공학석사학위논문

Infrared Small Target Detection Using Attention Multiscale Feature Fusion U-Net

주의집중 멀티 스케일 특징 융합 U-Net을 이용한 적외선 소형 표적 탐지 기법

2023년 8월

서울대학교 대학원

항공우주공학과

정 원 영

Infrared Small Target Detection Usion Attention Multiscale Feature Fusion U-Net

주의집중 멀티 스케일 특징 융합 U-Net을 이용한 적외선 소형 표적 탐지 기법

지도교수 박찬국

이 논문을 공학석사 학위논문으로 제출함

2023 년 06 월

서울대학교 대학원

항공우주공학과

정원영

정원영의 공학석사 학위논문을 인준함

2023 년 06 월

위원장 김유단 (인)

부위원장 박찬국 (인)

위원 김현진 (인)

Abstract

Infrared Small Target Detection Using Attention Multiscale Feature Fusion U-Net

Won Young Chung
Department of Aerospace Engineering
The Graduate School
Seoul National University

In order to improve detection performance in a U-Net-based Infrared Small Target Detection(IRSTD) algorithms, it is crucial to fuse low-level and high-level features. Conventional algorithms perform feature fusion by adding a convolutional layer to the skip pathway of the U-net and by connection the skip connection densely. However, with the added convolutional operation, the number of parameters of the network increase, hence the inference time increases accordingly. Therefore, in this paper, a UNet3+ based full-scale skip connection U-Net is used as a based network to lower the computational cost by fusing the feature with a small number of parameters. Moreover, this paper propose an effective encoder and decoder structure for improved IRSTD performance. A residual attention block is applied to each layer of the encoder for effective feature extraction. As for the decoder, a residual attention block is applied to the feature fusion section to effectively fuse the hierarchical information obtained from each layer. In addition, learning is performed through full-scale deep supervision to reflect all the information obtained from each layer. The proposed algorithm, coined Attention Multiscale feature Fusion U-Net(AMFU-Net), can hence guarantee effective target detection performance and a lightweight structure. (mIoU: 0.7512, FPS: 86.1)

Keywords: Infrared Small Target Detection(IRSTD), attention, multiscale feature fusion, UNet3+

Student number: 2021-26553

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

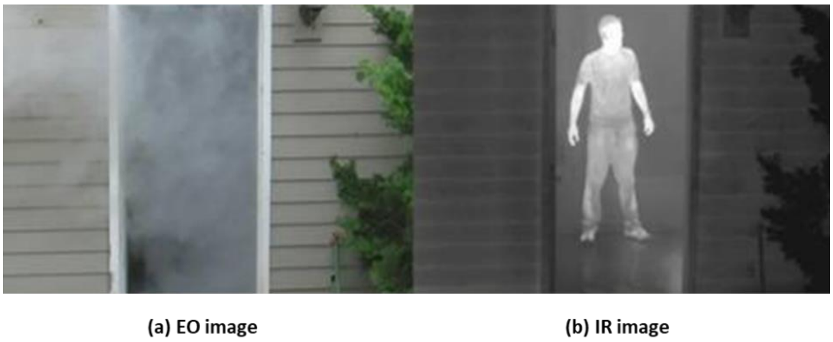

Figure 1.1 Infrared image

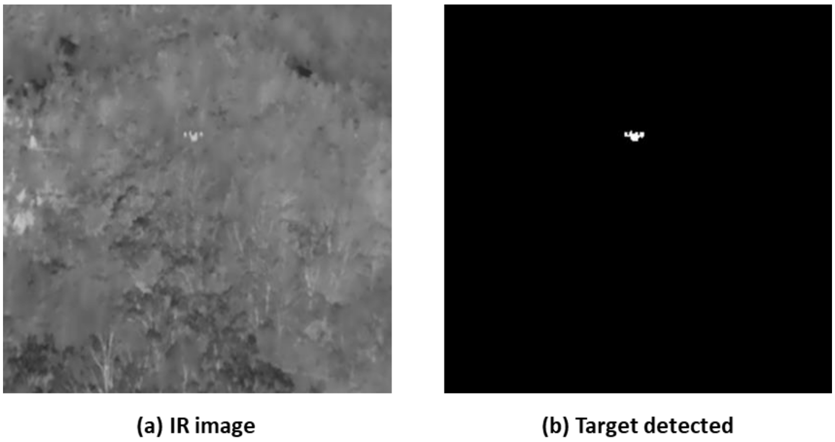

Figure 1.2 Infared small target detection result

Infrared(IR) images show robust characteristics for environmental factors compared to Electro-Optical (EO) images. Therefore, infrared image-based small target detection is widely applied to surveillance and reconnaissance systems, early warning systems, and remote sensing. However, IR small target detection has some challenging problems. First, it is difficult to effectively detect when the size of the target becomes too small. Second, IR images have harsh noise and interferences by the clutter around the target. Two types of approaches have been conducted to effectively perform IR small target detection in such challenging situations: 1) model-based and 2) data-driven [1].

First, model-based IR small target detection algorithms include image filter-based, local information-based, and data structure-based methods. The image filter-based methods, such as MaxMean, MaxMedian, and Top Hat filters, assume that the target is less static than the background and that adjacent pixels are highly correlated. However, this assumption is often not satisfied due to clutter and noise in practical uses. Local information-based methods consider the targets as locally salient and perform target detection through various local contrast filters. As a representative algorithm, Local Contrast Measurement (LCM) proposed by Chen et al. [2] obtains contrast information of fixed size window through a sliding window and performs detection using contrast difference between small targets and the background. The LCM algorithms has various variations, such as ILCM [3], RLCM [4], and HB-MLCM [5]. Gao et al. proposed Infrared Patch Image model (IPI) [6], a representative data structure-based method. Structural-based methods such as the IPI assume IR image as a combined model of background, target, and noise. Here, the background is assumed to be a low-rank matrix and the target as a sparse matrix. Based on these assumptions, small target detection is solved as an optimization problem that divides the combined matrix (IR image) into a low-rank matrix (background) and a sparse matrix (target). However, these model-based methods perform poorly when noise and clutter are severe or when the background of the image is complex.

Second, various data-driven methods have been recently proposed to overcome the abovementioned limitations of model-driven methods. Wang et al. [7] designed a GAN-based IRSTD network that lowers miss detection and false alarms through a two-path conditional Generative Adversarial Network (cGAN) of two generators and one discriminator. Dai et al. [8] proposed Asymmetric Contextual Modulation (ACM), a feature fusion module that can be used for various IRSTD network structures. Li et al. [9] proposed DNA-net, a network using dense nested U-net (UNet++) [10] with attention modules. The network performs effective small target detection by fusing feature maps obtained from each layer stages. Inspired by the DNA-net, we designed an IRSTD network using attention modules and UNet3+ [11], a U-net-based network which fuses features through a full-scale skip connection between the encoder and the decoder without using dense convolution.

1.2 Objectives and contributions

In this thesis, the lightweight and effective IR small target detection algorithm is propsed. Through the use of UNet3+ architecture and the attention mechanisim, it is shown that the effective infrared small target detection performance can be achieved with a small number of parameters.

Since the proposed network structure is designed for infrared small target detection from IR images, this thesis verified its performance using the IRSTD open dataset.

The main contributions of this theis are givne as follows:

1) We propose Attention Multiscale Feature Fusion U-Net(AMFU-Net), a lightweight infrared small target detection network sturcutre based on UNet3+, which outperforms the state-of-the-art methods on mIoU and still achieves on-line inference speed on an embedded system.
2) A residual attention block-based encoder is proposed to ensure robust feature extraction.
3) A novel method of utilizinf resdiaul attention blocks in every decoder stage is implemented, improving performance of multiscale feature fusion.

Chapter 2

Related Works

2.1 Model based Infrared Small Target Detection

In this section, briefly review the major works in model-based infrared small target detection algorithms. Model-based algorithms can be divided into three categories, each consisting of filter-based methods, local contrast-based methods and data structure-based methods.

2.1.1 Filter-based methods

Various types of image filters are used in filter-based methods to detect dim targets in infrared images and facilitate their detection through highlighted images. One such method is the high-pass filter method, initially proposed by Peng and Zhou [12], which applies six high-pass filters to suppress the background clutter by utilizing the differences in gray values between the targets and backgrounds. The high-pass filter kernel possesses several characteristics, such as a sum of weights equal to zero, a large and positive weight near the center, and a small and negative weight near the edges. This enables the targets and isolated noise to pass through the kernel while suppressing the slowly changing background. However, the effectiveness of this method may be limited in complex scenes, as it performs better in scenarios with smooth backgrounds.

The Median filter-based method is widely used for detecting small targets in infrared images. It is a traditional nonlinear spatial filter that employs sorting and replacing techniques for edge preservation. Deshpande et al [13] subsequently introduced the MaxMean/MaxMedian filter method, which incorporated direction information. However, these Median filter-based algorithms are highly susceptible to performance degradation when dealing with background clutter. However, since these Median filter-based algorithms are sensitive to background clutter, if there are many clutter, performance degradation can be severe.

In 1998, Tomasi et al. [14] proposed a method for detecting small targets in infrared images using the bilateral filter approach, which comprehensively considers the relationship between space and intensity. The bilateral filter comprises two Gaussian filter kernels that assign higher weights near the center based on Euclidean distance and gray values. It smooths the image while preserving edges and exhibits desirable characteristics such as nonlinearity, non-iterativeness, and simplicity. However, users often set the two Gaussian filter kernels to constant standard deviations. To achieve adaptive adjustment of standard deviation, novel methods have been proposed. For instance, Bae et al. [15] employed the target similarity index threshold to determine whether a pixel was a target or not. At the same time, Arnold [16] combined the dual-window circular structure template with the bilateral filter to detect small targets.

2.1.2 Local contrast-based methods

The visual system encodes contrast as the most significant quantity, enabling small targets with high intensity in complex backgrounds to capture attention quickly. In light of these observations, several local contrast-based approaches have been developed for infrared small target detection. Generally, these algorithms operate in a sequential manner: firstly, a patch window is moved pixel by pixel from left to right and top to bottom of the original image; secondly, the patch window is divided into central and surrounding parts; thirdly, the saliency image is computed based on intensity differences among these parts; and finally, target segmentation with adaptive threshold is applied to the saliency image for obtaining the final detection outcome. The Local Contrast Measure (LCM) introduced by Chen et al. is the most representative local contrast-based method. LCM assumes that the target is brighter than its neighbors. To mitigate the high false alarm rate in infrared small target detection, Han et al. proposed an improved LCM (ILCM) using an adaptive contrast mechanism. However, the sliding window in ILCM should be approximated to the target, which is difficult to predict. Subsequently, Qin et al. [17] proposed a novel LCM, wherein the sliding window needs only to be larger than the target. Moreover, a relative LCM (RLCM) was introduced to detect targets of varying sizes.

2.1.3 Subspace structure-based methods

Subspace structure-based methods have been proposed to distinguish between targets and backgrounds based on their distinct structural characteristics. In infrared images, the background is highly correlated, while the target is perceived as a disruptor of this correlation. Therefore, the detection of small infrared targets can be accomplished by recovering the low-rank matrix. However, this is an NP-hard problem, and rank minimization is not always feasible. As a solution, the nuclear norm is commonly used as an alternative to the rank function.

The subspace structure-based methods consist of four steps. First, the original infrared image is divided into a sequence of local image patches using a sliding-window strategy. Each local image patch is vectorized as a column of a novel image. Subsequently, the background patch images and target patch images are obtained through diverse algorithms based on the characteristics of the low-rank background and sparse target. Next, the background image and target image are reconstructed from the corresponding patch images. Finally, the adaptive segmentation method is applied to obtain the detection result.

One of the most representative subspace structure-based methods is the Infrared Patch Image (IPI), proposed by Gao et al. This method extends the traditional infrared image model to an IPI model, which seeks the low-rank background subspace structure and sparse target structure. However, the sparsity measurement based on the L1-norm may result in the mis-detection of strong edges (false alarm). To address this issue, Dai et al. [18] proposed a weighted IPI (WIPI) model that allocates an adaptive weight to the target patch image. Nevertheless, the inaccurate estimation of the background patch image still remains a problem. To overcome this issue, a nonnegative IPI approach based on partial sum minimization of singular values was introduced [19]. Additionally, other methods, such as the re-weighted IPI model [20] and principal component pursuit (PCP)-based method [21], have also been proposed.

2.2 Deep learning-based Infrared Small Target Detection

In recent years, the field of computer vision has witnessed a remarkable progress in deep learning-based algorithms, which have been extensively applied to infrared image small target detection. Compared to traditional model-based methodologies, deep learning-based algorithms have demonstrated superior performance and have shown the capability to achieve high accuracy even in the presence of noise and cluttered images.

The main process of deep learning-based algorithms for small target detection in infrared images involves the preparation of data containing small targets to train the deep learning model. This data is typically acquired in real-world scenarios with small targets present. Subsequently, the deep learning model utilizes various methods to detect small targets in infrared images. This involves an encoding process that extracts feature maps from the input image through multiple convolutional and pooling layers. The encoded information is then used to detect small targets in the infrared image via dense connection layers and softmax functions or by decoding the encoded feature map.

Dai et al. proposed an asymmetric contextual modulation (ACM) module that can be applied to CNN-based networks for target detection. The ACM module effectively combines low-level features and high-level features obtained from CNN layers through bottom-up local attentional modulation and top-down global attentional modulation.

Wang et al. introduced a novel approach called MDvsFA-cGAN for image segmentation that departs from the conventional deep learning-based methods that rely on a single objective to minimize the overall segmentation error. The MDvsFA-cGAN method employs a conditional Generative Adversarial Network, which includes two generator models and one discriminator model, and decomposes the segmentation task into two subtasks. Through an adversarial training process, both generator models are optimized to minimize the miss detection and false alarm loss functions, respectively. This approach provides an alternative solution that can reduce the complexity of the model and network design by incorporating multiple loss functions.

Li et al. designed a Dense Nested Attention U-Net (DNA-Net) for infrared small target detection using U-Net++, which is a variation of the U-Net-based network. To effectively fuse low-level and high-level features, they incorporated convolution layers into the skip connection pathway of U-Net and employed the U-Net++ structure with additional skip connections and convolutions. While U-Net++ can merge features effectively with additional convolutions and skip connections, excessive convolutions may dilute the feature map information. Therefore, DNA-Net applied attention modules to each convolutional neural network to minimize information dilution and achieve state-of-the-art performance.

Chapter 3

Attention Multiscale Feature Fusion U-Net

3.1 U-Net like networks

In this section, we discuss the U-Net-based network. U-Net employs an encoder to extract features from images and a decoder to reconstruct the encoded information. The feature maps obtained from the encoder contain contextual information, carrying the positional information of targets within the image. The information obtained through decoding serves as localization information, representing the shape details of the targets. These two types of information are fused using skip connections in U-Net. The subsequent U-Net-like network structures, U-Net++ and U-Net3+, which will be described later, reconfigure the skip connections for effective information fusion.

3.1.1 U-Net

U-Net [22] is a neural network architecture proposed by Ronneberger et al. in 2015 for the purpose of image segmentation. U-Net employs an encoder-decoder architecture, where the encoder performs down-sampling on the input image using convolutional neural networks, and the decoder performs up-sampling on the encoded image using deconvolutional neural networks.

By passing the input image through the encoder of U-Net, which consists of multiple convolutional neural networks, the context information of the input image can be obtained. Subsequently, through up-sampling and deconvolution from the encoded feature map, the spatial resolution of the image is increased, and the detailed information of the image is recovered, allowing for the acquisition of localization information.However, if only simple up-sampling and deconvolution are performed, information loss can occur. This is because the dimension of the image decreases through the CNN layers, and up-sampling is performed using linear interpolation from the reduced-dimension image. To prevent such information loss, U-Net utilizes skip connections between the encoder and decoder. Skip connections involve concatenating feature maps from the same level of the encoder and decoder. By concatenating the feature map from the encoder, which retains information prior to the occurrence of information loss, with the feature map from the decoder, which has undergone up-sampling but may have information loss, the reduction of information loss can be achieved. Through this approach, U-Net ensures effective performance.

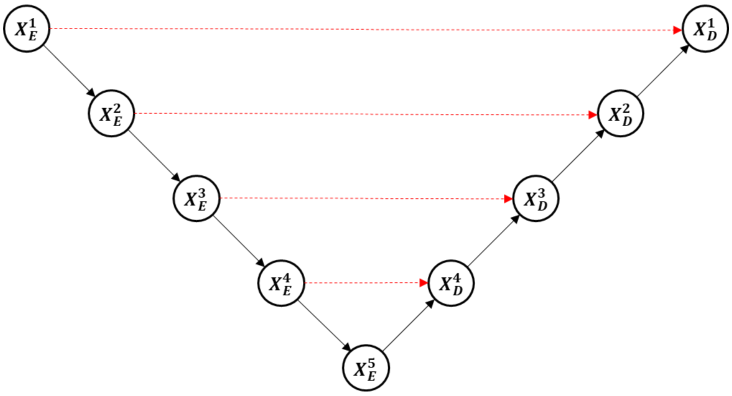

Figure 3.1 U-net structure

3.1.2 U-Net ++

U-Net++ is a network that shares a similar overall structure with U-Net but introduces changes to the skip connection architecture for effective feature fusion. In the original U-Net, feature maps from the same level are fused through skip connections, which allowed for excellent performance. However, this approach limits feature fusion across multiple layers.

To address this limitation, U-Net++ incorporates convolutional neural networks into the pathway of skip connections and modifies the skip connections to be dense. By using densely nested skip connections, U-Net++ consists of multiple sub U-Nets with varying depths, where each sub U-Net is connected through skip connections. This architecture enables the fusion of information obtained from multi-level layers. Consequently, U-Net++ performs more effective feature fusion compared to the original U-Net.

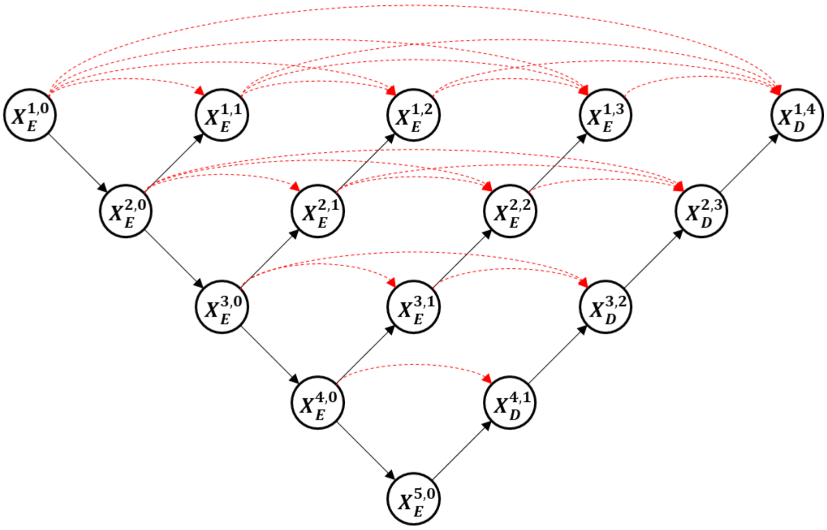

Figure 3.2 U-net++ structure

3.1.3 U-Net3+

UNet 3+ is a network architecture designed to ensure high performance similar to UNet++ while also maintaining a low number of parameters. UNet++ minimizes information loss in skip connections by adding multiple convolutional neural networks to the skip connection pathway and making the skip connections dense. However, this approach introduces a drawback of a significant increase in the number of parameters due to the addition of multiple convolutional neural networks.

To address this issue, UNet 3+ introduces the full-scale skip connection structure. The full-scale skip connection enables the fusion of multi-scale features without additional convolutional operations. It connects all layers of the encoder and decoder through skip connections, and additionally adds skip connections between decoders to incorporate relationships among decoders. This minimizes information loss during the up-sampling process in the decoder and allows for the generation of feature maps while minimizing information loss.

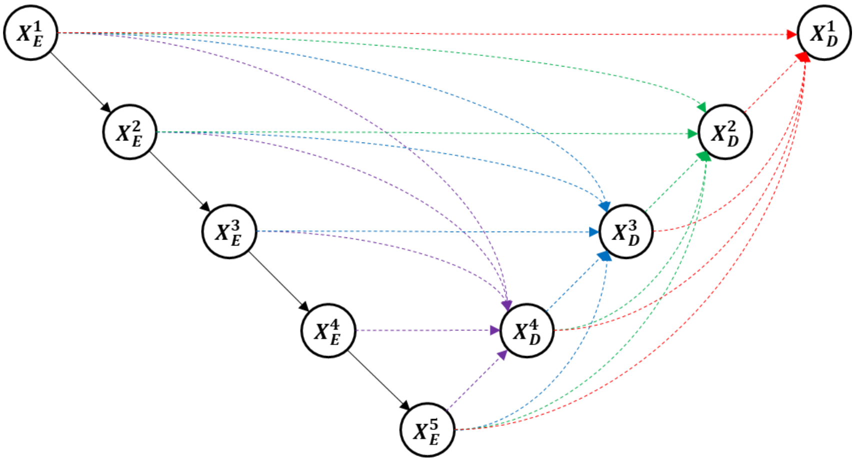

Figure 3.3 U-net3+ structure

3.1.4 Number of parameters

In thesis, we selected U-net3+ as the base network for IR small target detection. U-net3+ ensures effective detection performance while having fewer network parameters compared to U-net++. This results in lower power consumption and enables efficient operation on low-computing-power embedded systems.

In this section, an analysis of the number of parameters is presented for U-net, U-net++, and U-net3+ to explain how U-net3+ achieves a lower parameter numbers. When comparing the parameters of each network, U-net, U-net++, and U-net3+, we observe that they have the same number of parameters in the encoder since they share the same structure. However, the decoder parameters differ for each network.

The number of decoder parameters for U-net can be calculated as follows:

$$P^{i}_{U-De} = D_f \times D_f \times [d\left(X^{i+1}_{De}\right) \times d\left(X^{i}_{De}\right) + d\left(X^{i}_{De}\right)^2 \\ + d\left(X^{i}_{En} + X^{i}_{De}\right) \times d\left(X^{i}_{De}\right)] \tag{3.1}$$

In equation 3.1, D_f is the size of the convolution kernel, $\mathrm{d}(\cdot)$ is the depth of the nodes, X_{De} is the feature map from decoder, X_{En} is the feature map from encoder.

The number of decoder parameters for U-net++ can be calculated as follows:

$$P^{i}_{U^{++}-De} = D_f \times D_f \times [d\left(X^{i+1}_{De}\right) \times d\left(X^{i}_{De}\right) + d\left(X^{i}_{De}\right)^2 \\ + d\left(X^{i}_{En} + \sum_{k=1}^{N-1-i} X^{i,k}_{Me} + X^{i}_{De}\right) \times d\left(X^{i}_{De}\right)] \tag{3.2}$$

In equation (), D_f is the size of the convolution kernel, $\mathrm{d}(\cdot)$ is the depth of the nodes, X_{De} is the feature map from decoder, X_{En} is the feature map from encoder, and X_{Me} is the feature map from dense convolution section.

For UNet3+, feature maps from each decoder are derived from the number of channels in the first encoder, $N_{1st\_channel}$, and the scale N, yielding $N_{1st\_channel} \times N$ channels. Hence, the number of parameters of i^{th} decoder in UNet3+, $P^{i}_{U^{3+}-De}$ is calculated as:

$$P^{i}_{U^{3+}-De} = D_f \times D_f \times [(\textstyle\sum_{k=1}^{i} d(X^{k}_{En}) + \sum_{k=i+1}^{N} d(X^{k}_{de})) \times N_{1st\_channel} + d(X^{i}_{De})^{2}] \tag{3.3}$$

3.2 Residual attention block

In this section, we discuss the residual attention block, a network block applied in the proposed network. Neural networks have greatly contributed to improving accuracy in the field of artificial intelligence. There are various types of neural networks, with CNN (Convolutional Neural Network) being primarily used for image processing and RNN (Recurrent Neural Network) for natural language processing. While these neural networks yield excellent results, they are not without drawbacks. One of the challenges arises when the depth of the neural network increases, meaning the layers become deeper, making it difficult to train the network. This difficulty stems from the occurrence of gradient vanishing, where the gradient becomes effectively zero due to the continuous multiplication of derivatives of the neural network's activation functions. This renders gradient-based learning methods, such as backpropagation, ineffective. To address this issue, several techniques have been proposed, including improving activation functions and utilizing batch normalization. However, the most popular approach for mitigating the gradient vanishing problem involves modifying the network structure to incorporate skip connections between layers, as seen in ResNet (Residual Network). Section 3.2.1 of this paper explains the causes of gradient vanishing, while section 3.2.2 describes techniques used to prevent gradient vanishing.

3.2.1 Causes of gradient vanishing

Before understanding the issue of gradient vanishing, it is necessary to grasp how artificial neural networks learn. An artificial neural network is a layered structure consisting of nodes and connections between nodes, as illustrated in Figure 3.2. The connections between nodes are assigned weights, and the learning process of an artificial neural network involves continuously optimizing these weights to their optimal values through mathematical operations. To update the weights, a criterion is needed, which is defined by a cost function. Updating the weights based on the cost function means minimizing the difference between the output obtained by passing the input through the artificial neural network, where the weights and biases are summed and processed through activation functions, and the target or expected output. Since the cost function is often a high-dimensional equation that is difficult for us to comprehend, it can be represented graphically as shown in the figure3.3 and 3.4, with the parameters and the cost function as axes.

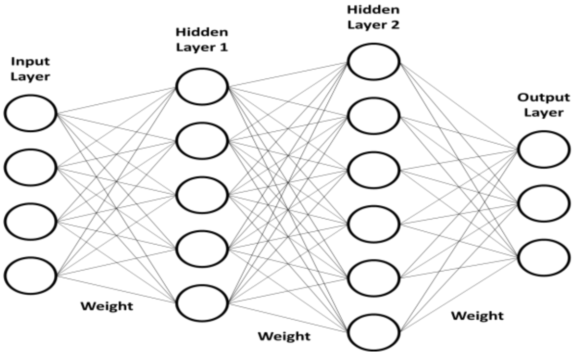

Figure 3.4 Fully connected layer

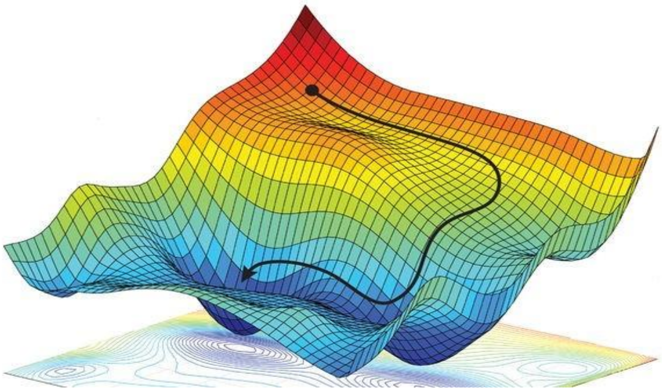

Figure 3.5 Cost function

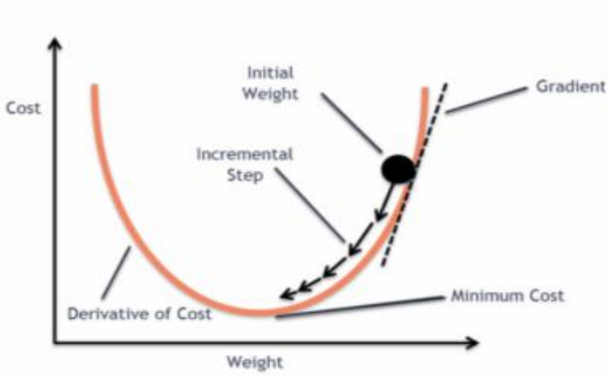

Figure 3.6 Cost function – 2D projection

Understanding such high-dimensional equations is difficult to grasp, and the direction to minimize the cost function can also be ambiguous. Therefore, gradient descent is used to determine this direction. Gradient descent is similar to descending a mountain while blindfolded. When descending a mountain blindfolded, since one cannot see the direction, they explore all the areas with lower slopes from their current position and move towards the direction of the lowest slope. By repeating this process multiple times, they eventually descend the mountain.

Applying this concept to gradient descent for neural network learning, the first step is to compute the derivatives of the cost function to find the direction of the lowest slope. Since the cost function is a high-dimensional equation, these derivatives consist of partial derivatives with respect to each variable, and this collection of partial derivatives is called the gradient. Taking a step in the direction of the lowest slope is equivalent to adjusting the parameters and biases towards the direction where the gradient of the cost function is minimized. By continuously updating the weights and biases, the neural network strives to minimize the error. This iterative process is referred to as neural network training.

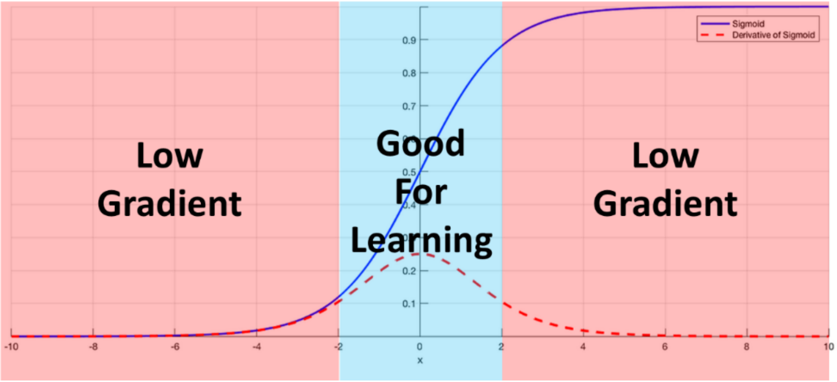

Figure 3.7 Sigmoid activation function

Figure3.2 represents an example of a deep learning network structure. With numerous weights and biases, updating all of them requires propagating the error values from the output layer to the input layer, updating the weights and biases along the way. This process is called error backpropagation. The backpropagation process involves continuously multiplying the partial derivatives of the activation functions as it progresses from the output layer towards the input layer. The vanishing gradient problem arises due to this structural issue. In the field of artificial intelligence, the sigmoid function has been widely used as an activation function. The reason for using the sigmoid function as an activation function is to mimic human learning processes. The initial artificial neural network, the perceptron, has a discontinuous structure where it sends a signal if the input exceeds a certain threshold and does not transmit a signal otherwise. Learning with such a discontinuous structure makes deep learning impossible.

Human learning is gradual and continuous. Similar to how humans gradually acquire knowledge over a long period of time, artificial neural networks introduced continuous activation functions to mimic this gradual learning process, and the sigmoid function was one of them. Using the sigmoid function allowed artificial neural networks to learn in a manner similar to humans. However, as the depth of artificial neural networks increased, the problem of gradient vanishing emerged. The figure3.5 represents the sigmoid function, and from the graph of the activation function, we can observe that the gradient is close to zero in most regions, and the region where learning is facilitated is not significantly wider than the other regions. In the process of error backpropagation, where the gradient of the activation function is continuously multiplied as the neural network goes through layers, the multiplied values are likely to be zero or very small. As a result, as the network gets deeper, the backpropagated error values converge to zero, preventing weight updates. In summary, as the network deepens, the learning of the artificial neural network is hindered. To solve more complex real-world problems, deeper artificial neural networks are needed. However, as the network becomes deeper, the problem of gradient vanishing arises, eventually preventing effective learning.

3.2.2 Prevent gradient vanishing problem

The problem of gradient vanishing occurs due to the structural issue in artificial neural networks where the error converges to zero as it is backpropagated towards the input layer. To address this problem, it is necessary to ensure that the error can be effectively transmitted to the input layer without loss. Nowadays, various approaches exist to tackle this problem, and they can be broadly categorized into two approaches.

The first approach is to change the activation function. By using activation functions that alleviate the gradient vanishing issue, such as Rectified Linear Unit (ReLU) or variants of it, the network can maintain a larger gradient flow during backpropagation, allowing for better learning in deep networks. The second approach is to modify the structure of the artificial neural network. One popular solution is the introduction of skip connections, as seen in Residual Neural Networks (ResNets), where shortcuts are added to allow direct paths for information flow between layers. This helps mitigate the vanishing gradient problem and enables effective learning in deeper networks.

By adopting these approaches, researchers have made significant progress in addressing the issue of gradient vanishing and enabling the training of deeper artificial neural networks.

3.2.2.1 Change activation function

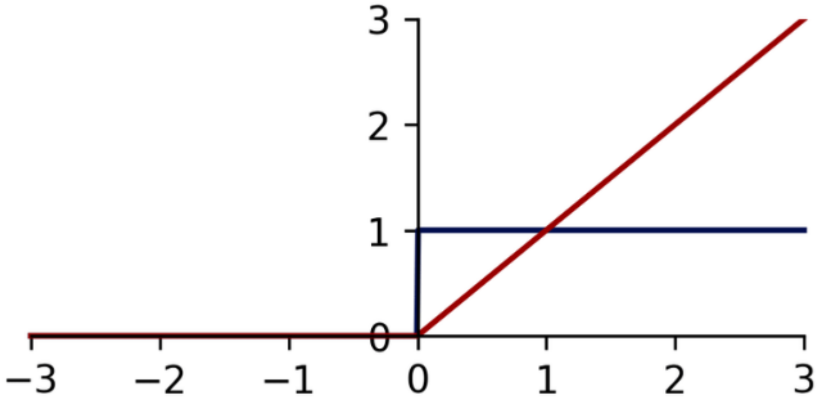

Figure 3.8 Rectified Linear Unit activation function

In order to minimize the gradient vanishing problem in deep artificial neural networks, the first approach is to change the activation function. The existing sigmoid function is replaced with the Rectified Linear Unit (ReLU) function, which is a very simple function. The ReLU function outputs 0 for negative input values and the input value itself for positive values, and it can be defined by the equation 3.4.

$$\mathrm{ReLU(x)} \triangleq \max(0, \mathrm{x}) \tag{3.4}$$

Replacing the sigmoid function with the ReLU function as the activation function in artificial neural networks brings two prominent advantages. Firstly, it significantly reduces the likelihood of encountering the gradient vanishing problem compared to the sigmoid function. The sigmoid function only takes values between 0 and 1, whereas the ReLU function outputs values from 0 to infinity. This means that during error backpropagation, there is a higher probability of multiplying by non-zero values, mitigating the risk of gradient vanishing.

$$\mathrm{sigmoid(x)} = \frac{1}{1+\mathrm{e}^{-\mathrm{x}}} \qquad (3.5)$$

The second advantage is that the ReLU function significantly reduces computational complexity compared to the sigmoid function. The sigmoid function, as expressed in equation 3.5, involves a fractional function with exponentials. On the other hand, the ReLU function takes a simple linear form for values greater than 0. When computing the gradients during backpropagation, the computational difference between these two functions is considerable. If the neural network is shallow and simple, the difference between sigmoid and ReLU may be negligible. However, in modern deep and complex neural networks, the lower computational complexity of the ReLU function becomes a significant advantage.

3.2.2.2 Using residual network

In a typical artificial neural network, the structure consists of nodes in each layer connected sequentially. This sequential structure makes it challenging for the neural network to overcome the vanishing gradient problem as it gets deeper. Not only does the error need to pass through multiple layers to reach the input layer, but if the network has small weights, effective learning becomes difficult. To address the gradient vanishing issue caused by this sequential structure, the concept of skip connections has been utilized in artificial neural networks.

Skip connections, also known as shortcut connections or residual connections, introduce additional connections that bypass one or more layers in the network. By allowing the gradient to flow directly from a later layer to an earlier layer, skip connections provide an alternative path for information and gradients to propagate through the network. This helps mitigate the vanishing gradient problem and allows for more effective training of deep neural networks.

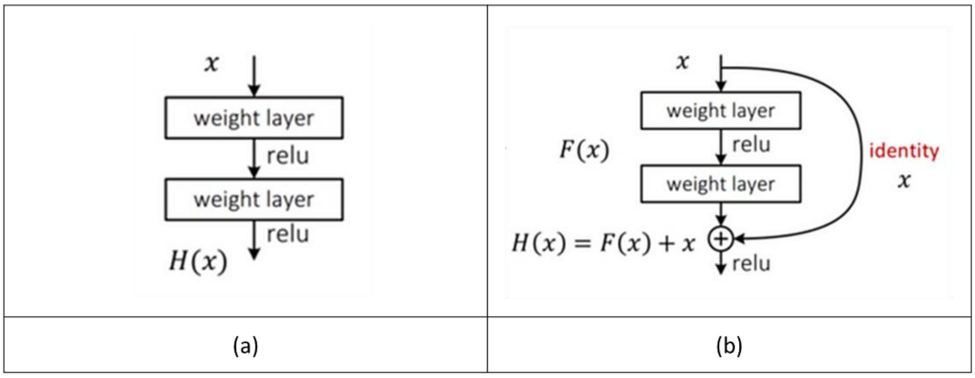

Figure 3.9 Network structure (a) plain network (b) residual network

The existing structure of an artificial neural network, as depicted in figure 3.7-(a), aims to learn H(x) for the input x. On the other hand, the neural network with skip connections, represented by Figure 3.7-(b), has a structure that optimizes F(x) by including the identity mapping (residual) of x in the target H(x) for the input x. The skip connections allow the residual to be passed along the network, enabling more efficient learning. Residual networks can be represented by equations such as equation 3.6 and equation 3.7.

$$y_1 = h(x_l) + \mathcal{F}(x_l, W_l) \tag{3.6}$$

$$x_{l+1} = f(y_l) \tag{3.7}$$

In the equations, x_l and x_{l+1}represent the input and output of the l-th unit, respectively, while $h(\cdot)$ represents the function that provides the skip connection. $\mathcal{F}(\cdot)$ represents the residual function, and $f(\cdot_l)$ represents the activation function, which can be represented by the ReLU function.

In order to demonstrate the ability of the residual network to minimize gradient vanishing, we can restate equation 3.6 and obtain equation 3.8.

$$x_{l+1} = x_l + \mathcal{F}(x_l, W_l) \tag{3.8}$$

The function $h(\cdot)$ represents the identity mapping, and the activation function is ReLU (where $y_l > 0$). Therefore, equation 3.6 can be expressed as equation 3.8. Also, recursively $\mathrm{x}_{l+2} = \mathrm{x}_{l+1} + \mathcal{F}(\mathrm{x}_{l+1}, W_{l+1}) = \mathrm{x}_l + \mathcal{F}(\mathrm{x}_l, W_l) + \mathcal{F}(\mathrm{x}_{l+1}, W_{l+1})$ we will have:

$$x_l = x_l + \textstyle\sum_{i=l}^{L-1} \mathcal{F}(x_l, W_l) \tag{3.9}$$

For any deeper unit L and any shallower unit l.

Also, equation 3.9 leads to nice backpropagation properties. Denoting the loss function as ϵ, from the chain rule of backpropagation we have:

$$\frac{\partial \epsilon}{\partial x_l} = \frac{\partial \epsilon}{\partial x_L}\frac{\partial x_L}{\partial x_l} = \frac{\partial \epsilon}{\partial x_L}\left(1 + \frac{\partial}{\partial x_l}\textstyle\sum_{i=l}^{L-1} \mathcal{F}(x_l, W_l)\right) \tag{3.10}$$

Looking at the first term of equation 3.10, we can observe the addition of a constant value of 1. This value is consistently propagated through backpropagation, regardless of the depth of the layer. As a result, it guarantees a minimum gradient, thereby addressing the issue of gradient vanishing.

3.3 Residual attention block

Section 3.2 describes the use of a residual network combined with an attention module to address the issue of gradient vanishing. In this section, the focus is on the residual attention block, which not only minimizes gradient vanishing but also performs feature refinement. This combination enables effective Infrared small target detection by incorporating both the benefits of the residual network and the attention module.

3.3.1 Attention module

In this section, a simple yet effective convolutional block attention module is described. This module operates in two separate stages, sequentially, when given a feature map. It generates attention maps for both channel attention and spatial attention. These attention maps are then multiplied with the original input feature map to perform adaptive feature refinement. The CBAM (Convolutional Block Attention Module) [23] used as the attention module in this thesis is lightweight yet versatile, allowing it to be attached to various CNN architectures and trained end-to-end.

When an image passes through multiple CNN layers, its dimensions decrease due to convolutional operations and pooling, while the number of channels increases.

These increased channels are obtained from multiple random kernels and are used for subsequent target detection tasks. However, these expanded channels are the result of random kernels and convolutional operations, meaning that certain channels may contain important information for target detection, while others may hold irrelevant or meaningless information. Therefore, to achieve effective target detection performance, it is necessary to emphasize important channels and attenuate unnecessary ones. This can be accomplished through channel attention.

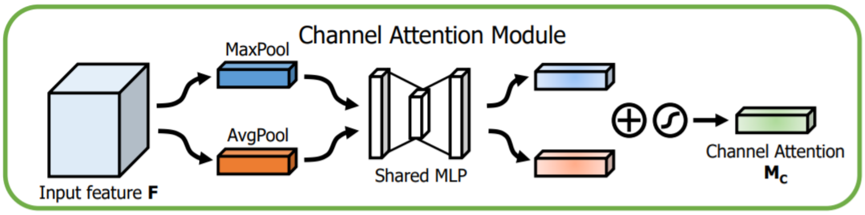

Figure 3.10 Channel attention process

Figure 3.8 illustrates the process of generating the channel attention vector required to refine the feature map in a channel-wise manner using CNN. Initially, the input feature map F undergoes global max pooling and global average pooling, resulting in two 1x1xC-sized vectors. These two vectors are then passed through a multilayer perceptron with shared weights, introducing non-linearity. The two vectors with non-linearity are summed together, and the resulting sum undergoes a sigmoid function to obtain a probability-weighted encoding. This process can be represented by equation 3.11 and equation 3.12.

$$M_c(X) = \sigma\Big(MLP\big(AvgPool(X)\big) + MLP\big(MaxPool(X)\big)\Big) \qquad (3.11)$$

$$X' = M_c(X) \otimes X \tag{3.12}$$

The term $M_c(\cdot)$ in equation 3.11 and 3.12 represents the channel attention vector, which is a probability-weighted representation indicating the importance of different feature maps among the C feature maps. The symbol σ denotes the sigmoid function, and the symbol $\otimes$ represents element-wise multiplication. The channel attention process is performed by element-wise multiplication between the generated channel attention vector M_c(X) and the input feature X.

After performing channel attention to emphasize important channels for target detection, spatial attention is now performed to highlight the locations of targets within the image.

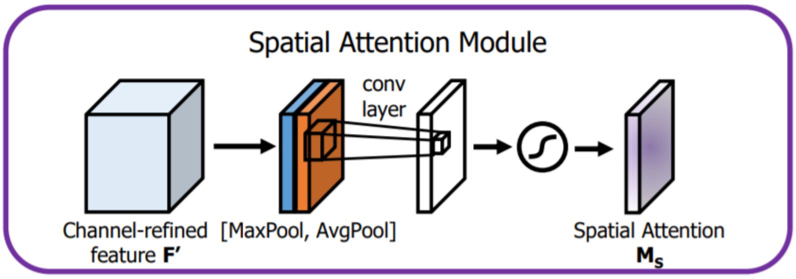
Spatial Attention Module
conv layer
Channel-refined feature **F'**
[MaxPool, AvgPool]
Spatial Attention **M_s**

Figure 3.11 Spatial attention process

The figure 3.9 illustrates the process of generating the spatial attention matrix for performing spatial attention. The channel-refined feature X' is used as the input. Global average pooling and global max pooling are applied along the channel axis of the channel-refined feature to create two matrices of size WxH each. These two matrices are then concatenated and subjected to convolution with a 7x7 kernel size to reduce the channel dimension. Subsequently, the channel-reduced matrix goes through the sigmoid activation function to generate the spatial attention matrix. This process can be expressed by equation 3.13.

$$M_s(X') = \sigma\big(f^{7\times7}([AvgPool(X')\,;MaxPool(X')])\big) \tag{3.13}$$

$$X'' = M_s(X')\ \otimes X' \tag{3.14}$$

The equation 3.13 represents the spatial attention matrix $M_s(\cdot)$, which expresses the importance of WxH pixels as probabilities. σ denotes the sigmoid function, $f^{7\times7}$ represents the convolution with a 7×7 kernel size, and [;] indicates concatenation. Spatial attention is performed by element-wise multiplication between the channel-refined feature and the spatial attention matrix. The CBAM (Convolutional Block Attention Module) applied in this thesis performs feature refinement by sequentially applying channel attention and spatial attention, as shown in the figure 3.10

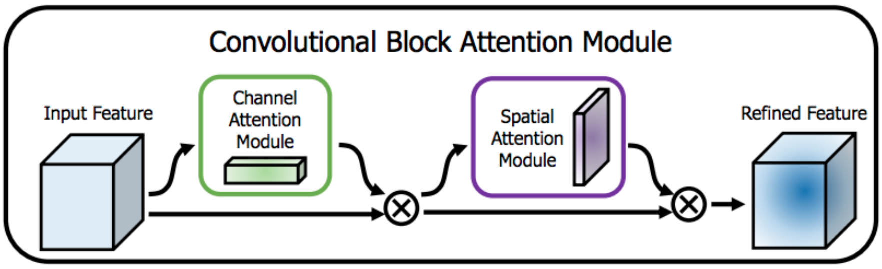

Figure 3.12 Covolutional block attention module process

3.3.2 Residual network with attention module

The Residual Attention Block is a structure that combines the residual network architecture with the Convolutional Block Attention Module (CBAM), as shown in figure 3.11. This configured Residual Attention Block can be applied to the encoder and decoder sections of the U-net for infrared small target detection.

By applying the Residual Attention Block to the encoder and decoder, it addresses the gradient vanishing issue inherent in residual networks and performs feature refinement through the attention module. This enables effective learning even with deep layers in the network for infrared small target detection. Additionally, feature refinement allows for assigning more weight to meaningful features relevant to the detection task and suppressing less important features, leading to improved detection performance.

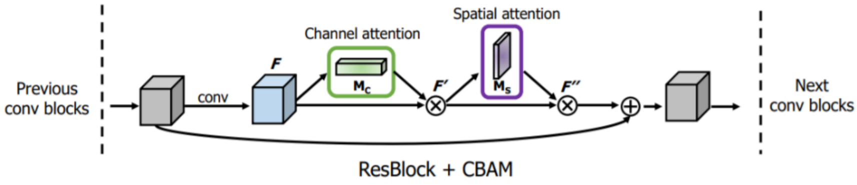

Figure 3.13 Residual attention block

3.4 Proposed Network Structure

3.4.1 Encoder section

The proposed network applies residual attention block to each encoder layer, as shown in figure 3.12. Main components of the block are the residual structure and the attention module. The former prevents gradient vanishing, whereas the latter performs adaptive feature refinement that gives weights to the features when learning the network. The attention module is comprised of the channel attention and the spatial attention. The channel attention can be expressed as:

$$\mathrm{M_c}(X) = \sigma\left[MLP(P_{avg}(X) + MLP(P_{max}(X))\right] \tag{3.15}$$

$$X' = \mathrm{M_c}(X) \otimes X \tag{3.16}$$

where σ denotes sigmoid function, and ⊗ denotes element-wise multiplication. The spatial attention can be expressed as:

$$M_s(X') = \ \sigma\left[f^{3\times3}\left(\left[P_{avg}(X'); P_{max}(X')\right]\right)\right] \tag{3.17}$$

$$X'' = \mathrm{M_s}(X') \otimes X' \tag{3.18}$$

where $f^{3\times3}$ denotes convolution operation with kernel size 3×3, and $[\,;]$ denotes concatenation.

The attention process is sequentially performed as follows:

1) Feature map $X \in \mathbb{R}^{H\times W\times C}$ is used as an input to the channel attention process. A 1D attention map $\mathrm{M_c}(X) \in \mathbb{R}^{1\times1\times C}$ is generated and is element-wise-multiplied with X to create a channel attention feature X'.
2) The output of the previous step, X', is used as an input to the spatial attention process. The spatial attention process creates a 2D spatial attention map, which is element-wise-multiplied with X' to create spatial attention feature X''.

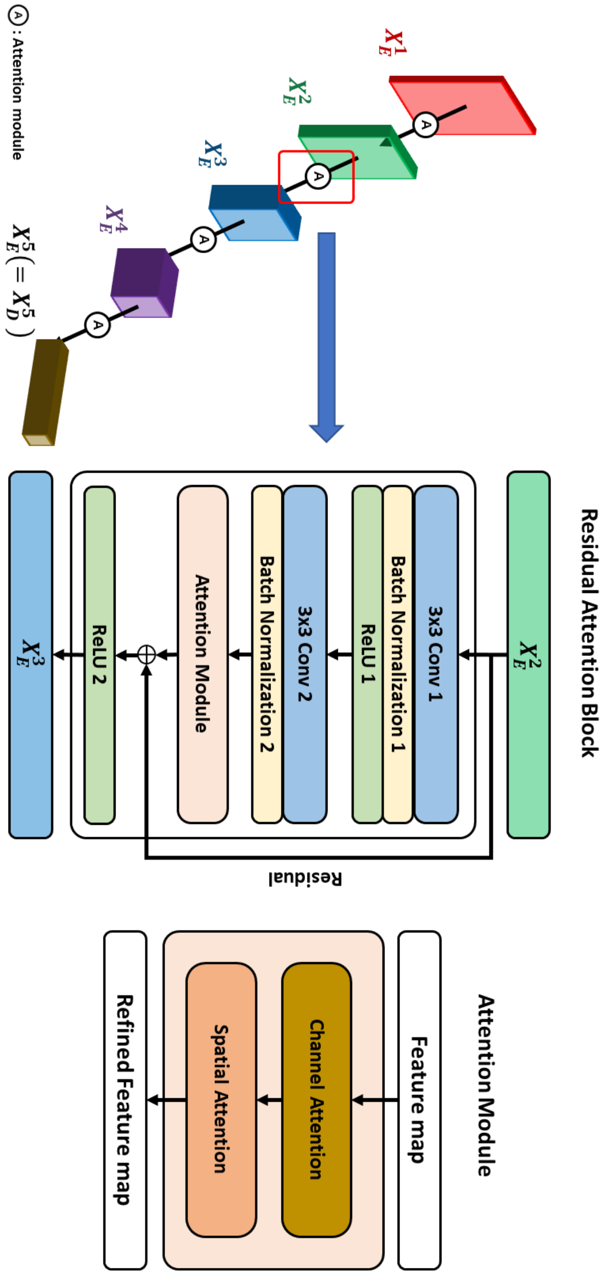

Figure 3.14 Encoder section of proposed network

3.4.2 Decoder section

For the decoder, our network adopts full-scale skip connection and the residual attention block to perform multiscale feature fusion. The process of multiscale feature fusion of the decoder is shown in figure 3.13-3.16. As an example, the process of multiscale feature fusion of the second stage decoder is shown in Fig. 3.14. First, by concatenating the feature maps X_D^3, X_D^4 and X_D^5 obtained from the decoder and feature maps X_E^1 and X_E^2 obtained from the encoder, the exquisite information from the shallow layer and the semantic information from the deep layer are seamlessly merged. The concatenated feature map is used as an input for the residual attention block and is channel-wise and pixel-wise refined through the attention module. Then, a feature map X_D^2 of the decoder can be obtained through 3×3 convolution, batch normalization, and ReLU activation function. Through this process, the refined and fused feature map can reflect both low-level and high-level information, apt for an effective IRSTD.

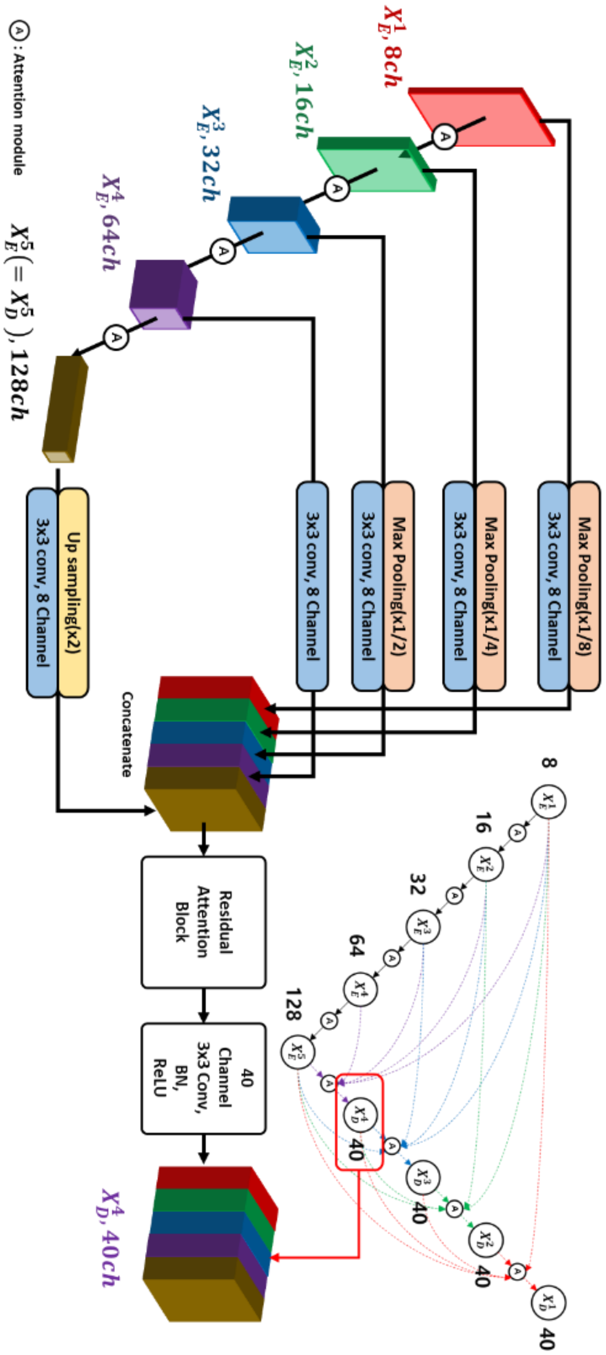

Figure 3.15 Decoder of proposed network #4

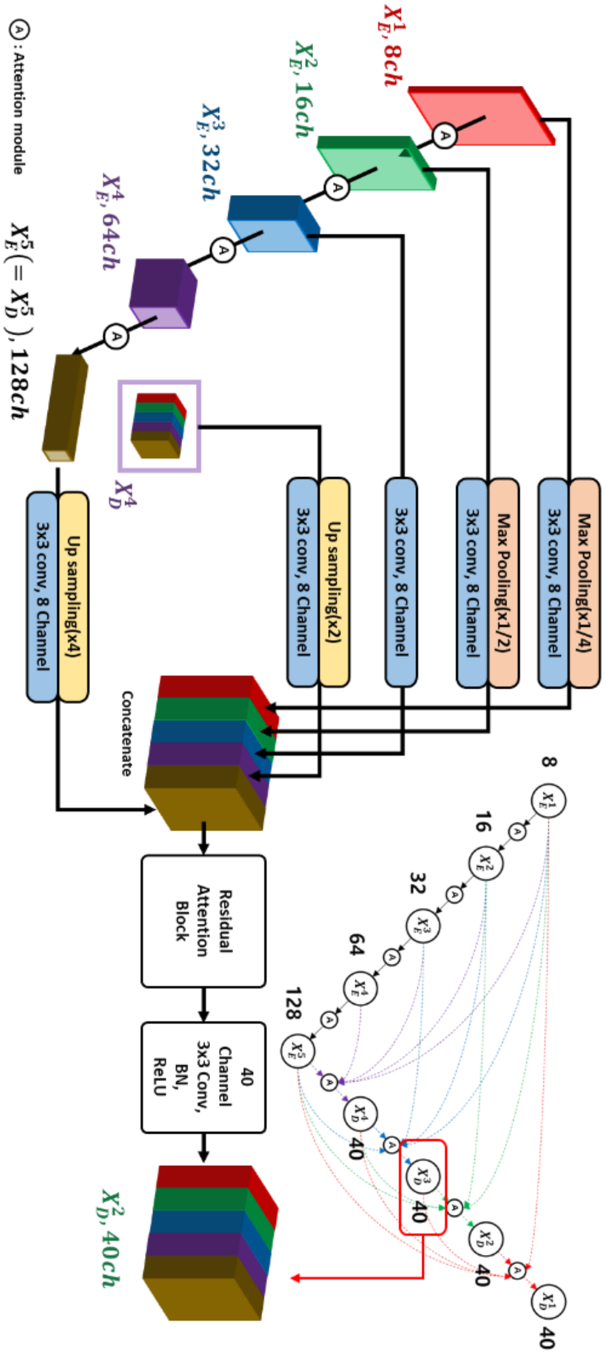

Figure 3.16 Decoder of proposed network #3

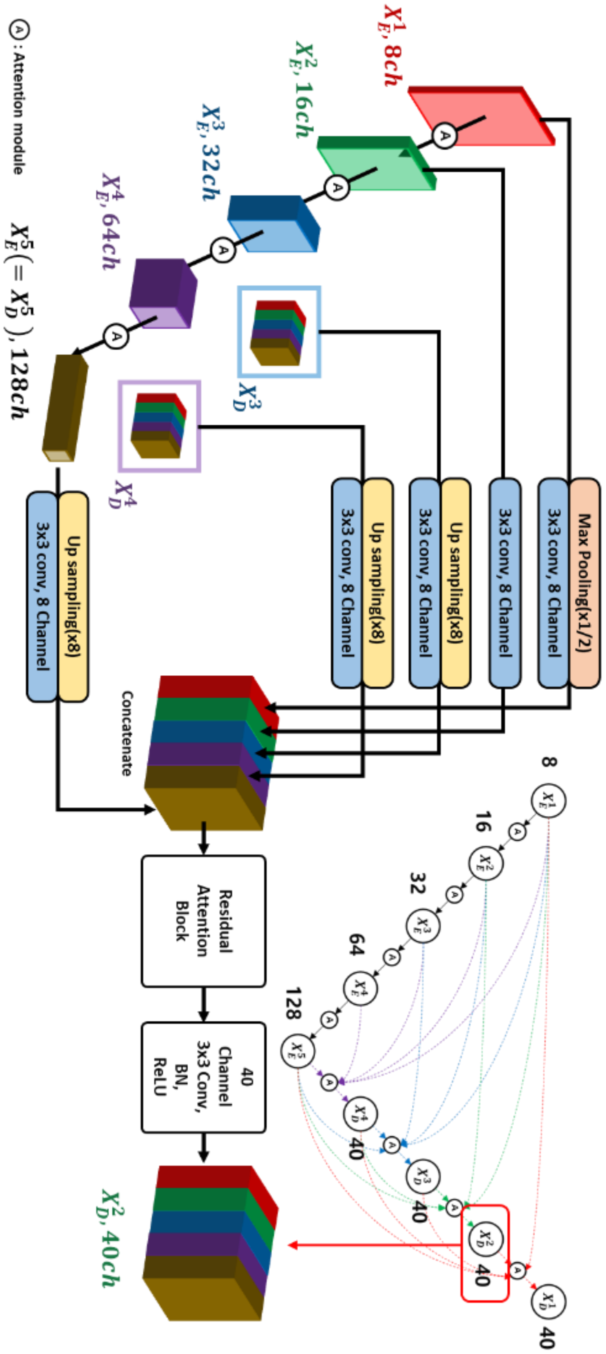

Figure 3.17 Decoder of proposed network #2

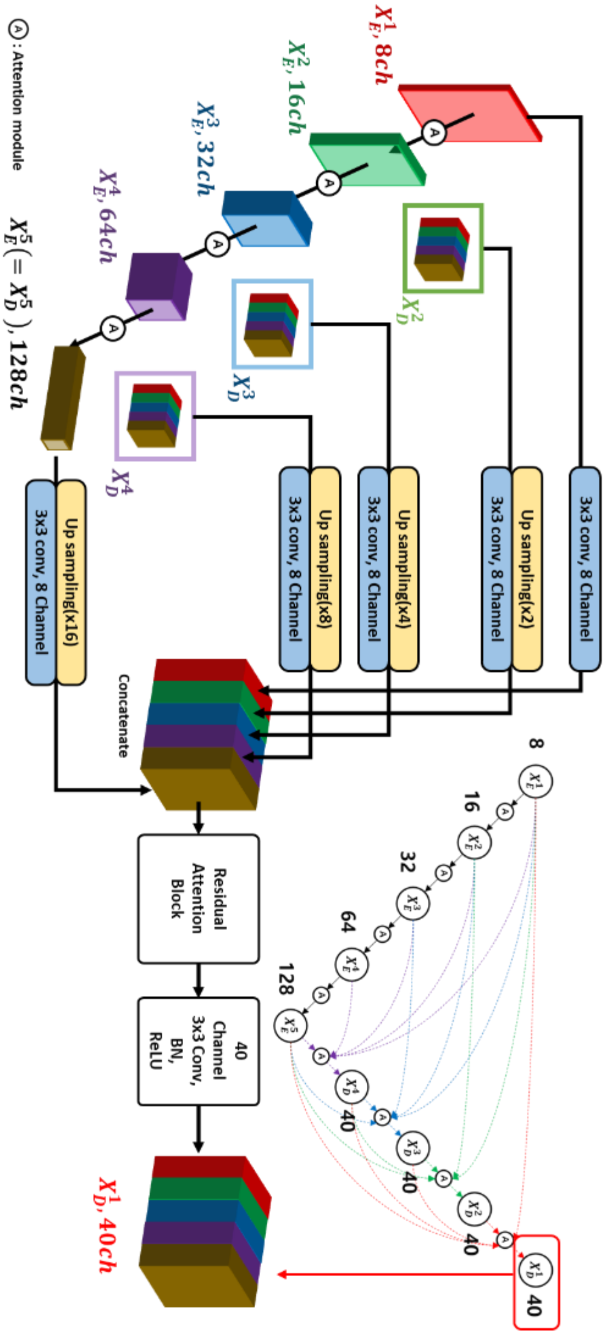

Figure 3.18 Decoder of proposed network #1

3.4.3 Loss function with deep supervision

We apply full-scale deep supervision to AMFU-net for hierarchical information learning and adopt Soft-IoU loss as the loss function.

$$Soft - IoU = \frac{intersection+smooth}{union+smooth} \tag{3.19}$$

$$Loss_{soft-IoU} = 1 - soft - IoU \tag{3.20}$$

For full-scale deep supervision, AMFU-net outputs feature map at every decoder stage, which is supervised by the ground truth. In addition to the five decoder outputs, the full-scale deep supervision of our network additionally uses the average of the 256×256-rescaled outputs to reflect information from every decoder stage. Full-scale deep supervision is achieved by averaging the losses from the six abovementioned outputs. As a result, training reflects all the essential information: the information from the shallow layer, the deep layer, and the overall.

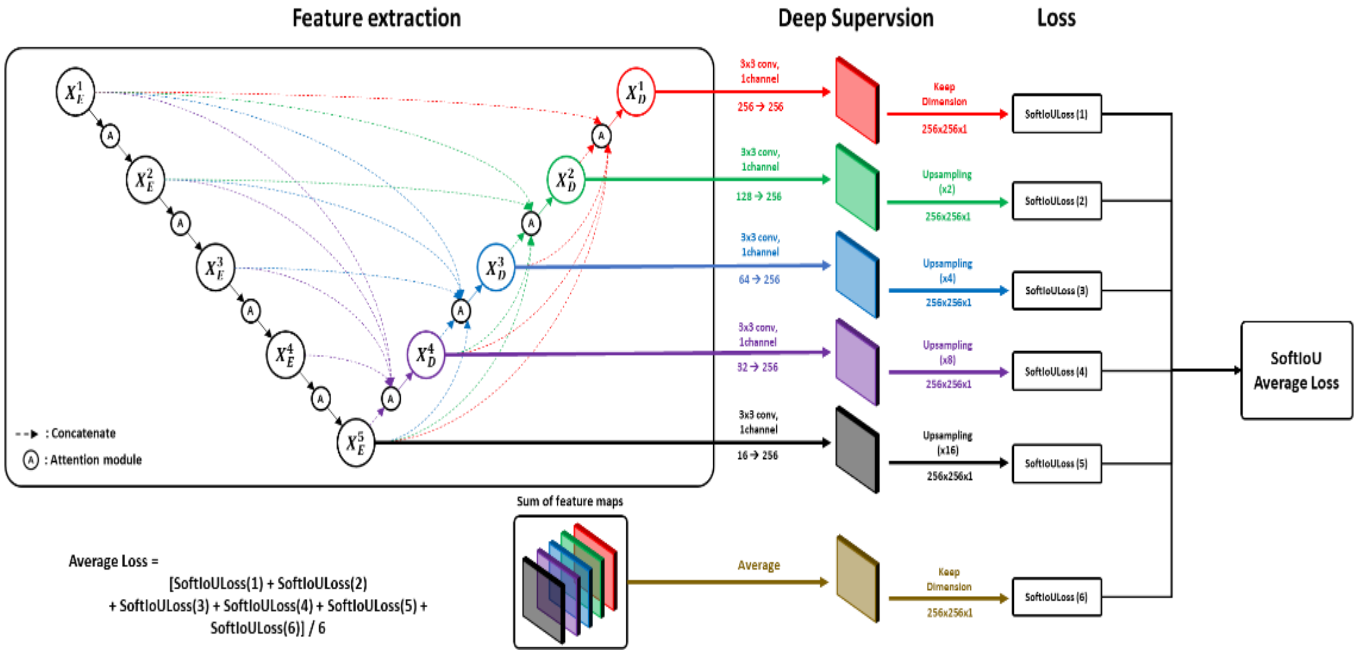

Figure 3.19 Deep supervision with soft IoU loss

Chapter 4

Performance Analysis of Proposed Network

In Chapter 3, we presented the novel network architectures designed for Infrared (IR) small target detection. The focus was on three distinct network structures: U-net, U-net++, and U-net3+. Each architecture was carefully examined and evaluated to identify the most suitable choice for our proposed method. Through a comprehensive analysis of network parameters, it was determined that U-net3+ offered a compelling advantage due to its ability to deliver effective performance while maintaining a reduced parameter count. This characteristic is particularly advantageous as it enables efficient operation even on resource-constrained systems with limited computing power.

Furthermore, in order to enhance the capabilities of the network in detecting small targets, we made a significant modification by substituting the conventional encoder and decoder blocks with residual attention blocks. This architectural refinement aimed to leverage the benefits of attention mechanisms and facilitate more efficient and accurate small target detection.

Building upon these advancements, we proceeded to introduce our proposed Infrared small target detection network, named AMFU-net (Attention Multiscale Feature fusion U-net). This network was specifically designed to address the challenges of detecting small targets in IR imagery. To evaluate its performance, we conducted simulations using open-source datasets that are widely accepted within the research community. The obtained results were subjected to both quantitative and qualitative analysis, allowing for a comprehensive assessment of the network's capabilities.

4.1 Simulation results

In this thesis, we have conducted a comprehensive evaluation of our proposed method using the ACM dataset [8]. The ACM dataset consists of infrared small target images captured using an IR camera in real-world environments. The dataset includes both images with simple backgrounds and images with complex backgrounds, providing a diverse range of scenarios for effective training.

To perform our experiments, we utilized a high-performance computing setup comprising an Intel I7-6700K processor clocked at 4.0GHz, an NVIDIA RTX 3090 graphics card with 24GB of video RAM, and 64GB of RAM. This configuration ensured sufficient computational power and memory capacity to handle the demanding nature of deep learning tasks. To facilitate the training process, the ACM dataset was split into two sets: the training set and the test set. The division was made with a ratio of 1:1, ensuring an equal distribution of data between the two sets.

In preparation for network training, all images in the dataset were normalized and resized to a resolution of 256x256. This preprocessing step helped ensure consistent input dimensions and improved the convergence of the network during training. Furthermore, it is important to mention the hyperparameter settings used for network learning. These settings play a crucial role in determining the model's performance and training dynamics. The specific hyperparameters employed in our experiments were carefully chosen to strike a balance between model complexity and generalization capabilities. The hyperparameter settings for network learning are as follows: 1) optimizer: Adagrad, 2) initialization method: Xavier, 3) batch size: 16, 4) learning rate: 0.05 and 5) epochs: 1500.

Overall, this experimental setup, including the ACM dataset, hardware specifications, dataset partitioning, image preprocessing, and hyperparameter settings, allowed us to conduct a rigorous evaluation of our proposed method for infrared small target detection.

4.1.1 Open source IRSTD dataset

In this thesis, the ACM dataset [8] was employed as the training dataset for the network. The ACM dataset consists of a collection of infrared (IR) small target images captured using an IR camera in real-world scenarios. The dataset was designed to encompass various conditions encountered in small target detection tasks, including different backgrounds and levels of complexity. To ensure the effectiveness of the network training process, the ACM dataset was carefully divided into two subsets: a train set and a test set. The train set and the test set were balanced, with an equal ratio of 1:1 in terms of the number of samples.

The train set comprises a diverse range of IR small target images, including both simple background scenarios and complex background scenarios. The images with simple backgrounds provide a clear and uncluttered context, facilitating the network's learning of the target characteristics. On the other hand, the images with complex backgrounds simulate challenging real-world conditions, where the small targets may be surrounded by noise, clutter, or other distracting elements. Similarly, the test set includes IR small target images with both simple and complex backgrounds. This enables a comprehensive evaluation of the trained network's performance across various scenarios and provides insights into its generalization capability.

By utilizing the ACM dataset, this thesis aims to develop and assess a robust IR small target detection network that can effectively handle diverse background conditions and accurately detect small targets amidst complex visual environments.

The figure 4.1 is an example of dataset.

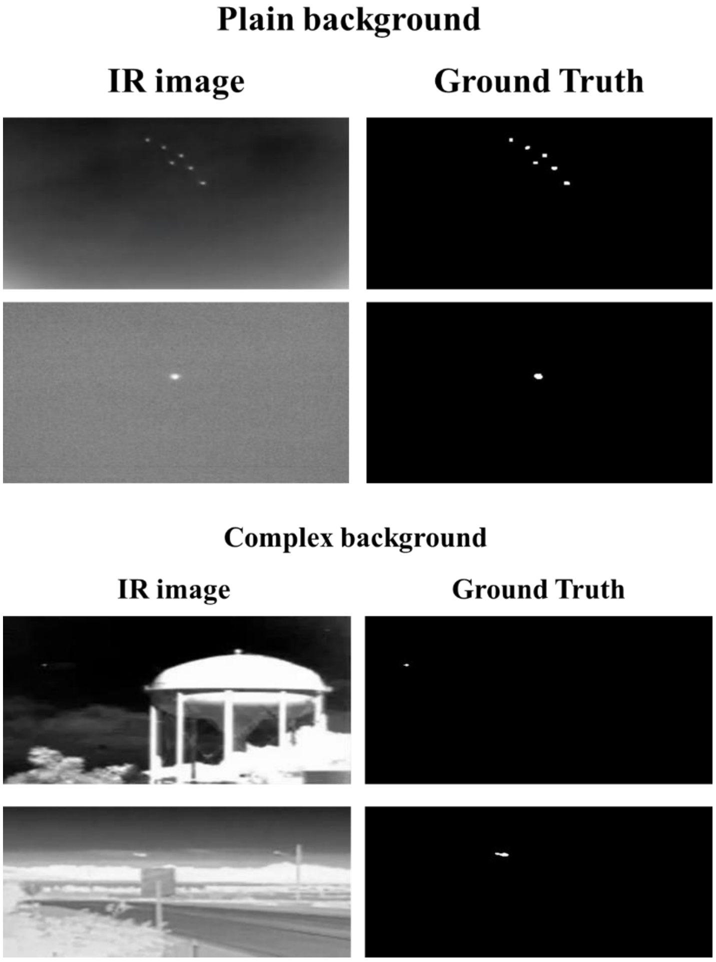

Figure 4.1 Infrared small target detection dataset

4.1.2 Qualitative analysis of IR small target detection

To evaluate the performance of the proposed AMFU-net, a comparative analysis was conducted against conventional deep learning-based Infrared Small Target Detection (IRSTD) algorithms. The hyperparameters of the comparative group were set based on the guidelines provided in the referenced papers [8] and [9].

Figures 4.2 to 4.5 present the detection results of the proposed AMFU-net and the comparison group. In Figure 4.2, the robust target detection performance of AMFU-net is evident compared to the existing algorithms. False alarms, represented by yellow boxes, can be observed in the results of DNA(VGG10) and DNA(Res10). In figure 4.3 and 4.4 show that the proposed network is more accurate in terms of segmented shape compared to existing algorithms. Furthermore, Figure 4.5 showcases an instance of a false alarm occurrence in the ACM dataset and a missed detection case in DNA(VGG10). These visual examples highlight the advantages of the proposed AMFU-net in terms of improved detection accuracy and reduced false alarm rates. By providing these comprehensive comparative results, we validate the efficacy and superiority of the proposed AMFU-net over conventional IRSTD algorithms. The findings demonstrate the network's ability to effectively detect infrared small targets while minimizing false alarms and enhancing the accuracy of target segmentation.

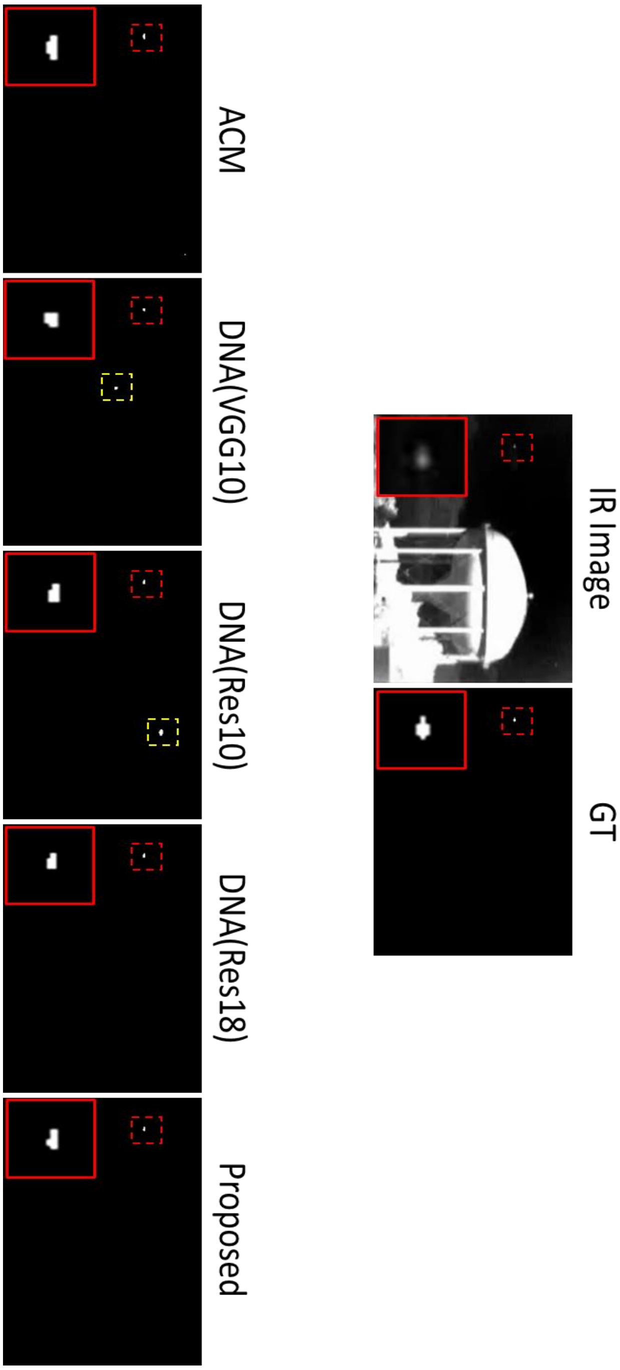

Figure 4.2 IR small target detection results #1

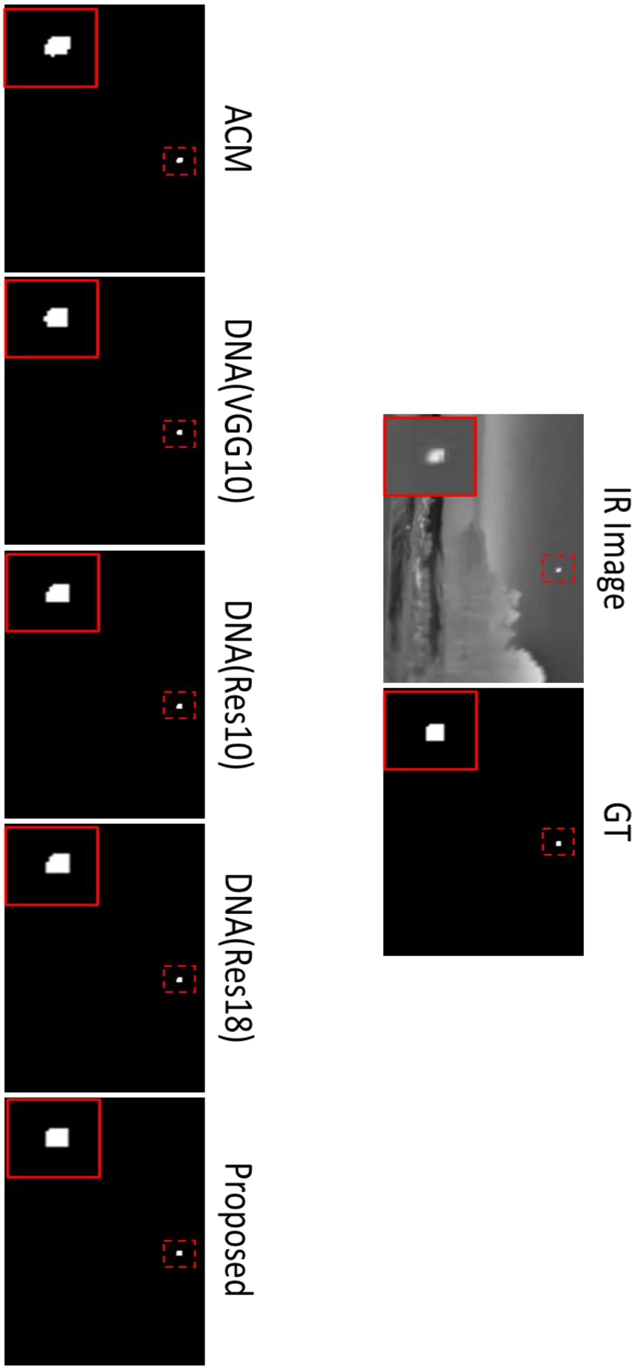

Figure 4.3 IR small target detection results #2

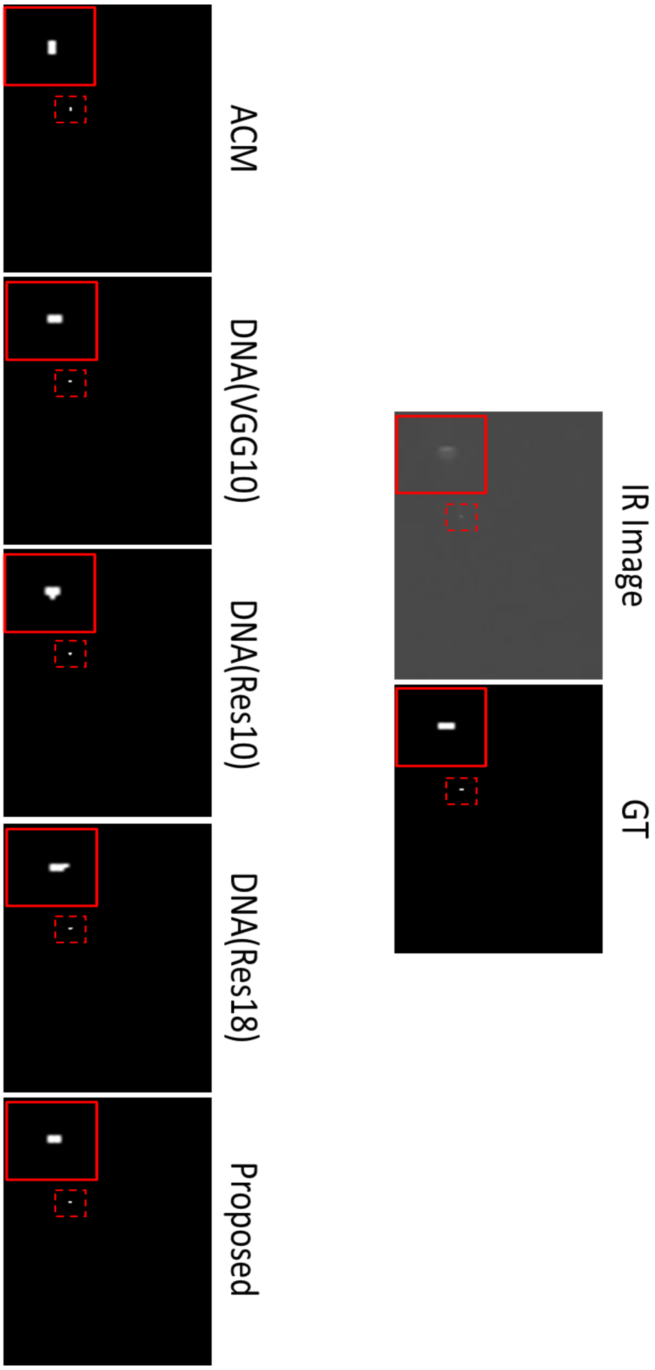

Figure 4.4 IR small target detection results #3

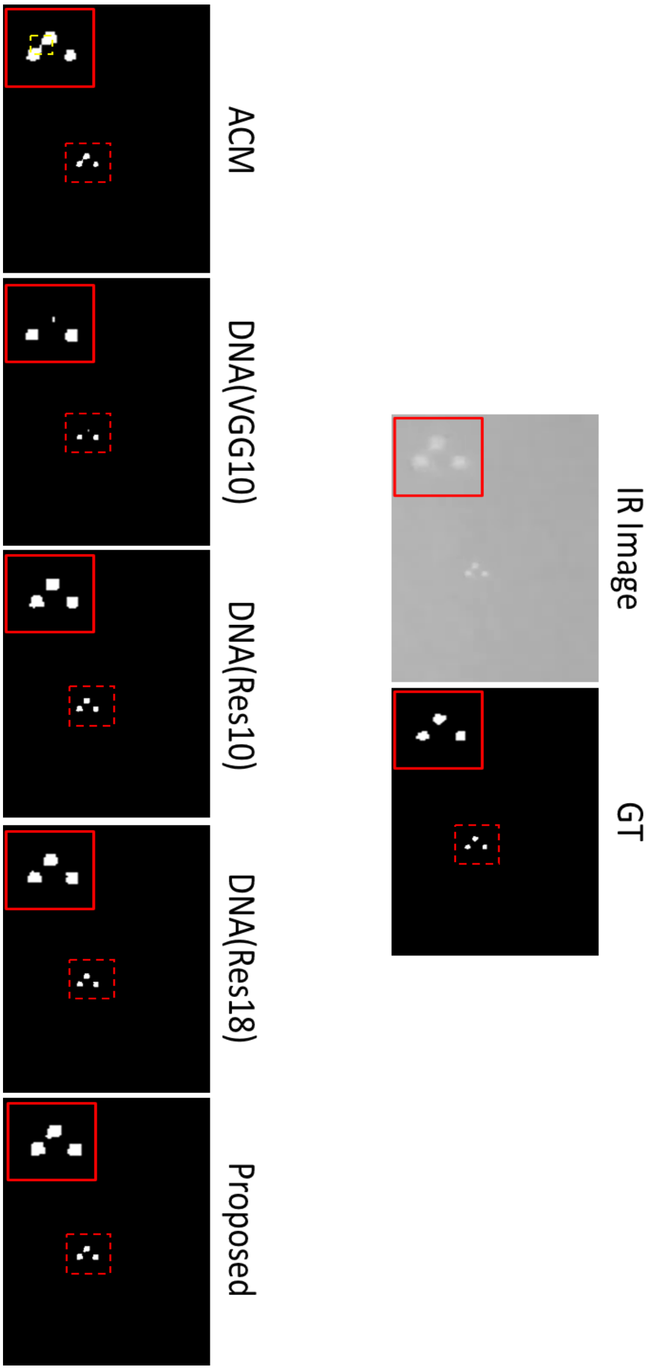

Figure 4.5 IR small target detection results #4

4.1.3 Quantitative analysis of IR small target detection

Table 4.1. is a quantitative analysis of the deep learning-based IRSTD algorithms. To analyze the detection performance of each network, mIoU was selected as a metric, which can be expressed as:

$$\mathrm{IoU_i} = \frac{TP_i}{TP_i+FP_i+FN_i} \tag{4.1}$$

$$\mathrm{mIoU} = \frac{1}{N}\sum_i^N IoU_i \tag{4.2}$$

The evaluation results reveal that AMFU-net achieves the highest detection performance among the compared algorithms. It demonstrates a significant improvement in mean Intersection over Union (mIoU) compared to ACM, DNA (VGG10), DNA (ResNet10), and DNA (ResNet18). Specifically, AMFU-net increases the mIoU by 10.6% (from 0.6791 to 0.7512) compared to ACM, 4.0% (from 0.7219 to 0.7512) compared to DNA (VGG10), 1.8% (from 0.7380 to 0.7512) compared to DNA (ResNet10), and 1.3% (from 0.7411 to 0.7512) compared to DNA (ResNet18).

Table 4.1 Quantitative analysis of different methods

Method	ACM	DNA (VGG10)	DNA (Res10)	DNA (Res18)	Proposed
mIoU	0.6791	0.7219	0.7380	0.7411	0.7512
Parameters (MB)	1.48	9.42	10.13	18.24	2.17
FPS	178.0	73.3	66.2	45.8	86.1

Furthermore, the computational efficiency of the algorithms was assessed in terms of Frames Per Second (FPS). It was observed that ACM achieves the fastest performance in terms of FPS. However, this comes at the cost of compromised detection performance, as it exhibits the lowest mIoU among all algorithms. On the other hand, the proposed AMFU-net achieves the second fastest performance while maintaining the highest mIoU. This indicates that AMFU-net strikes a balance between computational efficiency and detection accuracy.

In summary, the quantitative analysis substantiates that AMFU-net outperforms the compared algorithms in terms of detection performance, with notable improvements in mIoU. Additionally, despite its high detection accuracy, AMFU-net demonstrates competitive computational efficiency, positioning it as a promising solution for Infrared Small Target Detection tasks.

4.1.4 Analysis of Receiver Operating Characteristics(ROC)

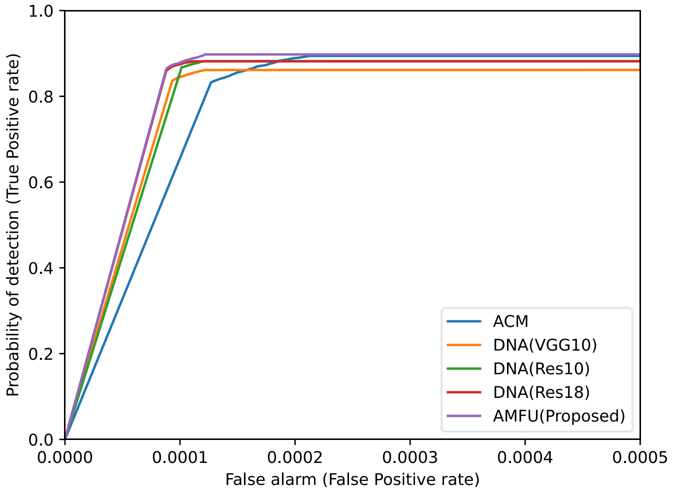

Figure 4.6 ROC curves of various deep learning algorithms

Figure 4.6 presents the Receiver Operating Characteristic (ROC) curves obtained from the selected deep learning-based algorithms. The ROC curves provide insights into the trade-off between the detection probability (True Positive Rate) and the false alarm rate (False Positive Rate).

From the graph, it is evident that both DNA (ResNet18) and the proposed AMFU-net exhibit superior robustness compared to the other tested IRSTD algorithms. This is indicated by their ROC curves being closer to the top-left corner of the graph, which represents higher detection probabilities and lower false alarm rates. Although DNA (ResNet18) and AMFU-net exhibit similar false alarm rates, the proposed algorithm demonstrates a higher probability of detection. This indicates that AMFU-net achieves a better balance between correctly detecting small targets and minimizing false alarms. Based on the analysis of the ROC curves, it can be concluded that AMFU-net is the most robust algorithm among the tested IRSTD algorithms. Its superior performance in terms of both detection probability and false alarm rate makes it a promising choice for accurate and reliable infrared small target detection applications.

4.2 Ablation study

To evaluate the impact of the residual attention block in AMFU-net, an ablation study was conducted by comparing it against two variants of the network: 1) AMFU-net without the residual attention block entirely, and 2) AMFU-net with the attention module removed.

Table 4.2 presents the comparison results of these variants in terms of detection performance and Frames Per Second (FPS). The variant without the residual attention block exhibited a degradation in detection performance of 5%, indicating that the residual attention block plays a crucial role in improving the network's ability to detect small targets accurately. On the other hand, the variant with only the attention module removed showed a degradation of 2% in detection performance. This suggests that while the attention module contributes to performance improvement, the overall impact is relatively smaller compared to the residual attention block. Considering the computational efficiency, the variant without the residual attention block achieved a faster FPS of 85%, while the variant with only the attention module removed achieved a slower FPS of 65%.

Table 4.2 Ablation study results

Method	AMFU-net w/o residual attention block	AMFU-net w/o attention module	AMFU-net
mIoU	0.7124	0.7328	0.7512
Parameters (MB)	1.63	1.67	2.17
FPS	159.5	142.8	86.1

This trade-off between detection performance and computational efficiency highlights the significance of the residual attention block in achieving a balance between accuracy and speed.

Based on these findings, it is recommended to utilize AMFU-net with the residual attention block intact. This configuration demonstrates the highest detection performance while still maintaining a relatively fast FPS of 86.1. The presence of the residual attention block ensures effective feature refinement and target detection, making it a preferred choice for IR small target detection tasks.

4.3 Embedded system applications

AMFU-net demonstrates its capability to perform effective Infrared Small Target Detection (IRSTD) with a reduced number of parameters, enabling its operation in low computational environments. In this context, the algorithms were validated on an embedded system with limited computational resources, providing a practical assessment of their performance.

The NVIDIA Jetson Orin, equipped with the NVIDIA Ampere architecture featuring 2048 CUDA cores and 64 Tensor cores, along with 32GB of RAM and running Ubuntu 20.04, was used for the evaluation. It is important to note that all networks were tested using the same trained weights as used in producing Table 4.1, ensuring consistency in terms of mean Intersection over Union (mIoU) and parameter size. The results presented in Table 4.3 confirm that the proposed algorithm delivers on-line inference speed, achieving an impressive Frames Per Second (FPS) of 29.5 on the embedded system. This demonstrates the algorithm's ability to effectively detect small targets even in low-power and low-computational environments.

The successful validation of the proposed algorithm on the embedded system highlights its practical viability and suitability for real-world applications that require efficient IRSTD on resource-constrained platforms. The algorithm's ability to operate effectively in low computational environments enhances its potential for deployment in various scenarios, including embedded systems and other edge computing devices.

Table 4.3 Inference time on embedded systems

Method	ACM	DNA (VGG10)	DNA (Res10)	DNA (Res18)	Proposed
mIoU	0.6791	0.7219	0.7380	0.7411	0.7512
Parameters (MB)	1.48	9.42	10.13	18.24	2.17
FPS	43.9	23.8	21.7	16.1	29.5

Chapter 5

Conclusion

5.1 Conclusion and summary

This theis proposed AMFU-net, an efficient IRSTD network. The network, based on Unet3+ with the residual attention block, can effectively fuse the output feature map obtained from each stage of the network with only few parameters through a full-scale skip connection. In addition, the proposed network prevents gradient vanishing by applying residual blocks to the encoder and the decoder, and performs effective feature extraction using the attention module. Comparative evaluation showed that even with 88% less parameters than the runner-up, the designed AMFU-net outperformed the state-of-the-art IRSTD networks on detection performance with the mIoU of 0.7512, and maintained a fast FPS of 86.1. Moreover, we proved that our lightweight network achieves online inference speed (FPS: 29.5) even on an embedded system with a low computational setup.

5.2 Future works

Detected results obtained from the proposed IR small target detection in this thesis can be effectively utilized in IR small target tracking research. To achieve this, the results obtained from the network can be used as measurements in tracking, enabling effective tracking even in noisy IR imagery.

Furthermore, research efforts can be directed towards enhancing detection performance. While the proposed network architecture in this thesis is based on Convolutional Neural Networks (CNNs), recent trends in research indicate active exploration of Transformer-based algorithms for target detection and segmentation. Transformer-based algorithms have shown impressive performance in various tasks. Therefore, it is deemed feasible to improve detection performance further by employing Transformer-based algorithms rather than CNN-based methods. However, it is important to note that Transformer-based algorithms typically have a higher number of parameters compared to CNN-based networks and require large-scale training datasets. Therefore, research in these areas would be necessary to address these challenges.

Bibliography

[1] M. Zhao, W. Li, L. Li, J. Hu, P. Ma, and R. Tao, "Single-frame infrared small-target detection: A survey," IEEE Geoscience and Remote Sensing Magazine, vol. 10, no. 2, pp. 87-119, 2022.

[2] C. P. Chen, H. Li, Y. Wei, T. Xia, and Y. Y. Tang, "A local contrast method for small infrared target detection," IEEE transactions on geoscience and remote sensing, vol. 52, no. 1, pp. 574-581, 2013.

[3] 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 Geoscience and Remote Sensing Letters, vol. 11, no. 12, pp. 2168-2172, 2014.

[4] 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 Geoscience and Remote Sensing Letters, vol. 15, no. 4, pp. 612-616, 2018.

[5] Y. Shi, Y. Wei, H. Yao, D. Pan, and G. Xiao, "High-boost-based multiscale local contrast measure for infrared small target detection," IEEE Geoscience and Remote Sensing Letters, vol. 15, no. 1, pp. 33-37, 2017.

[6] 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 transactions on image processing, vol. 22, no. 12, pp. 4996-5009, 2013.

[7] H. Wang, L. Zhou, and L. Wang, "Miss detection vs. false alarm: Adversarial learning for small object segmentation in infrared images," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 8509-

8518.

[8] Y. Dai, Y. Wu, F. Zhou, and K. Barnard, "Asymmetric contextual modulation for infrared small target detection," in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, pp. 950-959.

[9] B. Li et al., "Dense nested attention network for infrared small target detection," IEEE Transactions on Image Processing, 2022.

[10] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: Redesigning skip connections to exploit multiscale features in image segmentation," IEEE transactions on medical imaging, vol. 39, no. 6, pp. 1856-1867, 2019.

[11] H. Huang et al., "Unet 3+: A full-scale connected unet for medical image segmentation," in ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020: IEEE, pp. 1055-1059.

[12] J. Peng and W. Zhou, “Infrared background suppression for segmenting and detecting small target,” Acta Electronica Sinica, vol. 27, no. 12, pp. 47–52, 1999.

[13] S. D. Deshpande, M. H. Er, R. Venkateswarlu, and P. Chan, “Max-mean and max-median filters for detection of small targets,” in Proc. Conf. Signal Data Process. Small Targets, 1999, pp. 74–83, doi: 10.1117/12.364049.

[14] C. Tomasi and R. Manduchi, “Bilateral filtering for gray and color images,” in Proc. Int. Conf. Comput. Vis., 1998, pp. 839– 846, doi: 10.1109/ICCV.1998.710815.

[15] T.-W. Bae, S.-H. Lee, and K.-I. Sohng, “Small target detection using the bilateral filter based on target similarity index,” IEICE Electron. Exp., vol. 7,

no. 9, pp. 589–595, 2010, doi: 10.1587/elex.7.589.

[16] J. Arnold, “Detection and tracking of low-observable targets through dynamic programming,” in Proc. Signal Data Process. Small Targets, 1990, vol. 1305, pp. 207–217, doi: 10.1117/12.2321762.

[17] Y. Qin and B. Li, “Effective infrared small target detection utilizing a novel local contrast method,” IEEE Geosci. Remote Sens. Lett., vol. 13, no. 12, pp. 1890–1894, 2016, doi: 10.1109/ LGRS.2016.2616416.

[18] Y. Dai, Y. Wu, and Y. Song, “Infrared small target and background separation via column-wise weighted robust principal component analysis,” Infrared Phys. Technol., vol. 77, pp. 421– 430, Jul. 2016, doi: 10.1016/j.infrared.2016.06.021.

[19] Y. Dai, Y. Wu, Y. Song, and J. Guo, “Non-negative infrared patchimage model: Robust target-background separation via partial sum minimization of singular values,” Infrared Phys. Technol., vol. 81, pp. 182–194, Mar. 2017, doi: 10.1016/j.infrared.2017.01.009.

[20] J. Guo, Y. Wu, and Y. Dai, “Small target detection based on reweighted infrared patch-image model,” IET Image Process., vol. 12, no. 1, pp. 70–79, 2017, doi: 10.1049/iet-ipr.2017.0353.

[21] X. Wang, Z. Peng, D. Kong, P. Zhang, and Y. He, “Infrared dim target detection based on total variation regularization and principal component pursuit,” Image Vis. Comput., vol. 63, pp. 1–9, Jul. 2017, doi: 10.1016/j.imavis.2017.04.002.

[22] O. Ronneberger, P. Fischer, and T Brox, “U-net: Convolutional networks for biomedical image segmentation,” in Proc. Medical Image Comput. Comp.-Assis. Interv. – MICCAI, Navab N., Hornegger J.,Wells W., FrangiA. eds. Lecture Notes in Computer Science, vol. 9351, Cham: Springer,2015.

[23] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in Proceedings of the European conference on computer vision (ECCV), 2018, pp. 3-19.

국문초록

U-Net 기반의 IRSTD(Infrared Small Target Detection)알고리즘에서 탐지 성능을 향상시키기 위해서는 저차원의 특징과 고차원의 특징을 융합하는 것이 중요하다. 기존 알고리즘은 U-Net의 스킵 경로에 컨볼루션 레이어를 추가하고 스킵 연결을 보다 조밀하게 연결하여 저차원 – 고차원 특징 융합을 수행한다. 그러나 컨볼루션 연산이 추가되면 네트워크의 매개 변수 수가 증가하므로 추론시간이 그에 따라 증가하게 된다. 따라서 본 논문에서는 풀 스케일 스킵 연결(Full-scale skip connection) U-Net을 기반 네트워크로 사용하여 적은 수의 매개 변수만으로 저차원 – 고차원의 특징을 융합함으로써 계산 비용을 낮춘다. 또한, 본 논문은 높은 수준의 IRSTD 결과를 보장하기 위해 효과적인 인코더 및 디코더 구조를 제안한다. 잔여 주의 블록(Residual attention block)은 효과적인 특징 추출을 위해 인코더의 각 레이어에 적용된다. 디코더에서는 네트워크의 각 계층으로부터 얻어진 저차원 – 고차원 정보 융합을 효과적으로 수행하기 위해 특징 융합 모듈에 잔여 주의 블록을 적용하였다. 또한 네트워크 학습 수행 시, 각 계층에서 얻어진 모든 특징을 반영하여 학습을 진행하기 위하여 심층 감독(Deep supervision)을 통해 손실 함수를 계산한다. 제안된 알고리즘인 주의집중 멀티스케일 특징 융합 U-Net(Attention Multiscale Feautre Fusion U-Net, AMFU-Net)은 효과적인 표적 탐지 성능과 경량 구조를 보장할 수 있다.

주요어: 적외선 소형 표적 탐지, 주의집중 알고리즘, 특징 융합, U-Net3+

학번: 2021-26553