MTU-Net: Multilevel TransUNet for Space-Based Infrared Tiny Ship Detection

Tianhao Wu, Boyang Li, Yihang Luo, Yingqian Wang, Chao Xiao, Ting Liu, Jungang Yang, Wei An, and Yulan Guo, *Senior Member, IEEE*

***Abstract*— Space-based infrared tiny ship detection aims at separating tiny ships from the images captured by Earth-orbiting satellites. Due to the extremely large image coverage area (e.g., thousands of square kilometers), candidate targets in these images are much smaller, dimer, and more changeable than those targets observed by aerial- and land-based imaging devices. Existing short imaging distance-based infrared datasets and target detection methods cannot be well adopted to the space-based surveillance task. To address these problems, we develop a space-based infrared tiny ship detection dataset (namely, NUDT-SIRST-Sea) with 48 space-based infrared images and 17 598 pixel-level tiny ship annotations. Each image covers about 10 000 km^2 of area with 10 000 × 10 000 pixels. Considering the extreme characteristics (e.g., small, dim, and changeable) of those tiny ships in such challenging scenes, we propose a multilevel TransUNet (MTU-Net) in this article. Specifically, we design a vision Transformer (ViT) convolutional neural network (CNN) hybrid encoder to extract multilevel features. Local feature maps are first extracted by several convolution layers and then fed into the multilevel feature extraction module [multilevel ViT module (MVTM)] to capture long-distance dependency. We further propose a copy–rotate–resize–paste (CRRP) data augmentation approach to accelerate the training phase, which effectively alleviates the issue of sample imbalance between targets and background. Besides, we design a FocalIoU loss to achieve both target localization and shape description. Experimental results on the NUDT-SIRST-Sea dataset show that our MTU-Net outperforms traditional and existing deep learning-based single-frame infrared small target (SIRST) methods in terms of probability of detection, false alarm rate, and intersection over union. Our code is available at https://github.com/TianhaoWu16/Multi-level-TransUNet-for-Space-based-Infrared-Tiny-ship-Detection**

***Index Terms*— FocalIoU loss, infrared ship detection, space-based detection, tiny ship, vision Transformer (ViT).**

I. Introduction

SPACE-BASED infrared tiny ship detection aims at separating tiny ships from the images captured by various (e.g., low, middle, and geostationary) Earth-orbiting satellites [1], [2]. Due to the much longer imaging distance, the targets of space-based infrared images of ocean scenes exhibit several different characteristics [1], [2] (e.g., larger image size, more complex background, more suspicious targets, extremely small targets, and multiscale targets) from that of previous land- and aerial-based single-frame infrared small target (SIRST) detection [3], [4], [5].

To detect tiny targets under complex scenes by an infrared band, numerous traditional methods have been proposed, including filtering-based methods [6], [7], local contrast-based methods [8], [9], [10], [11], [12], [13], and low-rank-based methods [5], [14], [15], [16], [17], [18], [19]. Although promising progress has been achieved, these methods essentially rely on handcrafted features and fixed hyperparameters. When the scenes (e.g., land, ocean, ports, and clouds background) change dramatically, these methods suffer from a significant decrease in probability of detection (P_d) and an increase in false alarm rate (F_a).

With the advances of deep learning, numerous convolutional neural network (CNN)-based methods [20], [21], [22], [23], [24], [25], [26], [27], [28] have been proposed recently, introducing significant performance improvement in SIRST. Dai et al. [23] proposed the first segmentation-based SIRST detection network (ACM). Then, Dai et al. [24] improved ACM by introducing a dilated local contrast measure and developed an ALC-Net. Moreover, Wang et al. [25] used a conditional generative adversarial network (MFvsFA-cGAN) to achieve a tradeoff between miss detection and false alarm for infrared small target detection. Li et al. [27] proposed a dense nested attention network (DNA-Net) to extract high-level information of small targets. However, the above CNN-based methods are designed for the short-distance imaging SIRST detection task (e.g., land- and aerial-based SIRST). The candidate targets in the space-based tiny ship detection task are much smaller,

Manuscript received 18 August 2022; revised 27 September 2022 and 23 December 2022; accepted 27 December 2022. Date of publication 9 January 2023; date of current version 18 January 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 61972435, Grant 61401474, Grant 61921001, and Grant 62001478; and in part by the National Key Research and Development Program of China under Grant 2021YFB3100800. *(Corresponding author: Jungang Yang.)*

Tianhao Wu, Boyang Li, Yihang Luo, Yingqian Wang, Chao Xiao, Ting Liu, Jungang Yang, and Wei An are with the College of Electronic Science and Technology, National University of Defense Technology (NUDT), Changsha 410073, China (e-mail: wutianhao16@nudt.edu.cn; liboyang20@nudt.edu.cn; luoyihang@nudt.edu.cn; wangyingqian16@nudt.edu.cn; xiaochao12@nudt.edu.cn; liuting@nudt.edu.cn; yangjungang@nudt.edu.cn; anwei@nudt.edu.cn).

Yulan Guo is with the College of Electronic Science and Technology, National University of Defense Technology (NUDT), Changsha 410073, China, and also with the School of Electronics and Communication Engineering, Sun Yat-sen University, Guangzhou 510275, China (e-mail: yulan.guo@nudt.edu.cn).

Digital Object Identifier 10.1109/TGRS.2023.3235002

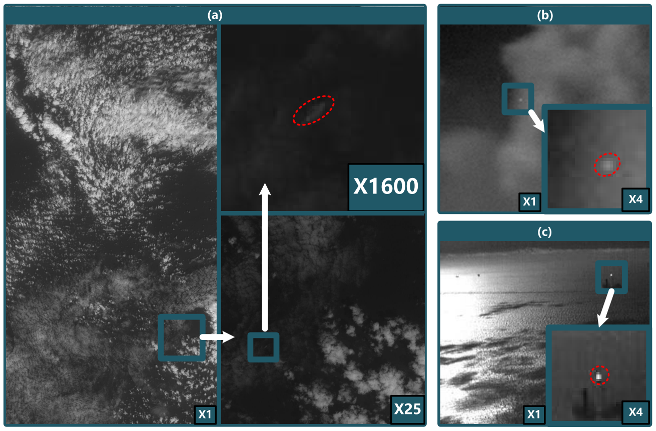


Fig. 1. Comparison of space-, air-, and land-based infrared images. (a) Typical image zoomed in 1, 25, and 1600 times in our space-based SIRST dataset. (b) Typical image zoomed in one and four times in the aerial-based SIRST dataset. (c) Typical image zoomed in one and four times in land-based SIRSTs. The targets are highlighted by red dotted circles.

dimer, and more changeable than those targets observed by aerial- and land-based imaging devices. These methods cannot be well adopted to handle such challenges in space-based SIRST tiny ship detection.

To address the above problems, we first develop a space-based infrared tiny ship dataset, NUDT-SIRST-Sea. It contains 17 598 tiny annotated ships and 48 images with 10 000 × 10 000 pixels captured by cameras mounted on the low Earth-orbiting satellite. As shown in Fig. 1, the tiny ships in space-based ocean scenes are visually nonsalient in local regions compared to those targets in the land- and aerial-based SIRST datasets. General CNN-based detection methods are not good at capturing long-distance dependency between targets and background. Therefore, it is necessary to further exploit the contextual relationship to achieve the improved detection performance.

Inspired by the success of the vision Transformer (ViT) [29] structure in generic object detection, we first design a multilevel ViT CNN hybrid encoder. Specifically, we design a multilevel ViT module (MVTM) to achieve coarse-to-fine feature extraction. In our multilevel ViT CNN hybrid encoder, multilevel features are first extracted by CNN. Then, these features are refined by MVTM to capture long-distance dependency. Due to the sparsity of tiny ships in space-based infrared images, the foreground targets and background are extremely imbalanced. To address this issue, we propose a novel copy–rotate–resize–paste (CRRP) data augmentation approach to increase the ration of candidate targets in the training phase and ultimately accelerate the convergence of the network. Moreover, we find that existing intersection over union (IoU)-like loss overly focuses on producing complete shape of the target but lacks the ability to locate small-scale targets. The focal loss focuses on hard samples but lacks the ability to produce the complete shape of the target. Therefore, we design a novel FocalIoU loss to accurately localize tiny targets and completely produce the shape of the target.

To the best of our knowledge, this is the first deep learning-based work to achieve space-based infrared tiny ship detection. The contributions of our work can be summarized as follows.

1) To our knowledge, NUDT-SIRST-Sea is the largest manually annotated dataset with a wide variety of categories in space-based infrared observation. The 17 598 high-precision bounding boxes and pixel-level annotations are introduced to support the development and evaluation of various target detectors in space-based infrared images.
2) We propose a novel Transformer CNN hybrid architecture [i.e., multilevel TransUNet (MTU-Net)] for space-based infrared tiny ship detection. With the help of multilevel ViT CNN hybrid encoder, the long-distance dependency of tiny ships can be well incorporated and fully exploited by coarse-to-fine feature extraction and multilevel feature fusion.
3) A CRRP data augmentation method and a FocalIoU loss are proposed to alleviate the foreground–background imbalance problem and achieve "double-win" of target localization and shape description.
4) Experimental results show that space-based infrared tiny ship detection is a challenging task, and previous land- and aerial-based SIRST methods cannot well handle those challenges (e.g., extremely small and dim targets) introduced by this task. Our method can achieve state-of-the-art (SOTA) results in three metrics: probability of detection (P_d), false alarm rate (F_a), and IoU.

This article is organized as follows. In Section II, we briefly describe the importance of our task and present the statistical characteristics and challenges of our NUDT-SIRST-Sea dataset in detail. In Section III, we briefly review the related work. In Section IV, we introduce the architecture of our MTU-Net, CRRP data augmentation approach, and our FocalIoU loss in detail. The experimental results are represented in Section V. Section VI gives the conclusion.

II. Analysis of the NUDT-SIRST-Sea Dataset

A. Importance

Space-based infrared tiny ship detection is one of the most important tasks in the SIRST detection family, which generally includes land-based [25], aerial-based [23], [28], and space-based SIRST detection tasks. As shown in Table I, a space-based infrared image usually covers about 10 000 km^2 of area with 10 000 × 10 000 pixels, thousands of times larger than the size of other land- and aerial-based images. Due to the much longer imaging distance, the targets of space-based infrared images exhibit extreme characteristics (e.g., small, dim, and changeable) than other SIRST images. Due to the lack of sufficient high-quality annotated datasets, existing deep learning-based methods cannot work well on the space-based long-distance detection tasks. To alleviate this problem, we propose a space-based SIRST dataset (namely, NUDT-SIRST-Sea). Specifically, we collect 48 real images captured by sensors mounted on the low Earth-orbiting satellite from both near-infrared (i.e., 845–885-nm wavelength) and short-infrared waveband (i.e., 1650–1660-nm wavelength). The field of view for both wave lengths images is completely overlapped; 41 images are used for training and the rest seven

TABLE I

MAIN CHARACTERISTICS OF SEVERAL POPULAR SIRST DATASETS. NOTE THAT OUR NUDT-SIRST-SEA DATASET HAS IMAGES OF THE HIGHEST RESOLUTION, GROUND-TRUTH LABEL TYPE, THE LOWEST TARGET-TO-BACKGROUND RATIO, THE LARGEST TARGET NUMBER, THE LOWEST AVERAGE SNR, THE SMALLEST AVERAGE TARGET SIZE, AND THE LARGEST NUMBER OF GROUND-TRUTH ANNOTATIONS COMPARED WITH THE MAINSTREAM SIRST DATASETS

Datasets	Image Type	Resolution	Label Type	Scene Type	Target/Background ratio	Average SNR	Average Target size	Target number
NUAA-SIRST [23]	real	256 × 256	Coarse Label	Aerial-based	0.06097%	0.91	40	533
NUST-SIRST [25]	synthetic	320 × 240	Coarse Label	Land-based	0.94856%	0.93	151	10337
IRSTD-1k [28]	real	256 × 256	Ground Truth	Land-based	0.02594%	0.76	85	1001
NUDT-SIRST [27]	synthetic	256 × 256	Ground Truth	Aerial/Land-based	0.06778%	0.88	44	1901
NUDT-SIRST-Sea	**real**	**10000 × 10000**	**Ground Truth**	**Space-based**	**0.000029%**	**0.28**	**29**	**17598**

images are used for test. Each image covers about 10 000 km^2 of area with 10 000 × 10 000 pixels. Moreover, we take more than 500 h to manually generate 17 598 pixel-level tiny ship annotations. Both bounding boxes and pixel-level annotations are provided for the development and evaluation of various target detectors. The average target size of our dataset is 29 pixels, which is much smaller than the size of those images in other datasets. The target-to-background ratio of our NUDT-SIRST-Sea dataset is 0.000029%, which is hundreds of times smaller than the target-to-background ratios of other SIRST datasets.

B. Statistical Properties of NUDT-SIRST-Sea

1) Much Larger Image Size: Compared with existing SIRST datasets in Table II, each image of NUDT-SIRST-Sea covers about 10 000 km^2 of area with 10 000 × 10 000 pixels, thousands of times larger than the image sizes of NUDT-SIRST [27], NUST-SIRST [25], and NUAA-SIRST [23]. As shown in Fig. 2(a), a much larger image contains more different scenes (e.g., port, land, clouds, and sea). Besides, a much larger image size results in higher computational difficulties.

2) Much More Complex Background: As shown in Fig. 1, aerial- and land-based infrared images are much simpler than space-based infrared images due to the limited coverage area. As shown in Fig. 2(b), different scenes (e.g., clouds, tiny ships, port, land, and sea face) can form more types of complex scenes. Several scenes are considered as the difficult targets in NUDT-SIRST-Sea: urban inland river, cloud blocks, dense cluster targets, and targets in port. These complex scenes challenge the method's ability to capture long-distance context information.

3) Multitype Suspicious Targets: Fig. 2(c) shows that our NUDT-SIRST (sea) dataset has a rich variety of suspicious targets, including tiny clouds, port containers, reefs, and land bright spots. These suspicious targets are very easily confused with real ship targets in shape and brightness and thus cause false alarm.

4) Much Smaller Targets: As shown in Table I, the average target size of our NUDT-SIRST-Sea dataset is only 29 pixels, which is much smaller than the average target size of images in other mainstream SIRST datasets. The target-to-background ratio of our NUDT-SIRST-Sea dataset is 0.000029%, hundreds of times smaller than the target-to-background ratios of NUDT-SIRST [27], NUST-SIRST [25], and NUAA-SIRST [23]. As shown in Fig. 2(d), 76% targets cover less than 0.005% area in space-based images. Targets of other datasets [23], [25], [27], [28] mostly cover over 0.05% area in space-based images. Therefore, much smaller targets in NUDT-SIRST-Sea make this dataset more challenging than other datasets.

5) Much Dimmer Targets: As shown in Table I, our NUDT-SIRST-Sea has the much smaller average target SNR than other datasets [23], [25], [27], [28]. Detailed comparisons among these existing datasets are shown in Fig. 2(e). Datasets, such as NUDT-SIRST [27], NUST-SIRST [25], and NUAA-SIRST [23], mostly focus on bright targets. However, more than 20% of targets have a brightness smaller than 0.5 in our NUDT-SIRST-Sea. In contrast, less than 5% of targets have a brightness smaller than 0.5 in other aerial- and land-based datasets. Compared to other datasets, NUDT-SIRST-Sea is more challenging on dim targets.

6) Multiscale Targets: As shown in Fig. 2(f), the size of different types of ships (e.g., large cruise ships, medium-sized oil recovery wells, and small yachts) varies a lot, ranging from 2 to 500 pixels. Due to the large area occupied by space-based infrared images, targets with different scales often appear in the same scene. Detecting targets with different scales in the same scene is a fairly challenging task.

III. RELATED WORK

In this section, we briefly review the major works in space-based visible tiny ship detection, SIRST detection, and ViT.

A. Space-Based Visible Tiny Ship Detection

Space-based visible tiny ship detection aims to detect tiny ships in remote sensing visible images. Chen et al. [30] proposed a degraded reconstruction enhancement-based method for tiny ship detection in remote sensing images. They incorporated a cross-stage multihead attention module in the detector to further improve the feature discrimination by leveraging the self-attention mechanism and introduced a large-scale dataset. Wu et al. [31] proposed an effective tiny ship detector for low-resolution RSIs (namely, R-TSDet). LR-TSDet consisted of three key components: a filtered feature aggregation (FFA) module, a hierarchical-atrous spatial pyramid (HASP) module,

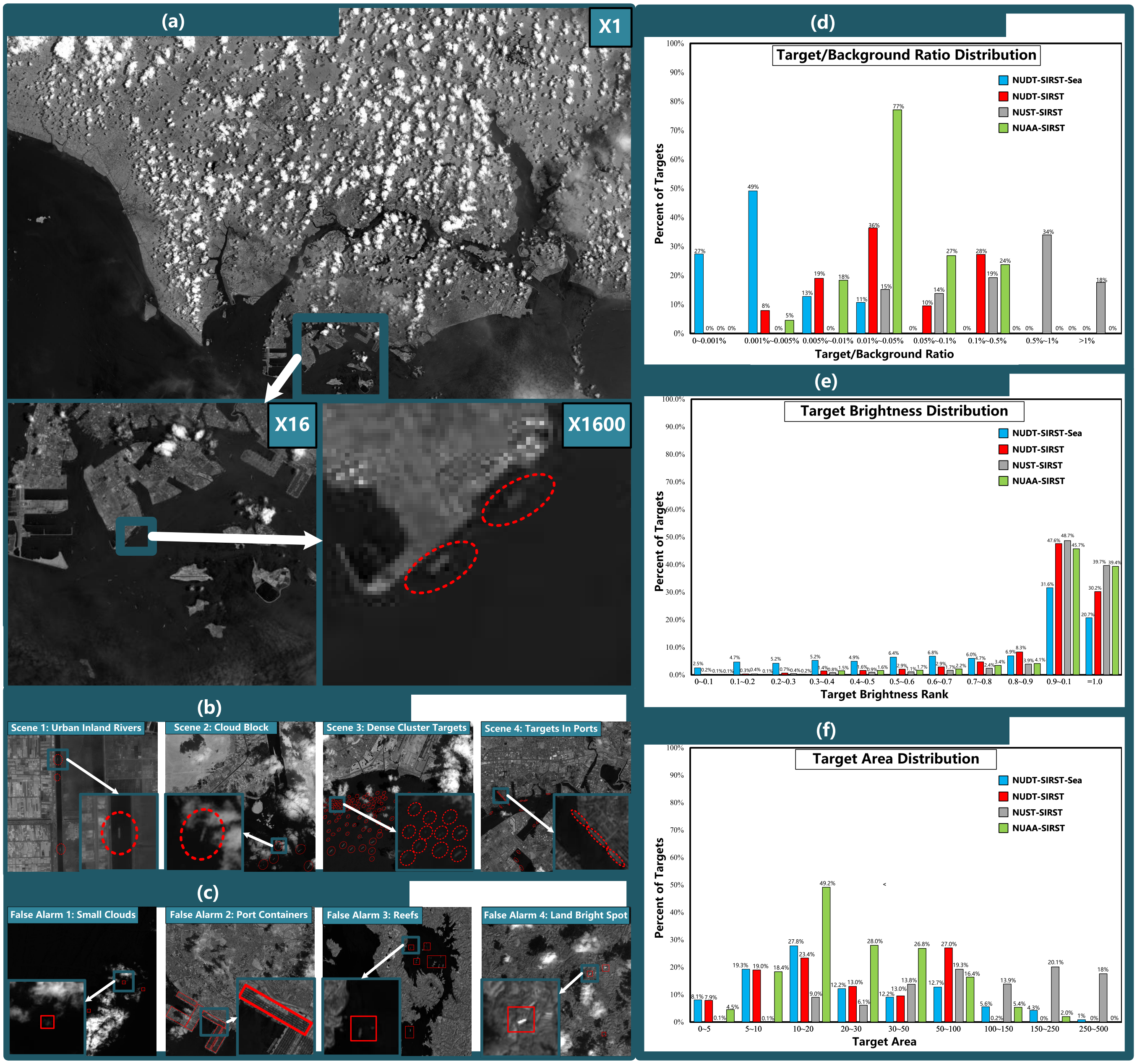


Fig. 2. Overall description of the NUDT-SIRST-Sea dataset. (a) Typical image zoomed in 1, 16, and 1600 times with tiny ships in the NUDT-SIRST-Sea dataset. (b) Illustration of the variety of background. (c) Illustration of the variety of suspicious targets. (d) Distribution of target number with the target-to-background ratio. (e) Distribution of target number with target brightness. (f) Distribution of target number with target area.

and an IoU-Joint loss. Furthermore, they introduced a new dataset called GF1-LRSD collected from the Gaofen-1 satellite for tiny ship detection in low-resolution RSIs. Li et al. [32] proposed a new SDVI algorithm, named enhanced YOLO v3 tiny network, for real-time ship detection. The algorithm can be used in video surveillance to achieve accurate classification and positioning of six types of ships (including ore carrier, bulk cargo carrier, general cargo ship, container ship, fishing boat, and passenger ship) in real time.

However, the above methods are designed for the visible tiny ship detection task. Tiny ships in infrared band are much dimmer and shapeless than those in RGB bands. Moreover, infrared images contain poorer contextual relation than visible images. Multitype suspicious targets (i.e., tiny clouds, port containers, reefs, and land bright spots) exhibit more texture and color differences from the tiny ships in visible images and are more easily confused with targets (i.e., ships under clouds, ships in port, and ships by reefs) in infrared images.

B. SIRST Detection

General SIRST detection tasks (e.g., aerial- and land-based SIRSTs) have been extensively investigated for decades. Many filter-based background methods [6], [7] were proposed. These methods used specifically designed filters for the background noise and clutter suppression. Considering that the small target is more visually salient than its surrounding background, human visual system local contrast-based methods [8], [9], [10], [11], [12], [13] have been proposed. However, clutters can be easily confused with targets in highlighted scenes. To solve this problem, low-rank-based methods were proposed [5], [14], [15], [16], [17], [18], [19]. Nonlocal self-correlation between background patches was used in infrared images to construct low-rank sparse decomposition model. After that, Xia et al. [33] proposed a mechanism named dynamic image structure evolution (DISE) and a DISE-derived single-frame IRSTD framework. Nevertheless, these previous filter-based methods, local contrast-based methods, and low-rank-based methods rely on fixed hyperparameters. When real scenes change dramatically, such as in clutter background, target shape, and target size, it is difficult to use fixed hyperparameters to handle such variations.

Different from traditional methods, CNN-based methods adopt a data-driven training manner to learn common characteristics among small targets. Due to the previous open-sourced SIRST datasets and the powerful detection models, CNN-based methods have achieved promising progress recently. Dai et al. [23] proposed the first segmentation-based CNN network. They designed an asymmetric contextual module to aggregate features from shallow layers and deep layers. Then, Dai et al. [24] further improved their ACM by introducing a dilated local contrast measure in their ALC-Net. Specifically, a feature cyclic shift scheme was designed to achieve a trainable local contrast measure. After that, Wang et al. [25] decomposed the infrared target detection problem into two opposed subproblems (i.e., miss detection and false alarm). They proposed a conditional generative adversarial network (MDvsFA) to achieve the tradeoff between miss detection and false alarm for infrared small target detection. Considering that pooling layers in the networks could lead to the loss of targets in deep layers, Li et al. [27] proposed a DNA-Net. With the help of their specifically designed dense nested interactive module (DNIM), high-level information of small targets can be extracted and the response of small targets can also be maintained in the deep CNN layers. Zhang et al. [28] proposed an infrared shape network. In their network, a Taylor finite difference (TFD)-inspired edge block and a two-orientation attention aggregation (TOAA) block were devised to address the problem of submerging of infrared targets in the background of heavy noise and clutter.

Benefited from these specifically designed architectures and modules, the above CNN-based methods have achieved promising results in the land- and aerial-based SIRST detection tasks. However, space-based SIRST images are quite different from the aerial- and land-based SIRST ones. Smaller, dimer, and more changeable targets make it difficult to achieve high-performance detection under limited receptive fields introduced by CNN architectures. Moreover, poor long-distance dependency capture ability of traditional CNN architectures may result in more false alarm. Therefore, it is necessary to introduce more long-distance information capture modules to further exploit the correlation of targets and background.

C. Vision Transformer

Inspired by the success of the Transformer architectures in the NLP area [34], some works try to apply them in the computer vision area. ViT [29] is the first work to apply Transformer to computer vision. It uses nonoverlapping medium-sized image patches in Transformer to achieve high-precision image classification. With the success of ViT, more promising ViT works emerged. Various ViT-based structures are proposed to handle various high-level tasks (e.g., object detection and semantic segmentation). For example, Carion et al. [35] proposed the first end-to-end object detection network with Transformers (namely, DETR). After that, Chen et al. [36] proposed TransUNet and argued that Transformers can serve as powerful encoders for medical image segmentation tasks with the combination of U-Net [37] to enhance finer details by recovering localized spatial information. In the SIRST detection field, Liu et al. [38] proposed the first work to explore the ViT to detect infrared small-dim targets and achieved promising performance in SIRST. They first used CNN to extract local features. Then, they adopted ViT to learn high-level information of target localization from local features. Next, Qi et al. [39] proposed a fusion network architecture of transformer and CNN (FTC-Net), which consists of two branches. The CNN-based branch uses a U-Net with skip connections to obtain low-level local details of small targets.

Although achieving promising performance, the above transformer-based works are not designed for spaced-based SIRST detection tasks. Specifically, DETR is designed for normal scale object detection and cannot well capture the features of tiny targets. TransUNet mainly focuses on the whole image segmentation performance but pays less attention to local information of tiny targets. The ViT for SIRST method proposed by Liu et al. [38] is mainly designed for aerial- and land-based SIRSTs. However, the space-based SIRST detection task requires both high-level information for target localization and low-level information for shape description. Their single-level ViT structure only applied on the features extracted by the last CNN layer. Thus, their method cannot fully capture low-level information for shape description and easily confuse the real targets with suspicious targets (e.g., tiny clouds, port containers, reefs, and land bright spots). To address the above problems, our MTU-Net combines MVTM and CNN in a multilevel ViT CNN hybrid encoder. CNN extracts multilevel features. Then, MVTM refines the features to capture the long-distance dependency of multilevel features.

IV. Methodology

In this section, we introduce our MTU-Net (Sections IV-A–IV-D), CRRP data augmentation method (Section IV-E), and FocalIoU loss (Section IV-F) in detail.

A. Overall Architecture

As shown in Fig. 3, our MTU-Net takes a single image as its input and sequentially consists of a multilevel ViT CNN hybrid encoder (Section IV-B), a U-shape decoder

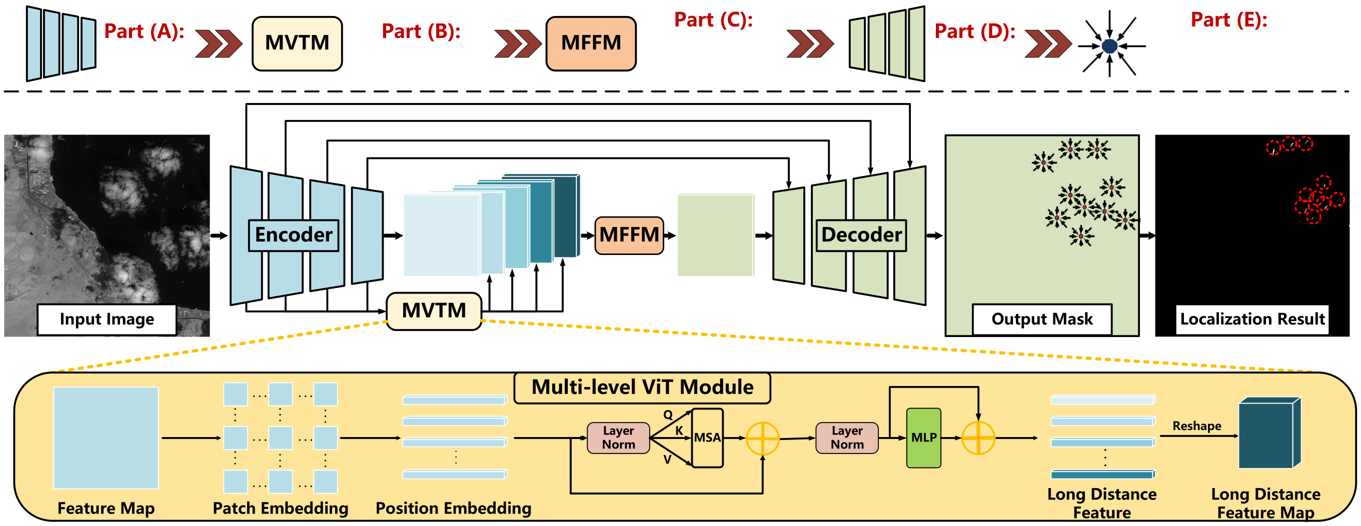


Fig. 3. Illustration of the proposed MTU-Net in this article. (a) Encoder. The input image is fed into the CNN encoder to coarsely extract multiscale features. (b) MVTM. Then, features of different levels go through the MVTM to extract long-distance features. (c) MFFM. The multilevel features are fed into the MFFM, where these features are concatenated and fused to incorporate long-distance information. (d) Decoder. Features with multilevel long-distance information are fed into the U-shape decoder and fused at the nodes of skip connection to generate the final predicted probability map. (e) Eight-connected neighborhood clustering module. The predicted probability map is clustered and the centroid of each target region is finally determined.

(Section IV-C), and an eight-connected neighborhood clustering module (Section IV-D) to generate the pixel-level localization and classification results.

Section IV-B introduces our multilevel ViT CNN hybrid encoder. The input image is first cut into image patches $\boldsymbol{I} \in \mathbb{R}^{C\times H\times W}$, where C, H, and W denote the channel, width, and height of image patch, respectively. Image patches $\boldsymbol{I}$ are preprocessed before being fed into the CNN to coarsely extract multiscale features $\boldsymbol{F_i} \in \mathbb{R}^{C_i\times H_i\times W_i}$ $(i \in \{1, 2, \ldots, k\})$, where k denotes the level numbers of CNN. Then, each feature of different levels F_i $(i \in \{1, 2, \ldots, k-1\})$ goes through the MVTM to obtain $\boldsymbol{V}_i \in \mathbb{R}^{C_i\times H_k\times W_k}$ $(i \in \{1, 2, \ldots, k-1\})$. The multilevel features $\{\boldsymbol{V_i}\}$ $(i \in \{1, 2, \ldots, k-1\}$ and $\boldsymbol{F_k}$ are fed into the multilevel feature fusion module (MFFM). In MFFM, features are concatenated and fed to a 1×1 convolution to generate the features $\boldsymbol{M_k} \in \mathbb{R}^{C_k\times H_k\times W_k}$ with multilevel long-distance information. Section IV-C introduces the U-shape decoder. Features $\boldsymbol{M_k}$ with multilevel long-distance information are fed into the decoder and fused with $\boldsymbol{F_i}$ $(i \in \{k-1, k-2, \ldots, 1\})$ at the nodes of skip connection to generate $\boldsymbol{M_i}$ $(i \in \{k-1, k-2, \ldots, 1\})$ and final predicted probability map $\boldsymbol{P}$. Section IV-D elaborates the eight-connected neighborhood clustering module. The final predicted probability maps $\boldsymbol{P}$ are fed into this module to calculate the spatial locations of target centroid, which are then used for comparison in Section V-C.

B. Multilevel ViT CNN Hybrid Encoder

1) Motivation: As shown in Fig. 3, our MTU-Net consists of a multilevel ViT CNN hybrid encoder, a U-shape decoder, and an eight-connected neighborhood clustering module to generate the pixel-level localization and classification results. To achieve efficient feature extraction for extremely large images (e.g., $10\,000 \times 10\,000$ resolution), the images are first cut into 1024×1024 patches and then fed into the ResNet-18 [40] to extract multiscale local features. To distinguish multitype suspicious targets in complex backgrounds, more long-distance information is required. The proposed MVTM refines the extracted multiscale local features. In this way, the long-distance dependency of suspicious targets in complex background is captured from high-level features. Moreover, multiscale infrared small targets are significantly different in their sizes, ranging from 1 pixel (i.e., point targets) to tens of pixels (i.e., extended targets). With the increase of network layers, high-level information of target localization is obtained, while the shape description of extended targets is easily lost after multiple downsamplings. Therefore, we designed an MFFM to fuse multilevel features extracted by MVTM. In this way, high-level information for target localization and low-level information for shape description can be fused and enhanced by our MFFM.

2) Multilevel ViT Module: MVTM contains $(k-1)$ ViT branches, and all branches have the same structure. we adopt ResNet-18 [40] as the feature embedding module to extract multiscale local features $\boldsymbol{F_i} \in \mathbb{R}^{C_i\times H_i\times W_i}$ $(i \in \{1, 2, \ldots, k\})$. Features $\boldsymbol{F_i}$ are flattened into 2-D patches $\boldsymbol{E_{em}}^{(i)} \in \mathbb{R}^{N_i\times(P_i{}^2C_i)}$, where $N_i = (H_iW_i/P_i{}^2)$ is the number of patches and (P_i, P_i) is the resolution of each patch. After position embedding, we get $\boldsymbol{E}^{(i)}_{\mathbf{pos}}$. We obtain embedded tokens $\boldsymbol{E}^{(i)} = \boldsymbol{E}^{(i)}_{\mathbf{em}} + \boldsymbol{E}^{(i)}_{\mathbf{pos}}$, where n is the number of tokens and $n = (H_iW_i/P_i^2)$. Embedded tokens $\boldsymbol{E}^{(i)}$ are divided into m heads $\boldsymbol{E}^{(i)} = \{\boldsymbol{E}^{(i)}_1, \boldsymbol{E}^{(i)}_2, \ldots, \boldsymbol{E}^{(i)}_j, \ldots, \boldsymbol{E}^{(i)}_m\}$, $\boldsymbol{E}^{(i)}_j \in \mathbb{R}^{n\times(C_i/m)}$ $(j \in \{1, 2, \ldots, m\})$ and then fed into the multihead self-attention module MSA to obtain interaction tokens $\boldsymbol{E}^{(i)}_a$. We define these processes as

$$\boldsymbol{E}^{(i)}_a = \mathrm{MSA}[\mathrm{LN}(\boldsymbol{E}^{(i)})] + \boldsymbol{E}^{(i)} \tag{1}$$

where LN is the layer normalization.

In each head, the multihead self-attention module MSA defines three trainable weight matrices to transform queries $\boldsymbol{Q}^{(i)}$, keys $\boldsymbol{K}^{(i)}$, and values $\boldsymbol{V}^{(i)}$. Then, $\boldsymbol{E}_{\boldsymbol{a}}^{(i)}$ are fed into the MLP module to obtain final tokens $\boldsymbol{E}_{\boldsymbol{b}}^{(i)}$, and the result of MLP can be expressed as

$$\boldsymbol{E}_{\boldsymbol{b}}^{(i)} = \mathrm{MLP}\big[\mathrm{LN}\big(\boldsymbol{E}_{\boldsymbol{a}}^{(i)}\big)\big] + \boldsymbol{E}_{\boldsymbol{a}}^{(i)} \tag{2}$$

where $\boldsymbol{E}_{\boldsymbol{b}}^{(i)} \in \mathbb{R}^{N_i \times ({P_i}^2 C_i)}$.

Then, we reshape tokens $\boldsymbol{E}_{\boldsymbol{b}}^{(i)} \in \mathbb{R}^{{P_i}^2 C_i}$ into features $\boldsymbol{V}_i$, where $P_i = H_k = W_k$. The result of the ViT branch can be expressed as

$$\boldsymbol{V}_{\boldsymbol{i}} = \mathrm{Reshape}\big\{\mathrm{MLP}\big[\mathrm{LN}\big(\boldsymbol{E}_{\boldsymbol{a}}^{(i)}\big)\big]\big\} + \boldsymbol{E}_{\boldsymbol{a}}^{(i)}\big). \tag{3}$$

3) Multilevel Feature Fusion Module: In our MFFM, features are fused by capturing the long-distance dependency of these extracted high-level features. The multilevel features $\{\boldsymbol{V}_{\boldsymbol{i}}\}$ $(i \in \{1, 2, \ldots, k-1\}$ and CNN local features $\boldsymbol{F}_{\boldsymbol{k}}$ are first concatenated and then fed to an 1×1 convolution to fuse the features $\boldsymbol{M}_{\boldsymbol{k}} \in \mathbb{R}^{C_k \times H_k \times W_k}$. All multilevel features are fused by capturing the long-distance dependency of these extracted high-level features. The fusion features can be expressed as

$$\boldsymbol{M}_{\boldsymbol{k}} = \mathrm{Conv}[\mathrm{Concat}(\boldsymbol{F}_{\boldsymbol{k}}, \boldsymbol{V}_{\boldsymbol{k-1}}, \boldsymbol{V}_{\boldsymbol{k-2}}, \ldots, \boldsymbol{V}_{\boldsymbol{1}})]. \tag{4}$$

C. U-Shape Decoder

To obtain confidence maps of small targets, we adopt a decoder to upsample multilevel features $\boldsymbol{M}_{\boldsymbol{k}}$. Multilayer features $\{\boldsymbol{F}_{\boldsymbol{i}}\}$ $(i \in \{1, 2, \ldots, k-1\}$ in the encoder are concatenated with features obtained by upsampling operation through skip connection operation to generate $\boldsymbol{M}_{\boldsymbol{i}}$ $(i \in \{k-1, k-2, \ldots, 1\})$. The processing of the decoder can be expressed as

$$\boldsymbol{M}_{\boldsymbol{i-1}} = \mathrm{Conv}\{\mathrm{Concat}[\boldsymbol{F}_{\boldsymbol{i-1}}, \mathrm{Upsample}(\boldsymbol{M}_{\boldsymbol{i}})]\}. \tag{5}$$

A robust predicted probability map can be expressed as

$$\boldsymbol{P} = \mathrm{Sigmoid}(\boldsymbol{M}_{\boldsymbol{0}}). \tag{6}$$

D. Eight-Connected Neighborhood Clustering Module

After the U-shape decoder, we introduce an eight-connected neighborhood clustering module [41] to clutter all pixels and calculate the centroid of each target. If any two pixels (m_0, n_0) and (m_1, n_1) in feature maps $\boldsymbol{P}$ have intersection areas in their eight neighborhoods, i.e.,

$$\mathrm{N}_8(m_0, n_0) \cap \mathrm{N}_8(m_1, n_1) \neq 0 \tag{7}$$

where $\mathrm{N}_8(m_0, n_0)$ and $\mathrm{N}_8(m_1, n_1)$ represent the eight neighbor-hoods of pixel (m_0, n_0) and (m_1, n_1), respectively. Then, (m_0, n_0) and (m_1, n_1) are judged as adjacent pixels. If these two pixels have the same value, i.e.,

$$p(m_0, n_0) = p(m_1, n_1) \quad \forall (p(m_0, n_0), p(m_1, n_1)) \in \boldsymbol{P} \tag{8}$$

where $p(m_0, n_0)$ and $p(m_1, n_1)$ represent the value of pixel (m_0, n_0) and (m_1, n_1), respectively, and these two pixels are considered to belong to the same target area. Once all targets in the image are determined, centroid can be calculated according to their coordinates.

E. Data Augmentation

As mentioned in Section II, the distribution of the foreground targets and background is extremely imbalanced in our NUDT-SIRST-Sea. This foreground–background imbalance issue makes the network pay more attention to those uninformative background regions and thus hinders the quick convergence of the network. Copy–paste (CP) data augmentation is a powerful data augmentation method for instance segmentation [42]. Based on the CP data augmentation method, we further propose a CRRP data augmentation method (namely, CRRP) to manually increase the ratio of candidate targets in the training phase and thus accelerate the convergence of the network. Our CRRP data augmentation method copies both targets and target neighborhood background, while the CP data augmentation method only copy targets. In this way, our CRRP data augmentation method can well preserve the information of target itself and contextual information between targets and background. Otherwise, suspicious targets (e.g., tiny clouds, port containers, reefs, and land bright spots) are detected as targets without the contextual dependency. Therefore, our CRRP is a more suitable data augmentation method for space-based SIRST detection task compared to the CP method.

As shown in Fig. 4(a), we first collect images of the targets' neighborhood and randomly copy one target. Then, the selected targets are randomly rotated. After that, the target is randomly resized as a candidate target. Finally, we paste the candidate target into the background area of image background region. As shown in Fig. 4(b), the imbalance of the foreground targets and background distribution is relieved and the training time is also greatly reduced compared to previous simple data augmentation methods (e.g., rotate, translate, and color jitter).

F. FocalIoU Loss

Focal loss [43] focuses on hard samples (e.g., small-scale targets, edges of targets, and suspicious targets), which helps target localization. However, the focal loss causes more false alarm due to the high response in the background suspicious area. SoftIoU loss [44] focuses on large-scale targets and loses small-scale targets. This is because large-scale targets contribute much more in IoU than small-scale targets, resulting in the loss of small-scale targets. To achieve the "double-win" of target localization and shape description, we combine the SoftIoU loss and the focal loss to develop a FocalIoU loss. Our FocalIoU loss combines the advantages of the focal loss and the SoftIoU loss, with a low response in background areas, and focuses on small-scale targets. The formula of our FocalIoU loss function is expressed as (11)

$$\mathrm{FL}(p, y) = -y(1-p)^{\gamma}\log(p) - (1-y)p^{\gamma}\log(1-p) \tag{9}$$

$$\mathrm{SoftIoU} = \frac{\mathrm{smooth} + \sum p \times y}{\mathrm{smooth} + \sum p + \sum y - \sum p \times y} \tag{10}$$

$$\mathrm{FIoUL}(p, y) = 2(1 - \mathrm{SoftIoU})[\mathrm{FL}(p, y)]^{\frac{1+\mathrm{SoftIoU}}{2}} \tag{11}$$

where p denotes the probability of each pixel, y denotes the label of each pixel in probability map $\boldsymbol{P}$, and γ is an adjustable

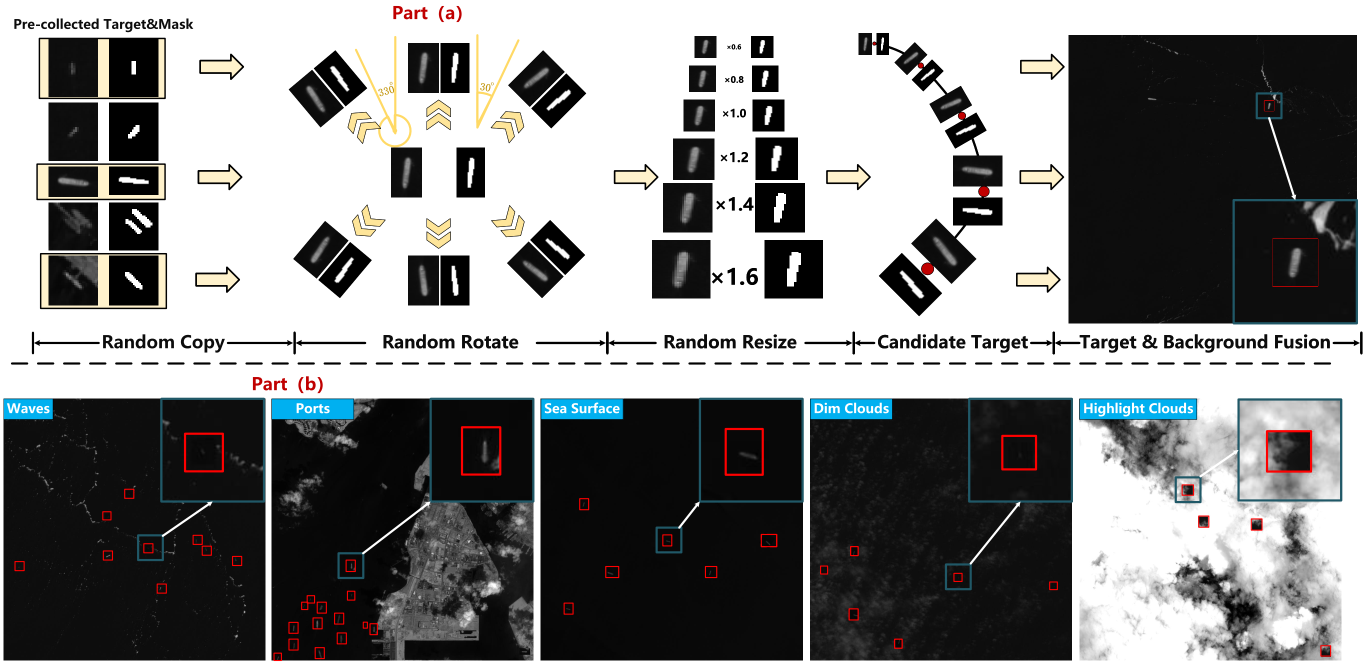


Fig. 4. Illustration of the CRRP data augmentation method. (a) CRRP data augmentation. Images of the targets' neighborhood are first collected and randomly copied. Then, the selected target is randomly rotated. After that, the target is randomly resized as a candidate target. Finally, the candidate target is pasted into the background area of image background region. (b) Samples of synthesized images.

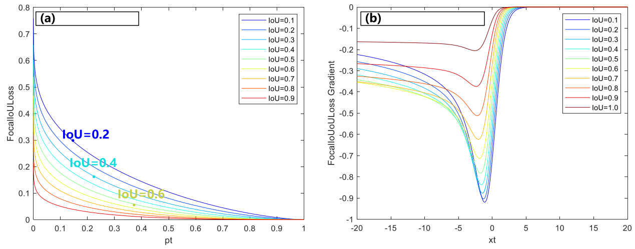


Fig. 5. FocalIoU loss analysis. (a) FocalIoU loss function curve. (b) FocalIoU loss function gradient curve.

factor to control the attention on hard samples. FL and FIoUL are the abbreviations of focal loss and FocalIoU loss, respectively. SoftIoU is a convergent IoU with an adjustable factor smooth to avoid infinity.

To further analyze our FocalIoU loss function, we derive the FocalIoU loss function in (13)

$$\frac{\partial \mathrm{FL}(p, y)}{\partial p} = -(1-y)\gamma\, p^{\gamma-1}\log(1-p)+(1-y)p^{\gamma}\frac{1}{1-p} + y\gamma\,(1-p)^{\gamma-1}\log(p) - y(1-p)^{\gamma}\frac{1}{p} \tag{12}$$

$$\frac{\partial \mathrm{FIoUL}(p, y)}{\partial x} = (1-\mathrm{SoftIoU}^2)\mathrm{FL}(p, y)^{\frac{1-\mathrm{SoftIoU}}{2}} \cdot \frac{\partial \mathrm{FL(p, y)}}{\partial p}\frac{\partial p}{\partial x} \tag{13}$$

where $p = \mathrm{Sigmoid}(x)$, in which x is the value of each pixel in the MTU-Net output map and p is the probability value of each pixel.

As shown in Fig. 5(a), samples with low IoU output result in a high FocalIoU loss and a sharp decrease of the FocalIoU loss. When IoU is small, the overall segmentation performance of this image is poor, and the FocalIoU loss focuses on difficult simple samples (e.g., large-scale targets) more than difficult samples. Consequently, F_a decreases, while IoU increases. When IoU is large, the FocalIoU loss performs like the focal loss and focuses more on difficult samples, which helps P_d to increase.

V. Experiment

In this section, we first introduce our evaluation metrics and implementation details. Then, we compare our MTU-Net to several SOTA SIRST detection methods. Finally, we present ablation studies to investigate our network.

A. Evaluation Metrics

Following the pioneering DNA-Net [27], we adopt probability of detection (P_d) and false alarm rate (F_a) to evaluate the localization performance and use the IoU to evaluate the shape description performance. Besides, we adopt a receiver operating characteristic (ROC) [45] analysis to further show the overall detection effectiveness, target detection ability, and background suppression ability

1) Probability of Detection: The probability of detection (P_d) is a target-level evaluation metric. It measures the ratio of correctly predicted target number T_{correct} over all target number T_{All}. P_d is defined as

$$P_d = \frac{T_{\mathrm{correct}}}{T_{\mathrm{All}}}. \tag{14}$$

If the centroid deviation of the target is smaller than the pre-defined deviation threshold D_{thresh}, we consider those targets as correctly predicted ones. We set the pre-defined deviation threshold as 3 in this article.

2) False Alarm Rate: False alarm rate (F_a) is another target-level evaluation metric. It is used to measure the ratio of falsely predicted pixels P_{false} over all image pixels P_{All}. F_a is defined as

$$F_a = \frac{P_{\text{false}}}{P_{\text{All}}}. \tag{15}$$

If the centroid deviation of the target is larger than the pre-defined deviation threshold, we consider those pixels as falsely predicted ones.

3) Intersection Over Union: It is a target-level evaluation metric. It evaluates the target description performance of an algorithm. IoU is calculated as the ratio of intersection and the union areas between the target predictions and target labels. The IoU is defined as

$$\text{IoU} = \frac{\text{Target}_{\text{inter}}}{\text{Target}_{\text{Union}}} \tag{16}$$

where $\text{Target}_{\text{inter}}$ and $\text{Target}_{\text{Union}}$ represent the interaction areas and union areas between target prediction and target label, respectively.

4) 3-D ROC: ROC [45] analysis is widely applied in object detection. Common ROC curves generally include a 3-D ROC curve specified by threshold τ, probability of detection P_d, and false alarm rate F_a, and three 2-D ROC curves of (τ, P_d), (τ, $P_d(\tau)$), and (τ, $F_a(\tau)$). Both $P_d(\tau)$ and $F_a(\tau)$ can be calculated in (14) and (15) by changing the predicted pixel threshold τ. The 2-D ROC curves of (F_a, P_d), (τ, P_d), and (τ, F_a) indicate overall detection effectiveness, target detection ability, and background suppression ability of different methods, respectively. A good detector has 2-D ROC curves of (F_a, P_d), (τ, P_d), and (τ, F_a) close to the top-left, top-right, and bottom-left corner of the coordinate axis, respectively. However, due to the existence of intersection between ROC curves, it is difficult to judge which one has a better performance for the closer ROC curve. To quantitatively evaluate all methods, AUC values of three 2-D ROC curves are introduced, which are expressed as $\text{AUC}_{(F_a,P_d)}$, $\text{AUC}_{(\tau,P_d)}$, and $\text{AUC}_{(\tau,F_a)}$. Higher $\text{AUC}_{(F_a,P_d)}$ and $\text{AUC}_{(\tau,P_d)}$ values represent better detection performance, while lower $\text{AUC}_{(\tau,F_a)}$ values denote better background supression capability. AUC_{OA} and AUC_{SNPR} further represent the overall accuracy (OA) and signal-to-noise probability ratio (SNPR) based on the above three AUC values, and the expression is given as follows:

$$\text{AUC}_{\text{OA}} = \text{AUC}_{(F_a,P_d)} + \text{AUC}_{(\tau,P_d)} - \text{AUC}_{(\tau,F_a)} \tag{17}$$

$$\text{AUC}_{\text{SNPR}} = \frac{\text{AUC}_{(\tau,P_d)}}{\text{AUC}_{(\tau,F_a)}}. \tag{18}$$

Similarly, higher AUC_{OA} and AUC_{SNPR} values denote better detection performance and background clutter suppression capability, respectively.

B. Implementation Details

The NUDT-SIRST-Sea dataset contains 41 images for training and seven images for test. These real images were captured by sensors mounted on a low Earth-orbiting satellite. All input images with a resolution of 10 000 × 10 000 were first cut into patches with a resolution of 1024 × 1024. Before training, all input images were first normalized. Then, these normalized images were sequentially processed by random image flip, Gaussian blurring, and CRRP for data augmentation before being fed into the network. ResNet-18 [40] was chosen as our segmentation backbone. The number of downsampling layers i was 4. Our network was trained using the FocalIoU loss function and optimized by the Adagrad method [46] with the CosineAnnealingLR scheduler. We initialized the weights and bias of our model using the Xavier method [47]. We set the learning rate, batch size, and epoch as 0.05, 8, and 1500, respectively. All models were implemented in PyTorch [48] on a computer with an AMD Ryzen 9 3950X @ 2.20-GHz CPU and an Nvidia RTX 3090 GPU.

C. Comparison to the SOTA Methods

To demonstrate the superiority of our method, we compare our MTU-Net with several SOTA methods, including traditional methods (filtering-based methods: Top-Hat [6] and Max-Median [7]; local contrast-based methods: TLLCM [10] and WSLCM [11]; local rank-based methods: NRAM [15], RIPT [16], and PSTNN [17]; and CNN-based methods, including DNA-Net [27], MDvsFA-cGAN [25], ACM [23], ALC-Net [24], and ResU-Net [49]) on the NUDT-SIRST-Sea dataset. For a fair comparison, we retrained all the CNN-based methods on our NUDT-SIRST-Sea dataset.

1) Qualitative Results: Qualitative results on our NUDT-SIRST-Sea are shown in Fig. 6. Compared with traditional methods, our method can generate more precise localization and classification results with smaller F_a. The results achieved by traditional methods easily lose dense small-scale targets [Fig. 6(a)] and targets in port [Fig. 6(b)]. Traditional methods generate bad shape segmentation in dim targets [Fig. 6(c)]. The CNN-based methods (MDvsFA-cGAN, ResU-Net, ACM, ALC-Net, and DNA-Net) perform much better than traditional methods. However, due to the extremely dim targets [Fig. 6(a) and (c)] in our NUDT-SIRST-Sea, MDvsFA-cGAN loses more targets. Our MTU-Net can generate better shape segmentation [Fig. 6(c)] than DNA-Net, ACM, and ALC-Net. Our MTU-Net can generate better target localization in scenes with port ships [Fig. 6(b)]. This is because our MTU-Net can effectively capture long-distance dependency with the help of coarse-to-fine MVTM.

2) Quantitative Results: Similar to DNA-Net [27], we first obtained their predicts and then performed noise suppression by setting a threshold to remove low-response areas for all the compared algorithms. Specifically, the adaptive threshold (T_{adaptive}) was calculated for traditional methods according to

$$T_{\text{adaptive}} = \text{Max}[0.7\text{Max}(\boldsymbol{P}), 0.5\sigma(\boldsymbol{P}) + \text{Avg}(\boldsymbol{P})] \tag{19}$$

where $\text{Max}(\boldsymbol{P})$ represents the largest value of output, T_{adaptive} is an adaptive threshold, and $\sigma(\boldsymbol{P})$ and $\text{avg}(\boldsymbol{P})$ denote the standard derivation and average value of output, respectively.

For deep learning-based methods, we followed their original papers and adopted their fixed thresholds (i.e., 0, 0, 0, 0, and 0.5 for DNA-Net [27], ResU-Net [49], ACM [23], ALC-Net [24], and MDvsFA-cGAN [25], respectively). We kept all remaining parameters the same as their original papers.

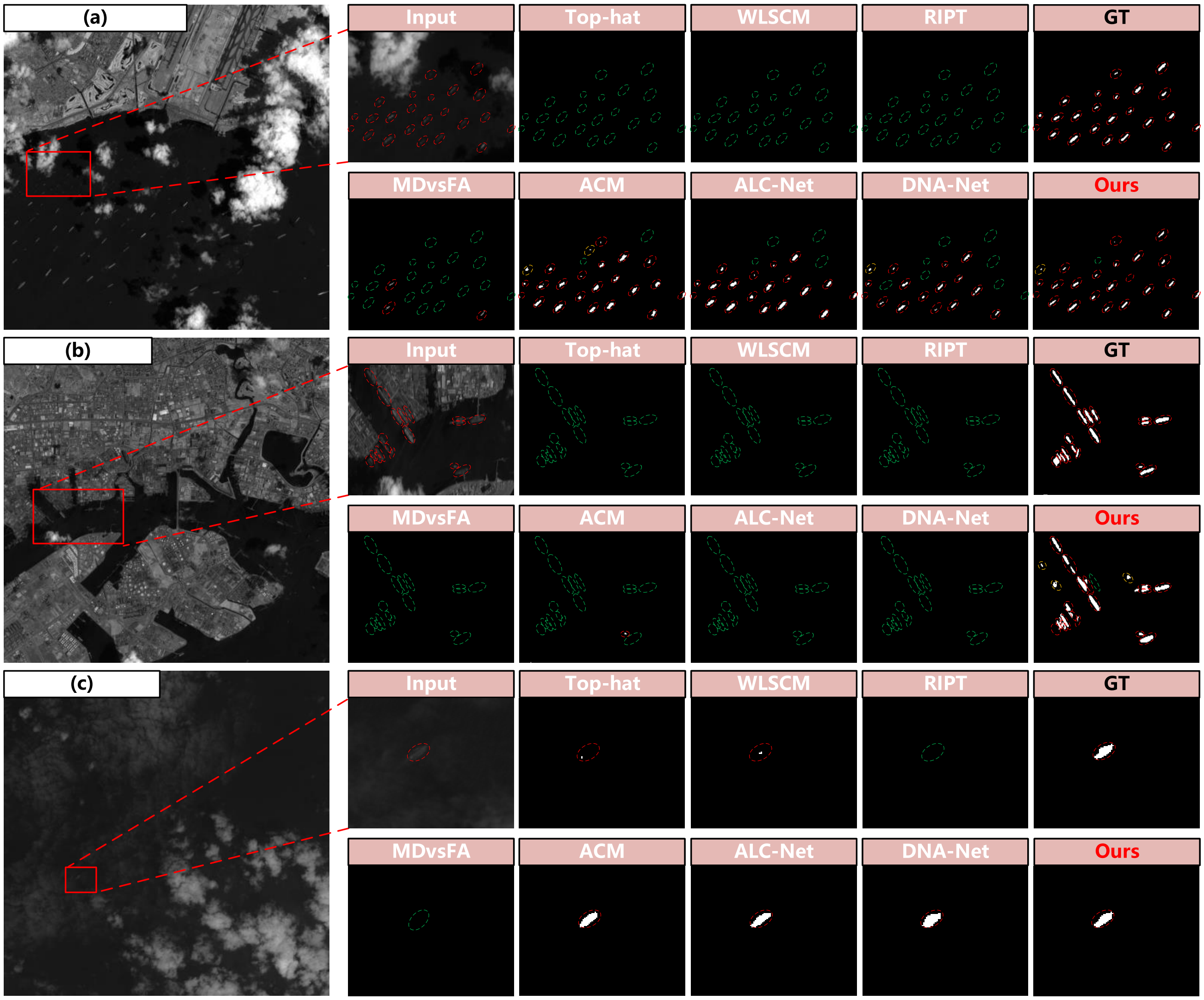


Fig. 6. Qualitative results achieved by different SIRST detection methods in three typical scenes: (a) dense targets, (b) port ships, and (c) dim targets. For better visualization, the target area is enlarged. The correctly detected target, false alarm, and miss detection areas are highlighted by red, yellow, and green dotted circles, respectively. Our MTU-Net can generate output with precise target localization and shape segmentation under a smaller F_a.

TABLE II

IoU, P_d, AND F_a VALUES ACHIEVED BY DIFFERENT TRADITIONAL AND CNN-BASED SOTA METHODS ON THE NUDT-SIRST-SEA DATASET. FOR IoU AND P_d, LARGER VALUES INDICATE HIGHER PERFORMANCE. FOR F_a, SMALLER VALUES INDICATE HIGHER PERFORMANCE. THE BEST RESULTS ARE IN RED AND THE SECOND BEST RESULTS ARE IN BLUE

Traditional methods				Deep learning based methods			
Method Description	IoU $(\times 10^{-2})$	P_d $(\times 10^{-2})$	F_a $(\times 10^{-6})$	Method Description	IoU $(\times 10^{-2})$	P_d $(\times 10^{-2})$	F_a $(\times 10^{-6})$
Filtering Based: Top-Hat [6]	1.17	10.39	95.37	CNN Based: ACM [23]	47.57	70.46	21.31
Filtering Based: Max-Median [7]	0.28	6.62	46.17	CNN Based: ALC-Net [24]	48.9	58.65	9.13
Local Contrast Based: WSLCM [11]	0.60	8.02	7.33	CNN Based: MDvsFA-cGAN [25]	0.37	0.18	92
Local Contrast Based: TLLCM [10]	0.59	10.27	14.41	CNN Based: ResU-Net [49]	46.05	60.18	7.92
Local Rank Based: MSLSTIPT [5]	0.33	6.68	6283	CNN Based: DNA-Net [27]	42.17	61.60	17.19
Local Rank Bas ed: NRAM [15]	0.39	5.89	**3.585**	**MTU-Net-ResNet10 (ours)**	59.98	81.22	16.64
Local Rank Based: RIPT [16]	0.36	6.32	**1.9186**	**MTU-Net-ResNet18 (ours)**	**64.14**	**85.44**	11.72
Local Rank Based: PSTNN [17]	1.50	7.59	15.44	**MTU-Net-ResNet34 (ours)**	**62.00**	**84.70**	12.12

Quantitative results are shown in Table II. Our MTU-Net outperforms traditional methods significantly. This is because NUDT-SIRST-Sea contains challenging images with various scales, orientations, and brightness of tiny ship targets. Our MTU-Net can effectively capture long-distance features between background and targets. Limited by manually selected

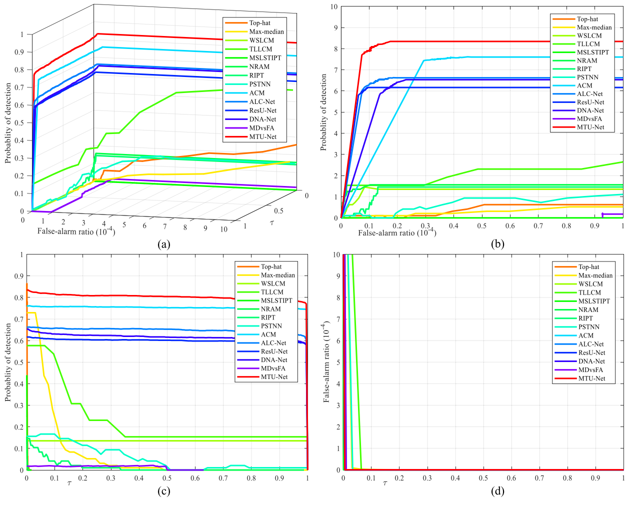


Fig. 7. 3-D ROC and three corresponding 2-D ROC curves of different comparing methods. (a) 3-D ROC curve. (b) 2-D ROC curve of (F_a, P_d). The curve of (F_a, P_d) closer to the top-left corner of the coordinate axis indicates a higher detection effectiveness. (c) 2-D ROC curve of (τ, P_d). The curve of (τ, P_d) closer to the top-right corner of the coordinate axis indicates a better target detection ability. (d) 2-D ROC curve of (τ, F_a). The curve of (τ, F_a) closer to the bottom-left corner of the coordinate axis indicates a better background suppression ability.

parameters, those model-driven traditional methods cannot well cope with such challenging scenes. It is worth noting that the improvements achieved by MTU-Net over other deep learning-based methods (i.e., MDvsFA-cGAN, ACM, ALC-Net, and DNA-Net) are obvious. Our MTU-Net achieves 64.14% on IoU, 85.44% on P_d, and 11.72×10^{-6} on F_a. Our MTU-Net outperforms other deep learning methods more than 15% on IoU and P_d. Besides, our MTU-Net only suffers a decrease of 3.82×10^{-6} in terms of F_a than ResU-Net. This is because our MTU-Net can effectively capture long-distance dependency with the help of coarse-to-fine MVTM. The long-distance dependency helps network detect small targets and generate less false alarm in suspicious targets.

3) 3-D ROC Analysis: Fig. 7 shows four types of ROC curves corresponding to the detection maps. It can be observed that the 2-D ROC curve of (F_a, P_d) of our MTU-Net is much closer to the top-left corner than that of other methods in Fig. 7(b). The 2-D ROC curves of (F_a, P_d) show that our MTU-Net has a better detection effectiveness than other methods. The 2-D ROC curve of (τ, F_a) of our MTU-Net in Fig. 7(c) achieves a much higher P_d when τ is smaller than 0.6. The 2-D ROC curves of (τ, F_a) in Fig. 7(b) show that our MTU-Net has a much better detection background suppression ability than traditional methods and the same ability as other deep learning-based methods. Our MTU-Net achieves the best overall performance on detection effectiveness, target detection ability, and background suppression ability.

To quantitatively evaluate the effectiveness of our method, we test five commonly used AUC indicators on NUDT-SIRST-Sea. The experimental results are listed in Table III. Table III shows that our MTU-Net obtains a good $\text{AUC}_{(\tau,F_a)}$, the best $\text{AUC}_{(\tau,P_d)}$, and the best $\text{AUC}_{(P_d,F_a)}$. It indicates that our MTU-Net has good background suppression ability and detection ability. Since it is difficult to judge which method has better detection performance by any of $\text{AUC}_{(F_a,P_d)}$, $\text{AUC}_{(\tau,P_d)}$, and $\text{AUC}_{(\tau,F_a)}$, we use AUC_{OA} and AUC_{SNPR} to fully evaluate the performance of all methods. The best AUC_{OA} and the second best AUC_{SNPR} obtained by MTU-Net represent that our MTU-Net can accurately detect targets and effectively suppress background clutters and noise.

D. Ablation Study

In this section, we compare our MTU-Net with several variants to investigate the potential benefits introduced by our MVTM, CRRP data augmentation, and FocalIoU loss function. The results are shown in Tables IV–VI.

1) Multilevel ViT Module: The MVTM is used to achieve coarse-to-fine feature extraction. In our MVTM, ViT refines the CNN features by capturing the long-distance dependency of these extracted high-level features. To demonstrate the effectiveness of our MVTM, we introduced the following network variants.

1) *MTU-Net w/o level **k** ViT:* Level k denotes that we removed $1 \sim k$ ViT branches from our MVTM. For a fair comparison, we made a model size comparable and retrained these variants in NUDT-SIRST-Sea.
2) *MTU-Net w/o MVTM:* We removed MVTM from our MTU-Net. For a fair comparison, we made a model size comparable and retrained this variant on NUDT-SIRST-Sea.

As shown in Table IV, MTU-Net w/o level 1 ViT suffers decreases of 3.93% and 2.85% and an increase of 0.86×10^{-6} in terms of IoU, P_d, and F_a values over MTU-Net on the NUDT-SIRST-Sea dataset. As the number of ViT branch decreases, the values of MTU-Net in IoU and P_d gradually decrease and the value of F_a gradually increases. This is because fewer multilevel features are extracted by MVTM in the multilevel ViT CNN hybrid encoder. Since fewer long-distance information is used, the performance is poor. Specifically, when all ViT branches are pruned and MVTM is removed, MTU-Net suffers decreases of 11.20% and 7.06% and an increase of 32.04×10^{-6} in terms of IoU, P_d, and F_a values.

2) CRRP Data Augmentation: Since the distribution of the labels and background of the dataset is extremely imbalanced. This problem misleads the network to focus more on the background region of the image. The imbalance causes more false alarm and reduces the convergence speed. Therefore, we use a data augmentation method as a parameter-free solution to alleviate this problem.

1) *MTU-Net w/o DA:* We removed CRRP data augmentation in this variant and retrained our MTU-Net on NUDT-SIRST-Sea.
2) *MTU-Net With CP-DA:* We used the CP data augmentation method in this variant and retrained our MTU-Net on NUDT-SIRST-Sea.
3) *MTU-Net With CRRP-DA:* We used the CRRP data augmentation method in this variant and retrained our MTU-Net on NUDT-SIRST-Sea.

TABLE III

DETECTION ACCURACY AND BACKGROUND SUPPRESSION CAPABILITY OF DIFFERENT COMPARISON METHODS ON THE NUDT-SIRST-SEA DATASET. THE BEST RESULTS ARE IN RED AND THE SECOND BEST RESULTS ARE IN BLUE

Method Description	$\mathrm{AUC}_{(F_a,P_d)}$	$\mathrm{AUC}_{(\tau,P_d)}$	$\mathrm{AUC}_{(\tau,F_a)}$	AUC_{OA}	AUC_{SNPR}
Filtering Based: Top-Hat [6]	**0.784**	0.071	0.0026	0.8524	26.83
Filtering Based: Max-Median [7]	0.4735	0.0163	1.132×10^{-4}	0.4896	143.8983
Local Contrast Based: WSLCM [11]	0.0952	0.0055	$\mathbf{4.1405 \times 10^{-7}}$	0.1006	1.3218×10^{4}
Local Contrast Based: TLLCM [10]	0.4522	0.0384	9.3276×10^{-6}	0.4905	4.113×10^{3}
Local Rank Based: MSLSTIPT [5]	0.2953	0.0195	0.0141	0.3006	1.3819
Local Rank Based: NRAM [15]	0.1312	0.0165	$\mathbf{4.19430 \times 10^{-7}}$	0.1477	3.927×10^{4}
Local Rank Based: RIPT [16]	0.038	0.0116	2.0218×10^{-6}	0.0496	5.7537×10^{3}
Local Rank Based: PSTNN [17]	0.1257	0.0164	1.7932×10^{-5}	0.1941	915.5587
CNN Based: ACM [23]	0.7541	**0.7537**	3.68×10^{-5}	**1.51**	2.05×10^{4}
CNN Based: ALC-Net [24]	0.6455	0.6489	1.34×10^{-5}	1.29	4.85×10^{4}
CNN Based: ResU-Net [49]	0.6003	0.6008	0.795×10^{-5}	1.20	$\mathbf{7.54 \times 10^{4}}$
CNN Based: DNA-Net [27]	0.6243	0.6178	1.723×10^{-5}	1.24	3.58×10^{4}
CNN Based: MDvsFA-cGAN [25]	0.0163	0.0091	9.27×10^{-5}	0.0252	97.68
MTU-Net (Ours)	**0.8091**	**0.8027**	1.09×10^{-5}	**1.611**	$\mathbf{7.36 \times 10^{4}}$

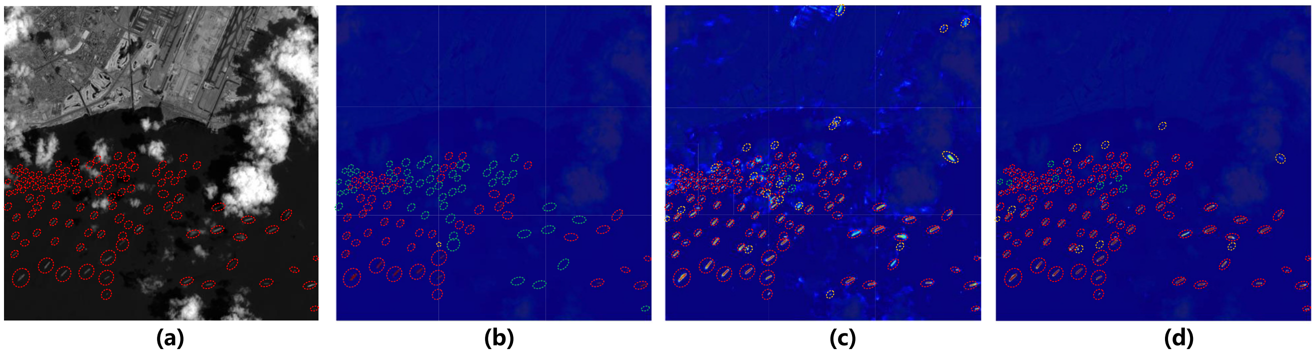


Fig. 8. Visualization maps of MTU-Net output. The output of MTU-Net is marked by a red solid frame. The correctly detected target, false alarm, and miss detection areas are highlighted by red, yellow, and green dotted circles, respectively. (a) Input Image. (b) Final probability density map from the output of MTU-Net (SoftIoU loss). The map shows the responses low in the background area and loses more small-scale targets. (c) Final probability density map from the output of MTU-Net (focal loss). The map shows the output of MTU-Net (focal loss) responses more in small-scale targets area and causes more false alarm. (d) Final probability density map of MTU-Net (FocalIoU loss). The map has a low response in the background area and focuses on small-scale targets.

TABLE IV

IOU, P_d, AND F_a VALUES ACHIEVED BY MAIN VARIANTS OF MTU-NET ON THE NUDT-SIRST-SEA DATASET

Model	IoU $(\times 10^{-2})$	P_d $(\times 10^{-2})$	F_a $(\times 10^{-6})$
MTU-Net	**64.14**	**85.44**	**11.72**
MTU-Net w/o level 1 ViT	60.21	82.59	12.58
MTU-Net w/o level 2 ViT	59.68	82.85	15.86
MTU-Net w/o level 3 ViT	53.29	81.96	28.93
MTU-Net w/o MVTM	52.94	78.38	43.76

TABLE V

IOU, P_d, AND F_a VALUES ACHIEVED BY MAIN VARIANTS OF DATA AUGMENTATION USED BY MTU-NET ON THE NUDT-SIRST-SEA DATASET. CP REPRESENTATIVES USING THE CP METHOD FOR DATA AUGMENTATION

Model	IoU $(\times 10^{-2})$	P_d $(\times 10^{-2})$	F_a $(\times 10^{-6})$
MTU-Net w/o DA	58.97	80.39	23.68
MTU-Net with CP DA	61.45	82.64	17.24
MTU-Net with CRRP DA	**64.14**	**85.44**	**11.72**

As shown in Table V, MTU-Net with CP-DA suffers decreases of 2.69% and 2.80% and an increase of 5.52×10^{-6} in terms of IoU, P_d, and F_a values over MTU-Net with CRRP-DA on our NUDT-SIRST-Sea dataset. MTU-Net w/o DA suffers decreases of 5.17% and 5.05% and an increase of 11.96×10^{-6} in terms of IoU, P_d, and F_a values over MTU-Net with CRRP-DA on our NUDT-SIRST-Sea dataset. Note that, F_a of the MTU-Net drops a lot using our CRRP data augmentation method. This is because there are a large number of highlighted complex backgrounds and suspicious targets. These highlighted complex backgrounds and suspicious targets occupy much more area than real targets in the space-based infrared image. Without data augmentation, MTU-Net causes more false alarms on highlighted complex backgrounds and suspicious targets. Our CRRP data augmentation method can preserve the long-range information and contextual information of targets. Thus, MTU-Net can better learn the long-range

TABLE VI

IoU, P_d, AND F_a VALUES ACHIEVED BY DIFFERENT LOSS FUNCTIONS USED WITH MTU-NET ON THE NUDT-SIRST-SEA DATASET

Loss Function	$IoU(\times10^{-2})$	$P_d(\times10^{-2})$	$F_a(\times10^{-6})$
Focal loss [43]	53.16	**86.18**	33.93
SoftIoU loss [44]	62.00	75.22	**9.42**
FocalIoU loss (ours)	**64.14**	85.44	11.72

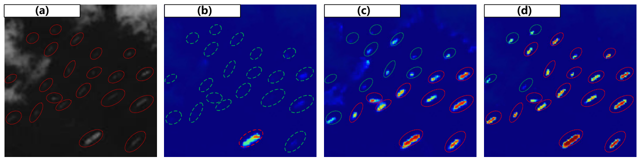


Fig. 9. FocalIoU loss analysis. (a) Input image. (b)–(d) Visualization maps of MTU-Net output in different IoU's. The output of MTU-Net is marked by a red solid frame. The correctly detected target, false alarm, and miss detection areas are highlighted by red, yellow, and green dotted circles, respectively. MTU-Net shifts from focusing on large-scale targets to focusing on small-scale targets as IoU rises from 0.2 to 0.6.

information of targets and achieve a better performance on IoU, P_d, and F_a.

3) FocalIoU Loss: The FocalIoU loss helps MTU-Net focus more on images with low IoU and reduces the weights of difficult samples relative to simple samples when IoU is small. FocalIoU loss achieves the "double-win" of target localization and shape description. To demonstrate the effectiveness of our FocalIoU loss, we retrained our MTU-Net using the SoftIoU loss and the focal loss for a fair comparison.

Visualization maps shown in Fig. 8 also demonstrate the effectiveness of our FocalIoU loss. The focal loss focuses on hard samples (e.g., small-scale targets and the edges of targets). However, the focal loss causes a higher response in the background area, resulting in more false alarm. The SoftIoU loss focuses on large-scale targets and loses small-scale targets because large-scale targets contribute much more in IoU than small-scale targets, resulting in miss detection of small-scale targets. The FocalIoU loss combines the advantages of both focal loss and SoftIoU loss, with a low response in the background area, and focuses on small-scale targets.

As shown in Table VI, MTU-Net with the focal loss suffers a decrease of 11.02% and an increase of 22.21×10^{-6} in terms of IoU and F_a values over MTU-Net with the FocalIoU loss. MTU-Net with the focal loss achieves an increase of 0.74% in P_d value. This is because the focal loss focuses on difficult positive samples (e.g., small-scale targets) but leads to a high F_a value and a low IoU value. MTU-Net with the SoftIoU loss suffers decreases of 2.14% and 10.22% in terms of IoU and P_d values over MTU-Net with the FocalIoU loss. MTU-Net with the SoftIoU loss achieves a decrease of 2.3×10^{-6} in F_a value. This is because the SoftIoU loss is calculated by the IoU of output, leading to more focus on large-scale targets. Numerous small-scale targets contribute less to IoU, resulting in higher IoU and smaller F_a but smaller P_d.

As shown in Fig. 9, MTU-Net shifts from focusing on large-scale targets to focusing on small-scale targets as IoU rises. The above results demonstrate that our FocalIoU loss can achieve the "double-win" of target localization and shape description.

VI. CONCLUSION

In this article, we propose the first and largest manually annotated dataset for space-based infrared tiny ship detection. Besides, we propose a novel pipeline for space-based infrared tiny ship detection, which contains MTU-Net, the CRRP data augmentation method, and FocalIoU loss. Specifically, a multilevel feature extraction module is designed to adaptively extract multilevel long-distance features in our MTU-Net. The CRRP data augmentation method is designed to alleviate the imbalance between target and background samples. The FocalIoU loss is proposed to achieve accurate target localization and shape description. Experimental results on the NUDT-SIRST-Sea dataset show that the proposed MTU-Net model outperforms traditional SIRST methods and existing deep learning-based SIRST methods in a set of evaluation metrics.

REFERENCES

[1] J. Tang, C. Deng, G. B. Huang, and B. Zhao, "Compressed-domain ship detection on spaceborne optical image using deep neural network and extreme learning machine," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 3, Mar. 2015.

[2] N. Wang, B. Li, X. Wei, Y. Wang, and H. Yan, "Ship detection in spaceborne infrared image based on lightweight CNN and multisource feature cascade decision," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, May 2021.

[3] M. Teutsch and W. Kruger, "Classification of small boats in infrared images for maritime surveillance," in *Proc. Int. WaterSide Secur. Conf.*, Nov. 2010, pp. 1–7.

[4] H. Deng, X. Sun, M. Liu, C. Ye, and X. Zhou, "Small infrared target detection based on weighted local difference measure," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 7, Jul. 2016.

[5] Y. Sun, J. Yang, and W. An, "Infrared dim and small target detection via multiple subspace learning and spatial–temporal patch-tensor model," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, May 2020.

[6] J.-F. Rivest and R. Fortin, "Detection of dim targets in digital infrared imagery by morphological image processing," *Opt. Eng.*, vol. 35, no. 7, pp. 1886–1893, Jul. 1996.

[7] S. D. Deshpande, M. H. Er, R. Venkateswarlu, and P. Chan, "Max-mean and max-median filters for detection of small targets," *Proc. SPIE*, vol. 3809, pp. 74–83, Oct. 1999.

[8] C. L. P. Chen, H. Li, Y. Wei, T. Xia, and Y. Y. Tang, "A local contrast method for small infrared target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 1, Jan. 2014.

[9] J. Han, Y. Ma, B. Zhou, F. Fan, K. Liang, and Y. Fang, "A robust infrared small target detection algorithm based on human visual system," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 12, pp. 2168–2172, Dec. 2014.

[10] J. Han, S. Moradi, I. Faramarzi, C. Liu, and Q. Zhao, "A local contrast method for infrared small-target detection utilizing a tri-layer window," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 10, pp. 1822–1826, Dec. 2019.

[11] J. Han et al., "Infrared small target detection based on the weighted strengthened local contrast measure," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 9, pp. 1670–1674, Sep. 2021.

[12] S. Kim and J. Lee, "Scale invariant small target detection by optimizing signal-to-clutter ratio in heterogeneous background for infrared search and track," *Pattern Recognit.*, vol. 45, no. 1, pp. 393–406, Jan. 2012.

[13] X. Wang, G. Lv, and L. Xu, "Infrared dim target detection based on visual attention," *Infr. Phys. Technol.*, vol. 55, no. 6, pp. 513–521, Nov. 2012.

[14] C. Q. Gao, D. Meng, Y. Yang, Y. Wang, X. Zhou, and A. G. Hauptmann, "Infrared patch-image model for small target detection in a single image," *IEEE Trans. Image Process.*, vol. 22, no. 12, pp. 4996–5009, Dec. 2013.

[15] L. Zhang, L. Peng, T. Zhang, S. Cao, and Z. Peng, "Infrared small target detection via non-convex rank approximation minimization joint $l_{2,1}$ norm," *Remote Sens.*, vol. 10, no. 11, p. 1821, Nov. 2018.

[16] Y. Dai and Y. Wu, "Reweighted infrared patch-tensor model with both nonlocal and local priors for single-frame small target detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 8, pp. 3752–3767, Aug. 2017.

[17] L. Zhang and Z. Peng, "Infrared small target detection based on partial sum of the tensor nuclear norm," *Remote Sens.*, vol. 11, no. 4, p. 382, Feb. 2019.

[18] H. Zhu, S. Liu, L. Deng, Y. Li, and F. Xiao, "Infrared small target detection via low-rank tensor completion with top-hat regularization," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 2, Oct. 2020.

[19] Y. Dai, Y. Wu, Y. Song, and J. Guo, "Non-negative infrared patch-image model: Robust target-background separation via partial sum minimization of singular values," *Infr. Phys. Technol.*, vol. 81, pp. 182–194, Mar. 2017.

[20] J. Shermeyer, T. Hossler, A. V. Etten, D. Hogan, R. Lewis, and D. Kim, "RarePlanes: Synthetic data takes flight," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 207–217.

[21] M. Liu, H. Y. Du, Y. J. Zhao, L. Q. Dong, and M. Hui, *Image Small Target Detection Based on Deep Learning With SNR Controlled Sample Generation*, 2018, pp. 211–220.

[22] B. McIntosh, S. Venkataramanan, and A. Mahalanobis, "Infrared target detection in cluttered environments by maximization of a target to clutter ratio (TCR) metric using a convolutional neural network," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 57, no. 1, pp. 485–496, Feb. 2021.

[23] Y. Dai, Y. Wu, F. Zhou, and K. Barnard, "Asymmetric contextual modulation for infrared small target detection," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 949–958.

[24] Y. Dai, Y. Wu, F. Zhou, and K. Barnard, "Attentional local contrast networks for infrared small target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 11, Nov. 2021.

[25] H. Wang, L. Zhou, and L. Wang, "Miss detection vs. false alarm: Adversarial learning for small object segmentation in infrared images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8508–8517.

[26] Y. Guo, M. Choi, K. Li, F. Boussaid, and M. Bennamoun, "Soft exemplar highlighting for cross-view image-based geo-localization," *IEEE Trans. Image Process.*, vol. 31, pp. 2094–2105, 2022.

[27] B. Li et al., "Dense nested attention network for infrared small target detection," *IEEE Trans. Image Process.*, early access, Aug. 22, 2022, doi: 10.1109/TIP.2022.3199107.

[28] M. Zhang, R. Zhang, Y. Yang, H. Bai, J. Zhang, and J. Guo, "ISNet: Shape matters for infrared small target detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 877–886.

[29] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021, pp. 437–446.

[30] 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," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022.

[31] J. Wu, Z. Pan, B. Lei, and Y. Hu, "LR-TSDet: Towards tiny ship detection in low-resolution remote sensing images," *Remote Sens.*, vol. 13, no. 19, p. 3890, Sep. 2021.

[32] H. Li, L. Deng, C. Yang, J. Liu, and Z. Gu, "Enhanced YOLO v3 tiny network for real-time ship detection from visual image," *IEEE Access*, vol. 9, pp. 16692–16706, 2021.

[33] C. Xia, S. Chen, X. Zhang, Z. Chen, and Z. Pan, "Infrared small target detection via dynamic image structure evolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022.

[34] H. Ye et al., "Contrastive triple extraction with generative transformer," in *Proc. AAAI Conf. Artif. Intell.*, 2021, vol. 35, no. 16, pp. 14257–14265.

[35] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds., 2020, pp. 213–229.

[36] J. Chen et al., "Transunet: Transformers make strong encoders for medical image segmentation," 2021, *arXiv:2102.04306*.

[37] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds., 2015, pp. 234–241.

[38] F. Liu, C. Gao, F. Chen, D. Meng, W. Zuo, and X. Gao, "Infrared small-dim target detection with transformer under complex backgrounds," 2021, *arXiv:2109.14379*.

[39] M. Qi et al., "FTC-Net: Fusion of transformer and CNN features for infrared small target detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 8613–8623, 2022.

[40] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[41] K. Wu, E. J. Otoo, and A. Shoshani, "Optimizing connected component labeling algorithms," *Proc. SPIE*, vol. 5747, pp. 1965–1976, Apr. 2005.

[42] G. Ghiasi et al., "Simple copy-paste is a strong data augmentation method for instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 2917–2927.

[43] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, Feb. 2020.

[44] Y. Huang, Z. Tang, D. Chen, K. Su, and C. Chen, "Batching soft IoU for training semantic segmentation networks," *IEEE Signal Process. Lett.*, vol. 27, pp. 66–70, 2020.

[45] C.-I. Chang, "An effective evaluation tool for hyperspectral target detection: 3D receiver operating characteristic curve analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 6, Jun. 2020.

[46] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *J. Mach. Learn. Res.*, vol. 12, no. 7, pp. 1–39, 2011.

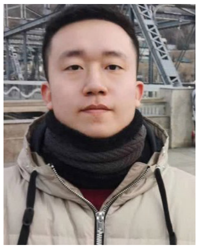
[47] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. Int. Conf. Artif. Intell. Statist. (ICAIS)*, 2010, pp. 249–256.

[48] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2019, pp. 8026–8037.

[49] F. I. Diakogiannis, F. Waldner, P. Caccetta, and C. Wu, "ResUNet—A: A deep learning framework for semantic segmentation of remotely sensed data," *ISPRS J. Photogramm. Remote Sens.*, vol. 162, pp. 94–114, Apr. 2020.

Tianhao Wu received the B.E. degree in electronic engineering from the National University of Defense Technology (NUDT), Changsha, China, in 2020, where he is currently pursuing the M.E. degree with the College of Electronic Science and Technology.

His research interests include infrared small target detection, light field imaging, and camera calibration.

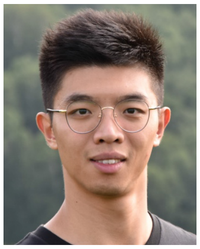


Boyang Li received the B.E. degree in mechanical design manufacture and automation from Tianjin University, Tianjin, China, in 2017, and the M.S. degree in biomedical engineering from the National Innovation Institute of Defense Technology, Academy of Military Sciences, Beijing, China, in 2020. He is currently pursuing the Ph.D. degree in information and communication engineering with the National University of Defense Technology (NUDT), Changsha, China.

His research interests include infrared small target detection, weakly supervised semantic segmentation, and deep learning.

Yihang Luo received the B.E. degree in communication engineering from Hunan Normal University, Changsha, China, in 2020. She is currently pursuing the M.E. degree with the College of Electronic Science and Technology, National University of Defense Technology (NUDT), Changsha.

Her research interests include infrared image denoising and infrared small target detection.



Yingqian Wang received the B.E. degree in electrical engineering from Shandong University (SDU), Jinan, China, in 2016, and the M.E. degree in information and communication engineering from the National University of Defense Technology (NUDT), Changsha, China, in 2018, where he is currently pursuing the Ph.D. degree with the College of Electronic Science and Technology.

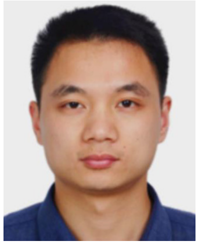
His research interests focus on low-level vision, particularly on light field imaging and image super-resolution.

Chao Xiao received the B.E. degree in communication engineering and the M.E. degree in information and communication engineering from the National University of Defense Technology (NUDT), Changsha, China, in 2016 and 2018, respectively, where he is currently pursuing the Ph.D. degree with the College of Electronic Science.

His research interests include deep learning, small object detection, and multiple object tracking.

Ting Liu received the B.E. degree in electrical engineering and automation from the Hunan Institute of Engineering, Xiangtan, China, in 2017, and the M.E. degree in control engineering from Xiangtan University (XTU), Xiangtan, in 2020. She is currently pursuing the Ph.D. degree with the College of Electronic Science, National University of Defense Technology (NUDT), Changsha, China.

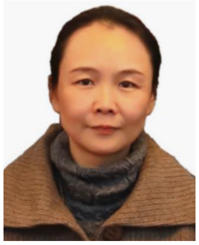
Her research interests focus on signal processing, target detection, and image processing.



Jungang Yang received the B.E. and Ph.D. degrees from the National University of Defense Technology (NUDT), Changsha, China, in 2007 and 2013, respectively.

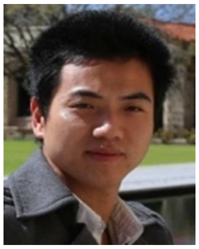
He was a Visiting Ph.D. Student with The University of Edinburgh, Edinburgh, U.K., from 2011 to 2012. He is currently an Associate Professor with the College of Electronic Science, NUDT. His research interests include computational imaging, image processing, compressive sensing, and sparse representation.

Dr. Yang received the New Scholar Award of Chinese Ministry of Education in 2012 and the Youth Innovation Award and the Youth Outstanding Talent of NUDT in 2016.



Wei An received the Ph.D. degree from the National University of Defense Technology (NUDT), Changsha, China, in 1999.

She was a Senior Visiting Scholar with the University of Southampton, Southampton, U.K., in 2016. She is currently a Professor with the College of Electronic Science and Technology, NUDT. She has authored or coauthored over 100 journal and conference publications. Her research interests include signal processing and image processing.



Yulan Guo (Senior Member, IEEE) received the B.E. and Ph.D. degrees from the National University of Defense Technology (NUDT), Changsha, China, in 2008 and 2015, respectively.

He has authored over 100 papers at highly referred journals and conferences. His research interests focus on 3-D vision, particularly on 3-D feature learning, 3-D modeling, 3-D object recognition, and scene understanding.

Dr. Guo is a Senior Member of ACM. He served as the Area Chair for CVPR 2021, ICCV 2021, and ACM Multimedia 2021. He also served as an Associate Editor for IEEE TRANSACTIONS ON IMAGE PROCESSING, *IET Computer Vision*, *IET Image Processing*, and *Computers & Graphics*. He organized several tutorials, workshops, and challenges in prestigious conferences, such as CVPR 2016, CVPR 2019, ICCV 2021, 3DV 2021, CVPR 2022, ICPR 2022, and ECCV 2022.