International Journal of Control, Automation, and Systems 22(11) (2024) 3295-3302
http://dx.doi.org/10.1007/s12555-024-0044-8
ISSN:1598-6446 eISSN:2005-4092
http://www.springer.com/12555

IS-YOLO: A YOLOv7-based Detection Method for Small Ship Detection in Infrared Images With Heterogeneous Backgrounds

Indah Monisa Firdiantika and Sungho Kim*

Abstract: Ship detection from infrared images occupies an important role in maritime search and tracking applications. When compared with methods to process daytime RGB photos, processing infrared images has challenges due to the reduced signal-to-clutter ratio (SCR), indistinct outlines, and inadequate spatial resolutions. In addition, the detection of small targets remains challenging owing to their size variations and unclear edges, leading to missed detections and low accuracy. This work suggests the use of infrared-ship YOLO (IS-YOLO), a model to recognize small ships in infrared images. The proposed technique, based on YOLOv7, enhances the ability to detect infrared objects in heterogeneous scenarios. First, we improve the ability of the YOLOv7 backbone to extract features by introducing a new structure for the efficient layer aggregation network (ELAN) with a two convolutions and GhostConv module. Secondly, the max pooling pyramid-ELAN is introduced to integrate multi-scale information. Furthermore, we capture an infrared small ship dataset using the FLIR T620 camera. The experimental results demonstrate that the IS-YOLO model had the best performance in small ship detection from infrared images compared to several state-of-the-art models, as shown by optimal metrics that include average precision (AP@.5, AP@.5:.95, the number of parameters, and the model size): AP@.5, 88.9%; AP@.5:.95, 38.2%; 32.8 M; and 63.1 Mb, respectively. The proposed approach can serve as a valuable reference for the development of small-ship detection methods with infrared images.

Keywords: Deep learning, infrared image, maritime, small ship detection.

1. INTRODUCTION

Currently, the maritime transportation industry is progressing rapidly, with the corresponding growth in shipping traffic leading to a rise in maritime violations. Automated ship detection can aid in acquiring ship allocation data [1]. Ship detection is a significant area of study in remote sensing because it has a wide range of practical uses, including the control of fishing activities, conducting naval warfare, and salvaging vessels. Preliminary research has shown a variety of methods for detecting objects, with a particular focus on synthetic aperture radar [2]. Compared to radar, an infrared object detection system has the ability to overcome barriers like haze and other atmospheric conditions, as well as capturing images regardless of lighting conditions. Due to their strong anti-interference capabilities and the ability to operate in all weather conditions, they are well-suited for a wide range of applications [3].

Apart from the aforementioned advantages of infrared imaging, the challenges associated with object detection in these images encompass several crucial factors inherent to the infrared imaging technique. In infrared images, the signal power of small ship targets is typically weak, lacking sufficient texture and structural information, especially at long distances. In addition, complicated sea clutter such as sparkles of sunlight on the water, islands, trailing waves, and fog along the sea skyline is frequently irregular and has inconsistent shapes, thereby diminishing the precision of ship recognition. Additionally, the irregular shapes and varying sizes of ship targets limit the robustness of detection. Owing to these factors, ship detection using infrared images has garnered significant attention from researchers, leading to the development of numerous ship target detection techniques [4,5].

There are two approaches to identifying a small object: absolute size and relative size. In the former, an object is classified as small if its dimensions are smaller than 32×32 pixels, as indicated by the COCO dataset [6]. In the second method, an object is small if the area of the bounding box is smaller than 0.03, or the ratio of its width and height to the image's width and height is smaller

Manuscript received January 15, 2024; revised May 7, 2024; accepted August 1, 2024. Recommended by Senior Editor Yongsoon Eun. This work was supported by a National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. IRIS RS-2023-00219725). This work was also supported by 2024 Yeungnam University Research Grants.

Indah Monisa Firdiantika and Sungho Kim are with the Department of Electronic Engineering, Yeungnam University, 280 Daehak-ro, Gyeongsan-si 38541, Korea (e-mails: {indahmonisaf, sunghokim}@yu.ac.kr).
* Corresponding author.

than 0.1 [7]. Small-object detection methods have inspired many researchers to propose detection models appropriate for small objects in infrared images. Convolutional neural network (CNN)-based models (the most well-known deep learning framework) exhibit exceptional feature extraction capabilities, leading to increased accuracy and robustness. Modern object detection algorithms are primarily based on deep learning approaches, which are divided into two main categories: two-stage and one-stage models. A two-stage detection technique entails analyzing object frames or images in two stages. Existing methods for detection include the Mask R-CNN [8], Fast R-CNN [9], and Faster R-CNN [10]. At first, the two-stage algorithm seeks potential targets and their characteristics. Subsequently, every candidate becomes input for the network carrying out classification and regression to accurately determine the position, dimensions, and category of each region. Although the two-stage detector has high accuracy, inference speed is slow and unsuitable for high-speed real-time applications. A one-stage method employs one network to perform classification and regression in a single step, as exemplified by YOLO [11,12]. The one-stage network allows for immediate prediction of item position, size, and categorization without the need to construct candidate regions. Therefore, it demonstrates greater efficiency in comparison to the two-stage algorithm.

With the continuous improvement of object detection, especially in one-stage models, several studies related to small target detection from infrared images have been conducted. Ye *et al*. [13] came up with CAA-YOLO for ocean ship detection. CAA-YOLO applies a high-resolution feature layer (P2), triplet attention, and a combined attention mechanism to preserve features of small targets. Gao *et al*. [14] proposed a lightweight model (YOLOv5mobile) for small ship detection in IR images. The backbone of YOLOv5 is replaced by that of MobileNetV3 to make the detection model lightweight. Guo *et al*. [15] presented the multi-attention pyramid context network (MAPC-Net), which incorporates a scale attention mechanism into the original network's multi-scale feature pyramid for small ship detection from infrared images. Furthermore, in [16], the authors proposed ship detection with a one-stage detection-and-attention mechanism that lets it successfully detect small ships in infrared images. Despite the success of the above examples, it has been difficult to further improve detection performance.

Based on the above observations, we further developed small-ship detection from [16] with the contributions that can be summarized as

1) We propose several new modules, namely, the efficient layer aggregation network (ELAN) two convolutions and GhostConv (E-C2G), ResC3, and the max pooling pyramid-ELAN (MPPELAN).
2) We increase the number of infrared small ship images in our previous dataset to improve the performance of object detection in a real environment.
3) We demonstrate the performance of our proposed method and compare it to other methods. Furthermore, we conduct experiments on the publicly available single-frame infrared small target (SIRST v2) dataset. The experimental results show that our proposed network outperforms existing methods both quantitatively and qualitatively.

2. RELATED WORK

2.1. Efficient layer aggregation networks

To improve the accuracy of infrared small object detection against complex backgrounds, it is crucial to implement more efficient, high-quality network architectures that enhance network performance. Highly efficient network architectures exhibit more diverse combinations of gradients, but the cross stage partial (CSP) module in the YOLOv5l [17] backbone network achieves model scaling by stacking residual blocks. Nevertheless, the act of stacking residual layers solely amplifies the longest lines of a gradient, while leaving the shortest paths unaffected. On the other hand, the ELAN [18] achieves a longer shortest gradient path for the entire network by reducing the number of transition layers. The ELAN uses a technique that enhances the efficiency of deep models by regulating the shortest and longest gradient path, hence facilitating faster learning and convergence. Implementing this design method improves the learning capabilities of the network, enhancing the expressiveness of the training model. The YOLOv7 model proposed in [12], introduced extended ELAN (E-ELAN) technique, which is designed to be compatible with models that have an unlimited number of stacked computational blocks. The E-ELAN enhances the network's learning by merging cardinality without compromising the original gradient path.

2.2. Spatial pyramid pooling

The SPP mechanism described in [19] is widely recognized for its exceptional performance when extracting features of various sizes. It achieves this by effectively maintaining the spatial information of an image. The cross stage partial darknet (CSPDarknet) [20] is employed as the primary network, while Spatial Pyramid Pooling is utilized for the fusion of features, marking the first instance of its application. Subsequently, the path aggregation network structure [21] is employed as the bottleneck of the model to perform a more comprehensive integration of feature maps in YOLOv4 [22]. Jocher *et al*. [17] suggested a YOLOv5 network model that incorporates adaptive anchor frame automated learning from the training set, together with the utilization of Leaky ReLU and sigmoid as activation functions. The spatial pyramid pooling fast (SPPF) layer was suggested as a replacement for

the initial SPP layer, with an algorithm capable of achieving rapid and precise detection. Wang *et al*. introduced a novel approach for object detection. Their algorithm introduced the spatial pyramid pooling cross stage partial conv (SPPCSPC) module as a replacement for the original SPPF module [12]. The new module was designed to improve feature fusion. Wang and colleagues [23] introduced SPPELAN, which integrates spatial pyramid pooling into the ELAN framework. The process begins with a convolutional layer that modifies the dimensions of the channels, and is then followed by a sequence of spatial pooling operations to capture contextual information at many scales. The results are combined and fed into an additional convolutional layer to enhance the network's ability to extract detailed characteristics from different spatial hierarchies.

3. PROPOSED FRAMEWORK

3.1. The IS-YOLO architecture

In this section, the YOLOv7-based network [12] called IS-YOLO is modified for small object detection from infrared images. The IS-YOLO network consists of four basic components: the input, the backbone main layer, the neck feature fusion layer, and the head output layer. The overall architecture is shown in Fig. 1. The main layer of the backbone consists of several components: the CBS module for convolution (Conv) and batch normalization (BN) with a sigmoid linear unit (SiLU) activation function, max pooling (MP), the E-C2G module, and the MP-PELAN module.

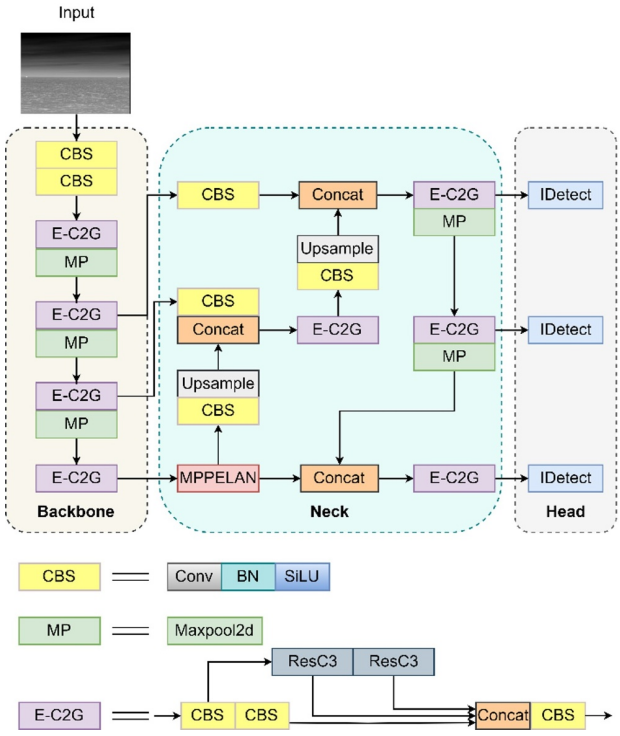

Fig. 1. The structure of IS-YOLO model architecture.

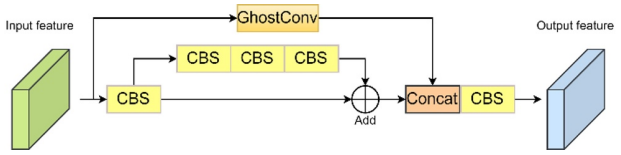

Fig. 2. Schematic diagram of ResC3 module.

Ships release considerable quantities of heat when sailing, which causes the targets in the infrared image to appear quite bright. As a result, the brightness of a ship's area is greater than the background. Nevertheless, great brightness from mountains and buildings in infrared photos has an adverse effect on the ability to detect ships. These factors affect the network's ability to extract features. The E-ELAN module in YOLOv7 [12] is inefficient in capturing the image's features. Furthermore, the ELAN structure incorporates a greater number of convolutional modules and residual connections, resulting in increased computational complexity and decreased inference performance. Therefore, we improved E-ELAN for use in the E-C2G module for the backbone and the MPPELAN module.

The E-C2G module is comprised of the CBS module and the ResC3 module. The input from the initial convolutional layer is divided into two paths, each of which is processed separately through a sequence of ResC3 and convolutional layers. Finally, the outputs from these paths are combined. This approach of using two paths simultaneously allows for effective propagation of gradients and reuse of features, resulting in a notable improvement in the model's ability to learn and in its inference speed. This is achieved by maintaining a deep architecture without incurring the usual computational cost associated with increasing complexity.

As a feature extraction module, ResC3 combines the image features that were extracted by the CBS convolutional layer in order to provide a more comprehensive feature representation. As shown in Fig. 2, when the feature map enters ResC3, it will be processed in two ways. The GhostConv [24] module in ResC3 decreases the dimensionality of the feature map to enhance the convolution kernel's comprehension of the feature information. It then increases the dimensionality to extract more comprehensive feature information. In the end, features are extracted using a residual structure that combines the input and output in order to eliminate redundant gradient information. Therefore, the ELAN in E-C2G module has the ability to splice and aggregate the intermediate information, while the ResC3 module can effectively decrease the parameter count.

The MPPELAN is an enhanced version of SPPELAN from [2] as shown in Fig. 3. The SPPELAN converts feature maps of varying scales to a uniform scale via Spatial

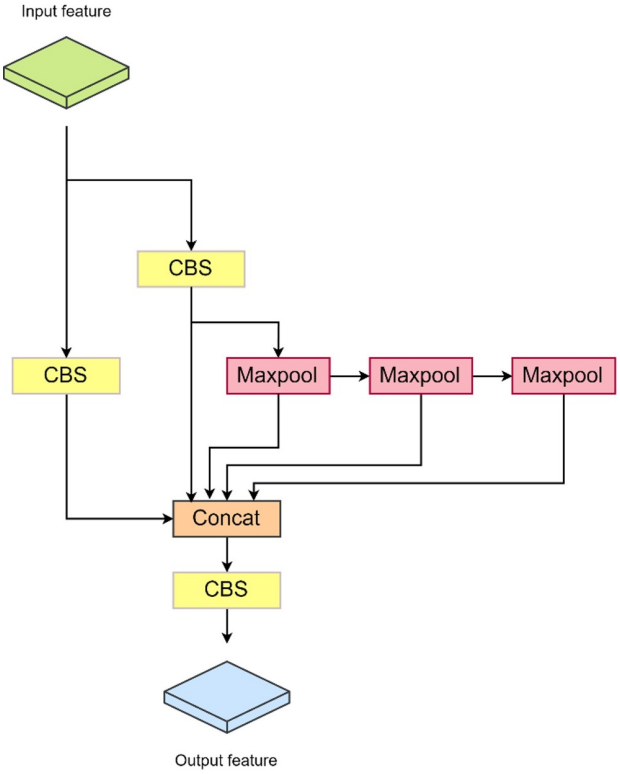

Fig. 3. Schematic diagram of MPPELAN module.

Pyramid Pooling. These converted feature maps are then combined to generate a feature representation that is well-suited to further processing. The MPPELAN module integrates features from different scales to enhance the network's ability to capture both local details and global context effectively. Comprising a sequence of convolutional layers followed by max pooling operations, the module first transforms the input feature map into a higher-dimensional space, enabling it to capture diverse aspects of the input information. Subsequently, max pooling operations with a customizable kernel size downsample the feature maps, capturing features at various scales. These scaled features are then concatenated along the channel dimension, and they pass through a final convolutional layer that further refines the representation while reducing the number of channels. Through this process, the MPPELAN module fosters robust feature fusion, crucial for tasks such as object detection and semantic segmentation, where capturing multi-scale information is pivotal for accurate and comprehensive analysis of the input data.

4. EXPERIMENT

In this section, we first introduce the experimental dataset, the experiment environment, related evaluation metrics, and the results.

4.1. Dataset

Used in this study was a dataset from Yeungnam University of small ships in infrared images. More images were captured and combined with a previous dataset [16] to increase the amount of data. The FLIR T620 camera was used to capture ship targets in several regions. This dataset contains 1370 training images and 120 val images in YOLO labeling format. Samples from the dataset are in Fig. 4.

The distribution of target numbers per image is shown in Fig. 5. There are from one to nine targets per image, and most of the images have two targets.

4.2. Experiment environment

We conducted the experiments on a personal computer with an AMD Ryzen 9 7950X 16-core processor, NVIDA's GeForce RTX 4090 GPU, and the Ubuntu 23.04 operating system running a compilation environment with Python 3.8.19, PyTorch 2.2.1, and CUDA 11.2. Parameters applied in the experimental training process are in Table 1.

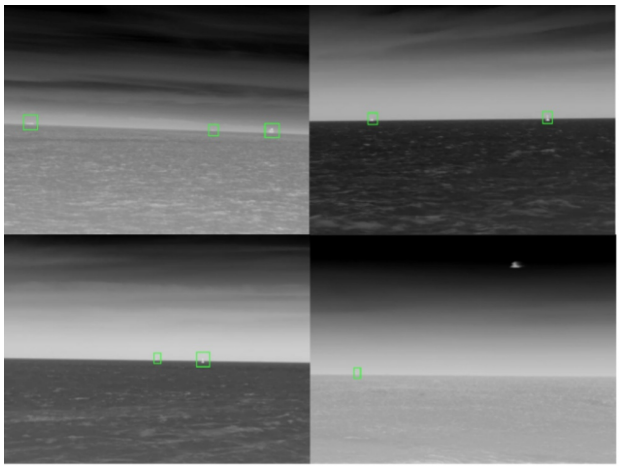

Fig. 4. Dataset sample.

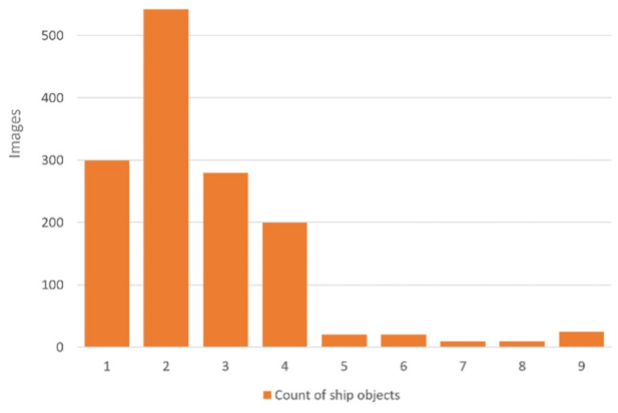

Fig. 5. The number of ship target in dataset.

Table 1. Training parameters.

Parameter	Value	Parameter	Value
Learning rate	0.01	Image size	640 × 640
Batch size	1	Optimizer	Adam
Worker	1	Epoch	50

4.3. Evaluation metrics

The algorithm's performance was evaluated using and average precision, the number of parameters, model size, GFLOPs, inference time. AP is the mean of the highest precision values across various recall conditions for one category of target. The PR Curve is summed up by converting AP into a single scalar number. Across a range of confidence threshold values, average precision is high when both recall and precision are high, and it is low when either of them is low. The calculation formula is stated as follows:

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

The precision-recall curve is expressed by $p(r)$, and dr shows a very small change in the recall value r as we integrate precision function $p(r)$ over a recall value ranging between 0 to 1. The term AP@.5 indicates the mean average precision is calculated with a threshold greater than 0.5, while AP@.5:.95 indicates the mean average precision computed at various intersection over union (IoU) thresholds ranging from 0.5 to 0.95 with increments of 0.05.

4.4. Experiment results

Five sets of ablation experiments using YOLOv7, YOLOv7 with MPPELAN, E-C2G with SPPCSPC, E-C2G with SPPELAN, and our method (IS-YOLO) were performed to verify the effectiveness of the improvements proposed in this paper. Table 2 illustrates the change in the performance of the object detection model during the construction process of IS-YOLO.

The first experiment uses the original YOLOv7 [12] as a benchmark. The MPPELAN module was introduced into YOLOv7 to replace the SPPCSPC module for the second experiment. Its AP@.5 was lower than with YOLOv7, but the AP@.5:.95 was higher than YOLOv7. Moreover, both the number of parameters and the model size were the lowest of the other experiments. In the third, fourth, and fifth experiments, the SPPCSPC module, the SPPELAN module, and the MPPELAN module were used with E-C2G to analyze the effectiveness of the MPPELAN. In Table 2, we can see that integrating SPPCSPC yielded higher parameters. Both SPPCSPC and SPPELAN experiments with the E-C2G module did not increase the AP score. However, the experiment using the MPPELAN with the E-C2G module (IS-YOLO) achieved the highest scores for AP@.5 and AP@.5:.95 (88.9% and 38.2%, respectively). Although YOLOv7 with the MPPELAN module achieved the lowest parameters and the smallest model size, IS-YOLO gained lower parameters and a smaller model size than the baseline YOLOv7.

Furthermore, to validate the effectiveness of the proposed algorithm for detecting small-ship targets in infrared images, it was compared with YOLOv7 [12], YOLOv8l [25], YOLOv9 [23], Faster R-CNN [10], FCOS [26], and RetinaNet [27] to verify its efficacy. We used the MM detection toolbox [28] to evaluate our model and compare with [10,26,27]. Table 3 presents the findings, highlighting the results representing the most favorable outcomes in general.

The aforementioned findings indicate that the enhanced IS-YOLO detection model exhibited a notably higher level of sophistication in identifying small-ship targets in infrared images. This is evident from the data presented in Table 3, where the average detection accuracy of the proposed algorithm in this study surpasses all other advanced algorithms listed.

To ensure the rigorousness of the results, we completed comparative experiments on the SIRST v2 dataset [29], which is the second version of SIRST v1 [30]. In contrast to SIRST v1, SIRST v2 encompasses a greater number of urban vistas cluttered with streetlights, cranes, and other non-target background interference that necessitates a sophisticated semantic understanding to differentiate them. There were 512 images in the dataset: 360 images for training, 103 for validation, and 51 for testing. The hardware and software environments for the experiments were

Table 2. Ablation experiment.

Model	AP@.5 (%)	AP@.5:.95 (%)	Params (M)	Model size (Mb)	GFLOPs	Inference time (ms)
YOLOv7 (ELAN + SPPCSPC)	86.5	34.9	36.4	71.3	103.2	3.6
ELAN + MPPPELAN	84.6	35.5	**28.8**	**59.6**	**97.1**	**3.4**
E-C2G + SPPCSPC	83.9	33.5	40.4	81.5	127.3	4.1
E-C2G + SPPELAN	84.1	35.6	32.8	63.1	121.2	3.9
IS-YOLO (E-C2G + MPPELAN)	**88.9**	**38.3**	32.8	63.1	121.2	3.9

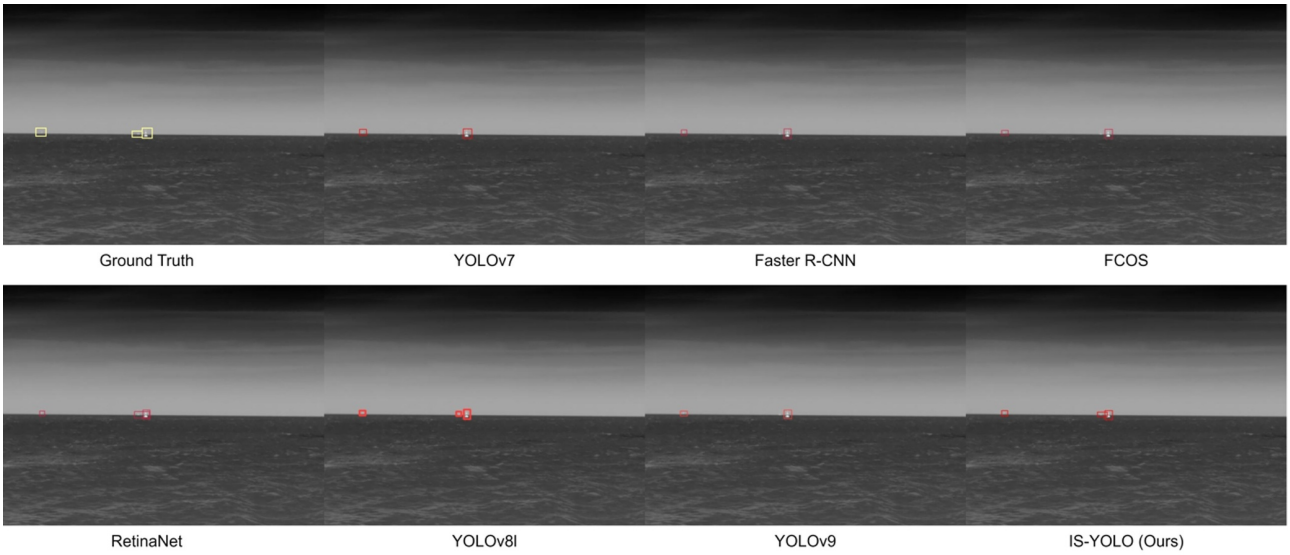

Fig. 6. Detection results of advanced detection models.

Table 3. Several outstanding algorithms are chosen to be compared with our model.

Method	AP@.5 (%)	AP@.5:.95 (%)	Params (M)	Model size (M)
YOLOv7	86.5	34.9	36.4	71.3
Faster R-CNN	84	30.9	-	533
FCOS	77.3	28.8	-	244
RetinaNet	79.60	29.10	-	278
YOLOv8l	84.9	35.2	43.6	83.7
YOLOv9	83.9	32	60.4	116
IS-YOLO	**88.9**	**38.3**	**32.8**	**63.1**

Table 4. Results of SIRST-V2 dataset.

Model	P (%)	R (%)	AP@.5 (%)	AP@.5:.95 (%)
YOLOv7	89.1	65.7	71.6	30.7
IS-YOLO	**88.3**	**77.4**	**80.3**	**32.6**

consistent with those in Subsection 4.2.

The second dataset was an IRSTD-1k dataset [31] consisting of 1,000 manually labeled, realistic infrared images with various target shapes, different target sizes, and rich background clutter in diverse scenes. The dataset covers different kinds of small targets, such as drones, creatures, vessels, and vehicles, captured at various positions from a long distance. The images are 512×512 with diverse backgrounds, such as the sea, a river, a field, a mountain area, a city, and cloud, with heavy clutter and noise. Experiment results on both datasets are presented in Tables 4 and 5.

We can see from the above tables that IS-YOLO provided good performance in small-ship detection from infrared images. The proposed model's precision, recall, and mAP reached higher scores than the YOLOv7 baseline. To verify the visual effects of the proposed algorithm in infrared small ship detection, several test images were selected to evaluate the effectiveness of IS-YOLO. The outcomes from using the YOLOv7 baseline, several state-of-the-art models, and IS-YOLO are showcased in Fig. 6. Furthermore, the small targets in our results are magnified to enhance their visibility and can be seen in Fig. 7.

Table 5. Results of IRSTD-k dataset.

Model	P (%)	R (%)	AP@.5 (%)	AP@.5:.95 (%)
YOLOv7	85.3	78.3	82.2	32.5
IS-YOLO	**84.6**	**79.7**	**82.8**	**33.6**

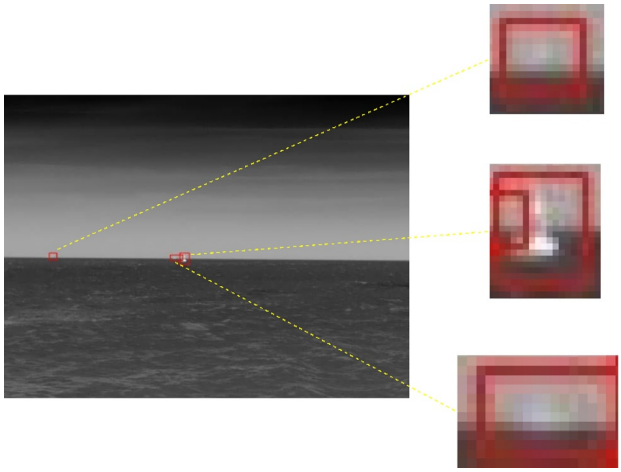

Fig. 7. The small target in the testing image.

5. CONCLUSION

To address the challenges of small object detection from infrared images, this paper proposed the IS-YOLO model based on YOLOv7, which can accurately detect small ships in a complex environment. First, we introduced the ELAN two convolutions and GhostConv network, which significantly improves the ability of the backbone to acquire more features. Second, by introducing the max pooling pyramid-ELAN, we not only improved feature extraction but helped the model to acquire features of different scales, thus enhancing accuracy. The IS-YOLO model outperformed the baseline model's AP@.5 and AP@.5:.95 by achieving improvements of 2.4%, and 3.3%, respectively.

CONFLICTS OF INTEREST

The authors declare they have no competing financial interests or personal relationships that could appear to influence the work in this paper.

REFERENCES

[1] X. Nie, M. Duan, H. Ding, B. Hu, and E. Wong, "Attention mask R-CNN for ship detection and segmentation from remote sensing images," *IEEE Access*, vol. 8, pp. 9325-9334, 2020.

[2] F. Yang, Q. Xi, B. Li, and Y. Ji, "Ship detection from thermal remote sensing imagery through region-based deep forest," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 3, pp. 449-453, 2018.

[3] X. Tong, B. Sun, J. Wei, Z. Zuo, and S. Su, "EAAU-Net: Enhanced asymmetric attention U-net for infrared small target detection," *Remote Sensing*, vol. 13, no. 16, 3200, 2021.

[4] X. Wun, D. Hong, Z. Huang, and J. Chanussot, "Infrared small object detection using deep interactive U-Net," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1-5, 2022.

[5] Y. Li, Z. Li, Y. Zhu, B. Li, W. Xiong, and Y. Huang, "Thermal infrared small ship detection in sea clutter based on morphological reconstruction and multi-feature analysis," *Applied Sciences*, vol. 9, no. 18, 3786, 2019.

[6] T.-Y. Lin, M. Maire, S. Belongie, H. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," *Proc. of 13th European Conference on Computer Vision–ECCV 2014*, Zurich, Switzerland, September 2014.

[7] C. Chen, M.-Y. Liu, O. Tuzel, and J. Xiao, "R-CNN for small object detection," *Proc. of 13th Asian Conference on Computer Vision–ACCV 2016*, Taipei, Taiwan, November 2016.

[8] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," *Proc. of the IEEE International Conference on Computer Vision*, pp. 2980-2988, 2017.

[9] R. Girshick, "Fast R-CNN," *Proc. of the IEEE International Conference on Computer Vision*, pp. 1440-1448, 2015.

[10] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137-1149, 2016.

[11] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[12] C.-Y. Wang, A. Bochkovsky, and H.-Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[13] J. Ye, Z. Yuan, C. Qian, and X. Li, "CAA-YOLO: Combined-attention-augmented YOLO for infrared ocean ships detection," *Sensors*, vol. 22, no. 10, 3782, 2022.

[14] Z. Gao, Y. Zhang, and S. Wang, "Lightweight small ship detection algorithm combined with infrared characteristic analysis for autonomous navigation," *Journal of Marine Science and Engineering*, vol. 11, no. 6, 1114, 2023.

[15] F. Guo, H. Ma, L. Li, M. Lv, and Z. Jia, "Multi-attention pyramid context network for infrared small ship detection," *Journal of Marine Science and Engineering*, vol. 12, no. 2, 345, 2024.

[16] I. M. Firdiantika and S. Kim. "One-stage infrared ships detection with attention mechanism," *Proc. of 23rd International Conference on Control, Automation and Systems (ICCAS)*, IEEE, 2023.

[17] G. Jocher, A. Stoken, J. Borovec, *et al*., "Ultralytics/YOLOv5: v3.1 - Bug fixes and performance improvements," *Zenodo*, 2020.

[18] C.-Y. Wang, H.-Y. M. Liao, and I.-H. Yeh, "Designing network design strategies through gradient path analysis," arXiv preprint arXiv:2211.04800, 2022.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1904-1916, 2015.

[20] C.-Y. Wang, H.-Y. M. Liao, I.-H. Yeh, Y.-H. Wu, P.-Y. Chen, and J.-W. Hsieh, "CSPNet: A new backbone that can enhance learning capability of CNN," *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020.

[21] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[22] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," arXiv preprint arXiv:2004.10934, 2020.

[23] C.-Y. Wang, I.-H. Yeh, and H.-Y. M. Liao, "YOLOv9: Learning what you want to learn using programmable gradient information," arXiv preprint arXiv:2402.13616, 2024.

[24] K. Han, Y. Wang, Q. Tian, J. Gio, C. Xu, and C. Xu, "GhostNet: More features from cheap operations," *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[25] G. Jocher, A. Chaurasia, and J. Qiu, YOLO by Ultralytics, https://github.com/ultralytics/ultralytics, Ultralytics. 2023.

[26] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: A simple and strong anchor-free object detector," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 4, pp. 1922-1933, 2022.

[27] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *Proc. of the IEEE International Conference on Computer Vision*, pp. 2999-3007, 2017.

[28] K. Chen, J. Wnag, J. Pang, *et al*., "MMDetection: Open MMLab detection toolbox and benchmark," arXiv preprint arXiv:1906.07155, 2019.

[29] Y. Dai, X. Li, F. Zhou, Y. Qian, Y. Chen, and J. Yang, "One-stage cascade refinement networks for infrared small target detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1-17, 2023.

[30] Y. Dai, Y. Wu, F. Zhou, and K. Barnard, "Asymmetric contextual modulation for infrared small target detection," *Proc. of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 949-958, 2021.

[31] M. Zhang, R. Zhang, Y. Yang, H. Bai, J. Zhang, and J. Guo, "ISNet: Shape matters for infrared small target detection," *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 867-876, 2022.

Indah Monisa Firdiantika was born in Bandar Lampung, Indonesia, in 1998. She received her B.S. degree in electrical engineering from Universitas Muhammadiyah Yogyakarta, in 2020. She is currently pursuing a master's degree with the Department of Electronic Engineering. Her research interests include object segmentation and infrared small target detection.

Sungho Kim graduated from the College of Engineering, Korea University, in February 2000. He received his B.S. and Ph.D. degrees from the School of Electrical and Electronic Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea, in 2002 and 2007, respectively. Since 2007, he has been with the Agency for Defense Development. From 2007 to 2010, he worked as a Senior Researcher with the Defense Science Research Institute (ADD). Since 2011, he has also been working as a Professor with the Department of Electronic Engineering, Yeungnam University, Korea. His research interests include hyperspectral, infrared, multi-sensor fusion, and deep learning.