Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 9
AIDML 2025

A Review on Semantic Understanding of Road Incidents Based on Vehicle-mounted Vision Systems

Zesheng Ma

School of Ecological and Environmental Sciences, East China Normal University, Shanghai, China

10164304127@stu.ecnu.edu.cn

Abstract. Road incidents pose serious threats to driving safety. In recent years, thanks to rapid advances in vehicle-mounted hardware and artificial intelligence technologies, it has become possible to apply these technologies for real-time road monitoring. This paper reviews methods for detecting and semantically understanding road incidents based on vehicle-mounted vision systems, focusing specifically on the applications and challenges of object detection and semantic segmentation in traffic scenarios. The paper introduces commonly used object detection models (e.g., YOLO, Faster R-CNN) and semantic segmentation models (e.g., DeepLab, PSPNet), along with their optimization methods. These optimization methods propose various improvement strategies such as attention mechanisms, feature fusion, and lightweight design to balance detection accuracy, real-time performance, and computational resources in complex traffic environments. Finally, the paper discusses future research directions, including improving model robustness under extreme conditions, multimodal data fusion, and hardware acceleration, aiming to promote broader applications of vehicle-mounted vision systems in intelligent transportation.

Keywords: Vehicle-mounted Vision System, Semantic Understanding, Deep Learning, Intelligent Transportation.

1. Introduction

Various unexpected incidents frequently occur on roads, such as traffic accidents, temporary construction activities, and road damage. These incidents not only reduce traffic efficiency but also directly threaten driving safety. With rapid urban development, responding quickly to these incidents in real-time has become a significant challenge in traffic management. Traditional methods mainly rely on manual patrol reports or fixed sensors, both of which have notable drawbacks. Firstly, manual patrols are time-consuming and slow in response. Secondly, fixed sensors have limited coverage, often leading to monitoring blind spots.

In recent years, autonomous driving technology has advanced rapidly, and an increasing number of vehicles are equipped with onboard cameras and edge computing devices, providing fundamental conditions for obtaining real-time road information. At the same time, the rapid progress in artificial intelligence, especially computer vision and deep learning, has made real-time monitoring and analysis of road conditions using vehicle-mounted equipment practically feasible.

The use of onboard devices for real-time traffic monitoring and analysis has several important benefits. Firstly, these devices can promptly detect incidents and warn vehicles behind, thereby preventing accident escalation and reducing the risk of secondary accidents. Secondly, autonomous driving systems receiving information about road abnormalities can automatically adjust driving routes, optimize navigation paths, and alleviate congestion. Additionally, replacing traditional manual inspections with intelligent devices can save human resources and reduce traffic management costs. Finally, this research helps build smarter and more efficient traffic management systems, improves overall urban transportation efficiency, and supports the development of smart cities.

Computer vision-based analysis methods typically consist of two main parts: object detection and semantic segmentation. Centering on these two components, this paper introduces commonly used

models and recent studies aimed at improving them. Furthermore, it briefly summarizes the current challenges facing research in this area and provides insights into future research directions.

2. Analysis of Object Detection Models

2.1. Commonly Used Object Detection Models

2.1.1. YOLO Model

YOLO model is a widely used deep learning-based object detection model. YOLO stands for "You Only Look Once," a concise yet clear description highlighting the model's key feature: it can simultaneously perform object localization and classification tasks in just a single forward pass, achieving fast and accurate detection results.

The detection process of the YOLO model (illustrated in Figure 1) can be divided into three steps [1]. First, the input image is divided into an S×S grid. Second, predictions are made in two parts: the first part predicts bounding boxes along with their confidence scores, indicating the position, size, and the presence of objects within each grid cell; the second part predicts class probabilities, indicating the likelihood that each grid cell belongs to a particular object category. Lastly, bounding boxes and class probabilities from all grid cells are combined and post-processed using Non-Maximum Suppression (NMS) to remove overlapping boxes, leaving only the boxes with the highest confidence as the final detection results.

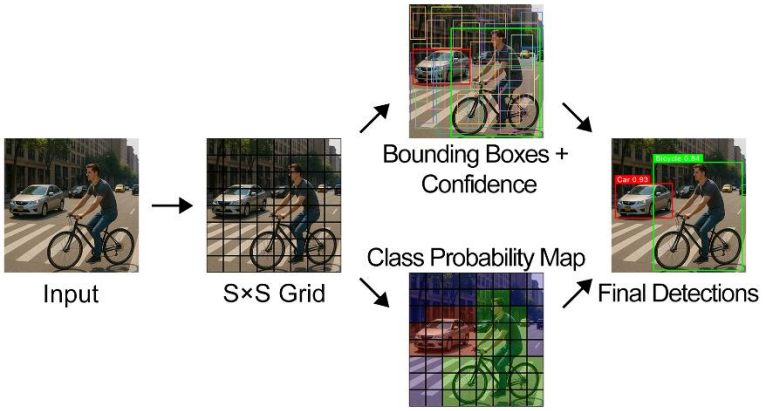

Figure 1. Detection process of the YOLO model (Picture credit: Original)

The main features of the YOLO model are as follows [2]: Firstly, YOLO completes all detection tasks, including localization and classification, in a single forward pass. Unlike traditional methods, YOLO does not require repeated computation of candidate regions; instead, it uses a single network to predict bounding boxes and classes for all objects. Secondly, YOLO is much faster than other object detection algorithms, such as Faster R-CNN, due to its single forward pass, making it suitable for real-time applications like autonomous driving and surveillance. Lastly, YOLO's network structure leverages global information, capturing better contextual understanding from images.

2.1.2. Faster R-CNN Model

Faster R-CNN (Region-based Convolutional Neural Network) is a deep learning object detection algorithm optimized from traditional R-CNN methods, especially through improvements in the region proposal stage.

Faster R-CNN has three key characteristics [1, 2]. Firstly, its main innovation is the introduction of the Region Proposal Network (RPN). Traditional object detection methods depend on external algorithms (like selective search) to generate candidate regions, whereas Faster R-CNN integrates RPN into an end-to-end training framework, greatly improving efficiency. The RPN directly generates candidate regions from input images automatically. Secondly, Faster R-CNN is a two-stage detection model: the first stage generates candidate regions using RPN, and the second stage classifies

objects and refines bounding boxes based on these regions. Since the candidate regions are learned through a convolutional network, they are highly accurate. Lastly, Faster R-CNN shares convolutional features between RPN and Fast R-CNN, improving computational efficiency by avoiding repetitive calculations.

The advantages and disadvantages of YOLO and Faster R-CNN are compared in Table 1 below:

Table 1. Comparison of Object Detection Models

Model	Advantages	Disadvantages
YOLO	High efficiency, real-time capability, global context modeling	Poor detection of small objects, lower accuracy, localization errors
Faster R-CNN	Higher accuracy, automated region proposals	High computational load, complex training, unsuitable for high-performance real-time applications

From the comparison shown in Table 1, Faster R-CNN is more suitable for offline and high-precision tasks, while YOLO and other lightweight algorithms are better for real-time applications.

2.2. Improvements in Object Detection Models

2.2.1. Optimizing Faster R-CNN with Attention Mechanism

When humans process visual data, attention is initially focused on important objects while ignoring irrelevant ones. This selective processing, known as the attention mechanism, reduces the amount of visual data processed, allocating limited computational resources to critical targets to enhance visual task complexity, efficiency, and accuracy. The principle of the attention mechanism is illustrated in Figure 2.

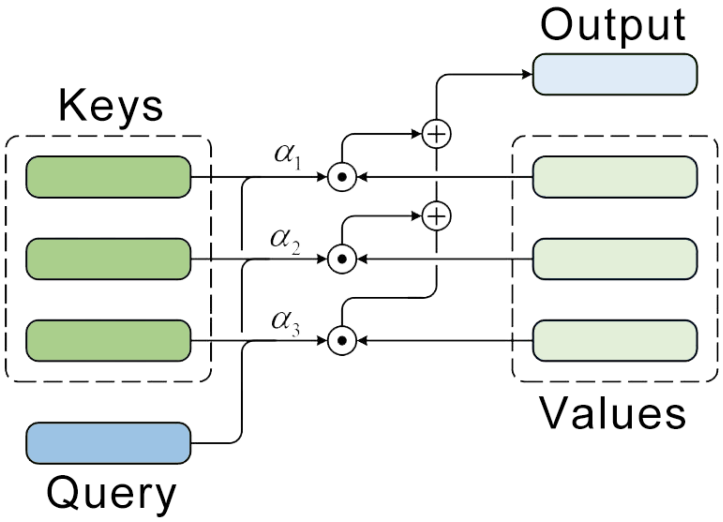

Figure 2. Principle of the attention mechanism [3]

Liu proposed introducing attention mechanisms into Faster R-CNN to enhance its semantic understanding capability [3]. By introducing scene semantic attention to dynamically weight scene-related objects (such as vehicles, pedestrians, and traffic lights), the optimized model generates descriptions more aligned with the scene's characteristics. Experimental results showed improvements in evaluation metrics BLEU-4 and CIDEr, scoring 37.5 and 121.9 respectively, outperforming other models.

However, the research also highlighted two drawbacks. First, it did not deeply explore the spatiotemporal relationships among objects (e.g., vehicle trajectories). Second, the real-time performance of this method was not verified, potentially limiting its applicability to vehicle-mounted systems.

2.2.2. Optimized YOLO Algorithms

Zhou and colleagues proposed a pothole detection method based on the YOLOv5s algorithm, with additional improvements aimed at enhancing detection accuracy and generalization while maintaining real-time performance [4]. Their optimization approach introduced two key mechanisms: Squeeze-and-Excitation (SE) and Bidirectional Feature Pyramid Network (BiFPN).

The SE mechanism improves model performance by squeezing and exciting input features, assigning different weights to feature maps, and directing the model's attention toward critical parts, thus enhancing feature extraction. The BiFPN addresses issues in multi-resolution feature fusion seen in traditional methods by weighting and combining multi-scale features, further improving performance.

Experimental results indicated significant improvements in accuracy, rising from 81.1% to 95.0% (an increase of 17.1%), and mean Average Precision (mAP) improving from 48.2% to 49.5% (an increase of 2.69%). The detection speed remained almost unchanged, fulfilling real-time detection requirements. Despite achieving a good balance between accuracy and speed, this optimization method still faces limitations, including insufficient robustness in complex environments, high annotation costs, and increased model complexity potentially causing overfitting. The researchers suggest future optimization could involve expanding datasets and exploring lightweight attention mechanisms.

In addition, Xia et al. further improved YOLOv5s to tackle real-time detection challenges in complex traffic environments [5]. Their improvements included four main aspects:

First, the introduction of the M-ELAN (Multi-scale Efficient Layer Aggregation Networks) module, enhancing multi-scale feature extraction while significantly reducing model parameters (81.4% fewer), thereby increasing detection speed without sacrificing accuracy. Second, integrating the SimAM (Simple Attention Module) enabled the network to compute channel and spatial attention weights, enhancing feature extraction without adding additional parameters. Third, applying a parameter-balancing strategy optimized the parameter ratio between the backbone and neck of the network, improving detection accuracy, especially for small objects. Lastly, replacing the CIoU loss function with the EIoU loss function accelerated convergence and improved detection accuracy.

Experiments on the SODA10M dataset showed that the improved TTD-YOLO model outperformed standard YOLOv5-S and other detection algorithms such as Faster R-CNN and YOLOX-S. Compared to standard YOLOv5-S, the improved model's mAP increased from 44% to 45.1% (a 2.5% improvement), with inference speed increased by 4.8%.

3. Analysis of Semantic Segmentation Models

3.1. Commonly Used Semantic Segmentation Models

3.1.1. DeepLab Model

The DeepLab series, developed by Google, primarily addresses the issue of capturing multi-scale objects in semantic segmentation. The model's main characteristics include three modules: Atrous Convolution, Atrous Spatial Pyramid Pooling (ASPP), and an improved encoder-decoder structure. Atrous convolution expands the receptive field of convolutions without decreasing spatial resolution, allowing for richer contextual information capture. The ASPP module simultaneously employs atrous convolutions at various sampling rates to effectively capture multi-scale contextual information. Finally, the DeepLabv3+ version uses an encoder-decoder architecture, significantly improving detailed segmentation, particularly at object boundaries [6].

3.1.2. PSPNet Model

PSPNet employs a Pyramid Pooling Module (PPM) for effective multi-scale context feature fusion, enabling better extraction of global semantic information. The PPM divides feature maps into multiple scales (e.g., 1×1, 2×2, 3×3, 6×6), performs pooling at each scale, and then merges the results

to capture both global and local contextual information effectively. The PSPNet model, shown in Figure 3, is particularly suitable for complex scenes (e.g., urban environments) due to its strong ability to understand overall scene layouts. Additionally, PSPNet exhibits stable training processes and robust generalization performance, making it effective for large-scale image segmentation tasks [7].

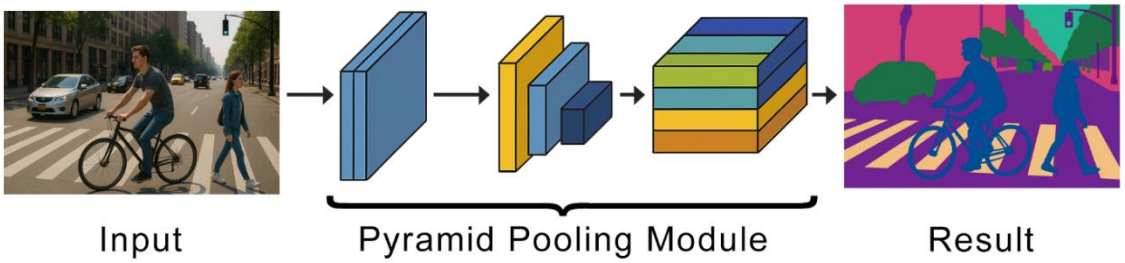

Figure 3. Workflow of the PSPNet Model (Picture credit: Original)

3.1.3. SegNet Model

SegNet, proposed by the University of Cambridge, is based on the classic encoder-decoder structure designed for efficient pixel-level segmentation. The encoder generally uses the VGG16 network architecture to extract high-dimensional features. Its decoder mirrors the encoder, progressively restoring spatial dimensions through upsampling to achieve fine-grained pixel-level segmentation. SegNet employs pooling indices from max-pooling layers in the encoder to facilitate nonlinear upsampling in the decoder, greatly reducing computational load and preserving accurate boundary information [8].

SegNet features a simple structure with fewer parameters, making it suitable for real-time segmentation or applications on resource-limited devices, despite relatively lower accuracy.

Table 2 summarizes the advantages and disadvantages of the semantic segmentation models mentioned above:

Table 2. Comparison of Semantic Segmentation Models

Model	Advantages	Disadvantages
DeepLab	Strong multi-scale context capturing capability, fine boundary segmentation	Complex model structure, high computational resource demand
PSPNet	Good global contextual capturing, precise boundary segmentation	Slightly complex model, high computational cost
SegNet	Lower memory consumption, higher efficiency	Lower accuracy, weaker detail capturing

3.2. Improvements in Semantic Segmentation Models

3.2.1. Optimized DeepLab Model

Wang improved the DeepLabV3+ model, aiming to enhance segmentation accuracy in traffic scenarios while maintaining real-time processing [9]. This research tackled issues of inadequate detailed representation and segmentation quality caused by complex urban road environments, particularly regarding variations in lighting, weather conditions, and road surfaces. The improvements consisted of two key parts:

The first part, IDE_DeepLabV3+, introduced a Coordinate Dual Attention Mechanism (CDAM) to enhance feature representation and designed a Parallel Multi-branch ASPP (PM_ASPP) module to better integrate multi-scale features. The decoder employed modules for boundary refinement and Multiscale Context Extraction (MSCE), improving edge segmentation accuracy and handling occluded road areas better.

The second part, IDEL_DeepLabV3+, focused on model lightweighting based on IDE_DeepLabV3+. It replaced the original backbone with MobileNetV2 and used depthwise separable convolutions within the ASPP module, significantly reducing model parameters and improving inference speed while maintaining segmentation accuracy.

The IDEL_DeepLabV3+ model achieved improvements over the original DeepLabV3+, increasing mIoU from 66.53 to 69.22 (4.04% improvement), and mPA from 82.14 to 83.70 (1.89% improvement).

3.2.2. Optimized PSPNet Model

Liu proposed combining superpixel segmentation with multi-level and multi-scale feature fusion for image semantic segmentation [3]. By extracting multi-level and multi-scale features and integrating superpixel segmentation results, the approach more precisely extracted semantic information, improving object recognition accuracy. This method, built upon PSPNet, enhanced segmentation capability for small objects and fuzzy edges in complex traffic scenes, leading to improved accuracy. Compared with PSPNet, the improved model increased mIoU from 79.7 to 80.2 (0.63% improvement) and mPA from 88.7 to 89.6 (1.01% improvement).

3.2.3. Semantic Segmentation Model Based on Knowledge Distillation

Xie et al. proposed a "boundary-aware guided multi-level feature knowledge distillation" semantic segmentation algorithm for traffic scenarios [10]. It aimed to address problems such as detail loss in segmentation results and large model parameter size. The algorithm enhanced segmentation through multi-level feature fusion, attention mechanisms, and knowledge distillation techniques.

The main innovations of this method include three points: First, an adaptive multi-level feature fusion module combined shallow spatial details with deeper semantic information, selectively emphasizing object boundaries and main body information to improve boundary segmentation performance. Second, an interactive attention fusion module proposed a new attention mechanism, combining long-range dependencies in spatial and channel dimensions to optimize information interaction, enhancing segmentation results. Lastly, a boundary loss function based on candidate boundaries distilled boundary-aware knowledge from a complex teacher network to a simpler student network, significantly improving segmentation accuracy, particularly for small or slender objects.

Experimental results on Cityscapes and CamVid datasets showed that this lightweight algorithm maintained high segmentation performance. On the CamVid dataset, the improved model raised mIoU from 67.1 to 76.8 (14.5% improvement) and PA from 91.8 to 93.2 (1.52% improvement) compared to DeepLabV3+. On the Cityscapes dataset, the model improved mIoU from 69.1 to 76.3 (10.4% improvement) and PA from 93.2 to 96.1 (3.11% improvement) compared to PSPNet.

4. Challenges and Future Outlook

4.1. Challenges

4.1.1. Accuracy of Object Detection in Complex Scenes

Road incidents usually occur in complex and dynamic traffic environments. To effectively identify situations like traffic accidents and road obstacles, vehicle-mounted cameras must quickly and accurately detect targets such as other vehicles, pedestrians, and construction signs. However, changes in weather (e.g., rain, snow, fog), variations in lighting conditions (day and night), and different road conditions (dry or muddy) can significantly reduce the accuracy of object detection and semantic segmentation. Additionally, heavy traffic often leads to overlapping and occlusion between objects, further decreasing detection accuracy. Small targets, such as falling rocks or distant pedestrians, are also challenging due to their minimal appearance and indistinct features in the images.

4.1.2. Balance Between Real-time Performance and Computational Resources

Vehicle-mounted systems must respond quickly, especially when sudden braking or obstacles appear unexpectedly. Deep learning models, especially convolutional neural networks, often require substantial computational power, which is limited on vehicle-mounted equipment. Achieving fast and efficient deep learning inference under resource-constrained conditions remains a significant challenge. High-precision models like DeepLabV3+ and Faster R-CNN typically have heavy computational demands, making real-time performance difficult when dealing with multi-scale feature extraction and complex post-processing tasks.

4.1.3. Dataset Issues

High-quality training data is crucial for model effectiveness, yet data for road incidents typically suffer from imbalanced class distributions. For example, hazardous chemical leakage events are rare and produce limited image data, making it difficult for models to accurately recognize these incidents. Additionally, traffic conditions vary greatly by region, weather, and time, causing existing data to quickly become outdated. Improving the model's generalization to apply across various scenarios is thus a critical research challenge.

4.1.4. Multi-task Learning and End-to-End System Integration

Vehicle-mounted cameras often need to perform multiple tasks simultaneously, such as object tracking, behavior recognition, and scene understanding. Currently, designing a unified framework capable of efficiently handling multiple tasks without interference is challenging. Moreover, implementing deep learning models in real vehicle systems must also consider system stability, reliability, and scalability. Optimizing model performance under limited hardware conditions and network bandwidth to ensure real-time responsiveness remains an open issue.

4.2. Application Scenarios

The application scenarios of this research primarily include Advanced Driving Assistance Systems (ADAS), Intelligent Transportation Systems (ITS), and Vehicle-to-Everything (V2X) technology.
In ADAS, vehicle-mounted cameras can detect sudden road incidents in real time, such as sudden braking by vehicles ahead or traffic accidents. The system can immediately issue warnings and provide driving assistance suggestions, such as collision warnings, lane-change alerts, automatic speed adjustments, or even stopping the vehicle, enhancing driving safety.

For ITS, vehicle-mounted cameras can not only identify accidents or abnormal vehicle behaviors but also monitor traffic flow and congestion levels. Accurate real-time monitoring enables the system to optimize traffic signals dynamically, reducing traffic congestion.

V2X technology, a crucial component of autonomous driving systems, facilitates communication between vehicles, infrastructure, networks, and cloud platforms. Incident detection models installed on vehicles can instantly upload accident information to the cloud, enabling quick route adjustments for other vehicles, thus reducing risks and improving the efficiency and safety of the entire traffic system.

4.3. Future Outlook

Future research will focus on two major directions: adapting to advances in hardware technology and further optimization of models themselves.

From the hardware perspective, future vehicle-mounted systems will integrate not only cameras but also various sensors like LiDAR and millimeter-wave radar due to advancements in sensor technology. Therefore, research into multimodal data fusion is essential, enhancing system adaptability under harsh conditions, and improving detection accuracy and robustness. Additionally, with the widespread adoption of advanced hardware such as GPUs, TPUs, and FPGAs, inference speed for deep learning models will significantly improve. Coupled with research into lightweight

models and model compression techniques (e.g., pruning, quantization, and knowledge distillation), future systems can effectively meet the dual demands of real-time performance and computational efficiency.

From the model perspective, current detection models still struggle under severe conditions such as rain, snow, fog, or nighttime. Future research will enhance model robustness, enabling accurate detection even with impaired visual information. Additionally, adaptive learning and transfer learning will become increasingly important, allowing models to automatically adjust parameters based on regional, climatic, and road condition differences. Transfer learning can also help mitigate data scarcity, enabling quick model adaptation to new environments through online and incremental learning methods.

5. Conclusion

This paper reviewed current research and developments in semantic understanding of road incidents based on vehicle-mounted vision systems, analyzing in detail the major models for object detection and semantic segmentation, along with their optimization techniques. Although significant advancements have been achieved, numerous challenges remain. Future research must enhance the adaptability of models, integrate multimodal data, and explore model optimization and hardware acceleration to achieve higher real-time performance and accuracy. It is expected that future research will resolve these issues, supporting the development of smarter, safer, and more efficient traffic management systems, and further advancing the construction of smart cities.

References

[1] V. Mandal, A. R. Mussah, P. Jin, et al, Artificial intelligence-enabled traffic monitoring system, Sustainability, 12.21 (2020) 9177.

[2] M. Kiac, P. Sikora, L. Malina, et al, ADEROS: artificial intelligence-based detection system of critical events for road security, IEEE Systems Journal, 17.4 (2023) 5073-5084.

[3] S. Liu, Research on Understanding of Vehicle Road Collaborative Traffic Scene Based on Image Semantics and Deep Learning, Shenzhen University, (2023)

[4] J. Zhou, Y. Hu, J. Shi, et al, Road Pothole Detection for Autonomous Vehicles Based on Deep Learning. Computer Science and Application, 14 (2024) 29.

[5] W. Xia, P. Li, H. Huang, et al, TTD-YOLO: A real-time traffic target detection algorithm based on YOLOV5, IEEE Access (2024)

[6] Y. Deng, S. Mo, H. Gan, et al, Based on DeepLab v3+ model to realize the road scene semantic segmentation, 2022 IEEE 10th Joint International Information Technology and Artificial Intelligence Conference (ITAIC). Vol. 10. IEEE, (2022)

[7] Q. Yang, L. Yu, Recognition of taxi violations based on semantic segmentation of PSPNet and improved YOLOv3, Scientific Programming 2021.1 (2021) 4520190.

[8] R. Kavya, K. M. Z. Hussain, N. Nayana, et al, Lane Detection and Traffic Sign Recognition from Continuous Driving Scenes using Deep Neural Networks, 2021 2nd International Conference on Smart Electronics and Communication (ICOSEC). IEEE, (2021)

[9] Y. Wang, Improved DeepLabV3+ for Semantic Segmentation of Traffic Scenes, Chang'an University, (2024)

[10] X. Xie, Z. Duan, C. Luo, et al, Traffic Scene Semantic Segmentation Algorithm with Knowledge Distillation of Multi-level Features Guided by Boundary Perception, Pattern Recognition and Artificial Intelligence, 37.09 (2024) 770-785.