Research Advances in YOLO-Based Obstacle Detection Algorithms

Yuhan Huang[1,*]

[1]School of Computer Science, Hubei University, Wuhan, Hubei province,430064, China

Abstract. With recent breakthroughs in artificial intelligence and computer vision, obstacle detection has become an essential capability for enhancing safety and enabling automation across various industrial sectors. Among the range of methods available, YOLO-series algorithms have gained widespread adoption in practical applications due to their effective balance between inference speed and detection accuracy. This paper presents a systematic literature review of studies on YOLO-based obstacle detection published between 2023 and early 2025. It focuses on ten representative works that highlight significant advances in areas such as multimodal sensor fusion, lightweight model deployment, and scenario-specific optimizations—including applications in underground mining, agricultural robotics, and low-altitude UAV missions. By summarizing the technological evolution and comparing model performance across different constraint conditions, this study addresses the current absence of comprehensive survey work in this rapidly evolving field. Furthermore, it provides a structured analysis of improvements in network architectures, training strategies, and evaluation protocols, delivering valuable insights for researchers and practitioners working toward efficient and robust obstacle detection systems.

1 Introduction

The rapid advancement of artificial intelligence and computer vision technologies has significantly promoted the widespread application of obstacle detection techniques. Currently, obstacle detection has become a core supporting technology in various fields such as inspection robots, intelligent transportation systems, agricultural automation, and safety management in mining areas. In recent years, vision-based detection methods using deep learning have achieved breakthrough progress. Among them, the YOLO (You Only Look Once) series of algorithms have been widely adopted in industrial practices including rail transit safety, agricultural automation, low-altitude UAV inspection, and mining area safety monitoring, owing to their effective balance between detection speed and accuracy. In light of this, this article provides a systematic review of the latest research progress in YOLO-based algorithms for obstacle detection.

This study focuses on 10 recently published representative papers to reflect the current research status and advances in YOLO-based obstacle detection technology. In terms of

* Corresponding author: 202231116020148@stu.hubu.edu.cn

multimodal data fusion, Qiu Gang et al. [1] proposed an inspection robot obstacle recognition method based on image matching, which improves detection accuracy in low-light conditions by fusing infrared and visible light. Yang Zhifang [2] and Sun et al. [3] enhanced environmental perception through sensor complementarity in underground coal mine and general scenarios, respectively, by integrating radar and vision. Regarding lightweight model deployment, Ruan Shunling et al. [4] designed an efficient detection model for edge devices in mining areas, balancing computational resources and accuracy. Dong Yibing et al. [5] improved a UAV vision detection network based on YOLOv8, achieving lightweight deployment in low-altitude scenarios. For specific scenario optimization, Zhao Hongliang et al. [6] and Yang Haolin et al. [7] adapted YOLOv5 for complex backgrounds in rail transportation and farmland, respectively. Han Keli et al. [8] further optimized the YOLOv8n model to enhance the detection of small obstacles in cotton fields. In emerging tasks, Cao et al. [9] improved YOLOv7 performance in dense pedestrian detection through weight optimization. Assemlali et al. [10] proposed a CNN-driven novel classification framework to address real-time detection and classification of dynamic obstacles.

This study adopts a systematic literature review methodology to conduct a comprehensive review and analysis of representative achievements in vision-based obstacle detection published between 2023 and 2025. The article systematically summarizes the evolution of technological developments, addressing the current lack of survey work in this field (as shown in Table 1) and thereby provides researchers with a clear roadmap of technological progress.

Table 1. Overview comparison.

Overview	Time	Technology	Lightweight	Multimodal Fusion	Robustness
A Review of Obstacle Detection Technologies for Autonomous Vehicles	2010-2022	Information integration	Partially involved	Partially involved	Partially involved
A Review of Research Methods for Obstacle Detection in Intelligent Vehicles	2007	Basic Technology	Not involved	Partially involved	Partially involved
This paper	2023-2025	Obstacle detection targeting vertical fields	Key coverage	Key coverage	Key coverage

2 Related Work

2.1 Dataset

Research on obstacle detection technology highly relies on high-quality and representative datasets. Depending on the application scenario, various studies employ diverse datasets to train and validate their models.

In extremely harsh environments such as underground coal mines, publicly available datasets are scarce, and researchers often resort to self-constructed datasets. For example, Yang Zhifang [2], while studying radar and vision fusion technology, built a dataset of underground tunnel scenes containing synchronized radar point clouds and camera images. This dataset needed to cover complex working conditions such as sudden illumination

changes, dust interference, and equipment occlusion. Similarly, Ruan Shunling et al. [4], focusing on edge computing in mining areas, also used a dataset derived from real mining scenarios, which included various obstacles such as mining trucks, excavators, ore piles, pedestrians, as well as potholes and rolling stones. Its characteristics include complex backgrounds, multi-scale targets, and significant inter-class variations.

In the field of rail transportation, the research by Zhao Hongliang et al. [6], which improved YOLOv5 for obstacle detection, typically utilized datasets sourced from monitoring equipment installed along railway lines. These datasets include trains, pedestrians, vehicles, signalling equipment, and illegally intruding animals or debris. The background is relatively structured, but the detection requirements for small and long-distance targets are extremely high.

Obstacle detection in agricultural field environments faces unique challenges. Yang Haolin et al. [7] and Han Keli et al. [8] both addressed complex field environments, with datasets primarily consisting of images captured between crop rows in cotton fields, cornfields, etc. The obstacle categories mostly include stones, plastic film, weeds, and abandoned agricultural machinery. Such datasets are characterized by complex background textures, target colors similar to the soil background, and rapidly changing lighting conditions.

For applications such as UAV inspection and robotics, the perspective and tasks of the datasets are more dynamic. The research by Dong Yibing et al. [5] on low-altitude UAV visual target detection used datasets composed of aerial images captured by drones, with targets mostly including vehicles, pedestrians, and building rooftops. The characteristics include a top-down perspective and a large span of target scales. Qiu Gang et al. [1], based on infrared and visible light image matching, necessarily employed datasets containing strictly paired infrared and visible image pairs to overcome conditions with poor visible light imaging, such as nighttime and haze. The obstacle categories include equipment, pedestrians, or animals with distinct thermal features.

Furthermore, some studies focus on challenges in general scenarios, such as dense pedestrian detection. The work by Jie Cao et al. [9] required handling highly occluded and crowded scenes, using datasets derived from public pedestrian detection datasets with targeted enhancements. The research by Assemlali et al. [10] placed more emphasis on the classification of dynamic obstacles, necessitating datasets containing image sequences to reflect the target's motion information.

In summary, obstacle detection datasets are highly scenario-dependent and task-oriented. From underground mines and fields to aerial environments, the construction of these datasets aims to replicate and address the core challenges specific to their applications. A summary of these datasets, along with their classes, Annotation Granularity, and core challenges, is provided in Table 2.

Table 2. Dataset.

Dataset Name	Resolution /Modality	Classes	Annotation Granularity	Core Challenges	Model Performance (mAP@0.5, etc.)
Paired Infrared-Visible Dataset [1]	Infrared images + visible light images	Equipment, pedestrians, animals	Bounding boxes + class labels	Heterogeneous image registration, cross-modal information complementarity, nighttime/haze conditions	Nighttime Recall and Precision far exceed visible-light-only models

Underground Coal Mine Environment Dataset [2]	Radar point cloud + visible light images	Equipment, tunnel structures, pedestrians, obstacles	Bounding boxes + class labels	Low illumination, dust interference, multimodal fusion	Fusion method accuracy significantly higher than vision-only models
Camera-Radar Fusion Dataset [3]	Radar point cloud + visible light images	Generic obstacles	Bounding boxes	Spatiotemporal sensor synchronization, data alignment	Overall mAP improvement of fusion model, reliable detection maintained especially in vision-failure scenarios
Mining Area Obstacle Dataset [4]	Visible light images	Mining trucks, excavators, ore piles, pedestrians, potholes, rolling stones	Bounding boxes	Multi-scale variations, occlusion, small objects (potholes), complex background	Parameters reduced by 44%, inference speed increased by 34%
Low-Altitude UAV Perspective Dataset [5]	Visible light images	Vehicles, pedestrians, buildings	Bounding boxes	Perspective pitch, large target scale variation	minimal accuracy drops with significantly reduced parameters
Rail Transit Obstacle Dataset [6]	Visible light images	Trains, pedestrians, vehicles, signaling equipment, animals	Bounding boxes	Small objects, long-distance targets, structured background	Improved YOLOv5 accuracy
Complex Field Environment Dataset [7]	Visible light images	Stones, plastic film, weeds, agricultural machinery	Bounding boxes	Complex background texture, target color similar to soil, lighting variations	Improved YOLOv5 mAP@0.5 significantly increased
Cotton Field Obstacle Dataset [8]	Visible light images	Stones, residual film, weeds	Bounding boxes	mall objects, similarity to cotton plant background	parameters reduced by 46.5%
Dense Pedestrian Dataset [9]	Visible light images	Pedestrians	Bounding boxes	Severe occlusion, crowd density	Weight-optimized YOLOv7 showed improved performance on dense pedestrian detection
Dynamic Obstacle Dataset [10]	Visible light images	Dynamic obstacles	Bounding boxes + sequence information	Motion information extraction, temporal modeling	CNN model performed well in dynamic obstacle detection and classification

2.2 Evaluation Indicators

To objectively evaluate the performance of different models, the aforementioned studies generally adopt widely accepted evaluation metrics in the field of computer vision, primarily using Average Precision (AP) and mean Average Precision (mAP) as the core metrics for assessing detection accuracy, while also considering detection speed to meet real-time requirements.

In underground coal mine scenarios, the radar-vision fusion method employed by Yang Zhifang [2] demonstrates its core advantage in enhancing robustness under harsh conditions. The evaluation metrics showed that the fusion model achieved a significantly higher mAP on their self-built underground dataset compared to a vision-only model, particularly on subsets with interference such as dust and low light, proving the effectiveness of the fusion approach. The obstacle detection model for mining areas by Ruan Shunling et al. [4] was designed for deployment on edge computing devices. Therefore, in addition to reporting a high mAP, their study emphasized the reduction in the model's parameter count and computational complexity, as well as maintaining an acceptable FPS (Frames Per Second) under high computational constraints, achieving a balance between accuracy and efficiency.

Models improved based on the YOLO series reported excellent performance across various domains. The rail transportation model by Zhao Hongliang et al. [6], the field obstacle model by Yang Haolin et al. [7], and the cotton field model by Han Keli et al. [8] all achieved significant improvements in mAP on their respective datasets. For instance, the improved YOLO models in [7] and [8] reached high mAP levels, far surpassing the baseline models and effectively addressing the issues of missed detections and false alarms for small targets and complex backgrounds in field environments. The lightweight LMUAV-YOLOv8 network by Dong Yibing et al. [5] focused on significantly reducing model complexity while maintaining accuracy and markedly increasing FPS to meet the computational constraints of UAV platforms.

For multi-modal fusion approaches, such as the camera-radar information fusion technology by Sun Q H et al. [3], the evaluation metrics not only included an improvement in overall mAP but also demonstrated performance gains across different distance ranges and weather conditions. This proved that the fusion model could maintain reliable detection using radar when visual sensors failed. The research on infrared and visible light image matching by Qiu Gang et al. [1] highlighted that their method significantly outperformed visible-light-only models in recall and precision during nighttime scenarios, effectively reducing the risks associated with nighttime operations.

Studies such as the dense pedestrian detection by Cao Jie et al. [9] and the dynamic obstacle classification by Assemlali et al. [10] placed greater emphasis on performance in highly occluded and crowded scenarios.

In summary, the evaluation frameworks of these studies collectively indicate that an excellent obstacle detection model must achieve an optimal trade-off between accuracy (mAP) and efficiency (e.g., FPS, computational load), with a specific focus tailored to the application scenario, such as resisting interference, handling small objects, or enabling real-time deployment. A summary of the core evaluation metrics employed in these studies—including their names, definitions, and practical significance—is systematically presented in Table 3.

Table 3. Evaluation indicators.

Metric Name	Definition	Significance
AP@50	AAverage Precision calculated at a fixed IoUthreshold of 0.5	Measures the detection capability under lenient localization requirements. It is the

		most widely used accuracy metric in the industry.
AP@[.5:.95] (or mAP)	Mean Average Precision. AP is calculated at 10 IoU thresholds from 0.5 to 0.95 (step size 0.05) and then averaged.	A core metric for the COCO dataset. Comprehensively evaluates localization accuracy across varying strictness levels, better reflecting the model's precise localization capability.
Params	Number of model parameters, usually measured in millions (M) or billions (B).	Directly related to model size and complexity. A key metric for evaluating model lightweightness and deployment difficulty.
FLOPs	Floating Point Operations.	Measures computational complexity. Strongly correlated with inference speed. Lower values indicate theoretically lower energy consumption and faster speed.
FPS	Frames Per Second.	A core metric for inference speed. Directly determines whether the system can meet real-time requirements.
Latency	The average time (in milliseconds, ms) required to process a single image.	Reciprocal of FPS. More intuitively reflects the time consumption of a single inference.

3 Equations and mathematics

3.1 Accuracy Enhancement (Unimodal)

This paradigm primarily aims to improve the recognition accuracy and robustness under a single visual modality. Typical technical approaches include introducing channel or spatial attention mechanisms atop the YOLO series baselines, redesigning the feature pyramid structure, and making targeted modifications to the loss function.

Works based on the YOLOv5 baseline are particularly common. Studies by Zhao Hongliang et al. [6] and Yang Haolin et al. [7] are representative examples. By incorporating attention mechanisms, designing more efficient feature fusion networks, and optimizing the loss function, they enhanced the model's ability to represent targets in complex backgrounds (such as rail transportation and farmland environments), significantly improving the detection accuracy for small and occluded objects.

The advantage of this strategy lies in its relatively concise model structure, which can significantly enhance performance while maintaining efficient end-to-end inference. However, its limitation is that performance improvements are often achieved at the cost of increased model complexity (e.g., adding extra modules), leading to higher parameter counts and computational overhead, thereby imposing greater demands on the computational resources and memory of deployment devices.

3.2 Lightweight and Acceleration

To address the challenge of deploying the models on resource-constrained devices, the lightweight and acceleration paradigm has emerged. The core objective of this direction is to significantly reduce model complexity and inference latency while minimizing accuracy loss, through various model compression and acceleration techniques.

Typical techniques include the adoption of depth wise separable convolutions, the design of lighter backbone networks (e.g., GhostNet), model pruning, and knowledge distillation. The work by Dong Yibing et al. [5] is an exemplary effort in this direction. They constructed the LMUAV-YOLOv8 lightweight network based on YOLOv8, achieving efficient deployment on UAV platforms through architectural optimization. Similarly, the model developed by Ruan Shunling et al. [4] for edge devices in mining areas also incorporates similar lightweight design strategies.

The advantage of this approach is its ability to reduce the model's dependency on hardware computational power, thereby meeting real-time requirements. However, its limitation lies in the fact that extreme compression may impair the model's representational capacity, particularly its perception of subtle features such as small objects or targets with indistinct textures, resulting in a performance ceiling.

3.3 Multimodal Fusion

To overcome the inherent limitations of single visual sensors, which are susceptible to variations in lighting and weather conditions, multimodal information fusion has become a key focus of cutting-edge research. This approach enhances the perceptual robustness of systems in harsh environments by fusing data from cameras with other sensors, leveraging complementary information.

Based on the level of fusion, it can be categorized into data-level, feature-level, and decision-level fusion. The research by Yang Zhifang [2] and Qinhao Sun et al. [3] focuses on feature-level fusion of radar and vision, with innovations in designing sophisticated data alignment and feature fusion networks that effectively combine visual texture information with radar depth/velocity data. Qiu Gang et al. [1] employed pixel-level matching and fusion of infrared and visible light images to enhance the inspection robots' ability to recognize heat-emitting targets in low-light conditions.

The advantage of this strategy is its ability to improve system stability and reliability under complex working conditions. Its main challenge lies in the high system complexity, involving preprocessing steps such as multi-sensor temporal synchronization, spatial calibration, and heterogeneous data alignment, which significantly increases the difficulty and cost of engineering implementation.

3.4 Scenario Specialization

In addition to general strategies, scenario-specific research addressing challenges such as dense occlusions, dynamic obstacles, and small targets is equally important. In dense crowd/occlusion scenarios, Jie Cao et al. optimized label/sample assignment and loss reweighting based on YOLOv7 to alleviate matching confusion for dense targets and improve recall in crowded areas [9]. Assemlali et al. focused on the detection and classification of dynamic obstacles, emphasizing that incorporating temporal consistency and trajectory-level features can enhance the stability of moving target detection [10]. For small targets, common practices include preserving/enhancing high-resolution branches, refining the Feature Pyramid Network (FPN), increasing dense anchors, or incorporating super-resolution/amplification strategies. However, these must be carefully balanced with the lightweight strategies discussed in § 3.2 to avoid excessive latency and memory consumption.

4 Conclusion

In summary, the development of obstacle detection model technology demonstrates a clear evolutionary trajectory: from enhancing accuracy in unimodal approaches, to optimizing lightweight design and efficiency, and further to improving robustness through multimodal fusion.

In terms of unimodal accuracy, as illustrated in Figure 1, this direction builds upon well-established baseline models such as YOLO. Through strategies like introducing attention mechanisms, redesigning feature pyramids, and optimizing loss functions, it aims to maximize the representational capability and detection accuracy of a single visual modality in specific scenarios. Its advantage lies in its relatively concise model architecture and ability to maintain efficient inference. However, performance improvements often come at the cost of increased model complexity and computational demands, placing higher requirements on deployment hardware.

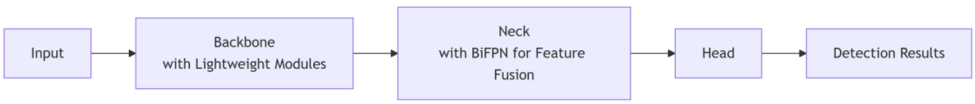

Fig. 1. Accuracy Enhancement.

In terms of lightweight design and acceleration, as illustrated in Figure 2, this direction serves as a direct solution to address the increasing complexity of unimodal models and meet practical deployment requirements. By employing techniques such as depth wise separable convolutions, lightweight backbone networks, model pruning, and knowledge distillation, it significantly reduces the model's parameter count, computational load, and inference latency while maintaining acceptable accuracy loss. This enables real-time performance on resource-constrained platforms such as drones and edge computing devices. The core challenge lies in balancing model compression with the preservation of its ability to perceive small targets and subtle features.

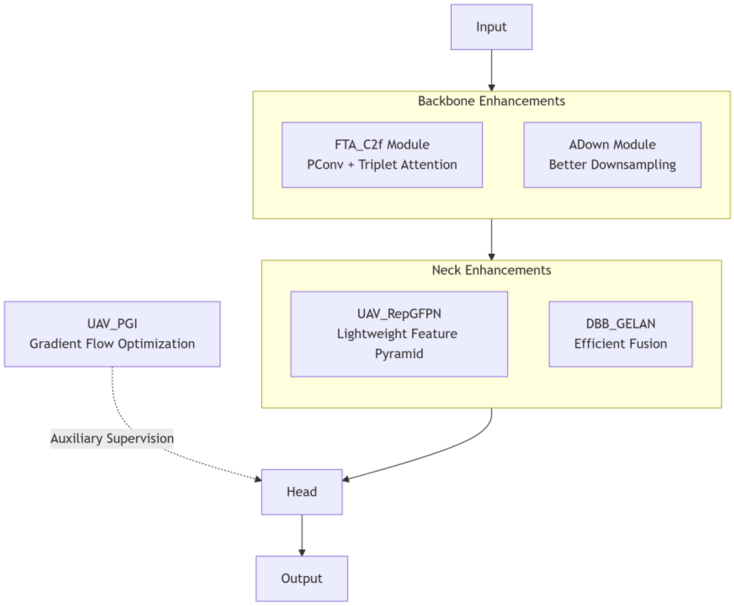

Fig. 2. Lightweight and Acceleration.

In the aspect of multimodal fusion, as depicted in Figure 3, this direction fundamentally addresses the inherent limitation of single visual sensors being susceptible to environmental interference. By integrating data from heterogeneous sensors such as cameras, radar, and infrared, it leverages complementary information to enhance the system's perceptual robustness under complex working conditions like sudden illumination changes and adverse weather. While its advantages are significant, the trade-off involves a sharp increase in system complexity, posing engineering challenges such as spatiotemporal synchronization of multiple sensors, calibration, and alignment and fusion of heterogeneous data.

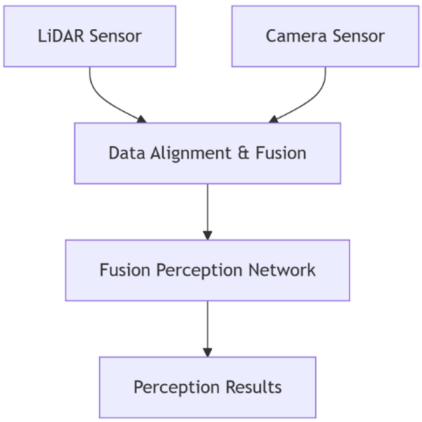

Fig. 3. Multimodal Fusion.

These three dimensions are not isolated but rather complementary and co-evolutionary. The future development trend will involve the organic integration of all three: constructing efficient and lightweight model architectures, embedding advanced modules for accuracy enhancement, and building upon this foundation to elegantly integrate multimodal information. Ultimately, this approach will achieve an optimal balance among accuracy, efficiency, and robustness, promoting the practical application of obstacle detection technology across various real-world scenarios.

References

1. Qiu, G., Zhang, N. L., Bai, C., Tan, X., Chen, J., & Gao, S. (2025). Obstacle recognition and classification for inspection robots based on infrared and visible image matching. Infrared Technology, *47*(1), 81 – 88.
2. Yang, Z. F. (2023). Dual-modal environmental perception technology for underground coal mines based on radar and vision fusion. Industry and Mine Automation, *49*(11), 67 – 75.
3. Sun, Q. H., Yao, G. S., He, S., Duan, X. C., & Ma, M. J. (2025). Obstacle detection technology based on camera and radar information fusion. Electro-Mechanical Engineering, *41*(1), 46 – 51.
4. Ruan, S. L., Wang, J., Gu, Q. H., & Lu, C. W. (2024). Research on obstacle detection model in mining areas for edge computing. Coal Science and Technology, *52*(11), 141 – 152.
5. Dong, Y. B., Zeng, H., & Hou, S. J. (2025). LMUAV-YOLOv8: A lightweight network for visual target detection in low-altitude UAVs. Journal of Computer Engineering & Applications, *61*(3).
6. Zhao, H. L., Guo, Y. M., Wang, J. X., & Yang, J. (2024). Rail transit obstacle detection algorithm based on improved YOLOv5. Electronic Measurement Technology, *47*(1).
7. Yang, H. L., Wang, Q. H., Li, H. B., Geng, D. Y., Wu, J. D., & Yao, Y. C. (2024). Obstacle detection in complex field environments based on improved YOLOv5. Journal of Chinese Agricultural Mechanization, *45*(6), 216.
8. Han, K. L., Wang, Z. K., Yu, Y. F., Liu, S. P., Han, S. J., & Hao, F. P. (2025). Obstacle detection method in complex cotton field environments based on improved YOLOv11n model. Transactions of the Chinese Society for Agricultural Machinery, *56*(5), 111 – 120.
9. Cao, J., Niu, Y., & Liang, H. P. (2025). Dense pedestrian detection algorithm based on YOLOv7 with optimized weights. Chinese Journal of Liquid Crystals and Displays, *40*(3), 505 – 515.
10. Assemlali, H., Bouhsissin, S., & Sael, N. (2025). Deep learning-driven CNN model for detection and classification of dynamic obstacles. Green Energy and Intelligent Transportation, 100334.