remote sensing

MDPI

Article

UCDnet: Double U-Shaped Segmentation Network Cascade Centroid Map Prediction for Infrared Weak Small Target Detection

Xiangdong Xu [1,2], Jiarong Wang [1,*], Ming Zhu [1], Haijiang Sun [1], Zhenyuan Wu [1,2], Yao Wang [1,2], Shenyi Cao [3] and Sanzai Liu [1,2]

[1] Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun 130033, China; xuxiangdong21@mails.ucas.ac.cn (X.X.); zhuming@ciomp.ac.cn (M.Z.); sunhj@ciomp.ac.cn (H.S.); wuzhenyuan22@mails.ucas.ac.cn (Z.W.); wangyao222@mails.ucas.ac.cn (Y.W.); liusanzai22@mails.ucas.ac.cn (S.L.)

[2] University of Chinese Academy of Sciences, Beijing 100049, China

[3] Faculty of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an 710000, China; caoshenyi@stu.xjtu.edu.cn

* Correspondence: wangjiarong@cust.edu.cn; Tel.: +86-138-4406-4774

Abstract: In recent years, the development of deep learning has brought great convenience to the work of target detection, semantic segmentation, and object recognition. In the field of infrared weak small target detection (e.g., surveillance and reconnaissance), it is not only necessary to accurately detect targets but also to perform precise segmentation and sub-pixel-level centroid localization for infrared small targets with low signal-to-noise ratio and weak texture information. To address these issues, we propose UCDnet (Double U-shaped Segmentation Network Cascade Centroid Map Prediction for Infrared Weak Small Target Detection) in this paper, which completes "end-to-end" training and prediction by cascading the centroid localization subnet with the semantic segmentation subnet. We propose the novel double U-shaped feature extraction network for point target fine segmentation. We propose the concept and method of centroid map prediction for point target localization and design the corresponding Com loss function, together with a new centroid localization evaluation metrics. The experiments show that ours achieves target detection, semantic segmentation, and sub-pixel-level centroid localization. When the target signal-to-noise ratio is greater than 0.4, the IoU of our semantic segmentation results can reach 0.9186, and the average centroid localization precision can reach 0.3371 pixels. On our simulated dataset of infrared weak small targets, the algorithm we proposed performs better than existing state-of-the-art networks in terms of semantic segmentation and centroid localization.

Keywords: infrared weak small target; target detection; semantic segmentation; centroid localization; sub-pixel-level localization

check for updates

Citation: Xu, X.; Wang, J.; Zhu, M.; Sun, H.; Wu, Z.; Wang, Y.; Cao, S.; Liu, S. UCDnet: Double U-Shaped Segmentation Network Cascade Centroid Map Prediction for Infrared Weak Small Target Detection. *Remote Sens.* **2023**, *15*, 3736. https://doi.org/10.3390/rs15153736

Academic Editors: Yanni Dong, Xiaochen Yang and Qian Du

Received: 5 July 2023
Revised: 23 July 2023
Accepted: 25 July 2023
Published: 27 July 2023

1. Introduction

Infrared small target detection technology, as the main technical support for surveillance and reconnaissance [1,2], precise localization [3–5], and attitude estimation [6], has been widely applied in various fields. In the field of surveillance and reconnaissance, such as drone tracking and search and rescue operations, not only the rough detection of infrared targets is required but also precise segmentation and centroid localization of the targets. This is essential for effective prediction of target motion trajectories, thereby enabling early warning and appropriate measures. In the field of the Internet of Things (IoT), such as smart transportation and smart agriculture, precise detection is a prerequisite for achieving perception and decision-making. In the aforementioned application scenarios, targets often occupy a small number of pixels in the image and suffer from issues such as texture loss and low signal-to-noise ratio.

Challenges:

On one hand, as the distance between the imaging device and the target increases, the size of the target in the image becomes smaller, and the target appears in a faint and weak state. Existing semantic segmentation algorithms perform well in segmenting large-sized or clearly bounded targets, but they show limited effectiveness in segmenting infrared small targets with low signal-to-noise ratio and weak texture information.

On the other hand, existing infrared weak small target detection methods focus on bounding box detection and pay less attention to the problem of accurate centroid localization of these targets. When the infrared small targets have irregular shapes and undergo continuous changes in posture, the center of the bounding box cannot effectively represent the centroid of the target. Existing deep learning-based methods for predicting target centroids are still limited to pixel-level predictions, only able to predict which pixel the centroid belongs to in the image. In certain specific scenarios, a deviation of one pixel in predicting the target centroid can result in several meters of actual distance, and this deviation may be further amplified as the process continues.

Existing Methods:

To date, there are various traditional methods for infrared small target detection. Common approaches include: (1) filter-based methods [7–9], (2) methods based on the human visual system [10–14], and (3) methods based on low-rank sparse recovery [15,16]. However, filter-based detection algorithms can only suppress uniform background to a certain extent and fail to address complex background problems, resulting in low detection performance and poor robustness. Methods based on the human visual system rely on contrasting the differences between small target regions and their surrounding areas, making them unsuitable for detecting weak small targets. Low-rank sparse recovery methods can improve detection precision by applying certain constraints. However, in infrared images with weak small targets and complex backgrounds, some strong clutter signals may be as sparse as the target signal, leading to increased false alarm. Overall, traditional algorithms heavily rely on manual craftsmanship, lack adaptability, and are difficult to be widely applied in infrared target detection in various application scenarios.

Recent approaches have turned to deep learning methods to address these issues. For example, Liu et al. [17] developed a CNN-based method for infrared small target detection. Y. Dai et al. [18] encoded spatial information and proposed an Asymmetric Context Module (ACM). They also introduced the SIRST dataset, which has been widely used in the field of infrared small target detection. Other researchers, such as Q. Hou et al. [19], proposed a robust infrared small target detection network (RISTDnet) based on deep learning. They constructed a feature extraction framework that combines handcrafted features and convolutional neural networks, established a mapping network between the feature map and the likelihood of small targets in the image, and applied threshold segmentation to identify real targets. X. Zhou et al. [20] unfolded the sparse low-rank regularization model into a deep neural network. They used the infrared patch-image (IPI) model to convert the original infrared image into patch images through local patch construction. Zhao et al. [21] proposed a detection pattern based on generative adversarial networks (GAN) to automatically learn the features of the target and directly predict the intensity of the target.

Our Contribution:

In this paper, to address the sub-pixel-level accurate centroid localization of infrared weak small targets, we propose UCDnet (Double U-shaped Segmentation Network Cascade Centroid Map Prediction for Infrared Weak Small Target Detection), and its prediction results are shown in Figure 1. The algorithm takes the input infrared image and passes it through a semantic segmentation subnet for feature extraction, feature fusion, and dense prediction, outputting pixel-level predictions (i.e., semantic segmentation results) of the targets (Figure 1a). Meanwhile, the input infrared image is also fed into the centroid localization subnet, which generates a centroid map with target centroid position information. The pixel-level predictions from the semantic segmentation subnet are mapped onto the centroid map, allowing the centroid map to retain only the results corresponding to the target's pixel locations. Finally, through decoding operations, accurate centroid results of the targets can be obtained (Figure 1b). Additionally, we have created an infrared simulation dataset specifically tailored for weak small targets.

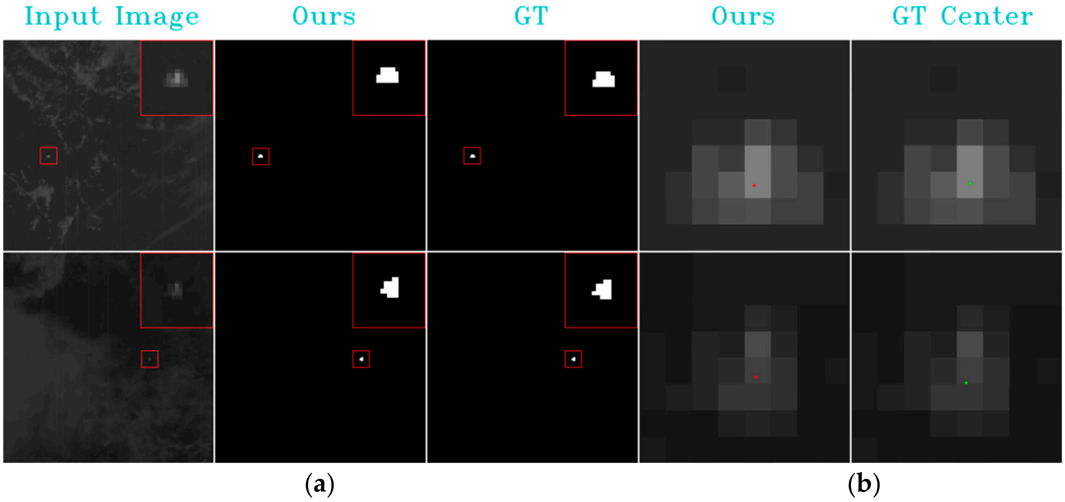


Figure 1. The performance of our algorithm on our simulated datasets for semantic segmentation and centroid localization. (**a**) The **left column** shows our UCDnet semantic segmentation results, and the **right column** shows the corresponding ground truth. (**b**) The **left column** shows our UCDnet centroid localization results, and the **right column** shows the corresponding ground truth.

Our contributions can be summarized as follows:

We have designed a novel sub-pixel-level centroid localization method for infrared weak small targets. We innovatively propose the concept and method of centroid map prediction, addressing the challenge of accurately locating the centroid of point targets. Our method utilizes a lightweight fully convolutional structure and incorporates a novel Com loss function. We also introduce a new evaluation metric for centroid localization. Experimental results demonstrate that our average centroid localization precision achieves 0.3371 pixels when the signal-to-noise ratio of the targets is above 0.4.

We have proposed a feature extraction network for semantic segmentation of infrared weak small targets. Our method employs a novel double U-shaped structure, which effectively combines the feature extraction module and attention module to finer predict the easily overlooked edge pixels, which is crucial for precise centroid localization. Experimental results show that our semantic segmentation IoU achieves 0.9186 when the signal-to-noise ratio of the targets is above 0.4.

We have simulated a series of infrared weak small target datasets for training, validation, and testing purposes. These datasets include various forms of infrared small targets. To replicate real-world scenarios, we have introduced multiple types of vertical line noise and Gaussian noise in the images. We have also specifically simulated the dim and weak states of targets when they pass through cloud, which pose challenges for detection algorithms.

2. Materials and Methods

The main idea of UCDnet is to construct an infrared small target detection network that achieves dense prediction and sub-pixel-level centroid localization. Its network structure is shown in Figure 2 and consists of two parts: the semantic segmentation subnet and the centroid localization subnet. Section 2.1 introduces the simulation content of the dataset. Section 2.2 presents the proposed semantic segmentation method, while Section 2.3 describes the proposed sub-pixel-level centroid localization method. Finally, Section 2.4 discusses the loss function and training method used during network training.

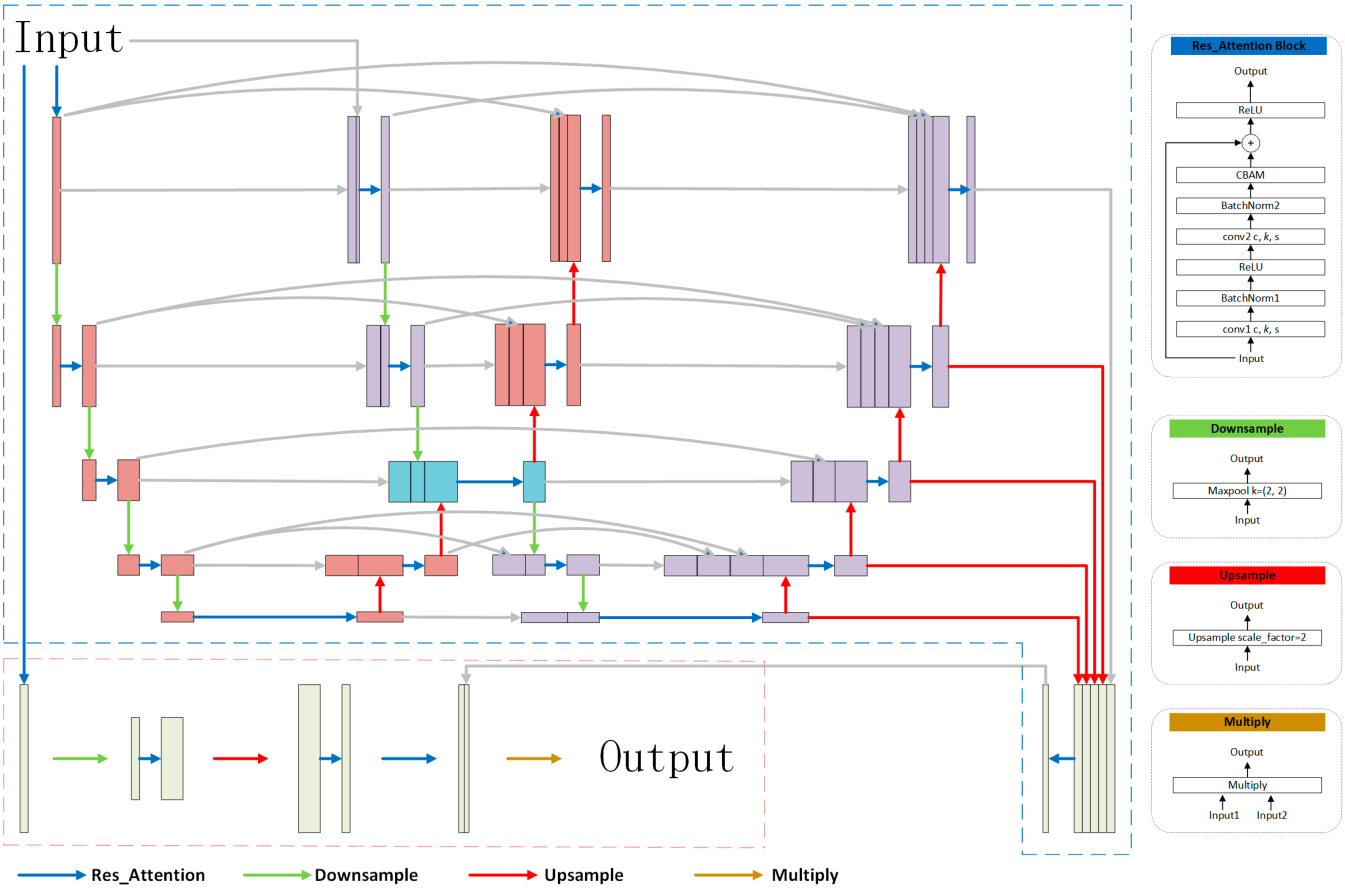


Figure 2. The network structure diagram of our proposed algorithm, UCDnet, consists of a semantic segmentation subnet and a centroid localization subnet. In the semantic segmentation subnet, a double U-shaped backbone network, which embeds attention modules, is used to extract features from the input infrared images. A feature fusion prediction branch is then used to perform pixel-wise binary classification based on the fused features at different levels, generating a semantic segmentation mask. In the centroid localization subnet, a "floating-point" centroid map is predicted based on the input infrared images. The semantic segmentation mask generated by the semantic segmentation subnet is mapped into the centroid map to narrow down the predicted range of the centroid, enabling accurate calculation of the target's centroid. The details of the Res_Attention block, Downsample block, Upsample_n block, and Multiply block are shown in Table A1 in Appendix A.

2.1. Dataset Simulation

The dataset used in this paper is simulated based on satellite imagery (with a size of around 10,000 × 10,000 pixels) collected by meteorological satellites. The simulated targets are generated based on the grayscale distribution of real targets, and a total of 125 targets were simulated. The extraction of target features from infrared images is closely related to background information and noise [22]. The image data from different backgrounds exhibit significant differences. For meteorological satellite images, cloud cover is the main source of interference, and the presence of high-intensity noise further complicates target detection. We simulated 8 to 14 frames of motion trajectories for each target individually. In order to make the dataset more representative of real-world scenarios, slight variations in the shape and pixel values of the targets were introduced within each trajectory, resulting in a total of 1306 images. Finally, random cropping was performed on these images to obtain sub-images of size 256 × 256 pixels containing the targets (the simulation process of the infrared weak small target dataset is illustrated in Figure A1 in Appendix A). The size of the targets ranged from 6 × 4 pixels to 10 × 13 pixels, and the signal-to-noise ratio (SNR) of the images varied from 0.4 to 6.4. It can also be observed from Figure 3 that the SNR of the image complies with the central limit theorem. Figure 4 illustrates the range of target sizes, with blue dots representing the number of targets for each size. Figure 5 showcases the approximate shapes of the targets.

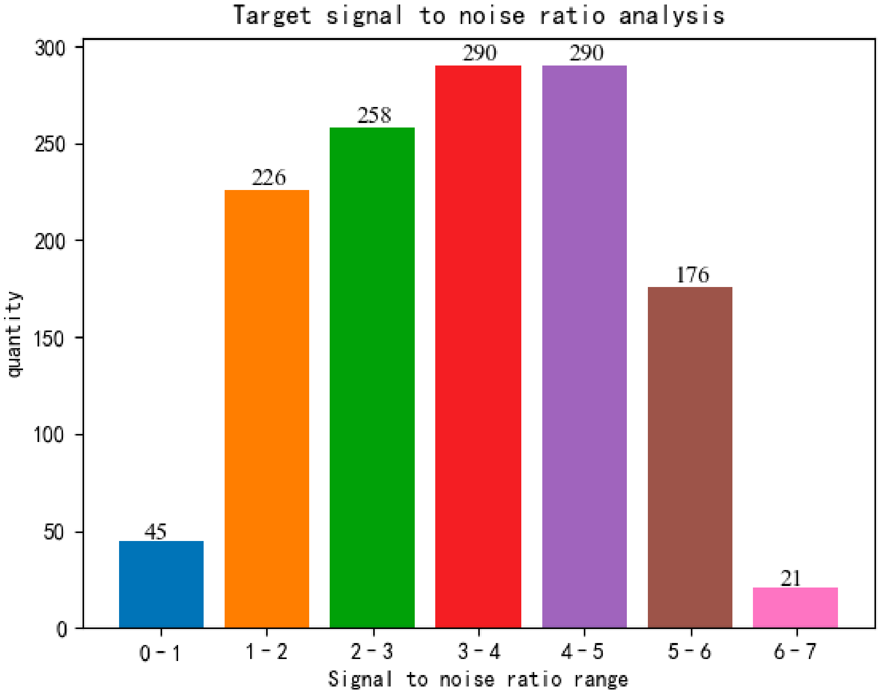


Figure 3. Histogram of the target signal-to-noise ratio distribution.

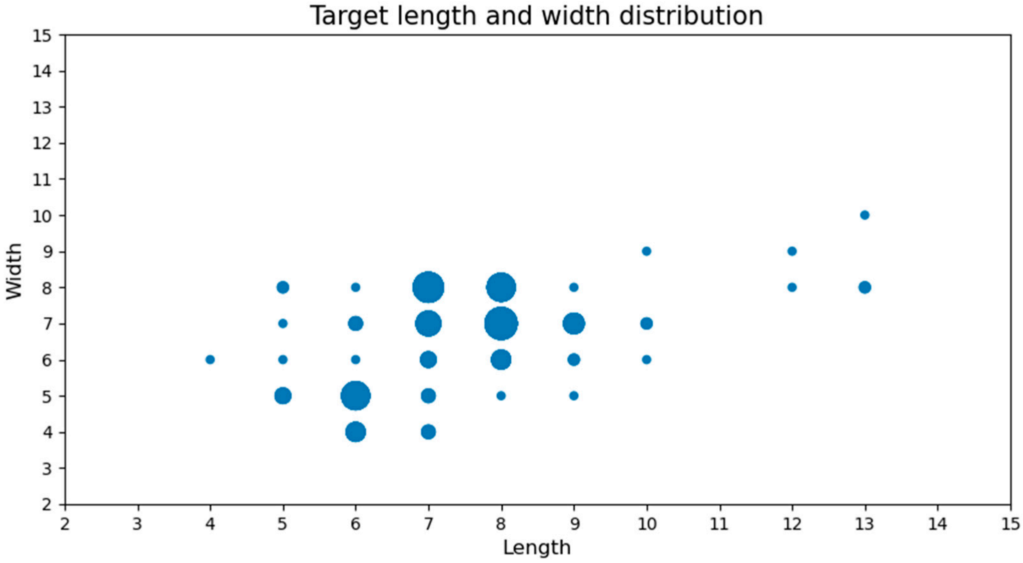


Figure 4. Distribution of target's length and width, along with the quantity distribution for each size.

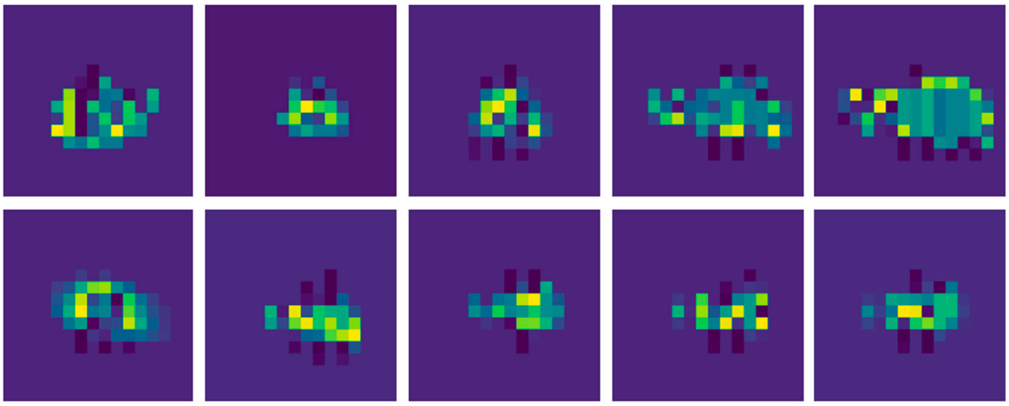


Figure 5. The schematic representation of the shapes of certain targets in the dataset used in this paper.

2.2. Semantic Segmentation Subnet

The semantic segmentation subnet consists of a feature extraction backbone network and a feature fusion prediction branch. The feature extraction backbone network adopts a double U-shaped structure, focusing on extracting edge features of the targets to handle low signal-to-noise ratio situations. It incorporates embedded attention modules to enhance segmentation precision. The feature fusion prediction branch performs feature fusion at different hierarchical levels and conducts per-pixel binary classification based on the fused results to generate semantic segmentation masks.

The double U-shaped network structure is shown in Figure 6. Inspired by the U-Net [23] and DNAnet [24] network structures, we have made improvements to the main feature extraction network. It follows the overall Encoder–Decoder structure, which is similar to the combination of two large U-shaped networks and one small U-shaped network. The two large U-shaped networks perform four downsampling operations, where the shallow network extracts more target texture information, and the deep network allows for a larger field of view and captures deeper semantic information [25]. The small U-shaped network just performs two downsampling operations, reserving more target positional information than bigger ones. Compared to the U-Net network, our approach uses more feature concatenation to achieve precise extraction of edge features.

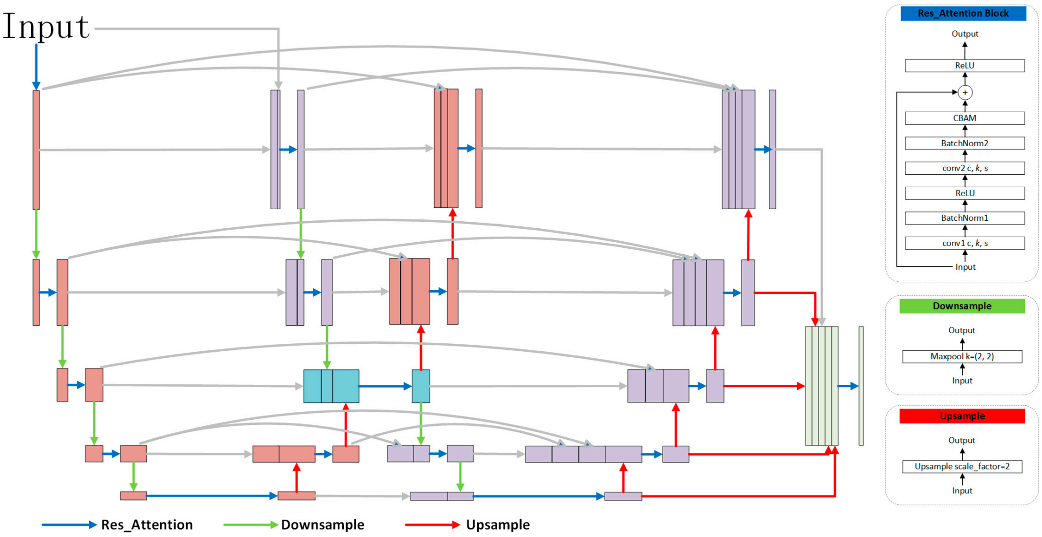


Figure 6. The structure of semantic segmentation subnet. Firstly, the input infrared image is processed by a double U-shaped backbone network with embedded attention modules to extract features. Then, the feature fusion prediction branch utilizes the fused features at different hierarchical levels to perform per-pixel binary classification and generate semantic segmentation masks.

In contrast to most existing infrared small target semantic segmentation methods, inspired by the DNAnet network structure, each layer of our main feature extraction network incorporates a specific attention module [26]. Each attention module in a layer includes channel attention and spatial attention components, enabling adaptive learning of the importance of different channels and spatial positions in the image. The channel attention dynamically adjusts the impact of each channel in the output feature map, while the spatial attention dynamically searches for the precise location of the targets on the feature map. The combination of two attentions helps the network pay more attention to the positional information of small targets and some detailed edge information in each feature map. The feature extraction network determines the overall position of the infrared small targets based on their grayscale and contrast information, while the attention modules focus more on the edge information of the targets.

The basic and core component of the entire network is the Res_Attention Block. In each layer, the input feature map undergoes a series of operations, including a 3 × 3 convolution, normalization, activation function, another 3 × 3 convolution, and normalization, followed by a hybrid attention mechanism. The output of this series of operations is then added to the original input to obtain the output of the Res_Attention Block. The hybrid attention mechanism consists of a channel attention mechanism followed by a spatial attention mechanism. For an input F, the equations for the channel attention mechanism (Mc(F)) and the spatial attention mechanism (Ms(F)) can be expressed as:

$$M_c(F) = \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) \tag{1}$$

$$M_s(F) = \sigma\left(f^{7\times7}([AvgPool(F); MaxPool(F)])\right) \tag{2}$$

where σ represents the sigmoid operation, MLP represents a convolution operation with a kernel size of 1 × 1, AvgPool represents average pooling, MaxPool represents maximum pooling, [;] represents channel concatenation, and $f^{7\times7}$ represents a convolution with a kernel size of 7 × 7.

2.3. Centroid Localization Subnet

The centroid localization subnet (shown in Figure 7) consists of the centroid feature extraction network and the mapping module. The feature extraction network includes a hybrid attention mechanism, downsampling, and upsampling structures. The hybrid attention mechanism focuses on learning the relationship between the image's grayscale, contrast, and target centroids. The mapping module performs pixel-wise multiplication, multiplying the binary segmentation result with the centroid map. This operation filters out the parts that do not belong to the target, resulting in more accurate centroid localization.

The centroid localization subnet predicts the "floating-point" centroid map (as shown in Figure 8) based on the semantic segmentation results to calculate the target centroid. The centroid map, which is different from the heatmap proposed in CenterNet [27], is first introduced and applied by us. In the centroid map, the value of each pixel represents the proximity to the target centroid. To achieve sub-pixel-level centroid localization, we do not designate a specific pixel as the centroid in both the ground truth label map and the predicted map. Instead, we employ an indirect constraint method to obtain the floating-point centroid position (sub-pixel-level). Specifically, during the training phase, the labels for the centroid localization subnet are calculated using a two-dimensional Gaussian function [28]. When the target centroid position is (x_0, y_0), the label value at any position in the centroid map can be obtained using the two-dimensional Gaussian function as follows:

$$V_{ij} = \mathrm{e}^{-\frac{(x_0-j)^2+(y_0-i)^2}{2}} \tag{3}$$

where j represents the current pixel's column, i represents the current pixel's row, and V_{ij} represents the corresponding label value at that position. As shown in Figure 8, when the centroid position is $x_0 = 3.3$, $y_0 = 4.6$, we can calculate the values of V_{43}, V_{53}, and V_{54} as follows:

$$V_{43} = \mathrm{e}^{-\frac{(3.3-3)^2+(4.6-4)^2}{2}} \approx 0.79852 \text{ (yellow pixel)}$$

$$V_{53} = \mathrm{e}^{-\frac{(3.3-3)^2+(4.6-5)^2}{2}} \approx 0.88250 \text{ (red pixel)}$$

$$V_{54} = \mathrm{e}^{-\frac{(3.3-4)^2+(4.6-5)^2}{2}} \approx 0.72253 \text{ (blue pixel)}$$

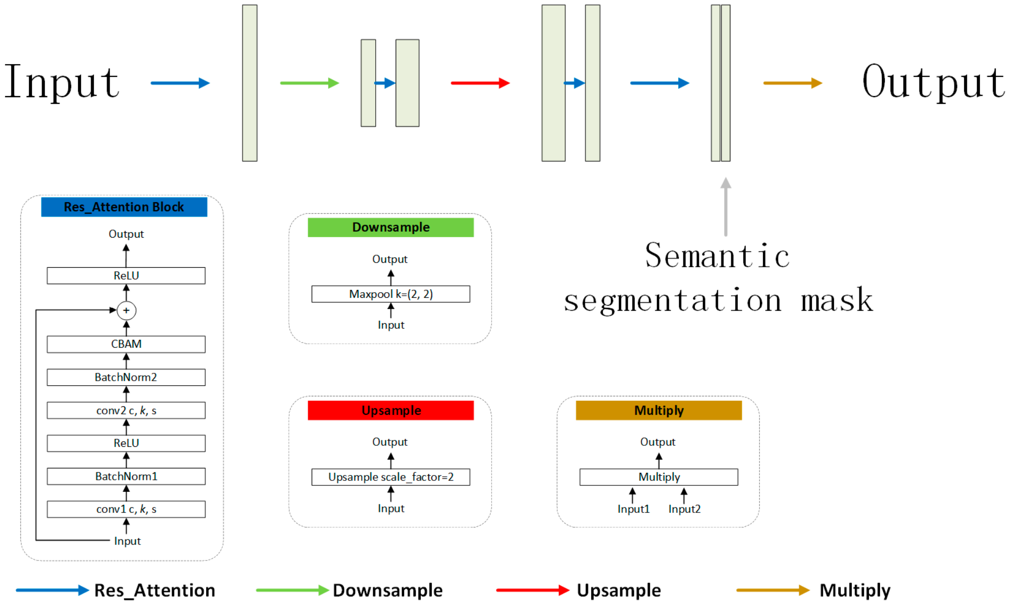
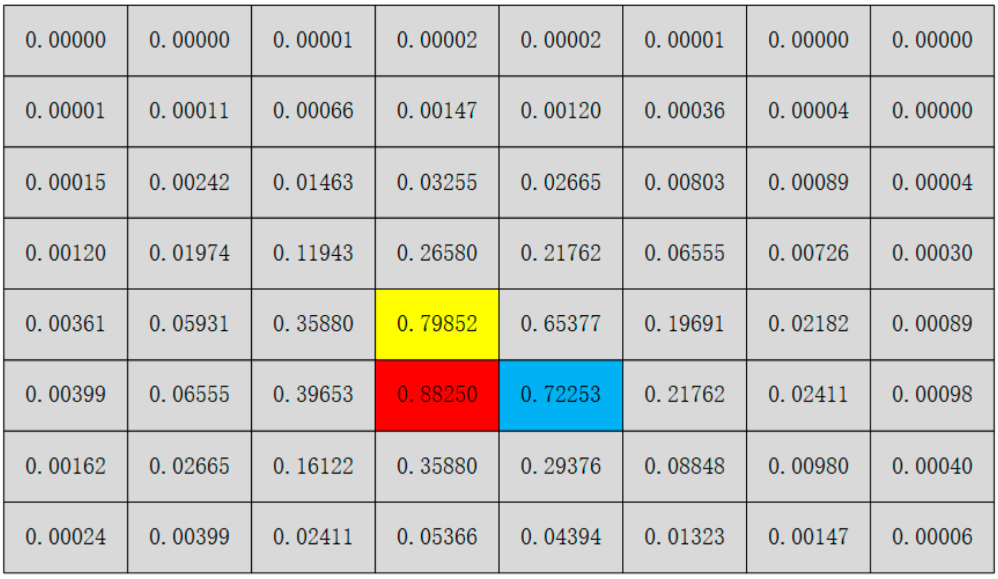


Figure 7. The structure of centroid localization subnet. Firstly, the feature extraction module with embedded attention modules is used to extract the regions where all suspected target centroids are located in the image. Then, the more accurate pixel-level prediction results of the targets generated from semantic segmentation subnet are mapped to the centroid map. This process eliminates false target centroid regions and preserves the true target centroid regions (i.e., centroid map).



0. 00000	0. 00000	0. 00001	0. 00002	0. 00002	0. 00001	0. 00000	0. 00000
0. 00001	0. 00011	0. 00066	0. 00147	0. 00120	0. 00036	0. 00004	0. 00000
0. 00015	0. 00242	0. 01463	0. 03255	0. 02665	0. 00803	0. 00089	0. 00004
0. 00120	0. 01974	0. 11943	0. 26580	0. 21762	0. 06555	0. 00726	0. 00030
0. 00361	0. 05931	0. 35880	0. 79852	0. 65377	0. 19691	0. 02182	0. 00089
0. 00399	0. 06555	0. 39653	0. 88250	0. 72253	0. 21762	0. 02411	0. 00098
0. 00162	0. 02665	0. 16122	0. 35880	0. 29376	0. 08848	0. 00980	0. 00040
0. 00024	0. 00399	0. 02411	0. 05366	0. 04394	0. 01323	0. 00147	0. 00006

Figure 8. The label for training centroid localization subnet and also the predicted centroid map.

During the prediction phase, the actual centroid positions are calculated based on the predicted centroid map, following these steps:

1. Firstly, determine the number of targets N in our predicted centroid map, which corresponds to N clusters of pixels.
2. For each cluster of pixels mentioned above, select the maximum value $V_{i_1j_1}$ from our predicted results as the centroid reference point $A_{i,i\in[1,\ N]}$ (represented by the red position in Figure 8).
3. Take the centroid reference point A_i as the center and find the maximum value $V_{i_2j_2}$ within the vertical neighborhood $\mathrm{N}_4(A_\mathrm{i})$ to serve as the centroid vertical correction point $B_{i,i\in[1,\ N]}$ (represented by the yellow position in Figure 8).
4. Similarly, take the centroid reference point A_i as the center and find the maximum value $V_{i_3j_3}$ within the horizontal neighborhood $\mathrm{N}_4(A_\mathrm{i})$ to serve as the centroid horizontal correction point $C_{i,i\in[1,\ N]}$ (represented by the blue position in Figure 8).
5. Combining the centroid reference point A_i, the vertical correction point B_i, and the horizontal correction point C_i, calculate the sub-pixel-level precise positions x_0 and y_0 of each centroid using the following equations:

$$V_{i_1j_1} = \mathrm{e}^{-\frac{(x_0-j_1)^2+(y_0-i_1)^2}{2}} \tag{4}$$

$$V_{i_2j_2} = \mathrm{e}^{-\frac{(x_0-j_2)^2+(y_0-i_2)^2}{2}} \tag{5}$$

$$V_{i_3j_3} = \mathrm{e}^{-\frac{(x_0-j_3)^2+(y_0-i_3)^2}{2}} \tag{6}$$

where i_1 = i_3 and j_1 = j_2, the above three equations can be used to obtain the following:

$$x_0 = \frac{2\ln\frac{V_{i_3j_3}}{V_{i_1j_1}} - \left(j_1^2 - j_3^2\right)}{2(j_3 - j_1)} \tag{7}$$

$$y_0 = \frac{2\ln\frac{V_{i_2j_2}}{V_{i_1j_1}} - \left(i_1^2 - i_2^2\right)}{2(i_2 - i_1)} \tag{8}$$

6. Repeat the above steps 2 to 5 until all N predicted targets have been traversed.

In summary, compared to directly predicting the centroid positions (x_0, y_0) using the network, predicting the centroid map makes the network training more easily convergent (due to the presence of more constraints). Furthermore, compared to predicting bounding boxes and then post-processing to obtain the centroids, the proposed method of predicting the centroid map and then extracting the centroids is more efficient and accurate.

2.4. Loss Function

The loss function used in our method consists of two parts. For the semantic segmentation part, as shown in Equation (9), we employ Dice Loss as the loss function. For the centroid localization part, as shown in Equation (10), we have developed our own loss function based on the MSE Loss, which we refer to as the Center of Mass Loss (Com Loss).

$$\text{Dice Loss} = 1 - 2\frac{|A \cap B|}{|A| + |B|} \tag{9}$$

$$\text{Com Loss} = \begin{cases} (y-p)^2 & if\ y > 0.1 \\ (y-p)^2(1-y)^4 & other \end{cases} \tag{10}$$

In Equation (9), A and B represent the semantic segmentation label (per-pixel) and the predicted semantic segmentation results (per-pixel), respectively. In Equation (10), y represents the centroid map label (per-position), and p represents the predicted centroid map (per-position).

To improve the algorithm's performance, we first optimize the semantic segmentation results. During the first 10 epochs, we only calculate the loss value for semantic segmentation. After the 10th epoch, we add to calculate the loss value for the centroid localization module. The weights assigned to the loss values of the semantic segmentation and centroid localization parts follow specific function distributions: $\frac{1}{1+e^{-0.1(60-epoch)+0.5}} + 1$ for the semantic segmentation part and $\left(1 - \frac{1}{1+e^{-0.1(60-epoch)+0.5}}\right) \times 0.8 + 0.2$ for the centroid localization part (shown in Figure 9). This approach improves the overall efficiency of the network by filtering out unnecessary computations and further enhances the performance of the centroid localization module.

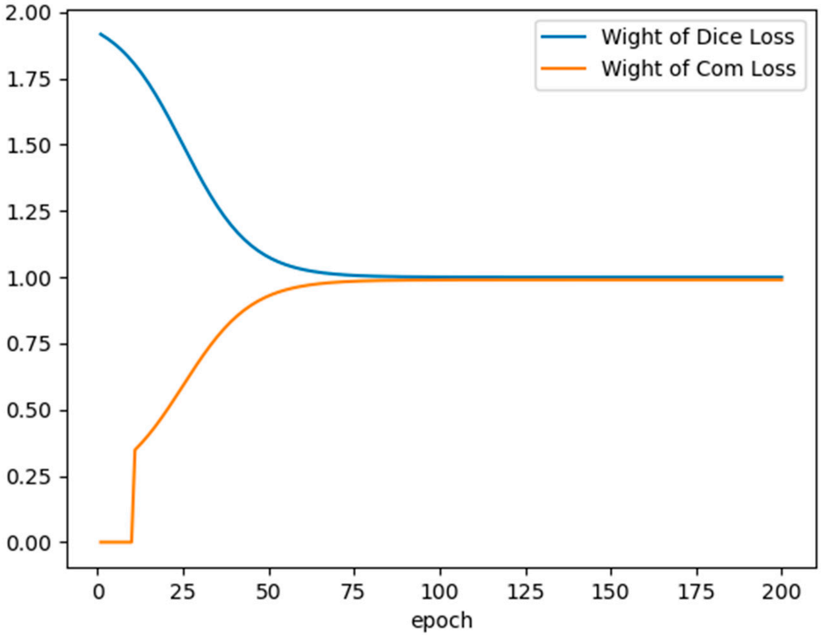


Figure 9. Illustration of the weights of Dice Loss and Com Loss in the overall loss.

3. Results

This section presents the experimental results of the proposed method in this paper, along with some comparative results, which demonstrate that our method outperforms other approaches. Section 3.1 provides a detailed introduction to the evaluation metrics used in this study. Section 3.2 presents all the experimental procedures and analyses.

3.1. Evaluation Metrics

3.1.1. Semantic Segmentation Evaluation Metrics

Since there is only one semantic segmentation class in this paper, the main evaluation metrics used are IoU, Precision, and Recall. In semantic segmentation, the definition of IoU is the ratio of the intersection to the union of the ground truth and predicted sets. This ratio can be expressed as the ratio of True Positive (TP) to the sum of TP, False Positive (FP), and False Negative (FN), which represents the overall segmentation precision. Precision is defined as the ratio of the intersection to the predicted segmentation set. This ratio can be expressed as the ratio of TP to the sum of TP and FP, reflecting the precision of the predicted results. Recall is defined as the ratio of the intersection to the ground truth segmentation set. This ratio can be expressed as the ratio of TP to the sum of TP and FN, reflecting the recall or sensitivity of the predicted results. TP, FP, FN, and TN are shown in Figure 10.

$$\mathrm{IoU} = \frac{\mathrm{TP}}{\mathrm{FP} + \mathrm{FN} + \mathrm{TP}} \tag{11}$$

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{FP} + \mathrm{TP}} \tag{12}$$

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{FN} + \mathrm{TP}} \tag{13}$$

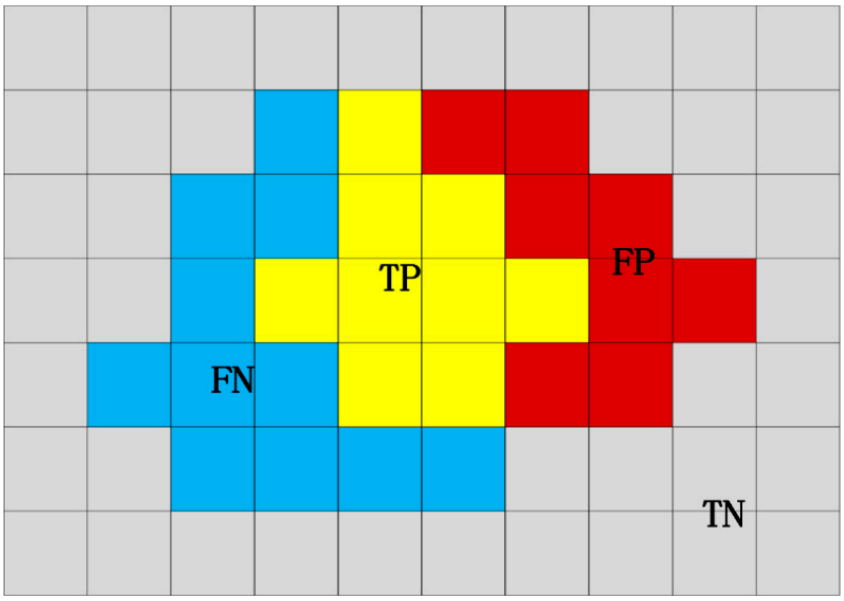


Figure 10. Illustration of IoU evaluation metric. TP (True Positive): the predicted result is positive, and it is actually positive. FP (False Positive): the predicted result is positive, but it is actually negative. FN (False Negative): the predicted result is negative, but it is actually positive. TN (True Negative): the predicted result is negative, and it is actually negative.

3.1.2. Centroid Localization Evaluation Metrics

If there is only one target in each image, we can directly use the Euclidean distance between the predicted centroid and the actual centroid of the target as the evaluation metric for centroid localization. However, considering the cases of multiple targets or false alarms, this paper introduces a novel centroid localization evaluation metric. Specifically, clustering is performed with the real centroids as centers, and all detected centroids are assigned to the nearest real centroid. Then, the sum of distances between all predicted centroids and their closest real centroids is calculated. This distance sum is divided by the smaller value between the number of real centroids and the number of detected centroids. The calculation equation is as follows:

$$\mathrm{mDis} = \frac{\sum_{i=0}^{m} \min\limits_{0 \ll j \ll n} \left(\sqrt{\left(x_i - x_j\right)^2 + \left(y_i - y_j\right)^2} \right)}{\min(m, n)} \tag{14}$$

where n represents the number of actual target centroids, m represents the number of detected target centroids, $\mathrm{x_i}$ represents the horizontal coordinate of the detected target centroid, $\mathrm{x_j}$ represents the horizontal coordinate of the real target centroid, $\mathrm{y_i}$ represents the vertical coordinate of the detected target centroid, and $\mathrm{y_j}$ represents the vertical coordinate of the real target centroid.

3.2. Experimental Results and Analysis

3.2.1. Experimental Setup

In this experiment, the input image size of the network is 256 × 256 pixels, and the batch size is set to 16. The training and testing dataset are divided approximately in a ratio of 5:1. The semantic segmentation subnet in the experiment utilizes Dice Loss as the loss function, while the centroid localization subnet employs Com Loss as the loss function. The optimization method for the loss function is Adaptive Gradient (AdaGrad), which is a first-order optimization method [29]. The advantage of using AdaGrad is that it allows the manual adjustment of learning rate to be replaced by setting hyperparameters. The hardware device mainly used in the experiment is an NVIDIA RTX 3090 graphics card with 24 GB of VRAM. The software programs primarily used include the deep learning framework PyTorch, the plotting tool Matplotlib, the image processing library OpenCV, and the scientific computing library NumPy.

3.2.2. Comparison with Other State-of-the-Art Methods

In this section, we evaluate our implementation by comparing it with traditional image processing method MLCM [30], classic semantic segmentation network U-Net [23], and the latest semantic segmentation networks DNAnet [24], HRNet [31], and MTU-Net [32]. We perform a quantitative and qualitative analysis of the experimental results.

(1) Quantitative Analysis

Table 1 compares the performance of our algorithm and the other five state-of-the-art algorithms in terms of evaluation metrics such as IoU, Precision, Recall, F1-score, and GFLOPs. The experimental results demonstrate that our method ranks first in all evaluation metrics.

Table 1. Comparison of our semantic segmentation algorithm with five other algorithms.

Semantic Segmentation Algorithm	IoU	Precision	Recall	F1-Score	GFLOPs
MLCM [30]	0.1810	0.1810	1.0000	0.3065	-
U-Net [23]	0.4905	0.6435	0.5774	0.6087	46.02
HRNetv2-w18 [31]	0.7216	0.8307	0.8443	0.8374	4.64
MTU-Net [32]	0.8712	0.9403	0.9229	0.9315	5.48
DNAnet [24]	0.8862	0.9566	0.9236	0.9398	14.05
UCDnet (ours)	0.9186	0.9673	0.9480	0.9576	15.56

(2) Qualitative Analysis

We visualized the semantic segmentation results of three classical algorithms and our UCDnet. As shown in Figure 11, yellow annotations represent false alarms, red annotations represent missed detections, green annotations represent correctly predicted targets, and blue annotations indicate zoomed-in views of the targets. Due to the influence of factors such as complex background in infrared images and noise resembling target shapes, the existing algorithms and network detection performance are unsatisfactory. From the prediction results, it can be observed that the traditional MLCM algorithm has significant limitations, with excessive false alarms and almost complete loss of target edge information. The U-Net misses some targets in particular situations but retains some edge information compared to traditional algorithms. The DNAnet's prediction results show a good segmentation of the edges, but still have some errors compared to the ground truth. However, it is evident that our algorithm's predicted target edges closely match the ground truth, demonstrating high robustness without false alarms or missed targets. Our algorithm outperforms other comparative approaches in terms of detection and segmentation performance.

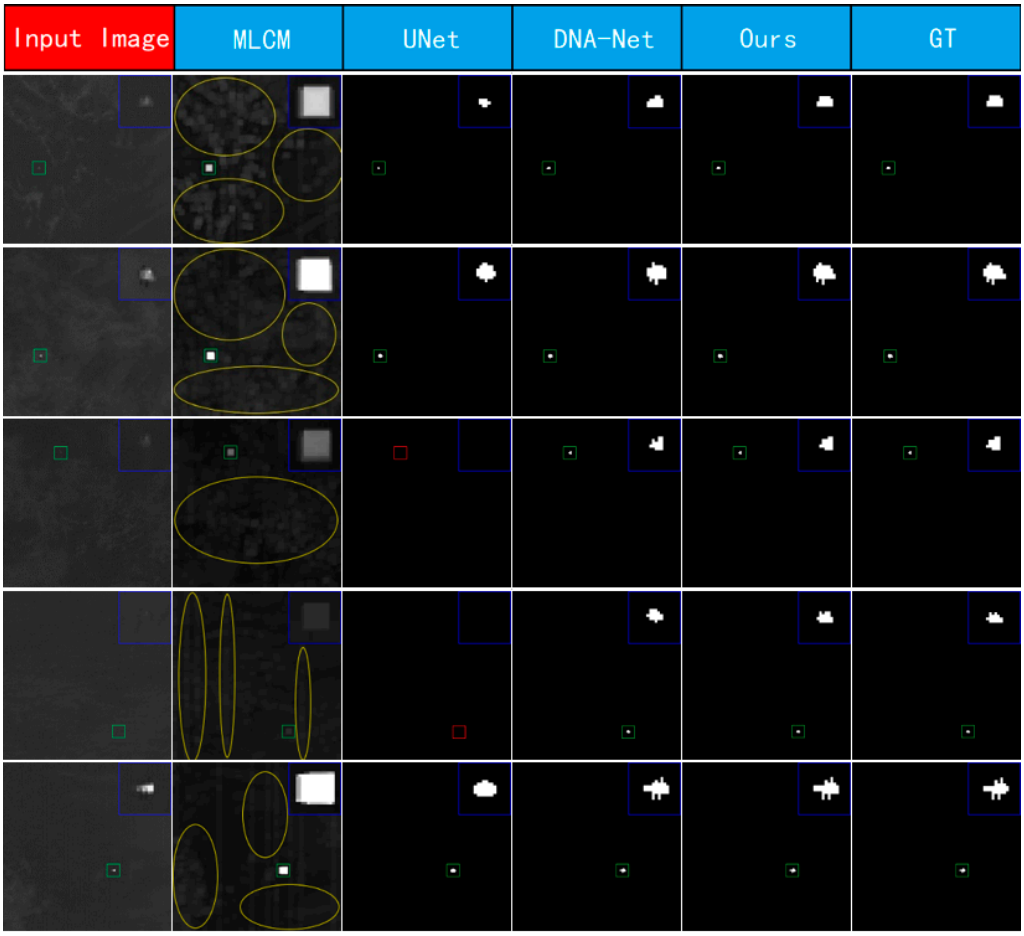


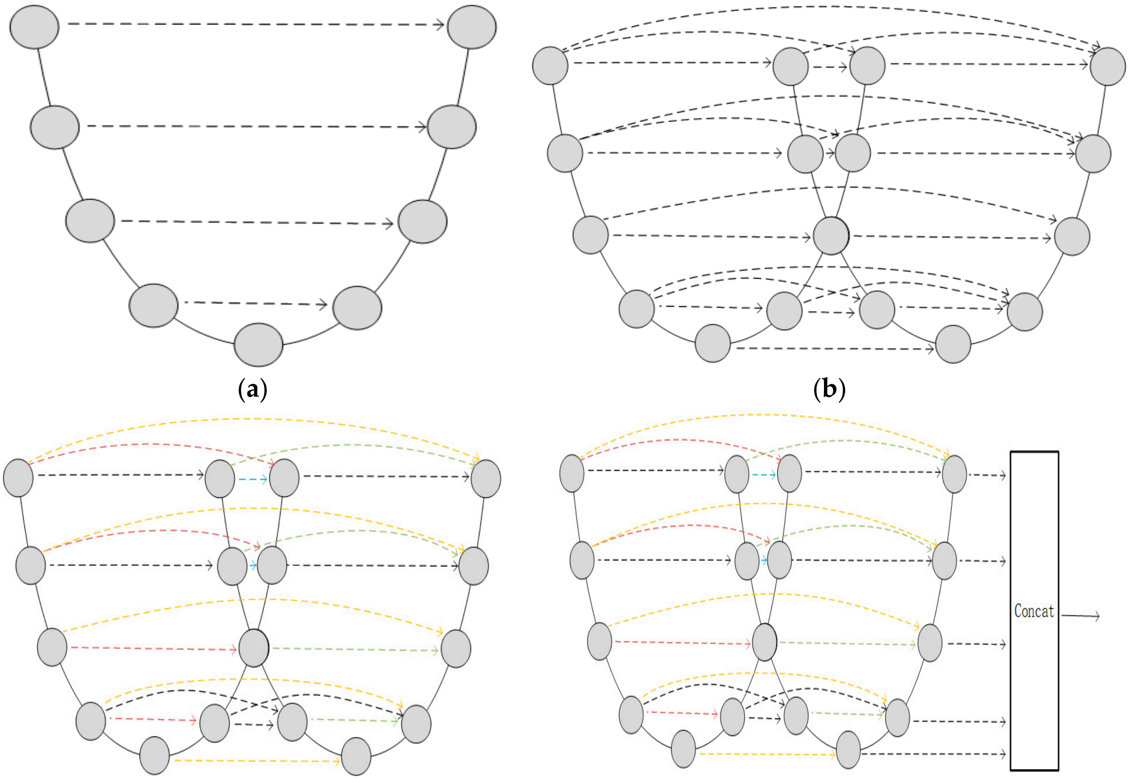
Figure 11. Comparison of our semantic segmentation results with the semantic segmentation results of three classical algorithms.

3.2.3. Analysis of UCDnet Feature Extraction Network

(1) Quantitative Analysis

To validate the superior performance of the proposed network structures in feature extraction compared to other similar network structures, we conducted the following ablation experiments on the main feature extraction network of the semantic segmentation subnet. Throughout the ablation experiments, we kept the structure of the centroid localization subnet unchanged and investigated the impact of variations in the structure of the semantic segmentation subnet on the network performance. As illustrated in Figure 12, "ablation study1" represents the network structure of a single U-shaped structure shown in Figure 12a, "ablation study2" represents a network structure similar to double U-shaped structure shown in Figure 12b, "ablation study3" represents a network structure based on Figure 12b with an added attention module shown in Figure 12c, and "ablation study4" represents a network structure with the addition of a feature fusion strategy based on Figure 12c. Training each of the four networks for 100 epochs yielded the results shown in Table 2 and Figure 13.

From Table 2, it can be observed that the double U-shaped network structure achieves higher IoU, Precision, and Recall compared to the single U-shaped network structure. By incorporating attention modules into the network, the IoU of the network in the test results improves by two percentage points, while Recall increases by three percentage points. Additionally, the introduction of a feature fusion strategy leads to an IoU improvement of over one percentage point. Considering the high IoU requirement for the output results in this study, the network structure depicted in Figure 12d is adopted.



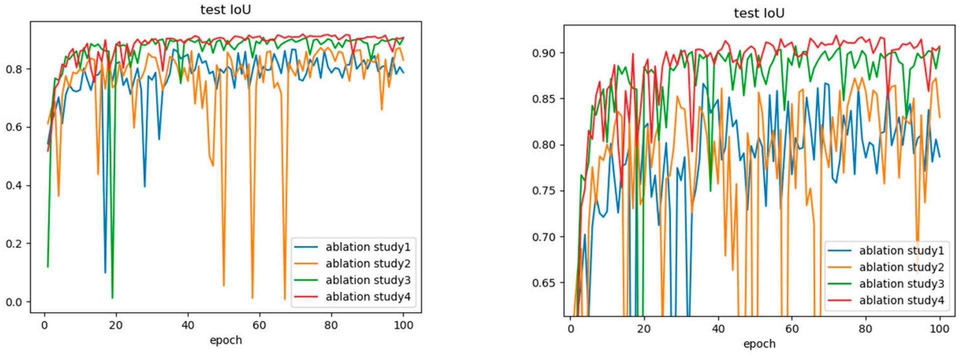
(**a**) (**b**)

(**c**) (**d**)

Figure 12. Architecture diagrams of the four networks in the ablation experiments, where (**d**) represents the network structure proposed in this paper. The figure (**a**) represents a single U-shaped structure without an attention module, the figure (**b**) represents a double U-shaped structure without an attention module, and the figure (**c**) represents a double U-shaped structure with an added attention module. The colored dotted lines in the figures indicate the presence of the attention module in the corresponding operation step.

Table 2. Comparison of results from four ablation studies.

Semantic Segmentation Algorithm	IoU	Precision	Recall
single U-shaped structure (ablation study1)	0.8665	0.9419	0.9155
double U-shaped structure (ablation study2)	0.8728	0.9600	0.9057
double U-shaped structure + attention module (ablation study3)	0.9070	0.9599	0.9428
double U-shaped structure + attention module + feature fusion (ours)	0.9186	0.9673	0.9480



(**a**) (**b**)

Figure 13. The curve depicting the variation of IoU for each ablation experiment across epochs. (**b**) presents a zoomed-in view of (**a**).

According to the curve shown in Figure 13, it can be observed that the network structure proposed in this paper consistently outperforms the other three ablation study network structures in terms of the evaluation metric IoU. Furthermore, the IoU values exhibit minimal fluctuations with increasing epochs. Considering the comprehensive set of ablation experiments, the proposed network structure in this paper demonstrates stronger feature extraction capabilities and better stability, while exhibiting superior learning ability for capturing detailed target features.

(2) Qualitative Analysis

From Figure 14, it can be observed that in the encoding process of the first U-shaped structure, as the downsampling iterations increase, the network pays more attention to the brighter regions in the image. In the decoding process, with the increase in upsampling iterations, the network can further distinguish between the target and the background. In the encoding process of the second U-shaped structure, FM (2,1) serves both as an encoding and decoding component, enabling more accurate discrimination between the background and the target. With an increase in downsampling iterations during the encoding process, FM (4,1) is able to fully differentiate the background regions surrounding the target. Comparing FM (0,2) and FM (0,3), although the rough shape of the target can already be observed at the output of the first U-shaped structure, the output of the second U-shaped structure is closer to the actual target.

3.2.4. Analysis of Centroid Localization Network

(1) Quantitative Analysis

As shown in Table 3, we compared the performance between three classical centroid localization methods and our method. Our approach demonstrated the best localization precision, outperforming even the suboptimal squared weighted centroid method by 0.2009 pixels.

Table 3. Comparison of four centroid localization methods.

Centroid Localization Methods	Localization Precision (Pixels)
center of bounding box	0.6096
gray value centroid method	0.5775
squared weighted centroid method	0.5380
UCDnet (Ours)	0.3371

The solution method for the gray value centroid method is as follows: For image block B, determine the center of the minimum bounding rectangle enclosing the target. Establish a Cartesian coordinate system with this center as the origin. Then, calculate the centroid using the following equations:

$$\text{m00} = \sum_{x,y \in B} I(x,y) \tag{15}$$

$$\text{m10} = \sum_{x,y \in B} x \times I(x,y) \tag{16}$$

$$\text{m01} = \sum_{x,y \in B} y \times I(x,y) \tag{17}$$

$$\text{C} = \left(\frac{m10}{m00}, \frac{m01}{m00} \right) \tag{18}$$

where $I(x,y)$ represents the pixel value at position (x, y), and C represents the coordinates of the centroid obtained on the Cartesian coordinate system established with the center of the image block.

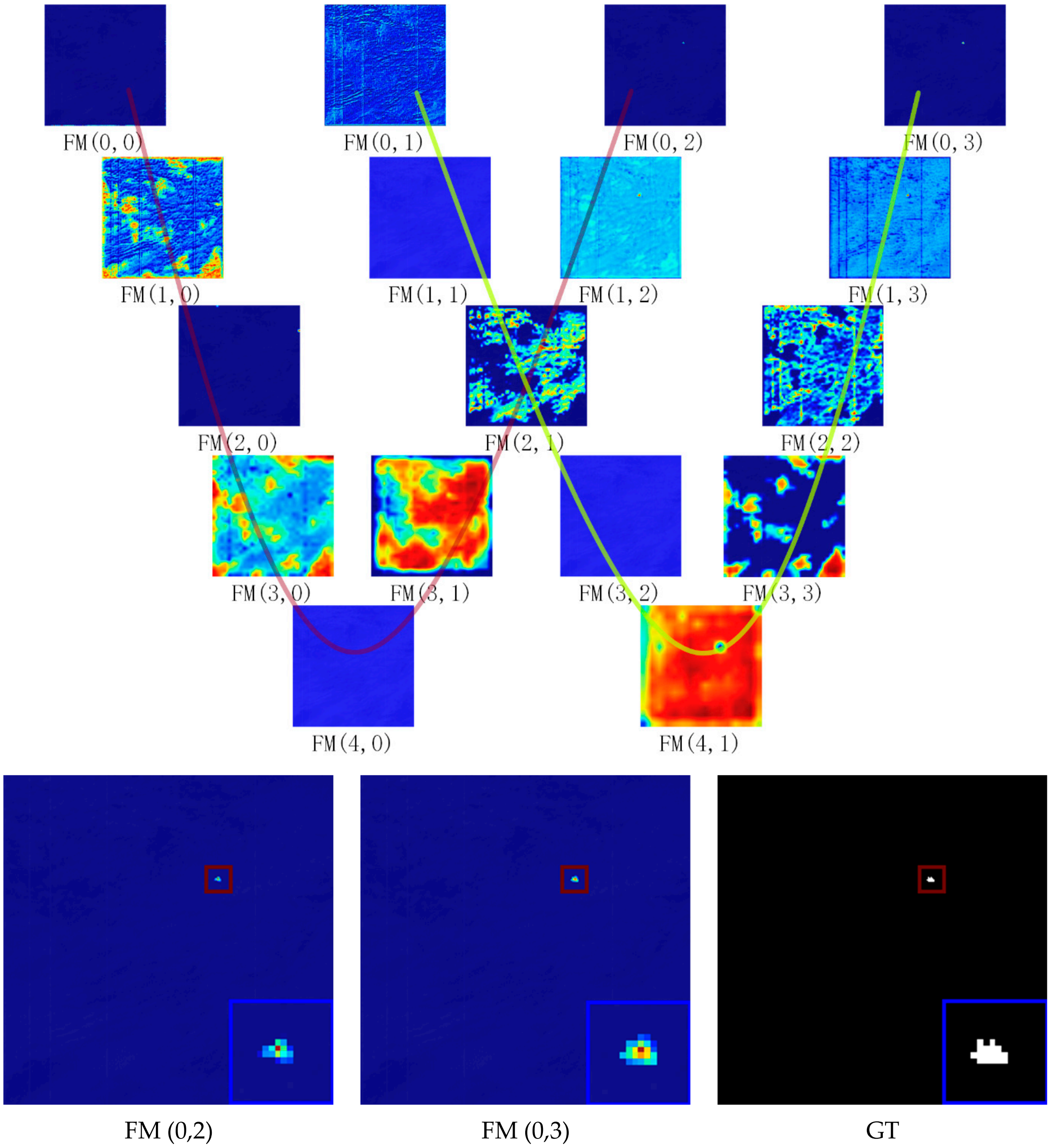


Figure 14. This is a heatmap generated using Grad-CAM. It can represent the contribution of each part of the network to the output.

The solution method for the squared weighted centroid method is as follows: For image block B, determine the center of the minimum bounding rectangle enclosing the target. Establish a Cartesian coordinate system with this center as the origin. Then, calculate the centroid using the following equations:

$$\mathrm{m00} = \sum_{x,y \in B} I^2_{(x,y)} \tag{19}$$

$$\mathrm{m10} = \sum_{x,y \in B} x \times I^2_{(x,y)} \tag{20}$$

$$\mathrm{m01} = \sum_{x,y \in B} y \times I^2_{(x,y)} \tag{21}$$

$$\mathrm{C} = \left(\frac{m10}{m00}, \frac{m01}{m00} \right) \tag{22}$$

where $I(x, y)$ represents the pixel value at position (x, y), and C represents the coordinates of the centroid obtained in the Cartesian coordinate system established with the center of the image block.

(2) Qualitative Analysis

We selected two images from the dataset and fed them into our UCDnet. We saved the final centroid localization results. Then, we extracted the target pixel blocks from the images, enlarged them eight times, and mapped our centroid localization results onto the magnified images (as shown in Figure 15). The centroid localization precision achieved sub-pixel-level precision.

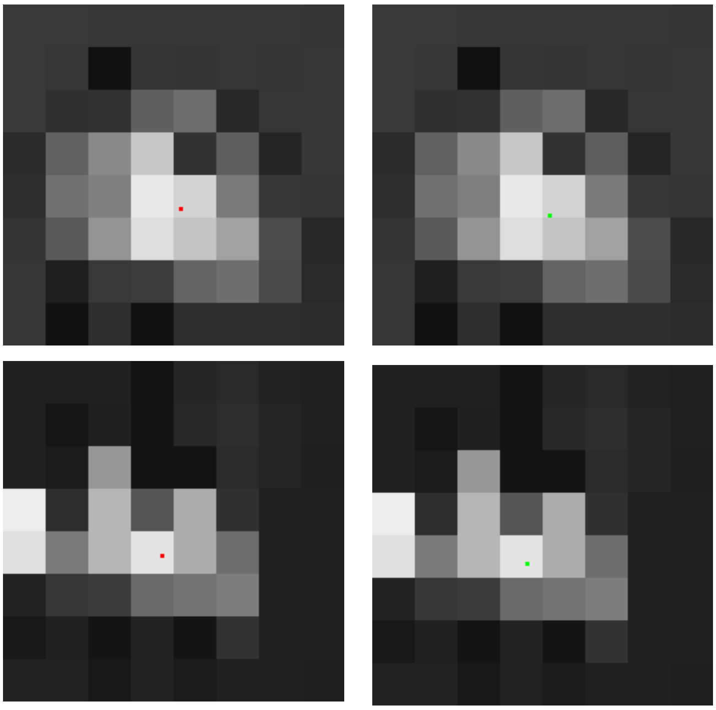


Figure 15. Centroid localization result image, with our detected results **on the left** and ground truth **on the right**. The red dots indicate the centroids of the detected targets, and the green dots represent the true positions of the centroids.

3.2.5. Discussion

Currently, numerous classical target detection algorithms have been applied in the field of infrared weak small target detection. We selected five representative algorithms, namely, MLCM, U-Net, HRNet, MTU-Net, and DNANet, for comparative experiments with our proposed method and obtained outstanding results. The IoU of our semantic segmentation results reached a leading level because of the proposed double U-shaped structure with attention mechanism. Furthermore, we innovatively introduced a deep learning-based approach for sub-pixel-level centroid localization, significantly improving the centroid localization precision compared to existing methods. Extensive experiments validated that our proposed algorithm exhibits favorable performance and robustness.

Additionally, to verify the applicability and generalization ability of our proposed algorithm beyond the field of surveillance and reconnaissance, we evaluated our algorithm on three different publicly available datasets from diverse fields. On an industrial field dataset for motor magnetic tile defect detection, our algorithm demonstrated excellent segmentation results even when facing weak texture information of motor magnetic defects (see experimental results in the first row of Figure A2 in Appendix A). On a medical field dataset for cell semantic segmentation [33,34], our algorithm accurately extracted edges of irregularly shaped cells (see experimental results in the second row of Figure A2 in Appendix A). On an agricultural field dataset for olive fruit semantic segmentation, our algorithm achieved precise segmentation for small-sized olive fruits, remaining unaffected by factors like crowding and occlusion (see experimental results in the third row of Figure A2 in Appendix A). These supplementary experiments further demonstrate the high robustness of our proposed algorithm. Our algorithm proves to be effective not only on our simulated dataset but also on various datasets from other fields, showcasing its excellent performance.

The proposed algorithm in this paper also has some limitations: (1) Point targets in infrared images are prone to confusion with high-frequency noise and similar objects. The method proposed in this paper focuses on fine segmentation and localization but requires further improvement in false alarm removal. (2) Infrared weak small targets have low signal-to-noise ratio and can easily be overwhelmed by background clutter. The algorithm in this paper may experience missed detections when the target signal-to-noise ratio is below 1. This is unacceptable for applications such as security monitoring and autonomous driving, which demand high safety requirements. These limitations will be the focus of our future work.

4. Conclusions

This paper presents a novel method called UCDnet for infrared weak small target detection and centroid localization. The innovation of our proposed semantic segmentation subnet lies in the improved U-Net and DNAnet structures, with the integration of attention modules. The double U-shaped backbone feature extraction network in our approach enables more accurate segmentation of target edges, while the addition of attention modules improves the capturing of target positional information. The innovation of our centroid detection subnet lies in the ability to overcome the constraint of unit pixel size in the original image, achieving sub-pixel-level centroid localization and minimizing the difference between predicted and ground truth centroid positions. Extensive comparative experiments demonstrate the superiority of our proposed semantic segmentation and centroid localization methods in terms of detection precision and robustness compared to existing mainstream methods.

In the future, we plan to expand our research in three directions: (1) How to use the motion information of targets in sequential images to reduce false alarms. The algorithm proposed in this paper is based on single-frame image for target detection, which may encounter challenges in removing false alarms in certain low-quality images or complex scenes. We will explore the temporal characteristics of targets in sequential images, combining spatial and temporal features to improve target discriminability. (2) How to construct an integrated network for enhancement and detection to address the issue of missed detections when the target is extremely weak. In our future work, we will embed more efficient enhancement modules or super-resolution reconstruction modules into the detection network to enhance target features and capture target details, aiming to achieve or even surpass the human eye's detection limit. (3) How to utilize multi-source image information fusion for handling weak small targets [35]. The infrared images used in this paper are obtained from thermal radiation imaging. Due to the technological limitations of sensors, issues such as loss of target texture information, high noise, and low signal-to-noise ratio may exist. In the future, we will develop a weak small target detection network based on multi-source data, incorporating data from multiple platforms and payloads to achieve stable and continuous detection and localization capabilities.

Author Contributions: Methodology, X.X., J.W., M.Z., H.S. and S.C.; Software, X.X., Z.W., Y.W. and S.L.; Writing—original draft, X.X., Z.W., Y.W., S.C. and S.L.; Writing—review & editing, J.W., M.Z. and H.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Science and Technology Department of Jilin Province, China, under grant number 20210201137GX.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

In Appendix A, we provide additional technical details and more experimental results.

Appendix A.1. Simulation Details of Infrared Weak Small Target Dataset

The production process of the proposed infrared weak small target dataset is shown in Figure A1. The overall steps include simulating background images, adding noise, de-signing target motion trajectories, adding targets to the background, randomly cropping image patches with targets, and generating semantic segmentation labels.

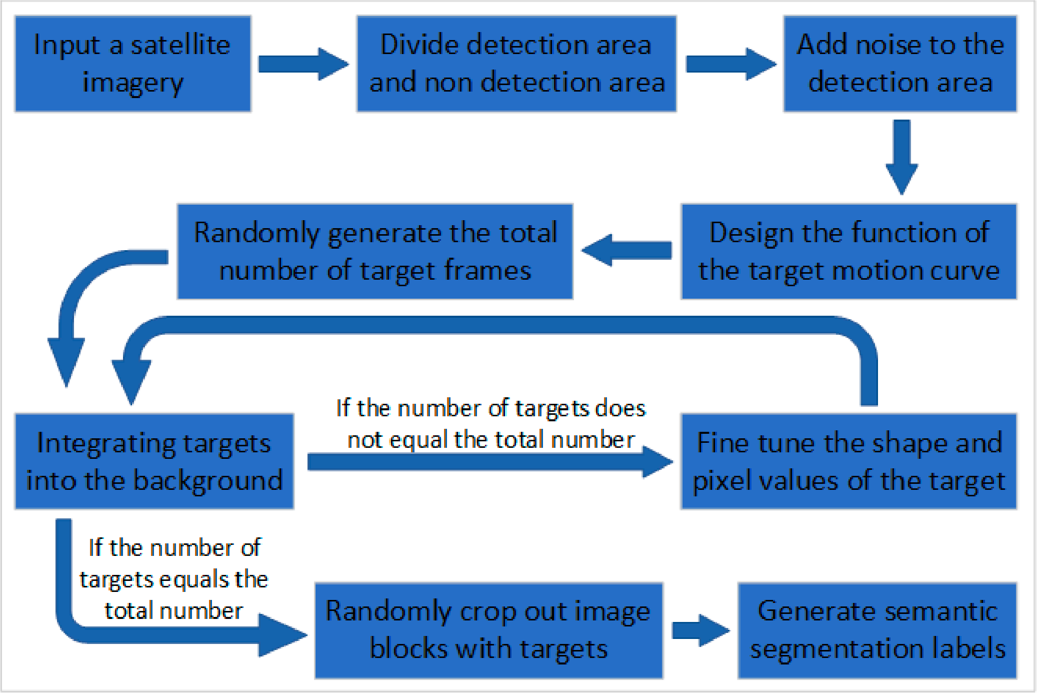

Input a satellite imagery
Divide detection area and non detection area
Add noise to the detection area
Design the function of the target motion curve
Randomly generate the total number of target frames
If the number of targets does not equal the total number
Integrating targets into the background
Fine tune the shape and pixel values of the target
If the number of targets equals the total number
Randomly crop out image blocks with targets
Generate semantic segmentation labels

Figure A1. The schematic diagram of the simulation process for the proposed dataset of infrared weak small targets in this paper.

Appendix A.2. Details of Network Architectures

In Table A1, we list the details of the Res_Attention block, Downsample block, Upsample_n block, and Multiply block used in our proposed network. These details include kernel size, stride, padding, upsampling and downsampling ratios, among others.

Table A1. Constituent elements of each block in the proposed network architecture.

Block	Layer	Input Channels	Output Channels	Kernel Size	Stride	Padding
Res_Attention	Conv2d	in_channels	out_channels	3	1	1
	BatchNorm2d	out_channels	out_channels	-	-	-
	ReLU	-	-	-	-	-
	Conv2d	out_channels	out_channels	3	1	1
	BatchNorm2d	out_channels	out_channels	-	-	-
	CBAM	out_channels	out_channels	-	-	-
	ReLU	-	-	-	-	-
Downsample	MaxPool2d	in_channels	in_channels	2	2	-
Upsample_n	Upsample	in_channels	in_channels	-	n	-
Multiply	-	1, 1	1	-	-	-

Appendix A.3. Experimental Results in Other Field

In Figure A2, we demonstrate the semantic segmentation results of our UCDnet in the industrial field (top row), medical field (middle row), and agricultural field (bottom row). The results reveal the good performance of our algorithm in these fields.

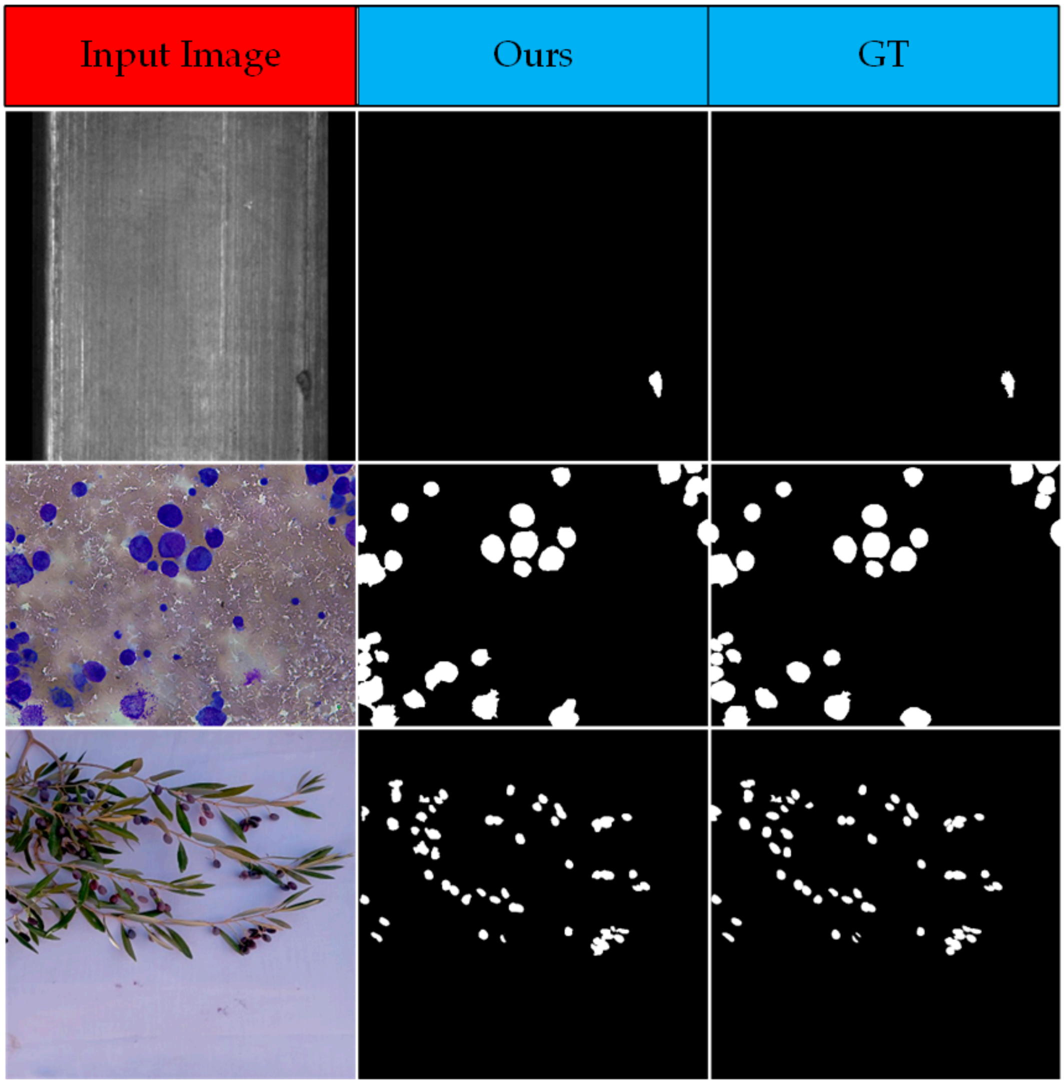


Figure A2. The first row in the figure shows the semantic segmentation results of motor magnetic tile defects, **the second row** presents the semantic segmentation results of cells, and **the third row** exhibits the semantic segmentation results of olive fruits.

References

1. Chen, G.; Wang, H.T.; Chen, K.; Li, Z.J.; Song, Z.D.; Liu, Y.L.; Chen, W.K.; Knoll, A. A Survey of the Four Pillars for Small Object Detection: Multiscale Representation, Contextual Information, Super-Resolution, and Region Proposal. *IEEE Trans. Syst. Man Cybern.-Syst.* **2022**, *52*, 936–953. [CrossRef]
2. Sun, X.L.; Guo, L.C.; Zhang, W.L.; Wang, Z.; Yu, Q.F. Small Aerial Target Detection for Airborne Infrared Detection Systems Using LightGBM and Trajectory Constraints. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **2021**, *14*, 9959–9973. [CrossRef]
3. Sun, Y.; Yang, J.G.; An, W. Infrared Dim and Small Target Detection via Multiple Subspace Learning and Spatial-Temporal Patch-Tensor Model. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 3737–3752. [CrossRef]
4. Liu, C.; Xie, F.Y.; Dong, X.M.; Gao, H.X.; Zhang, H.P. Small Target Detection From Infrared Remote Sensing Images Using Local Adaptive Thresholding. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **2022**, *15*, 1941–1952. [CrossRef]
5. Yu, W.B.; Liu, C.; Yang, H.T.; Wang, G.X. A method for improving the detection accuracy of subpixel edge. In Proceedings of the Chinese Automation Congress (CAC), Hangzhou, China, 22–24 November 2019; pp. 158–162.
6. Kosari, A.; Sharifi, A.; Ahmadi, A.; Khoshsima, M. Remote sensing satellite's attitude control system: Rapid performance sizing for passive scan imaging mode. *Aircr. Eng. Aerosp. Technol.* **2020**, *92*, 1073–1083. [CrossRef]

7. Li, Y.S.; Li, Z.Z.; Zhang, C.; Luo, Z.F.; Zhu, Y.; Ding, Z.Q.; Qin, T.Q. Infrared maritime dim small target detection based on spatiotemporal cues and directional morphological filtering. *Infrared Phys. Technol.* **2021**, *115*, 19. [CrossRef]
8. Li, Y.S.; Li, Z.Z.; Shen, Y.; Li, J. Infrared Small Target Detection Based on 1-D Difference of Guided Filtering. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 5. [CrossRef]
9. Zhang, L.; Lin, Z. Infrared Small Target Detection Based on Anisotropic Contrast Filter. In Proceedings of the 2020 IEEE 5th International Conference on Signal and Image Processing (ICSIP), Nanjing, China, 23–25 October 2020.
10. Han, J.H.; Ma, Y.; Huang, J.; Mei, X.G.; Ma, J.Y. An Infrared Small Target Detecting Algorithm Based on Human Visual System. *IEEE Geosci. Remote Sens. Lett.* **2016**, *13*, 452–456. [CrossRef]
11. Yang, L.N.; Jia, B.; Liu, Q. Adaptive Small Target Detection Based on Least Squares and Human Visual System. In Proceedings of the IEEE 17th International Conference on Communication Technology (ICCT), Chengdu, China, 27–30 October 2017; pp. 1565–1568.
12. Qiang, W.; Liu, H.K. An Infrared Small Target Fast Detection Algorithm in the Sky Based on Human Visual System. In Proceedings of the 4th Annual International Conference on Network and Information Systems for Computers (ICNISC), Wuhan, China, 20–22 April 2018; pp. 176–181.
13. Faramarzi, I.; Han, J.H.; Chen, Y.Y. Infrared Dim and Small Targets Detection Based on Multi-scale Local Contrast Measure Utilizing Efficient Spatial Filters. In Proceedings of the 6th International Conference on Signal Processing and Intelligent Systems (ICSPIS), Sadjad Univ, Mashhad, Iran, 23–24 December 2020.
14. Han, J.H.; Moradi, S.; Faramarzi, I.; Liu, C.Y.; Zhang, H.H.; Zhao, Q. A Local Contrast Method for Infrared Small-Target Detection Utilizing a Tri-Layer Window. *IEEE Geosci. Remote Sens. Lett.* **2020**, *17*, 1822–1826. [CrossRef]
15. He, Y.J.; Li, M.; Zhang, J.L.; An, Q. Small infrared target detection based on low-rank and sparse representation. *Infrared Phys. Technol.* **2015**, *68*, 98–109. [CrossRef]
16. Wei, H.Y.; Tan, Y.H.; Lin, J. Robust Infrared Small Target Detection Via Temporal Low-rank and Sparse Representation. In Proceedings of the 3rd International Conference on Information Science and Control Engineering (ICISCE), Beijing, China, 8–10 July 2016; pp. 583–587.
17. Liu, M.; Du, H.Y.; Zhao, Y.J.; Dong, L.Q.; Hui, M. Image Small Target Detection based on Deep Learning with SNR Controlled Sample Generation. In Proceedings of the 2nd International Conference on Computer Science and Mechanical Automation (CSMA), Wuhan, China, 10–12 November 2016; pp. 211–220.
18. Dai, Y.M.; Wu, Y.Q.; Zhou, F.; Barnard, K. Asymmetric Contextual Modulation for Infrared Small Target Detection. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), Electr Network, Waikoloa, HI, USA, 5–9 January 2021; pp. 949–958.
19. Hou, Q.Y.; Wang, Z.P.; Tan, F.J.; Zhao, Y.; Zheng, H.L.; Zhang, W. RISTDnet: Robust Infrared Small Target Detection Network. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 5. [CrossRef]
20. Zhou, X.Y.; Zhang, Y.; Hu, Y. Infrared Small Target Detection Via Learned Infrared Patch-Image Convolutional Network. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Kuala Lumpur, Malaysia, 17–22 July 2022; pp. 867–870.
21. Zhao, B.; Wang, C.P.; Fu, Q.; Han, Z.S. A Novel Pattern for Infrared Small Target Detection With Generative Adversarial Network. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 4481–4492. [CrossRef]
22. Cao, S.Y.; Lian, R.Y.; Zhang, Y.W.; Wu, F.Y.; Peng, Z.M. Infrared dim target detection via hand-crafted features and deep information combination. In Proceedings of the 17th Conference on Industrial Electronics and Applications (ICIEA), Chengdu, China, 16–19 December 2022; pp. 632–635.
23. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Munich, Germany, 5–9 October 2015; pp. 234–241.
24. Li, B.Y.; Xiao, C.; Wang, L.G.; Wang, Y.Q.; Lin, Z.P.; Li, M.; An, W.; Guo, Y.L. Dense Nested Attention Network for Infrared Small Target Detection. *IEEE Trans. Image Process.* **2023**, *32*, 1745–1758. [CrossRef] [PubMed]
25. Hu, H.G.; Zheng, Y.X.; Zhou, Q.W.; Xiao, J.; Chen, S.Y.; Guan, Q. MC-Unet: Multi-scale Convolution Unet for Bladder Cancer Cell Segmentation in Phase-Contrast Microscopy Images. In Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM), San Diego, CA, USA, 18–21 November 2019; pp. 1197–1199.
26. Zhao, Z.; Chen, K.; Yamane, S. CBAM-Unet++:easier to find the target with the attention module "CBAM". In Proceedings of the 2021 IEEE 10th Global Conference on Consumer Electronics (GCCE), Las Vegas, NY, USA, 12 October 2021.
27. Zhou, X.; Wang, D.; Krähenbühl, P. Objects as Points. *arXiv* **2019**, arXiv:1904.07850.
28. Oakley, J.P. Statistical properties of local extrema in two-dimensional Gaussian random fields. *IEEE Trans. Signal Process.* **1998**, *46*, 130–140. [CrossRef]
29. Fang, J.K.; Fong, C.M.; Yang, P.; Hung, C.K.; Lu, W.L.; Chang, C.W. AdaGrad Gradient Descent Method for AI Image Management. In Proceedings of the 7th IEEE International Conference on Consumer Electronics—Taiwan (ICCE-Taiwan), Taoyuan, Taiwan, 28–30 September 2020.
30. Heydarian, M.; Doyle, T.E.; Samavi, R. MLCM: Multi-Label Confusion Matrix. *IEEE Access* **2022**, *10*, 19083–19095. [CrossRef]

31. Sun, K.; Xiao, B.; Liu, D.; Wang, J.D.; Soc, I.C. Deep High-Resolution Representation Learning for Human Pose Estimation. In Proceedings of the 32nd IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 5686–5696.
32. Wu, T.H.; Li, B.Y.; Luo, Y.H.; Wang, Y.Q.; Xiao, C.; Liu, T.; Yang, J.G.; An, W.; Guo, Y.L. MTU-Net: Multilevel TransUNet for Space-Based Infrared Tiny Ship Detection. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 15. [CrossRef]
33. Cui, Z.M.; Li, C.J.; Chen, N.L.; Wei, G.D.; Chen, R.N.; Zhou, Y.F.; Shen, D.G.; Wang, W.P. TSegNet: An efficient and accurate tooth segmentation network on 3D dental model. *Med. Image Anal.* **2021**, *69*, 12. [CrossRef] [PubMed]
34. Hossain, M.S. Microc alcification Segmentation Using Modified U-net Segmentation Network from Mammogram Images. *J. King Saud Univ.-Comput. Inf. Sci.* **2020**, *32*, 1225–1226. [CrossRef]
35. Ghaderizadeh, S.; Abbasi-Moghadam, D.; Sharifi, A.; Tariq, A.; Qin, S.J. Multiscale Dual-Branch Residual Spectral-Spatial Network With Attention for Hyperspectral Image Classification. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **2022**, *15*, 5455–5467. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.