

Robust Detection of Infrared Dim and Small Targets under Complex Backgrounds

Mengke Zhang^{1,*}, Lisha Yao²

¹School of Big Data and Artificial Intelligence, Anhui Xinhua University, Hefei, Anhui, China

²School of Big Data and Artificial Intelligence, Anhui Xinhua University, Hefei, Anhui, China

* Corresponding author

Abstract: This paper is based on the inspection scenario of photovoltaic power stations and has constructed an optimized technical system for precise detection. It focuses on breaking through the key technical difficulties of precise target detection for unmanned aerial vehicles in complex terrain environments, providing a systematic solution to improve the efficiency and accuracy of photovoltaic power station inspections. For detection problems such as the weak texture features of photovoltaic modules, a target detection algorithm based on the improved YOLOv8 network is proposed. Firstly, the Swin Transformer module is introduced to construct hierarchical feature representation and sliding window mechanism, effectively capturing multi-scale information in the image. Secondly, a progressive feature pyramid network is designed, gradually transforming and integrating information from low-level features to high-level features, solving the difficulties in detecting small targets of photovoltaic modules and the problem of blurred feature loss. Finally, through simulation and experiments, the effectiveness of the proposed path planning algorithm and target detection algorithm is verified.

Keywords: Photovoltaic Modules, Small-Target Detection, YOLOv8 Network.

1. Introduction

The unmanned aerial vehicle (UAV) inspection system has established a new model for intelligent inspection of photovoltaic power stations, providing innovative solutions to the problems in the operation and maintenance system of photovoltaic power stations. However, it also faces many challenges and difficulties. In response to the problems of low efficiency and high safety risks in the traditional manual inspection mode, this paper, based on UAV technology, proposes a photovoltaic module target detection method based on the improved YOLOv8 model. To address the detection problems caused by weak texture features on the surface of photovoltaic modules, small target detection difficulties, and background similarity interference, a photovoltaic module target detection algorithm based on the improved YOLOv8 model is proposed. Firstly, the Swin Transformer module is used to replace the Transformer module of the main network of YOLOv8, enhancing the feature extraction performance of the main network of YOLOv8 in photovoltaic module infrared images,

significantly improving the model's ability to extract weak texture features in infrared images. Secondly, a progressive feature pyramid network is designed to optimize the fusion process of multi-scale features, solve the problem of feature degradation, and improve the detection ability for small targets. To further verify the performance of the model, comparative experiments and ablation experiments were conducted on the photovoltaic module dataset in this chapter.

2. YOLOv8 Network

The network architecture of YOLOv8 is composed of three core modules : backbone network, feature fusion network and detection head. Its network architecture provides a more accurate and efficient solution for various detection scenarios by effectively dealing with complex background interference and target scale differences through interaction and optimization between modules while processing in real time. The network architecture of YOLOv8 is shown in Figure 2-1, which shows its structure and the relationship between modules.

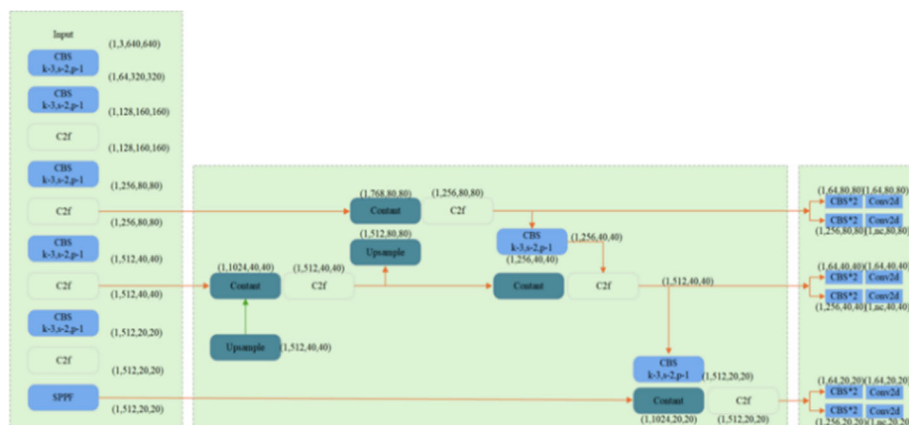


Figure 2-1. Architecture of the Progressive Feature Pyramid Network

The backbone network of YOLOv8 is a key component of its architecture, which is responsible for extracting multi-scale feature information from input images and providing rich semantic and spatial information for subsequent object detection tasks. The input image size of the network is $640 \times 640 \times 3$. Firstly, it is preliminarily processed by convolution layer and pooling layer.

At this point, the network extracts the low-level features of the picture via convolution operations and reduces spatial dimensions by pooling operations. Hence, the complexity of computation is decreased. After the preliminary processing, the input feature maps are passed to a variety of convolutional layers and to the feature fusion (C2F) module. C2F module comprises several convolutional layers and feature fusion layers. Progressive reduction of spatial resolution and growth of the number of channels help the network to extract the higher level semantic information [1].

The feature fusion layer also mixes the feature information extracted from the different layers and therefore increases the representation capabilities of the network. After the last C2F module, YOLOv8 adds the spatial pyramid pooling layer to the model for multi-scale feature fusion. The SPP layer extracts the contextual information at different scales via the pooling operations applied to the feature maps with different resolutions. This helps to make the network more robust to multi-scale objects. This layer extracts the multi-level spatial information at different scales and fuses it into the unified feature map. With the help of this design, YOLOv8 manages to keep high detection accuracy and robustness towards both small and large objects [2].

When it comes to the YOLOv8, the backbone network uses certain techniques such as deep convolution and large kernel convolution. It makes sure that feature extraction is highly efficient. Deep convolution enables the extraction of high-level semantics from the input with the help of conv layers. However, large kernel convolution helps the network to capture a wider context.

Besides increasing the efficiency of the feature extraction process, these designs make sure that the adaptability of the model to various target scales is increased. YOLOv8 is extremely fast and accurate at the same time. With the use of multilevel convolution and feature fusion, the backbone of YOLOv8 creates a highly efficient feature extraction module.

3. Improved YOLOv8 Network

(1) Swin Transformer Module

In this work, the conventional Transformer model in YOLOv8 is replaced with Swin Transformer model which improves the efficiency of YOLOv8 in target detection of photovoltaic components.

By introducing hierarchical feature representation and sliding window mechanism, Swin Transformer shows strong performance in a variety of visual tasks, especially in dealing with complex background and fine-grained features and scale changes [3]. One advantage of Swin Transformer is its hierarchical structure. It can gradually extract from detail features to abstract features by layer-by-layer downsampling, so that the model can handle image features with different resolutions. This can effectively capture information from local to global, so it can maintain high detection performance

for targets with weak texture features such as photovoltaic modules.

Traditional CNN is easy to lose global information when dealing with complex backgrounds and multi-scale targets, because it relies on local convolution operations to extract image features. The Swin Transformer uses a sliding window self-attention mechanism to learn the long-distance dependencies and global context in the image layer by layer, so as to effectively capture the multi-scale information in the image and more accurately identify the object features [4]. Traditional convolutional neural networks are usually optimized only at fixed scales, while Swin Transformer can extract features at different scales at the same time through flexible hierarchical design, which makes it stable in the face of photovoltaic modules with different sizes and weak texture features. Especially in the case of complex background, weak and blurred target features, Swin Transformer shows stronger robustness and higher accuracy [5].

In order to reduce the computational complexity and improve the processing efficiency, Swin Transformer will use a slicing module to divide the input image into multiple non-overlapping blocks after inputting the original RGB image. These tiles are visually similar to 'local regions', and each tile is regarded as a 'token' that will carry the local information of the image.

Afterward, the RGB values in each block are concatenated into a feature vector. The original features will be projected into a high dimensional space through the linear embedding layer to make feature processing and feature fusion more efficient. After completing the segmentation and feature embedding process, a series of Swin Transformer modules will be applied to process the features in those segments. The Swin Transformer makes use of the local attention mechanism which eliminates the computational costs brought by the global self-attention operations. In addition, the local self-attention mechanism makes Swin Transformer able to achieve reduction in computational costs while retaining the ability of global information capturing. Through the hierarchical architecture design, the Swin Transformer can process features in different scales for each layer.

Block merging layer is another essential part of the Swin Transformer architecture. At the beginning of the training process, two adjacent 2×2 blocks are merged into one block, which is then processed by a linear transformation layer. As a result, the size of the final feature map is $4C$. Such an architecture not only increases the feature representation but also reduces the number of tokens used. In fact, the resolution of the output feature map becomes half the resolution of the input feature map.

However, with increasing depth of the network architecture gradually, the Swin Transformer continues to transform the lower level of the detailed features of the image into high-level abstract features. Thus, owing to such hierarchical transformation of features, the Swin Transformer can handle multiscale target objects while integrating lower level details with higher level features simultaneously. In other words, we get a hierarchical feature representation of the image by means of the Swin Transformer [6]. It is this mechanism that makes it so efficient in detecting objects in complicated real-world environments like defect inspection on photovoltaic modules.

Thus, this simple addition becomes rather useful when it comes to detection of both localized and contextual features of the photovoltaic module surface texture. Indeed, due to the hierarchical down-sampling architecture, the network not only learns the local texture features of the modules in its shallow layers but also takes care of the global structure of the modules in its deeper layers.

In this work, the conventional Transformer model in YOLOv8 is replaced with Swin Transformer model which improves the efficiency of YOLOv8 in target detection of photovoltaic components.

(2) Improved Feature Pyramid Network

However, in regard to the fusion of non-adjacent layers, the existing feature fusion techniques have often failed due to the issue of information loss. In order to address such problems, we present our AFPN (Asymmetric Feature Pyramid Network) in this research. The design of AFPN successfully solves the problem of large semantic gap of non-adjacent layers by merging the features from the different layers bit by bit, thus dramatically improving the performance and precision of the feature fusion process. First, the AFPN merges adjacent low-level features, then, slowly merges high-level features. With

the help of its layer-by-layer fusion strategy, AFPN successfully reduces the semantic gap of different layers [7].

Thus, the quality of feature fusion is improved, and the detection performance of small targets is enhanced, so that the network can provide more accurate target detection when dealing with complex scenes.

AFPN also adds adaptive spatial fusion operation, which further optimizes the adaptive spatial fusion operation of feature fusion process. In the complex environment where multiple objects are intertwined, the information difference is effectively reduced, and the characteristics of each spatial position can accurately reflect the real situation of the object. The feature extraction process of AFPN is similar to many FPN-based detection methods. Features are extracted from different levels of the backbone network, and then feature fusion is performed [8]. Figure 3-1 shows the architecture of AFPN. By gradually integrating different levels of features, it shows the bottom-up feature extraction process in the backbone network. How to gradually integrate low-level features, high-level features and top-level features [9], feature information can be fully represented in the entire network.

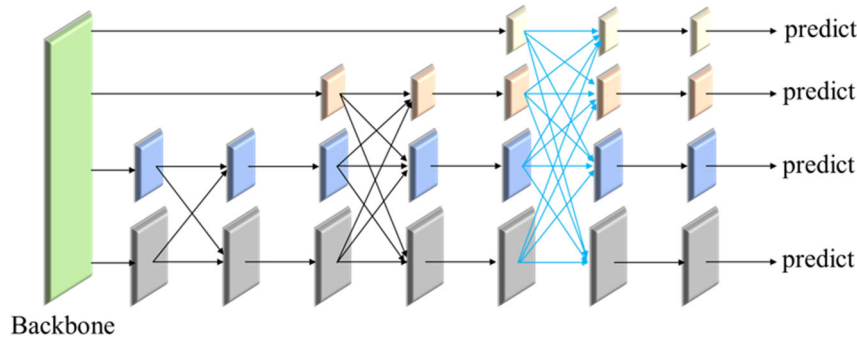


Figure 3-1. AFPN Network

To ensure that features from different levels can be effectively aligned, AFPN employs 1×1 convolution and bilinear interpolation techniques for upsampling [10]. The 1×1 convolution adjusts the number of channels of the feature map without changing its spatial dimensions, ensuring dimensional matching of features at different scales. Bilinear interpolation, on the other hand, smooths the image by taking a weighted average of surrounding pixels, effectively improving the quality and representational capability of the feature maps. Through these fusion and upsampling operations, AFPN progressively reduces the size of the feature maps at each stage and extracts more refined feature information through multiple rounds of fusion.

During the down-sampling stage, AFPN adjusts the size and stride of the convolution kernel as needed, reducing the spatial resolution of the feature map while preserving key information. The down-sampling operation causes the size of the feature map to decrease progressively, thereby enabling better processing of large-scale image information and enhancing the model's global perception capability.

After completing feature fusion, AFPN continues further feature learning through residual units similar to those in ResNet. Each residual unit consists of multiple convolution operations and residual connections. This approach effectively mitigates the vanishing gradient problem in deep

networks and ensures the efficient flow of information within the network. By transforming and fusing information from low-level features to high-level features, AFPN significantly improves its detection capability for small targets. Employing a bidirectional fusion pathway that combines top-down and bottom-up approaches, along with an adaptive spatial fusion operation, it is able to accurately localize small targets and suppress background noise.

4. Experiments and Conclusions

(1) Experimental Evaluation Indicators

Precision: The fraction of detections predicted as positive by the model that were actually positive. The precision rate calculates the proportion of objects that are accurately detected out of all instances that are detected as objects.

Recall: This computes the proportion of all true objects that can be accurately detected as objects. Recall reflects the ability of the model to cover all the actual existing targets, that is, the proportion of all the real targets that the model successfully identifies.

If an algorithm is able to detect all the actual objects present in the image, this suggests that its recall rate is high. An improvement in recall rate implies that the algorithm is becoming more effective at reducing missed detections. In

most object detection tasks, evaluating the overall performance of the algorithm requires taking both precision and recall into account, since these two measures together provide a more complete picture.

IoU, which stands for intersection over union, is a measure of how much a predicted bounding box overlaps with the corresponding ground-truth box. Suppose we set the IoU threshold at 0.7. In that case, we call a prediction a true positive (TP) if its IoU exceeds 0.7 and its class label is correct. If the IoU falls below 0.7 or the class label is incorrect, we refer to that prediction as a false positive (FP). The specific calculation formulas for precision and recall are presented in Equations 4-(1) and 4-(2):

Precision:

$$\text{Precision} = \frac{TP}{TP + FP} \quad 4-(1)$$

Recall:

$$\text{Recall} = \frac{TP}{TP + FN} \quad 4-(2)$$

In theory, we would like both precision and recall to be as high as possible. In practice, however, these two measures tend to work against each other; it is uncommon to see them both peak together. That is why relying on precision alone or recall alone is not enough for a thorough evaluation of the model. This is where the F1 score steps in. The F1 score

combines precision and recall and finds a middle ground, letting us push precision higher while still keeping recall at a decent level.

The calculation formula for the F1-score is presented in Equation 4-(3):

F1-Score:

$$F1 = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}} \quad 4-(3)$$

(2) Comparative Experimental Results and Analysis

Two hundred infrared images of photovoltaic modules, spanning a variety of types, were chosen for the experiments in this section. One hundred twenty of these were designated as the test set, while the remaining eighty were used for training. Data augmentation was applied to the original images; this step was intended to increase the diversity of the training samples and to strengthen the robustness of the resulting model. For the training process, we set the input image size to 512×512 pixels, used a batch size of 12, and ran the optimization for 200 epochs. After training concluded, we saved the final model weights and evaluated the model's performance on the test set.

This study selects YOLOv7, Faster-RCNN, and the original YOLOv8 as baseline models, and conducts comparative experiments centered on three core metrics: Precision, Recall, and F1-Score. The detection results of each model are presented in Table 4-1.

Table 4-1. Detailed Detection Results of Photovoltaic Module Detection Models

Method	Precision	Recall	F1-Score
YOLOv7	0.925	0.915	0.920
Faster-RCNN	0.927	0.932	0.930
YOLOv8	0.930	0.936	0.933
Improved YOLOv8	0.957	0.947	0.953

From the table, it can be observed that the original YOLOv8 model achieves a precision of 0.930, demonstrating the advantages of the YOLO series. However, there is still room for improvement in the model's balance between precision and recall, as well as in its detection accuracy. In contrast, the improved YOLOv8 model proposed in this paper, while retaining the efficient structure of the original network, further introduces the Swin Transformer module and the progressive feature pyramid network, enabling more accurate recognition of photovoltaic module features, enhancing the quality of feature fusion in complex scenes, and effectively strengthening the algorithm's ability to process samples. The experimental results show that the precision of the improved model is increased to 0.957.

5. Conclusion

This paper proposes a photovoltaic module target detection method based on the improved YOLOv8 model. For the infrared images of photovoltaic modules, the characteristics of deep learning methods in target detection are systematically analyzed. Traditional image detection methods generally rely on manual feature extraction and image processing techniques, and have better recognition effects in a single scene, but lack

certain universality and efficiency, and cannot meet the detection requirements of large-scale power stations. In response to the problems of traditional detection methods, this paper proposes a photovoltaic module target detection algorithm based on the improved YOLOv8 model. Firstly, the Swin Transformer module is used to replace the Transformer module of the main network of YOLOv8 model, enhancing the feature extraction performance of the main network of YOLOv8 in infrared images of photovoltaic modules, significantly improving the model's ability to extract weak texture features in infrared images. Secondly, a progressive feature pyramid network is designed to further optimize the fusion process of multi-scale features, solving problems such as feature degradation. To further verify the performance of the model, comparative experiments were conducted on the photovoltaic module dataset. The algorithm was verified to have good detection accuracy, effectively solving the problems of environmental adaptability and detection robustness of traditional methods, providing a precise technical solution for intelligent inspection of photovoltaic power stations.

Acknowledgements

This work is supported by the 2026 University-Level Natural Science Research Project of Anhui Xinhua University (Grant No. 2026zrqyb06), the 2025 Provincial College Student Innovation and Entrepreneurship Training Program (Grant No. S202512216095), the 2023 Anhui Provincial Scientific Research Preparation Plan Project (Grant No. 2023AH051806), and the 2025 University Young and Middle-Aged Teacher Training Action Program - Outstanding Young Teacher Cultivation Project (Grant No. YQYB2025061). The author would like to express their sincere gratitude to all the funding agencies for their financial support.

References

- [1] Ren, R., Sun, H., Zhang, S., Wang, N., Lu, X., Jing, J., Xin, M., & Cui, T. (2023). Intelligent detection of lightweight “Yuluxiang” pear in non-structural environment based on YOLO-GEW. *Agronomy*, 13(9), 2418. <https://doi.org/10.3390/agronomy13092418>.
- [2] Ju, R.-Y., & Cai, W. (2023). Fracture detection in pediatric wrist trauma X-ray images using YOLOv8 algorithm. *Scientific Reports*, 13(1), 20077. <https://doi.org/10.1038/s41598-023-47481-3>
- [3] Baek, J.-W., & Chung, K. (2023). Swin transformer-based object detection model using explainable meta-learning mining. *Applied Sciences*, 13(5), 3213. <https://doi.org/10.3390/app13053213>
- [4] Peng, Y., Ren, J., Wang, J., & Shi, M. (2023). Spectral-swin transformer with spatial feature extraction enhancement for hyperspectral image classification. *Remote Sensing*, 15(10), 2696. <https://doi.org/10.3390/rs15102696>
- [5] Huo, D., Kastner, M. A., Liu, T., Kawanishi, Y., Hirayama, T., Komamizu, T., & Ide, I. (2023). Small object detection for birds with swin transformer. 2023 18th International Conference on Machine Vision and Applications (MVA), 1–5. <https://doi.org/10.23919/MVA57639.2023.10216093>
- [6] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10012–10022. <https://doi.org/10.1109/ICCV48922.2021.00986>
- [7] Wang, B., Ji, R., Zhang, L., & Wu, Y. (2023). Bridging multi-scale context-aware representation for object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(5), 2317–2329. <https://doi.org/10.1109/TCSVT.2022.3221755>
- [8] Zhang, D., Zhang, H., Tang, J., Wang, M., Hua, X., & Sun, Q. (2020). Feature pyramid transformer. *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII*, 323–339. Springer International Publishing.
- [9] Yang, G., Lei, J., Zhu, Z., Cheng, S., Feng, Z., & Liang, R. (2023). AFPN: Asymptotic feature pyramid network for object detection. 2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2184–2189.
- [10] Jing, X. (2024). Research on pedestrian detection and tracking methods based on deep learning [Master’s thesis]. Shenyang Ligong University.