50.7	236.7	37.2	78.7	38.9
YOLOv9-C-RepELAN	200	39.5	202.3	53.8	81.2	39.1

Overall, replacing the original feature extraction modules with RepELAN significantly reduces the parameter count and FLOPs across all models while also improving inference speed. Moreover, the AP values for most models are either maintained or slightly improved. For example, in YOLOv8-N, replacing C2f with RepELAN results in a 24.7% reduction in FLOPs (from 8.1 G to 6.1 G), with almost no change in accuracy (AP50:95 increases from 37.3% to 37.2%). The experimental results show that the RepELAN module performs better than other state-of-the-art feature extraction modules in both model efficiency and detection performance, proving its design is effective.

3.3. Multi-Scale Kernel Design of the CMI Module

Table 3 illustrates the effect of various convolution kernel configurations on the performance of the CMI module. The results show that using the 3×3 kernel alone leads to poor performance due to its limited ability to extract texture information. The model employing the multi-scale kernel structure (5, 7, 9, and 11) attained optimal performance. Moreover, as the kernel size expands and the stride increases to 4, the parameter count and computational cost rise by 0.7% and 0.9%, respectively, whereas performance reduces by 3.8%. The findings indicate that larger convolution kernels may introduce background noise, subsequently impacting the model's performance. The kernel designs listed in Table 3 were inspired by the experimental configurations presented in [34].

Table 3. Multi-scale kernel design.

Kernel Design	Params (M)	FLOPs (G)	AP50:95
(3, 3, 3, 3)	14.4	44.7	41.6
(5, 5, 5, 5)	14.4	44.8	42.0
(7, 7, 7, 7)	14.5	44.9	41.4
(3, 5, 7, 9)	14.4	44.9	40.4
(5, 7, 9, 11)	14.5	45.0	42.6
(5, 9, 13, 17)	14.6	45.4	41.0

Next, we examined the effect of the number of convolution kernels in the CMI module, as illustrated in Table 4. The utilization of merely two convolution kernels (5×5 and 7×7)

constrained the scale information obtained by the network. As the quantity of convolution kernels increased, the network's performance was enhanced, achieving optimal results with four kernels employed. However, when the number of kernels was increased to five, performance decreased. This suggests that larger or additional convolution kernels may introduce extra background noise or redundant features, which negatively affect the model's accuracy.

Table 4. Comparison of different numbers of kernels.

Kernel Number	Params (M)	FLOPs (G)	AP50:95
2	14.4	44.8	40.9
3	14.4	44.9	41.2
4	14.5	45.0	42.6
5	14.5	45.2	41.1

3.4. Verifying the Effectiveness of the EHMencoder Module for Abandoned Object Detection

This section incrementally integrates the core components TFusion and CMI of the EHMencoder into the baseline model RT-DETR-R18 to verify their effectiveness in detecting abandoned objects. Table 5 presents the experimental results.

Table 5. Performance comparison of the efficient EHMencoder.

Model	Params (M)	FLOPs (G)	AP50:95	AP50	AP75	AP_s	AP_m	AP_l
1.RT-DETR-R18	19.9	57.0	40.0	82.5	31.7	20.1	43.3	34.9
2.RT-DETR-R18 + TFusion	19.9	58.2	40.7	84.1	33.4	21.2	43.6	36.8
3.RT-DETR-R18 + CMI	23.9	66.8	41.2	83.8	35.3	19.6	44.2	37.4
4.RT-DETR-R18 + TFusion + CMI	24.0	68.1	41.6	84.3	35.6	18.9	44.8	37.7

Compared to the baseline model RT-DETR-R18, Model 2 exhibits significant improvements in all accuracy metrics after implementing TFusion, even though there is only a minor increase in the number of parameters and FLOPs. Integrating the CMI component into Model 3 brings about an increased number of parameters and computational load. However, it improves performance on AP50:95 and AP75 by 3% and 11.4%, respectively, significantly enhancing the model's ability to detect abandoned objects. Although Model 4, which incorporates both TFusion and CMI, has the highest computational cost, it achieves further enhancements in all AP metrics, except for a slight decrease in AP_s. These results verify the EHMencoder's efficiency in detecting multi-scale abandoned objects on highways.

3.5. Ablation Study

This subsection incrementally introduces the RLNet based on RepELAN, the CMI module, and the TFusion module to explore the contribution of each component to the performance of abandoned object detection. The experimental results, as shown in Table 6, clearly demonstrate the performance enhancements contributed by each module.

Initially, the model experienced a 47.7% and 41.9% reduction in the number of parameters and computational load, respectively, a 16% increase in FPS, and a 2% increase in both AP50:95 and AP50 metrics by replacing the backbone network of RT-DETR-R18 with the RepELAN-based RLNet. This significant performance enhancement is primarily attributed to the simple and efficient structure of RepELAN, which effectively extracts crucial information during training and employs fewer residuals and a faster cascade architecture during inference, thereby optimizing overall model performance. Further, introducing the TFusion module in the encoder improves the model's detailed information, resulting in a 3.5% and

8.2% increase in AP50:95 and AP75, respectively. Subsequently, integrating the CMI module at the beginning of the encoder resulted in a 2.9% increase in AP50:95, demonstrating the necessity of multi-scale feature interactions in complex highway scenarios. This module employs a variety of sizes of DW convolution to expand multi-scale feature representations under different receptive fields. Finally, the collaborative effects of these modules enable the LE-DETR model to achieve real-time capabilities while reducing the number of parameters and computational load by 27.1% and 21.1%, respectively, compared to the baseline model. The model improves detection accuracy across various scales, with increases of 6.5% in AP50:95, 2.2% in AP50, and 18% in AP75.

Table 6. Results of ablation experiments; √ means that this component has been applied. Our baseline is indicated by Bas.

Bas.	RLNet	TFusion	CMI	Params (M)	FLOPs (G)	FPS (bs = 8)	AP 50:95	AP50	AP75	AP_s	AP_m	AP_l
√				19.9	57.0	86.2	40.0	82.5	31.7	20.1	43.3	34.9
	√			10.4	33.1	100	40.8	84.2	32.9	21.6	44.1	35.4
	√	√		10.4	35.2	91.7	41.4	84.2	34.3	21.4	44.5	36.9
	√	√	√	14.5	45.0	79.3	42.6	84.3	37.4	21.3	45.7	38.6

3.6. Comparison to State-of-the-Art Models

To evaluate the proposed method's performance advantages, we conducted experiments using the LE-DETR model on the HAOD and compared its performance to that of Faster R-CNN [37], DETR [38], Deformable-DETR [39], DINO-Deformable-DETR [40], YOLOv5, YOLOv7, YOLOv8, YOLOv9, and RT-DETR.

The experimental results in Table 7 demonstrate that LE-DETR attained a comprehensive accuracy (AP50:95) of 42.6, surpassing all other models in the comparison and showcasing its superior detection performance. Specifically, the AP50 of LE-DETR was 84.3, which was higher than the 4.6 and 3.4 percentage points of the end-to-end detection model DINO-Deformable-DETR-R50 and the classical two-stage model Faster R-CNN-R50, respectively. Furthermore, LE-DETR displayed balanced and exceptional performance in detecting abandoned objects of various sizes, particularly excelling in medium-sized object detection by outperforming the suboptimal model by 5.5%. Notably, LE-DETR, a model based on the DETR architecture, not only guarantees real-time capabilities but also exhibits a higher level of detection accuracy than the YOLO series of real-time detection models. These experiments validate the effectiveness and superiority of LE-DETR in abandoned object detection. Its efficient performance and economical computational costs render it more suitable for practical applications and deployment.

Table 7. Comparison to state-of-the-art models. The inference speeds of the Faster R-CNN-R50, DETR-R18, Deformable-DETR-R50, and DINO-Deformable-DETR-R50 were not tested because they are not real-time detectors.

Model	Epochs	Params (M)	FLOPs (G)	FPS (bs = 8)	AP50:95	AP50	AP75	AP_s	AP_m	AP_l
Faster R-CNN-R50	12	41.294	202	-	38.0	82.9	27.4	20.6	41.0	33.4
DETR-R18	500	28.832	48.372	-	31.4	77.4	18.2	9.0	34.6	29.0
Deformable-DETR-R50	50	40.101	187	-	38.1	82.9	27.4	22.3	40.5	35.4
DINO-Deformable-DETR-R50	12	47.559	266	-	39.2	82.9	30.5	18.1	42.8	32.9
RT-DETR-R18	150	20.0	57.3	86.2	40.0	82.5	31.7	20.1	43.3	34.9

Table 7. *Cont.*

Model	Epochs	Params (M)	FLOPs (G)	FPS (bs = 8)	AP50:95	AP50	AP75	AP_s	AP_m	AP_l
YOLOv5-M	112	21.2	49.0	133.3	37.5	80.3	28.2	25.5	41.0	33.9
YOLOv7	200	36.9	104.7	75.8	39.2	81.6	30.5	15.7	42.1	37.4
YOLOv8-S	188	11.2	28.6	169.5	38.5	79.1	34.5	15.0	41.2	35.3
YOLOv8-M	152	25.9	78.9	101	37.5	77.6	31.4	15.9	40.7	34.1
YOLOv9-C	200	25.3	102.1	37.7	38.9	78.7	29.4	18.2	42.1	35.4
LE-DETR	150	14.5	45.0	79.3	42.6	84.3	37.4	21.3	45.7	38.6

3.7. Visualization and Analysis

The detection performance of RT-DETR-R18 and LE-DETR is compared in Figure 8 under various highway environments. The figure demonstrates that LE-DETR outperforms RT-DETR-R18 in terms of detection performance. It is evident from the first row of images that LE-DETR exhibits a higher level of detection confidence than RT-DETR-R18 in the presence of external disturbances. The second row demonstrates that LE-DETR accurately identifies abandoned objects, even when they are obscured by vehicles and in extreme weather conditions, whereas RT-DETR-R18 cannot detect them. In the third row, due to the limited feature extraction capability of RT-DETR-R18, its prediction boxes overlap in sparse regions, whereas LE-DETR is able to more accurately identify sparse areas. In the fourth row, the objectives are not recognized by RT-DETR-R18 for smaller-sized abandoned objects, whereas LE-DETR successfully identifies them. Overall, LE-DETR demonstrates exceptional detection capabilities for abandoned objects of different scales in various complex environments, making it highly suitable for highway abandoned object detection tasks.

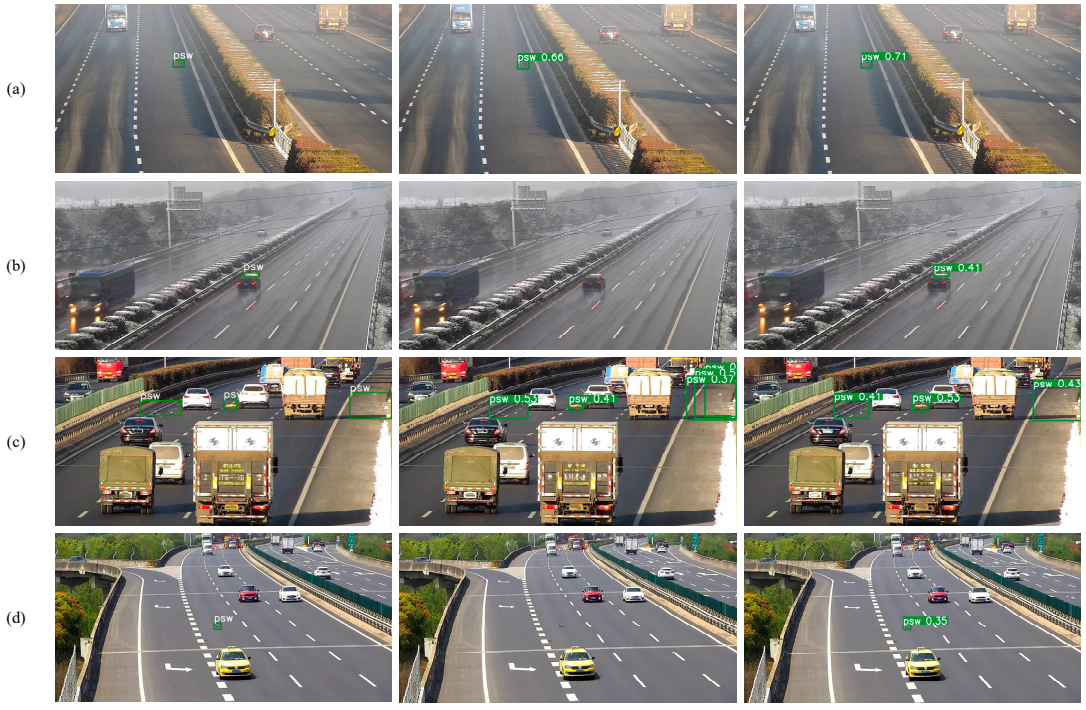


Figure 8. Visualization of detection results compared to the baseline model. In the figure, psw denotes abandoned objects on the highway. (**a**) Detection in the presence of external disturbances; (**b**) Detection of abandoned objects obscured by vehicles and in extreme weather conditions; (**c**) Detection in sparse regions; (**d**) Detection of smaller-sized abandoned objects.

4. Conclusions

This study focuses on the existing challenges with abandoned object detection technology, including computational resource consumption, scale variability of abandoned objects, and interference from complex backgrounds. We propose LE-DETR, a lightweight end-to-end model for detecting abandoned objects on highways. Initially, we designed the real-time feature extraction module RepELAN, which surpasses comparable modules regarding the number of parameters, floating-point operation times, inference speed, and detection accuracy. Building upon the RepELAN module's success, we constructed the RLNet backbone network, effectively reducing computational load and improving parameter utilization. Additionally, we designed the TFusion module, which effectively enhances the representation of detailed features by combining high-level semantic information with low-level spatial information. Subsequently, we proposed the CMI module, which improves the multi-scale feature representation of abandoned objects by utilizing large kernels of varying sizes in depth-wise convolution to extract features from different receptive fields in parallel. The EHMencoder was proposed by combining the TFusion and CMI modules, which brought about a significant improvement in the accuracy of detection for multi-scale abandoned objects. Ultimately, we constructed a novel dataset of abandoned highway objects and conducted experiments. The LE-DETR model is more appropriate for resource-constrained devices while maintaining real-time performance, as evidenced by comprehensive results that indicate a 6.5% and 2.2% improvement in AP50:95 and AP50, respectively, and a 27.1% and 21.1% reduction in the number of parameters and FLOPs compared to baseline models. LE-DETR's ability to effectively detect abandoned objects on highways under various complicated circumstances was also demonstrated in the experiments. Although the model proposed in this paper effectively improves the detection efficiency of abandoned objects on highways, it has certain limitations. The dataset primarily includes daytime highway scenes, while nighttime and extreme weather conditions have not been sufficiently considered. The model is only able to identify a portion of abandoned object types that are present in real-world scenarios as a result of the dataset's restricted size. Consequently, future research should concentrate on improving the algorithm's generality to identify various abandoned objects in various scenarios.

Author Contributions: Conceptualization, B.Z. and C.W.; Methodology, B.Z.; Validation, B.Z.; Formal analysis, B.Z. and C.W.; Writing—original draft, B.Z.; Writing—review and editing, C.W. and J.W.; Investigation, C.W. and J.W.; Supervision, J.W. All authors have read and agreed to the published version of the manuscript.

Funding: This work is partly supported by the Postgraduate Research & Practice Innovation Program of Jiangsu Province (SJCX24_1327).

Data Availability Statement: The data presented in this study are available upon request from the corresponding author due to privacy concerns.

Acknowledgments: The authors wish to thank the Jinling Institute of Technology (JIT) for providing computational and storage resources.

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

References

1. Fu, H.; Xiang, M.; Ma, H.; Ming, A.; Liu, L. Abandoned object detection in highway scene. In Proceedings of the 2011 6th International Conference on Pervasive Computing and Applications, Port Elizabeth, South Africa, 26–28 October 2011; pp. 117–121. [CrossRef]
2. Zivkovic, Z. Improved adaptive Gaussian mixture model for background subtraction. In Proceedings of the 17th International Conference on Pattern Recognition, 2004, ICPR 2004, Cambridge, UK, 26–26 August 2004; Volume 2, pp. 28–31. [CrossRef]

3. Szwoch, G. Extraction of stable foreground image regions for unattended luggage detection. *Multimed. Tools Appl.* **2016**, *75*, 761–786. [CrossRef]
4. Zeng, Y.; Lan, J.; Ran, B.; Gao, J.; Zou, J. A Novel Abandoned Object Detection System Based on Three-Dimensional Image Information. *Sensors* **2015**, *15*, 6885. [CrossRef]
5. Lin, C.-Y.; Muchtar, K.; Yeh, C.-H. Robust techniques for abandoned and removed object detection based on Markov random field. *J. Vis. Commun. Image Represent.* **2016**, *39*, 181–195. [CrossRef]
6. Park, H.; Park, S.; Joo, Y. Detection of Abandoned and Stolen Objects Based on Dual Background Model and Mask R-CNN. *IEEE Access* **2020**, *8*, 80010–80019. [CrossRef]
7. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. *arXiv* **2018**, arXiv:1703.06870. [CrossRef]
8. Song, H.; Wang, J.; Zhang, Y. Detection of abandoned objects based on Yolov9 and background differencing. *Signal Image Video Process.* **2024**, *19*, 54. [CrossRef]
9. Zivkovic, Z.; van der Heijden, F. Efficient adaptive density estimation per image pixel for the task of background subtraction. *Pattern Recognit. Lett.* **2006**, *27*, 773–780. [CrossRef]
10. Wang, C.-Y.; Yeh, I.-H.; Liao, H.-Y.M. YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information. *arXiv* **2024**, arXiv:2402.13616. [CrossRef]
11. Qian, P.; Yuan, K.; Yao, J.; Fan, C.; Zhang, H.; Liu, Y.; Lu, X. Residual-Network-Leveraged Vehicle-Thrown-Waste Identification in Real-Time Traffic Surveillance Videos. *IEEE Trans. Intell. Transp. Syst.* **2021**, *22*, 1817–1826. [CrossRef]
12. Uijlings, J.R.R.; Van De Sande, K.E.A.; Gevers, T.; Smeulders, A.W.M. Selective Search for Object Recognition. *Int. J. Comput. Vis.* **2013**, *104*, 154–171. [CrossRef]
13. van de Sande, K.E.A.; Uijlings, J.R.R.; Gevers, T.; Smeulders, A.W.M. Segmentation as selective search for object recognition. In Proceedings of the 2011 International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; pp. 1879–1886. [CrossRef]
14. Qiu, S.; Wen, G.; Deng, Z.; Liu, J.; Fan, Y. Accurate non-maximum suppression for object detection in high-resolution remote sensing images. *Remote Sens. Lett.* **2018**, *9*, 237–246. [CrossRef]
15. Neubeck, A.; Van Gool, L. Efficient Non-Maximum Suppression. In Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06), Hong Kong, China, 20–24 August 2006; pp. 850–855. [CrossRef]
16. Liu, W.; Li, J.; Liu, Y.; Ma, C.; Xu, J.; Liu, C. Research on the lightweight detection network of abandoned objects in freeway based on video. In Proceedings of the Sixth International Conference on Traffic Engineering and Transportation System (ICTETS 2022), Guangzhou, China, 23–25 September 2023; SPIE: Bellingham, WA, USA; pp. 269–274. [CrossRef]
17. Zhou, L.; Xu, J. Enhanced Abandoned Object Detection through Adaptive Dual-Background Modeling and SAO-YOLO Integration. *Sensors* **2024**, *24*, 6572. [CrossRef] [PubMed]
18. Zhao, Y.; Lv, W.; Xu, S.; Wei, J.; Wang, G.; Dang, Q.; Liu, Y.; Chen, J. DETRs Beat YOLOs on Real-time Object Detection. In Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 17–21 June 2024; pp. 16965–16974. [CrossRef]
19. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. *arXiv* **2023**, arXiv:1706.03762. [CrossRef]
20. Ding, X.; Zhang, X.; Ma, N.; Han, J.; Ding, G.; Sun, J. RepVGG: Making VGG-style ConvNets Great Again. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 19–25 June 2021; pp. 13728–13737. [CrossRef]
21. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2015**, arXiv:1409.1556.
22. Han, K.; Wang, Y.; Tian, Q.; Guo, J.; Xu, C.; Xu, C. GhostNet: More Features From Cheap Operations. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 1577–1586. [CrossRef]
23. Kang, M.; Ting, C.-M.; Ting, F.F.; Phan, R.C.-W. ASF-YOLO: A novel YOLO model with attentional scale sequence fusion for cell instance segmentation. *Image Vis. Comput.* **2024**, *147*, 105057. [CrossRef]
24. Chen, Y.; Zhang, P.; Li, Z.; Li, Y.; Zhang, X.; Qi, L.; Sun, J.; Jia, J. Dynamic Scale Training for Object Detection. *arXiv* **2021**, arXiv:2004.12432. [CrossRef]
25. Shamsolmoali, P.; Zareapoor, M.; Chanussot, J.; Zhou, H.; Yang, J. Rotation Equivariant Feature Image Pyramid Network for Object Detection in Optical Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–14. [CrossRef]
26. Zhou, Y.; Yang, X.; Zhang, G.; Wang, J.; Liu, Y.; Hou, L.; Jiang, X.; Liu, X.; Yan, J.; Lyu, C.; et al. MMRotate: A Rotated Object Detection Benchmark using PyTorch. In Proceedings of the 30th ACM International Conference on Multimedia, in MM '22, Lisbon, Portugal, 10–14 October 2022; Association for Computing Machinery: New York, NY, USA; pp. 7331–7334. [CrossRef]
27. Zhang, Y.; Yuan, Y.; Feng, Y.; Lu, X. Hierarchical and Robust Convolutional Neural Network for Very High-Resolution Remote Sensing Object Detection. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 5535–5548. [CrossRef]

28. Lin, Z.; Ji, K.; Leng, X.; Kuang, G. Squeeze and Excitation Rank Faster R-CNN for Ship Detection in SAR Images. *IEEE Geosci. Remote Sens. Lett.* **2019**, *16*, 751–755. [CrossRef]
29. Liang, X.; Zhang, J.; Zhuo, L.; Li, Y.; Tian, Q. Small Object Detection in Unmanned Aerial Vehicle Images Using Feature Fusion and Scaling-Based Single Shot Detector With Spatial Context Analysis. *IEEE Trans. Circuits Syst. Video Technol.* **2020**, *30*, 1758–1770. [CrossRef]
30. Wang, G.; Zhuang, Y.; Chen, H.; Liu, X.; Zhang, T.; Li, L.; Dong, S.; Sang, Q. FSoD-Net: Full-Scale Object Detection From Optical Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5602918. [CrossRef]
31. Wang, A.; Chen, H.; Liu, L.; Chen, K.; Lin, Z.; Han, J. Yolov10: Real-time end-to-end object detection. *arXiv* **2024**, arXiv:2405.14458.
32. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception Architecture for Computer Vision. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826. [CrossRef]
33. Yu, W.; Zhou, P.; Yan, S.; Wang, X. InceptionNeXt: When Inception Meets ConvNeXt. In Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 17–21 June 2024; pp. 5672–5683. [CrossRef]
34. Cai, X.; Lai, Q.; Wang, Y.; Wang, W.; Sun, Z.; Yao, Y. Poly Kernel Inception Network for Remote Sensing Detection. In Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 17–21 June 2024; pp. 27706–27716. [CrossRef]
35. Xu, S.; Wang, X.; Lv, W.; Chang, Q.; Cui, C.; Deng, K.; Wang, G.; Dang, Q.; Wei, S.; Du, Y.; et al. PP-YOLOE: An evolved version of YOLO. *arXiv* **2022**, arXiv:2203.16250. [CrossRef]
36. Wang, C.-Y.; Bochkovskiy, A.; Liao, H.-Y.M. YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023; pp. 7464–7475. [CrossRef]
37. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef] [PubMed]
38. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-End Object Detection with Transformers. In *Computer Vision—ECCV 2020*; Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M., Eds.; Springer International Publishing: Cham, Switzerland, 2020; pp. 213–229. [CrossRef]
39. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable DETR: Deformable Transformers for End-to-End Object Detection. *arXiv* **2021**, arXiv:2010.04159. [CrossRef]
40. Zhang, H.; Li, F.; Liu, S.; Zhang, L.; Su, H.; Zhu, J.; Ni, L.M.; Shum, H.-Y. DINO: DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection. *arXiv* **2022**, arXiv:2203.03605. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.