Woody International Publish Limited
An Multidisciplinary Academic Journal Publisher

Journal of Theory and Practice in Engineering and Technology, Volume 1, Issue 2, 2024
https://www.woodyinternational.com/

Enhanced YOLOv8 Infrared Image Object Detection Method with SPD Module

Zheng Ren

College of Computing, Georgia Institute of Technology, North Avenue, Atlanta, GA 30332
superpy001@gmail.com

Abstract: ***This study proposes an improved deep learning-based method for object detection in low-resolution infrared images in factory environments. Considering the impact of factors such as fog, smoke and insufficient illumination on the detection accuracy, multispectral imaging techniques, especially infrared thermography, are introduced in this study to enhance the recognition capability of the system. We developed a novel algorithm based on YOLOv8 by improving the existing deep learning model and integrated the SPD module, which significantly improves the recognition accuracy of low-resolution images and small objects.The SPD module employs a unique downsampling method on the input feature maps, boosting the model's ability to detect low-resolution images and small targets while also improving inference speed. Our experimental results on LLVIP and VEDAI datasets demonstrate the superior performance of the proposed method for target detection in infrared images.***

Keywords: YOLOv8; Low Resolution Infrared Images; Object Detection.

Cited as: Ren, Z. (2024). Enhanced YOLOv8 Infrared Image Object Detection Method with SPD Module. *Journal of Theory and Practice in Engineering and Technology*, *1*(2), 1–7. Retrieved from https://woodyinternational.com/index.php/jtpet/article/view/21.

1. Introduction

In contemporary industrial environments, object recognition and pedestrian detection face numerous challenges, particularly in scenarios with insufficient lighting, the presence of smoke or fog, and occluded objects[1]. These conditions significantly impact the accuracy and reliability of traditional recognition algorithms. Methods that rely heavily on visible light sensors often fail to perform satisfactorily in complex factory settings, where such challenging conditions are common[2].

To address these issues, infrared imaging has emerged as a highly effective multispectral imaging technique due to its independence from visible light[3]. Infrared thermal imaging can produce clear images even in the absence of light or under low-light conditions, making it an invaluable tool for overcoming the limitations of traditional methods. However, despite its advantages, infrared image processing and object detection still face several unresolved issues, such as maintaining high detection accuracy in low-resolution and complex background conditions[4].

In recent years, deep learning techniques have made significant strides in image processing and target detection. Algorithms based on deep convolutional neural networks[5], such as YOLO[6] and Faster R-CNN[7], have demonstrated high effectiveness across various domains[8]. Yet, applying these advanced techniques directly to infrared detection often fails to yield satisfactory results due to the unique characteristics of infrared images and the complexity of real-world application scenarios.

This study aims to tackle the specific challenges of infrared target detection in industrial environments by proposing an enhanced detection method that leverages cutting-edge deep learning frameworks and the unique properties of infrared images. We introduce an innovative SPD module within the YOLOv8[11] architecture, specifically designed to enhance detection accuracy for low-resolution images and small objects, while also accelerating inference speed. This novel approach allows the model to maintain high performance even under the challenging conditions typical of industrial settings.

The integration of the SPD module improves the model's ability to handle low-resolution infrared images, providing more accurate detection of small objects amidst complex backgrounds and enhancing the inference speed. Our experimental results demonstrate that the proposed model achieves state-of-the-art performance, highlighting its potential for practical application in industrial environments. This study not only advances the field of infrared target detection but also provides a robust solution to the persistent challenges faced in real-world industrial scenarios.

2. Related Work

In recent years, researchers have introduced a variety of enhanced deep learning algorithms to tackle the challenges associated with infrared target detection. In 2020, Hongyan Cao[14] made significant strides by improving the YOLOv3 algorithm. By incorporating multi-scale feature extraction and integrating batch normalization (BN) with convolutional layers, Cao enhanced the detection accuracy of long-range infrared targets from 66% to 88% . Following this, in 2021, Gu Josiah[15] presented an improved Faster R-CNN algorithm. This enhanced model utilized techniques such as multi-scale fusion and an optimized loss function, leading to a 3.95% increase in model accuracy . In the same year, Wei Cai proposed the YOLO-FCSP target detection framework, which significantly advanced the accuracy of infrared weak target detection to 92.6%. Wei Cai[16] achieved this improvement by reducing the number of downsampling operations, incorporating cross-stage local networks, and employing multipath aggregation. These advancements highlight the ongoing efforts and innovative approaches researchers have undertaken to refine infrared target detection, demonstrating the potential of deep learning to overcome the unique challenges posed by infrared imagery in various practical applications.

Deep learning target detection algorithms can be broadly categorized into two main classes: two-stage and one-stage algorithms. Two-stage algorithms, such as those in the R-CNN family[12], operate by first generating candidate regions and then performing object classification and localization within these regions. These algorithms, including Faster R-CNN, are renowned for their high detection accuracy. However, their detection speed is slower, making them less suitable for real-time applications, which limits their practical use in dynamic environments. Conversely, one-stage algorithms, represented by the YOLO series, are designed for fast detection with good real-time performance and high accuracy[17-19]. The YOLO series, including variants like YOLOv3, YOLOv4, and YOLOv5, process images in a single pass, significantly enhancing detection speed while maintaining competitive accuracy. This makes them particularly valuable in applications requiring rapid and accurate object detection[20]. For instance, Mingdi[21] improved the Faster R-CNN algorithm, achieving high accuracy on challenging datasets such as Foggy Cityscapes[22] and Rain Vehicle Color-24[23]. This demonstrates the potential of two-stage algorithms in specific scenarios despite their slower performance. On the other hand, single-stage algorithms like YOLO and RetinaNet are celebrated for their rapid recognition capabilities. Xianglin[25], for example, developed the SwinFocus module to enhance the feature extraction capabilities of YOLOv5[20,26] , resulting in impressive detection accuracy. These advancements underscore the ongoing evolution of deep learning-based target detection algorithms, highlighting the trade-offs between accuracy and speed[13]. By balancing these factors, researchers continue to refine and adapt these algorithms to meet the diverse needs of various practical applications, from industrial environments to autonomous driving.

3. Algorithm And Model

3.1 Overview

To validate the effectiveness of our proposed method, we extended the YOLOv8 architecture by integrating a novel SPD module. The SPD module, a convolutional neural network component, is specifically designed to address the challenges posed by low-resolution images and the detection of small objects. This module not only enhances the model's detection accuracy for these challenging targets but also significantly improves the inference speed. Importantly, it achieves these improvements without compromising the quality of the training process. By incorporating the SPD module, our enhanced YOLOv8 architecture demonstrates superior performance, making it a robust solution for accurate and efficient target detection in complex industrial environments.

3.2 SPD

The SPD layer down samples the input feature map in a unique manner that preserves fine-grained information. Specifically, the SPD layer divides the input feature map into several sub-maps, each representing a portion of the original feature map. These sub-maps are then down sampled according to a specified ratio, ensuring that critical details are maintained during the process. This down sampling technique allows the model to retain essential features necessary for accurate detection, especially for low-resolution images and small objects. The process can be described by the following equation:

$$f_{x,y} = X[i + x{:}S{:}scale, i + y{:}S{:}scale] \tag{1}$$

where, $f_{x,y}$ represents a subgraph, and X denotes the original feature map. The indices $i + x$ and $i + y$ indicate that the original feature map can be evenly divided according to the scale factor, ensuring that each subgraph $f_{x,y}$ is composed of blocks segmented from the original feature map X based on this scale factor $scale$. For instance, when the scale factor $scale = 2$, the original feature map X is divided into four subgraphs: $f_{0,0}$ $f_{1,0}$, $f_{0,1}$, and $f_{1,1}$. Each subgraph has a shape of $(\frac{S}{scale}, \frac{S}{scale}, C)$, where S represents the spatial dimensions of the original feature map, and C denotes the number of channels. This unique downsampling method allows the preservation of fine-grained information within each subgraph, thereby enhancing the model's ability to detect low-resolution images and small objects more effectively. By segmenting and downsampling the feature map in this manner, the SPD layer contributes to maintaining high detection accuracy while also improving inference speed.

Next, by concatenating these sub-feature maps along the channel dimension, we obtain the intermediate feature map $X^{'}$. The spatial dimensions of $X^{'}$ are reduced by a factor of $scale$ in each direction, while the channel dimension is expanded by a factor of $scale^2$ compared to the original feature map. This segmentation and downsampling process results in each subgraph containing only a portion of the original feature map, allowing for more focused processing of local features. This targeted processing helps preserve fine-grained information, crucial for detecting small objects and maintaining high detection accuracy.

Moreover, the downsampling ratio enables these subgraphs to capture information at different scales. By processing local features at various scales, the model can capture a richer set of multi-scale information, enhancing its ability to detect objects of different sizes and in varied contexts. This multi-scale processing is particularly beneficial in complex industrial environments, where objects can vary significantly in size and appearance.

After the SPD feature transformation layer, a convolution layer with C_2 filters is added, where $C_2 < scale^2 C_1$, using a non-stride (i.e., stride=1) convolution. This process further transforms the intermediate feature map $X^{'}$from $(\frac{S}{scale}, \frac{S}{scale}, scale^2 C_1)$ to $X^{''}(\frac{S}{scale}, \frac{S}{scale}, scale^2 C_2)$.

The use of non-stride convolution is essential to preserve as much discriminative feature information as possible. Using a convolution with a stride greater than 1 could potentially lose critical information. For example, a 3x3 convolution with a stride of 3 would significantly "shrink" the feature map, sampling each pixel only once. If the stride were 2, asymmetric sampling could occur, leading to inconsistent sampling times for even and odd rows and columns, thereby resulting in information loss.

By applying a non-stride convolution to $X^{'}$ we obtain a new feature map $X^{''}$, which retains the same spatial dimensions as $X^{'}$, but with the channel dimension reduced from $scale^2 C_1$ to C_2. This careful handling of feature dimensions and sampling ensures that fine-grained details captured in the earlier stages are preserved, while also refining the feature map for subsequent processing. Additionally, this approach enhances the inference speed of the model, thus improving both the efficiency and the overall detection performance.

4. Experiments

4.1 Datasets

The LLVIP[9] dataset is a comprehensive collection designed for pedestrian detection using paired visible light and infrared images. It includes a total of 33,672 images, forming 16,836 pairs of visible light and infrared images. Notably, most of these images were captured in low-light environments, making the dataset particularly valuable for studying detection techniques under challenging lighting conditions. In this study, we focus exclusively on the infrared images from the LLVIP dataset for both training and testing purposes. The training set comprises 12,025 infrared images, providing a robust foundation for model development. The test set includes 3,464 infrared images, which are used to evaluate the performance and generalizability of the proposed detection method.

The VEDAI[10] dataset is an extensive collection of aerial multispectral images specifically designed for vehicle detection tasks. This study utilizes only the infrared modality of the VEDAI dataset to train and test the proposed detection algorithm. The dataset encompasses nine distinct vehicle classes, offering a diverse range of targets for detection. It consists of over 1,200 images, with each image containing an average of approximately 5.5 targets, ensuring a varied and comprehensive training set. The training set includes 1,089 infrared images, which are used to train the detection model. The test set, comprising 121 infrared images, serves to assess the model's accuracy and effectiveness in identifying vehicles under different conditions.

By leveraging the rich and diverse image data provided by the LLVIP and VEDAI datasets, this study aims to enhance the capabilities of infrared target detection algorithms. The extensive training and testing sets enable thorough evaluation and validation of the proposed methods, ensuring that the improvements in detection accuracy and speed are well-supported by empirical evidence.

4.2 Evaluation metrics

In this study, the mean Average Precision (mAP) was employed as the primary evaluation metric to assess the performance of the proposed model. The mAP is a widely recognized and robust metric used in object detection tasks, providing a comprehensive measure of the model's accuracy across all target classes.

The calculation process begins with determining the Average Precision (AP) for each class. This involves evaluating precision and recall at various confidence threshold levels. Precision is defined as the ratio of true positive detections to the total number of positive detections (both true positives and false positives). Recall, on the other hand, is the ratio of true positive detections to the total number of actual positive instances (true positives and false negatives). A precision-recall curve is then plotted by varying the confidence threshold, and the AP is computed as the area under this curve (AUC).

AP is computed individually for each class in the dataset, ensuring a comprehensive evaluation of the model's performance across different object categories. This detailed assessment provides insights into how well the model detects various types of objects.

To calculate the mean Average Precision (mAP), we use the following formula:

$$mAP = \frac{1}{n}\sum_{i=0}^{n} AP_i = \frac{1}{n}\sum_{i=0}^{n} \int_0^1 P_i(r)\, dr \tag{2}$$

By utilizing mAP as the evaluation metric, this study ensures a rigorous and standardized assessment of the model's performance. The detailed calculation of AP for each class, followed by the averaging process, highlights the model's strengths and weaknesses across different categories, offering a nuanced understanding of its detection capabilities. This approach not only facilitates comparison with other models but also aids in identifying areas for further improvement, ultimately advancing the field of infrared target detection in complex industrial environments.

4.3 Results

Table 1: LLVIP comparative experimental results.

Method	mAP@0.5(%)	mAP@0.5:0.95(%)	Training Time(h)	Inference Speed(ms)
Faster RCNN	85.3	52.5	8.82	5.2
YOLOv3	93.3	60.4	7.63	4.1
YOLOv5	94.6	61.8	7.02	3.5
YOLOv8	94.8	62.9	6.51	3.3
CFT	95.9	63.5	6.12	3.0
Ours	**96.1**	**63.8**	**5.95**	**2.4**

To comprehensively evaluate the effectiveness of the proposed method, we conducted a detailed comparison with several mainstream object detection algorithms, including Faster R-CNN, YOLOv3, YOLOv5, YOLOv8, as well as the improved method CFT. As shown in Table 1, our proposed algorithm significantly outperforms these commonly used algorithms in the domain of infrared image target detection.

Specifically, in terms of key performance metrics such as mAP@0.5 and mAP@0.5:0.95, our method achieved improvements of 10.8%, 2.8%, 1.5%, 1.3%, and 0.2% over Faster R-CNN, YOLOv3, YOLOv5, YOLOv8, and CFT respectively. Additionally, for mAP@0.5:0.95, the improvements were 11.3%, 3.4%, 2.0%, 0.9%, and 0.3% respectively. These results clearly demonstrate the superior accuracy of our algorithm in detecting targets within infrared images.

The introduction of the SPD module has effectively enhanced the model's ability to perceive objects at different scales, particularly when dealing with low-resolution images. This ensures comprehensive and precise feature representation, thereby optimizing overall detection performance. The significant improvements in detection accuracy validate the efficacy of our proposed approach, highlighting its potential to advance the field of infrared target detection.

In addition to enhancing detection accuracy, our algorithm also demonstrates significant improvements in terms of training time and inference speed. We conducted our experiments by training the algorithm for 100 epochs on a 4090Ti GPU with a batch size of 32.

Regarding training time, our algorithm showed remarkable efficiency compared to the five baseline algorithms: Faster R-CNN, YOLOv3, YOLOv5, YOLOv8, and CFT. Specifically, our method reduced training time by 48.24%, 28.07%, 18.07%, 9.41%, and 2.86%, respectively. This considerable reduction in training time underscores the efficiency of our approach, making it more practical for real-world applications where training time is a critical factor.

When evaluating inference speed, which measures the time taken to complete inference for a single image, our algorithm also performed exceptionally well. Our method takes only 2.4ms per image, representing a significant reduction of 2.8ms, 1.7ms, 1.1ms, 0.9ms, and 0.6ms compared to Faster R-CNN, YOLOv3, YOLOv5, YOLOv8, and CFT, respectively.

These improvements in both training time and inference speed are attributable to the efficient design of our algorithm, including the integration of the SPD module. The SPD module not only enhances the model's detection accuracy by preserving fine-grained details and improving multi-scale feature representation but also contributes to faster processing times by streamlining the feature extraction process.

To further validate the robust performance of our method, we also conducted experiments on the VEDAI dataset. As shown in Table 2, our method consistently demonstrated strong effectiveness. The results reaffirmed the efficacy of our approach, maintaining high accuracy and efficiency across different datasets. Specifically, our method achieved remarkable mAP50 metrics, confirming its versatility and reliability in various application scenarios. These additional experiments on the VEDAI dataset underscore the generalizability and robustness of our proposed method for infrared image target detection.

Table 2: VEDAI comparative experimental results.

Method	VEDAI	
	mAP@0.5(%)	mAP@0.5:0.95(%)
YOLOv8	62.9	46.5
CFT	63.3	51.5
Ours	**63.7**	**52.1**

Overall, our algorithm's ability to significantly reduce training time and inference speed, while still improving detection accuracy, highlights its robustness and suitability for deployment in time-sensitive industrial environments. These performance gains demonstrate that our proposed method is not only more accurate but also more efficient, making it a valuable contribution to the field of infrared image target detection.

5. Conclusion

We propose an improved YOLOv8-based object detection algorithm specifically designed for detecting cars and pedestrians in low-resolution infrared images. By introducing the innovative SPD module, this algorithm significantly enhances both detection accuracy and efficiency, while also improving inference speed. The SPD module effectively preserves fine-grained information through its unique downsampling method, which also contributes to faster inference. Experimental results on the LLVIP and VEDAI datasets demonstrate the outstanding performance of our method. These results highlight the superior capability of our approach in infrared image target detection. Our research provides an effective solution for target detection in low-resolution infrared images, showing remarkable advantages, especially in handling complex scenarios such as car and pedestrian detection.

References

[1] Everingham, M., Van Gool, L., Williams, C. K., Winn, J., & Zisserman, A. (2010). The pascal visual object classes (voc) challenge. International journal of computer vision, 88, 303-338.

[2] Geiger, A., Lenz, P., Stiller, C., & Urtasun, R. (2013). Vision meets robotics: The kitti dataset. The International Journal of Robotics Research, 32(11), 1231-1237.

[3] Qingyun, F., Dapeng, H., & Zhaokui, W. (2021). Cross-modality fusion transformer for multispectral object detection. arXiv preprint arXiv:2111.00273.

[4] Seifert, C., Aamir, A., Balagopalan, A., Jain, D., Sharma, A., Grottel, S., & Gumhold, S. (2017). Visualizations of deep neural networks in computer vision: A survey. Transparent data mining for big and small data, 123-144.

[5] Zhou, F. Y., Jin, L. P., & Dong, J. (2017). Review of convolutional neural network.

[6] Jiang, P., Ergu, D., Liu, F., Cai, Y., & Ma, B. (2022). A Review of Yolo algorithm developments. Procedia computer science, 199, 1066-1073.

[7] Girshick, R. (2015). Fast r-cnn. In Proceedings of the IEEE international conference on computer vision (pp. 1440-1448).

[8] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C. (2016). Ssd: Single shot multibox detector. In Computer Vision – ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11 – 14, 2016, Proceedings, Part I 14 (pp. 21-37). Springer International Publishing.

[9] Jia, X., Zhu, C., Li, M., Tang, W., & Zhou, W. (2021). LLVIP: A visible-infrared paired dataset for low-light vision. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 3496-3504).

[10] Razakarivony, S., & Jurie, F. (2016). Vehicle detection in aerial imagery: A small target detection benchmark. Journal of Visual Communication and Image Representation, 34, 187-203.

[11] Jocher, G., Chaurasia, A., & Qiu, J. (2023). Ultralytics YOLO (Version 8.0. 0)[Computer software]. URL: https://github. com/ultralytics/ultralytics.

[12] Sunkara, R., & Luo, T. (2022, September). No more strided convolutions or pooling: A new CNN building block for low-resolution images and small objects. In Joint European conference on machine learning and knowledge discovery in databases (pp. 443-459). Cham: Springer Nature Switzerland.

[13] Ding, X., Zhang, X., Ma, N., Han, J., Ding, G., & Sun, J. (2021). Repvgg: Making vgg-style convnets great again. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 13733-13742).

[14] Cao, H. Y., Shen, X. L., & LIU, C. (2020). Improved infrared target detection algorithm of YOLOv3. Journal of Electronic Measurement and Instrumentation, 34(8), 188-194.

[15] Jiaojiao, G. U., Bingzhen, L. I., Ke, L. I. U., & Wenzhi, J. I. A. N. G. (2021). Infrared ship target detection algorithm based on improved faster R-CNN. Infrared Technology, 43(2), 170-178.

[16] Cai, W., Xv, P. W., Yang, Z. Y., Jiang, X. H.,& Jiang, B.(2021). Dim target detection in infrared image with complex background Applied Optics 42(4) pp 643-50

[17] Tan, F., Mu, P.A., & Ma, Z.X. (2021).Multi-target tracking algorithm based on YOLOv3 detection and feature point matching Acta Metrologica Sinica 42(02) pp 157-62

[18] Chen, Z., Guo, H., Yang, J., Jiao, H., Feng, Z., Chen, L., & Gao, T. (2022). Fast vehicle detection algorithm in traffic scene based on improved SSD. Measurement, 201, 111655.

[19] Saxena, S., Dey, S., Shah, M., & Gupta, S. (2023). Traffic sign detection in unconstrained environment using improved YOLOv4. Expert Systems with Applications, 121836.

[20] Ge, H., & Wu, Y. (2023). An Empirical Study of Adoption of ChatGPT for Bug Fixing among Professional Developers. Innovation & Technology Advances, 1(1), 21–29. https://doi.org/10.61187/ita.v1i1.19

[21] Hu, M., Wu, Y., Yang, Y., Fan, J., & Jing, B. (2023). DAGL-Faster: Domain adaptive faster r-cnn for vehicle object detection in rainy and foggy weather conditions. Displays, 79, 102484.

[22] Sakaridis, C., Dai, D., & Van Gool, L. (2018). Semantic foggy scene understanding with synthetic data. International Journal of Computer Vision, 126, 973-992.

[23] Hu, M., Wang, C., Yang, J., Wu, Y., Fan, J., & Jing, B. (2022). Rain rendering and construction of rain vehicle color-24 dataset. Mathematics, 10(17), 3210.

[24] Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In Proceedings of the IEEE international conference on computer vision (pp. 2980-2988).

[25] Meng, X., Liu, Y., Fan, L., & Fan, J. (2023). YOLOv5s-Fog: an improved model based on YOLOv5s for object detection in foggy weather scenarios. Sensors, 23(11), 5321.

[26] Jocher, G., Chaurasia, A., Stoken, A., Borovec, J., Kwon, Y., Michael, K., ... & Mammana, L. (2022). ultralytics/yolov5: v6. 2-yolov5 classification models, apple m1, reproducibility, clearml and deci. ai integrations. Zenodo.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of Woody International Publish Limited and/or the editor(s). Woody International Publish Limited and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.