IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 6 November 2024, accepted 28 November 2024, date of publication 4 December 2024, date of current version 16 December 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3510746

TOPICAL REVIEW

A Review of Intrusion Detection for Railway Perimeter Using Deep Learning-Based Methods

JIN WANG[1,2], (Member, IEEE), HONGYANG ZHAI[2], YANG YANG[1], NIUQI XU[2], HAO LI[2], AND DI FU[2]

[1]Beijing Key Laboratory of Traffic Engineering, Beijing University of Technology, Beijing 100124, China
[2]Beijing Engineering Research Center of Urban Transport Operation Guarantee, Beijing University of Technology, Beijing 100124, China

Corresponding author: Yang Yang (yyangscholar@bjut.edu.cn)

This work was supported in part by Beijing Natural Science Foundation under Grant L221026 and Grant 8232005, and in part by the National Natural Science Foundation of China under Grant 41801380.

ABSTRACT Efficiently detecting intrusions on a railway perimeter is crucial for ensuring the safety of railway transportation. With the development of computer vision, researchers have been actively exploring methods for detecting foreign object intrusion via image recognition technology. This article reviews the background and importance of detecting railway perimeter intrusion, summarizes the limitations of traditional detection methods, and emphasizes the potential of improving detection accuracy and efficiency in image recognition with deep learning models. Further, it introduces the development of deep learning in image recognition, focusing on the principles and progress of key technologies such as convolutional neural networks (CNNs) and vision transformers (ViTs). In addition, the application status of semantic segmentation and object detection algorithms based on deep learning in detecting railway perimeter intrusion is explored, including the classification, principles, and performance of the algorithms in practical applications. Finally, it highlights the primary challenges faced in railway perimeter intrusion detection and projects future research directions to resolve these challenges, including multisource data fusion, large-scale dataset construction, model compression, and end-to-end multitask learning networks. These studies support the accuracy and real-time detection of railway perimeter intrusion, and provide technical guarantees for railway transportation monitoring tasks.

INDEX TERMS Railway, semantic segmentation, object detection, foreign object intrusion, railway safety.

I. INTRODUCTION

The railway is a pivotal infrastructure that fulfills the extensive requirements for passenger mobility and freight transportation. With the continuous expansion of railway networks, railway perimeter intrusion has become a major threat. The railway perimeter [1] is the boundary of the railway line area that needs to be physically or electronically protected (Figure 1). Railway perimeter intrusion [2] involves the entry of unauthorized individuals, animals, vehicles, or other foreign objects (such as falling rocks) into the railway perimeter (Table 1). According to [3], more than half of railway-related fatalities are due to such intrusions in countries such as the United States, Australia, Belgium, and Croatia. These intrusions not only cause significant train delays and operational inefficiencies, but also result in substantial human casualties and property damage. Therefore, railway perimeter intrusion detection (RPID) is crucial for ensuring railway transportation safety and efficiency.

The methods for RPID can be divided into contact and non-contact approaches [4]. Table 2 summarizes the pros and cons of each approach. The contact approaches rely on direct physical interaction between the sensor and an intruding object. Popular techniques include electronic fences and vibrating fiber optics installed on fences/walls along railways [5]. They have a low probability of false positive rates; however, they have a limited detection range and cannot determine the size and type of incursions. Non-contact approaches mainly

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

TABLE 1. Overview of railway intrusion types [1], [2].

Intrusion types	Intrusion reasons	Risk level
Human	Taking shortcuts, suicide attempts, theft, vandalism, etc.	High. Certain behaviors can critically endanger railway safety.
Vehicles	Accidents at level crossings or road-rail junctions.	High. Slow response may lead to train collisions.
Animals	Searching for food or habitat in railway areas.	Low.
Natural factors.	Rockfalls, landslides, and mudslides in mountainous regions	Extremely High. Collisions with trains can have catastrophic outcomes.
Others	Falling cargo, debris from railway structures, or construction sites.	Low impact because of small size and limited influence.

TABLE 2. PROS/CONS of various railway intrusion detection methods [1].

Methods	Pros	Cons
	Contact methods	
Vibrating fiber detector	Low false negative rate. Strong adaptability to the environment.	High false alarm rate. Poor positioning accuracy.
Electronic fences	Applicable to walls, sensitive	Inconvenient, easy misjudgment
	Non-contact methods	
LiDAR detector	High-ranging accuracy. Not affected by light.	Fog weather recognition performance decreases. Costs are higher.
Radar detector	Strong adaptability to the environment. Long detection distance for moving objects.	Point clouds are sparse and data processing is difficult.
Infrared detector	Good nighttime detection performance. Large viewing range.	Poor low-temperature differential detection. Unable to identify detailed features.
Video surveillance	Strong recognition ability. Easy to deploy and maintain.	Poor generalization capabilities for intrusion targets. Poor adaptability to the weather.

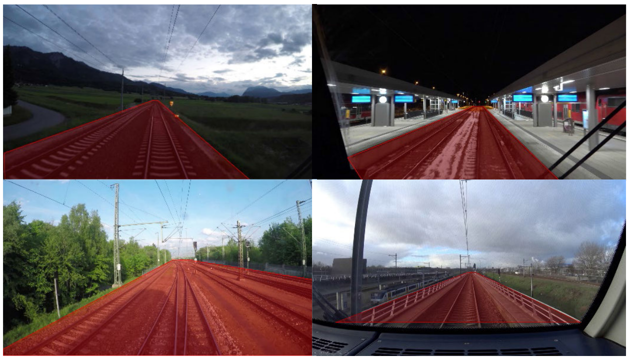

FIGURE 1. Typical railway perimeter scenarios.

include video surveillance, radar, light detection and ranging (LiDAR), and infrared barrier detection [6]. These methods are characterized by massive amounts of data and extensive coverage; however, data processing efficiency and detection accuracy are key concerns in applications.

Images from video surveillance are widely utilized in RPID, urban monitoring, and traffic monitoring because of their excellent visibility and cost-effectiveness. However, traditional video processing methods, such as the frame difference and optical flow methods, rely on frame-by-frame analysis and manual intervention, making them time-consuming. As the video data volume increases, there is a growing demand for higher levels of automation in video processing. Deep learning methods provide promising solutions for learning patterns from diverse image datasets. The trained models can automatically detect the locations and types of intrusions in images. In recent years, deep learning methods have been applied in urban monitoring [7], [8] and traffic monitoring scenarios [10], [11], [12]. Compared with urban and traffic monitoring (Table 3), RPID involves additional challenges in enhancing resilience under harsh weather conditions, improving the detection ability of small targets, and real-time processing capabilities. In response to these challenges, studies have focused on improving deep learning networks to RPID. However, a systematic review of these research efforts remains limited. To address this gap, this paper reviews the deep learning methods in RPID and summarizes the challenges.

TABLE 3. Comparisons among RPID, URBAN, and TRAFFIC monitoring systems.

Viewpoint	RPID	Urban Monitoring	Traffic Monitoring
Objective	Detect unauthorized intrusions near railways to prevent accidents and system damage.	Monitor human activities such as crime or unusual behavior.	Monitor vehicle and pedestrian behavior to prevent traffic accidents.
Environmental Complexity	Dynamic railway environments with trains, vegetation, and varying weather.	Complex urban environments with dense human activity, vehicles, and buildings.	Complex road networks with various vehicle types and pedestrian interactions.
Real-Time Requirements	High urgency to prevent collisions or sabotage.	High, focused on crime prevention and rapid response to incidents.	Real-time monitoring to prevent accidents and manage traffic flow.
Context of Intrusions	Intruders can be humans, animals, or debris, presenting diverse scenarios.	Focus on complex human behavior such as crimes or suspicious activities.	Primarily focused on vehicle-related anomalies such as accidents or traffic violations.
Risk and Consequences	High risk, as intrusions can lead to fatal accidents or railway damage.	Lower risk, mostly property damage or injuries.	Significant risk of accidents, but response time is generally longer compared to RPID.
Automation Level	Highly automated, often deployed in remote areas without human oversight.	Typically, supplemented by human operators for complex threat assessments.	Highly automated, but human intervention may be required in certain cases.
Detection Complexity	Challenging because of environmental factors such as trains, vegetation, and weather.	Complex because of the need to monitor various human behaviors and activities.	Requires detection of various traffic entities and adherence to traffic rules.

The remainder of this paper is organized as follows: Section II provides an overview of semantic segmentation and object detection algorithms utilized in image processing. Section II-B1 examines the applications of these algorithms for RPID. Finally, Section III concludes the paper with a summary and recommendations for further research.

II. CURRENT STATE OF DEEP LEARNING NETWORKS IN IMAGE PROCESSING

This section introduces the development of deep learning networks, and reviews semantic segmentation and object detection networks. These methods are pivotal for the accurate detection of railway perimeter intrusions.

A. THE BIRTH, DECLINE, AND RESURGENCE OF DEEP LEARNING

Deep learning focuses on the autonomous extraction and learning of complex data features. It employs multi-layered neural network models to enhance classification, regression, and clustering tasks. Originating in the 1940s [13], deep learning models were initially designed to emulate the neural system of the human brain via artificial neural networks [14]. With the progress of computer technology, the backpropagation algorithm [15] has facilitated distributed representations of neural networks. However, by the mid-1990s, limited by hardware, artificial neural networks were difficult to train and performed ineffectively. Moreover, classic machine learning algorithms, such as support vector machines [16], began to take the lead owing to their short training time and interpretability.

Deep learning was introduced in 2006. The deep belief network [17] addresses the "vanishing gradient" issue through pre-training and subsequent fine-tuning. Using its innovative eight-layer convolutional neural network (CNN) architecture, AlexNet [18] won the ImageNet Large-Scale Visual Recognition Challenge. The classification accuracy and robustness efficiency achieved by AlexNet were superior to those of conventional techniques. AlexNet has marked the rise of deep learning networks and has demonstrated its advantages in various applications, such as intelligent transportation, facial recognition, medical image analysis, and natural language processing [19].

B. OVERVIEW OF CNNS AND VISION TRANSFORMER

1) CONVOLUTIONAL NEURAL NETWORK

A CNN is a classic architecture of image-based deep learning networks. It effectively captures detailed spatial features and optimizes weight parameters through data training. A standard CNN architecture (Figure 2) typically includes convolutional layers, pooling layers, activation functions, and fully connected layers. The convolutional layers extract local features via a series of learnable filters. Feature maps are generated by sliding each filter over the input image. Pooling layers, such as max pooling, reduce the spatial dimensions of feature maps. Activation functions such as ReLU allow the learning of more complex features and address vanishing gradient problems. The fully connected layers map the extracted high-level features to the output, such as the target categories in image classification tasks.

On the basis of the aforementioned CNN architecture, several classic deep learning networks are summarized as follows (Table 4):

- AlexNet [18]: This method achieved success in the ImageNet competition by introducing deep hierarchies and ReLU activation functions.

TABLE 4. Summary of typical backbones based on CNN.

Network	Year	Conclusion
LeNet [41]	1998	A five-layer neural network with a straightforward architecture, it is one of the earliest convolutional neural networks.
AlexNet [18]	2012	A five-layer convolutional and three-layer fully connected network, the first champion of CNN in the ImageNet competition.
VGGNet [20]	2014	Replaces 5×5 or 7×7 convolutional kernels with smaller 3×3 kernels.
InceptionNet [22]	2014	Uses multiple convolutional kernels of varying sizes to extract features at different levels.
ResNet [23]	2016	Introduces residual connections to address the gradient vanishing and explosion that occur in extremely deep neural networks.
SqueezeNet [32]	2016	Lightweight CNN that reduces the number of channels using 1×1 convolutional kernels and eliminates the fully connected layer, considerably reducing the model's size.
MobileNet [31]	2017	Lightweight CNN that uses depthwise separable convolution to reduce the number of parameters and computational complexity.
ResNeXt [42]	2017	Improves the model's accuracy and efficiency by connecting multiple small convolutional kernels in parallel.
ShuffleNet [43]	2018	Uses grouped convolutions and channel shuffling to reduce the model's parameter count and computational load effectively.
EfficientNet [44]	2019	Scales uniformly across different network dimensions (depth, width, resolution) using compound scaling parameters to achieve better performance.
GhostNet [45]	2020	Lightweight CNN that applies low-cost operations on basic feature maps to generate additional feature maps, effectively increasing the network's receptive field and expressive power without considerably increasing the computational burden.
VAN [46]	2023	Proposes a novel linear attention mechanism Large Kernel Attention (LKA), which breaks down large kernel convolution into depthwise, depthwise dilated, and 1×1 convolution. It effectively enhances the model's accuracy and reduces computational requirements by combining the strengths of CNNs and Transformers.
FasterNet [47]	2023	Lightweight CNN that introduces a novel partial convolution, which more efficiently extracts spatial features by simultaneously reducing redundant computations and memory access.

- VGGNet [21]: VGGNet simplifies the network architecture and enhances the representational power by adopting uniform 3 × 3 convolutional kernels.
- InceptionNet [22]: By designing parallel convolutional operations, this network strengthens the ability to learn and capture multi-scale features.
- ResNet [23]: By designing residual connections, it addresses the vanishing gradient issue during network training, and allows a deep network architecture.

On the basis of these classic networks, current research has focused on designing deeper/wider structures [24], attention mechanisms [25], refined activation functions [26], and regularization algorithms [27]. Additionally, methods such as model compression [28], knowledge distillation [29], and pruning [30] have been explored to increase training efficiency and model applicability.

The training process of the CNN basically includes forward propagation, loss-function calculation, backpropagation, and parameter updating. In forward propagation, the input image is transformed to feature maps through convolutional layers, pooling layers, and activation functions. The loss function evaluates the discrepancy between the network's output and the ground truth. The backpropagation operation calculates the gradients of each parameter. Finally, the network parameters are updated by optimizing the minimum of the loss function.

However, CNNs also have limitations, such as high computational resource requirements and sensitivity to the input size. Thus, to address these issues, lightweight CNN architectures, such as MobileNet [31] and SqueezeNet [32], are proposed to reduce the model size and computational demands.

2) VISION TRANSFORMER

The transformer model, which was originally designed for natural language processing, uses a self-attention mechanism

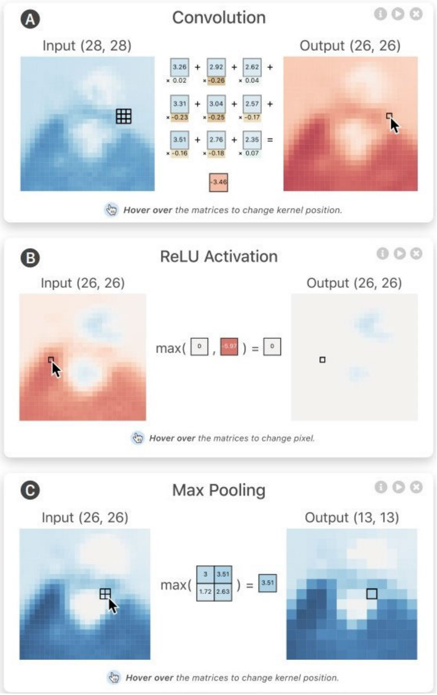

FIGURE 2. **Visualization of convolution [20].**

that effectively handles image data. This mechanism enables the model to establish connections directly between different positions in a sequence. By dividing the image into 16 × 16 small patches, the popular Vision Transformer (ViT) [33] uses a self-attention mechanism and eliminates the network's reliance on local features. Then, a series of improved networks, such as the Swin-transformer [34] and CSwin-transformer [35], have subsequently emerged. They incorporate mechanisms of local attention, hybrid local-global features, and shifted windows between consecutive self-attention layers. Thus, they enhance the learning ability of long-range dependencies while retaining local detail information. In total, both CNNs and ViTs are important structures in semantic segmentation and object detection.

C. DEEP LEARNING-BASED SEMANTIC SEGMENTATION ALGORITHMS

Semantic segmentation assigns category labels to each pixel in an image. Classic semantic segmentation methods, such as threshold-based segmentation [36] and edge detection [37], require human intervention. These methods are more scenario-specific and lack robustness. Deep learning networks of semantic segmentation networks can be grouped as follows (Table 5): encoder-decoder networks, dilated convolutions, bilateral structures, and transformer-based segmentation networks.

TABLE 5. **Summary of semantic segmentation networks with different structures.**

Structures	Instances
Encoder-decoder based model	FCN, U-Net, SegNet, *etc.*
Dilated convolutional based model	DeepLabV1, DeepLabV2, DeepLabV3, DeepLabV3+, *etc.*
Dual-branch structure	BiseNetV1, Fast-SCNN, STDC, BiseNet V2, DDRNet, *etc.*
Transformer based segmentation	SETR, Segmenter, Topformer, Seaformer, DeMT, MobileViTV1, MobileViTV3, *etc.*

1) ENCODER-DECODER BASED MODEL

By replacing the fully connected layers in a classic CNN with convolutional layers, an FCN allows the input of images of any size and generates corresponding dense pixel-level predictions [38]. It provides a feasible solution for end-to-end training in semantic segmentation tasks. Following this development, a series of FCN-based networks, such as U-net [39] and SegNet [40], have been proposed to improve the capabilities of segmentation. This type of structure achieves pixel-level predictions by extracting features through the encoder component. The features are then mapped back to the size of the raw image (Figure 3).

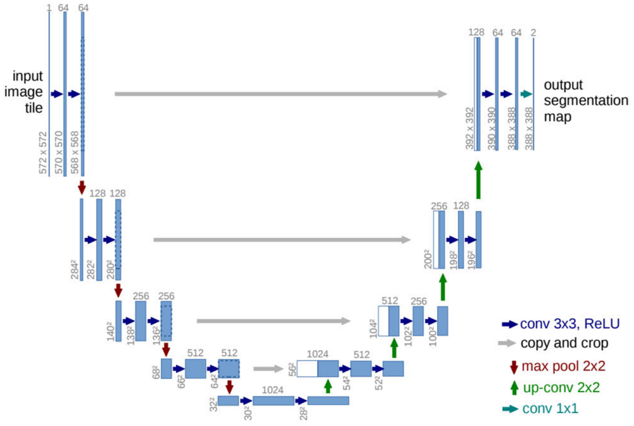

FIGURE 3. **The structure of U-net [39].**

The encoder-decoder structure is characterized by its straightforward design. It facilitates flexible adjustments and expansions to accommodate various tasks in both the encoder and decoder stages. However, the decoder component incurs a computational load because of the consecutive transposed convolutions in the up-sampling stage. This limits its applicability in scenarios that require real-time segmentation.

2) DEEPLAB SERIES OF NETWORKS BASED ON DILATED CONVOLUTIONS

The DeepLab series proposed by Google is a semantic segmentation method based on deep convolutional networks. The original DeepLab [48] introduced a dilated convolution [49] to replace traditional convolution operations. By incorporating a dilation rate into the standard convolution operation, the dilated convolution broadens the receptive field of the kernel without a proportional increase in the number of parameters. Figure 4 shows the receptive field for the dilated convolution at S = 1, 2, and 3 dilation rates. DeepLabV2 [50] subsequently introduced Conditional Random Fields to refine segmentation boundaries. DeepLabV3 [51] adopted multi-scale feature fusion, and DeepLabV3+ [52] optimized the Xception backbone network.

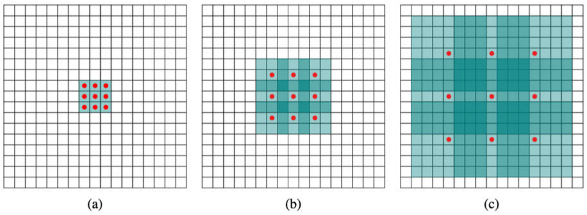

FIGURE 4. Visual representation of the receptive field for dilated convolution at S = 1, 2, and 3 dilation rates [49].

The training time of the DeepLab series is often longer because of the substantial number of learnable parameters. Although the dilated convolutions effectively expand the receptive fields, they may lead to information loss within local areas. Furthermore, inappropriate selection of the dilation rate may induce the checkerboard phenomenon. In applications, the optimal performance of the networks requires more time and fine-tuning of the hyperparameters.

3) REAL-TIME SEMANTIC SEGMENTATION NETWORK WITH A DUAL-BRANCH STRUCTURE

Scholars have focused on lightweight, real-time semantic segmentation models. In 2016, ENet [53] was considered a milestone in real-time semantic segmentation. It uses dilated convolutions, lightweight encoders, and reduced down-sampling. However, ENet encounters reduced channel numbers and limited input sizes.

To address the issues of significant loss of spatial information in E-Net, BiSeNetV1 [54] uses a dual-branch structure to extract semantic and detailed features individually. It effectively captures spatial information while maintaining real-time performance (Figure 5). A Learning to Down-Sample module from Fast-SCNN [55] extracted shallow features, and then the features were shared between semantic and detail branches. On the basis of BiSeNetV1, STDC [56], BiSeNetV2 [57], and DDRNet [58] were developed to optimize the shallow layers of the networks and enhance the fusion effect. Because the direct fusion of different feature maps may lead to imbalanced information in dual-branch structures, PIDNet [59] uses a three-branch structure to capture and fuse features at different levels.

Overall, the dual-branch structure effectively balances the real-time performance and accuracy of semantic segmentation. It is suitable for scenarios that demand immediate responses. However, it is still a challenge to guarantee an optimal balance between the performance and efficiency of the network in practical applications.

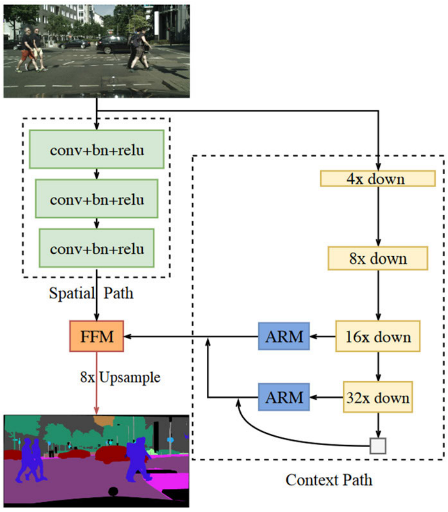

FIGURE 5. Structure of BiSeNetV1 network [54].

4) TRANSFORMER BASED SEGMENTATION

On the basis of a transformer structure, ViT [60] proposes a self-attention mechanism to capture global contextual information in images. Using the transformer's global self-attention mechanism, SETR [61] and Segmenter [62] also successfully captured global contextual information from serialized images. They achieved satisfactory results on the ADE20K, Pascal Context, and Cityscapes datasets. However, these models face challenges in resource-constrained devices owing to the large number of parameters.

With the global modeling capability of transformers, MobileViT [63], [64] combines the local receptive field of CNNs on semantic segmentation. Other lightweight models include Topformer [65], Seaformer [66], and DeMT [67]. Topformer achieves high efficiency and immediate response by adopting a global self-attention mechanism and lightweight decoder.

Future research should focus on compressing the model sizes, enhancing the acceleration efficiency of the hardware, and developing more efficient attention mechanisms. Moreover, these lightweight models should be expanded to more tasks and practical applications.

D. DEEP LEARNING-BASED OBJECT DETECTION ALGORITHMS

Object detection involves identifying and locating objects in one or more images. Current deep learning-based object detection networks can be divided into classification-based two-stage networks and regression-based one-stage networks.

1) TWO-STAGE OBJECT DETECTION NETWORKS

The R-CNN series (Figure 6) is a typical two-stage algorithm that divides the detection problem into two phases: region proposal and candidate region classification. The R-CNN algorithm proposed by Girshick et al. [68] generates object

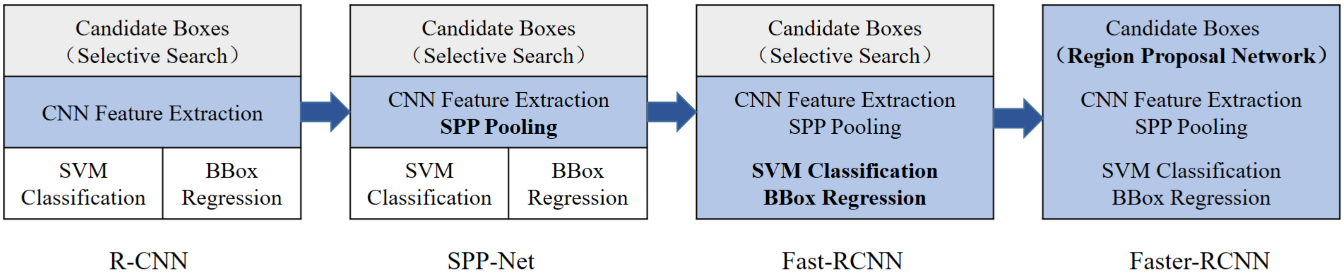

FIGURE 6. Development history of the R-CNN series.

candidate regions on the basis of a selective search method. It extracts the image features from each region via a CNN. Subsequently, it distinguishes the target classes and finally corrects the location of the detection frame with a regressor. Although the R-CNN algorithm has considerably improved the accuracy of networks, many overlapping candidate frames and redundant features obviously reduce the computational efficiency.

In the same year, a spatial pyramid pooling network (SPPNet) proposed by He et al. [69] avoids substantial redundant computations. It requires only a single feature-mapping computation for the entire image. Without loss of detection accuracy, the detection speed is more than 100 times faster than that of the R-CNN. By performing classification and regression simultaneously, the Fast-RCNN proposed by Girshick et al. [70] integrates the one-time feature extraction of SPPNet and changes the classifier and regressor structure of R-CNN in parallel. However, the selective search method, which generates candidate regions, limits the efficiency of object detection. Therefore, the following question arises: "Can a CNN model be used to replace the selective search method for generating region proposals?" Faster R-CNN [71] subsequently addressed this issue. In Faster R-CNN, a region proposal network (RPN) based on a CNN is applied to generate high-quality detection frames. This design improves the detection speed and enables end-to-end implementation of object detection for the first time.

Two-stage object detection networks achieve high accuracy by extracting candidate regions, and subsequently classifying and regressing the bounding boxes for each region. However, they struggle to meet the requirements of real-time and end-to-end object detection. Consequently, these networks are preferred in applications that focus on high accuracy, not real-time performance. This means that two-stage networks are often not optimal for intrusion detection of railway perimeters.

2) ONE-STAGE OBJECT DETECTION NETWORKS

Unlike the two-stage algorithm, which requires regional proposals, the one-stage algorithm can directly predict the position, category, and confidence of a target in a single phase. The high efficiency You Only Look Once (YOLO) [72] series are representative networks, because they transform the detection task into a regression problem. YOLO divides the image into a grid, and replaces the extraction of the candidate region. It predicts the bounding boxes and categories for each grid by applying a neural network to the entire image. Finally, YOLO achieves end-to-end training by simultaneously predicting the bounding boxes, target confidence, and categories for all the grids.

To data, nine versions of the YOLO series (Table 6) have been released, and each version has contributed to refining the network architecture and optimizing the loss functions. They have addressed issues such as the multi-scale problem, imbalance of samples, and poor detection results for small objects. Thus, the YOLO series detects objects more accurately and efficiently. For example, multi-scale detection and the use of more anchor boxes contribute to better adaptation to various object shapes and sizes. It provides a more reliable solution for real-time object detection.

In addition to the YOLO series, single shot multibox detector (SSD) [73], RetinaNet [74], EfficientDet [75], CenterNet [76], CornerNet [77], and the Transformer-based DETR (Detection Transformer) [78] are also popular one-stage object detection networks (Table 6).

III. DEEP LEARNING NETWORKS FOR INTRUSION DETECTING IN RAILWAY PERIMETERs FROM IMAGES

With the development of image-processing methods, RPID can be divided into two independent tasks: track area segmentation and foreign object detection. Multiple task learning methods have been subsequently proposed to perform object detection and track area segmentation simultaneously via a shared feature extraction network [79], [80]. Figure 7 illustrates the two intrusion detection methods in the railway perimeter, where the left image shows the two tasks, and the right image depicts the process of multi-task learning.

A. SEMANTIC SEGMENTATION-BASED RAILWAY TRACK AREA SEGMENTATION

1) COMMONLY USED RAILWAY TRACK SEMANTIC SEGMENTATION NETWORKS

Generally, images captured by cameras on railways have a large field depth and a broad area. They often include targets outside the railway perimeter, such as pedestrians, vehicles,

TABLE 6. Summary of popular one-stage object detection networks.

Algorithm	Year	Conclusion
YOLOv1	2016	YOLOv1 proposes a novel approach that transforms the object detection task into a regression problem, and segments the image into a grid where each cell is responsible for detecting objects within it.
SSD	2016	SSD proposes a method that predicts the category of the object and its location simultaneously within a single convolutional network, predicting targets on feature maps of various scales.
RetinaNet	2017	RetinaNet uses Focal Loss to address the class imbalance in object detection. It reduces the weight of easily classified samples by adjusting the weight distribution of the loss function, thereby increasing the contribution of difficult-to-classify samples to the overall loss. This enhances the model's ability to learn from challenging samples, consequently improving the performance of object detection.
YOLOv2	2017	Building upon YOLOv1, it enhances detection accuracy and speed through improvements, including the use of Anchor Boxes and Batch Normalization.
YOLOv3	2018	YOLOv3 further refines the YOLOv2, enhancing detection capabilities by implementing multi-scale detection and using more anchor boxes to accommodate objects of various shapes, particularly small target objects.
CornerNet	2018	CornerNet simplifies the model design by predicting the coordinates of the top-left and bottom-right corners of the target to complete the object detection task, thereby avoiding the use of Anchor Boxes and directly predicting the position of the box.
EfficinetDet	2019	EfficientNet enhances the accuracy and speed of object detection by integrating the efficient feature extraction capabilities of EfficientNet, the feature fusion mechanism of BiFPN, and the advantages of the Swish activation function.
CenterNet	2019	CenterNet streamlines the object detection process by directly predicting the center of the target and regressing the width and height to accomplish the object detection task.
YOLOv4	2020	YOLOv4 significantly enhances detection performance and speed by introducing new technologies, such as CSPDarknet53 as the backbone, Spatial Pyramid Pooling (SPP), and Path Aggregation Network (PANet).
YOLOv5	2020	YOLOv5 integrates the ideas from EfficientDet and PANet via a lightweight network architecture using multi-level feature fusion along with a cross-stage feature pyramid approach to enhance object detection performance.
DETR	2020	DETR is the first object detection algorithm based on the Transformer architecture, transforming the object detection task into a sequence-to-sequence problem. It uses the attention mechanism to establish global contextual information and introduces target position encoding and self-attention mechanisms to detect objects.
YOLOv6	2022	YOLOv6 abandons the anchor box mechanism used from YOLOv1 to v5, stacking multiple RepVGGBlocks in the backbone network to enhance the network's representational capacity and complexity. At the output end, the detection head is decoupled, separating the bounding box and the category classification processes.
YOLOv7	2022	YOLOv7 uses an efficient long-range attention network, leveraging the expansion, shuffling, and aggregation of features to continuously enhance the network's learning capabilities. It introduces a coarse-to-fine guided head labeling strategy, improving the model's recognition accuracy.
YOLOv8	2023	YOLOv8, from the same group as YOLOv5, features a newly designed backbone network and adheres to an anchor-free philosophy. It uses a positive sample allocation strategy during loss function computation, performing excellently in object detection and instance segmentation tasks.
YOLOv9	2024	YOLOv9 introduces programmable gradient information to address the changes required for deep networks to predict multiple targets, and designs a new lightweight network architecture based on gradient path planning for general efficient layer aggregation. The combination of both innovations has achieved excellent performance in lightweight models.

farmlands, and buildings. Thus, segmenting the track area as a region of interest (ROI) is crucial in detecting foreign objects.

Research on railway segmentation employs classical image processing techniques and deep learning-based methods.

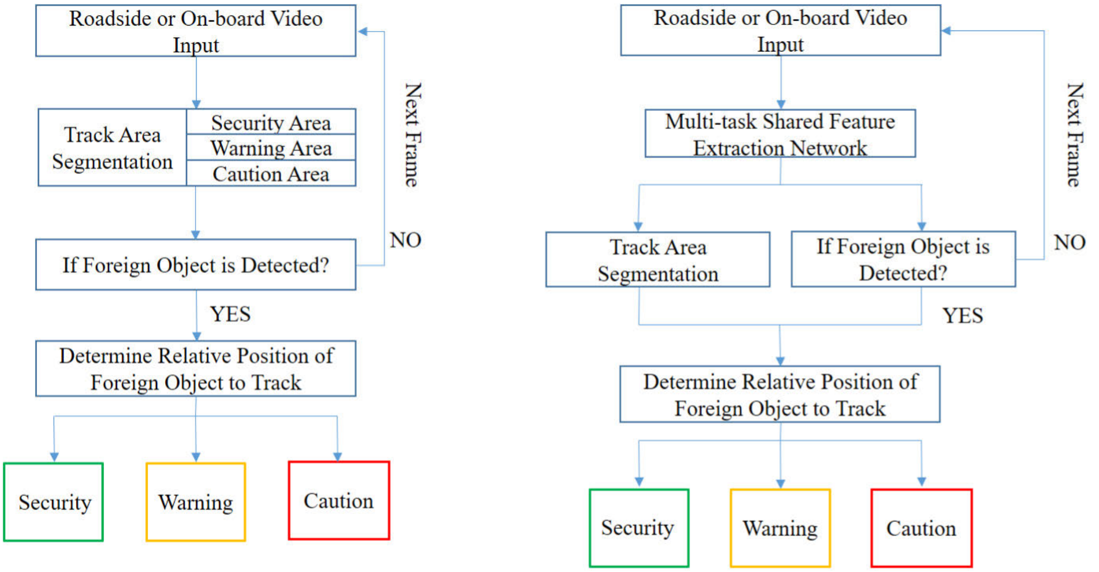
Roadside or On-board Video Input
Track Area Segmentation
Security Area
Warning Area
Caution Area
If Foreign Object is Detected?
Next Frame
NO
YES
Determine Relative Position of Foreign Object to Track
Security
Warning
Caution
Roadside or On-board Video Input
Multi-task Shared Feature Extraction Network
Track Area Segmentation
If Foreign Object is Detected?
Next Frame
NO
YES
Determine Relative Position of Foreign Object to Track
Security
Warning
Caution

FIGURE 7. Basic intrusion detection processes on railway tracks, where the left image shows the two tasks of track area segmentation and foreign object detection, and the multi-task learning detection process is depicted on the right image.

Classical image processing techniques such as Canny edge detection [81], Sobel edge detection [82], morphological operations [83], and threshold-based segmentation methods [84] are applied to extract the edges of railway tracks. These methods are effective but are more dependent on the choice of parameters, and may result in low efficiency. Furthermore, their performance may be affected by environmental factors such as illumination, weather conditions, and background.

Deep learning-based approaches exhibit distinct advantages in terms of automatic feature extraction. They demonstrate high accuracy and robustness in railway track segmentation. Through dilated cascade connections and polygon-fitting optimization, Wang et al. [85] proposed an efficient railway area detection method based on a CNN. It achieves high-precision on railway area identification under various lighting conditions and complex environments. On the basis of the DeepLab architecture, DFNet [86] achieves a high intersection-over-union (IoU) in a self-built dataset but with a relatively lower frame per second (FPS). RFODLab [87] introduces a channel attention mechanism and optimized loss function on the basis of the DeepLabV3+. This network effectively enhances the precision and recall of foreign object detection on high-speed railway tracks. However, the size of the network is also large. For similar references, refer to DFA-UNet [88] and Mask-RCNN [89]. Weng et al. [90] introduced an edge detection module and attention mechanism based on DLinkNet with high accuracy. However, the edge detection module increases the number of parameters and the computation time. This limits its application to vehicle-mounted or roadside equipment.

In total, despite the high accuracy of these networks, they often have large parameters and long computation times. Moreover, the processing ability of hardware is often limited. Consequently, lightweight segmentation networks, such as ERTNet [91] with an encoder-decoder structure and RailSegViTNet [92] based on ViT, have been developed to extract railway track areas. ERTNet achieves a balance between segmentation accuracy and computational efficiency with only 0.5M parameters and 0.92G FLOPs for floating-point operations. RailSegViTNet integrates lightweight bottleneck blocks, separable self-attention mechanisms, and feature aggregation. It achieves an average IoU of 91.43%, with 2.01G FLOPs and a parameter of 1.4 M. LRseg [93] achieved 18 FPS on a Jetson TX2 and 94 FPS on a computer equipped with an Intel Core i9-12900 CPU. Considering the balance between real-time performance and accuracy, LRseg is more suitable for onboard equipment.

Recent research has focused on reducing the model parameters and computational loads to balance time efficiency and detection accuracy. They provide solutions for limited computational resources without compromising segmentation accuracy. Future research should use hardware acceleration technology and explore multi-source data fusion methods to overcome the adverse effects of varying lighting and weather conditions.

2) SEMANTIC SEGMENTATION DATASET FOR RAILWAY

The success of deep learning relies on good-performance networks and big data. To the best of the authors' knowledge, the public railway scenario datasets currently available are Railsem19 and MRSI. Railsem19, constructed by Zendel et al. [94] in 2019, consists of 8500 images of railway scenes from different cities. The MRSI dataset [95] is collected from various sensors installed on locomotives. A total of 27,000 images of freight and subway tracks are recorded from track scenes under different lighting and weather conditions. Using the Railsem19 dataset, Table 8 shows the mean IoU (mIoU), parameters, and FLOPs of mainstream semantic segmentation networks.

TABLE 7. Comparisons of mainstream semantic segmentation models using the railsem19 dataset [93].

Model	MIoU/%	Parameters/M	FLOPs/G
BiseNet V2	91.5	3.34	12.26
STDC	92.4	12.3	11.75
LRASPP	90.2	3.22	1.98
PIDNet	90.9	7.72	5.94
Mobile-PSPNet	92.2	2.65	10.27
Mobile-DeepLabV3	92.6	3.23	22.61
Segmenter	86.9	26.0	37.36
Segformer	91.0	3.72	7.89
PSPNet	92.9	12.92	67.51
DeepLabV3	93.2	13.6	85.97
DDRNet	91.9	20.15	17.87
Topformer	88.6	5.02	1.61
Segnext	91.3	13.93	15.72
LRseg	92.4	0.78	1.47

In addition to the two public datasets, a railway simulation framework named TrainSim [96] automatically generates realistic railway scenarios and produces labeled datasets from simulated sensors such as LiDAR and cameras. This framework is effective for data expansion. It provides effective testing and training data for RPIDs.

Therefore, under the supervised learning paradigm, semantic segmentation networks require a large amount of labeled data for training. Alleviating the reliance on labeled data or making full use of unlabeled data is important to alleviate data scarcity and high annotation costs. Future research should explore the application of few-shot, weakly supervised, and self-supervised learning algorithms in track segmentation scenarios.

B. OBJECT DETECTION-BASED RAILWAY TRACK INTRUSION DETECTION METHOD

1) COMMONLY USED NETWORKS FOR OBJECT DETECTION OF RAILWAY TRACK INTRUSIONS

With the development of computer vision, visual detection has been popularly applied to detect railway obstacles. A multi-camera parallel monitoring system identifies the entrance of a person or obstacle into the designated warning area [97]. However, this system is sensitive to environmental changes, leading to a relatively high rate of false positives. Moreover, the selection of a threshold is crucial to guarantee the accuracy of the algorithm. Uribe et al. [98] analyzed the trajectories of objects in consecutive frames and tracked potential obstacles via an optical flow method. Sriwardene et al. [99] also used the optical flow method to detect obstacles and launch warnings to drivers when obstacles invade dangerous areas. However, the optical flow method has high computational requirements and may not satisfy real-time demands. Using a background subtraction algorithm, Li et al. [100] extracted intrusion targets from foreground images. However, the background subtraction algorithm is more suitable for detecting moving objects. These methods are generally limited in their ability to extract image features adaptively in complex railway environments.

With respect to the two-stage deep learning networks, He et al. [101] introduced a new up-sampling parallel structure and context extraction module based on an R-CNN. The network is trained through transfer learning to detect foreign objects on railway tracks. The accuracy of the improved R-CNN is achieved with a score of 90.6%, and it is 3.5% higher than that of the R-CNN. However, the detection speed was only 11 FPS. As discussed in Subsection II.D, two-stage object detection networks struggle to satisfy real-time processing requirements. Therefore, one-stage object detection networks such as SSD and YOLO are often selected as optimal networks.

On the basis of SSD, Xu et al. [102] replaced the backbone network with ResNet, and rapidly detected railway intruders with residual and feature fusion modules. Cong and Li [103] proposed an obstacle-detection network based on YOLOv3. This network demonstrated effectiveness in experiments with various scenarios and complex environments. However, in early studies, how to balance accuracy and speed was challenging, and how to guarantee the detection precision of small objects was insufficient in complex environments. To optimize the network, a lightweight feature extraction and adaptive feature fusion network, SEF-Net [104], is proposed to integrate image and primitive features. This network achieves good performance on detecting small targets with high efficiency. A RailDet network [105] was use to resolve the false positives in foreign object detection with high speed computation. Qin et al. [106] proposed an improved RetinaNet algorithm for efficient obstacle detection. The algorithm uses focal and global knowledge distillation strategies [107]. It enhances feature extraction and modeling capabilities without increasing the computational burden of deployment. It also addresses the difficulty of locating small obstacles through side-aware boundary localization [108]. Through experimental testing, RFA-Net improved the detection accuracy of small obstacles by 18.6%. It achieves a balance between accuracy and speed, with a detection accuracy of 92.7%, and a mean average precision (mAP) at an inference speed of 40.4 FPS. Table 8 presents comparisons of mainstream object detection algorithms on a self-built railway dataset. The mAP and FPS are selected as metrics to evaluate the performance of networks.

In summary, one-stage object detection networks, such as the YOLO series, generally meet real-time requirements. For the detection of railway track intrusions, most researchers are focused on increasing the detection accuracy for small targets.

In applications, low light and adverse weather conditions severely affect image quality. Similar to tracking segmentation tasks, detection from all-weather real-time images is a challenge. Therefore, integrating multi-source sensor data such as visible light images, infrared images, and LiDAR point clouds is a major topic for detecting foreign intrusions on railways.

TABLE 8. Comparisons among object detection networks [104].

Models	mAP/%	FPS
Faster-RCNN	77.57	10
DFF-Net	90.12	54
SSD	87.85	47
YOLOv3	90.80	67
YOLOv4	94.28	55
YOLOv5x	94.73	42
Mobilenetv2-SSD	76.31	68
Mobilenetv2-YOLOv3	82.83	74
LFD-Net	90.24	209
YOLOv3-tiny	84.52	205
YOLOv4-tiny	88.24	184
YOLOv5s	90.76	102
SEF-Net	94.75	81

2) OBJECT DETECTION DATASET FOR RAILWAY INTRUSION

The training samples of railway perimeter intrusion are low-probability events. The solutions for obtaining samples typically include simulating intrusion behaviors, creating synthetic datasets, and using transfer learning to publicly available datasets.

Guan et al. [109] obtained video sequences from six railway sections. They set up cameras along tracks from multiple shooting positions to capture the perspective-induced variations in obstacle sizes in the same scene. The experiments were subsequently performed to simulate different behaviors of individuals and groups crossing and staying on the tracks. Vehicle intrusion events were also captured at level crossings. The COCO dataset and selected pedestrians, vehicles, and animals are introduced as intrusion objects for training [105]. The test dataset from railway surveillance videos was composed of a total duration of approximately 30 mins and an image resolution of 1280 × 720. The surveillance videos contained 3,030 frames with simulated pedestrian intrusion on railway power outages. The videos included different weather conditions: sunny, cloudy, foggy, and snowy.

With respect to railway perimeter intrusion, potential intruders are diverse and random. Small or zero intruder samples, such as landslides, mudslides, and falling rocks on track lines, need to be constructed to support sufficient feature learning of networks.

C. MULTI-TASK LEARNING FOR RAILWAY INTRUSION DETECTION

The accuracy and real-time performance of foreign object detection and track area segmentation tasks have improved, however, two shortcomings exist in both tasks: (1) The preprocessing and inference of the two kinds of networks often consume considerable computational resources and time. (2) The features extracted by individual tasks are not fully shared. Therefore, multi-task learning has been investigated to rapidly detect the intrusion of foreign objects on railway tracks in an end-to-end manner.

1) Emulti-task learning

Multi-task learning performs well in terms of computational efficiency by sharing an encoder to learn the feature representations of the input data [79], [80], [110], [111]. It has become a paradigm that leverages information from multiple related tasks. Specifically, the encoder transforms the input data into a low-dimensional representation, and captures important data features. These features are shared by multiple decoders, eliminating the need for redundant feature extraction. Each decoder is responsible for a task, such as image classification, object detection, or semantic segmentation. By jointly training multiple decoder networks, the model simultaneously learns multiple tasks and optimizes the loss function. The multi-tasking learning approach enhances the generalization ability of the model.

2) application of multi-task learning in railway foreign object intrusion detection

For autonomous driving, Wu et al. [112] proposed a panoramic driving perception network called YOLOP. It uses a parallel multi-task learning network capable of performing three detection tasks simultaneously: detecting traffic objects, detecting drivable areas, and segmenting traffic lanes. Song et al. [113] combined semantic segmentation and depth estimation tasks to achieve real-time detection of obstacles in autonomous driving. Inspired by autonomous driving applications, Pan et al. [79] and Pang et al. [80] designed multi-task intrusion detection models for track area segmentation and foreign object detection in railway scenarios. Figure 8 illustrates the encoder-decoder architecture of multi-task learning for detecting foreign objects on tracks. The feature-sharing encoder is composed of a backbone and a neck network. It extracts robust and universal features from an input image. The backbone network is designed on the basis of the enhanced CSPResStage. It incorporates an effective squeeze-excitation module and a dilated convolution. The neck network consists of a feature pyramid network (FPN) and a path aggregation network (PAN), and it effectively fuses multiscale features from the backbone module. The decoder is divided into two parts: one for foreign object detection and the other for track area segmentation. Three detection heads in the decoder stage learn multi-level features from the neck network and predict the sizes of foreign objects with large, medium, and small sizes. The segmentation decoder restores the low-resolution feature map to raw image sizes and obtains the masks of the track areas.

The backbone network of DSORnet [114], similar to YOLOv5s, integrates the focus, SPP, and BottlenCSP

TABLE 9. Summary of segmentation and detection networks for RPID.

Ref.	Years	Methods	Datasets	Performance
Track area segmentation				
[85]	2018	SegNet, dilated convolution, and polygon fitting	Self-built dataset (5617 images)	MIoU=98.46%; FPS=38

[86]	2022	CBAM and SRM	Self-built dataset (21379 images)	MIoU=86.5%;
				RIoU=86.6%

[87]	2021	DeepLab v3+, ResNet 50, Combination of focal loss and dice loss	Self-built dataset (1217 images)	Recall=96.70%;
				Precision=91.52%

[88]	2023	Improved U-Net, and DFA attention module	Public RailSem19 dataset	MIoU=86.62%;
				Recall=90.57%; FPS=49.61

[89]	2022	ME Mask R-CNN, SSwin-Le Transformer, and ME-PAPN	Self-built dataset (3000 images)	mAP=91,3%; FPS=4.2

[90]	2024	Channel-spatial dual-attention mechanism, edge detection module	Self-built dataset (6990 images)	Accuracy=97.9%;
				MIoU=87.7%

[91]	2023	Sandglass-type feature extraction unit, feature-matching-based cross-fusion decoder, knowledge distillation	Self-built dataset (5280 images)	MIoU=92.4%;
				R.IoU=88.36%;
				0.5M parameters;
				0,92G FLOPs

[92]	2024	Hybrid architecture based on ViT and CNN, Lightweight self-attention mechanism	Self-built dataset (3000 images) and RailSem19 dataset	MIoU=91,43%; 2.01G FLOPs 1.4M parameters

[93]	2024	BSConv, SC-FFM, PPM, and knowledge distillation	Self-built dataset (2920 images) and RailSem19 dataset	MIoU=92.4%;
				R.IoU=86.2%;
				0.78M parameters
				1.47G FLOPs

Intrusion detection				

[102]	2019	Improved SSD, ResNet.	Self-built dataset (6384 images)	mAP=91.61%;
				FPS=26

[104]	2022	Improved YOLO, K-means clustering, ASFF, CIoU loss function	Self-built dataset (8776 images)	mAP=94.75%;
				FPS=81

[105]	2022	Depthwise separable convolutions, KNN, RailDet	COCO dataset	Accuracy=96.90%;
				FPS=23.26

[106]	2024	Improved RetinaNet, knowledge distillation, side-aware boundary localization	Self-built dataset (4246 images)	mAP=92.7%
				FPS=40.4

[109]	2022	Fast region proposal, Improved YOLO-tiny network	Self-built dataset (83664 images)	mAP=80.3%
Intrusion detection by multi-task learning (segmentation & detection)				

[79]	2022	Track line detection method based on row classification, multi-objective optimization method	Self-built dataset (2400 images)	MIoU=73.69%;
				FPS=40;
				Accuracy=68.6%

[80]	2023	Feature sharing encoder, two specific decoders for det & seg, squeeze-excitation module	RailSem19 and MRSI dataset	mAP=93.93%;
				MIoU=96.38%;
				FPS=31.53

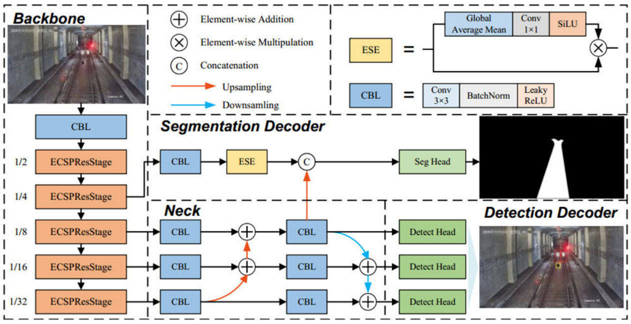

FIGURE 8. Architecture of multi-task learning for detecting railway foreign objects [80].

modules. It also uses an FPN to aggregate feature layers from the backbone network. The branch of semantic segmentation applies the lane segmentation head from the YOLOP network, whereas the branch of object detection adopts the decoding head of YOLOv5s. Experiments conducted on station throat areas confirmed that multi-task learning offers superior efficiency and accuracy in segmenting track areas and detecting foreign objects. However, multi-task learning faces the challenges of label annotation and complex network structures. More specifically, multi-task learning requires fine-grained annotation labels of locations and categories of foreign objects, and detailed segmentation information of track areas. However, the design and optimization of network structures are also limited in sharing features effectively, resolving potential conflicts in optimization, and balancing the requirements of different tasks. Therefore, designing an efficient network architecture is a major challenge that can fully utilize the advantages of shared representations and overcome interference among multiple tasks. Intelligent upgrades of annotation tools and the development of deeper optimization strategies are expected to address this challenge.

IV. CONCLUSION AND FUTURE WORK

Investigating efficient, high-precision, and real-time foreign object intrusion detection methods is crucial for ensuring train safety under various environmental conditions. Deep learning-based image recognition models effectively increase the accuracy and efficiency of intrusion detection while reducing labor costs. In recent years, deep learning-based semantic segmentation and object detection algorithms have provided robust technical support for detecting foreign objects on railways (Table 9).

Beginning with the development of deep learning, the advantages of CNNs and ViT in image processing are introduced. Deep learning-based semantic segmentation and object detection networks are summarized because they are the foundation for detecting foreign objects and segmenting track areas. The popular deep learning networks for RPID are categorized into three groups: semantic segmentation of railway track areas, detection of foreign objects, and multi-task learning-based detection of foreign objects. The main conclusions are as follows:

(1) Semantic segmentation based on deep learning networks achieves precise delineation of the track areas. Currently, the track-area related segmentation algorithms focus on reducing the number of parameters and computational loads without losing accuracy. However, semantic segmentation tasks for railway tracks face challenges such as data scarcity and the high cost of pixel-level annotation. Future research should explore strategies such as few-shot, semi-supervised, and self-supervised learning for track segmentation scenarios.

(2) Deep learning-based object detection networks have achieved rapid and accurate detection results for foreign objects of various types and geometric sizes. The detection of small targets is certainly improved. However, low-light conditions and adverse weather severely affect image quality. The integration of multi-source sensor data, such as visible-light image data, infrared image data, and LiDAR point clouds, should be considered when detecting foreign objects on railway tracks.

(3) Multi-task learning frameworks provide new solutions for detecting foreign objects. Sharing features extracted by the network enhances the model's generalization performance and effectively reduces computational costs. This enables the simultaneous processing of track area segmentation and foreign object detection. However, the applications of multi-task learning are lacking on railways. Balancing the optimization weights and maintaining learning efficiency is the main issue in cases of limited labeling samples on different tasks.

In summary, the detection and segmentation of deep learning-based railway perimeter intrusions have developed rapidly. Future work is needed in terms of generalizability, acceleration of the inference process, and adaptability to varying environments. Leveraging weakly supervised and unsupervised learning methods can also reduce the reliance on high-quality annotations. Investigating multi-source data fusion and image denoising may solve the challenges of low-light and adverse weather conditions. Further research should also investigate how to address zero false negatives and low false positives when detecting foreign objects in railway scenarios.

REFERENCES

[1] T. Shi, P. Guo, R. Wang, Z. Ma, W. Zhang, W. Li, H. Fu, and H. Hu, "A survey on multi-sensor fusion perimeter intrusion detection in high-speed railways," *Sensors*, vol. 24, no. 17, p. 5463, Aug. 2024.

[2] Z. Cao, Y. Qin, L. Jia, Z. Xie, Y. Gao, Y. Wang, P. Li, and Z. Yu, "Railway intrusion detection based on machine vision: A survey, challenges, and perspectives," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 7, pp. 6427–6448, Jul. 2024.

[3] S. Grabušić and D. Barić, "A systematic review of railway trespassing: Problems and prevention measures," *Sustainability*, vol. 15, no. 18, p. 13878, Sep. 2023.

[4] M. Di Summa, M. E. Griseta, N. Mosca, C. Patruno, M. Nitti, V. Renò, and E. Stella, "A review on deep learning techniques for railway infrastructure monitoring," *IEEE Access*, vol. 11, pp. 114638–114661, 2023.

[5] Q. Liu, Y. Qin, Z. Xie, T. Yang, and G. An, "Intrusion detection for high-speed railway perimeter obstacle," in *Proc. 3rd Int. Conf. Elect. Inf. Technol. Rail Transp. (EITRT)*, vol. 483, 2017, pp. 465–473.

[6] Y.-L. Zhang, Z.-Q. Zhang, G. Xiao, R.-D. Wang, and X. He, "Perimeter intrusion detection based on intelligent video analysis," in *Proc. 15th Int. Conf. Control, Autom. Syst. (ICCAS)*, Busan, Korea (South), Oct. 2015, pp. 1199–1204.

[7] K. Dharany and M. Trinita, "Enhancing urban safety: The role of object detection in smart city surveillance systems," *ITEJ*, vol. 8, no. 2, pp. 63–72, Dec. 2023.

[8] R. Arroyo, J. J. Yebes, L. M. Bergasa, I. G. Daza, and J. Almazán, "Expert video-surveillance system for real-time detection of suspicious behaviors in shopping malls," *Exp. Syst. Appl.*, vol. 42, no. 21, pp. 7991–8005, Nov. 2015.

[9] R. Du, P. Santi, M. Xiao, A. V. Vasilakos, and C. Fischione, "The sensable city: A survey on the deployment and management for smart city monitoring," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 2, pp. 1533–1560, 2nd Quart., 2019.

[10] A. Fedorov, K. Nikolskaia, S. Ivanov, V. Shepelev, and A. Minbaleev, "Traffic flow estimation with data from a video surveillance camera," *J. Big Data*, vol. 6, no. 1, p. 73, Dec. 2019.

[11] A. Pramanik, S. Sarkar, and J. Maiti, "A real-time video surveillance system for traffic pre-events detection," *Accident Anal. Prevention*, vol. 154, May 2021, Art. no. 106019.

[12] S. W. Khan, Q. Hafeez, M. I. Khalid, R. Alroobaea, S. Hussain, J. Iqbal, J. Almotiri, and S. S. Ullah, "Anomaly detection in traffic surveillance videos using deep learning," *Sensors*, vol. 22, no. 17, p. 6563, Aug. 2022.

[13] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *Bull. Math. Biophys.*, vol. 5, no. 4, pp. 115–133, Dec. 1943.

[14] S. Ding, H. Li, C. Su, J. Yu, and F. Jin, "Evolutionary artificial neural networks: A review," *Artif. Intell. Rev.*, vol. 39, no. 3, pp. 251–260, Mar. 2013.

[15] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, vol. 1, D. E. Rumelhart and J. L. McClelland, Eds., Cambridge, MA, USA: MIT Press, 1986, pp. 318–362.

[16] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Sep. 1995.

[17] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, Jul. 2006.

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 25, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds., Red Hook, NY, USA: Curran Associates, 2012, pp. 1097–1105.

[19] M. Liu, J. Shi, Z. Li, C. Li, J. Zhu, and S. Liu, "Towards better analysis of deep convolutional neural networks," *IEEE Trans. Vis. Comput. Graphics*, vol. 23, no. 1, pp. 91–100, Jan. 2017.

[20] Z. J. Wang, R. Turko, O. Shaikh, H. Park, N. Das, F. Hohman, M. Kahng, and D. H. P. Chau, "CNN explainer: Learning convolutional neural networks with interactive visualization," *IEEE Trans. Vis. Comput. Graphics*, vol. 27, no. 2, pp. 1396–1406, Feb. 2021.

[21] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[22] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[24] Z. Wu, C. Shen, and A. van den Hengel, "Wider or deeper: Revisiting the ResNet model for visual recognition," *Pattern Recognit.*, vol. 90, pp. 119–133, Jun. 2019.

[25] Z. Niu, G. Zhong, and H. Yu, "A review on the attention mechanism of deep learning," *Neurocomputing*, vol. 452, pp. 48–62, Sep. 2021.

[26] A. Apicella, F. Donnarumma, F. Isgrò, and R. Prevete, "A survey on modern trainable activation functions," *Neural Netw.*, vol. 138, pp. 14–32, Jun. 2021.

[27] R. Moradi, R. Berangi, and B. Minaei, "A survey of regularization strategies for deep models," *Artif. Intell. Rev.*, vol. 53, no. 6, pp. 3947–3986, Aug. 2020.

[28] Y. Cheng, D. Wang, P. Zhou, and T. Zhang, "Model compression and acceleration for deep neural networks: The principles, progress, and challenges," *IEEE Signal Process. Mag.*, vol. 35, no. 1, pp. 126–136, Jan. 2018.

[29] S. J. Cheng, Q. X. Zhao, X. Y. Zhang, N. Yadikar, and K. Ubul, "A review of knowledge distillation in object detection," *IEEE Access*, early access, Jun. 22, 2023, doi: 10.1109/ACCESS.2023.3288692.

[30] S.-K. Yeom, P. Seegerer, S. Lapuschkin, A. Binder, S. Wiedemann, K.-R. Müller, and W. Samek, "Pruning by explaining: A novel criterion for deep neural network pruning," *Pattern Recognit.*, vol. 115, Jul. 2021, Art. no. 107899.

[31] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.

[32] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5 MB model size," 2016, *arXiv:1602.07360*.

[33] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu, Z. Yang, Y. Zhang, and D. Tao, "A survey on vision transformer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 87–110, Jan. 2023.

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

[35] X. Dong, J. Bao, D. Chen, W. Zhang, N. Yu, L. Yuan, D. Chen, and B. Guo, "CSWin transformer: A general vision transformer backbone with cross-shaped windows," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 12114–12124.

[36] H. Mittal and M. Saraswat, "An optimum multi-level image thresholding segmentation using non-local means 2D histogram and exponential kbest gravitational search algorithm," *Eng. Appl. Artif. Intell.*, vol. 71, pp. 226–235, May 2018.

[37] M. Karakose, O. Yaman, M. Baygin, K. Murat, and E. Akin, "A new computer vision based method for rail track detection and fault diagnosis in railways," *Int. J. Mech. Eng. Robot. Res.*, vol. 6, no. 1, pp. 22–27, 2017.

[38] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[39] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*, vol. 9351, N. Navab, J. Hornegger, W. Wells, and A. Frangi, Eds., Cham, Switzerland: Springer, 2015, pp. 234–241.

[40] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder–decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[41] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[42] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5987–5995.

[43] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6848–6856.

[44] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.

[45] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, and C. Xu, "GhostNet: More features from cheap operations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 1577–1586.

[46] M.-H. Guo, C.-Z. Lu, Z.-N. Liu, M.-M. Cheng, and S.-M. Hu, "Visual attention network," *Comput. Vis. Media*, vol. 9, no. 4, pp. 733–752, Dec. 2023.

[47] J. Chen, S.-H. Kao, H. He, W. Zhuo, S. Wen, C.-H. Lee, and S.-H.-G. Chan, "Run, don't walk: Chasing higher FLOPS for faster neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 12021–12031.

[48] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected CRFs," 2014, *arXiv:1412.7062*.

[49] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2015, *arXiv:1511.07122*.

[50] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[51] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.

[52] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder–decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 801–818.

[53] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, "ENet: A deep neural network architecture for real-time semantic segmentation," 2016, *arXiv:1606.02147*.

[54] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "BiseNet: Bilateral segmentation network for real-time semantic segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 325–341.

[55] R. P. K. Poudel, S. Liwicki, and R. Cipolla, "Fast-SCNN: Fast semantic segmentation network," 2019, *arXiv:1902.04502*.

[56] M. Fan, S. Lai, J. Huang, X. Wei, Z. Chai, J. Luo, and X. Wei, "Rethinking BiSeNet for real-time semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 9716–9725.

[57] C. Yu, C. Gao, J. Wang, G. Yu, C. Shen, and N. Sang, "BiSeNet v2: Bilateral network with guided aggregation for real-time semantic segmentation," *Int. J. Comput. Vis.*, vol. 129, no. 11, pp. 3051–3068, Nov. 2021.

[58] H. Pan, Y. Hong, W. Sun, and Y. Jia, "Deep dual-resolution networks for real-time and accurate semantic segmentation of traffic scenes," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 3, pp. 3448–3460, Mar. 2023.

[59] J. Xu, Z. Xiong, and S. P. Bhattacharyya, "PIDNet: A real-time semantic segmentation network inspired by PID controllers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 19529–19539.

[60] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, and T. Unterthiner, "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021, pp. 0–10.

[61] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. S. Torr, and L. Zhang, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 6877–6886.

[62] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 7242–7252.

[63] S. Mehta and M. Rastegari, "MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer," 2021, *arXiv:2110.02178*.

[64] S. N. Wadekar and A. Chaurasia, "MobileViTv3: Mobile-friendly vision transformer with simple and effective fusion of local, global and input features," 2022, *arXiv:2209.15159*.

[65] W. Zhang, Z. Huang, G. Luo, T. Chen, X. Wang, W. Liu, G. Yu, and C. Shen, "TopFormer: Token pyramid transformer for mobile semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 12073–12083.

[66] Q. Wan, Z. Huang, J. Lu, G. Yu, and L. Zhang, "SeaFormer++: Squeeze-enhanced axial transformer for mobile visual recognition," 2023, *arXiv:2301.13156*.

[67] Y. Xu, Y. Yang, and L. Zhang, "DeMT: Deformable mixer transformer for multi-task learning of dense prediction," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, Jun. 2023, pp. 3072–3080.

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

[69] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015.

[70] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[71] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

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

[73] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. -Y. Fu, and A. C. Berg, "SSD: Single shot MultiBox detector," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 21–37.

[74] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2999–3007.

[75] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: Scalable and efficient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10778–10787.

[76] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, "CenterNet: Keypoint triplets for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6568–6577.

[77] H. Law and J. Deng, "CornerNet: Detecting objects as paired keypoints," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 734–750.

[78] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 213–229.

[79] H. Pan, Y. Li, H. Wang, and X. Tian, "Railway obstacle intrusion detection based on convolution neural network multitask learning," *Electronics*, vol. 11, no. 17, p. 2697, Aug. 2022.

[80] B. Pang, Q. Zhang, M. Chai, and H. Wang, "An efficient network for obstacle detection in rail transit based on multi-task learning," in *Proc. IEEE 26th Int. Conf. Intell. Transp. Syst. (ITSC)*, Sep. 2023, pp. 1795–1800.

[81] A. A. Shah, B. S. Chowdhry, T. D. Memon, I. H. Kalwar, and J. A. Ware, "Real time identification of railway track surface faults using Canny edge detector and 2D discrete wavelet transform," *Ann. Emerg. Technol. Comput.*, vol. 4, no. 2, pp. 53–60, Apr. 2020.

[82] T. Shi, J.-Y. Kong, X.-D. Wang, Z. Liu, and G. Zheng, "Improved Sobel algorithm for defect detection of rail surfaces with enhanced efficiency and accuracy," *J. Central South Univ.*, vol. 23, no. 11, pp. 2867–2875, Nov. 2016.

[83] M. Koohmishi and M. Palassi, "Evaluation of morphological properties of railway ballast particles by image processing method," *Transp. Geotechnics*, vol. 12, pp. 15–25, Sep. 2017.

[84] K. M. Pooja and R. Rajesh, "Image segmentation: A survey," in *Recent Advances in Mathematics, Statistics and Computer Science*. Singapore: World Scientific, 2015, pp. 521–527.

[85] Z. Wang, X. Wu, G. Yu, and M. Li, "Efficient rail area detection using convolutional neural network," *IEEE Access*, vol. 6, pp. 77656–77664, 2018.

[86] H. Yang, X. Li, Y. Guo, and L. Jia, "Discretization–filtering–reconstruction: Railway detection in images for navigation of inspection UAV," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–13, 2022.

[87] H. Song, S. Wang, Z. Gu, P. Dai, X. Du, and Y. Cheng, "Modeling and optimization of semantic segmentation for track bed foreign object based on attention mechanism," *IEEE Access*, vol. 9, pp. 86646–86656, 2021.

[88] Y. Zhang, K. Li, G. Zhang, Z. Zhu, and P. Wang, "DFA-UNet: Efficient railroad image segmentation," *Appl. Sci.*, vol. 13, no. 1, p. 662, Jan. 2023.

[89] D. He, R. Ren, K. Li, Z. Zou, R. Ma, Y. Qin, and W. Yang, "Urban rail transit obstacle detection based on improved R-CNN," *Measurement*, vol. 196, Jun. 2022, Art. no. 111277.

[90] Y. Weng, X. Huang, X. Chen, J. He, Z. Li, and H. Yi, "Research on railway track extraction method based on edge detection and attention mechanism," *IEEE Access*, vol. 12, pp. 26550–26561, 2024.

[91] Z. Chen, J. Yang, L. Chen, Z. Feng, and L. Jia, "Efficient railway track region segmentation algorithm based on lightweight neural network and cross-fusion decoder," *Autom. Construction*, vol. 155, Nov. 2023, Art. no. 105069.

[92] Z. Chen, J. Yang, and F. Zhou, "RailSegVITNet: A lightweight VIT-based real-time track surface segmentation network for improving railroad safety," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 36, no. 1, Jan. 2024, Art. no. 101929.

[93] Z. Feng, J. Yang, Z. Chen, and Z. Kang, "LRseg: An efficient railway region extraction method based on lightweight encoder and self-correcting decoder," *Exp. Syst. Appl.*, vol. 238, Mar. 2024, Art. no. 122386.

[94] O. Zendel, M. Murschitz, M. Zeilinger, D. Steininger, S. Abbasi, and C. Beleznai, "RailSem19: A dataset for semantic rail scene understanding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 1221–1229.

[95] Y. Chen, N. Zhu, Q. Wu, C. Wu, W. Niu, and Y. Wang, "MRSI: A multimodal proximity remote sensing data set for environment perception in rail transit," *Int. J. Intell. Syst.*, vol. 37, no. 9, pp. 5530–5556, Sep. 2022.

[96] G. D'Amico, M. Marinoni, F. Nesti, G. Rossolini, G. Buttazzo, S. Sabina, and G. Lauro, "TrainSim: A railway simulation framework for LiDAR and camera dataset generation," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 12, pp. 15006–15017, Dec. 2023.

[97] S. Oh, S. Park, and C. Lee, "A platform surveillance monitoring system using image processing for passenger safety in railway station," in *Proc. Int. Conf. Control, Autom. Syst.*, 2007, pp. 394–398.

[98] J. A. Uribe, L. Fonseca, and J. F. Vargas, "Video based system for railroad collision warning," in *Proc. IEEE Int. Carnahan Conf. Secur. Technol. (ICCST)*, Oct. 2012, pp. 280–285.

[99] A. L. S. Sriwardene, M. A. V. J. Muthugala, A. G. Buddhika, and P. Jayasekara, "Vision based smart driver assisting system for locomotives," in *Proc. IEEE Int. Conf. Inf. Autom. for Sustainability (ICIAfS)*, Dec. 2016, pp. 1–6.

[100] C. Li, Z. Xie, Y. Qin, L. Jia, and Q. Chen, "A multi-scale image and dynamic candidate region-based automatic detection of foreign targets intruding the railway perimeter," *Measurement*, vol. 185, Nov. 2021, Art. no. 109853.

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

[102] Y. Xu, C. Gao, L. Yuan, S. Tang, and G. Wei, "Real-time obstacle detection over rails using deep convolutional neural network," in *Proc. IEEE Intell. Transp. Syst. Conf. (ITSC)*, Oct. 2019, pp. 1007–1012.

[103] Z. Cong and X. Li, "Track obstacle detection algorithm based on YOLOv3," in *Proc. 13th Int. Congr. Image Signal Process., Biomed. Eng. Informat. (CISP-BMEI)*, Oct. 2020, pp. 12–17.

[104] T. Ye, Z. Zhao, S. Wang, F. Zhou, and X. Gao, "A stable lightweight and adaptive feature enhanced convolution neural network for efficient railway transit object detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 10, pp. 17952–17965, Oct. 2022.

[105] Z. Cao, Y. Qin, Z. Xie, Q. Liu, E. Zhang, Z. Wu, and Z. Yu, "An effective railway intrusion detection method using dynamic intrusion region and lightweight neural network," *Measurement*, vol. 191, Mar. 2022, Art. no. 110564.

[106] Y. Qin, D. He, Z. Jin, Y. Chen, and S. Shan, "An improved deep learning algorithm for obstacle detection in complex rail transit environments," *IEEE Sensors J.*, vol. 24, no. 3, pp. 4011–4022, Feb. 2024.

[107] Z. Yang, Z. Li, X. Jiang, Y. Gong, Z. Yuan, D. Zhao, and C. Yuan, "Focal and global knowledge distillation for detectors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 4633–4642.

[108] J. Wang, W. Zhang, Y. Cao, K. Chen, J. Pang, T. Gong, J. Shi, C. C. Loy, and D. Lin, "Side-aware boundary localization for more precise object detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 403–419.

[109] L. Guan, L. Jia, Z. Xie, and C. Yin, "A lightweight framework for obstacle detection in the railway image based on fast region proposal and improved YOLO-tiny network," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–16, 2022.

[110] Y. Zhang and Q. Yang, "A survey on multi-task learning," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 12, pp. 5586–5609, Dec. 2022.

[111] S. Vandenhende, S. Georgoulis, W. Van Gansbeke, M. Proesmans, D. Dai, and L. Van Gool, "Multi-task learning for dense prediction tasks: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 7, pp. 3614–3633, Jul. 2022.

[112] D. Wu, M.-W. Liao, W.-T. Zhang, X.-G. Wang, X. Bai, W.-Q. Cheng, and W.-Y. Liu, "YOLOP: You only look once for panoptic driving perception," *Mach. Intell. Res.*, vol. 19, no. 6, pp. 550–562, Dec. 2022.

[113] T.-J. Song, J. Jeong, and J.-H. Kim, "End-to-end real-time obstacle detection network for safe self-driving via multi-task learning," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 9, pp. 16318–16329, Sep. 2022.

[114] Z. Xiao, "Study on nonspecific foreign matter intrusion into the range of tracks based on deep learning," M.S. thesis, School Traffic Transp. Eng., Central South Univ., Changsha, Hunan, 2023.

JIN WANG (Member, IEEE) received the Ph.D. degree in geomatics engineering from the Leibniz University of Hannover, Germany, in 2013. She is currently an Associate Professor with Beijing Key Laboratory of Traffic Engineering, Beijing University of Technology, China. She has authored more than 40 research articles. Her research interests include image and LiDAR data processing, intelligent road maintenance, mobile and terrestrial laser scanning, and machine learning in transportation and civil engineering.

HONGYANG ZHAI received the bachelor's degree in transportation engineering from Ningxia University, Yinchuan, China, in 2023. He is currently pursuing the M.S. degree with Beijing Engineering Research Center of Urban Transport Operation Guarantee, Beijing University of Technology, Beijing, China. His current research interest includes railway safety.

YANG YANG received the B.S. degree in road and railway engineering and the Ph.D. degree in urban rail engineering from Beijing University of Technology, Beijing, China, in 2012 and 2018, respectively. She is currently a Lecturer with the College of Metropolitan Transportation, Beijing University of Technology. Her research interests include emergency management of urban rail transit, train timetable, and traffic big data.

NIUQI XU is currently pursuing the M.S. degree in transportation engineering from Beijing University of Technology, China. His research interests include intelligent processing of LiDAR point clouds, road sight distance access, driver's workload, and road safety evaluation in intelligent transportation systems.

HAO LI received the bachelor's degree from the School of Environmental Science and Engineering, Qingdao University. He is currently pursuing the M.S. degree with Beijing Engineering Research Center of Urban Transport Operation Guarantee, Beijing University of Technology, Beijing, China. His main research interest includes the field of transportation safety.

DI FU received the bachelor's degree from the College of Automotive and Transportation Engineering, Wuhan University of Science and Technology, in 2023. She is currently pursuing the master's degree with Beijing Engineering Research Center of Urban Transport Operation Guarantee, Beijing University of Technology, Beijing, China. Her current research interest includes vehicle driving safety.