2025 International Conference on Advanced Mechatronics and Intelligent Energy Systems

# The Impact of Quadrotor UAV on Urban Rail Transit Inspection

AIPCP25-CF-AMIES2025-00080 | Article

# The Impact of Quadrotor UAV on Urban Rail Transit Inspection

Yu Han

*School of Urban Railway Transportation, Shanghai University of Engineering Science, Shanghai, 201600, China*

a137218@correo.umm.edu.mx

**Abstract.** With the rapid expansion of urban rail transit networks, the traditional manual inspection model is facing severe challenges in terms of efficiency, safety, and detection accuracy. This paper explores the application of image recognition algorithms applicable to unmanned aerial vehicles (UAVs) in the intelligent inspection of urban rail transit. Specifically, it discusses the application of intelligent inspection technologies based on quadrotor UAVs and image recognition algorithms in urban rail transit. Through improved YOLO-series algorithms, U-Net models, and multi-source data fusion strategies, the inspection efficiency and fault recognition accuracy have been significantly enhanced. Practical experiments show that the use of UAVs can increase inspection efficiency by 3-5 times, and the fault recognition rate is 40% higher than that of manual inspection. The key technological breakthroughs include complex light compensation, small-target detection optimization, and multi-source data fusion strategies, which effectively address the issues of dynamic blur and false alarms in mobile inspections. Future research will focus on multi-modal sensor collaboration, edge computing deployment, and self-learning model iteration to provide technical support for constructing an intelligent rail transit operation and maintenance system.

## INTRODUCTION

Urban rail transit inspection can identify potential problems and faults, thereby avoiding safety accidents and safeguarding the lives and property of personnel. As of 2023, a total of 338 urban rail transit lines have been put into operation in 59 cities across China, with an operating mileage of 10,975.8 kilometers [1]. With the increase in operational and under-construction lines, the number of inspectors and inspection costs will rise sharply. Meanwhile, the management difficulty of inspections increases, the quality is hard to guarantee, and safety accidents are prone to occur, making the use of unmanned aerial vehicles (UAVs) for inspection particularly important. According to practical data, in the case of 560 km of operational and under-construction lines, deploying only 3 UAVs can increase the inspection frequency from 2-3 times per week to 6-8 times per week, significantly reducing inspection costs. As the line mileage increases, the costs will gradually decrease, and the inspection efficiency can be improved by more than 10 times [2].

At present, the image recognition algorithms for UAVs used in urban rail transit inspection mainly include two types: traditional computer image processing and machine learning and computer vision technologies based on convolutional neural networks [3]. In recent years, research on convolutional neural networks has included the following: in 2020, Lei et al. studied a vision-based concrete crack detection method, including images collected by UAVs, preprocessing algorithms, crack center point methods, and a classifier based on a support vector machine model [4]. Lin et al. proposed a region-based fast convolutional neural network that integrates visible light and thermal imaging data, achieving automatic identification and hierarchical classification of both surface and internal corrosion characteristics in steel structures [5]. Subsequently, Ramandi's research team developed an intelligent crack detection framework for digital rock core image analysis. This framework employs primary visual preprocessing to locate potential crack regions, followed by advanced image interpretation algorithms to precisely extract planar cracks from 3D grayscale computed tomography images [6]. For urban rail transit inspection, image recognition technology has now relatively mature development, but the navigation and obstacle avoidance

capabilities of UAVs in complex environments still need to be improved, as well as how to enhance UAV stability in extreme environments. These issues are of great significance for UAV inspection.

This paper aims to explore some UAV algorithms applicable to urban rail transit inspection and provide some suggestions.

## IMAGE RECOGNITION TECHNOLOGY

### YOLO

The object identification model You Only Look Once (YOLO) is based on deep learning. By converting the object identification job into a one-stage regression issue, its main idea is to use a single forward propagation to directly forecast the categories and position information of every item in an image. The input picture is divided into an S×S grid of cells using YOLO, and each grid cell is in charge of predicting a certain number of bounding boxes. The central coordinates, width, height, and confidence are all included in each bounding box. In the meantime, every grid cell forecasts the probability distribution of every category, and the product ultimately determines the object's categorisation. Meanwhile, each grid cell also predicts the probability distribution of all categories, and the category of the object is finally determined by the product of the confidence and the category probability.

The YOLO framework utilizes hierarchical convolutional layers to hierarchically extract visual patterns from input data, thereby establishing an integrated feature representation mechanism. To address scale variation challenges in object detection models, the architecture incorporates cross-resolution feature aggregation through multi-level pyramid connections, which contributes to improved recognition performance across varying object scales while maintaining real-time processing efficiency. This design paradigm effectively balances computational complexity with detection accuracy in practical vision systems. The model's loss function integrates localization errors (coordinates and dimensions), confidence errors (whether an object is present), and classification errors, optimizing overall performance through end-to-end training. Compared with two-stage detection models (such as Faster R-CNN), YOLO significantly improves detection speed by omitting the region proposal step, making it suitable for real-time applications. However, its early versions had lower detection accuracy for small objects and dense scenes. Subsequent improved versions (such as YOLOv3 and the subsequent YOLOv4-v8) have gradually optimized performance balance by introducing mechanisms like Anchor Boxes and Feature Pyramid Networks (FPN).

### U-NET

U-Net is a medical image segmentation model based on the Fully Convolutional Network (FCN). Its core idea is to achieve end-to-end pixel-level prediction through a symmetric encoder-decoder structure. The encoder is composed of repeated convolutional layers and max-pooling layers, gradually extracting high-level semantic features of the image while compressing spatial dimensions. The decoder progressively restores spatial resolution through transposed convolution or interpolation operations, and establishes skip connections with feature maps from corresponding encoder layers to fuse shallow detail information (such as edges and textures) with deep semantic information, thereby enhancing the precision of segmentation boundaries. The network exhibits a U-shaped symmetric structure overall, with the final layer outputting a segmentation mask matching the input image size through a 1×1 convolution. The innovation of U-Net lies in its design of an efficient contextual information transmission mechanism: Skip connections alleviate the loss of spatial information caused by downsampling, enabling high-precision segmentation even with limited annotated data (e.g., medical images), especially for objects with complex morphologies like cells and organs. The model is optimized using a cross-entropy loss function, and data augmentation techniques (such as elastic deformation) are typically integrated during training to enhance robustness.

### MRC - YOLOv8

For the demand of cement crack detection on a certain road in Hunan Province, Chen et al. proposed the MRC-YOLOv8 algorithm based on the improved YOLOv8 framework. The model architecture consists of three parts: data preprocessing module, feature extraction network, and detection head. In the preprocessing stage, mosaic enhancement technology, adaptive anchor box optimization, and grayscale adaptive filling method are adopted. The

feature extraction network innovatively uses a depthwise separable residual module to replace the standard C2f component, and improves the original C3 module by integrating the ELAN architecture to simplify the configuration of convolutional layers, optimize the gradient transfer efficiency, and enhance the utilization rate of the bottleneck structure, expanding the gradient feature capture capability while maintaining a compact architecture [7]. The spatial attention mechanism is introduced in the network design, and cross-scale feature interaction is realized through hierarchical deployment of extended residual modules and lightweight inverse residual modules: the multi-scale context fusion mechanism is used for high-level feature processing, and the calculation efficiency of small receptive fields is optimized for low-level feature extraction. The encoder-decoder architecture includes the trunk network, the low-level processing stage of lightweight inverse residual modules, and the high-level feature fusion stage of two extended residual modules [8]. Empirical evaluations demonstrate that the refined algorithm exhibits enhanced performance compared to the baseline framework, achieving statistically significant advancements in recognition precision while simultaneously reducing computational latency during implementation.

## U-Net

In terms of model architecture, researchers implemented structural optimization based on the classic U-Net framework: adjusting the original single-channel input to color image input (channel number D=3), and reducing the input size from 512×512 to 256×256. The encoder part adopts 3×3 edge padding convolution operations combined with ReLU activation functions, and the number of feature channels is expanded to 64 after the first layer of convolution. Feature dimensionality reduction is achieved through max pooling with a stride of 2, and the number of convolution channels doubles after each downsampling. It is worth noting that a regularization mechanism is introduced in the deep network structure, and Dropout layers are added in the last two downsampling stages to prevent overfitting. The decoding path realizes feature upsampling through 2×2 transposed convolution, performs skip connections with the features of the corresponding layers of the encoder, and completes pixel-level classification through 1×1 convolution combined with the Sigmoid function after two convolution operations [9].

To assess the model's performance, the team prepared a low-reinforcement ratio concrete specimen with a cross-sectional size of 200×150×203mm (length×width×height), and simulated the structural stress state through a four-point loading test. A multi-scale monitoring system was constructed by combining mobile intelligent terminals and UAV platforms to collect data on the evolution of structural surface damage. The improved U-Net network can effectively identify the crack area on the concrete surface, convert the original image into a damage distribution map, and extract crack geometric parameters (including extension length, expansion width, distribution area, and main direction angle) through morphological analysis. Experimental results show that this method can not only accurately quantify the characteristics of surface cracks but also effectively track the development path of structural damage through the analysis of damage mode evolution (the process of bending crack initiation and its transformation into shear cracks), providing an important basis for evaluating structural safety. The U-Net model can be used for early warning of surface cracks to prevent safety hazards. The extraction of crack direction and size can carry out integrity assessment before damage occurs, which is the theme of future research [10].

## RDD-YOLOv5

Jiang proposed an improved road defect detection framework, RDD-YOLOv5, which achieves efficient identification of road surface anomalies through multi-dimensional network optimization. The architectural design mainly introduces three innovations: (1) In the reconstruction of the backbone network, a self-attention calculation unit based on a movable window mechanism is adopted to replace the traditional convolutional structure, and the SW module is integrated in the minimum scale feature generation stage, effectively enhancing the model's ability to model global contextual information; (2) In the optimization of the feature fusion layer, an Enhanced Visual Feature Encoder (EVC) is introduced into the FPN+PAN architecture, realizing effective fusion of deep semantic information and shallow detail features through a cross-scale feature interaction mechanism; (3) In terms of improving nonlinear expression capabilities, the activation function of the benchmark model is replaced with the Gaussian Error Linear Unit (GELU), whose improved mathematical expression enhances the network's ability to characterize complex features. Experiments show that this adjustment slightly affects the training convergence speed but significantly improves the overall model performance [11].

Ablation experiments demonstrate significant positive synergistic effects among the improved modules. The jointly optimized model achieves a detection accuracy of 91.5% on the road defect dataset. In terms of network

compression, through parameter redistribution and computational graph optimization, the final model storage volume is controlled within 24MB, meeting the deployment requirements for embedded devices. To enhance practical application efficiency, the research team has also established a mathematical model of UAV aerial photography parameters and defect detection accuracy. Without increasing computational resources, optimizing the combination of flight altitude and speed parameters improves the system's comprehensive performance in dynamic detection scenarios by 17.6%. This method ensures detection accuracy while significantly improving the engineering applicability of the algorithm, providing an effective technical solution for intelligent road inspection.

## Improved YOLOv8

Zhu et al. proposed an enhanced YOLOv8 architecture to systematically improve the balance between feature extraction and computational efficiency in object detection tasks. In terms of the core module of the algorithm, three main innovations are implemented: (1) at the level of network structure optimization, the C2F module in the basic framework is reconstructed into a composite structure of local convolution and efficient multi-stage attention mechanism (EMA), which achieves network scale compression by reducing the redundancy of full convolution operations, and optimizes the interaction efficiency of feature channels using attention weights; (2) In terms of feature enhancement mechanism, intelligent agent attention units are deployed in the backbone network, combined with normalized exponential function and linear attention dual path calculation method, to achieve selective enhancement of key feature dimensions; (3) In terms of multi-scale modeling, a bimodal feature fusion module (BiFormer) is designed to implement spatial sparse sampling based on dynamic perception mechanism, and achieve collaborative representation of global context and local details through adaptive weight allocation strategy.

In the experimental stage, mainstream detection models such as FSF Net, SCGCN, and GS-YOLOv5 were selected as comparison benchmarks. The improved YOLOv8 achieved a detection accuracy of 93.57% and a recall rate of 88.51% on the standard test set. In terms of operational efficiency, the model achieved a real-time detection speed of 64.5 FPS, with optimized computation and parameter count to 10.81G FLOPs and 30.42M, respectively, reducing the original model by 25% and 19.7%. The ablation experiment showed that each improved module effectively improved the adaptability of the model to complex scenes while reducing the computational load by 14.3%. The optimal balance was achieved in the three dimensions of detection accuracy, inference speed, and hardware resource consumption, providing a feasible solution for embedded device deployment [12].

## Improved YOLOv3

The improved YOLOv3 model has undergone multiple optimizations based on the original YOLOv3 to better meet the needs of unmanned aerial vehicle inspection for defect detection of electrical circuit components. Firstly, to address the issue of insensitivity in small object detection, the K-means++ algorithm is adopted instead of the traditional K-means algorithm. By optimizing the selection of initial clustering centers, a more reasonable anchor box size is generated, thereby improving the localization ability for small-sized defect targets such as insulators. Secondly, to address the model bias caused by imbalanced positive and negative samples, the Focal Loss loss function is introduced to dynamically adjust the weights of difficult and easy samples, suppress the dominant role of easy to distinguish samples in training, and enhance the model's attention to difficult to distinguish samples (such as edge blur or occlusion defects). In addition, in terms of activation function, the original Leaky ReLU was replaced with the Mish activation function, which reduces information loss in feature extraction due to its smooth gradient characteristics and further improves the classification accuracy of the model. To further enhance the feature extraction capability, the model embeds the SENet attention mechanism in the Darknet53 backbone network, dynamically adjusts the weights of feature channels, highlights key feature information, and combines multi-scale detection (extended to four detection scales) to effectively alleviate the problem of small target feature loss in deep networks. Finally, by comprehensively optimizing the network structure and training strategy, the improved model achieved an average detection accuracy of 94.40% and a single frame detection speed of 0.079 seconds in unmanned aerial vehicle inspection scenarios. Compared with the original YOLOv3 and other mainstream methods such as Faster R-CNN and SSD, it showed significant advantages in accuracy and real-time performance, and can meet the actual needs of automatic defect detection for transmission lines [13].

## DISCUSSION

In the field of image recognition technology, every algorithm possesses distinct advantages while inevitably exhibiting shortcomings.

The MRC-YOLOv8 algorithm can perform multi-scale detection of target cracks, with small memory, faster frame rate, and training speed. However, its recognition accuracy for horizontal cracks, oblique cracks, and expansion joints is not sufficient, which is something that needs to be improved in the future for the MRC-YOLOv8 algorithm. The U-Net model can predict the future development trend of cracks, which is beneficial for effective prevention before accidents occur. However, the downside is that the prediction accuracy still needs to be improved. The RDD-YOLOv5 algorithm can recognize road defects with an accuracy of 91.5% and only occupies 24MB of space. However, the collective model of the improved YOLOv5 algorithm is too large and needs to be lightweight. The improved YOLOv8 algorithm with a Block structure, a balance between computational efficiency and accuracy, has been achieved through a lightweight design. However, its improvements in small object detection and reducing false alarm rates still cannot surpass YOLOv8. Nevertheless, the improved YOLOv3 model is highly reliable in small object detection and can be used to inspect small circuit devices.

## CONCLUSION

This article summarizes the increasingly important role of quadcopter drones in the inspection of urban rail transit infrastructure. According to existing research, drones equipped with high-resolution cameras and sensors can significantly improve detection efficiency and be able to enter dangerous or difficult to reach areas such as tunnel tops or elevated tracks. It is worth noting that AI based defect recognition algorithms, such as YOLOv5 and YOLOv8 used for crack detection, can achieve an accuracy of over 90%, indicating their potential to replace traditional manual inspections. RMC-VOLOv8, RDD-YOLOv5 and other image recognition algorithms can identify cracks during inspections, and using drones can more conveniently inspect inspection dead points such as tunnel tops. Improved YOLOv3 can be used for circuit device detection. Improved VOLOv8 can recognize flames and smoke, which can be detected and measures taken in a timely manner in case of a fire on the operating route. The U-Net algorithm can predict the future development trend of cracks, which can prevent accidents in a timely manner. Although drone recognition technology has made great progress, there are still some limitations: for example, it is difficult for drones to operate in environments without GPS signals, and they are susceptible to electromagnetic interference from the track power supply system, which can affect stability; Future research directions can lean towards the application of multimodal sensor fusion to unmanned aerial vehicles.

## REFERENCES

1. L. Zhang, S. Peng, Q. Feng et al., J. East China Jiaotong Univ. (2025).
2. Y. Chen, F. Luo, and W. Bai, Eng. Technol. Res. **5**(06), 11-12 (2020).
3. Y. Jiang, H. Yan, Y. Zhang, K. Wu, R. Liu, and C. Lin, Sensors **23**, 8241 (2023).
4. B. Lei, Y. Ren, N. Wang, L. Huo, and G. Song, Struct. Health Monit. **19**, 1871-1883 (2020).
5. H. J. Lim, S. Hwang, H. Kim, and H. Sohn, Struct. Health Monit. **20**, 3424-3435 (2021).
6. H. L. Ramandi, S. Irtza, T. Sirojan, A. Naman, R. Mathew, V. Sethu, and H. Roshan, J. Hydrol. **607**, 127482 (2022).
7. K. Y. Wong, YOLOv7 (2023). Available: https://github.com/WongKinYiu/yolov7
8. X. Chen, C. Wang, C. Liu, X. Zhu, Y. Zhang, T. Luo, and J. Zhang, Sensors **24**, 4751 (2024).
9. Ronneberger, P. Fischer, and T. Brox, in Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (Springer, Cham, Switzerland, 2015), pp. 234-241.
10. S. Bhowmick, S. Nagarajaiah, and A. Veeraraghavan, Sensors **20**, 6299 (2020).
11. Y. Jiang, H. Yan, Y. Zhang, K. Wu, R. Liu, and C. Lin, Sensors **23**, 8241 (2023).
12. W. Zhu, S. Niu, J. Yue, and Y. Zhou, Sci. Rep. **15**, 2399 (2025).
13. X. Ye, J. Sun, Y. Gan, Q. Ran, D. Wu, and Z. Lü, Electr. Meas. Instrum. **60**(5), 85-91 (2023).