IST-YOLO: Infrared Small Target Detector based on Improved YOLOv8

Ruoyu Wang 2023106231@zut.edu.cn
Bicao Li[*] lbc@zut.edu.cn
Bei Wang wangbei@zut.edu.cn
Danting Niu 2024106245@zut.edu.cn
Zhongyuan University of Technology

Yongzhao Wang wangyongzhao1987@126.com
Anyang Normal University

Editors: Vu Nguyen and Hsuan-Tien Lin

Abstract

Compared with natural images, the target of a single-frame infrared small target image occupies fewer pixels, has fuzzy imaging, less shape and texture information, and a more complex background. This leads to lower detection accuracy and makes it difficult to achieve accurate target localization. Therefore, in this paper, an infrared small target detection algorithm, IST-YOLO, is proposed based on yolov8. First, our algorithm improves the structure of standard model by adding an upsampling layer and a higher resolution detection head, which has a better ability to detect small targets. Second, we designed the Adaptive Residual Module (ARM) by combining the residual structure with the frequency adaptive dilated convolution to enhance the capacity of extracting deep small target position information while retaining the rich semantic information in the shallow layers. Finally, the Local and Globa Fusion (LGFusion) module is designed to enhance the information interaction between local and global features of the model. Experiments show that the accuracy of IST-YOLO outperforms both standard and popular algorithms.

Keywords: Infrared Small Targets Detection; YOLOv8; Adaptive Residual Module; Local and Global Fusion;

1. Introduction

In recent years, the rapid development of digital image processing technology has promoted the research about infrared target detection. The International Society for Optical Engineering (SPIE) defines an infrared small target as a target occupying no more than 9 × 9 pixels in a 256 × 256 pixel infrared image with low target resolution Zhang et al. (2003). Single-frame type detection algorithms detect small targets within a single image Zhao et al. (2022). There are four main types of algorithms for common single-frame type infrared small target detection methods: filter-based Bae et al. (2012), human visual system-based Nasiri et al. (2017), image data structure-based Xia et al. (2022) and deep learning-based Wu et al. (2021). Bai and Zhou (2010) proposed the classical Top Hat Operational Transform (Top Hat), which applies morphological methods to the detection of small targets. Hou and Zhang (2007) proposed a significance detection method based on spectral residuals, which is simple to operate and easy to implement, but the detection effect of the infrared images

[*] Corresponding author

with complex backgrounds is not ideal, and it is prone to misdetection. Gao et al. (2013) proposed a new Infrared Patch-Image (IPI) model, which utilizes a local block construction method and a stable principal component tracking method to recover the target sparse component and the background low-rank component in the data matrix, and finally uses image reconstruction and segmentation to complete the detection of small targets.

With the development of artificial intelligence, small target detection algorithms based on deep learning have been increasingly emphasized by researchers at home and abroad. Deep learning networks currently used for object detection are categorized into two-stage object detection algorithms and single-stage object detection algorithms. The two-stage object detection algorithms are represented by R-CNN Girshick et al. (2014) and its variants Fast R-CNN Girshick (2015), Faster R-CNN Ren et al. (2015). The one-stage object detection algorithms are represented by the SSD Liu et al. (2016), YOLO family Redmon et al. (2016); Redmon and Farhadi (2017, 2018); Bochkovskiy et al. (2020); Jocher (2020); Li et al. (2022); Wang et al. (2023, 2024b,a) of algorithms. The Faster R-CNN algorithm first presets prediction boxes with different aspect ratios, then classifies the prediction boxes using image features, filters out the candidate frames related to the background so as to generate sparse region candidate frames, and finally classifies the objects in the generated candidate frames in the second stage and predicts the target location. The detection framework of SSD consists of two parts, the front-end network is a convolutional neural network VGG16 used for the initial extraction of features from the image, and the back-end is a multi-scale feature detection network used to extract features at different scales from the feature maps generated by the front-end network. After the image is processed by SSD network, the category probability and position coordinates of the object can be obtained directly, and the final result can be obtained directly after a single detection, which is characterized by fast detection speed. The YOLO algorithm views object detection as a regression problem, it trains directly on the whole image as input, treating the target and the background as a whole, rather than using a sliding window, which greatly improves the detection speed, but there is a certain degree of degradation in the detection accuracy.

However, since infrared imaging systems rely on temperature difference imaging, compared with visible light images, infrared images often have problems such as small targets, blurred imaging , and little shape and texture information Hsieh et al. (1997). The current deep learning networks for object detection, such as Faster R-CNN, SSD, YOLO series algorithms, etc., have better detection results for medium and large targets in general scenes where the targets are dispersed and do not have overlapping or occlusion phenomena, but the average detection accuracy of small targets in infrared images is poor, and the detection accuracy is much lower than that of medium and large targets in general scenes Nasrabadi (2013). Therefore, finding a more suitable network for infrared small and small target detection is still a worthwhile research problem.

To address the above problems, we propose IST-YOLO, an infrared small target detection network based on yolov8. Our proposed method has three improvements. The improved model structure incorporates an upsampling layer and a higher resolution detection head to detect small targets with smaller pixel occupancy. The Adaptive Residual Module (ARM) combines the idea of frequency adaptive dilated convolution with residual structure He et al. (2016) to extract the deep small target positional information in the feature extractor while preserving the shallow rich semantic information. The Local and

Global Fusion (LGFusion) module models local and global features, fuses shallow and deep features of the model, and enhanced information interaction between local and global features of small targets. In summary, the contributions of this paper can be summarized as follows:

(1) An object detection network with a high-resolution detection head is introduced to improve the accuracy of small target detection. This network adds upsampling layers and a high-resolution detection head to the original network, and a feature map with larger size is employed for small target detection. This method provides better detection of small targets.

(2) An ARM is proposed that combines frequency adaptive dilation convolution with residual structure to enhance the ability of extracting deep positional information in the feature extractor while retaining the rich semantic information in the shallow layer, so as to make a better effect of doing cross-layer connectivity during up-sampling.

(3) The LGFusion module is designed, motivated by the idea of convolutional attention fusion, which models local and global features and enhance information interaction between local and global features of small targets, and this approach achieve the good detection performance of the small target.

(4) Experimental results on the NUAA-SIRST dataset show that these enhancements make IST-YOLO outperform the original algorithm and other algorithms object detectors in terms of detection accuracy.

2. Method

2.1. IST-YOLO

The overall architecture of our proposed IST-YOLO is shown in Fig. 1. An upsampling layer and a higher resolution detection head are introduced for better detection of small targets.The backbone network is passed through the C2f to extract features. ARM is passed before SPPF in order to extract advanced position information. Deep and shallow features are spliced during upsampling. The LGFusion module is added before the detection head to model local and global features, enhance information interaction between local and global features of small targets. Finally, the multi-scale detection head is then used to detect targets of different sizes.

2.2. Improvement of Model Structure

The yolov8 contains three detector heads with resolutions of 20×20, 40×40, and 80×80. The low resolution detection head is used to detect objects with a large pixel ratio, and the high resolution detection head is used to detect objects with a small pixel ratio. Since the target of a single-frame infrared small target image accounts for a small pixel ratio of the whole image, it is difficult for the original detection head to detect it, so we improved the network structure of yolov8 by adding upsampling layers, increasing the size of the feature map from 80×80 to 160×160, and adding a detection head with a resolution of 160×160. This method can better detect small targets, and the experimental results show that the improved structure compared to the standard model without degrading mAP_{50} while the

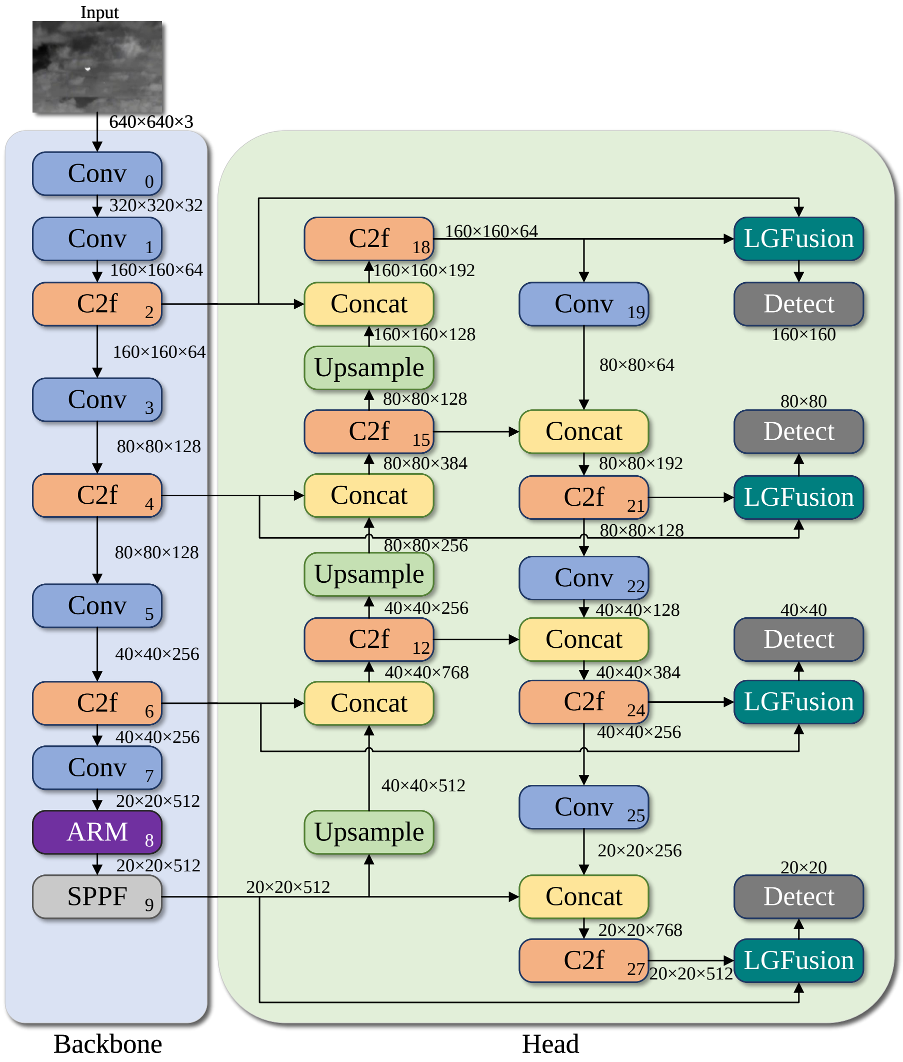


Figure 1: The proposed IST-YOLO, consisting of two parts: backbone and head. ARM and LGFusion are our self-developed modules. The parameters in the figure are the ones when the model size is s.

mAP_{50-95} is improved from 0.358 to 0.383. The standard model and the improved structure of the model network are shown in Fig. 2.

2.3. Adaptive Residual Module (ARM)

Dilated convolution Yu and Koltun (2015) was initially proposed to solve problems in image segmentation to obtain richer positional information and aggregate multi-scale contextual information from an expanded sense field. The widely used extended convolution can be

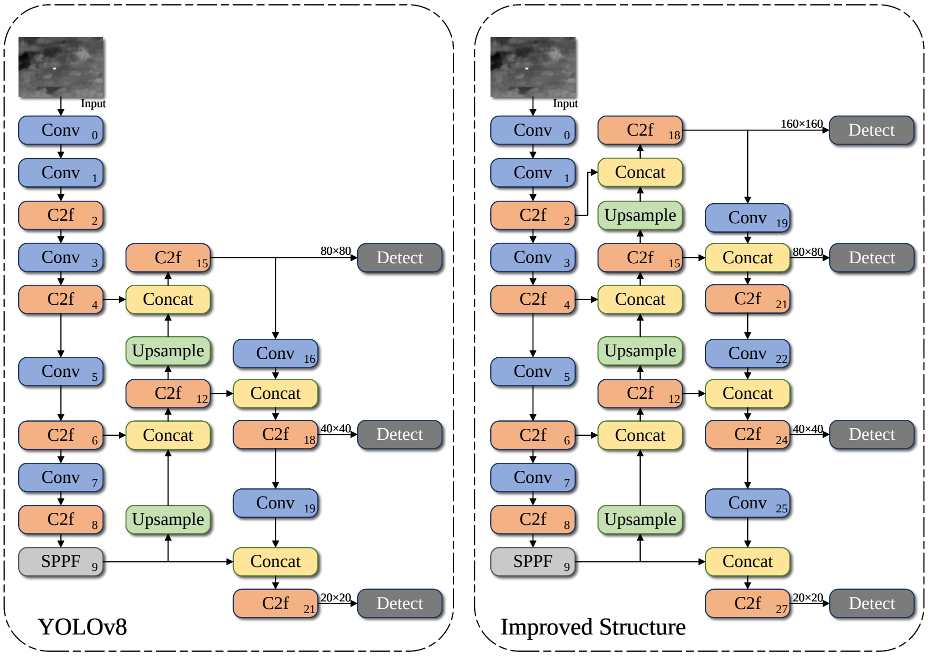


Figure 2: Structure of the standard and improved models. We added upsampling layers and a high-resolution detection head to the standard model.

represented by the following Eq. 1:

$$Y\left(p\right)=\sum_{i=1}^{K\times K}W_{i}X\left(p+\mathit{\Delta}p_{i}\times D\right) \tag{1}$$

where $Y(p)$ is the pixel value at position p in the output feature map, K is the kernel size, W_i is the weight parameter of the kernel, and $X\left(p+\mathit{\Delta}p_{i}\times D\right)$ is the pixel value at the position corresponding to the position where p is offset by $\mathit{\Delta}p_i$ in the input feature map. The receptive field can be expanded by increasing the expansion rate. However, the increase of expansion rate to expand the receptive field also leads to the degradation of the frequency information capturing ability. Chen et al. (2024a) proposed a Frequency Adaptive Dilated Convolution, which achieves a balance between the effective bandwidth and the receptive field by dynamically adjusting the expansion rate. As shown in Eq. 2, $\hat{D}(p)$ can be predicted using a convolutional layer with parameter θ.

$$Y\left(p\right)=\sum_{i=1}^{K\times K}W_{i}X\left(p+\mathit{\Delta}p_{i}\times\hat{D}\left(p\right)\right) \tag{2}$$

The optimization of θ can be written as Eq. 3. $HP+$ and $HP-$ denote pixels with the highest/lowest high-frequency power, respectively.

$$\theta = \max_{\theta}\left(\sum_{p\in HP-}\hat{D}\left(p\right) - \sum_{p\in HP+}\hat{D}\right) \tag{3}$$

For a static convolutional kernel, the weights W can be decomposed into Eq. 4. Here $\bar{W}$represents the kernel-wise averaged W. The term $\hat{W}$ denotes the residual part, extracting the high-frequency components.

$$W = \bar{W} + \hat{W} \tag{4}$$

Adaptive kernel dynamically adjusts the high-frequency and low-frequency components, which can be formalized as Eq. 5. where μ_l, μ_h are the dynamic weights of each channel, predicted by a simple lightweight global pooling + convolutional layer.

$$W = \mu_l\bar{W} + \mu_h\hat{W} \tag{5}$$

In the yolov8 algorithm downsampling process, the loss of position information leads to low detection accuracy. This problem is partially solved by adding a high-resolution detection head, which improves the detection accuracy. However, the feature extractor is strong in

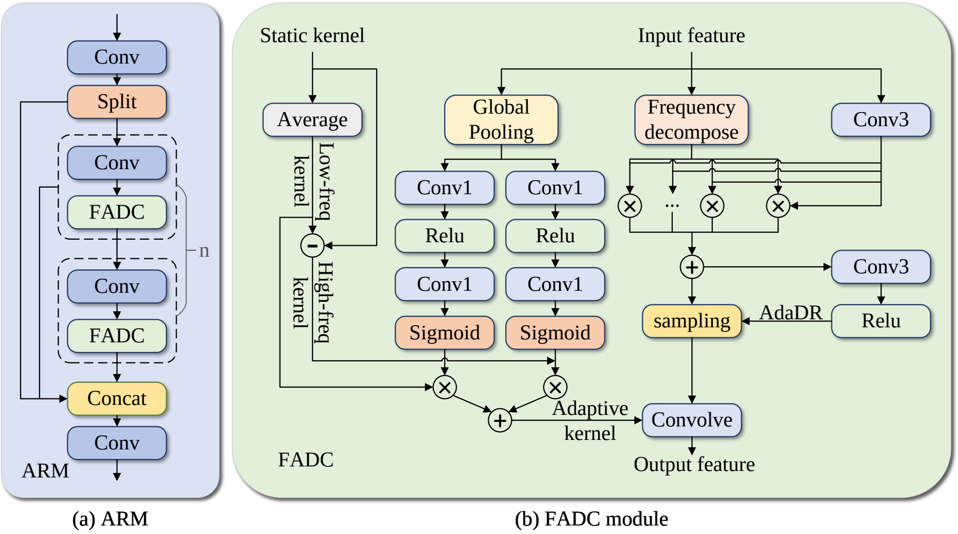


Figure 3: (a) is the proposed ARM. For the input feature map, firstly go through 1×1 convolution layer to reduce the number of channels to reduce the amount of computation, and secondly after split operation to input into the residual structure, and then after convolution for channel splicing, and finally through the 1×1 convolution layer to reduce the dimensionality. (b) is the FADC module in ARM, AdaDR is an adaptive dilation rate.

semantic information extraction and poor in location information extraction. Inspired by Chen et al. (2024a), we designed the Adaptive Residual Module (ARM), which expands the deep sensing field of the feature extractor and extracts more location information while conveying rich semantic information. It makes up for the shortcomings of the backbone network where small targets are rich in semantic information but insufficient in location information. The module is shown in Fig. 3. We add the ARM to the eighth layer in the backbone network, as shown in Fig. 1, to retain the rich semantic information in the shallow layer while better extracting the small target location information.

2.4. Local and Global Fusion (LGFusion)

Convolutional operations are limited by local properties and receptive fields, and are deficient in modeling global features. On the contrary, attention mechanisms excel at extracting global features and capturing remote dependencies. Based on these two points Hu et al.

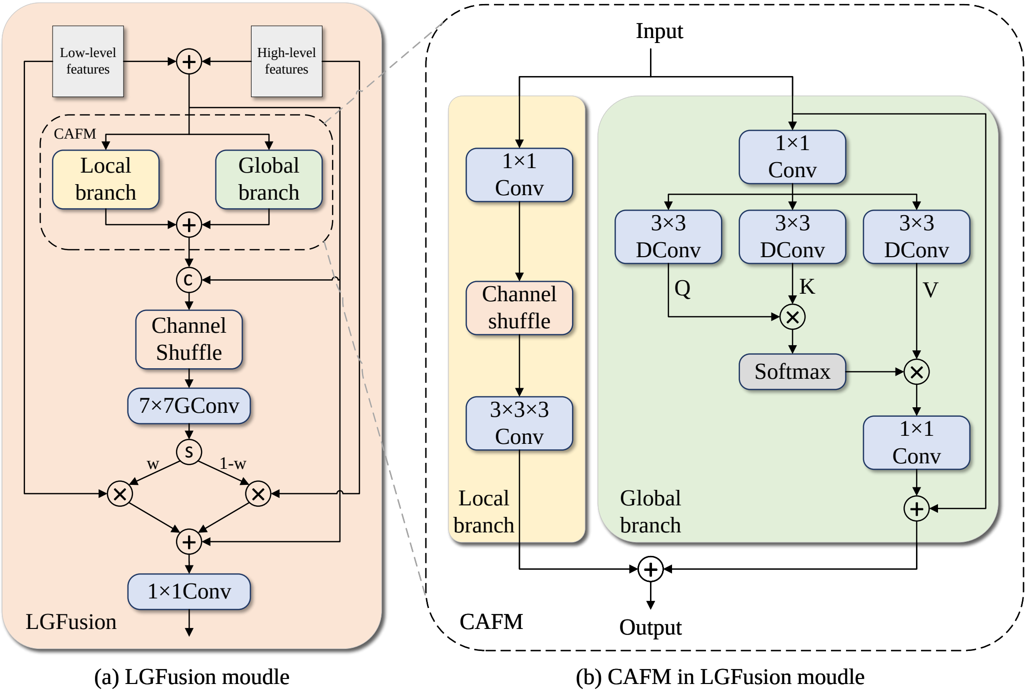


Figure 4: (a) is the proposed LGFusion module, c stands for Channel-wise Concatenation and s for Sigmoid. Firstly the input feature maps are preliminarily fused, and the local and global features are extracted and fused by the CAFM module. Secondly, channel cascading, channel shuffling and activation are performed to obtain weights W, which fully mix the weights of global and local branches to ensure information interaction. Finally the weights W are then used to further fuse the input feature map. (b) is the CAFM module in LGFusion. It consists of two branches, Local branch and Global branch.

(2024) proposed a convolution and attention fusion module (CAFM). In the CAFM a self-attention mechanism is used in the global branch to capture a wider range of information about the hyperspectral data, while the local branch focuses on extracting local features for comprehensive denoising. The module explores the problem of global and local modeling using a combination of convolution and attention.The network of yolov8 extracts local features more adequately for small targets and less for global features. This leads to poor detection of small targets. Inspired by Hu et al. (2024) and combining the idea of content-guided attention Chen et al. (2024b) we designed an LGFusion module, as shown in Fig. 4. The module employs an attentional mechanism to obtain global dependencies while extracting local features, enhancing cross-channel interactions and facilitating the integration of local and global information for small targets. The LGFusion module is added to the model before the detection header, connecting layers 2 and 18, 4 and 21, 6 and 24, 9 and 27, respectively, as shown in Figure 1. The network model with LGFusion added obtains global dependencies and models local and global features based on shallow and deep layers, enhance information interaction between local and global features of small targets.

3. Experiments and Analysis

3.1. Datasets

Publicly available datasets of infrared small target imagery are very limited, and most traditional methods are evaluated on their internal datasets.The infrared small target images studied in this paper are from the NUAA-SIRST dataset, which is the first high-quality single-frame infrared small target dataset constructed by Dai et al. (2021).

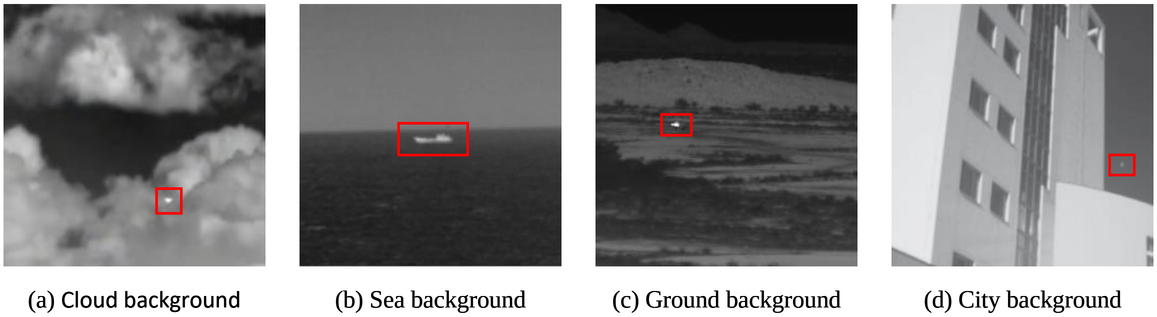

(a) Cloud background
(b) Sea background
(c) Ground background
(d) City background

Figure 5: Example of a typical scenario for the NUAA-SIRST dataset. They are cloud background, sea background, ground background and city background.

The dataset contains 427 infrared images with 480 targets, constructed by selecting the most representative images from hundreds of image sequences. The targets have weak brightness, low contrast, lack of color and texture information, and are hidden in complex background environments such as clouds, cities, and oceans, which are interfered with by strong light sources, sheeted clouds, the sea surface, and the edge of the sky, and many of them are even difficult to be recognized by the human eye. These targets are difficult to detect. Fig. 5 illustrates the typical scene images contained in the above dataset, with a preponderance of images with clouds in the background and fewer of the other three.

3.2. Evaluation Metrics

In this study, two metrics, mAP_{50} and mAP_{50-95}, are used to evaluate the performance of the network. mAP is the mean Average Precision, which is the average value of AP. A target detection model is usually described in terms of the precision (mAP) metric to describe the strengths and weaknesses, and the higher the mAP value, the better the detection of that object detection model on a given dataset. For a dataset containing N classes, the mAP is represented as in Eq. 6.

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \tag{6}$$

The mAP_{50-95} is the mAP threshold of 50% to the mAP threshold of 95% at 5% intervals,10 mAP values were obtained and then these ten values were averaged as in Eq. 7.

$$mAP_{50-95} = \frac{1}{10}\sum_{n=0}^{9} mAP_{50+5n} \tag{7}$$

AP is determined by Precision(P), Recall(R), and the thresholds of the intersection over union(IoU) defined as Eq. 8.

$$AP = \sum_{K} \left(R_K - R_{k-1}\right) P_k \tag{8}$$

P_k and R_k denote the ratio of true-positive objects at the kth threshold to all detected true-true objects and ground true-true objects, respectively. They are calculated as in Eq. 9 and 10.

$$P = \frac{TP}{TP + FP} \tag{9}$$

$$R = \frac{TP}{TP + FN} \tag{10}$$

where TP denotes correctly categorized objects with IoU above a particular threshold. FP is a predicted bounding box with IoU below a certain threshold, or an incorrectly detected object. FN denotes a missed ground truth object.

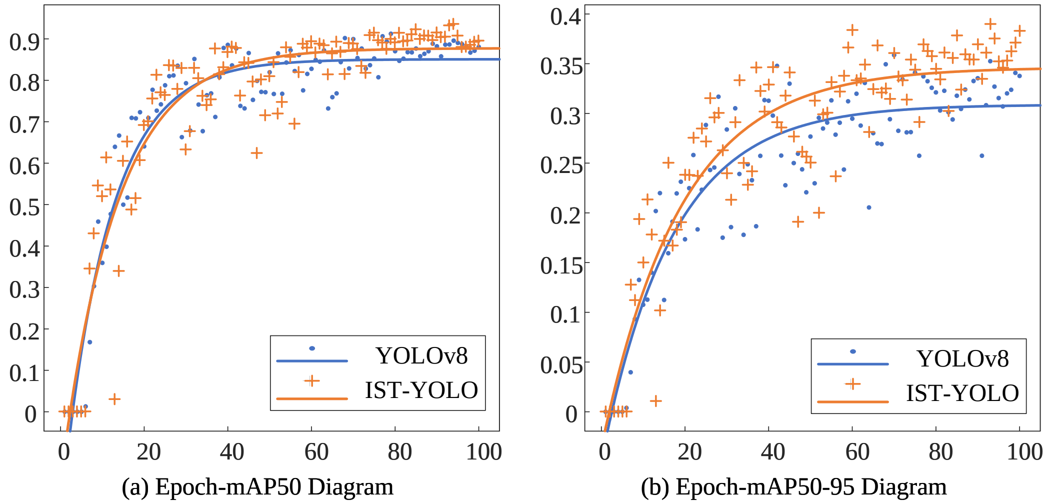
3.3. Ablation Experiments

In order to compare the improved algorithm proposed in this paper with the YOLOv8 algorithm and to ensure the reliability of the experiments, all the training and testing data are trained in the same training environment and the epoch of each training is guaranteed to be 100. after the training is completed, the optimal model weights are used for testing. So as to fully verify the applicability of the detection network designed in this chapter to different complex scenarios and the effectiveness of the detection of infrared small targets, ablation experiments are carried out on the NUAA-SIRST dataset, and the detection results are shown in Tab. 1.

Table 1: Ablation experiments for each module on the NUAA-SIRST dataset.

Structural improvements	ARM	LGFusion	mAP$_{50}$	mAP$_{50\text{-}95}$
			0.899	0.358
✓			0.899	0.383
	✓		0.916	0.382
		✓	0.912	0.377
✓	✓	✓	**0.929**	**0.384**

As shown in Tab. 1, the improvement of structure improves mAP_{50-95} by 0.25 without degrading mAP_{50}. mAP_{50} and mAP_{50-95} are both significantly improved by adding the ARM module and LGFusion module, respectively. Finally, the accuracy of the model is optimized by adding the three improvement points simultaneously. Compared with Baseline, mAP_{50} increases by 0.3 and mAP_{50-95} increases by 0.26, indicating that the model structure of IST-YOLO is more reasonable for the learning of infrared small targets. The mAP_{50}, mAP_{50-95} during training is shown in Fig. 6. The bounding box loss, PR curve is shown in Fig. 7.



(a) Epoch-mAP50 Diagram

(b) Epoch-mAP50-95 Diagram

Figure 6: Comparison of accuracy between IST-YOLO and YOLOv8. (a) is a comparison plot of mAP_{50} curves and (b) is a comparison plot of mAP_{50-95} curves.

It can be seen that IST-YOLO's mAP_{50} training is not much different from YOLOv8 in the early stage, but it is significantly better than YOLOv8 in the later stage. and IST-YOLO's mAP_{50-95} is consistently much better than YOLOv8. Thus, it can be seen that IST-YOLO's learning of the single-frame infrared small targets is more reasonable.

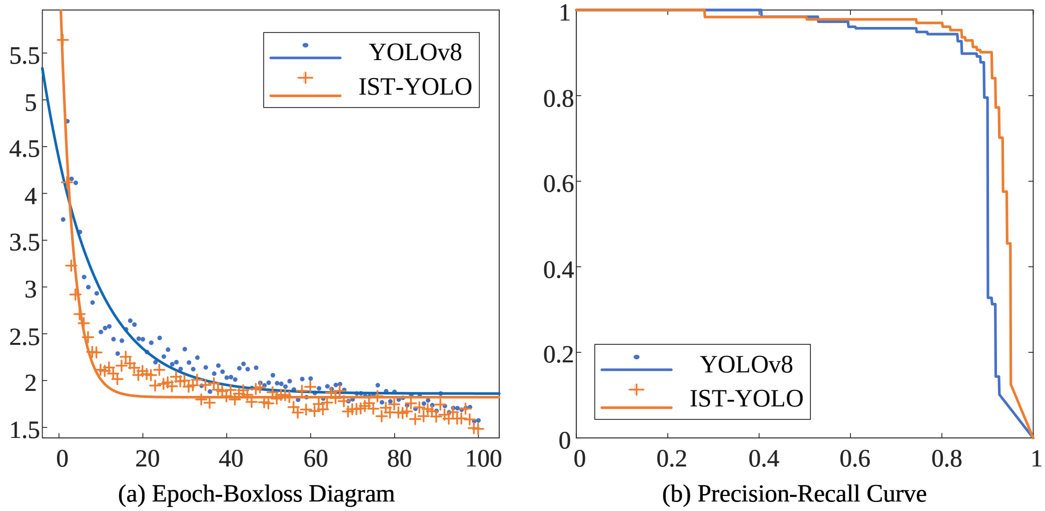


Figure 7: Loss and PR curves. (a) is the loss of the prediction box during training, and (b) is a comparison of the PR curves of the models.

As shown in Fig. 7 (a), the bounding box loss of IST-YOLO converges faster and the loss is lower. The PR curve is the curve of Precision and Recall for different IOU thresholds, and the larger the area enclosed by the PR curve, the better the model is for that dataset. As shown in (b), IST-YOLO is better than YOLOv8 on the single-frame infrared small target dataset.

3.4. Comparison Experiment

We compared IST-YOLO with different algorithms on the NUAA-SIRST dataset, and the results are shown in Tab. 2. All methods in the table have a training epoch of 100.

Table 2: Compare results with different models on NUAA-SIRST dataset.

Method	mAP_{50}	$\text{mAP}_{50\text{-}95}$
YOLOv3	0.843	0.339
YOLOv5s	0.901	0.374
YOLOv6	0.901	0.336
YOLOv7-tiny	0.798	0.256
YOLO-world	0.876	0.344
YOLOv9c	0.866	0.371
RT-DETR-resnet50	0.887	0.365
RT-DETR-resnet101	0.881	0.357
YOLOv10s	0.872	0.356
IST-YOLO(Ours)	**0.929**	**0.384**

It can be seen that both mAP_{50} and mAP_{50-95} have objective improvements compared to previous versions of YOLOv3 Redmon and Farhadi (2018), YOLOv5 Jocher (2020), YOLOv6 Li et al. (2022) and YOLOv7 Wang et al. (2023). Compared with the latest YOLO-World Cheng et al. (2024), YOLOv9 Wang et al. (2024b), RT-DETR-resnet50, RT-DETR-resnet101 Zhao et al. (2024), and YOLOv10 Wang et al. (2024a), mAP_{50} improves by 5.3%, 6.3%, 12.5%, 4.8%, and 5.7%, respectively, and mAP_{50-95} improved by 4%, 1.3%, 4.1%, 2.7%, and 2.8%, respectively. The effectiveness of the proposed network model for infrared small target detection in complex background is fully verified. The PR curves of all the above methods are shown in Fig. 8.

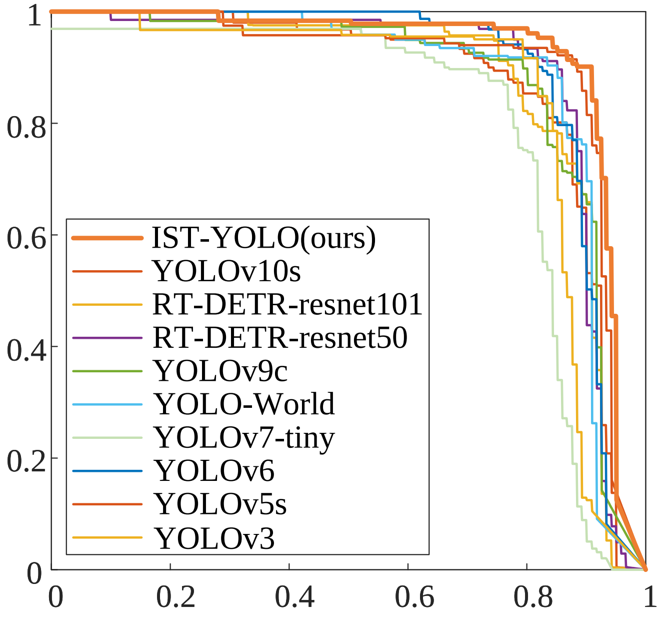


Figure 8: PR curve for IST- YOLO and other methods. PR curves are Precision and Recall curves for different IOU thresholds, the larger the area enclosed by the PR curve, the better the model for the dataset.

The PR curve of IST-YOLO is significantly better than the other methods. The detection comparison between IST-YOLO and other models methods on the NUAA-SIRST dataset is shown in Fig. 9. The three infrared small targets in the first row of data YOLOv8 detects only one, YOLOv10 detects two, and YOLOv9 does not detect them, while IST-YOLO detects all three small targets. It can be seen that our algorithm detects dense small targets better than other algorithms. In the second row of data are targets with a small percentage of pixels, only our IST-YOLO detects the small targets accurately. In the third row of data our algorithm detects better, both YOLOv8 and YOLOv9 algorithms incorrectly detect the background as a target, which shows that our algorithm can reduce the false alarm rate. In the fourth row of data our algorithm has the best detection of targets, while YOLOv8 has false alarms and incorrectly recognizes the background as a target.

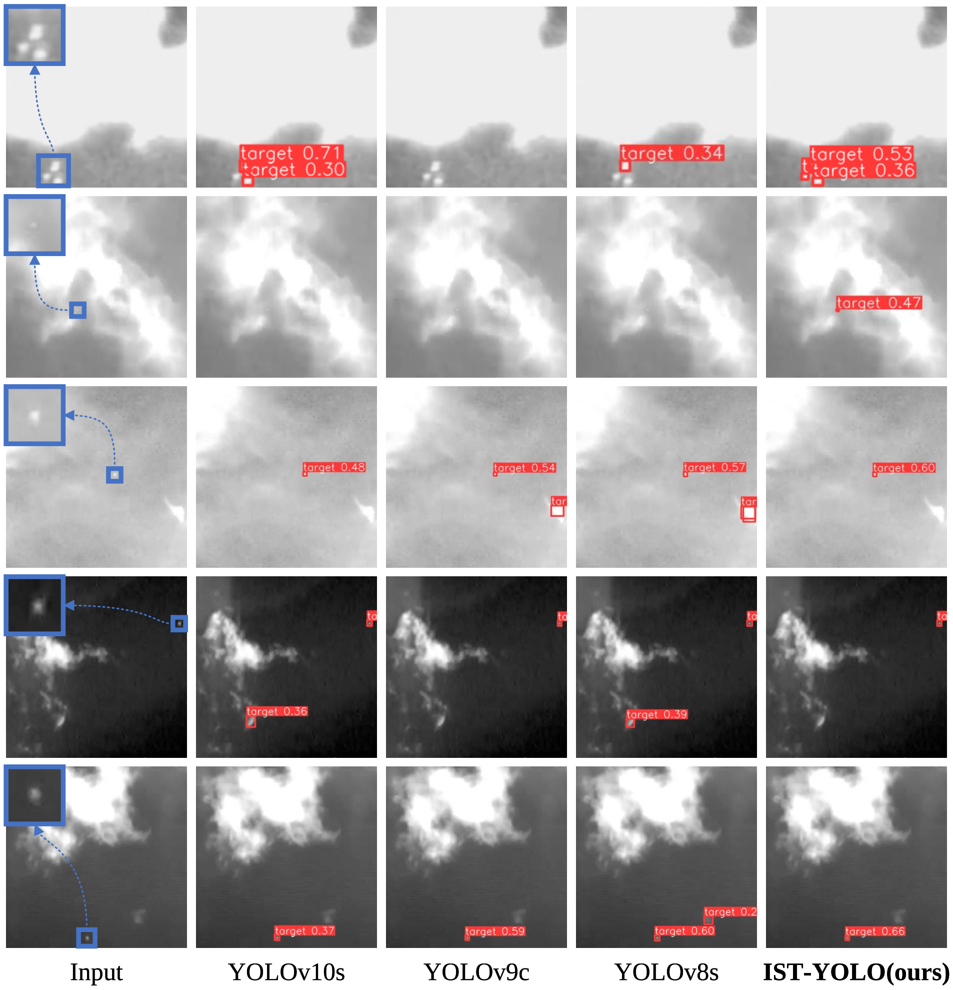


Figure 9: Prediction results for IST-YOLO and other networks on the NUAA-SIRST dataset. The blue box in the first column zooms in on the target.

4. Conclusion

Single-frame infrared small targets are difficult to detect due to small target size, blurred imaging, and little shape and texture information. This paper proposes the IST-YOLO algorithm to better detect small targets by adding a high-resolution detection head. The ARM module is introduced to enhance the position information of infrared small targets for more accurate localization of small targets. The LGFusion module is proposed to establish the dependence of local and global information to increase the detection accuracy of infrared small targets. The experimental results show that the IST-YOLO algorithm outperforms other algorithms. The improvement in accuracy illustrates that our algorithm can be considered as an effective method for infrared small target detection. IST-YOLO achieves high accuracy but suffers from slower inference due to additional layers. Future research aims to develop the faster and more effective frameworks for infrared small target detection.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (No.61901537, NO.62301623), Science and technology innovation talent project of Education Department of Henan Province (No. 23HASTIT030), Henan province key science and technology research projects (232102211002), Training Program of Young Master's supervisor in Zhongyuan University of Technology (No. SD202207).

References

Tae-Wuk Bae, Fei Zhang, and In-So Kweon. Edge directional 2d lms filter for infrared small target detection. *Infrared Physics & Technology*, 55(1):137–145, 2012.

Xiangzhi Bai and Fugen Zhou. Analysis of new top-hat transformation and the application for infrared dim small target detection. *Pattern Recognition*, 43(6):2145–2156, 2010.

Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020.

Linwei Chen, Lin Gu, Dezhi Zheng, and Ying Fu. Frequency-adaptive dilated convolution for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3414–3425, 2024a.

Zixuan Chen, Zewei He, and Zhe-Ming Lu. Dea-net: Single image dehazing based on detail-enhanced convolution and content-guided attention. *IEEE Transactions on Image Processing*, 2024b.

Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. Yolo-world: Real-time open-vocabulary object detection. *arXiv preprint arXiv:2401.17270*, 2024.

Yimian Dai, Yiquan Wu, Fei Zhou, and Kobus Barnard. Asymmetric contextual modulation for infrared small target detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 950–959, 2021.

Chenqiang Gao, Deyu Meng, Yi Yang, Yongtao Wang, Xiaofang Zhou, and Alexander G Hauptmann. Infrared patch-image model for small target detection in a single image. *IEEE transactions on image processing*, 22(12):4996–5009, 2013.

Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Xiaodi Hou and Liqing Zhang. Saliency detection: A spectral residual approach. In *2007 IEEE Conference on computer vision and pattern recognition*, pages 1–8. Ieee, 2007.

Chih-Cheng Hsieh, Chung-Yu Wu, Far-Wen Jih, and Tai-Ping Sun. Focal-plane-arrays and cmos readout techniques of infrared imaging systems. *IEEE Transactions on Circuits and Systems for Video Technology*, 7(4):594–605, 1997.

Shuai Hu, Feng Gao, Xiaowei Zhou, Junyu Dong, and Qian Du. Hybrid convolutional and attention network for hyperspectral image denoising. *IEEE Geoscience and Remote Sensing Letters*, 2024.

Nishimura K. Mineeva T. Vilarino Jocher, G. Yolov5. *GitHub repository*, 2020.

Chuyi Li, Lulu Li, Hongliang Jiang, Kaiheng Weng, Yifei Geng, Liang Li, Zaidan Ke, Qingyuan Li, Meng Cheng, Weiqiang Nie, et al. Yolov6: A single-stage object detection framework for industrial applications. *arXiv preprint arXiv:2209.02976*, 2022.

Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 21–37. Springer, 2016.

Mahdi Nasiri, Mohammad Reza Mosavi, and Sattar Mirzakuchaki. Ir small target detection based on human visual attention using pulsed discrete cosine transform. *IET Image Processing*, 11(6):397–405, 2017.

Nasser M Nasrabadi. Hyperspectral target detection: An overview of current and future challenges. *IEEE Signal Processing Magazine*, 31(1):34–44, 2013.

Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7263–7271, 2017.

Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

Ao Wang, Hui Chen, Lihao Liu, Kai Chen, Zijia Lin, Jungong Han, and Guiguang Ding. Yolov10: Real-time end-to-end object detection. *arXiv preprint arXiv:2405.14458*, 2024a.

Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7464–7475, 2023.

Chien-Yao Wang, I-Hau Yeh, and Hong-Yuan Mark Liao. Yolov9: Learning what you want to learn using programmable gradient information. *arXiv preprint arXiv:2402.13616*, 2024b.

Yifan Wu, Feng Pan, Qichao An, Jiacheng Wang, Xiaoxue Feng, and Jingying Cao. Infrared target detection based on deep learning. In *2021 40th Chinese Control Conference (CCC)*, pages 8175–8180. IEEE, 2021.

Chaoqun Xia, Shuhan Chen, Xiaoqin Zhang, Zhaomin Chen, and Zhiyong Pan. Infrared small target detection via dynamic image structure evolution. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–18, 2022.

Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.

Wei Zhang, Mingyu Cong, and Liping Wang. Algorithms for optical weak small targets detection and tracking. In *International Conference on Neural Networks and Signal Processing, 2003. Proceedings of the 2003*, volume 1, pages 643–647. IEEE, 2003.

Mingjing Zhao, Wei Li, Lu Li, Jin Hu, Pengge Ma, and Ran Tao. Single-frame infrared small-target detection: A survey. *IEEE Geoscience and Remote Sensing Magazine*, 10 (2):87–119, 2022.

Yian Zhao, Wenyu Lv, Shangliang Xu, Jinman Wei, Guanzhong Wang, Qingqing Dang, Yi Liu, and Jie Chen. Detrs beat yolos on real-time object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16965–16974, 2024.