International Journal of Innovative Computing, Information and Control
Volume **22**, Number **1**, February **2026**

 ISSN 1349-4198
pp. **193–207**

SURFACE DIRT DETECTION ALGORITHM OF PHOTOVOLTAIC PANEL BASED ON LIGHTWEIGHT RT-DETR MODEL

Wei Liu[1], Yun Sun[1], Yuchen Zhang[2,*], Hongwei Chang[1], Xiaochen Wang[1] and Xinnan Fan[3]

[1]Shandong Electric Power Engineering Consulting Institute Corp., Ltd.
No. 106, Minziqian Road, Jinan 250013, P. R. China
{ liuwei29; sunyun; changhongwei; wangxiaochen }@spic.com.cn

[2]College of Information Science and Engineering
[3]Jiangsu Provincial Key Laboratory of Power Transmission and Distribution Equipment Technology
Hohai University
No. 1915, Hohai Avenue, Jintan District, Changzhou 213200, P. R. China
fanxn@hhu.edu.cn
*Corresponding author: 231323010017@hhu.edu.cn

Received April 2025; revised July 2025

Abstract. *Real-time detection of surface dirt accumulation on photovoltaic (PV) panels is crucial for enhancing power generation efficiency. However, existing deep learning models face challenges in practical deployment due to high computational complexity, substantial implementation costs, and suboptimal real-time performance. This paper proposes a lightweight-enhanced RT-DETR framework (LRT-DETR) specifically optimized for PV dirt detection tasks. First, the StarNet backbone network is adopted to reduce redundant computations by 43% through streamlined architectural design and accelerate the computing speed. Second, a multi-scale multi-head self-attention module is developed to enable parallel attention mechanisms across distinct feature subspaces, thereby enhancing model expressiveness and adaptability to complex scenarios involving occlusions and illumination variations. Finally, to improve small target detection accuracy, a GSConv module is designed with three-stage optimizations: channel compression, depthwise convolution, and shuffle reorganization. It is a more light-weight module with a better performance compared with normal convolution. Experimental validation on a self-constructed PV dirt accumulation dataset demonstrates that the proposed algorithm significantly enhances detection performance, achieving a mean average precision (mAP) of 77.2% with an improvement of +5.9% over baseline, while simultaneously reducing model parameters by 43.6% (to only 11.4 M) and increasing inference speed to 27 frames per second (FPS). This combination of high-precision, lightweight architecture, and real-time capability provides a highly effective and cost-efficient intelligent visual inspection solution for PV power station operation and maintenance. The framework exhibits extensibility potential for other resource-constrained industrial inspection scenarios.*

Keywords: RT-DETR, Multi-scale features, Self-attention, Small target detection

1. **Introduction.** Against the backdrop of the global energy structure's low-carbon transition, photovoltaic (PV) power generation has emerged as a core pillar of renewable energy systems. In the first quarter of 2025, China's installed capacity of wind and solar power reached 1.482 billion kilowatts, historically surpassing thermal power capacity and marking a substantial shift toward a clean energy-dominant power system. By early 2025, China's cumulative PV installed capacity exceeded 900 gigawatts (GW), with 39.47 GW added in the first two months alone. However, PV systems operating long-term outdoors

DOI: 10.24507/ijicic.22.01.193

face severe surface contamination challenges – deposits such as dust, bird droppings, and snow significantly attenuate light absorption efficiency [1]. Empirical studies indicate that contamination can reduce PV system power generation efficiency by 20%-30%, resulting in staggering economic losses. According to the International Energy Agency (IEA), global annual revenue losses due to PV module contamination have exceeded 3.2 billion and continue to worsen, projected 6-10 billion by 2025 [2].

This efficiency loss directly erodes power plant economics. Contamination-induced power loss equates to reduced asset utilization, significantly extending the investment payback period. Considering the levelized cost of electricity (LCOE) amplifies this impact: while China's PV LCOE has fallen below ¥0.30/kWh, efficiency losses from contamination can increase actual LCOE by ¥0.06-0.09/kWh, substantially undermining PV's cost advantage over coal power.

While current deep learning-based detection models (e.g., YOLO series [3-7], and Faster R-CNN [8]) demonstrate exceptional performance in general object detection tasks, they face unique challenges when applied to photovoltaic panel contamination detection. Firstly, contaminants exhibit highly irregular morphologies and low contrast against panel backgrounds (e.g., light-colored stains on transparent glass surfaces), demanding models with enhanced capabilities in fine-grained feature extraction. Secondly, as PV power stations are typically deployed in remote areas, detection algorithms must operate in real time on low-computational-power edge devices. However, the high computational complexity and parameter volume of existing models hinder their practical deployment in such scenarios [9].

In 2020, the FaceBook team proposed the DETR (Detection Transformer) model [10], which achieved end-to-end object detection through global attention mechanisms. However, its high computational complexity limits real-time performance. In 2021, Wang et al. introduced Deformable DETR [11], which optimized the attention module by focusing on key points, thereby improving accuracy while reducing training iterations. In 2024, the Baidu PaddlePaddle team proposed the RT-DETR (Real-Time DETR) model [12], which eliminated non-maximum suppression (NMS) to reduce computational costs and enhance real-time detection capabilities. Although RT-DETR accelerates convergence through cross-scale feature interaction and query denoising, its model size and computational demands remain challenging for direct deployment on edge devices.

To address these limitations, this paper proposes a lightweight RT-DETR variant tailored for photovoltaic panel contamination detection. The framework aims to enhance inference speed without compromising accuracy and incorporates a multi-scale feature fusion module to improve detection precision for small targets. Compared to mainstream real-time detection models, the proposed approach demonstrates superior performance in balancing efficiency and accuracy, offering a viable solution for resource-constrained industrial applications.

2. **Lightweight RT-DETR Network Architecture for Photovoltaic Panel Surface Contamination Detection.**

2.1. **RT-DETR model.**

The RT-DETR model adopts a Transformer-based end-to-end real-time detection architecture, which ensures accuracy and real-time performance by processing high-dimensional features and integrating multi-scale feature fusion while significantly reducing computational complexity. The primary components of the RT-DETR framework include a backbone network, an efficient hybrid encoder, a decoder, and a prediction head. The overall architecture is illustrated in Figure 1.

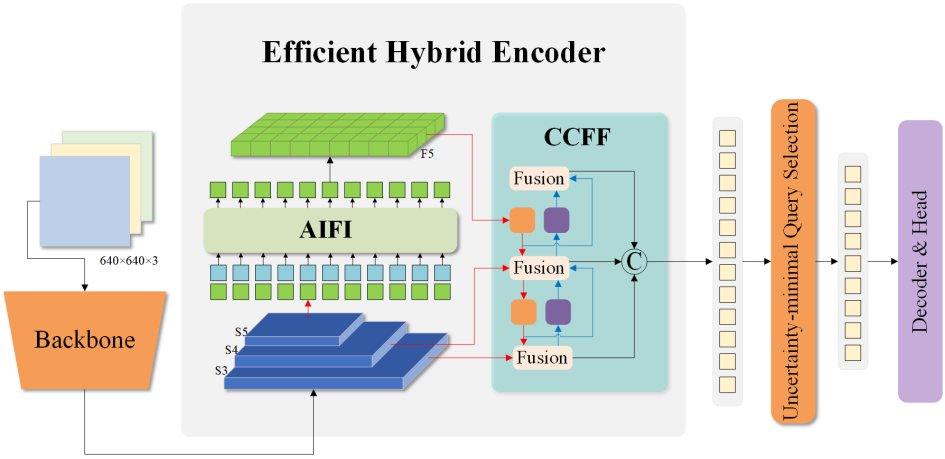


FIGURE 1. RT-DETR model structure

The RT-DETR model employs the classical deep residual network ResNet as its backbone network [13]. The last three output layers of the backbone (S3, S4, S5) retain the majority of extracted features, which serve as inputs to the efficient hybrid encoder. This encoder comprises two key modules: the attention-based intra-scale feature interaction (AIFI) module and the CNN-based cross-scale feature fusion (CCFF) module. Specifically, the S5 layer is exclusively fed into the AIFI module to learn deep-level features, while its output (F5), combined with the S3 and S4 layers, is processed by the CCFF module to enable multi-scale feature learning.

Subsequently, an uncertainty-aware minimal query selection mechanism is applied to selecting a fixed number of encoder features as the initial object queries for the decoder. Finally, the decoder, equipped with auxiliary prediction heads, iteratively refines these object queries to generate class labels and bounding boxes. This streamlined architecture ensures computational efficiency while maintaining robust feature representation capabilities, making it particularly suitable for edge-device deployment in photovoltaic inspection scenarios.

Photovoltaic power plants typically present challenging environments with complex lighting conditions and noise interference, which can significantly compromise the accuracy of machine vision detection. Additionally, surface contaminants on photovoltaic panels represent small targets that exhibit substantial similarity to the background panel surfaces, thereby increasing detection difficulty. The reflective properties of photovoltaic panels further introduce interference that adversely affects detection precision. To address these challenges, this paper proposes enhancements through backbone network modification and the integration of multi-head attention mechanisms, effectively improving the accuracy and stability of surface contamination detection on photovoltaic panels without increasing additional parameters or memory overhead. The specific improvements are as follows.

1) To overcome the limitations of low real-time performance and detection efficiency in photovoltaic panel contamination detection, we implement targeted adjustments to the backbone network. This optimization reduces model parameters while decreasing reliance on precise predictions, thereby enhancing robustness against noise, occlusion, and imperfect inputs. Both detection accuracy and speed demonstrate significant improvement.

2) For the critical challenge of detecting high-similarity contaminants against photovoltaic panel backgrounds [14], we introduce a multi-scale multi-head self-attention module within the hybrid encoder. This innovation achieves deep integration between the

multi-scale characteristics of CNN and the attention mechanism of Transformer. By employing dilated convolutions to construct a pyramidal receptive field [15], we implement low-resolution key-value (KV) pairs to reduce computational load while maintaining high-resolution queries (Q) for detail preservation. This architecture enhances the network's understanding of spatial relationships within input data, thereby improving model accuracy and robustness. The multi-head attention mechanism enables parallel processing of different feature subspaces, strengthening feature representation capability and adaptability to complex scenarios involving occlusion and illumination variations.

3) For detecting small targets such as bird droppings and dust particles, we incorporate a GSConv module [16] implementing three-stage optimization: channel compression, depthwise convolution, and shuffle reorganization. The depthwise separable convolution decomposes standard convolution into depthwise (channel-wise) and pointwise (1×1) convolutions, theoretically reducing computational parameters to 1/8-1/9 of standard convolution [17]. The channel compression strategy employs 1×1 convolution to halve channel dimensions ($c^2/2$), reducing subsequent operation parameters by 50% while preserving essential feature information. Finally, parameter-free channel shuffling enables cross-channel interaction through shuffle operations [18], replacing traditional 1×1 convolution-based channel reorganization and saving approximately (c^2) parameters. This approach achieves substantial computational resource savings while maintaining feature quality through enhanced cross-channel interaction, thereby improving recognition accuracy for small targets.

Compared with RT-DETR and Deformable DETR, our model focuses more on faster detection speed and the ability to recognize small targets. To achieve a faster detection speed, compared with the ResNet backbone network of RT-DETR and Deformable DETR, we designed a more lightweight backbone network. To enhance the recognition ability of small targets, we replaced the ordinary convolution in the multi-scale feature extraction module with GSConv, which reduces the number of parameters and computational cost while improving the recognition ability of small targets. After these improvements, our model has a lower computational cost and is suitable for deployment on edge terminals with limited computing resources. Compared with the YOLOv10/11 series, we designed a multi-scale multi-head attention feature extraction module to make up for the inability of traditional CNN architectures to focus on global information.

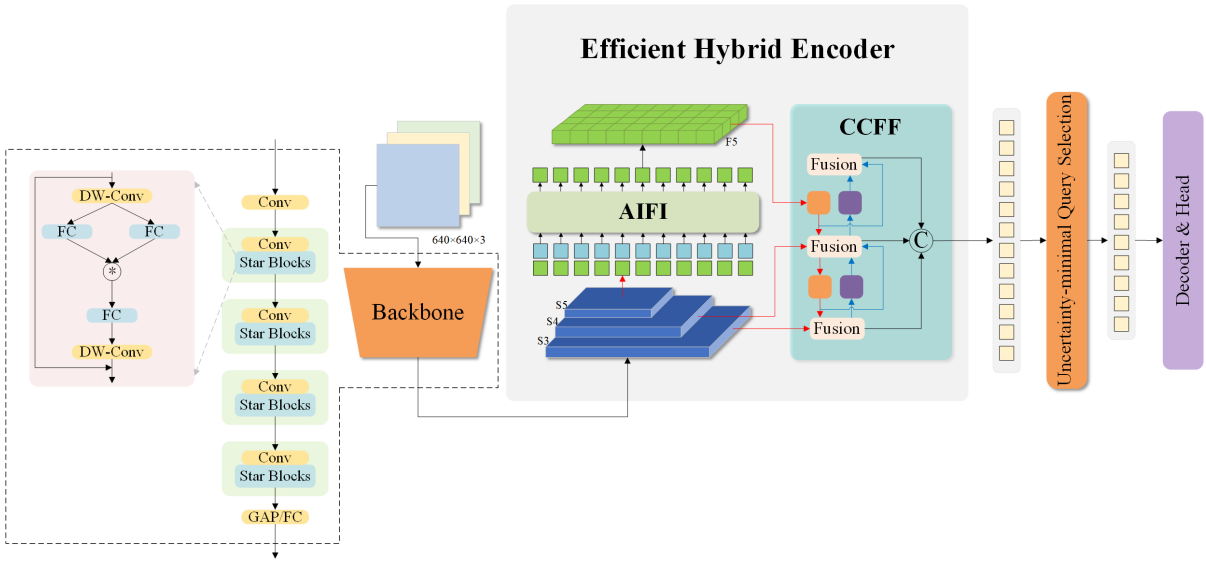


FIGURE 2. Structure diagram of the improved algorithm model

2.2. **StarNet framework.** To further enhance the performance of the algorithm while reducing computational complexity, we adopt the StarNet architecture [19] as a replacement for the original ResNet framework, as formalized in Equation (1). The StarNet introduces star operation defined by Equations (2) and (3), which achieves parameter consolidation by redefining the conventional weight matrix and bias term as a unified parametric entity. This operational innovation is realized through dimensional expansion of the input vector while maintaining mathematical equivalence, thereby enabling efficient feature transformation through the proposed star operation.

$$\omega_1^{\mathrm{T}} x * \omega_2^{\mathrm{T}} x = \left(\sum_{i=1}^{d+1} \omega_1^i x^i \right) * \left(\sum_{j=1}^{d+1} \omega_2^j x^j \right) = \sum_{i=1}^{d+1} \sum_{j=1}^{d+1} \omega_1^i \omega_2^j x^i x^j \tag{1}$$

$$\omega = \begin{bmatrix} W \\ B \end{bmatrix} \tag{2}$$

$$x = \begin{bmatrix} X \\ 1 \end{bmatrix} \tag{3}$$

The specific network architecture of StarNet is illustrated in Figure 3. StarNet is divided into four identical stages, each consisting of a convolutional layer followed by a Star Module. Within the Star Module, the input undergoes depthwise convolution and subsequently branches into two pathways: one pathway passes through a fully connected layer followed by activation function processing, while the other undergoes normalization before being processed by another fully connected layer. The outputs from both pathways are combined through Star Operation, and then passed through an additional fully connected layer and normalization. The integrated features undergo depthwise convolution again before being output via residual connection.

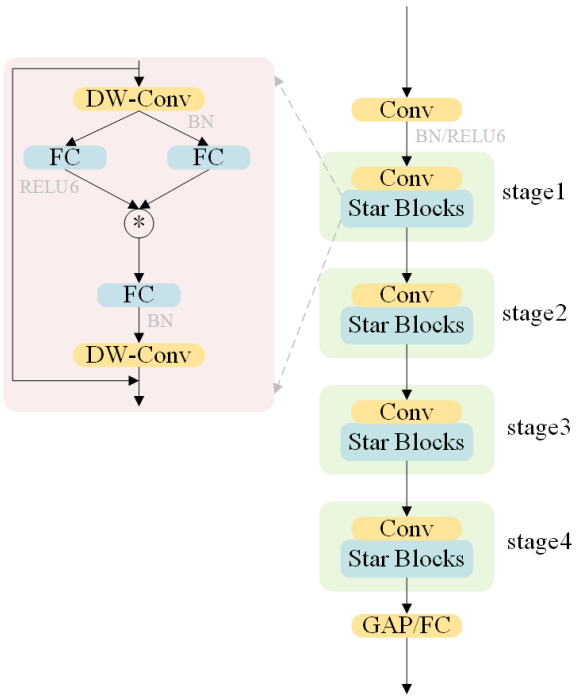


FIGURE 3. StarNet framework

Compared with ResNet, StarNet exhibits significantly smaller network scale and shallower depth, substantially reducing convolutional computations. As demonstrated in Equation (4), the hidden dimension of Star Operation follows an exponential scaling

strategy. This design enables simple stacking of network stages to achieve effects comparable to deeper networks. Consequently, StarNet not only reduces computational costs but also maintains or even enhances performance relative to conventional architectures.

$$\begin{aligned} S_1 &= \sum_{i=1}^{d+1}\sum_{j=1}^{d+1} \omega_{(1,1)}^{i}\omega_{(1,2)}^{j} x^i x^j \\ S_2 &= \mathrm{W}_{2,1}^{\mathrm{T}} S_1 \times \mathrm{W}_{2,2}^{\mathrm{T}} S_1 \\ S_3 &= \mathrm{W}_{3,1}^{\mathrm{T}} S_2 \times \mathrm{W}_{3,2}^{\mathrm{T}} S_2 \\ &\vdots \\ S_n &= \mathrm{W}_{n,1}^{\mathrm{T}} S_{n-1} \times \mathrm{W}_{n,2}^{\mathrm{T}} S_{n-1} \end{aligned} \tag{4}$$

2.3. **AIFI-MSMHSA module.** This work enhances the original Adaptive Intra-scale Feature Interaction (AIFI) module by modifying its intra-scale attention-based feature interaction with a Multi-Scale Multi-Head Self-Attention (Multiscal_MHSA) mechanism [20]. The detailed architecture of the proposed module is illustrated in Figure 4, while Figure 5 provides a schematic of the Multiscal_MHSA block. Traditional self-attention mechanisms may struggle to effectively handle objects of varying sizes, as a single attention head often fails to adequately model multimodal feature relationships in complex scenes. In contrast, the Multiscal_MHSA module introduces a multi-scale mechanism to

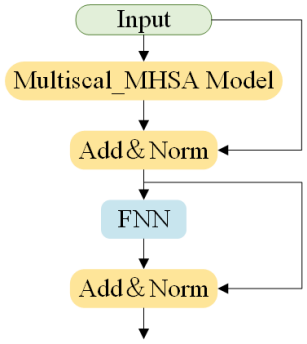


FIGURE 4. Structure diagram of AIFI-MSMHSA module

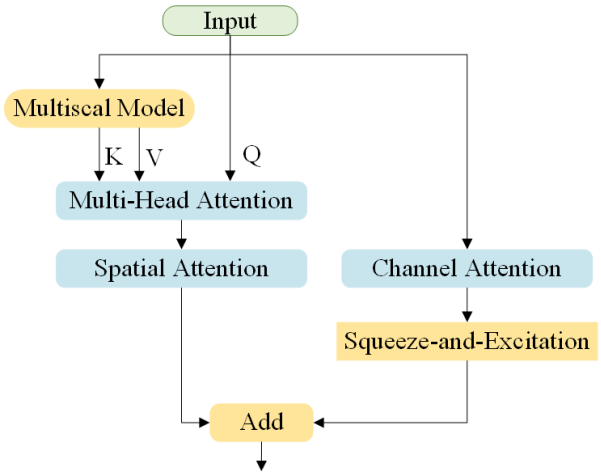


FIGURE 5. Structure diagram of Multi-Scale Multi-Head Self-Attention Module (Multiscal_MHSA Model)

capture hierarchical features. Specifically, the input features are projected into three sets of vectors (Q, K, V), with different attention heads focusing on feature interactions at distinct scales. This design enables the model to simultaneously attend to fine-grained details (e.g., edges), local structures (e.g., object parts), and global context (e.g., scene semantics). Additionally, the multi-head attention mechanism allows the model to focus on diverse feature subspaces in parallel, enhancing its representational capacity and adaptability to challenging scenarios such as occlusions and illumination variations. Crucially, the Multiscal_MHSA module reduces computational redundancy by computing attention within grouped feature scales rather than globally, thereby lowering computational complexity and improving real-time performance which is a critical requirement for real-time object detection systems.

As illustrated in Figure 6, the Multi-Scale Feature Extraction Module (Multiscal Model) processes input features through three parallel dilated convolutional pathways, followed by residual connection, normalization, and adaptive pooling operations to generate the final output. Compared to standard convolutions where each element only perceives local regions of the input feature map, dilated convolutions efficiently expand the receptive field by inserting "holes" (spaced pixels) between convolutional kernel elements. This enables coverage of broader contextual regions with minimal parameter overhead while capturing large-scale contextual information.

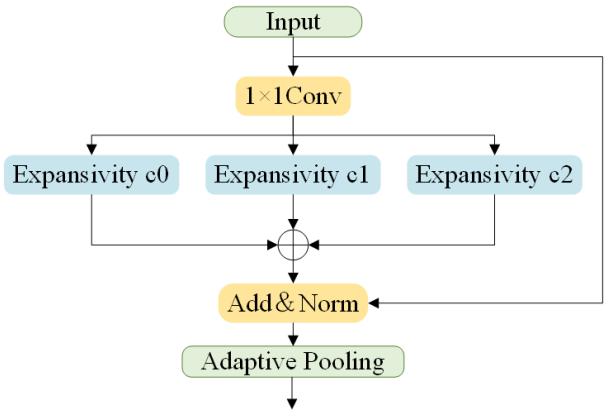


FIGURE 6. Structure diagram of Multi-Scale Feature Extraction Module (Multiscal Model)

The design inherently incorporates flexible multi-scale adaptability: By adjusting combinations of dilation rates, the module can be tailored to diverse task requirements. The three-branch architecture synergizes multi-granularity feature learning through hierarchical receptive fields, balancing local detail preservation and global context modeling. This parameter-efficient strategy avoids the computational burden of cascaded convolutions while achieving comparable or superior multi-scale representation capabilities.

2.4. **CCFF-GS module.** The original cross-scale feature fusion (CCFF) module employs 1×1 and 3×3 convolutions. As shown in Figure 7, to enhance computational efficiency and improve multi-scale feature fusion, we replace the 3×3 convolution with GSConv, a lightweight operator optimized through a three-stage refinement process: channel compression, depthwise convolution, and parameter-free channel shuffling. Compared to standard convolution, GSConv first applies depthwise separable convolution, decomposing the operation into depthwise (channel-wise) and pointwise (1×1) convolutions. This reduces computational complexity by theoretically lowering parameter counts to 1/8-1/9

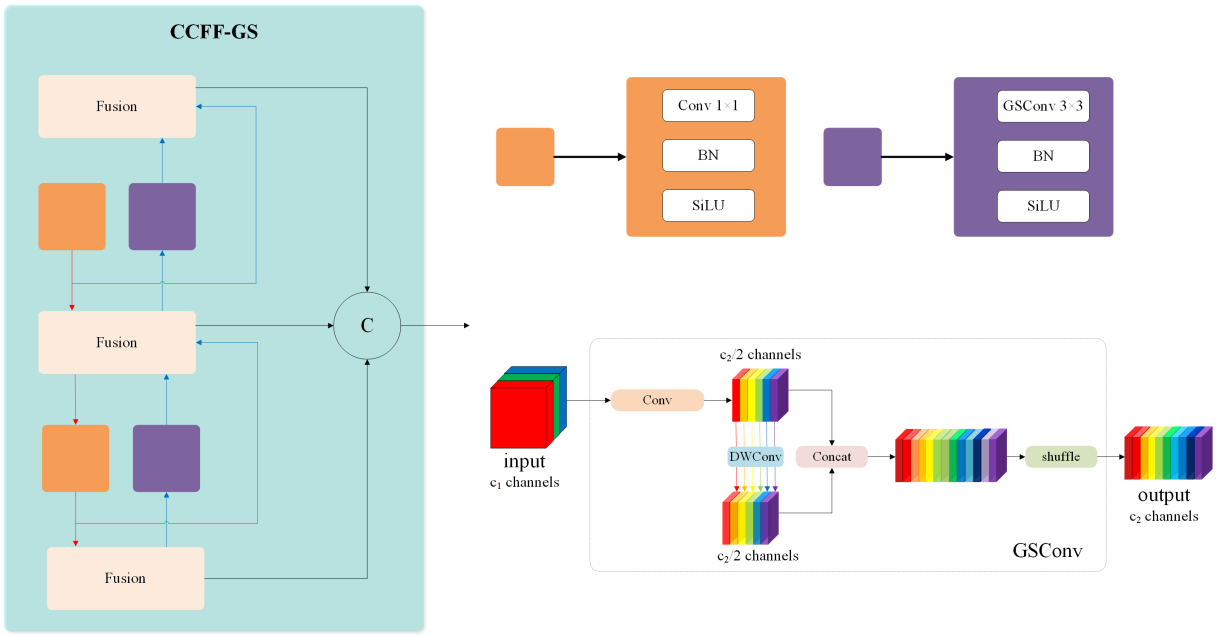


FIGURE 7. Structure diagram of CCFF-GS module

of standard convolution. Next, a channel compression strategy halves the channel dimension to $(c_2/2)$ via 1×1 convolution, directly reducing subsequent operation parameters by 50% while preserving critical feature information. Finally, a parameter-free channel shuffle mechanism enables cross-channel interaction through spatial rearrangement, replacing the channel recombination function of traditional 1×1 convolutions and saving approximately $O(c_2^2)$ parameter overhead. This design achieves significant computational savings while maintaining feature quality. By combining standard convolution (SC) and depthwise convolution (DWC) and using the shuffle operation for mixing, GSconv can better integrate feature information at different scales. This is particularly important for small object detection, as small objects usually retain detailed information only in the shallower feature maps. The branch design in the GSConv structure can generate multiple receptive fields of different sizes. Small targets usually require smaller receptive fields to capture local details, and GSConv precisely meets this requirement. By explicitly enhancing cross-channel interaction, the module strengthens feature representation for small objects, improving detection accuracy for challenging targets. The hierarchical optimization balances efficiency and performance, making it particularly suitable for resource-constrained scenarios requiring real-time inference.

3. **Evaluation Metrics.** For the photovoltaic panel surface contamination detection task, this study adopts average precision (AP), parameter count (Params), and frames per second (FPS) as evaluation metrics to assess algorithm performance.

Average precision (AP) is a widely used metric in object detection to quantify model accuracy across different target categories. A higher AP value indicates superior performance in detecting targets of the corresponding category. Confidence scores and ground-truth labels are assigned to predicted bounding boxes. Predictions are ranked by confidence scores, and true/false positives are determined using an Intersection-over-Union (IoU) threshold. Precision and Recall are computed as defined in Equations (5) and (6):

$$\text{Precision} = \frac{TP}{TP + FP} \tag{5}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{6}$$

where TP (true positives) denotes correctly detected targets, FP (false positives) represents incorrect detections, and FN (false negatives) indicates missed ground-truth targets. AP is derived as the area under the Precision-Recall curve, as formulated in Equation (7):

$$\text{AP} = \int_0^1 p(r)dr \tag{7}$$

where p and r denote Precision and Recall, respectively. Parameter count (Params) measures the total number of trainable parameters (e.g., weights, and biases) in the model. This metric reflects model complexity and scale, guiding decisions in model design and selection. Higher parameter counts typically imply greater computational and memory requirements. FPS (frames per second) quantifies real-time performance by calculating the number of images processed per second. It is computed as: FPS = 1/processing time per frame.

Higher FPS values indicate faster inference speeds, critical for real-time applications. Together, these metrics holistically evaluate detection accuracy, computational efficiency, and deployment feasibility.

3.1. **Dataset preparation and experiment environment.** This paper adopts a combination of online collection and self-production to construct a dataset of dirt on photovoltaic panel surfaces. The entire dataset mainly includes two types of samples: dust and bird droppings. These samples were manually labeled using Labelme, providing accurate bounding boxes and category labels for each sample. Subsequently, data augmentation techniques such as image cropping, rotation, noise addition, and contrast adjustment were applied to the dataset images to enhance its diversity. Through carefully designed data augmentation methods, the dataset can cover various situations and scenarios in the detection of substation instruments, making it suitable for model training, validation, and testing. The expanded dataset contains a total of 900 images. The dataset is first divided into training data and test set in an 8 : 1 ratio, and then within the training data, it is further subdivided into training set and validation set in a 9 : 1 ratio. This division ensures that we have sufficient data for training and validating the model's performance, while also reserving a portion of the data as a test set to evaluate the model's performance on unseen data.

3.2. **Ablation experiment.** To validate the effectiveness of the lightweight improvement strategies proposed in this study, ablation experiments were conducted on the self-constructed photovoltaic dirt dataset to evaluate the contributions of individual model modules. The baseline model was the original RT-DETR (ResNet18), allowing for a comparative analysis of the effects of different improvement strategies and verifying the efficacy of the three proposed modules. The StarNet backbone network, AIFI-MSMHSA module, and CCFF-GS module are denoted as A, B, and C, respectively. For each enhancement, the impact on detection speed and accuracy of the RT-DETR model was assessed on the custom dataset. All experiments were uniformly configured with an input resolution of 640×640, and the results are presented in Table 1.

As shown in the table, Model A – which replaces the backbone network with StarNet – achieves a 3.8% improvement in accuracy, reduces parameters by 7.9 M, and increases FPS by 26 frames compared to the baseline. This demonstrates that StarNet outperforms ResNet in maintaining high precision with lower computational costs, making it more suitable for photovoltaic panel surface dirt detection tasks.

TABLE 1. Ablation experiment results

Model	mAP_{50}/%	Params (M)	GFLOPs	FPS
RT-DETR (Baseline)	71.3	20.2	58.6	101
A	75.1	12.3	33.3	127
B	71.8	20.1	58.5	102
C	72.1	19.6	56.9	103
A + B	75.4	12.2	33.1	127
A + C	76.2	11.7	31.8	126
B + C	72.2	19.5	56.8	101
A + B + C (Our model)	77.2	11.4	31.5	128

For Model B, the integration of the AIFI-MSMHSA module introduces multi-scale multi-head self-attention mechanisms. This enhancement yields a 0.5% accuracy gain and a 1-frame FPS improvement while keeping parameter counts nearly unchanged, validating that the proposed module boosts detection speed without compromising accuracy or introducing computational overhead.

In Model C, the replacement of standard convolutions with the CCFF-GS module – featuring channel compression, depthwise convolution, and shuffle reorganization – achieves a 0.8% accuracy improvement, reduces parameters by 0.6 M, and increases FPS by 2 frames. This confirms the effectiveness of the three-stage optimization strategy in balancing computational efficiency and performance for small-target detection.

Finally, compared to the original RT-DETR baseline, our optimized model reduces parameters by half while achieving a 5.9% accuracy gain and a 27-frame FPS increase. These results underscore the substantial performance enhancement of the proposed framework, demonstrating its capability to perform real-time surface dirt detection on photovoltaic panels with both lightweight design and high precision.

3.3. **Comparative experimental analysis of different models.** To validate the superiority of the improved algorithm for photovoltaic panel surface dirt detection, comparative experiments were conducted against state-of-the-art object detection algorithms on the aforementioned custom dataset, with results summarized in Table 2.

TABLE 2. Comparison results of different models

Model	mAP_{50}/%	Params (M)	GFLOPs	FPS
Faster R-CNN	57.2	136.7	287.2	67
YOLOv10-m	67.5	15.4	59.1	135
YOLOv11-m	69.2	20.1	68.0	142
RT-DETR	71.3	20.2	58.6	101
Our model	77.2	11.4	31.5	128

As shown in Table 2, the proposed algorithm demonstrates significant improvements over Faster R-CNN in detection accuracy, model parameter count, and real-time performance. Compared to YOLOv10-m, it achieves a 9.7% increase in detection accuracy while reducing parameters by 4.0 M. When benchmarked against YOLOv11-m, the enhanced model exhibits an 8% accuracy gain alongside a 8.7 M parameter reduction. Notably, compared to the original RT-DETR, our algorithm achieves higher detection accuracy across multi-scale targets while maintaining a more compact parameter footprint, demonstrating superior adaptability to real-time inspection tasks in photovoltaic power plants.

These results collectively confirm that the improved algorithm optimally balances model lightweighting and detection performance. Its comprehensive capabilities surpass competing models, effectively meeting the dual requirements of lightweight design and high precision in photovoltaic power station scenarios. Crucially, the algorithm achieves efficient surface dirt detection even on low-cost, resource-constrained devices, highlighting its practical viability for industrial deployment.

To test the model's adaptability to different scenarios, the scene of snow covering the surface of photovoltaic panels was selected for testing. The Faster R-CNN, YOLOv10-m, YOLOv11-m, RT-DETR and the algorithm proposed in this paper were still selected for comparison. The comparison results are shown in Table 3. Although the accuracy gap between the models has decreased, the algorithm proposed in this paper still has a significant advantage over other models in terms of accuracy and computational cost.

Table 3. Comparison results on snow dataset

Model	mAP_{50}/%	Params (M)	GFLOPs	FPS
Faster R-CNN	76.8	136.7	287.2	67
YOLOv10-m	90.6	15.4	59.1	135
YOLOv11-m	91.8	20.1	68.0	142
RT-DETR	94.2	20.2	58.6	101
Our model	96.9	11.4	31.5	128

3.4. **Comparison of visualization results.** To validate the effectiveness of the enhanced algorithm, quantitative analyses were conducted in the aforementioned experiments. To more intuitively demonstrate its detection performance, this study selected challenging real-world application scenario images for visual presentation. Detection results were visualized using differently colored bounding boxes to highlight diverse target categories, along with corresponding labels and confidence scores, as illustrated in Figure 8.

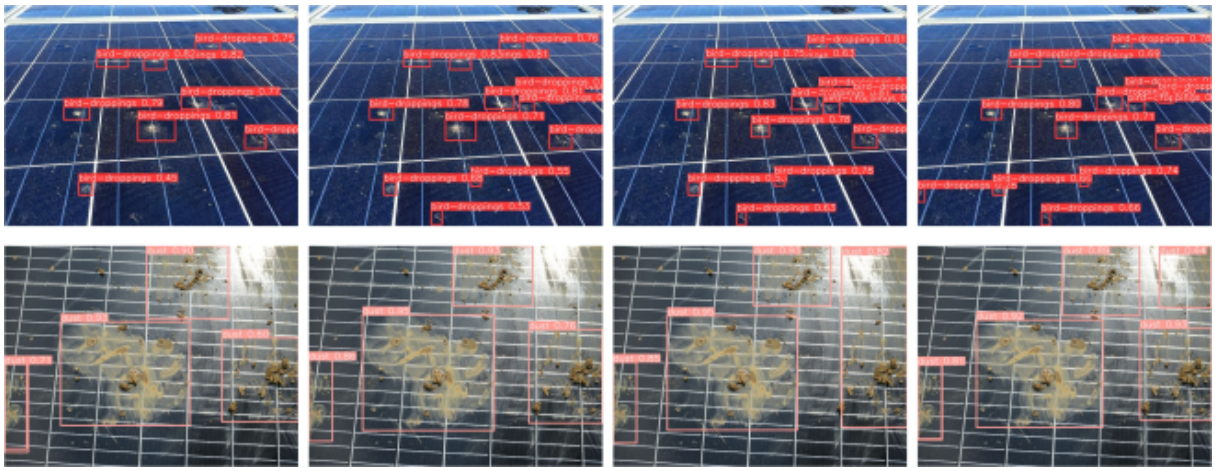


Figure 8. Comparison of detection results

To directly showcase the algorithm's practical detection capabilities in complex environments, a specialized visual validation dataset was constructed using industrially representative challenging inspection scenarios. The detection outcomes in Figure 8 encompass three typical scenarios: 1) Multi-scale target coexistence (featuring micro-defects ranging

from 0.5 to 50 pixels alongside standard-sized components), 2) Strong illumination interference (with specular reflections and shadow occlusion), and 3) High-density arrangements (component spacing less than 10% of target dimensions). These visualizations not only corroborate the quantitative findings but also vividly reveal the technical advantages of the improved algorithm in practical applications such as industrial quality inspection and security monitoring.

4. **Conclusions.** To effectively address the challenges of complex backgrounds and persistent false/missed detections for small targets in photovoltaic (PV) panel inspection, this study proposes a lightweight RT-DETR-based detection algorithm. Comprehensive validation on a dedicated PV dirt dataset demonstrates that the proposed approach achieves superior balance between model efficiency and detection performance compared to mainstream alternatives. Crucially, it attains high-precision recognition while significantly reducing computational costs, validating its robustness against occlusions and illumination variations. These results underscore its exceptional capability to reconcile accuracy with deployability in practical PV power stations. The demonstrated paradigm establishes a new benchmark for automated inspection systems (e.g., drones, and diagnostic platforms), enabling more reliable and sustainable PV maintenance.

Future improvements will focus on enhancing the model's generalization capabilities across diverse operational scenarios. Although the proposed algorithm demonstrates superior accuracy and computational efficiency under snow-covering conditions (Table 3), its performance gap with comparative models (e.g., Faster R-CNN, YOLOv10-m, YOLOv11-m, and RT-DETR) narrows in such edge cases. This suggests potential sensitivity to extreme environmental perturbations. To fortify robustness, we plan to 1) Extend validation to broader scenarios (e.g., dust storms, partial shading, and angular irradiation variance) to quantify domain adaptability; 2) Develop dynamic parameter optimization mechanisms that autonomously adjust to meteorological anomalies; 3) Integrate multimodal sensor fusion (e.g., thermal imaging, and weather data) to augment contextual awareness. Such enhancements will advance the model's deployment readiness in real-world photovoltaic monitoring systems.

Acknowledgment. This work is partially supported by Key Technologies for Improving the Efficiency of Photovoltaic Power Station Cleaning Systems Research and Application Project (Grant No. 37-2024-45-K0003). And this work is partially supported by Key Project of Jiangsu Provincial Key Laboratory of Power Transmission and Distribution Equipment Technology Team (Grant No. 2023JSSPD01). The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

[1] Chanchangi, N. Yusuf, A. Ghosh et al., Dust and PV performance in Nigeria: A review, *Renewable and Sustainable Energy Reviews*, vol.121, 2020.

[2] K. A. Abuqaaud and A. Ferrah, A novel technique for detecting and monitoring dust and soil on solar photovoltaic panel, *2020 Advances in Science and Engineering Technology International Conferences (ASET)*, Dubai, United Arab Emirates, pp.1-6, 2020.

[3] A. Wang, H. Chen, L. Liu et al., YOLOv10: Real-time end-to-end object detection, *arXiv Preprint*, arXiv: 2405.14458, 2024.

[4] W. Wu, H. Liu, L. Li et al., Application of local fully convolutional neural network combined with YOLO v5 algorithm in small target detection of remote sensing image, *PloS One*, 2021.

[5] H. Wu, K. Chen, M. Ni and L. Ma, Strip surface defect detection based on improved YOLOv7, *International Journal of Innovative Computing, Information and Control*, vol.20, no.5, pp.1493-1507, DOI: 10.24507/ijicic.20.05.1493, 2024.

[6] J. Terven, D. M. Cordova-Esparza and J. A. Romero-Gonzalez, A comprehensive review of YOLO architectures in computer vision: From YOLOv1 to YOLOv8 and YOLO-NAS, *Machine Learning and Knowledge Extraction*, vol.5, no.4, pp.1680-1716, 2023.

[7] X. Xie, J. Ren, Y. Zeng et al., HATSC-YOLOv10: Improved YOLOv10 for satellite remote sensing images of small object detection, *2024 China Automation Congress (CAC)*, Qingdao, China, pp.3795-3799, 2024.

[8] Z. Chen et al., An improved Faster R-CNN transmission line bolt defect detection method, *2022 World Automation Congress (WAC)*, San Antonio, TX, USA, pp.82-85, 2022.

[9] J. Xu, Q. Jia, T. Qiu, Y. Xu, M. Wu and B. Yang, Research and application of intelligent detection technology for bridge girder bottom appearance defects by suspended bridge inspection vehicle, *International Journal of Innovative Computing, Information and Control*, vol.20, no.1, pp.15-30, DOI: 10.24507/ijicic.20.01.15, 2024.

[10] N. Carion, F. Massa, G. Synnaeve et al., End-to-end object detection with transformers, *European Conference on Computer Vision*, pp.213-229, 2020.

[11] D. Wang, Z. Li, X. Du et al., Farmland obstacle detection from the perspective of UAVs based on non-local deformable DETR, *Agriculture*, vol.12, no.12, 2022.

[12] Y. Zhao, W. Lv, S. Xu et al., DETRs beat YOLOs on real-time object detection, *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.16965-16974, 2024.

[13] K. He, X. Zhang, S. Ren et al., Deep residual learning for image recognition, *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.770-778, 2016.

[14] H. Zhang, C. Hao, W. Song et al., Adaptive slicing-aided hyper inference for small object detection in high-resolution remote sensing images, *Remote Sensing*, vol.15, no.5, 1249, 2023.

[15] D. Ouyang, S. He, G. Zhang et al., Efficient multi-scale attention module with cross-spatial learning, *2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2023)*, Rhodes Island, Greece, pp.1-5, 2023.

[16] L. Ding, SlimEMA-YOLOv8: Enhanced traffic sign recognition for autonomous driving using EMA and Slim-neck in YOLOv8, *2024 6th International Conference on Internet of Things, Automation and Artificial Intelligence (IoTAAI)*, Guangzhou, China, pp.723-726, 2024.

[17] H. Zheng, J. Duan, Y. Dong et al., Real-time fire detection algorithms running on small embedded devices based on MobileNetV3 and YOLOv4, *Fire Ecology*, vol.19, no.1, 2023.

[18] C. Yu, B. Xiao, C. Gao et al., Lite-HRNet: A lightweight high-resolution network, *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA, pp.10435-10445, 2021.

[19] X. Ma, X. Dai, Y. Bai et al., Rewrite the stars, *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, pp.5694-5703, 2024.

[20] X. Dai, Y. Chen, B. Xiao et al., Dynamic head: Unifying object detection heads with attentions, *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA, pp.7369-7378, 2021.

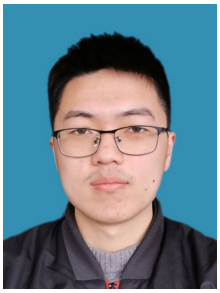
Author Biography

Wei Liu graduated from Changsha University of Science and Technology, China, in 2006.

Mr. Liu is currently a senior engineer in the New Energy Division of Shandong Electric Power Engineering Consulting Institute Corp., Ltd., China, and has been awarded the title of Excellent Project Manager. His main research interests include engineering research within the field of new energy: research and application of ecologically harmonious mountain wind power project construction, key technology research and application of construction organization for large-scale mountain wind power projects, and a waste to energy plant flue gas purification treatment system, as well as patent projects. He has published over 10 papers in well-known journals.

Yun Sun graduated from Shandong University of Finance, China, in 2010, passed the National Intermediate Economist Examination in 2016 and obtained the intermediate professional title, and participated in and passed the National Senior Economist Examination in 2024.

Ms. Sun is currently an engineer in the New Energy Division of Shandong Electric Power Engineering Consulting Institute Corp., Ltd., China. She published "How to Do a Good Job in Fee Management in the Current Situation of the New Energy Industry" in "Business Situation" and published "Research on Full Process Cost Control of New Energy EPC General Contracting Projects Based on Total Price Contracting Model" in "Economic Management" in December 2024. She research interests focus on the economic management and cost control mechanisms within the new energy industry. Her work specifically examines whole-process cost optimization in EPC turnkey projects under a lump-sum contract model, addressing challenges from initial bidding to final execution. She aims to develop practical frameworks that enhance cost efficiency and project sustainability in evolving energy markets.



Yuchen Zhang received the B.Eng. degree in Communication Engineering from Hohai University, China, in 2023 and is currently a postgraduate student at Hohai University.

His research focuses on developing object detection models. He aims to investigate the accuracy of these models and ensure real-time performance when deployed on edge devices through algorithm design.

Hongwei Chang obtained a Bachelor's degree in Agricultural Architecture and Environmental Engineering from Hebei Agricultural University, China, in 1994. In 2004, he was awarded the title of Senior Engineer, and in 2005, he obtained the title of First Class Constructor.

Mr. Chang is currently a senior engineer in the New Energy Division of Shandong Electric Power Engineering Consulting Institute Corp., Ltd., China, and has been engaged in the construction and management of power generation projects such as thermal power, wind power, and photovoltaics for 30 years. More than ten 135-1000 MW thermal power plants that participated in the construction have all achieved high standard production and won national, industry (ministerial), and provincial quality engineering awards. His research direction includes construction and management of power engineering. He is a chief editor of a book on architecture.

Xiaochen Wang graduated from Northwestern Polytechnical University with a bachelor's degree in 2007, and from University of Science and Technology Beijing, China, with a master's degree in 2010.

Ms. Wang is currently the deputy director of the First Department of New Energy of Shandong Electric Power Engineering Consulting Institute Corp., Ltd., China, and a senior engineer. She has been deeply involved in the new energy field and is currently mainly engaged in new energy technology management. As of 2025, she has participated in three provincial and national key projects, won multiple awards in the power industry, and her projects have been selected as first-level cases of digital technology application innovation in 2025. She has also participated in the compilation of several industry standards and group standards.

Xinnan Fan received the B.Eng. degree in Automation from Hohai University, China, 1987; the M.Sc. degree in Electrical Engineering and Its Automation from Hohai University, China, 1998; the Ph.D. degree in Water Conservancy Project from Hohai University, China, 2009.

Dr. Fan is responsible for the construction of the Jiangsu Provincial Key Laboratory of Power Transmission and Distribution Equipment Technology. The laboratory has undertaken more than 100 projects at or above the provincial and ministerial level, and has won 2 second prizes of the National Science and Technology Progress Award and 5 first prizes of the provincial and ministerial level Science and Technology Award. Multiple technological breakthroughs have been made in the research and development of high-speed railway traction power supply systems, resulting in the development of highly reliable intelligent gas insulated switchgear and 350 km/h high-speed train traction transformers that are at the world's advanced level. His research is dedicated to advancing power transmission and distribution equipment technology. His work focuses on technological innovation in the development of highly reliable intelligent gas insulated switchgear. A significant part of his research involves leading provincial key laboratories to bridge theoretical advancements with practical engineering applications, aiming to enhance the safety, intelligence, and reliability of critical power infrastructure.