Journal of Electronic Research and Application, 2025, Volume 9, Issue 6
http://ojs.bbwpublisher.com/index.php/JERA
ISSN Online: 2208-3510
ISSN Print: 2208-3502

Research on Optimization Algorithm for Video Recognition Accuracy in Smart Construction Sites

Xiang Gao[1,2], Lei Nie[1,2], Yinan Wang[1,2], Chi Zhang[1,2]

[1]POWERCHINA Hebei Electric Power Engineering Co., Ltd., Shijiazhuang 050031, Hebei, China
[2]Hebei Electric Power Design & Survey Technology Innovation Center, Shijiazhuang 050031, Hebei, China

**Author to whom correspondence should be addressed.*

Abstract: Aiming at the problems faced by construction site video management in the recognition of cigarette butts, reflective vests, and other objects, such as small target confusion, high-brightness false alarms, occlusion missed detections, and poor adaptability to complex environments, this study proposes a recognition accuracy optimization algorithm based on multimodal fusion. The research constructs a dataset containing three modalities of data: visible light, infrared, and millimeter-wave. The Dust-GAN algorithm is adopted to realize dust removal and enhancement of dusty images, and the SAA module is introduced into YOLOv8-s to improve the small target recall rate. Meanwhile, three-modal feature fusion is achieved, and channel pruning and quantization-aware training are used to realize algorithm lightweighting. The algorithm was deployed and operated on-site for 3 months, effectively reducing the construction site safety accident rate by 65%, which provides a solution for safety management and control in smart construction sites under complex environments.

Keywords: Smart construction site; Multimodal fusion; Edge computing; YOLOv8; SAA

Online publication: December 16, 2025

1. Introduction

Smart construction sites achieve full-lifecycle supervision of construction sites through front-mounted cameras and edge computing. However, in scenarios such as thermal power plants, power transmission and transformation, and petrochemical expansion, two typical target recognition algorithm scenarios still face severe challenges. One type of target is cigarette butts. Cigarette butt recognition is extremely challenging because the targets are often smaller than 20 pixels, their tiny size makes them frequently confused with welding slag, metal reflections, and dust light spots. Under infrared fill light at night, they appear as bright dots, with false alarm rates concentrated at over 20%. The other type of target is reflective vests. The bright strips on reflective vests produce specular reflections at different angles, making them easily misjudged as vehicle lights or warning barrels. Moreover, when workers bend over or the vests are blocked by safety belts, the miss rate reaches as high as 18%.

When the current YOLOv5/YOLOv8 series models migrate COCO pre-trained weights to construction site

scenarios, they exhibit low small target recall rates and high sensitivity to background noise. Furthermore, there is a lack of dedicated data for "dust-low illumination" conditions, leading to the coexistence of "false alarm fatigue" and "missed detection risk". Therefore, a set of algorithm systems tailored for complex environments is required to improve recognition accuracy.

2. Current status of video recognition algorithms at home and abroad

In the early days, both domestically and internationally, cigarette butts were detected based on color thresholding and morphological methods. The accuracy could reach 85% in simple backgrounds, but this approach was sensitive to changes in illumination. For reflective vest recognition, HSV color segmentation was relied on; however, the false alarm curve rose sharply as the vests became covered with dust.

After 2018, object detection algorithms such as Faster-RCNN, SSD, and YOLO were gradually applied to construction site safety helmet detection and intelligent cameras. In 2022, Jiang *et al.* proposed a smoking detection method based on YOLOv4-tiny, which achieved a mean Average Precision (mAP@0.5) of 86% [1]. Nevertheless, this method failed to resolve false alarms caused by high-brightness infrared spots at night. Wang *et al.* utilized the high-brightness feature of reflective strips and an attention mechanism, improving the mAP of vest recognition by 7%, yet missed detections still occurred in occluded scenarios [2].

Since 2022, visible-infrared fusion technology and visible-millimeter-wave radar fusion technology have become research hotspots. Infrared can penetrate smoke and dust, while millimeter-wave radar can measure distances to distinguish between metal reflections and fabrics. However, most existing methods remain limited to simulation environments or vehicle-mounted scenarios, and lack systematic verification under conditions specific to construction sites, such as dust, vibration, and complex backgrounds [3].

3. Related theories of multimodal AI

A tri-modal signal complementary model is established using visible light, infrared (IR) radiation, and millimeter waves (mmWaves) [4]. Among these, visible light provides color and texture details, and conducts comparative analysis of flame spectra. Infrared radiation is used to monitor abnormal temperatures and distinguish discrete thermal sources from environmental heat sources. The echo intensity of 77GHz radar is utilized to distinguish metallic luster from fabrics, while outputting the 3D coordinates of targets for occlusion reconstruction [5].

YOLOv8-s is a small-size object detection model featuring both high speed and high accuracy, it is the small-size version in the YOLOv8 model series [6]. When images are resized to 640×640 pixels and input into the YOLOv8-s model, under the strict criterion of "overlap between prediction boxes and ground truth boxes ≥ 50%", it can correctly detect an average of 41.2% of targets (mAP@0.5 = 41.2%). Moreover, in an environment equipped with an NVIDIA T4 graphics card and TensorRT acceleration, it only takes 2.5 milliseconds to process one image, which is equivalent to handling 400 images per second.

Cross-modal transformer (CMT) fuses sensor data (e.g., visible light, infrared radiation, and millimeter waves), unifies the data representation, and is used to achieve cross-modal alignment and fusion at the feature layer. First, it encodes the raw data of each modality into tokens carrying 3D positional information. It then calculates Q/K/V (the three core matrices in the Transformer attention mechanism) among these tokens [7]. Following that, it dynamically allocates weights, filters redundant information, and retains complementary features. Finally, it fuses the unified features and feeds them into downstream monitoring models, such as YOLOv8.

Channel pruning involves checking the γ coefficient (also known as the scaling factor) of the Batch Normalization (BN) layer in each convolutional layer; the smaller the γ value, the more insignificant the channel is [8]. Specifically, 30% of these "insignificant channels" are pruned. Another approach under channel pruning calculates the gradient contribution (sensitivity) of each layer to the final output, and directly prunes 20% of the layers with low contribution. This technique reduces the model size without compromising its accuracy.

4. Data collection and annotation

Images are captured using front-mounted cameras at the smart construction site, with continuous collection day and night for 20 days. Meanwhile, millimeter-wave radars are deployed. For small targets like cigarette butts, pre-annotated cigarette butts or reflective vests are extracted and pasted onto construction site background images with different positions, scales, and lighting conditions. Then, semi-transparent dust textures or masks are overlaid to simulate dust, haze, and lens smudges at the construction site, enabling the training model to recognize targets in polluted environments [9].

For reflective vests, the model is imported, and different perspectives and light sources (sunlight, vehicle lights, electric welding arc lights) are set to render extreme reflection effects such as high brightness, glare, and stripe flicker. A large number of high-brightness and dazzling vest images are obtained to prevent reflective vests from being misjudged as warning lights or vehicle lights in real scenarios [10]. Lastly, the number of images is expanded to 24,000, and the Smoky-Vest dataset is constructed. This dataset contains 39,000 bounding boxes for cigarette butts and 72,000 bounding boxes for reflective vests, with the class imbalance degree reduced from 12.4 to 2.7. This avoids significant data differences during the next stage of model training and improves the recognition effect.

When annotating photos and images, two bounding boxes are used to define the annotation rules for cigarette butts, where the "flame box" encloses the open flame or glowing area, and the "cigarette stick box" encloses the main body of the cigarette (annotation is still required even if there is no open flame). An annotation is considered correct only when the overlap rate between the predicted box and the manually drawn box exceeds 70%. Similarly, separate bounding boxes are used to define the annotation rules for reflective vests: the overall outline of the garment is defined as the "vest main body box", and the fluorescent strips are defined as the "reflective strip boxes". If a reflective vest is occluded by more than 50%, it is labeled with a "hard example" tag, and a lower weight is assigned during training to avoid misleading the model.

5. Model improvement and lightweight design

To optimize image quality in dusty scenarios, the Dust-GAN model algorithm is defined, and the model's adaptive dust removal capability is enhanced through algorithm improvements [11]. This method adopts a cyclic generative adversarial network (cyclic GAN) as its basic framework, converting dusty visible light images into clear ones to provide high-quality input for subsequent detection networks. During the model training phase, a dataset of 50,000 pairs of "dusty-clean" images is used for training. Experimental results show that the Peak Signal-to-Noise Ratio (PSNR) of the processed images is increased by 4.3 dB, and the false alarm rate in cigarette butt detection tasks is reduced by 9.8%, effectively improving image quality and detection reliability in complex scenarios.

Aiming at the low detection accuracy of small targets on construction sites (such as cigarette butts and fallen safety helmets), the Scale-Aware Attention (SAA) module is introduced into the Neck section of YOLOv8 [12].

Through deformable convolution, this module adaptively adjusts feature maps of different scales. Leveraging the minimum detection capability of the YOLOv8-s model, an attention module is inserted before the 160×160 detection head. This enables the model to learn which channels are more important when predicting small targets, amplify important features, suppress useless ones, and strengthen the edge and texture features of small targets such as cigarette butts and reflective vests. Experimental verification shows that the introduction of the SAA module significantly increases the small target recall rate from 78.2% to 91.4%, effectively improving the detection performance of small targets [13].

To further enhance the consistency and accuracy of multimodal detection, a cross-modal consistency verification mechanism is adopted. Using the 3D coordinate information provided by the millimeter-wave sensor, IoU-3D verification is performed on the detection results of visible light and infrared images (where 0 indicates no overlap between the prediction and the ground truth, and 1 indicates complete overlap of the two boxes) [14]. When the IoU-3D value of the visible light and infrared detection boxes at the same spatial position is less than 0.3, the "consistency recheck" module is triggered, and the detection results are re-scored by fusing tri-modal features. This method effectively reduces isolated false alarms by 42% and enhances the stability of the multimodal detection system.

For edge-side deployment, the model is lightweighted through the following measures:

(1) Channel pruning: Based on L1 regularization, the convolutional channels of the YOLOv8 backbone network are evaluated, with a channel pruning rate of 30%;

(2) Quantization-aware training: The model weights are quantized from 32-bit floating-point to 8-bit integer, and quantization errors are simulated during training to ensure controllable accuracy loss;

After quantization, the model storage size is reduced from 102MB to 26MB, and the inference speed is increased by 3 times. The lightweighted model (Dust-GAN + improved YOLOv8 + cross-modal verification) has a parameter count of 23.5M and a computational complexity (FLOPs) of 18.6G, meeting the computational power requirements of edge devices [15] .

6. Conclusion

The experiment was deployed at a smart construction site, with the deployment platform being an edge computing terminal NVIDIA Jetson AGX Orin (equipped with an 8-core ARM CPU, a 2048-core GPU, and 32GB of memory). It was connected to 8 monitoring devices (4 visible light channels + 2 infrared channels + 2 millimeter-wave channels) to achieve all-weather real-time monitoring.

Under normal weather conditions, the mean average precision (mAP) of the proposed algorithm reaches 90.2%, which is an increase of 5.3% compared with YOLOv8. In dusty weather, due to the dust removal and enhancement of Dust-GAN, the mAP reaches 85.7%, representing a maximum increase of 18.9% compared with the comparison methods. In nighttime scenarios, multi-modal fusion enables the mAP to reach 88.1%, which is significantly higher than that of single-modal visible light (65.3%). In terms of small target recall rate, the proposed algorithm achieves 91.4%, an increase of 22.1% compared with Faster R-CNN.

The system underwent a 3-month trial operation, during which a total of 236 unsafe behaviors were detected, including 102 cases of not wearing safety helmets, 48 cases of unauthorized smoking, and 37 cases of not wearing reflective vests. The alarm accuracy rate reached 92.8%, and false alarms were mainly caused by equipment interference under severe weather conditions. This verifies the practical application value of the proposed algorithm.

Disclosure statement

The authors declare no conflict of interest.

References

[1] Jiang X, Wang B, Xia Y, et al., 2022, Smoking Behavior Detection Based on Human Key Points and YOLOv4. Journal of Shaanxi Normal University (Natural Science Edition), 50(3): 96–103.

[2] Wang D, Bai C, Wu K, 2021, Review of Video Object Detection Based on Deep Learning. Journal of Computer Science and Exploration, 2021: 1–15.

[3] Varghese R, Sambath M, 2024, YOLOv8: A Novel Object Detection Algorithm with Enhanced Performance and Robustness. International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS), 2024.

[4] Chen S, Ma H, Wang T, et al., 2022, Video Sentiment Analysis Technology Based on Multimodal Fusion. Journal of Chengdu University of Information Technology, 2022(6): 656–661.

[5] Guo N, Jiang L, 2021, Processing of Multimodal Video Captions Based on Hard Attention Mechanism. Application Research of Computers, 38(3): 956–960.

[6] Pan W, Wei C, Qian C, et al., 2024, Improved YOLOv8s Model for Small Object Detection from UAV Perspective. Computer Engineering and Applications, 60(9): 142–150.

[7] Wang Y, Li M, Sun H, 2024, External Knowledge-Based VQA Integrating Cross-Modal Transformer. Science Technology and Engineering, 24(20): 8577–8586.

[8] He Y, Zhang X, Sun J, 2017, Channel Pruning for Accelerating Very Deep Neural Networks. Proceedings of the IEEE International Conference on Computer Vision, 2017: 1389–1397.

[9] Li H, Wang L, Zhang J, 2023, Research on Multimodal Data Acquisition and Synchronization System for Smart Construction Sites. Automation & Instrumentation, 2023(8): 145–149.

[10] Zhao Y, Wang T, Li T, 2021, Image Rendering and Data Augmentation Technology for Reflective Vests in Complex Lighting Environments. Journal of Graphics, 42(5): 825–832.

[11] Chen J, 2019, Design and Research of Dust Concentration Detection Based on Image Method, thesis, China Jiliang University.

[12] Zhang L, Tian Y, 2024, Multi-Scale Lightweight Vehicle Object Detection Algorithm Based on Improved YOLOv8. Computer Engineering and Applications, 60(3): 129–137.

[13] Yue M, Shu K, Zhang C, et al., 2024, Research on Infrared Small Target Detection Algorithm Based on Improved YOLOv8. Infrared Technology, 2024(11): 1286–1292.

[14] Ju R, Chien C, Chiang J, 2024, YOLOv8-ResCBAM: YOLOv8 Based on an Effective Attention Module for Pediatric Wrist Fracture Detection. arXiv. https://doi.org/10.48550/arXiv.2409.18826

[15] Wang M, Yao G, Yang Y, et al., 2023, Deep Learning-Based Object Detection for Visible Dust and Prevention Measures on Construction Sites. Developments in the Built Environment, 2023: 16.