

## RESEARCH ARTICLE

# CA-U2-Net: Contour Detection and Attention in U2-Net for Infrared Dim and Small Target Detection

LEIHONG ZHANG<sup>1</sup>, WEIHONG LIN<sup>1</sup><sup>2</sup>, ZIMIN SHEN<sup>2</sup>, DAWEI ZHANG<sup>1</sup>, BANGLIAN XU<sup>1</sup>,  
KAIMIN WANG<sup>1</sup>, AND JIAN CHEN<sup>1</sup><sup>3</sup>, (Senior Member, IEEE)

<sup>1</sup>School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China

<sup>2</sup>College of Communication and Art Design, University of Shanghai for Science and Technology, Shanghai 200093, China

<sup>3</sup>Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun 130033, China

Corresponding author: Weihong Lin (422170545@qq.com)


This work was supported in part by the National Natural Science Foundation of China under Grant 62275153 and Grant 62005165, in part by the Development Fund for Shanghai Talents under Grant 2021005, and in part by the Shanghai Industrial Collaborative Innovation Project under Grant HCXBCY-2022-006.

**ABSTRACT** With the development of infrared technology, infrared dim and small target detection plays an essential role in precision guidance and early warning systems. Due to the low contrast and signal-to-noise ratio that characterizes infrared dim and small target in images, the dim and small target can easily be drowned out by noise and background. A new infrared dim and small target detection network (CA-U2-Net) is proposed to address the challenge of infrared weak target detection and shape retention in complex backgrounds. Specifically, firstly, the U2-Net network structure has been improved to prevent the loss of shallow information due to increased network depth and to make it more suitable for detecting the dim and small target. Then, the upper and lower attention module was designed on the network to make the model more focused on dim and small target features while suppressing irrelevant information, further improving the detection rate. Finally, a contour detection branch was added to the top of the model to fuse the contour detection map with the feature map to get a better target shape. After experimental evaluation, the method achieved a detection rate of 97.17% and retained a more accurate infrared dim and small target shape. Compared to other advanced methods, our method performs better in detection rate, false detection rate and shape retention. In addition, a new infrared dim and small target dataset consisting of 10,000 images was constructed.

**INDEX TERMS** Attention mechanism, contour detection, infrared dim and small target detection.

## I. INTRODUCTION

The detection of infrared dim and small targets is widely used in fields such as surveillance [1], target warning [2], precise guidance [3], and forest fire prevention [4]. The image of an infrared dim and small target is made up of three parts: the weak target, the background and the noise. According to the definition of SPIE [5], the pixel size of a dim and small target is usually considered to be no larger than 81 (9 × 9), which is approximately 0.12% of an image with a

The associate editor coordinating the review of this manuscript and approving it for publication was Larbi Boubchir .

pixel size of 256 × 256. Backgrounds for infrared dim and small target include sky, buildings and sea. These complex backgrounds often have high contrast and greatly interfere with detection. Images taken with infrared detectors are also subject to some interference from noise, such as clutter and heat sources. Therefore, there are significant challenges in detecting dim and small target with low signal-to-noise ratios in complex scenes. Recently, more and more infrared dim and small target detection methods based on traditional methods and deep learning methods have been proposed.

Traditional methods for detecting infrared dim and small targets can be divided into three categories: filter-based

methods, human eye attention mechanism-based methods and low-rank sparse recovery-based methods. Filter-based methods focus on suppressing the background through differences in pixel-grey values. The most common methods are the Top-hat algorithm [6] and the Bilateral Filter (BF) [7]. Filter-based methods are capable of removing point noise and background interference that are not of the required size. However, it is too dependent on the relationship between each pixel point and its gray level difference, which is easy to cause false alarms. The method based on human eye attention mechanism mainly detects targets through significant features such as contrast, size, and shape. The most common methods include Relative Local Contrast Measure (RLCM) [8], Local Contrast Method (LCM) [9], Multi-scale Patch based Contrast Measure (MPCM) [10], Improved Local Contrast Method (ILCM) [11], and Weighted Local Difference Measure (WLDM) [12]. The method based on the human eye attention mechanism utilizes the presence of a salient region (contrast, size and shape) of the target in an infrared image for detection. However, when the image contrast is very low, it is easy to cause false alarms. The method based on low-rank sparse recovery mainly utilizes frequency feature differences for target detection. In recent years, methods proposed include the Infrared Patch Image Model (IPI) [13], Weighted Infrared Patch Image (WIPI) [14], Reweighted Infrared Patch-tensor Model (RIPI) [15], Non-Convex Rank Approximation Minimization Joint  $l_{2,1}$  Norm (NRAM) [16], Non-Convex Optimisation with  $L_p$ -Norm Constraint (NLOC) [17] and Self-Regularised Weighted Sparse Model (SRWS) [18]. The method based on low-rank sparse recovery can effectively decompose small targets and uniform backgrounds, but it is prone to generate false alarms for backgrounds with strong clutter edges, corners, and white spots. As traditional methods cannot extract deeper feature information from the target and are susceptible to clutter and noise, many methods have low detection rates and high false detection rates in scenes with complex backgrounds.

The infrared small target detection method based on deep learning uses Convolutional Neural Network (CNN) to achieve feature extraction, allowing for deeper semantic information to be obtained from the image. Therefore, deep learning-based methods are more robust than traditional methods. Based on CNN, Wang et al. [19] proposed MD vs. FA-cGAN, which utilizes generative adversarial networks to balance the missed detection rate and false alarm rate in image segmentation. The network achieves a balance between missed detection rate and false alarm rate through adversarial training of three sub-networks and improves the detection accuracy of small infrared targets. Dai et al. [20] proposed a segmentation-based network where an asymmetric background modulation module was designed to aggregate shallow and deep features. Subsequently, Dai et al. [21] further improved their network by expanding local contrast and designed a feature loop transformation scheme to achieve trainable local contrast measurement. Li et al. [22]

proposed a densely nested attention network (DNANET) with a densely nested interaction module and a cascaded channel and spatial attention module designed to implement the interaction between high-level and low-level features as well as adaptive enhancement of multi-level features, respectively. Wang et al. [23] proposed a coarse-to-fine internal attention-aware network (IAANET) that uses semantic contextual information of all pixels within a local region to classify each internal pixel. Deep learning-based algorithms for infrared dim and small image detection are less compared to traditional methods. Due to the insufficient number of datasets, which limits the training of deep models to a certain extent. Deep learning-based detection methods for infrared dim and small targets need to further improve detection performance.

Although many infrared small target detection methods have been proposed, these methods still have problems, such as poor robustness and insufficient detection performance. Therefore, accurately detecting infrared dim and small targets in complex backgrounds is still an urgent problem that needs further research. Inspired by U2-Net [24], this paper proposes the CA-U2-Net model using a U-shaped structure to interact with the left and right parts of the feature map information. Specifically, to make U2-Net more focused on infrared dim and small target, the top two coding and decoding layers are reduced in this model to prevent the loss of shallow features by reducing the number of down-sampling layers. To further improve detection rates, this model designed upper and lower attention modules to make the model more focused on small target features and enhance shallow semantic information, while suppressing irrelevant information. In order to improve the contour detection effect, a contour detection module is added to the top of the model, which combines contour detection with the output feature map to further optimize the edge information of infrared dim small targets. In summary, the contribution of this paper can be summarized as follows:

- To the best of our knowledge, we are the first to propose the use of the U2-Net for infrared dim and small target detection. By improving the network structure, it has been made more suitable for the detection of the infrared dim and small target. Detection rate improved from 82.86% to 85.87% and false detection rate reduced from 86.14% to 66.29%.
- We have designed the upper and lower attention module to effectively improve the detection rate of small targets and suppress irrelevant information. The detection rate was improved from 85.87% to 95.85% and the false detection rate was reduced from 66.29% to 25.42%.
- We have added a contour detection branch to fuse the contour detection map with the feature map of the target detection output, making edge shape retention significantly better. Compared with current methods, our method has significant advantages in detection rate, false detection rate and shape retention.

- We constructed a new infrared dim and small target dataset (CA-U2-Net IRST), which compensates for the lack of infrared dim and small target dataset and can improve the detection rate after training with a different model. The detection rate of our model improved from 95.85% to 97.17% and the false detection rate decreased from 25.42% to 23.35%.

## II. RELATED WORK

### A. EQUATIONS INFRARED DIM AND SMALL TARGET DETECTION

The target detection method based on deep learning mainly uses CNN to achieve feature extraction and obtain deeper semantic information of the image. Common methods include the two-stage target detection algorithm represented by R-CNN [25] and its improvement algorithms [26], [27], [28], which uses a candidate region-based method that requires candidate frames to be generated first, followed by classification and regression on the candidate frames. Although the use of a multi-level network for classification and localization gives the model greater classification and localization capabilities. However, too many sub-networks for classification and localization can also reduce the detection rate of the model. The other category is the single-stage target detection algorithm represented by YOLO [29] and its improvements [30], [31], [32]. The single-stage algorithm converts the problem of localization into a problem of regression without pre-generating candidate frames. However, it gives the probability and location coordinate values of the class to which the target belongs directly by calculating the convolutional layer. This type of algorithm only requires one detection to obtain the final detection result, which is relatively fast but has low accuracy.

Different from the above work, since infrared dim and small target detection only needs to distinguish foreground and background, we introduce the U2-Net network in semantic segmentation, and U2-Net can distinguish foreground and background well. By improving the U2-Net model, we design a CA-U2-Net network to solve the problem of infrared small target detection accuracy and shape, and improve the infrared dim and small target detection performance.

### B. ATTENTION MECHANISM

The attention mechanism was first proposed by Bahdanau et al. [33] in 2014 to solve the difficulty of accurately sequentially encoding long sentences in text to a fixed length. Attention Mechanism is widely used in image detection, speech recognition, natural language processing and other fields. In computer vision, by adding attention mechanisms, different parts of an image or feature map can be weighted to varying degrees, causing the neural network to pay different attention to different regions of the feature map, thereby better focusing the network on areas of interest.

In recent years, attention mechanisms have been introduced in infrared dim and small targets to improve model

detection performance. Dai et al. [20] used point-by-point attention modulation to retain and highlight the details of the infrared dim and small target. Tong et al. [34] proposed an enhanced asymmetric attention block that uses same-level feature information exchange and cross-level feature fusion to improve feature representation. Zhang et al. [35] designed a bidirectional attention aggregation block to compute low-level information and fuse it with high-level information to capture the shape features of the target and suppress noise.

Different from the above work, we design the upper and lower attention module to use the feature map of the coding layer with the corresponding decoding layer of the following layer after upper and lower attention modulation as the input to the decoding layer. The irrelevant regions in the input image are suppressed, while the salient feature of the infrared dim and small target is highlighted. Integrating the upper and lower attention modules into the U2-Net network structure not only has a low computational cost, but also improves the model's target detection and background suppression capabilities.

### C. CONTOUR DETECTION

Contour detection can be used in the fields of image segmentation [36], [37], target detection [38], [25], and semantic segmentation [39], [40], [41]. Early contour detection used contour edge detection operator (Canny) [42] to extract object contours. Since the rise of deep learning, various algorithms based on CNN have been proposed. The overall nested network HED proposed by Xie et al. [43] provided an idea for the design of edge detection networks, and subsequently algorithms such as Richer Convolutional Features (RCF) [44], Pixel Difference Networks (PidiNet) [45], Dense Extreme Inception Network (DexiNed) [46], and Deep Refinement Network (DRNet) [47] were proposed.

Contour is an external feature of the image target, and this feature is very important for our image analysis and target recognition. Especially in the field of infrared dim and small target detection, if the target edge shape can be accurately detected, the shape and curve can further identify the class of objects, which is more meaningful in practical applications. In recent years, PoolNet [48] and ISNet [35] have enhanced the edge detection effect by branching joint training. Although adding edge constraints during training can improve edge detection, some non-target edges will be retained, which affects the detection accuracy and also increases the computational effort. Unlike the above work, the model proposed in this paper already performs well in infrared dim and small target detection, but the edges are not clear enough. To address this problem, we add a contour detection fusion method to the top of the model. The infrared dim and small target picture is fused with the feature map after contour detection. This improves the detection shape effect and also reduces the computational effort than constrained edge detection.

#### D. DATASET

Infrared image datasets have long lacked public datasets, and due to the problem of high shooting cost, the existing infrared dim and small target datasets are MFIRST [19] and SIRST [20]. MFIRST contains 10,000 images, most of which are close-up shots with close and large target areas. A part of them are synthetic images, and many of them do not meet the definition of SPIE, which will affect the effect of training. SIRST contains 427 images with various scene types, but the number is too small and only suitable for detection not for training.

The publicly available datasets mentioned above have promoted the development of infrared dim and small target detection. However, many of them do not meet the definition and are costly and error-prone to manual labeling. Therefore, the problem of the insufficient training set of infrared dim and small target detection algorithm based on deep learning is solved by synthesizing the data set through the infrared simulation system. This paper uses an infrared simulation system to generate 10,000 infrared dim and small target images by simulating scene switching and small target motion. Our dataset is evaluated in Section IV.

### III. METHOD

In this section, we first review the U2-Net network and introduce the proposed new network in detail. Then details of the upper and lower attention block and contour detection block are presented, as well as the CA-U2-Net IRST dataset.

#### A. NETWORK STRUCTURE

##### 1) REVIEW U2-NET

The U2-Net network is mainly used for saliency target detection, which consists of residual U-blocks (RSU) that extract multi-scale features within a stage and an outer U-shaped structure that connects the RSU. This network design does not require the use of a backbone network for image classification and can be trained from scratch to achieve excellent results. The network can be made deeper to obtain high-resolution feature maps without increasing the computational cost as much as possible.

However, the structure of U2-Net network is deep and complex, and it can extract intra-level and inter-level information of different sizes of images through RSU and jump connections, but the retention of invalid features such as non-defective regions and the loss of edge information is easy to occur in the connections. Directly applying the U2-Net network for infrared dim and small target detection can detect most targets, but there are still some missed detections and a high false detection rate.

##### 2) OVERALL STRUCTURE OF CA-U2-NET

Inspired by the U2-Net network, this paper uses a U-shaped structure to realize the interaction of left and right partial

feature map information, and proposes the CA-U2-Net model. To make U2-Net focus more on the infrared dim and small target, we reduce the top two coding and decoding layers to prevent losing features in shallow layers by too much down-sampling. The improved model decreases much in size and improves the accuracy by 3.01% over U2-Net. To further improve the detection rate, we introduced the upper and lower attention module (ULA) to make the model more focused on small target features to reduce information loss and enhance shallow semantic information while suppressing irrelevant information. The detection accuracy of the model after increasing attention is as high as 95.85%, and the false detection rate is reduced to 25.42%. We design the contour detection module on top of the model to improve the contour detection effect. The fusion of contour detection with the feature map further optimizes the contour information of the infrared dim and small target. The overall structure of the network is shown in Figure 1, and the RSU module is shown in Figure 2.

As shown in Figure 1, CA-U2-Net is an encoding-decoding based infrared dim and small target detection network, which mainly consists of a U-shaped structure made of stacked RSUs of feature extraction structure, upper and lower attention module and edge detection branches. RSU mixes the feature maps of different scales and different receptive fields through the U shape structure, and can obtain more global information on different scales. En1, En2, De1, and De2 use the same RSU. As shown in Figure 2 (a), 5 represents the depth of block, and similarly the block of RSU4 is 4. For each feature map size in the process of down-sampling, an atrous convolution with a dilation rate of 1 and a convolution kernel size of  $3 \times 3$  is used to achieve the purpose of expanding the perceptual field, so as to achieve the function of extracting the feature information of the context and the neighborhood. En3, En4, and De3 use RSU4F, as shown in Figure 2(b). The structure of RSU4F and RSU4 are not the same, in RSU4F there is no down-sampling or up-sampling, but the atrous convolution replaces all the sampling layers with different dilation rates, which can extract multi-scale features without reducing the resolution of the feature map. Since the infrared dim and small targets are small, the resolution of the feature map is already very low when down-sampling to En\_2. If the down-sampling continues, smaller targets are easily lost in deeper nets. Therefore, down-sampling is no longer performed in RSU4F.

Multi-scale features are extracted from the progressively down-sampled feature maps and decoded into high-resolution feature maps by progressive up-sampling, Concatenation, convolution and attention mechanisms. This process mitigates the detail loss caused by the direct up-sampling of small-scale feature maps. The feature maps  $f_1$ ,  $f_2$ ,  $f_3$  and  $f_4$  output from De\_1, De\_2, De\_3, and En\_4 are collected, and then the feature maps with channel one are obtained by a  $3 \times 3$  convolution layer respectively. Then, we obtain F\_1, F\_2, F\_3, and F\_4 by bilinear interpolation scaling to the

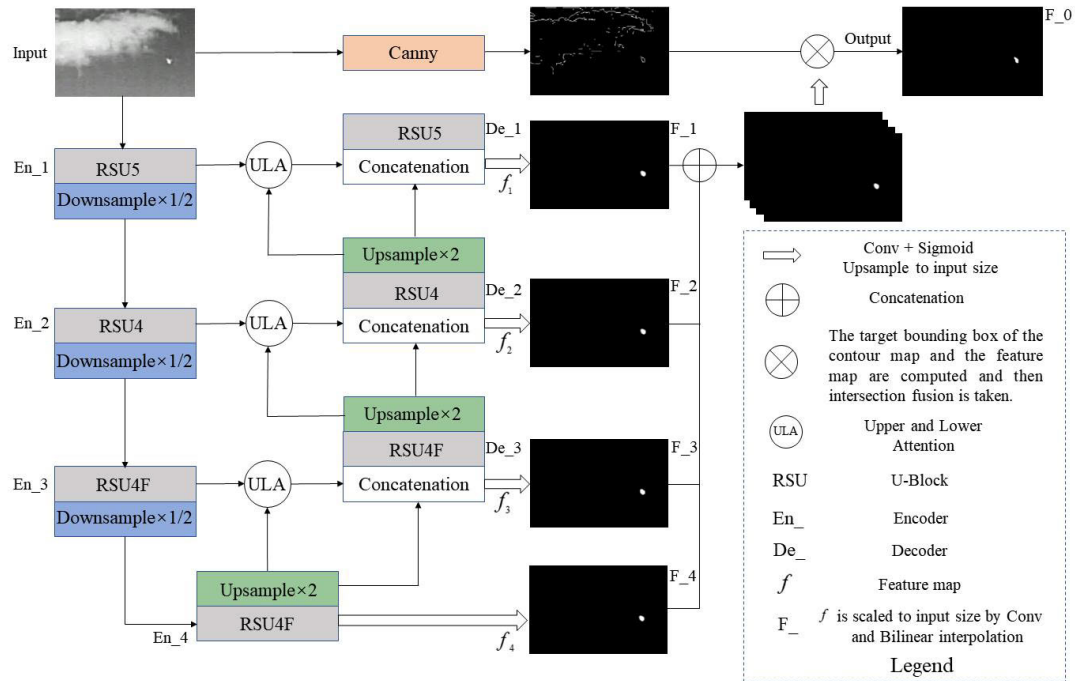


FIGURE 1. The overall structure of CA-U2-Net.

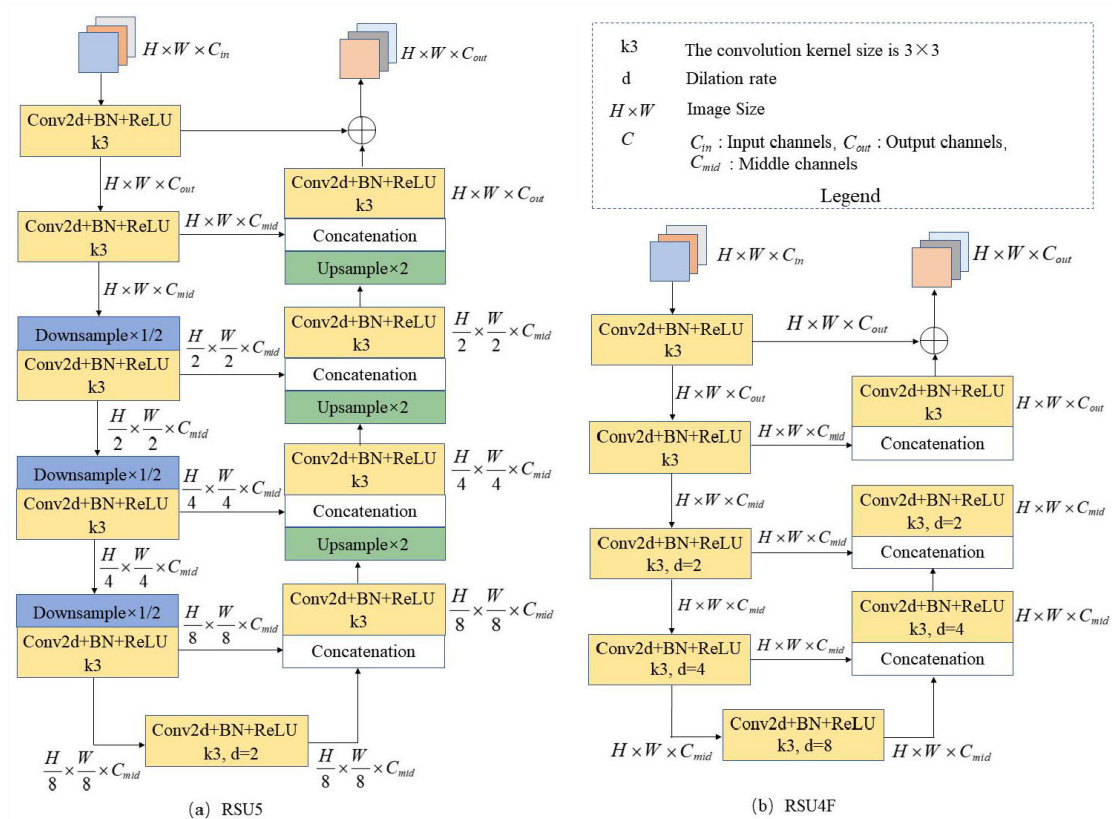


FIGURE 2. RSU module. (a) RSU5; (b) RSU4F.

input image size, and concatenate these four feature maps. Finally, the contour detection branch and feature fusion map are passed through a  $1 \times 1$  convolutional layer and a Sigmoid activation function to obtain the final predicted probability map  $F_0$ .

The network is trained by performing a Binary Cross-Entropy (BCE) calculation on each of the four outputs and ground truth, and then back-propagating the loss summation with the following equation:

$$\text{Loss} = \sum_{n=1}^N w_f^{(n)} l_f^{(n)} + w_F l_F \quad (1)$$

In Equation (1),  $N = 4$ ,  $w$  represents the weight of each loss, and  $l$  represents the binary cross-entropy loss function.  $w_f^{(n)} l_f^{(n)}$  is the four feature map outputs and the loss of ground true, and  $w_F l_F$  is the final fused image and the loss of ground truth.

### B. UPPER AND LOWER ATTENTION BLOCK

In order to obtain more feature information on small targets, in addition to adjusting the down-sampling strategy of the network, an upper and lower attention block is added. U2-Net originally concatenates the results of encoder and decoder up-sampling through jump connections, and using only simple jump connections to simulate the global multi-scale context easily leads to the loss of spatial information and thus the situation of missed detection. To address this problem, this paper designs an upper and lower attention block, which uses the feature map of the coding layer and the corresponding decoding layer of the following layer as the input of the decoding layer after upper and lower attention modulation. It can suppress the part of model learning which is irrelevant to the infrared dim and small target, and at the same time aggravate the feature learning with the infrared dim and small target. The internal of the ULA module is shown in Figure 3. The result of the  $1 \times 1 \times 1$  convolution of the feature map  $g$  of the same layer of the down-sampling layer is summed with the feature map of the previous layer of the up-sampling layer, and then passed through the ReLU activation function. Then the  $1 \times 1 \times 1$  convolution operation is used, and finally the attention coefficients ( $\alpha$ ) are obtained by Sigmoid.

### C. CONTOUR DETECTION BLOCK

RSU block enriches feature information by mixing receptive fields of different sizes and attention mechanisms to obtain multi-scale features. However, the deepening of network layers and pooling operations tend to ignore details such as edges and corners in the images, resulting in poor segmentation near edges. Therefore, this paper uses the Canny algorithm as a contour detection branch module. The Canny algorithm performs contour detection before using Gaussian smoothing filtering to suppress noise effectively. Then the gradient vector of each point in the image is calculated, and the gradient direction and gradient amplitude can be obtained based on the

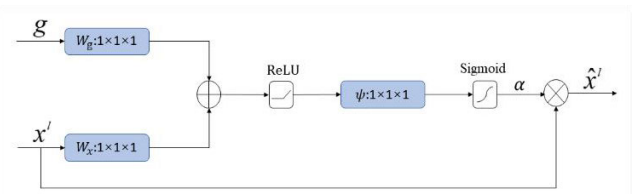


FIGURE 3. Upper and lower attention block.

gradient vector, and the places with large gradient changes are identified as contours. We traverse the results of contour detection and feature fusion separately to calculate the possible target bounding box  $B_1$  of the contour map and the target bounding box  $B_2$  of the fusion map. Exclude non-target bounding box  $B_1$  based on  $B_2$ , and then fill in the contours within  $B_1$ . If the number of the target bounding boxes  $B_1$  of the contour map is less than the target bounding box  $B_2$  of the fusion map, then the target within  $B_2$  is fused to the contour map to get the result. The process is shown in Figure 4.

### D. SYNTHETIC DATASET

By analyzing the MFIRST and SIRST datasets, we found that part of the small target is too large, which easily affects the feature learning. Most of them are single target, which is easy to be missed in the multi-target detection set. Therefore, we select the small target size of  $3 \times 3.9 \times 9$  as the small target size for the dataset construction of the subsequent images to meet the definition of small target. Table 1 shows the images of the dataset and ground truth.

In the infrared simulation system, we can switch the background, change the target size and motion trajectory, and also add noise. When we generate the infrared weak target image, we change the background to black and generate the ground truth image. Using infrared simulation system to synthesize the dataset not only reduces the cost and time, but also improves the accuracy of the dataset.

## IV. EXPERIMENTS

In this section, we introduce the Implementation Details, Evaluation Metrics, and Experimental Settings. In the Experimental Settings, we first introduce the training and test datasets, then verify the effect of the improved scheme step by step through ablation experiments and analyze the results. Finally, the CA-U2-Net network model is compared with other methods and analyzed.

### A. IMPLEMENTATION DETAILS

The experiments involved in this paper are trained and validated under Intel Core i9-10920X CPU @ 3.50 GHz, NVIDIA GeForce RTX 3090 environments. The experimental software is PyCharm 3.3, MATLAB 2020 and Unity3D 2020. the CA-U2-Net network model uses the PyTorch framework based on the Python language. The model is trained from scratch, and the whole training process is

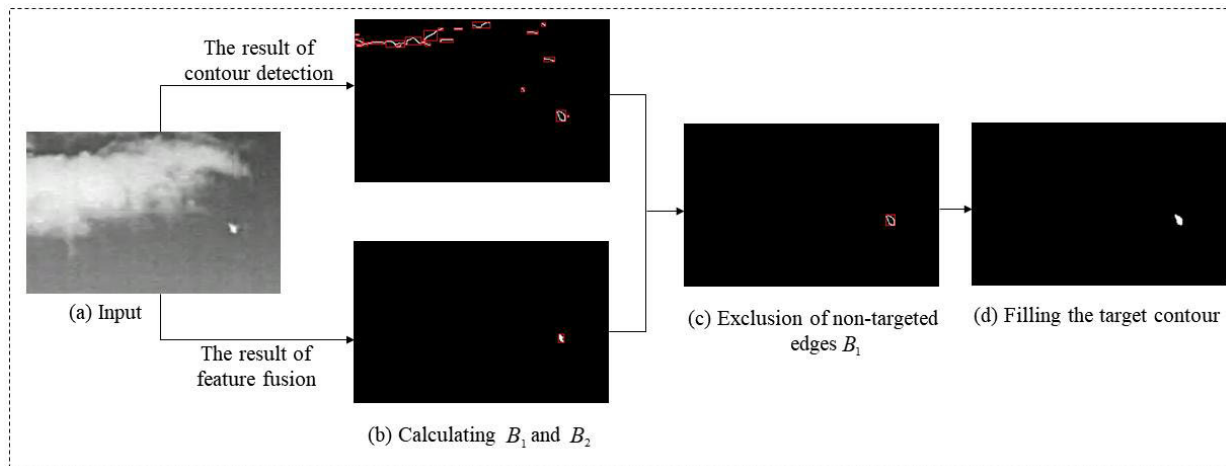


FIGURE 4. Contour detection block.

TABLE 1. Synthetic dataset.

	Near-infrared	Adding noise in the near-infrared	Ground truth
Single target			
Multiple targets			

100 epochs with 20,000 datasets, and the training consists of one batch of every 10 images.

**B. EXPERIMENTAL EVALUATION**

The most effective evaluation metrics for the infrared dim and small target are detection rate and false detection rate, and the detection rate  $P_d$  is defined as the ratio of the number of real detected targets to the number of actual targets. The higher the detection rate, the better the detection performance of the algorithm. The false alarm rate  $F_a$  is defined as the ratio of the number of false detected targets to the number of actual targets. The lower the false alarm rate, the better the detection

performance of the algorithm.

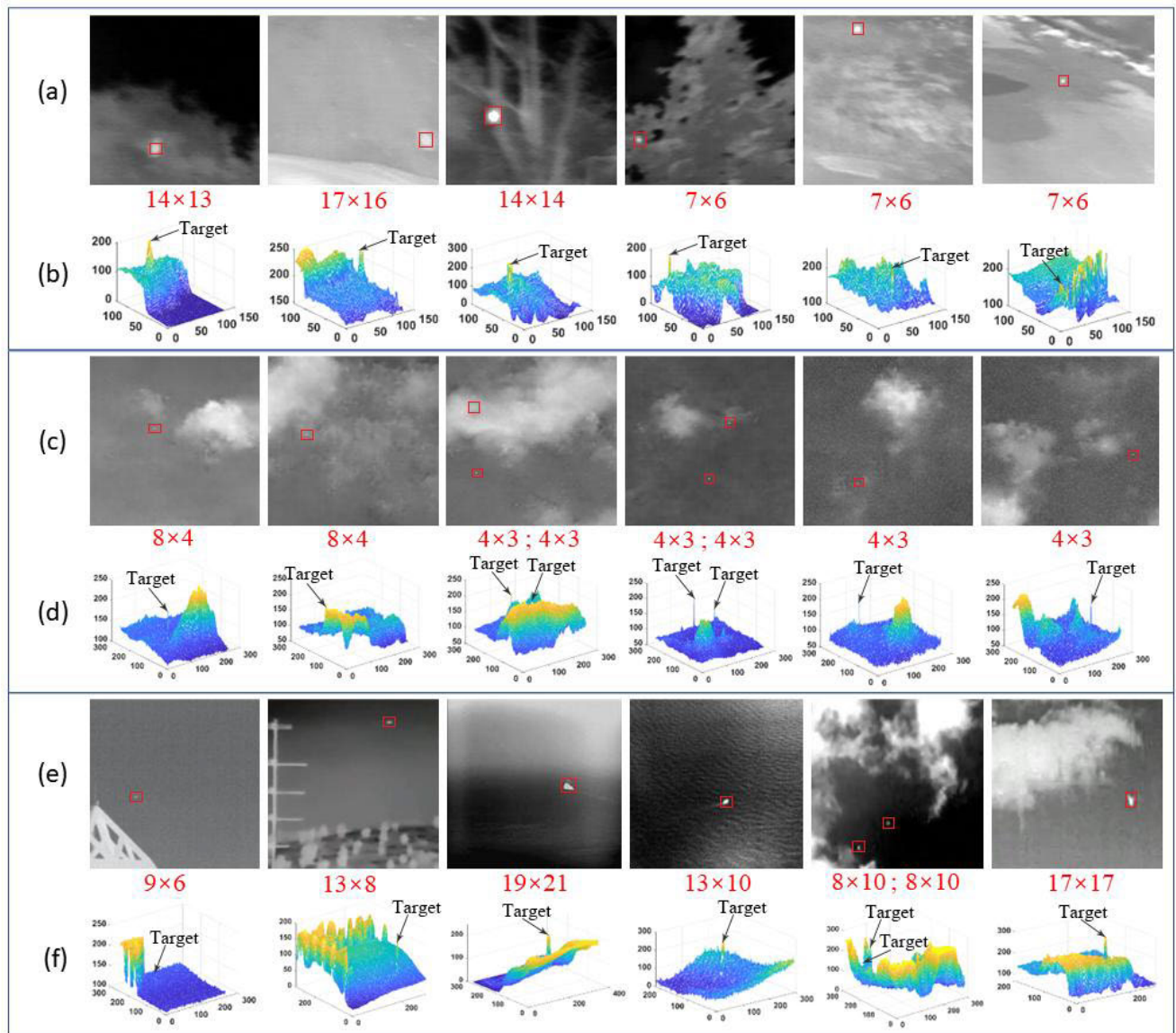
$$P_d = \frac{\text{Number of real targets detected}}{\text{Actual targets number}} \tag{2}$$

$$F_a = \frac{\text{Number of false targets detected}}{\text{Actual targets number}} \tag{3}$$

**C. EXPERIMENTAL SETTINGS**

1) INTRODUCTION TO TRAINING AND TESTING SETS

The training sets used in this paper are MFIRST and CA-U2-Net datasets. The MFIRST dataset consists of 10,000 infrared images with target sizes ranging from  $6 \times 6$  to  $20 \times 20$  pixels. The CA-U2-Net IRST dataset consists of



**FIGURE 5.** Three datasets. (a) is the MFIRST dataset. (b) is the 3-D plot of the MFIRST dataset. (c) is the CA-U2-Net dataset. (d) is the 3-D plot of the CA-U2-Net dataset. (e) is the SIRST dataset. (f) is the 3-D plot of the SIRST dataset.

**TABLE 2.** Detection rate and false detection rate of the model after training on different datasets.

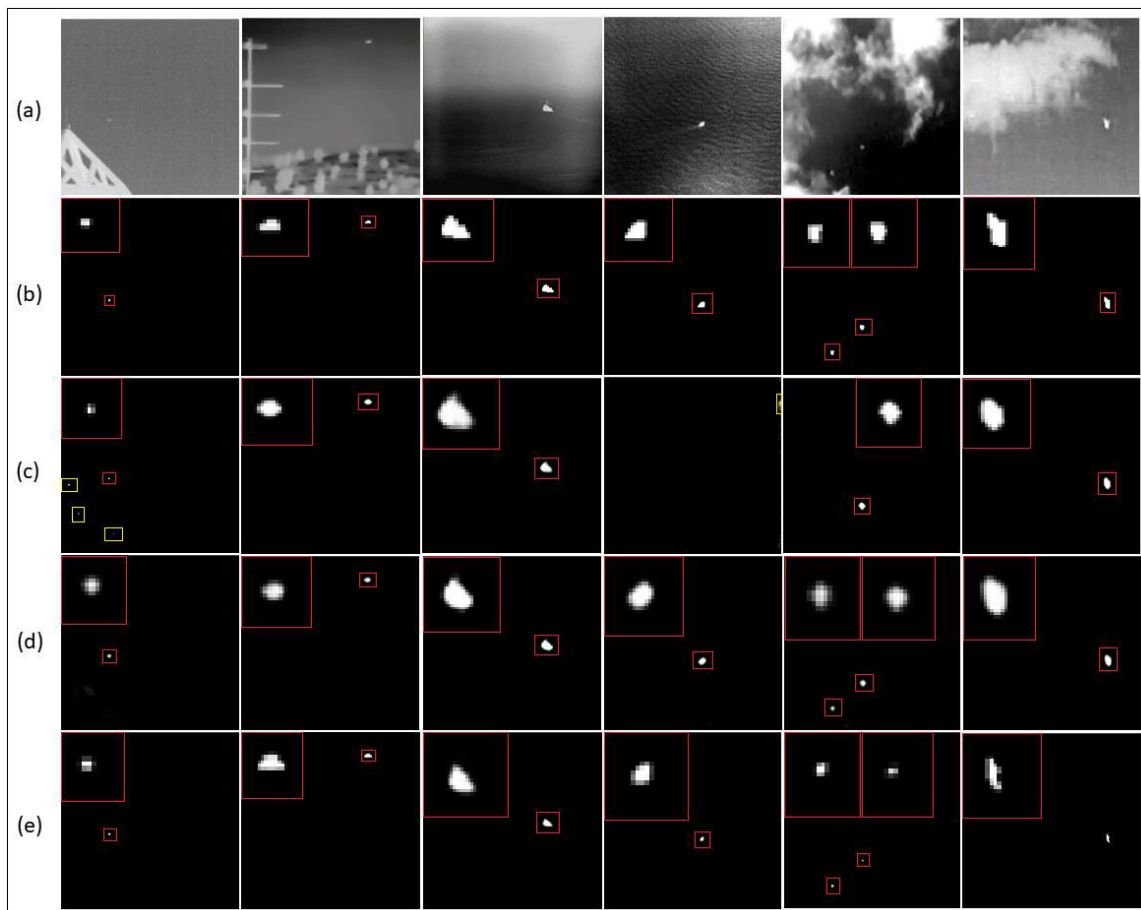
Training	MFIRST [19]		CA-U2-Net IRST		MFIRST+CA-U2-Net IRST	
	$P_d$	$F_a$	$P_d$	$F_a$	$P_d$	$F_a$
U2-Net	0.8286	0.8614	0.8380	0.8248	0.8587	0.8097
IAANET	0.9491	0.8474	0.9566	0.8267	0.9566	0.8116
DNANET	0.8531	0.6233	0.8700	0.5932	0.8832	0.5875
ACM	0.8474	0.3107	0.8625	0.2901	0.8662	0.2901
Ours	<b>0.9585</b>	<b>0.2542</b>	<b>0.9661</b>	<b>0.2335</b>	<b>0.9717</b>	<b>0.2335</b>

10,000 infrared images, composited with the sky as the background image, with target sizes ranging from  $3 \times 3$  to  $9 \times 9$  pixels.

The SIRST dataset includes a total of 427 infrared images and 531 targets. The backgrounds of the targets are complex and varied, including clouds, sea surface and buildings. The

target scales are diverse, with targets ranging in size from  $5 \times 5$  to  $20 \times 20$  pixels. There are both brighter targets and darker targets that are very close to the background brightness. For the experiments, 427 images were used for testing. Figure 5 shows the infrared images from the three datasets, using red boxes to mark the areas of the individual targets in the infrared





**FIGURE 6.** Ablation studies. (a) Images of SIRST; (b) Ground truth; (c) Detection results of the improved U2-Net model; (d) Detection results of the improved U2-Net after adding upper and lower attention; (e) Results of adding contour detection.

**TABLE 3.** Units for magnetic properties contribution of each improvement module in The Ca-u2-net model.

Method	$P_d$	$F_a$	Weight
U2-Net (Baseline)	0.8587	0.8097	176.3MB
Improved U2-Net	0.8719	0.6629	<b>112.4MB</b>
Improved U2-Net+Attention Mechanism	<b>0.9717</b>	<b>0.2335</b>	120.6MB
Improved U2-Net+Attention Mechanism+ Contour Detection	<b>0.9717</b>	<b>0.2335</b>	121.2MB

images for better visibility. The red font indicates the target size.

### 2) THE IMPACT OF CA-U2-NET DATASET ON PRECISION

In this paper, we apply the synthesized infrared dim and small target dataset to the infrared dim and small target detection method to verify the effectiveness of the synthesized dataset. Firstly, we trained the U2-Net, DNANET, ACM [2] and IAANET models were trained separately for the MFIRST dataset and this paper dataset, and then the models are trained together for the MFIRST dataset and this paper dataset. Finally, the target detection rate  $d$  and false detection rate  $a$  were tested on all datasets of SIRST, and the results are shown in Table 2.

From Table 2, it can be seen that different datasets train the model to obtain different accuracies, which indicates that the dataset has an impact on the detection model training. The accuracy of training with the MFIRST dataset is close to that of this paper, which indicates its effectiveness and feasibility, and it has a certain improvement for model detection accuracy. Combining the two datasets to train together can yield better results.

### 3) ABLATION STUDY

In this section, we study the contribution of each improvement module of the CA-U2-Net model. Training and testing were performed for each part of the ablation study using the same parameter settings. The ablation study for each part is

Infrared image	MPCM [10]	LIG [49]	SRWS [18]	DNANET [22]	IAANET [23]	ACM [2]	Ours	Ground truth

FIGURE 7. Comparison with other algorithms.

shown in Table 3, and the detection results of six test pictures are selected for display in Figure 6. Row (a) is the picture of SIRST. Row (b) is ground truth. Row (c) is the detection result of the improved U2-Net model. Row (d) is the detection result of the improved U2-Net with the addition of upper and lower attention. Row (e) is the result of the addition of contour detection.

We compare the  $P_d$ ,  $F_a$  and weight size in Table 3, and the improvement of U2-Net not only improves the detection rate and false reduction rate, but also is more suitable for the detection of infrared dim and small target, and the model weight decreases from 176.6 MB to 112.4 MB. The

model  $P_d$  and  $F_a$  are improved a lot with the addition of the attention module, and it can also be seen from Figure 5 that the previously undetected targets can be detected. The cost of adding attention is minimal, and the weight file only increases by 8.2 MB. Adding contour detection does not improve the detection accuracy, but the shape edge contour improves greatly, which is very similar to ground truth.

#### 4) COMPARATIVE EXPERIMENTS AND ANALYSIS

In order to verify the effectiveness of the proposed method, a comparative analysis was performed with existing infrared

**TABLE 4.** Units for magnetic properties detection rate and false detection rate Of 9 methods On 427 images.

Method	RLCM [8]	LCM [9]	MPCM [10]	ILCM [11]	IPI [13]	NRAM [16]	NLOC [17]	SRWS [18]	LIG [49]	IAAN ET[23]	DNAN ET[22]	ACM [2]	Ours
$P_d$	0.4161	0.6403	0.8794	0.6685	0.7532	0.8022	0.7334	0.8474	0.8945	0.9566	0.8832	0.8662	<b>0.9717</b>
$F_a$	5.6497	1.9020	0.2523	1.5254	4.8022	3.9171	4.1431	0.8775	0.5499	0.8116	0.5875	0.2901	<b>0.2335</b>

dim and small target detection methods in the SIRST dataset. The compared algorithms include RLCM, LCM, MPCM, ILCM, SRWS, Local Intensity and Gradient (LIG) [49], IPI, NRAM, NLOC, IAANET, DNANET and ACM. To ensure objectivity, 427 images of SIRST were tested, and all the deep learning-based comparison methods were retrained using the same training dataset as the proposed method. The detection rate and false detection rate are shown in Table 4. As shown in Figure 7, we select five of the better methods to further compare with our method by detection result graph.

It can be seen from Table 4 that the traditional methods LCM and IPI are the worst in terms of detection rate and false detection rate, and the detection process shows a large number of false targets, and many targets fail to be detected for complex scenes. Compared with other methods, our method is better at detecting dim and small targets from complex background images, with a detection rate of 97.17%. Also, the false detection rate is the lowest. Figure 7 shows the detection results based on traditional and deep learning detection algorithms. By comparing the detection results, it can be seen that the traditional algorithms, such as MPCM and LIG, can detect most of the dim and small targets. However, the target shape information is not obvious. The deep learning detection algorithm IAANET has the second highest detection rate, but the detected target shapes tend to lose information. DNANET is prone to many edges false detections and fails to detect the target in some images. ACM can detect the target but can only detect a point shape and not obtain better shape information. The CA-U2-Net detection algorithm proposed in this paper has a better detection effect for large size, small size and multiple targets, and can detect the shape information of the target well. The obtained target shape information is richer and the similarity with the real target is greater.

## V. CONCLUSION

In this paper, we propose the CA-U2-Net network for infrared dim and small target detection. First, the U2-Net network structure is improved to prevent the loss of shallow information due to increased network depth and make it more suitable for detecting dim and small targets. Then, an upper and lower attention block is added to the network to make the model more focused on small target features while suppressing irrelevant information to improve the detection rate further. Finally, a contour detection branch was added at the top of the model to fuse the contour detection map with the feature map of the target detection output to obtain a better target shape. The ablation study demonstrates the effectiveness of each

module in the CA-U2-Net network. Experiments show that the CA-U2-Net network has a higher detection rate in complex scenes and a lower false alarm rate than traditional and deep learning methods in recent years, and can better retain edge information of the dim and small target, which provides a basis for further image analysis. In addition, we construct a new infrared dim and small target dataset consisting of 10,000 images.

## REFERENCES

- [1] X. Ying, Y. Wang, L. Wang, W. Sheng, L. Liu, Z. Lin, and S. Zhou, "MocopNet: Exploring local motion and contrast priors for infrared small target super-resolution," 2022, *arXiv:2201.01014*.
- [2] T. Ma, Z. Yang, J. Wang, S. Sun, X. Ren, and U. Ahmad, "Infrared small target detection network with generate label and feature mapping," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022, doi: 10.1109/LGRS.2022.3140432.
- [3] 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, pp. 3737–3752, May 2021, doi: 10.1109/TGRS.2020.3022069.
- [4] S. Huang, Y. Liu, Y. He, T. Zhang, and Z. Peng, "Structure-adaptive clutter suppression for infrared small target detection: Chain-growth filtering," *Remote Sens.*, vol. 12, no. 1, p. 47, Dec. 2020.
- [5] T. Alexander, K. Skirmantas, and P. Anton, "Adaptive sequential algorithms for detecting targets in a heavy IR clutter," *Proc. SPIE*, vol. 3809, pp. 119–130, Oct. 1999.
- [6] V. Tom, T. Peli, M. Leung, and J. Bondaryk, "Morphology-based algorithm for point target detection in infrared backgrounds," *Proc. SPIE*, vol. 1954, pp. 2–11, Oct. 1993.
- [7] T.-W. Bae, "Small target detection using bilateral filter and temporal cross product in infrared images," *Infr. Phys. Technol.*, vol. 54, no. 5, pp. 403–411, Sep. 2011.
- [8] J. Han, K. Liang, B. Zhou, X. Zhu, J. Zhao, and L. Zhao, "Infrared small target detection utilizing the multiscale relative local contrast measure," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 4, pp. 612–616, Apr. 2018, doi: 10.1109/LGRS.2018.2790909.
- [9] 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, pp. 574–581, Jan. 2014, doi: 10.1109/TGRS.2013.2242477.
- [10] Y. Wei, X. You, and H. Li, "Multiscale patch-based contrast measure for small infrared target detection," *Pattern Recognit.*, vol. 58, pp. 216–226, Oct. 2016.
- [11] 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, doi: 10.1109/LGRS.2014.2323236.
- [12] 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, pp. 4204–4214, Jul. 2016, doi: 10.1109/TGRS.2016.2538295.
- [13] C. 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, doi: 10.1109/TIP.2013.2281420.
- [14] Y. Dai, Y. Wu, and Y. Song, "Infrared small target and background separation via column-wise weighted robust principal component analysis," *Infr. Phys. Technol.*, vol. 77, pp. 421–430, Jul. 2016.

- [15] Y. Dai and Y. Wu, "Reweighted infrared patch-tensor model with both non-local 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.
- [16] 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.
- [17] T. Zhang, H. Wu, Y. Liu, L. Peng, C. Yang, and Z. Peng, "Infrared small target detection based on non-convex optimization with  $l_p$ -norm constraint," *Remote Sens.*, vol. 11, no. 5, p. 559, Mar. 2019.
- [18] T. Zhang, Z. Peng, H. Wu, Y. He, C. Li, and C. Yang, "Infrared small target detection via self-regularized weighted sparse model," *Neurocomputing*, vol. 420, pp. 124–148, Jan. 2021.
- [19] 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)*, Seoul, South Korea, Oct. 2019, pp. 8508–8517, doi: [10.1109/ICCV.2019.00860](https://doi.org/10.1109/ICCV.2019.00860).
- [20] 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)*, Waikoloa, HI, USA, Jan. 2021, pp. 949–958, doi: [10.1109/WACV48630.2021.00099](https://doi.org/10.1109/WACV48630.2021.00099).
- [21] 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, pp. 9813–9824, Nov. 2021, doi: [10.1109/TGRS.2020.3044958](https://doi.org/10.1109/TGRS.2020.3044958).
- [22] B. Li, C. Xiao, L. Wang, Y. Wang, Z. Lin, M. Li, W. An, and Y. Guo, "Dense nested attention network for infrared small target detection," *IEEE Trans. Image Process.*, vol. 32, pp. 1745–1758, 2023, doi: [10.1109/TIP.2022.3199107](https://doi.org/10.1109/TIP.2022.3199107).
- [23] K. Wang, S. Du, C. Liu, and Z. Cao, "Interior attention-aware network for infrared small target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5002013, doi: [10.1109/TGRS.2022.3163410](https://doi.org/10.1109/TGRS.2022.3163410).
- [24] X. Qin, Z. Zhang, C. Huang, M. Dehghan, O. R. Zaiane, and M. Jagersand, "U2-net: Going deeper with nested U-structure for salient object detection," *Pattern Recognit.*, vol. 106, Oct. 2020, Art. no. 107404.
- [25] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015, doi: [10.1109/TPAMI.2015.2389824](https://doi.org/10.1109/TPAMI.2015.2389824).
- [26] Y. Zhang, J. Chu, L. Leng, and J. Miao, "Mask-refined R-CNN: A network for refining object details in instance segmentation," *Sensors*, vol. 20, no. 4, p. 1010, Feb. 2020, doi: [10.3390/s20041010](https://doi.org/10.3390/s20041010).
- [27] J. Chu, Z. Guo, and L. Leng, "Object detection based on multi-layer convolution feature fusion and online hard example mining," *IEEE Access*, vol. 6, pp. 19959–19967, 2018, doi: [10.1109/ACCESS.2018.2815149](https://doi.org/10.1109/ACCESS.2018.2815149).
- [28] H. Kaiming, G. Georgia, D. Piotr, and G. Ross, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2961–2969.
- [29] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 779–788, doi: [10.1109/CVPR.2016.91](https://doi.org/10.1109/CVPR.2016.91).
- [30] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6517–6525.
- [31] X. Zhou, L. Jiang, X. Guan, and X. Mou, "Infrared small target detection algorithm with complex background based on YOLO-NWD," in *Proc. 4th Int. Conf. Image Process. Mach. Vis. (IPMV)*, Mar. 2022, pp. 6–12.
- [32] X. Mou, S. Lei, and X. Zhou, "YOLO-FR: A YOLOv5 infrared small target detection algorithm based on feature reassembly sampling method," *Sensors*, vol. 23, no. 5, p. 2710, Mar. 2023.
- [33] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*.
- [34] X. Tong, B. Sun, J. Wei, Z. Zuo, and S. Su, "EAAU-net: Enhanced asymmetric attention U-net for infrared small target detection," *Remote Sens.*, vol. 13, no. 16, p. 3200, Aug. 2021.
- [35] 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)*, New Orleans, LA, USA, Jul. 2022, pp. 867–876, doi: [10.1109/CVPR52688.2022.00095](https://doi.org/10.1109/CVPR52688.2022.00095).
- [36] C. Rother, V. Kolmogorov, and A. Blake, "GrabCut: Interactive foreground extraction using iterated graph cuts," *ACM Trans. Graph.*, vol. 23, pp. 309–314, Aug. 2004.
- [37] P. Arbeláez, M. Maire, C. Fowlkes, and J. Malik, "Contour detection and hierarchical image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 898–916, May 2011, doi: [10.1109/TPAMI.2010.161](https://doi.org/10.1109/TPAMI.2010.161).
- [38] V. Ferrari, L. Fevrier, F. Jurie, and C. Schmid, "Groups of adjacent contour segments for object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 1, pp. 36–51, Jan. 2008, doi: [10.1109/TPAMI.2007.1144](https://doi.org/10.1109/TPAMI.2007.1144).
- [39] H. Zhang, K. Jiang, Y. Zhang, Q. Li, C. Xia, and X. Chen, "Discriminative feature learning for video semantic segmentation," in *Proc. Int. Conf. Virtual Reality Visualizat.*, Shenyang, China, Aug. 2014, pp. 321–326, doi: [10.1109/ICVRV.2014.65](https://doi.org/10.1109/ICVRV.2014.65).
- [40] Y. Wang, X. Zhao, Y. Li, and K. Huang, "Deep crisp boundaries: From boundaries to higher-level tasks," *IEEE Trans. Image Process.*, vol. 28, no. 3, pp. 1285–1298, Mar. 2019, doi: [10.1109/TIP.2018.2874279](https://doi.org/10.1109/TIP.2018.2874279).
- [41] G. Bertasius, J. Shi, and L. Torresani, "Semantic segmentation with boundary neural fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3602–3610.
- [42] J. Canny, "A computational approach to edge detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-8, no. 6, pp. 679–698, Nov. 1986, doi: [10.1109/TPAMI.1986.4767851](https://doi.org/10.1109/TPAMI.1986.4767851).
- [43] S. Xie and Z. Tu, "Holistically-nested edge detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Santiago, CL, USA, Dec. 2015, pp. 1395–1403, doi: [10.1109/ICCV.2015.164](https://doi.org/10.1109/ICCV.2015.164).
- [44] Y. Liu, M.-M. Cheng, X. Hu, J.-W. Bian, L. Zhang, X. Bai, and J. Tang, "Richer convolutional features for edge detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 1939–1946, Aug. 2019, doi: [10.1109/TPAMI.2018.2878849](https://doi.org/10.1109/TPAMI.2018.2878849).
- [45] Z. Su, W. Liu, Z. Yu, D. Hu, Q. Liao, Q. Tian, M. Pietikäinen, and L. Liu, "Pixel difference networks for efficient edge detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Montreal, QC, Canada, Oct. 2021, pp. 5097–5107, doi: [10.1109/ICCV48922.2021.00507](https://doi.org/10.1109/ICCV48922.2021.00507).
- [46] X. Soria, E. Riba, and A. Sappa, "Dense extreme inception network: Towards a robust CNN model for edge detection," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Snowmass, CO, USA, Mar. 2020, pp. 1912–1921, doi: [10.1109/WACV45572.2020.9093290](https://doi.org/10.1109/WACV45572.2020.9093290).
- [47] Y.-J. Cao, C. Lin, and Y.-J. Li, "Learning crisp boundaries using deep refinement network and adaptive weighting loss," *IEEE Trans. Multimedia*, vol. 23, pp. 761–771, 2021, doi: [10.1109/TMM.2020.2987685](https://doi.org/10.1109/TMM.2020.2987685).
- [48] J.-J. Liu, Q. Hou, M.-M. Cheng, J. Feng, and J. Jiang, "A simple pooling-based design for real-time salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 3912–3921, doi: [10.1109/CVPR.2019.00404](https://doi.org/10.1109/CVPR.2019.00404).
- [49] H. Zhang, L. Zhang, D. Yuan, and H. Chen, "Infrared small target detection based on local intensity and gradient properties," *Infr. Phys. Technol.*, vol. 89, pp. 88–96, Mar. 2018.



**LEIHONG ZHANG** was born in Shanghai, China. He received the B.S. and M.S. degrees from Jiangsu University and the Ph.D. degree from the Shanghai Institute of Optics and Mechanics, Chinese Academy of Sciences. He joined the Shanghai University of Engineering Science, for teaching, in 2009, and the University of Shanghai for Science and Technology, for teaching, in 2010. His current research interests include image processing, opto-fluidics optics, and medical instruments.



**WEIHONG LIN** was born in Fujian, China. He is currently pursuing the master's degree with the University of Shanghai for Science and Technology. His main research interests include image processing and object detection.



**ZIMIN SHEN** was born in Zhejiang, China. He is currently pursuing the master's degree with the University of Shanghai for Science and Technology. His main research interests include image processing and object detection.



**KAIMIN WANG** was born in Shanghai, China. He received the Ph.D. degree from the Beijing University of Posts and Telecommunications (BUPT), in 2016. He is currently a Master's Tutor with the University of Shanghai for Science and Technology (USST) and a member of the Ministry of Education and the Shanghai Key Laboratory of Modern Optical System, USST. His main research interests include optical signal processing and micro/nano-fabrication.



**DAWEI ZHANG** was born in Shanghai, China. He received the Ph.D. degree from the Shanghai Institute of Optics and Fine Mechanics, Chinese Academy of Sciences, in 2005. In 2005, he joined the School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, as a Professor. His current research interests include optical thin film, micro-nano optical devices, and opto-fluidics optics.



**BANGLIAN XU** was born in Shanghai, China. He received the Ph.D. degree in optical engineering from the Shanghai University of Technology, in June 2016. His main research interests include the design and preparation of new grating devices and optical anti-counterfeiting technology development.



**JIAN CHEN** (Senior Member, IEEE) was born in Changchun, China. He received the Ph.D. degree from the University of Chinese Academy of Sciences, in 2014. He is currently a Master's Tutor with the Changchun Institute of Optics, Fine Mechanics and Physics. His main research interests include optical imaging and image analysis.

...