IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received August 27, 2019, accepted September 16, 2019, date of publication October 1, 2019, date of current version October 18, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2944661

Detection of Infrared Small Targets Using Feature Fusion Convolutional Network

KAIDI WANG[1], SHAOYI LI[1], SAISAI NIU[2], AND KAI ZHANG[1]

[1]School of Astronautics, Northwestern Polytechnical University, Xi'an 710072, China
[2]Shanghai Institute of Spaceflight Control Technology, Shanghai 201109, China

Corresponding author: Kaidi Wang (matronta@163.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61703337, and in part by the Shanghai Aerospace Science and Technology Innovation Fund under Grant SAST2017-082.

ABSTRACT This paper proposes the feature extraction backbone network MNET that is specifically designed for the detection of infrared small targets. The overall network uses three down-sampling operations to adjust the size of the feature map, while preserving sufficient physical characteristics of the small infrared target to be used in the detection. In a next step, the dense connection is used to save the output of each layer of the network in the front channel of the feature map, to better integrate the location information of the shallow network and the semantic information of the deep network. In this way accurate network positioning and classification effects are achieved. As a last step, we introduce a feature attention mechanism to obtain the importance of each feature channel, and to enhance useful features according to their degree of importance. In this way we achieve an adaptive calibration of the feature channels. In order to train the proposed detection network MNET from scratch, the single-phase detection algorithm YOLO is adopted for the detection part. To verify the effectiveness of the proposed method, we captured images and created an infrared small target dataset. The experimental results show that MNET can accurately detect targets of 2 × 2 pixels size in infrared images of 640 × 512 pixels at a processing speed of up to 105 frames per second. MNET meets real-time requirements while providing high quality detection accuracy.

INDEX TERMS Infrared small target, target detection, convolutional neural network, feature fusion.

I. INTRODUCTION

In recent years, infrared imaging technology has witnessed rapid development and widespread use in various fields, such as surveillance, satellite reconnaissance, and remote sensing [1]. Many infrared applications, such as infrared search and tracking (IRST) [2], early warning systems, and missile tracking systems [3], require precise detection and localization of specific small targets. Therefore, infrared small target detection has attracted considerable research interest.

Although infrared small target detection has been studied for many years, it still has the following difficulties:

(1) Owing to the long imaging distance, the target occupies only a few pixels on the imaging plane. The signal is weak, and there is no obvious shape, size, or texture feature. Moreover, the image noise is strong and the signal-to-noise ratio is low.

(2) When the target moves, its imaging size changes in a certain range. The algorithm should be adaptable to such changes.

(3) The amount of data to be processed is large, leading to a contradiction between the complexity and real-time performance of the detection algorithm.

In recent years, infrared small target detection algorithms have attracted considerable interest with the development of convolutional neural networks (CNNs) [4]. Compared with traditional computer vision methods, a CNN directly takes an image as input and automatically learns the features that represent the deep nature of the data. Thus, the steps of feature extraction are simple and efficient. Moreover, compared with the manual selection of features and shallow features, a CNN has powerful data characterization ability [5]. In this case, people begin to study the use of CNN to solve the problem of infrared small target detection.

However, current research regarding the detection of infrared small targets based on convolutional neural networks

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

is mainly focused on fine-tuning detection networks that were designed for universal datasets and on performing infrared small target detection using them. Designing a new feature extraction backbone network for the characteristics of infrared small targets has not been the focus of current research. In case the general network relies on a backbone network that was pre-trained for the ImageNet classification task, certain limitations arise when solving specific problems such as infrared small target detection:

(1) Classification and detection tasks show different sensitivities in the feature map [6]. Classification tasks require multiple down sampling operations for the sake of performance when extracting deep semantic information from the network. However, local texture information plays a critical role in target detection, thus excessive down-sampling tends to affect the identification of small targets.

(2) A general CNN is designed for both large as well as medium targets [4], which inevitably leads to a poor detection of small targets. Moreover, the differences between individual, similar targets in the VOC data set are large. In order to obtain better detection results, the network design needs to be adapted and the network width needs to be increased, which results in a significant increase of the computational cost and a decrease in training speed and detection speed.

(3) Fine-tuning can alleviate the gap caused by the different distributions of data categories [7], but the infrared small target data set varies widely from general classification data sets or detection data sets. Thus, the effect of fine-tuning is minimal. In addition, if the target detector is directly fine-tuned using a pre-trained network, it is basically impossible to change the structure of the network itself [8].

Therefore, this paper proposes the feature extraction backbone network MNET that is specifically designed for the detection of infrared small targets and that addresses the problems mentioned above. In general, an inspection network uses six down-sampling steps to meet the detection needs of large targets. However, excessive down-sampling causes the disappearance of infrared small targets in the feature map. MNET adjusts the size of the feature map during only three down-sampling operations, and thus retains a large-scale feature map to ensure that the infrared small targets still hold enough physical characteristics for the detection. In the next step, the output of each layer of the network is saved in the front channel using a dense connection on the feature map. Compared with common methods that simply integrate the deep network into the shallow network [9], [10], our method can better combine the location information of the shallow network and the semantic information of the deep network. Thus, our network achieves a more accurate positioning and classification effect.

As the last step, we introduce a feature attention mechanism [11] to obtain the importance of each feature channel, and to enhance useful features according to the importance degree. In this way we achieve an adaptive calibration of the feature channel, while the information of each layer saved by the dense link is further refined. Since the classification data of ImageNet and infrared small target data are highly different in regards of the image channel and the target size, we do not use ImageNet training data for a pre-training, but use the infrared small target data set to directly train from scratch. In order to train the proposed detection network MNET from scratch, the single-phase detection algorithm YOLO [12] is adopted in the detection part. As YOLO does not need to create area suggestions in a first step, but can directly perform target detection, its detection speed is fast. In combination with the feature extraction backbone network MNET it can meet the real-time requirements of infrared small target detection.

In general, the main contributions of this work are:

(1) Since target classification varies from target detection, we specifically propose a feature extraction backbone network for the detection of infrared small targets, which shows higher sensitivity regarding location information of the target than commonly used classification backbone networks.

(2) Due to the small size, infrared small targets might disappear in the deep layers of the classification network. For this reason, we limit the down-sampling operations and obtain a large enough feature map.

(3) We use dense networks to store the output of each layer of the network and combine shallow and deep networks to extract semantic and location information. At the end of the network, we introduce a feature attention mechanism to achieve adaptive calibration of the feature channel and to further purifying the information of each layer saved by the dense network.

(4) Since there is a large difference between general classification data and infrared small target data, we do not use pre-trained networks, but use infrared small target data to train our network from scratch.

The remainder of this paper is organized as follows. Section 2 briefly introduces the related methods of infrared small target detection. Section 3 presents the details of infrared small target detection algorithm MNET based on the feature fusion convolution network designed in this paper. Section 4 describes experiments conducted on a number of real infrared images covering complex situations, such as sea-sky background, clouds, drastic changes in target scale, and target occlusion or flying out of the field of view. Section 5 analyses the experimental results. Finally, Section 6 concludes the paper.

II. RELATED WORK

Researchers have developed many robust methods for infrared small target detection, such as Bayesian estimation [13], Kalman filtering [14], morphological filtering [15], high-pass filtering [16], maximum mean/maximum median filtering [17] and its extension [18], wavelet-transform-based algorithm [19], [20], morphological method [21], principal component analysis (PCA)-based method [22], and top-hat transform [21], [23], [24].

The features that are used in the traditional target detection algorithms described above are manually designed.

The performance of such algorithms mainly depends on the prior knowledge of the designer, and on manual tuning. Thus, only a small number of parameters can appear in the design of the feature. However, deep learning accepts the original form of the data as the algorithmic input [25], and abstracts the original data layer by layer for the final feature representation. It ends with the mapping of features to targets.

In 2012, Hinton et al. won the ImageNet image classification competition with their proposed CNN AlexNet [4] and surpassed the classification rate of the algorithm that ranked second place by nearly 12%. Since then, the dominance of CNNs in the field of computer vision has continued. Subsequently, well-known networks such as VGGNet [26], GoogLeNet [27], and ResNet [4] emerged.

In the field of target detection, detection methods based on convolutional neural networks are mainly divided into two categories: one stage methods and two stage methods. Two stage methods were introduced first. Examples for two stage methods are networks such as Fast-RCNN [28], or Faster-RCNN [29]. Two stage methods first generate a series of candidate frames as samples, and then classify the samples using CNNs. Due to the large number of candidate frames to be generated, these methods are slow and cannot meet real-time detection requirements.

In order to address this problem, researchers have proposed one stage methods such as SSD [30] or YOLO [12]. One stage methods do not require the region proposal stage but can directly generate class probabilities and position coordinates for the target object. After a single test, the final detection result can be obtained directly, this increases the detection speed. SSD can use six differently sized feature maps for the prediction and can thus better adapt to the target size. For this reason, SSD is preferred by researchers. Since deep learning is widely used within practical applications, such as face detection, vehicle detection, and fruit detection, more and more people are also trying to combine deep learning with infrared small target detection.

Lin [31] *et al.* designed a seven-layer CNN for infrared small target detection. Wang [32] *et al.* proposed a single infrared image small target detection method based on depth convolution and used a neural network to extract the target features. Qi [33] *et al.* proposed a fast saliency detection model with a simple 5 × 5 convolutional kernel to obtain the saliency map of the input image, in which the targets are enhanced while the background is suppressed. Zhang [34] *et al.* use Fast R-CNN for long-wave infrared image detection. Redmon and Farhadi [35] use a darknet-19-based YOLOv2 for infrared small target detection.

However, all these approaches are rather straightforward, as they simply apply well-known detection methods to infrared small targets and perform some fine-tuning. Moreover, the scale of the targets is kept relatively large for these approaches. Thus, the difference between infrared small targets and general data is not considered. For this reason, the feature extraction backbone network MNET that is customized for infrared small targets is proposed in this paper

III. METHODOLOGY

The overall objective of this study is to define MNET, a new feature extraction backbone network for infrared small target detection that provides high accuracy, good real-time performance, and adaptability to scale changes. In this section, the structure of MNET is introduced first, subsequently important design principles are described in detail.

A. MNET

Early target detection algorithms based on convolutional neural networks have a poor detection effect on small targets. Later, FPNs [10] were proposed, while other algorithms added deep neural networks as well as shallow networks for solving this problem. The intention was to perform small target detection using a shallower network, but higher resolution in the layer. Because the semantic information of shallow networks is weak, they cannot correctly identify the target. Shallow networks and deep networks with strong semantic information are thus superimposed to improve the expressive ability of the shallow network. However, due to the excessive number of down-sampling operations that is needed, the small target is likely to disappear within the deep network (the size of the infrared small target is smaller than the smallest target that can be identified by the detection network). To solve this problem, MNET takes the original image with a size of 456 × 456 pixels as input and converts it into a feature image of size 57 × 57 pixels using three down-sampling and corresponding convolution operations. In this process, the shortcut residual structure is added to alleviate the degradation problem caused by the deep network. After this step, we do not perform any further down sampling operations, but use a dense network on the feature map of this size to preserve the output information of each layer network while deepening the network. Finally, the feature attention mechanism is introduced to realize the adaptive calibration of the feature channel, and to further refine the information of each layer that is saved by the dense network. Because the direct training detection network may not converge, YOLO is used in the detection part of the algorithm and batch normalization [36] is added in each layer, to allow for the network training to process faster and more stable. The structure of MNET is shown in Figure 1. Figure 2 shows the specific parameters of MNET. Each component and the corresponding design principles are described in detail below.

B. M MODULE

Most detection networks use VGG or the residual network ResNet as their infrastructure in order to take differently sized targets into account. When connecting the deep network to the shallow network, the output of the deep network is simply up-sampled and merged with the output of the shallow network. However, we intend to save as much information of each output layer on the scale of 57 × 57 pixels as possible, to better combine the shallow edge features with the deep semantic information. In this regard, we are inspired by the

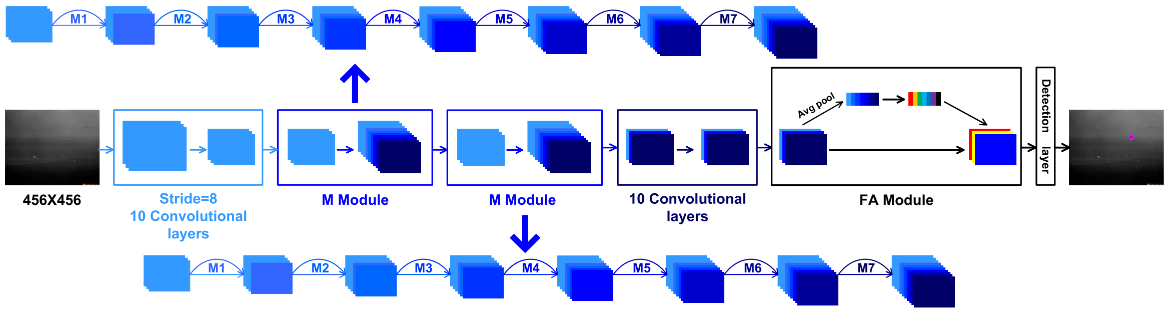
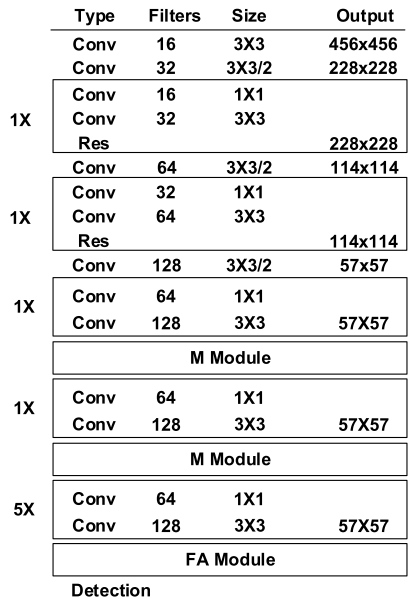


FIGURE 1. MNET structure.



	Type	Filters	Size	Output
	Conv	16	3X3	456x456
	Conv	32	3X3/2	228x228
1X	Conv	16	1X1	
	Conv	32	3X3	
	Res			228x228
	Conv	64	3X3/2	114x114
1X	Conv	32	1X1	
	Conv	64	3X3	
	Res			114x114
	Conv	128	3X3/2	57x57
1X	Conv	64	1X1	
	Conv	128	3X3	57X57
	M Module			
1X	Conv	64	1X1	
	Conv	128	3X3	57X57
	M Module			
5X	Conv	64	1X1	
	Conv	128	3X3	57X57
	FA Module			
	Detection			

FIGURE 2. MNET parameters.

dense connection in DenseNet [37]. In a dense connection, the feature map of each layer of the input network is the output feature map of all previous layers, while its own feature map serves as the input feature map for all subsequent layers. The feature fusion module used in this paper is named M Module and is shown in Figure 3.

The relationship between the output and input in the M Module is as follows:

$$X_n = M_c(\delta(CX_{n-1}) + X_{n-1}) \tag{1}$$

where X_{n-1} is the input feature map, X_0 to X_n describe the output of the n$^{\text{th}}$ layer, until X_7. C is a convolution operation, δ is a leaky activation function, $M_c(a, b) = Concat(a, b)$ indicates that two feature maps are concatenated.

The authors of DenseNet originally planned to use this dense connection to alleviate the problem of gradient disappearance, and to strengthen the network's expression while deepening the network. We have found that dense connections are suitable for the combination of shallow and deep information on the same scale and provide good recognition and localization abilities for infrared small target detection. We use two M Modules in MNET. On the one hand, it is because we want to use deeper network layers to improve the expression effect of the network. On the other hand, if only one M Module is used, the number of channels in the network will be too large and the speed will be slow. In addition, the dense connection provides a normalization effect, which is beneficial when the form of the detection network is directly trained, without using a pre-training network.

We design four different M Modules for experimental comparison.

As shown in Figure 4, the feature size of the input M Module is 57 × 57 pixels and 128 channels. Structure A consists of seven M-groups, each of which includes a Conv3-dense operation. Feature extraction is performed using a 3 × 3, 128-channel convolutional layer. The output of the set of convolutional layers is connected in series with the output of the upper set of dense layers by the dense layer. Each time the dense series is connected, the number of the feature map channels can be increased by 128. The final output is a feature map of 57 × 57 pixels and 1024 channels. Structure B adds a 1 × 1 convolution bottleneck layer to each group of the basic structure A. Since the series operation of the dense network increases the number of channels rapidly, adding the bottleneck layer before the feature extraction layer can significantly reduce the amount of calculations needed. In addition, the activation function can be used to integrate more nonlinearity into the network and to improve the ability of expression of the network. The final output is a feature map of 57 × 57 pixels and 1024 channels. Structure C replaces the bottleneck layer Conv1 of 128 channels in B with a 64-channel Conv1 bottleneck layer, and finally outputs a characteristic map of 57 × 57 pixels and 1024 channels. Structure D replaces the 128-channel convolution layer Conv3 in structure B with a 64-channel convolution layer Conv3. The final output is a 57 × 57 pixels and 576 channels feature map.

C. FA MODULE

After the feature fusion operation of the M Module, the characteristics of the shallow network and the deep network are preserved within each channel of the feature map. The next step is to add the feature attention mechanism to obtain the importance of each feature channel, and then to enhance

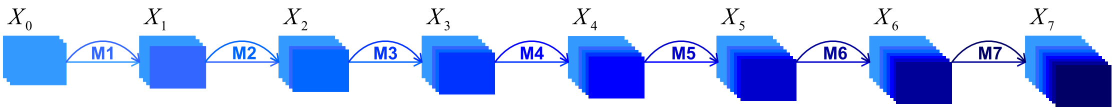
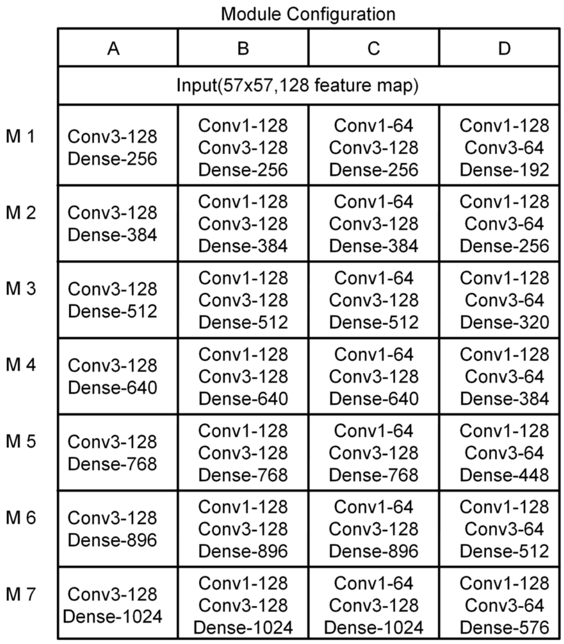


FIGURE 3. M Module structure.



Module Configuration

	A	B	C	D
	Input(57x57,128 feature map)			
M 1	Conv3-128 Dense-256	Conv1-128 Conv3-128 Dense-256	Conv1-64 Conv3-128 Dense-256	Conv1-128 Conv3-64 Dense-192
M 2	Conv3-128 Dense-384	Conv1-128 Conv3-128 Dense-384	Conv1-64 Conv3-128 Dense-384	Conv1-128 Conv3-64 Dense-256
M 3	Conv3-128 Dense-512	Conv1-128 Conv3-128 Dense-512	Conv1-64 Conv3-128 Dense-512	Conv1-128 Conv3-64 Dense-320
M 4	Conv3-128 Dense-640	Conv1-128 Conv3-128 Dense-640	Conv1-64 Conv3-128 Dense-640	Conv1-128 Conv3-64 Dense-384
M 5	Conv3-128 Dense-768	Conv1-128 Conv3-128 Dense-768	Conv1-64 Conv3-128 Dense-768	Conv1-128 Conv3-64 Dense-448
M 6	Conv3-128 Dense-896	Conv1-128 Conv3-128 Dense-896	Conv1-64 Conv3-128 Dense-896	Conv1-128 Conv3-64 Dense-512
M 7	Conv3-128 Dense-1024	Conv1-128 Conv3-128 Dense-1024	Conv1-64 Conv3-128 Dense-1024	Conv1-128 Conv3-64 Dense-576

FIGURE 4. M Module parameters.

useful features according to this importance. This means, the network may utilize the global information to selectively enhance beneficial features. In this way, the adaptive calibration of the feature channel can be realized, and the characteristics of each layer extracted by the M Module are further purified. Therefore, the FA Module is added at the end of the backbone network; its structure is shown in Figure 5. The characteristics of the FA Module are calculated as follows:

$$\begin{cases} X_1 = \delta(C_2(\delta(C_1(PX_0)))) \\ X_2 = X_0 + UX_1 \end{cases} \tag{2}$$

P is the global average pooling, δ is the leaky activation function, and C_1 is the convolution operation of 1×1, 8, C_2 is the convolution operation of 1×1, 128. U is an up-sampling operation with a magnification of 57.

The input X_0 of the FA Module is a 57×57, 128-channel feature map. The filters learned by each channel within a typical CNN operate on local receptive fields. Each feature map cannot utilize the context information of other feature maps. To solve this problem, each feature map is first compressed by a global average pooling, so that the feature map is transformed into a real number column of $1 \times 1 \times 128$. In theory, this number should have a global receptive field, which allows the feature map of the shallow network to also utilize global feature information. To take advantage of the information gathered during the extrusion operation, the excitation operation is used to fully capture channel dependencies. First, the dimensionality reduction operation of a 1×1, 8-channel convolutional layer and the dimensionality lifting operation of a 1×1, 128-channel convolutional layer are used. Here, the 1×1 convolutional layer can take over fully connecting and acquiring the importance of each feature channel. At the same time, the leaky activation function is added to the convolution operation to improve the nonlinearity of the module. Relative to the ReLU activation function, leaky retains a small negative value that can alleviate the ''dead'' ReLU problem. After up-sampling, to restore the size of 57×57 pixels, X is obtained. Finally, X_1 is added to the original input X_0 to enhance useful features in the original input and to obtain the final output.

D. WHY USING YOLO?

For the detection phase, the one-step algorithm YOLO is used as detection method. Two-step detection algorithms such as Fast-RCNN, or Faster-RCNN need to first obtain candidate regions using methods such as Regional Proposal Network (RPN). These candidate regions are then classified using high quality classifiers, which contributes to a large computational overhead. Thus, these methods are not suitable for real-time detection. YOLO combines the task of extracting candidate regions and classifying them within one network to convert detection problems into regression problems. It does not require the proposed area to provide bounding box coordinates and category probabilities. Using regression directly, the detection speed is faster, and is thus more suitable for real-time requirements of infrared small target detection. Moreover, studies have shown that only single-step algorithms can successfully converge without prior training [7]. This is supposed to be due to the RoI (Regions of Interest) pooling in the two-stage methods. RoI pooling generates features for each region proposal, which hinders the gradients from being smoothly back-propagated from a region-level to the convolutional feature maps. Such proposal-based methods work well with pre-trained network models because the parameter initialization is good for those layers before the RoI pooling. However, this is not true for a training from scratch.

We select YOLO as our one-step algorithm instead of the popular SSD. SSD uses six feature maps at different scales to detect targets of different sizes, which is significant for general data sets containing large, medium, and small targets. However, this is not of particular significance for infrared small target detection.

The YOLO detection algorithm is shown in Figure 6. For the YOLO network we divide the images of each training set

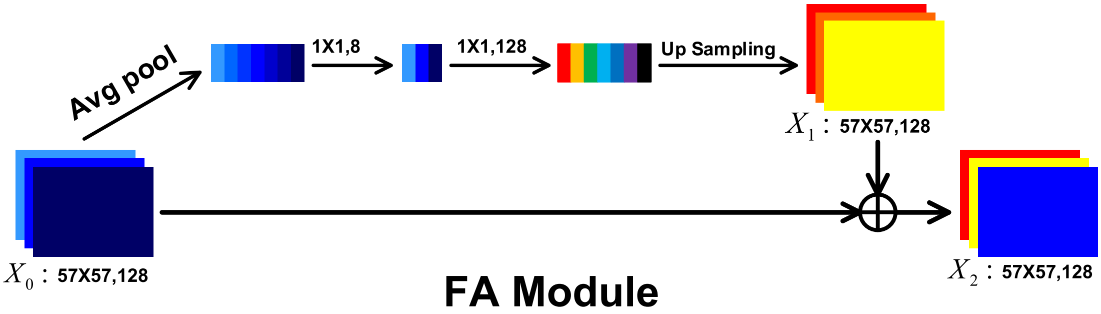


FIGURE 5. FA Module structure.

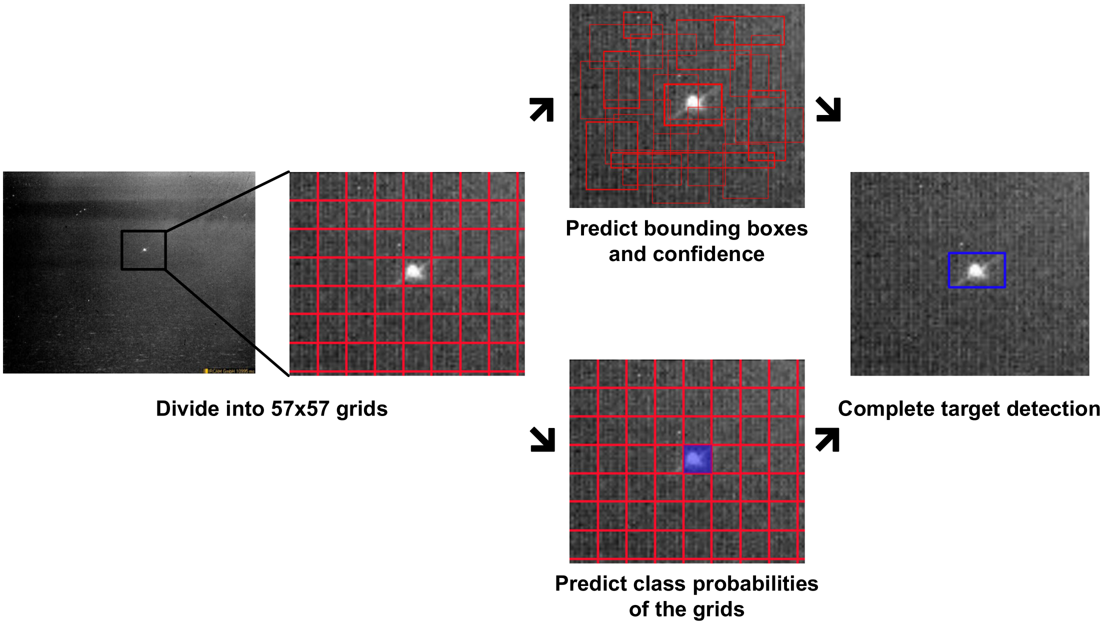

Divide into 57x57 grids
Predict bounding boxes and confidence
Complete target detection
Predict class probabilities of the grids

FIGURE 6. YOLO detection principle.

into a 57 × 57 grid. If the center of a target falls within a cell of the grid, then this grid is responsible for detecting the target. Each grid predicts three bounding boxes and corresponding confidences, as well as class probabilities. The confidence level is defined as follows:

$$Confidence = p_r(Object) \times IoU^{truth}_{pred} \qquad (3)$$

When a target falls within a cell of the grid, $p_r(Object)$ is 1, otherwise it is 0. IoU^{truth}_{pred} is used to indicate the coincidence between the reference and the prediction bounding boxes. Confidence reflects the accuracy of the prediction bounding box containing objects. When multiple bounding boxes detect the same target, YOLO uses the non-maximum suppression method to select the best bounding box.

IV. DATASET AND EXPRIMENT SETUP

In this section, the data sets used within the scope of this article as well as the network training parameter settings and the evaluation criteria are introduced.

A. IMAGE DATASET

In this study, multiple sets of infrared small target image sequences of 640 × 512 pixels are acquired at a beach using an infrared camera. A total of 6 groups of 29, 630 images are selected, and 5 representative images are selected from each sequence as shown in Figure 7. Further, 500 sheets are selected as the training set, while the others are used as test sets.

The statistics of the test set are summarized in Table 1. In the image sequence, the target flies out of the field of view or is occluded; hence, the number of images containing the target to be detected is slightly smaller than the total number of images. Sequences 1 to 5 comprise sea-sky backgrounds, including sea-sky-line and sea clutter. Sequence 6 comprises sea-sky and cloud backgrounds, including sea clutter, sea-sky-line, and cloud edge. The minimum size of the target is 2 × 2 pixels and the maximum size is around 25 × 25 pixels.

B. DATA AUGMENTATION

Deep learning requires a large amount of data to conduct appropriate training. Therefore, we use data augmentation technology to increase the amount of training data. Augmentation is a process of generating new instances from raw data through various transformation methods, such as rotation,

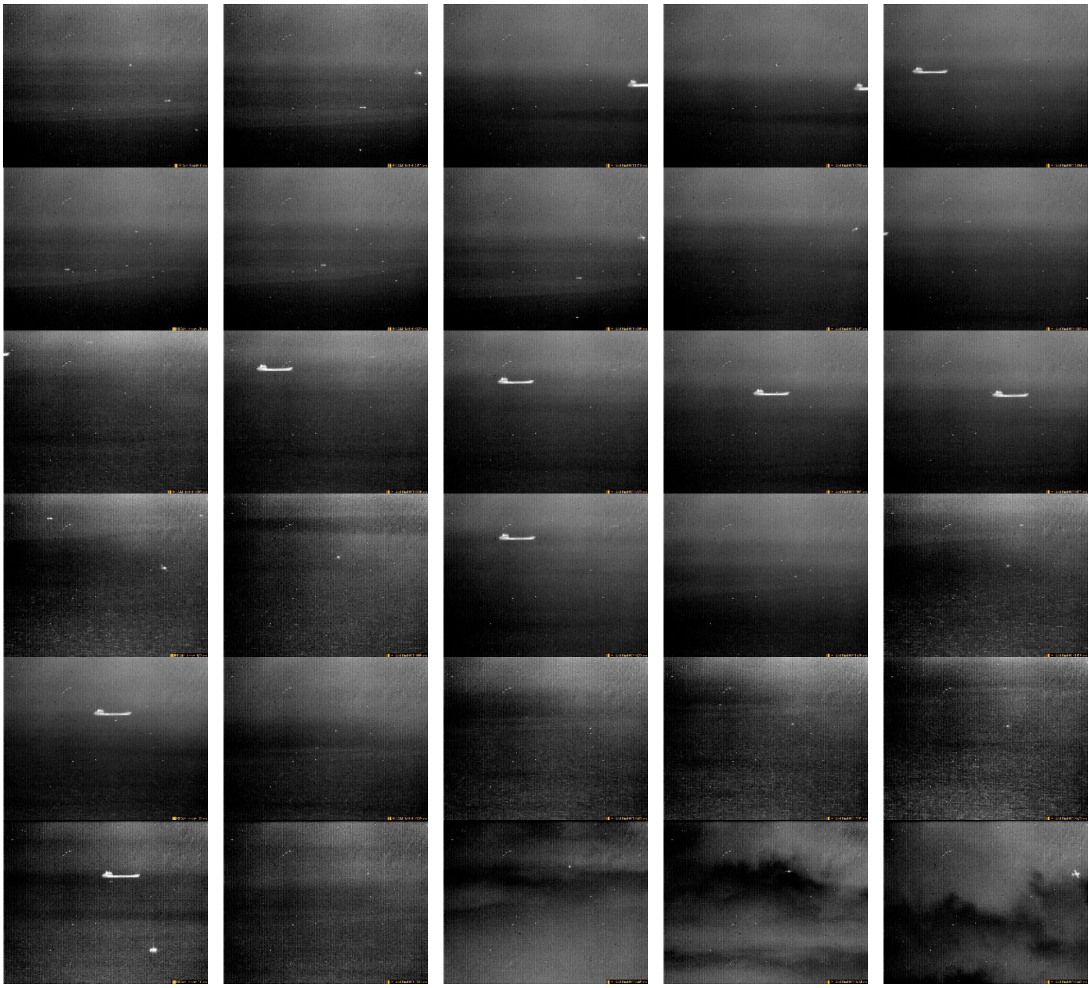


FIGURE 7. Partial infrared small target dataset used in this paper.

TABLE 1. Number and size range of actual targets for test sets.

Sequence	Total number	Number of target	Background	Scale of target
1	3543	2924	Sea and sky	2×2 ~ 22×26
2	4379	3774	Sea and sky	2×2 ~ 21×26
3	5022	4859	Sea and sky	2×2 ~ 18×15
4	8808	8259	Sea and sky	2×2 ~ 19×14
5	3520	3520	Sea and sky	2×2 ~ 11×10
6	3858	3688	Sea, sky and cloud	2×2 ~ 27×34

translation, and scaling. In this experiment, we rotate the datasets through 90°, 180°, and 270°, and we augment our datasets three times to obtain a small infrared target detection network with better detection performance.

C. HARDWARE CONFIGURATION

We use the following GPU to speed up the training process: Nvidia GeForce GTX 1080Ti. The entire program is written using the Darknet framework, and it is run in the Ubuntu environment.

D. TRAINING PARAMETERS

We train the infrared small target detection network using the dataset described above. Table 2 summarizes the parameter settings for training, such as the learning rate, batch size and momentum. In particular, 70000 training steps are performed to better analyze the training process. The model is trained after defining the training parameters. The learning rate decrease to 0.0001 after 40000 steps.

The choice of the batch size has a certain impact on the network performance. On the one hand, when the batch size

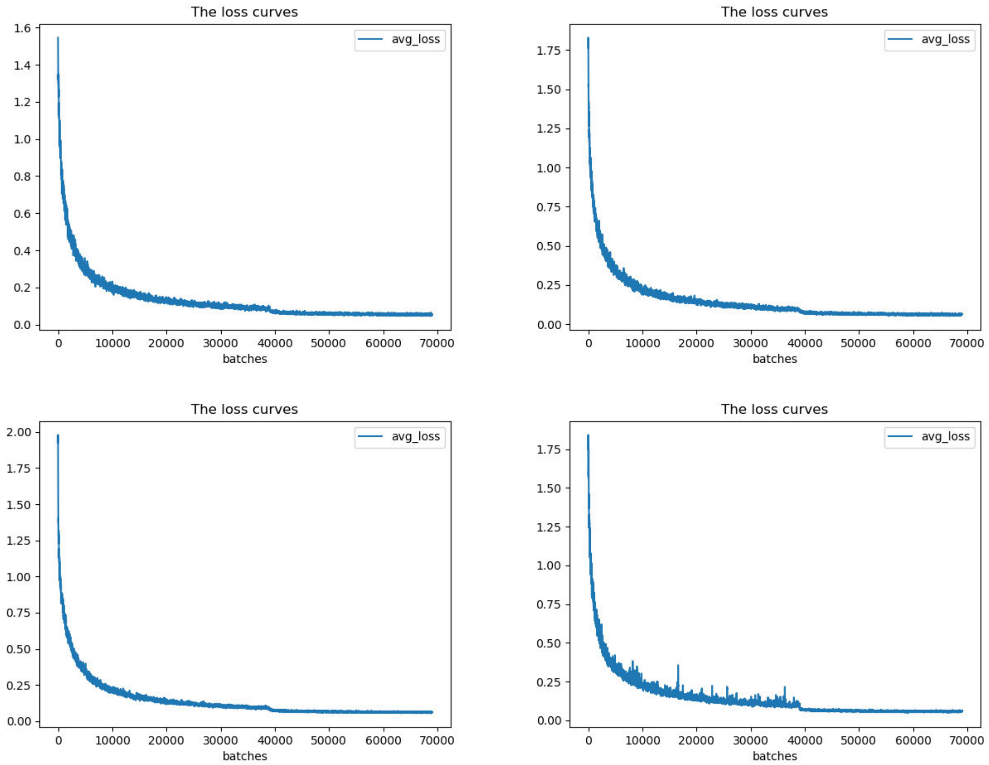


FIGURE 8. Loss functions of MNET-A, B, C, and D networks.

TABLE 2. Training parameters.

Learning rate	Batch	Momentum	Decay	Training steps
0.001	96	0.9	0.0005	70000

is too small, it is difficult to determine the direction of gradient correction; hence, convergence of the training process is difficult. On the other hand, a large batch size not only requires large storage space but also slows down parameter correction owing to the reduced number of iterations of an epoch. Therefore, choosing an appropriate batch size is of great significance for improving the convergence speed and accuracy of the network model. In this study, according to our hardware configuration, we finally chose a batch size of 96, divided into 8 sub-batches for training (equivalent to a batch size of 12).

E. EVALUATION METRICS

To verify the effectiveness of the proposed MNET detection network, we evaluate the algorithm from the perspective of accuracy, recall rate, and speed. For classification problems, the samples can be divided into four types: true positive (TP), false positive (FP), true negative (TN), and false negative (FN). Precision (P) and recall (R) are defined as follows:

$$P(precision) = \frac{TP}{TP + FP} \times 100\% \tag{4}$$

It represents the proportion of the number of correct samples to the number of positive samples.

Recall:

$$R(recall) = \frac{TP}{TP + FN} \times 100\% \tag{5}$$

It represents the proportion of the number of correct samples of positive class prediction to the total number of samples of positive class prediction.

In general, the algorithm cannot consider both the accuracy of the model and the recall. Improving the accuracy often reduces the recall and vice versa. To better evaluate the performance of the algorithm, we use F_1 values to consider both accuracy and recall. The F_1 value will increase only when the precision and recall are both extremely high.

F_1 is defined as

$$F_1 = \frac{2 \times R \times P}{R + P} \tag{6}$$

V. EXPERIMENTAL RESULTS AND ANALYSIS

In this section, our experiment based on the infrared small target dataset is presented and a comparison regarding the

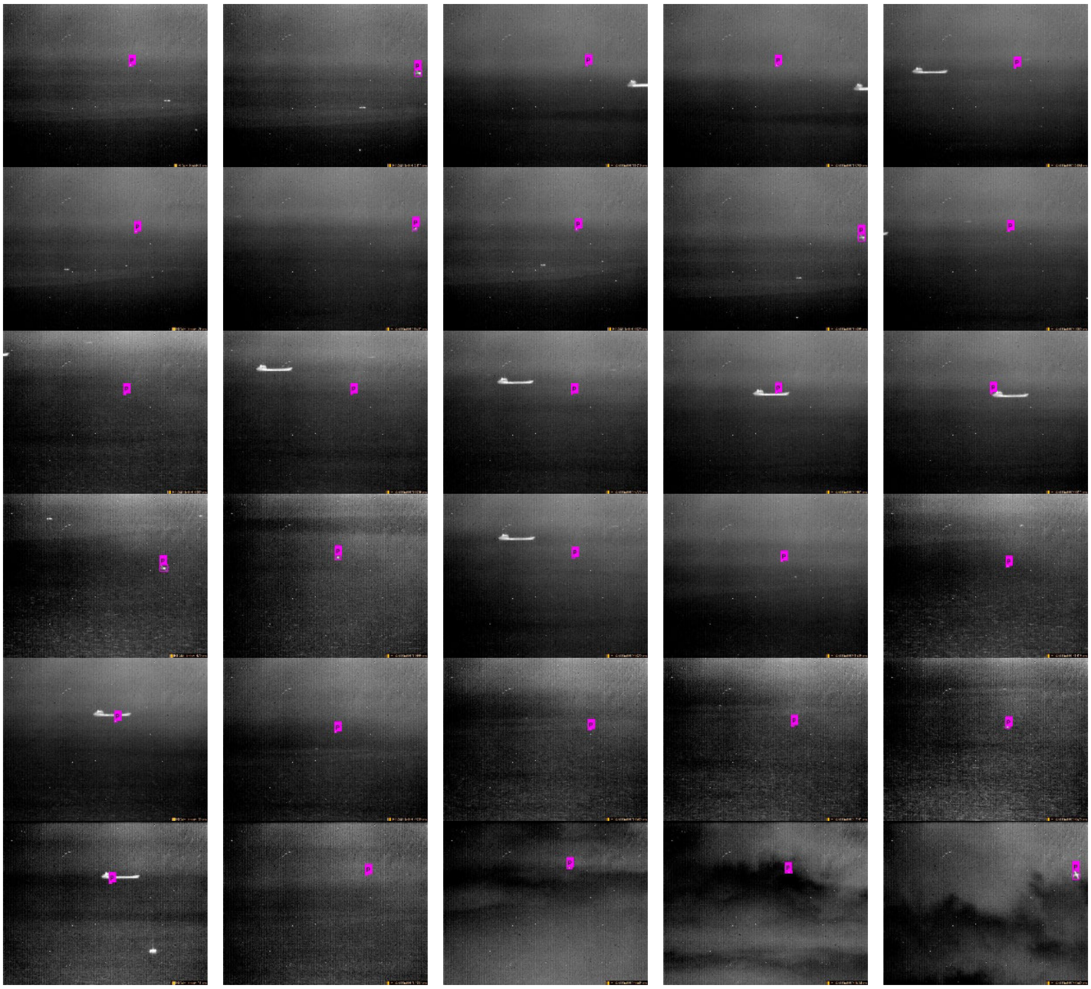


FIGURE 9. Test results for sequence 1-6.

detection performance of MNET with YOLOV2, YOLOV3 is provided.

A. TRAINING PROCESS

We use the dataset and the training parameters defined in the previous section to train the infrared small target detection networks MNET-A, B, C, and D designed in this study. Figure 8 shows the loss trend during training. The four networks eventually converged to approximately 0.05. The value of the loss function decrease, which means that the accuracy of the network increases with the training times.

B. DETECTION RESULT

The results of 30 representative images selected previously are shown in Figure 9 and Table 3. Table 4 summarizes the average confidence of five images in each sequence (0 for missing or multiple judgments).

YOLOv2 [25] and YOLOv3 [38] have shown excellent performance in target detection algorithms based on CNNs. YOLO series algorithms take into account both accuracy and speed. In particular, YOLOv3 has shown better detection effects for small targets compared with YOLOv2. Therefore, we compare MNET with these two algorithms. The training parameters of YOLOv2 and YOLOv3 are the same as those of MNET, and the experimental results are summarized in Table 4. We find that the four proposed networks have obvious advantages over YOLOv2 and YOLOv3 for datasets with small targets and large-scale transformation, especially when the targets are out of the field of view and have a sea-sky background. Their accuracy, recall, and F_1 values are improved by varying degrees, and their speed is even twice as high as that of YOLOv3. Thus, they meet the requirements of real-time detection while guaranteeing high accuracy. The average accuracy of MNET-C is 99.39%, the average recall is 99.80%, and the average F_1 is 0.996. Compared with YOLOv3, MNet-D has obvious advantages and it outperforms the other three MNET models. The speed of MNET-C is 105 frames per second. The detection part used by MNET is the same as YOLO, and the obvious performance improvement proves the importance and effectiveness of the targeted design feature extraction backbone network. Without pre-training models, training from scratch also avoids the

TABLE 3. Detection confidence of different networks.

	1	2	3	4	5	6
MNET-A	100%	80%	99.2%	99.8%	100%	100%
MNET-B	100%	97.4	80%	100%	100%	100%
MNET-C	100%	99.6%	99.8%	99.8%	100%	99%
MNET-D	100%	99%	79.4%	97.2%	100%	98.8%
YOLOv3	80%	79%	39.6%	37.8%	93%	71.85
YOLOv2	71%	73.45	38.8%	74.2%	54%	38.6%

TABLE 4. (a) Sequence 1 test results of different networks. (b) Sequence 2 test results of different networks. (c) Sequence 3 test results of different networks. (d) Sequence 4 test results of different networks. (e) Sequence 5 test results of different networks. (f) Sequence 6 test results of different networks.

(a)

	TP	FP	FN	P	R	F1	Fps
MNET-A	2911	27	11	99.08%	99.62%	0.993	82.8
MNET-B	2904	18	20	99.38%	99.31%	0.993	92.0
MNET-C	2919	21	5	99.29%	99.83%	0.996	105.3
MNET-D	2922	47	2	98.42%	99.93%	0.992	123.6
YOLOv3	2462	114	462	95.57%	84.20%	0.895	54.9
YOLOv2	2502	0	422	100%	85.57%	0.922	115.5

(b)

	TP	FP	FN	P	R	F1	Fps
MNET-A	3749	72	22	98.11%	99.34%	0.987	82.5
MNET-B	3767	59	7	98.45%	99.81%	0.991	92.7
MNET-C	3766	34	8	99.11%	99.79%	0.994	105.6
MNET-D	3685	157	89	95.96%	97.64%	0.968	123.0
YOLOv3	3496	580	278	85.77%	92.63%	0.891	55.0
YOLOv2	3327	1	447	99.97%	88.16%	0.937	115.7

(c)

	TP	FP	FN	P	R	F1	Fps
MNET-A	4837	84	22	98.29%	99.55%	0.989	82.4
MNET-B	4859	77	0	98.44%	100%	0.992	93.3
MNET-C	4856	77	3	98.44%	99.94%	0.992	104.9
MNET-D	4847	90	12	98.18%	99.75%	0.990	124.1
YOLOv3	3789	100	1070	97.43%	77.98%	0.866	55.0
YOLOv2	3025	0	1834	100%	62.26%	0.766	112.1

(d)

	TP	FP	FN	P	R	F1	Fps
MNET-A	8216	8	43	99.90%	99.48%	0.997	81.8
MNET-B	8230	4	29	99.95%	99.65 %	0.998	92.2
MNET-C	8232	7	27	99.92%	99.67%	0.998	105.4
MNET-D	8166	40	93	99.51%	98.87%	0.992	122.0
YOLOv3	6523	836	1736	88.64%	78.98%	0.835	56.4
YOLOv2	6178	1	2081	99.98%	74.80%	0.855	115.2

(e)

	TP	FP	FN	P	R	F1	Fps
MNET-A	3500	32	20	99.09%	99.43%	0.993	81.8
MNET-B	3507	57	13	98.40%	99.63 %	0.990	93.4
MNET-C	3504	14	16	99.60%	99.55%	0.996	105.3
MNET-D	3489	4	31	99.89%	99.12%	0.995	122.7
YOLOv3	3314	1	206	99.97%	94.15%	0.970	55.1
YOLOv2	3137	0	383	100%	89.12%	0.942	111.1

(f)

	TP	FP	FN	P	R	F1	Fps
MNET-A	3676	0	12	100%	99.67%	0.998	82.8
MNET-B	3670	0	18	100%	99.51%	0.998	93.2
MNET-C	3688	0	0	100%	100%	1	106.4
MNET-D	3684	3	4	100%	99.89%	0.999	123.1
YOLOv3	3459	24	229	99.92%	93.79%	0.968	56.5
YOLOv2	3508	3	180	99.91%	95.12%	0.975	111.6

interference of different types of data on infrared small target data. Moreover, the proposal of a high-precision real-time detection algorithm is of great significance for infrared small target detection on other devices.

TABLE 5. Test results with FA Module added or not.

	P	R	F1	Fps
MNET-C	99.39%	99.80%	0.995	105.5
MNET-CA	98.64%	99.50%	0.991	111.4

The average F_1 of MNET-B, MNET-C, and MNET-D is higher than that of MNET-A, which reflects the importance of adding a bottleneck convolutional layer in the M Module. It not only reduces the amount of computation and improves the speed of network detection but also adds more nonlinearity to the network to make the network provide better expression effects. The addition of the bottleneck layer can increase the computational efficiency by up to 50%. In the M Module, the bottleneck convolutional layer Conv1 is primarily used to condense the output of the upper M group, and the convolutional layer Conv3 is mainly used to extract features. The effect of MNET-C is better than that of MNET-D, which means that the number of channels in Conv3 should be less than that in Conv1 when designing the network. In other words, a wider network than the bottleneck layer should be used when extracting the features.

In the six sequences tested in this experiment, the target underwent drastic scale changes. The image sequence records the process of the target changing back and forth between the point target of 2 × 2 and the sub-imaging target of 25 × 25. By contrast, YOLOv3 is close to MNET in detecting the sub-imaging targets, but it is widened by MNET in detecting the point targets. When the point target occupies only 2 × 2 pixels in the 640 × 512-pixel images, YOLOv3 cannot detect the target, whereas MNET-C still performs the detection task well. Moreover, the excellent detection results of the MNET Model in the entire process reflect its adaptability to scale changes. The background of sequences 1 to 5 is a sea-sky background. Neither the sea-sky background nor the sea-skyline have any effect on the results of MNET. In addition, MNET shows excellent detection effects for sequence 6 with the cloud background.

The effect of the FA Module is shown in Table 5, where MNET-CA is the MNET-C with the FA Module removed. The FA Module can increase the detection index by adding only small amount of calculation operations. This demonstrates the effectiveness of further adaptive purification of each layers of features preserved in the M Module.

Note that, in deep learning tasks, the detection confidence threshold is usually set to 0.5 or 0.75. In this experiment, the confidence threshold is set to 0.5. The mean confidence of MNET and YOLOv3 is more than 90%. The selection of the threshold has a small effect on the results. However, YOLOv2 is not sensitive to small targets; hence, in this experiment, YOLOv2 is in an under-fitting state, and its confidence is mostly in the range of 60%∼83%. When the detection threshold is set to 0.75, the average recall of the six sequence detection results is 22.27%, which is much lower than that of other networks.

VI. CONCLUSION

This paper proposes MNET, a new feature extraction backbone network for small infrared target detection. The backbone network of common detection networks originates from classification networks. Since the target classification is different from the target detection, such a network is not sensitive to location information of small targets. Therefore, the presented approach combines a shallow network with the characteristic information of a deep network and further purifies it. Instead of using a pre-training model during the training process, the infrared small target data is used directly to train the network from scratch. The experimental results show that MNET can be used for the detection of infrared small targets in real-time under sea-sky backgrounds. Furthermore, MNET shows strong adaptability to scale changes of the targets. The proposed detection network has obvious advantages over YOLOv3 in terms of accuracy and speed. Future work will focus on the use of existing models for tracking small infrared targets in video data. In addition, research on anti-interference identification will be conducted.

REFERENCES

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

[2] X. Shao, H. Fan, G. Lu, and J. Xu, ''An improved infrared dim and small target detection algorithm based on the contrast mechanism of human visual system,'' *Infr. Phys. Technol.*, vol. 55, no. 5, pp. 403–408, Sep. 2012.

[3] X. Zhang, Q. Ding, H. Luo, B. Hui, Z. Chang, and J. Zhang, ''Infrared small target detection based on an image-patch tensor model,'' *Infr. Phys. Technol.*, vol. 99, pp. 55–63, Jun. 2019.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, ''ImageNet classification with deep convolutional neural networks,'' in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[5] M. D. Zeiler and R. Fergus, ''Visualizing and understanding convolutional networks,'' in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, vol. 8689. 2014, pp. 818–833.

[6] Z. Li, C. Peng, G. Yu, X. Zhang, Y. Deng, and J. Sun, ''DetNet: Design backbone for object detection,'' in *Proc. Eur. Conf. Comput. Vis*, Sep. 2018, pp. 334–350.

[7] R. Zhu, S. Zhang, X. Wang, L. Wen, H. Shi, L. Bo, and T. Mei, ''ScratchDet: Training single-shot object detectors from scratch,'' 2018, *arXiv:1810.08425*. [Online]. Available: https://arxiv.org/abs/1810.08425

[8] Z. Shen, L. Zhuang, J. Li, Y.-G. Jiang, Y. Chen, and X. Xue, ''DSOD: Learning deeply supervised object detectors from scratch,'' 2017, *arXiv:1708.01241*. [Online]. Available: https://arxiv.org/abs/1708.01241

[9] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg, ''DSSD: Deconvolutional Single Shot Detector,'' 2017, *arXiv:1701.06659*. [Online]. Available: https://arxiv.org/abs/1701.06659

[10] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, ''Feature pyramid networks for object detection,'' in *Proc. 30th IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2117–2125.

[11] J. Hu, L. Shen, and G. Sun, ''Squeeze-and-excitation networks,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 7132–7141.

[12] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, ''You only look once: Unified, real-time object detection,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 779–788.

[13] K. E. Matthews and N. M. Namazi, ''A Bayes decision test for detecting uncovered-background and moving pixels in image sequences,'' *IEEE Trans. Image Process.*, vol. 7, no. 5, pp. 720–728, May 1998.

[14] K. Huang and X. Mao, ''Detectability of infrared small targets,'' *Infr. Phys Technol.*, vol. 53, no. 3, pp. 208–217, May 2010.

[15] B. Ye and J. X. Peng, ''Application of order morphology filtering on detection of small target and point target,'' *Proc. SPIE*, vol. 4554, Sep. 2001, pp. 94–99.

[16] L. Yang, J. Yang, and K. Yang, "Adaptive detection for infrared small target under sea-sky complex background," *Electron. Lett.*, vol. 40, no. 17, pp. 1083–1085, Aug. 2004.

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

[18] W. Meng, T. Jin, and X. Zhao, "Adaptive method of dim small object detection with heavy clutter," *Appl. Opt.*, vol. 52, no. 10, pp. D64–D74, 2013.

[19] F. A. Sadjadi, "Infrared target detection with probability density functions of wavelet transform subbands," *Appl. Opt.*, vol. 43, no. 2, pp. 315–323, Feb. 2004.

[20] X. Wang, Z. M. J. I. Tang, and L. Engineering, "Combining wavelet packet with higher-order statistics for small IR targets detection," *Infr. Laser Eng.*, vol. 38, no. 5, pp. 915–920, 2009.

[21] X. Bai and F. Zhou, "Analysis of new top-hat transformation and the application for infrared dim small target detection," *Pattern Recognit.*, vol. 43, no. 6, pp. 2145–2156, 2010.

[22] C. Wang and S. Qin, "Adaptive detection method of infrared small target based on target-background separation via robust principal component analysis," *Infr. Phys. Technol.*, vol. 69, pp. 123–135, Mar. 2015.

[23] L. Deng, Z. Hu, Z. Quan, Q. Zhou, and Y. Li, "Adaptive top-hat filter based on quantum genetic algorithm for infrared small target detection," *Multimedia Tools Appl.*, vol. 77, no. 9, pp. 10539–10551, May 2018.

[24] M. Zeng, J. Li, and Z. Peng, "The design of top-hat morphological filter and application to infrared target detection," *Infr. Phys. Technol.*, vol. 48, no. 1, pp. 67–76, Aug. 2006.

[25] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. 30th IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 7263–7271.

[26] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: https://arxiv.org/abs/1409.1556

[27] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1–9.

[28] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[29] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[30] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.

[31] L.-K. Lin, S.-Y. Wang, and Z.-X. Tang, "Point target detection in infrared over-sampling scanning images using deep convolutional neural networks," *J. Infr. Millim. Waves*, vol. 37, no. 2, pp. 219–226, Apr. 2018.

[32] W. Wang, H. Qin, W. Cheng, C. Wang, H. Leng, and H. Zhou, "Small target detection in infrared image using convolutional neural networks," *Proc. SPIE*, vol. 10462, Oct. 2017, Art. no. 1046250.

[33] S. Qi, W. Zhang, and G. Xu, "Detecting consumer drones from static infrared images by fast-saliency and HOG descriptor," in *Proc. Int. Conf. Commun. Inf. Process.*, Nov. 2018, pp. 62–66.

[34] H. Zhang, C. Luo, Q. Wang, M. Kitchin, A. Parmley, J. Monge-Alvarez, and P. Casaseca-de-la-Higuera, "A novel infrared video surveillance system using deep learning based techniques," *Multimedia Tools Appl.*, vol. 77, no. 20, pp. 26657–26676, Oct. 2018.

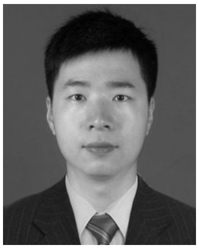
[35] J. Ryu and S. Kim, "Data driven proposal and deep learning-based small infrared drone detection," *J. Inst. Control, Robot. Syst.*, vol. 24, no. 12, pp. 1146–1151, Dec. 2018.

[36] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Int. Conf. Mach. Learn.*, Jun. 2015, pp. 448–456.

[37] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2016, pp. 4700–4708.

[38] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*. [Online]. Available: https://arxiv.org/abs/1804.02767

KAIDI WANG received the bachelor's degree in detection guidance and control engineering from Northwestern Polytechnical University, Xi'an, China, in 2016, where he is currently pursuing the master's degree with the School of Aerospace. He has authored or coauthored two articles in scientific journals and conference proceedings. His current research interest includes infrared target detection based on deep learning.



SHAOYI LI received the bachelor's degree in measurement and control technology from Southwest Jiaotong University Chengdu, China, in 2008, and the Ph.D. and master's degrees in navigation, guidance, and control from Northwestern Polytechnical University, Xi'an, China, in 2015 and 2011, respectively, where he is currently an Assistant Researcher with the School of Aerospace. He has authored or coauthored more than ten articles in scientific journals and conference proceedings. His current research interests include ATR based on infrared image, artificial intelligence, and image processing.

SAISAI NIU received the Ph.D. degree in aerospace manufacturing engineering from the Nanjing University of Aeronautics and Astronautics (NUAA), Nanjing, Jiangsu, China, in 2013. He is currently a Senior Engineer with the Shanghai Institute of Spaceflight Control Technology, China Aerospace Science and Technology Corporation. He has authored or coauthored more than 30 articles in peer-reviewed journals and conferences. His research interests include infrared target characteristic, infrared scene modeling, semi-physical simulation photoelectric detection guidance, and machine learning.

KAI ZHANG received the bachelor's, master's, and Ph.D. degrees in navigation, guidance, and control from Northwestern Polytechnical University, Xi'an, China, in 2009, 2004, and 2001, respectively, where he is currently an Associate Professor with the School of Aerospace. He has authored or coauthored more than 20 articles in scientific journals and conference proceedings. His current research interests include infrared imaging, artificial intelligence, and infrared scene modeling and simulation.