A Cross-modal Fusion Method for Multispectral Small Ship Detection

Yang Liu
Tsinghua Shenzhen International Graduate School
Tsinghua University
Shenzhen, China
yang-liu22@mails.tsinghua.edu.cn

Yu Liu
Tsinghua Shenzhen International Graduate School
Tsinghua University
Shenzhen, China
liuyu77360132@126.com

Xueqian Wang
Department of Electronic Engineering
Tsinghua University
Beijing, China
wangxueqian@mail.tsinghua.edu.cn

Linping Zhang
Department of Electronic Engineering
Tsinghua University
Beijing, China
zlp22@mail.tsinghua.edu.cn

Zhizhuo Jiang
Tsinghua Shenzhen International Graduate School
Tsinghua University
and Research Institute of Tsinghua University in Shenzhen
Shenzhen, China
jiangzhizhuo@sz.tsinghua.edu.cn

Yaowen Li
Tsinghua Shenzhen International Graduate School
Tsinghua University
Shenzhen, China
liyw23@sz.tsinghua.edu.cn

Chenggang Yan
Institute of Information and Control
Hangzhou Dianzi University
Hangzhou, China
cgyan@hdu.edu.cn

Ying Fu
School of Computer Science and Technology
Beijing Institute of Technology
Beijing, China
fuying@bit.edu.cn

Tao Zhang
Lishui Institute
Hangzhou Dianzi University
Hangzhou, China
tzhang@hdu.edu.cn

***Abstract*—The fusion module of RGB and infrared (IR) remote sensing images is the key of multispectral ship detection. Existing works have shown that the cross-attention-based feature fusion can achieve good performance by extracting the complementary information of RGB and IR modalities. However, the existing commonly used cross-attention mechanisms introduce lots of redundancy parameters and mainly focus on global feature interaction of multispectral images, ignoring local detail information that is also important for small ship detection. In this paper, we propose a novel multispectral ship detection approach named LoGFusion. In LoGFusion, we design the cross stage partial module with partial convolution (CSPMPC) to reduce feature redundancy and utilize the local cross-modal fusion module (LoCFM) and global cross-modal fusion module (GCFM) to capture both local and global cross-modal features. Furthermore, we introduce a Multispectral Small Ship Dataset (MSSD) containing over 5k ship targets for small target detection. Experiments on MSSD validate the effectiveness of our method in terms of small ship detection in multispectral images.**

***Index Terms*—Small ship detection, multispectral, cross-attention, feature fusion**

This work was supported by National Key R&D Program of China under Grant 2021YFA0715202, National Natural Science Foundation of China under Grants 62101303 and 62022092, and 62341130, and Autonomous Research Project of Department of Electronic Engineering at Tsinghua University. Corresponding author: Yu Liu (liuyu77360132@126.com).

I. Introduction

Remote sensing ship detection plays an essential role in civil and military fields such as marine monitoring, territorial security, and environment protection. Unfortunately, the detection of small ships is still a challenging task because of the weak features of small ships and complex ocean environment in practice.

Driven by the rapid development of deep learning (DL), DL-based detectors have become the mainstream of ship detection. Chen et.al [1] established the LEVIR-ship dataset and introduced the degraded reconstruction enhancement network (DRENet) for ship detection, helping the backbone pay more attention to the small ships. Xu et.al [2] proposed a low-resolution marine object detection model (LMO-YOLO) for low-resolution ship detection, preventing the weakening of small ship information with dilated convolution. A dual-supervised network is presented in [3] to enhance the quality of input images for ship detection and reduce the effect of complex scene disturbances. A novel ship detection network based on multi-feature transformation and fusion (MFTF-Net) was proposed in [4] to enhance small target detection performance via multifeatured transformation and fusion. The feature-enhanced structure (FES) and saliency prediction branch (SPB) were proposed in [5] to provide new insights

into ship detection assisted by target saliency. The hybrid spatial pyramid pooling (HSPP) was designed in [6] to fuse the local and global features, guiding the detection performance improvement of small ship targets.

Note that the aforementioned ship detection methods are mostly based on RGB optical remote sensing images. In addition to optical images, infrared (IR) band images still provide contour features in poor weather conditions that distinguish the ship targets from the background [7]. It has been proven in [8]–[10] that using multispectral remote sensing images can improve the detection performance by capturing complementary information between RGB and IR images. The cross-modal fusion transformer (CFT) was presented in [11] to fuse the features of the intra-modality and inter-modality of multispectral images, improving the quality of feature fusion across modalities. Shen et.al [12] proposed a dual-modal feature fusion (DMFF) module to capture complementary information across modalities from global perspectives, which addressed the issue of limited receptive fields when using convolutional neural network (CNN) for feature fusion. However, these methods [11], [12] employ standard attention mechanism during feature fusion, only modeling global feature interaction, while overlooking local feature interaction which are also vital for small ship targets.

To address the aforementioned issue, we introduce the cross stage partial module with partial convolution (CSPMPC) in the feature extraction stage, meanwhile utilize the local cross-modal fusion module (LoCFM) and global cross-modal fusion module (GCFM) to fusion local and global cross-modal features. Compared with existing works [11], [12], our network aggregates both local and global cross-modal features, instead of only focusing on global feature interaction. In addition, we establish a Multispectral Small Ship Dataset (MSSD)[1] using Sentinel-2 data [13], which contains 2494 visible-near-infrared (NIR) image patch pairs and more than 5500 small ship instances. Experiments demonstrate that our proposed method achieves better detection performance on MSSD in comparison with existing methods.

II. The Proposed Method

In this section, we first introduce the overall framework of the LoGFusion in Section II-A. Subsequently, we present the details of the CSPMPC for feature redundancy reduction in Section II-B. Lastly, we describe the LoCFM (Section II-C) and GCFM (Section II-D) for local and global cross-modal feature interaction, respectively.

A. Overall architecture

As depicted in Fig. 1, we propose a two-stream network to extract the RGB and NIR features by backbone1 and backbone2, respectively, both of which are based on CSPNet [14]. CSPNet [14] consists of convolution-batch-normalization-silu (CBS) and cross stage partial (CSP) modules for feature extraction. The spatial pyramid pooling (SPP) module [15] is utilized for deep feature extraction, comprising parallel max pooling modules with different kernel sizes. To reduce feature redundancy among different channels, we introduce the CSPMPC in the deep layers of the original CSPNet [14], where the features have high similarities among different channels [16]–[18]. Since NIR images usually contain less visual information than the corresponding RGB images, we construct a more lightweight backbone with fewer convolution blocks for the NIR branch than backbone1, as illustrated in Fig. 1. Then, the LoCFM and GCFM are used for the fusion of local and global cross-modal features, which are both significant for small ship detection. Global features usually cover richer semantic information, while local features contain the fine-grained details such as color and texture. After that, the fused features are fed into the neck, consisting of the feature pyramid network (FPN) [19] and path aggregation network (PANet) [20], to integrate semantic features and local texture features, and are ultimately processed by the heads for classification and regression. The neck and heads are set to be the same as YOLOv5s [15].

[1]https://github.com/kkkkkkb/MSSD

B. CSPMPC

CSPMPC is helpful to reduce redundant features among different channels. As shown in Fig. 2, the input feature maps are first routed into two parallel branches and then processed by the two CBS modules for feature extraction, respectively, where both the number of channels are halved. One of the two branches is connected to the end of the stage, while the other is fed into partial convolution [18] and the CBS module for further feature extraction. Partial convolution [18] only applies convolution operation on the first quarter of the input channels, keeping other channels untouched, which fully utilizes the redundancy of feature maps. Compared to regular convolution, partial convolution [18] offers a more focused and efficient approach to feature extraction by avoiding the redundancy of learning repeated features in most convolution kernels. Finally, the two branches are concatenated which is followed by a CBS module.

C. LoCFM

We design LoCFM to capture fine-grained cross-modal features from RGB and NIR bands with local window cross-attention, as shown in Fig. 3. Given input feature maps $\mathbf{F}_{RGB} \in \mathbb{R}^{H\times W\times C}$ and $\mathbf{F}_{NIR} \in \mathbb{R}^{H\times W\times C}$ and window size s, we first divide each feature map into local windows, which means the attention only focus on limited window regions. Then, we flatten each window into a set of tokens, $\mathbf{T}_{RGB\_s} \in \mathbb{R}^{s^2\times C}$ and $\mathbf{T}_{NIR\_s} \in \mathbb{R}^{s^2\times C}$. After that, we also employ multi-head mechanism [21] to help the model jointly learn the relationships between two modalities from different perspectives. The vectors $\{\mathbf{Q}_{RGB\_s}, \mathbf{K}_{RGB\_s}, \mathbf{V}_{RGB\_s}$ and $\mathbf{Q}_{NIR\_s}, \mathbf{K}_{NIR\_s}, \mathbf{V}_{NIR\_s}\} \in \mathbb{R}^{s^2\times D_h}$, where D_h is the number of hidden dimensions for one head, are produced from $\mathbf{T}_{RGB\_s}$ and $\mathbf{T}_{NIR\_s}$ separately as follows:

$$\begin{aligned}\mathbf{Q}_{RGB\_s} = \mathbf{T}_{RGB\_s}\mathbf{W}_{q1}, \mathbf{K}_{RGB\_s} = \mathbf{T}_{RGB\_s}\mathbf{W}_{k1},\\ \mathbf{V}_{RGB\_s} = \mathbf{T}_{RGB\_s}\mathbf{W}_{v1}\end{aligned} \tag{1}$$

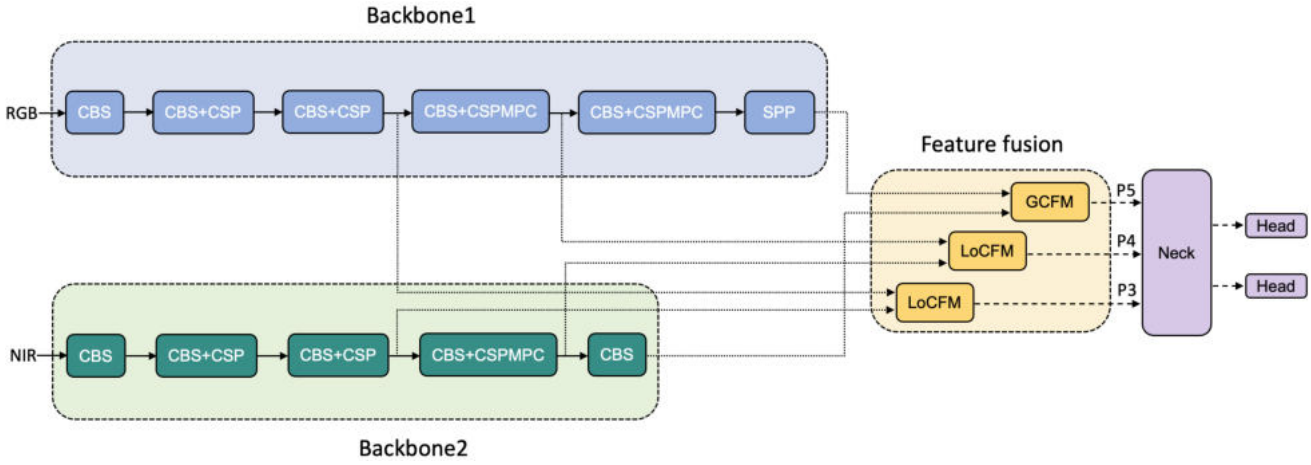

Fig. 1. Overall architecture of our proposed LoGFusion network.

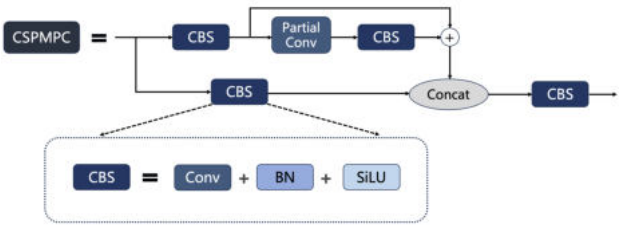

Fig. 2. Structure of the CSPMPC.

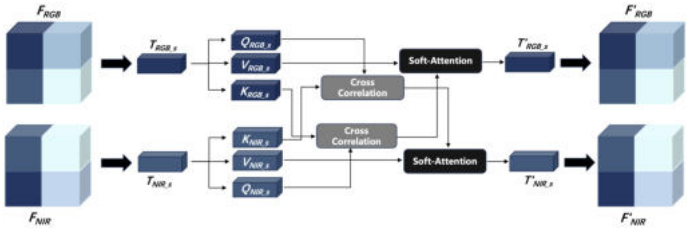

Fig. 3. Structure of the LoCFM.

$$\mathbf{Q}_{NIR\_s} = \mathbf{T}_{NIR\_s}\mathbf{W}_{q2}, \mathbf{K}_{NIR\_s} = \mathbf{T}_{NIR\_s}\mathbf{W}_{k2}, \\ \mathbf{V}_{NIR\_s} = \mathbf{T}_{NIR\_s}\mathbf{W}_{v2} \tag{2}$$

where $\{\mathbf{W}_{qi}, \mathbf{W}_{ki}$ and $\mathbf{W}_{vi}\} \in \mathbb{R}^{C\times D_h}(i = 1, 2)$ are learnable parameters.

The correlation matrix of RGB and NIR features is constructed using a dot-product operation and a softmax function, each element of which represents the correlation degree of two modalities. Next, the output of a cross-attention [21] head is a weighted sum over the value vectors.

$$\mathbf{head}_{RGB\_s} = \text{softmax}\left(\frac{\mathbf{Q}_{NIR\_s}\mathbf{K}^T_{RGB\_s}}{\sqrt{D_h}}\right)\mathbf{V}_{RGB\_s} \tag{3}$$

$$\mathbf{head}_{NIR\_s} = \text{softmax}\left(\frac{\mathbf{Q}_{RGB\_s}\mathbf{K}^T_{NIR\_s}}{\sqrt{D_h}}\right)\mathbf{V}_{NIR\_s} \tag{4}$$

Then, the outputs of all heads are concatenated, resulting in the final values $\mathbf{T}'_{RGB\_s} \in \mathbb{R}^{s^2\times C}$ and $\mathbf{T}'_{NIR\_s} \in \mathbb{R}^{s^2\times C}$, respectively. After that, $\mathbf{T}'_{RGB\_s}$ and $\mathbf{T}'_{NIR\_s}$ are combined and reshaped into the features $\mathbf{F}'_{RGB} \in \mathbb{R}^{H\times W\times C}$ and $\mathbf{F}'_{NIR} \in \mathbb{R}^{H\times W\times C}$, respectively.

The features $\mathbf{F}'_{RGB}$ and $\mathbf{F}'_{NIR}$ are added into the input features $\mathbf{F}_{RGB}$ and $\mathbf{F}_{NIR}$ through a residual connection, respectively, which can be expressed as:

$$\mathbf{F}^o_{RGB} = \mathbf{F}_{RGB} + \mathbf{F}'_{RGB} \tag{5}$$

$$\mathbf{F}^o_{NIR} = \mathbf{F}_{NIR} + \mathbf{F}'_{NIR} \tag{6}$$

The final output of LoCFM is the concatenation of $\mathbf{F}^o_{RGB}$ and $\mathbf{F}^o_{NIR}$ along the channel axis.

D. GCFM

To extract global cross-modal features, the GCFM computes attention scores in the whole feature maps. Given input feature maps $\mathbf{F}_{RGB} \in \mathbb{R}^{H\times W\times C}$ and $\mathbf{F}_{NIR} \in \mathbb{R}^{H\times W\times C}$, we directly flatten them into a set of tokens $\mathbf{T}_{RGB} \in \mathbb{R}^{HW\times C}$ and $\mathbf{T}_{NIR} \in \mathbb{R}^{HW\times C}$. Then, we also employ the multi-head mechanism [21] like in LoCFM. The subsequent operations are the same as LoCFM and are not be elaborated.

III. Results and Discussion

A. Multispectral ship detection dataset

We develop a multispectral small ship detection dataset from the Sentinel-2 satellite [13]. The dataset comprises 8 original satellite images with blue, green, red and NIR channels under different conditions, such as inshore, cloud, and muddy background cases. We crop the original images to obtain 2494 image patches with 512 × 512 pixels. This dataset contains 5584 annotated ships. The spatial resolution of images in our dataset is 10m × 10m. As shown in Fig. 5, most of the ships are smaller than 20 × 20 pixels in the MSSD.

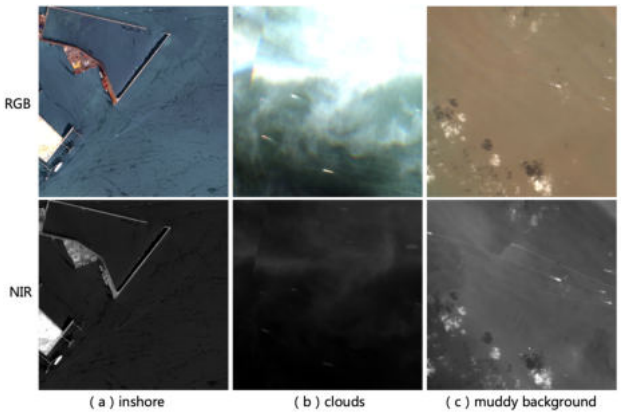

Fig. 4. Some multispectral image examples with different situations.

B. Results analysis

In this paper, F1-score (F1), AP_{50} and $\mathrm{AP}_{50:95}$ are used as metrics to evaluate the performance of our method, which are defined as follows:

$$F1 = 2 \times \frac{P \times R}{P + R} \tag{7}$$

$$AP = \int_0^1 P(R) dR \tag{8}$$

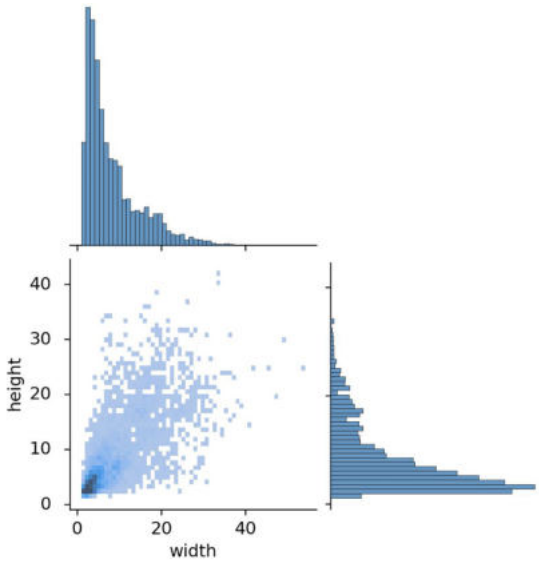

Fig. 5. The statistics of width/height (in terms of number of pixels) of ships in MSSD.

where P denotes precision and R denotes recall. AP_{50} is the average precision (AP) at intersection over union (IoU) = 0.50 in (8). Similarly, $\mathrm{AP}_{50:95}$ can be formulated as the mean of AP at IoU thresholds ranging from 0.50 to 0.95 with an interval of 0.05.

The baseline of our method is the YOLOv5s [15] network without the head for detecting large objects. Considering few ships are larger than 320m × 320m areas, the detection head with the feature maps whose size is 1/32 of the input image size is abandoned in the YOLOv5s [15] model. In the CSPMPC, we apply convolution operation on first half, first quarter, randomly selected half and randomly selected quarter of the input channels, respectively. And these four kinds of CSPMPC are denoted by CSPMPC(1/2), CSPMPC(1/4), random(1/2) and random(1/4). As shown in Table I, the proposed CSPMPC(1/4) achieves the $\mathrm{AP}_{50:95}$ with 33.4% and AP_{50} with 69.9%, respectively, and its parameter size is reduced by over 20% compared with the baseline. Table I implies that redundancy among different channels in the deep layers of the network may degrade the small ship detection performance.

Some ablation experiments about the proposed LoCFM and GCFM are completed in Table II. Compared with GCFM-only and LoCFM-only ways, "LoCFM+GCFM" produces better results, with $\mathrm{AP}_{50:95}$ of 34.4% and AP_{50} of 71.7% when fusion occurs in P3 and P4 stage (rows 1-3 in Table II) as shown in Fig 1. Furthermore, when we add the fusion of P5 stage (rows 4-7 in Table II), the ways of combining LoCFM and GCFM also achieve better performance than the GCFM-only and LoCFM-only ways. In comparison with the other ways, "2 LoCFM+1 GCFM" achieves better results, with $\mathrm{AP}_{50:95}$ of 35.1% and AP_{50} of 73.0%. The discussions above indicate both local details and global features are important for small ship detection.

Table III shows the experimental results of different methods on MSSD. Our method performs better than DRENet [4] in F1 (4.6%↑), $\mathrm{AP}_{50:95}$ (2.2%↑) and AP_{50} (2.8%↑), which is also designed for small ship detection. Note that the performance of pixel-level fusion (RGB+NIR) mode is worse than RGB mode for YOLOv6s [20]. It illustrates that pixel-level fusion may not always exploit the complementary information between modalities. Our method also surpasses previous multi-spectral networks (YOLOrs [8], CFT [11] and ICAFusion [12]) with fewer parameters.

Fig. 6 presents the examples of detection results to prove the superiority of our proposed model. The ship targets are tiny in these images, and some are surrounded by clouds. It can be observed that our method achieves improved results than other methods with fewer missed targets and false alarms.

IV. Conclusion

In this work, a novel cross-modal fusion model LoGFusion has been proposed for small ship detection of multispectral remote sensing images. First, we modify the baseline by introducing the CSPMPC to reduce redundant features among different channels. Second, we introduce the LoCFM and GCFM to extract local details and global features between RGB and IR modalities simultaneously, which are both vital for the detection of small ship targets. Experiments on MSSD have confirmed the superiority of our proposed method.

References

[1] J. Chen, K. Chen, H. Chen, Z. Zou and Z. Shi, "A Degraded Reconstruction Enhancement-Based Method for Tiny Ship Detection in Remote Sensing Images With a New Large-Scale Dataset," in *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, no. 5625014, pp. 1-14, 2022.

TABLE I
COMPARISION OF BASELINE AND CSPMPC

Methods	F1(%)	$AP_{50:95}$(%)	AP_{50}(%)	Params(M)
YOLOv5s	67.6	32.6	67.8	7.02
baseline	69.2	32.4	68.4	5.24
CSPMPC (1/2)	68.7	33.0	69.4	4.46
CSPMPC (1/4)	69.5	33.4	69.9	4.13
random (1/2)	69.1	32.8	69.5	4.46
random (1/4)	69.4	33.2	69.7	4.13

TABLE II
COMPARISION OF DIFFERENT FUSION WAYS

P3	P4	P5	F1(%)	$AP_{50:95}$(%)	AP_{50}(%)	Params(M)
G	G	–	71.0	33.9	70.8	5.45
L	L	–	71.5	34.3	71.5	5.45
L	G	–	71.5	34.4	71.7	5.45
G	G	G	71.1	34.5	71.8	7.21
L	G	G	72.1	35.0	72.7	7.21
L	L	G	**72.4**	**35.1**	**73.0**	7.21
L	L	L	71.6	34.5	72.2	7.21

[a]The letter G and L denote GCFM and LoCFM, respectively.

TABLE III
COMPARISION OF DIFFERENT METHODS

Methods	Modality	F1(%)	$AP_{50:95}$(%)	AP_{50}(%)	Params(M)
DRENet [4]	RGB	67.8	32.9	70.2	5.98
YOLOv6s [22]	RGB	65.5	32.1	62.4	18.50
YOLOv6s [22]	RGB+NIR	63.6	30.5	60.7	18.50
YOLOrs [8]	RGB+NIR	68.8	33.8	69.7	20.10
CFT [11]	RGB+NIR	71.9	34.7	71.7	44.54
ICAFusion [12]	RGB+NIR	71.8	33.4	71.2	23.24
Our method	RGB+NIR	**72.4**	**35.1**	**73.0**	7.21

[2] Q. Xu, Y. Li and Z. Shi, "LMO-YOLO: A Ship Detection Model for Low-Resolution Optical Satellite Imagery," in *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 4117-4131, 2022.

[3] Q. Xu, Y. Li, M. Zhang and W. Li, "COCO-Net: A Dual-Supervised Network With Unified ROI-Loss for Low-Resolution Ship Detection From Optical Satellite Image Sequences," in *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, no. 5629115, pp. 1-15, 2022.

[4] M. Zha, W. Qian, W. Yang and Y. Xu, "Multifeature Transformation and Fusion-Based Ship Detection With Small Targets and Complex Backgrounds," in *IEEE Geoscience and Remote Sensing Letters*, vol. 19, no. 4511405, pp. 1-5, 2022.

[5] Z. Ren, Y. Tang, Z. He, L. Tian, Y. Yang and W. Zhang, "Ship Detection in High-Resolution Optical Remote Sensing Images Aided by Saliency Information," in *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, no. 5623616, pp. 1-16, 2022.

[6] Z. Liu et al., "Improved YOLOv5s for Small Ship Detection With Optical Remote Sensing Images," in *IEEE Geoscience and Remote Sensing Letters*, vol. 20, no. 8002205, pp. 1-5, 2023.

[7] Y. Han, J. Liao, T. Lu, T. Pu and Z. Peng, "KCPNet: Knowledge-Driven Context Perception Networks for Ship Detection in Infrared Imagery," in *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, no. 5000219, pp. 1-19, 2023.

[8] M. Sharma et al., "YOLOrs: Object Detection in Multimodal Remote Sensing Imagery," in *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 1497-1508, 2021.

[9] J. Zhang, J. Lei, W. Xie, Z. Fang, Y. Li and Q. Du, "SuperYOLO: Super Resolution Assisted Object Detection in Multimodal Remote Sensing Imagery," in *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, no. 5605415, pp. 1-15, 2023.

[10] J. Zhu, X. Chen, H. Zhang, Z. Tan, S. Wang and H. Ma, "Transformer Based Remote Sensing Object Detection With Enhanced Multispectral Feature Extraction," in *IEEE Geoscience and Remote Sensing Letters*, vol. 20, no. 5001405, pp. 1-5, 2023.

[11] Q. Fang, D. Han, Z. Wang, "Cross-modality fusion transformer for multispectral object detection," *arXiv preprint arXiv:2111.00273*, 2021.

[12] J. Shen, Y. Chen, Y. Liu, X. Zuo, H. Fan and W. Yang, "ICAFusion: Iterative cross-attention guided feature fusion for multispectral object detection," *Pattern Recognition*, vol. 145, 2024.

[13] M. Drusch et al., "Sentinel-2: ESA's optical high-resolution mission for GMES operational services," *Remote Sensing of Environment*, vol. 120, pp. 25–36, 2012.

[14] C. -Y. Wang, H. -Y. Mark Liao, Y. -H. Wu, P. -Y. Chen, J. -W. Hsieh and I. -H. Yeh, "CSPNet: A New Backbone that can Enhance Learning Capability of CNN," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1571-1580, 2020.

[15] Glenn Jocher and Alex Stoken, "Yolov5," *https://github.com/ultralytics/yolov5*, 2021.

[16] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu and C. Xu, "GhostNet: More Features From Cheap Operations," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1577-1586, 2020.

[17] Qiulin Zhang, Zhuqing Jiang, Qishuo Lu, Jia'nan Han, Zhengxin Zeng, Shang-Hua Gao, and Aidong Men, "Split to be slim: An overlooked redundancy in vanilla convolution," *29th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 3195–3201, 2020.

[18] J. Chen et al., "Run, Don't Walk: Chasing Higher FLOPS for Faster Neural Networks," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12021-12031, 2023.

[19] T. -Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan and S. Belongie, "Feature Pyramid Networks for Object Detection," *2017 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 936-944, 2017.

[20] S. Liu, L. Qi, H. Qin, J. Shi and J. Jia, "Path Aggregation Network

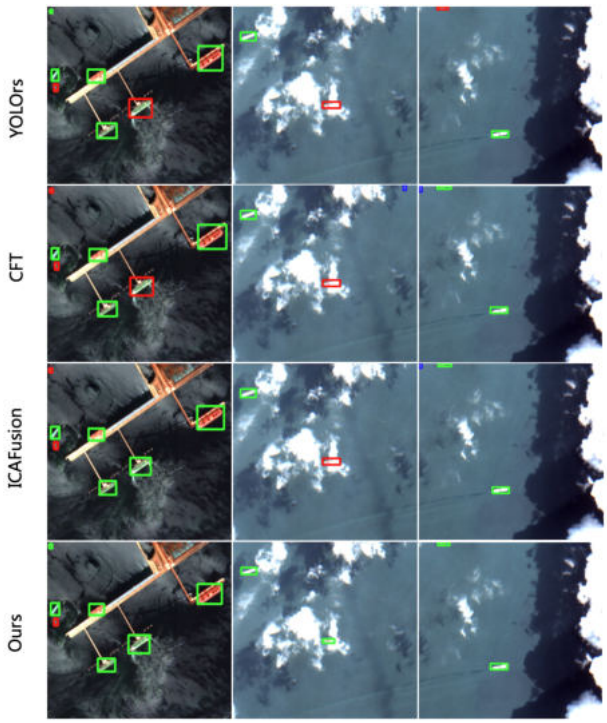

Fig. 6. The visual detection results of different methods. Green: correctly detected ships. Red: missed ships. Blue: false alarms.

for Instance Segmentation," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8759-8768, 2018.

[21] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, pp. 5998–6008, 2017.

[22] C. Li, L. Li, Y. Geng, H. Jiang, M. Cheng, B. Zhang, Z. Ke, X. Xu, and X. Chu, "Yolov6 v3. 0: A full-scale reloading," *arXiv preprint arXiv:2301.05586*, 2023.