IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 22 June 2024, accepted 23 July 2024, date of publication 30 July 2024, date of current version 8 August 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3435872

RESEARCH ARTICLE

A Strategy for Training Dim and Small Infrared Targets Detection Networks Under Sequential Cloud Background Images

WENXIN ZHAO[1,2,3], **XUEFENG LAI**[1,2], **XULONG ZHAO**[1,2], **YUCHENG XIA**[1,2], **AND JINMEI ZHOU**[1,2]

[1]Key Laboratory of Science and Technology on Space Optoelectronic Precision Measurement, CAS, Chengdu 610209, China
[2]Institute of Optics and Electronics, Chinese Academy of Sciences, Chengdu 610209, China
[3]School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100049, China

Corresponding author: Xuefeng Lai (laixuefeng@ioe.ac.cn)

This work was supported by the Institute of Optics and Electronics, Chinese Academy of Sciences.

ABSTRACT The existing open-source infrared dim and small target dataset have limited data capacity and insufficient scenarios. To improve the generalization and robustness of the network, it is usually needed to establish a new dataset for the specific application scenario. When using the conventional training method in the new dataset, there are problems such as low target-to-background proportion in full-sized images, inaccurate labels due to nonexpert annotators, and inadequate diversity of data, which affect the training efficiency and network performance. This article proposes a training strategy based on the newly established dataset to overcome these problems. Specifically, small-sized image transfer learning is proposed to increase the small target proportion, shorten training time, and improve training efficiency. Moreover, a label refinement method based on loss evaluation metrics is adopted to reduce the impact of inaccurate labeling on network training. In addition, an iterative training method is proposed by supplementing new false alarm and miss detection data into the dataset between each iteration to further improve the training performance. The experiments are carried out and the results show that the above method can effectively shorten training time, improve training efficiency and performance of infrared dim and small target detection networks.

INDEX TERMS Infrared dim and small targets, target detection, training strategy, sequential images, cloud background.

I. INTRODUCTION

Infrared dim and small target detection is one of the most important directions in the infrared research field. By accurately detecting infrared dim and small targets, it is possible to achieve monitoring and warning of long-distance targets, which is of great significance in ensuring the safety of aviation. In recent years, with the development and progress of artificial intelligence science, deep learning convolutional neural networks have rapidly gained a significant advantage over conventional methods in the field of computer vision. Deep learning has also gradually been applied to infrared dim and small target detection.

Liu et al. [1] proposed a correlation filter-based ensemble tracker with multi-layer convolutional features for thermal infrared tracking. A fusion method based on Kullback–Leibler divergence and a simple scale estimation strategy are provided to improve the tracking performance. Zhao et al. [2] designed a convolutional neural network TBC-Net for infrared dim and small target detection, which consists of a target extraction module (TEM) and a semantic constraint module (SCM). The SCM imposes a semantic constraint on TEM by combining the high-level classification task and solves the problem of the difficulty of learning features caused by class imbalance problem., Zhao et al. [2] designed a feature extraction network based on the YOLO detection framework [4]. The feature fusion network was improved by utilizing the idea of multi-path aggregation. The number of detection output layers was adjusted to enhance

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

the reuse of feature information. Zhang et al. [5] introduced CA-U2-Net, a refinement of the U2-Net tailored to make the network more focused on infrared dim and small targets. Bao et al. [6] improved DNA-Net by retaining the densely nested attention network structure in Dense Nested Attention Network (DNA-Net) and introducing a Swin-transformer in the feature extraction stage to enhance feature continuity resulting in better performance.

The above methods of infrared dim and small targets mainly focus on the network model [7], whereas some scholars have studied the training strategies of network models, including distributed learning [10], knowledge distillation [11], data augmentation, and optimization [12], and hyperparameter optimization [13]. Distributed learning and knowledge distillation usually require more training time and compute resources, whereas hyperparameters need to be optimized specifically for the given problem and model. Compared to other methods, augmenting and optimizing the dataset is more universally applicable.

Recently, scholars have also researched data augmentation methods. The Cutout [14] method randomly masks parts of the input image during the training process, which is simple and easy to implement. This method forces the model to learn more robust representations, which reduce overfitting and can be combined with other data augmentation methods. The Mixup [15] method mixes two input samples in proportion to generate a new sample. This method improves the model's generalization ability and reduces the impact of label noise. The GridMask [16] method randomly deletes subregions of the input image and fills them with patches from other training data. This method considers both global and local contexts, encouraging the learning of richer visual features and improving accuracy. Mosaic [17] concatenates and adjusts labels to generate diverse synthetic images with multiple data stitched together. This method increases data diversity, enriches context, and effectively utilizes resources.

The above methods mainly involve mixing, deleting, and other operations to the original image to enhance the generalization ability of the network model, mainly focusing on the final training performance whereas ignoring the analysis of training efficiency. The improvement of training efficiency can discover network problems timely, save computing resources, and reduce training time.

This article proposes a training strategy based on the newly established dataset to improve the training efficiency and network performance, including small-sized image transfer learning, label refinement based on loss evaluation metrics, and iterative training based on the updated dataset. The experiments are carried out and the results show that the above method can effectively shorten training time, improve training efficiency and performance of infrared dim and small target detection networks.

II. SEQUENTIAL INFRARED DIM AND SMALL TARGETS

According to the definition provided by SPIE (Society of Photo-Optical Instrumentation Engineers) [18], the size of infrared dim and small targets is less than 81 pixels, generally not exceeding 0.12% of the total number of pixels. The term "dim" indicates that there is a small difference between the target and the background, resulting in a low signal-to-noise ratio. The term "Small" indicates that the target occupies a small proportion, and in extreme cases, it can be as few as one or two pixels. Due to the limited number of pixels occupied by targets, these targets usually lack texture features, as shown in the Fig. 1.

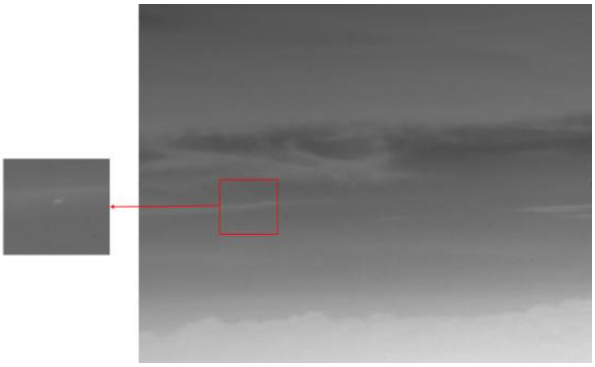

FIGURE 1. Infrared dim and small target.

Sequential images contain more motion information than single-frame images. In sky scenes, the motion characteristics of the target and background are different, and the image sequence contains the motion information of both. Through network learning, the motion information of small targets can be effectively utilized.

Image differencing [19] is a method of determining which regions have changed by comparing the pixel differences between consecutive frames. The infrared dim and small targets have relative motion characteristics with the moving platform, and the background's motion speed is usually smaller than the target's motion speed. This difference can be fully utilized for dim and small target detection. So, image differencing can be used to distinguish between the target and background, which facilitates the extraction of target information and is advantageous for subsequent target detection. Fig. 2 shows the original image and the differential image.

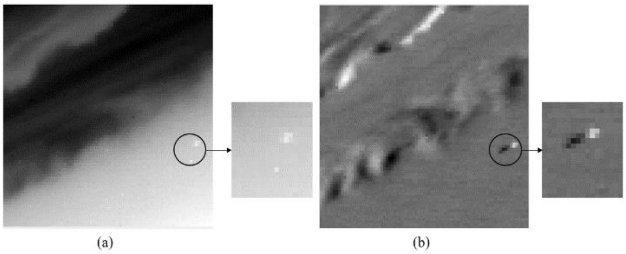

FIGURE 2. Original image (a) and differential image(b).

Therefore, this article uses differential operations to transform sequential images into sequential differential images, which improves the network's ability by extracting information of small and dim targets. In practical utilization, the first frame of the images is sequentially subtracted from the second, third, and fourth frames of the sequential images, as shown in Eq. (1). Where F is the original image sequence with a size of $\mathrm{M} \times \mathrm{N} \times 4$, G is the differential image sequence with a size of $\mathrm{M} \times \mathrm{N} \times 3$, and M and N are the width and height of the image respectively.

$$G = \begin{bmatrix} F(:,:,1) - F(:,:,2) \\ F(:,:,1) - F(:,:,3) \\ F(:,:,1) - F(:,:,4) \end{bmatrix} \tag{1}$$

Apart from the input method of images, suitable training strategies can make up the problems in the network training process, improve the accuracy and robustness of the network, and provide a reliable basis for subsequent decision-making.

III. INFRARED DIM AND SMALL TARGETS TRAINING STRATEGY

The performance of the network depends on the quality of the dataset. Different application scenarios have different characteristics and require different datasets. Although there are currently some open-source infrared target datasets, these datasets do not fully cover all scenarios and requirements. To better reflect the real situation, a dataset is built to collect infrared image data suitable for specific application scenarios. A good dataset should have the characteristics of diversity, coverage, high quality, balanced sample, and being updated timely. By optimizing the dataset, the accuracy, robustness, and adaptability of the model can be improved.

The impact of the dataset on training mainly includes three parts for the detection of infrared dim and small targets:

1) Low training efficiency of full-sized images. The pixels occupied by infrared dim and small targets are few, and most of the pixels belong to the background. The proportion of the target and background is unbalanced. Most of the time is spent on calculating the background during network training. Such a process slows down the learning of dim and small targets. In addition, the actual resources occupied by the targets are less, wasting storage space.

2) Inaccurate labeling. The correctness and accuracy of the labels have an impact on the network performance, which should be considered in the design of the training strategy. For network training, inaccurate labels will cause large loss and incorrect gradient direction of learning, resulting in training fluctuations. Therefore, timely refining inaccurate labels is significant for network training.

3) Incomplete dataset. The diversity and richness of data are crucial for training. In actual engineering, it is difficult to establish a completely sufficient and accurate dataset in the initial stage, thus continuous data supplementation is needed. In addition, false positive and false negative data is vital for training effectiveness, and how to mine and supplement false positive and false negative data in subsequent training is also a major challenge.

Therefore, this article conducts research on the above problems and proposes a training strategy related to dataset optimization, including small-sized image transfer learning, label refinement based on loss evaluation metrics, and iterative training based on updated dataset. The specific flowchart is shown as Fig. 3.

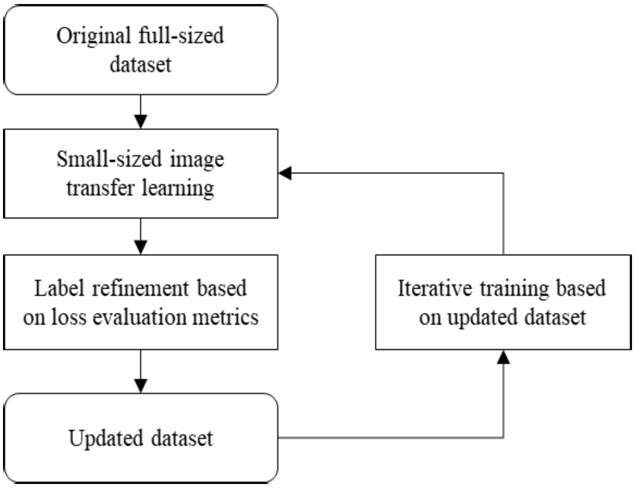
Original full-sized dataset
Small-sized image transfer learning
Label refinement based on loss evaluation metrics
Iterative training based on updated dataset
Updated dataset

FIGURE 3. **Flow chart of the training strategy.**

A. SMALL-SIZED IMAGE TRANSFER LEARNING

Full-sized images are frequently used to establish datasets for infrared dim and target detection algorithms, yet there are some problems with this approach.

First of all, it usually requires a large and rich dataset for the robustness of network. The scale of full-sized datasets increases with the continuous improvement of infrared image resolution, leading to increased storage requirements. Additionally, the computing time needed for the training network also increases with the increase of the scale of the dataset. For practical applications, processing a full-sized dataset requires more computing resources and training time.

Secondly, full-size images result in a low target-to-background proportion. In infrared dim and small target images, the number of pixels occupied by the target is few, thus the target occupies a small proportion in the full-sized images. The majority of the image is taken up by the background. The background is mainly trained and calculated during network training whereas the target is less trained, which lowers training efficiency.

Considering the above, we can choose to extract the region of interest, the area where the target is located when building a dataset. This can reduce irrelevant background information, improve the proportion of targets, and facilitate the learning of target information. Taking into account the parameter-sharing characteristic of convolutions and the small size of infrared targets, this article transforms full-sized images into small-sized images for transfer learning.

The same weights are used for different positions of the input data when convolution is carried out. This way, the network can learn the same feature representations regardless of the target position in the input images. Parameter sharing ensures that the small-sized images and full-sized images learn the same set of parameters. Moreover, even

TABLE 1. Comparison of parameters between full-sized images and small-sized images.

	Full-sized images	Small-sized images (1)	Small-sized images (2)	Small-sized images (3)	Small-sized images (4)
Images size	640x512x4	80x80x4	64x64x4	32x32x4	16x16x4
Single image pixels	327680	6400	4096	1024	256
Target size	3x3	3x3	3x3	3x3	3x3
Images scale	2.5MB	50KB	40KB	10KB	2.5KB
Target proportion	0.0275‰	1.4‰	2.2‰	8.8‰	35.2‰

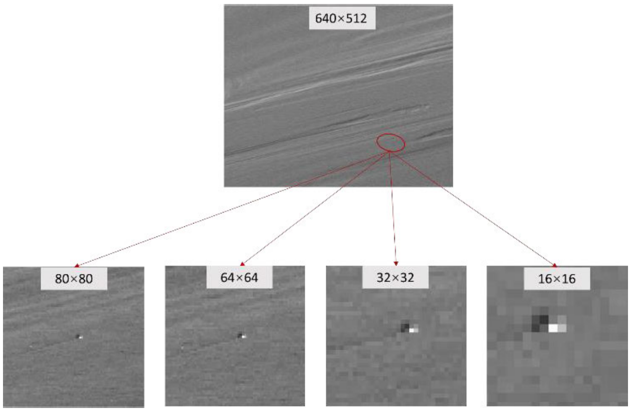

FIGURE 4. Full-sized and small-sized images.

with reduced size, the small-sized images can still contain small targets and sufficient background information due to the small proportion of infrared targets. The example of small-sized images and full-sized images are shown in Fig. 4. Table 1 shows the comparison of parameters between large-sized images and small-sized images.

B. LABEL REFINEMENT BASED ON LOSS EVALUATION METRICS

In the task of object detection, accurate and appropriate data labels are crucial. In the detection of dim and small infrared targets, there are challenges posed by interference factors such as electromagnetic interference, flickering pixels, and cloud-edge interference, as shown in Fig. 5. These interference factors have similar grayscale features to dim and small targets in infrared images, making them prone to false detections during the detection process. A multi-category label method is adopted to establish labels to reduce the impact of interference factors. This method treats all interference factors as an additional target category and learns them together with targets in the network. By increasing the label of the interference category, the network model can pay more attention to distinguishing subtle differences between targets and interference factors during the training process. The network model can identify correct targets more precisely and exclude interference factors, thereby further improving the accuracy of the detection.

Therefore, the loss in this article is divided into two parts: position loss ($\mathrm{Loss}_{\mathrm{loc}}$) and classification loss ($\mathrm{Loss}_{\mathrm{clc}}$), as shown in the Eq. (2). In Eq. (2), smoothL1 loss is used for the positional loss function and binary cross-entropy loss is chosen for the classification loss function. Then a and b respectively represent the corresponding weights for the two types of loss.

$$\mathrm{Loss} = a * \mathrm{Loss}_{\mathrm{loc}} + b * \mathrm{Loss}_{\mathrm{class}} \tag{2}$$

The position information of the infrared target includes four parameters: x (target center horizontal coordinate), y (target center vertical coordinate), w (target width), and h (target height). Therefore, the position loss is shown in Eq. (3). t_i is the predicted value, and g_i is the ground-truth value.

$$\mathrm{Loss}_{\mathrm{loc}} = \sum\nolimits_{i\in\{x,y,w,h\}} \mathrm{smooth}_{\mathrm{L1}}\left(t_i, g_i\right) \tag{3}$$

The classification loss calculation is as follows Eq. (4), where $c1$ and $c2$ are two categories of targets, and conf is the confidence value. p_i is the predicted value, and p_i^* is the ground-truth value.

$$\mathrm{Loss}_{\mathrm{class}} = \sum\nolimits_{i\in\{c1,c2,\mathrm{conf}\}} \mathrm{BCEloss}(p_i, p_i^*) \tag{4}$$

However, due to the limitations of manual operation, it is hard to label all images accurately. As a result, it is inevitable that mislabeled images will appear in the dataset. These inaccurate labels are classified into four categories: unlabeled targets, incorrectly positioned labeled targets, incorrectly categorized labeled targets, and incorrectly labeled backgrounds. The comparisons of these four types of inaccurate labels before and after refinement are shown in Fig. 6.

The inaccurate label is a major factor affecting training. In the actual training process, the inaccurate labels make the gradient learning direction wrong, which affects the training effect of the network model. The network continually learns error information due to inaccurate labels, thereby affecting its object detection capability.

Refining the inaccurate labels is necessary to improve network performance. However, manually judging massive labels is both consuming manpower and difficult to ensure accuracy. The key to solving this problem is how to select the inaccurate labels from the massive labels.

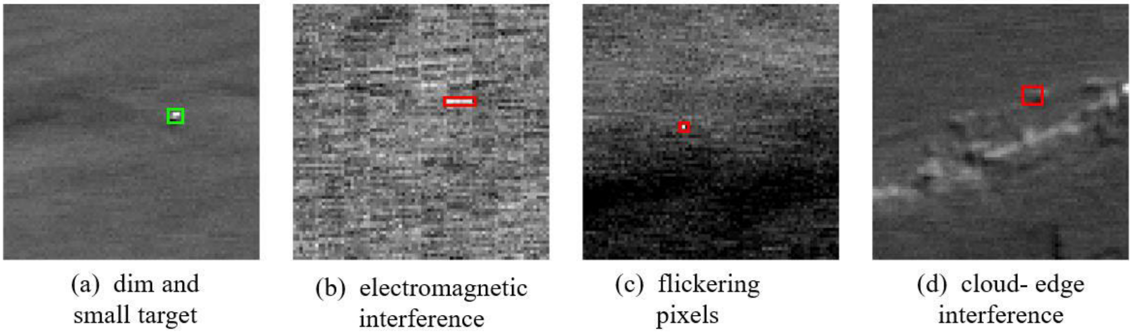
(a) dim and small target
(b) electromagnetic interference
(c) flickering pixels
(d) cloud- edge interference

FIGURE 5. Dim and small targets (a) and interfering factors(b), (c), (d).

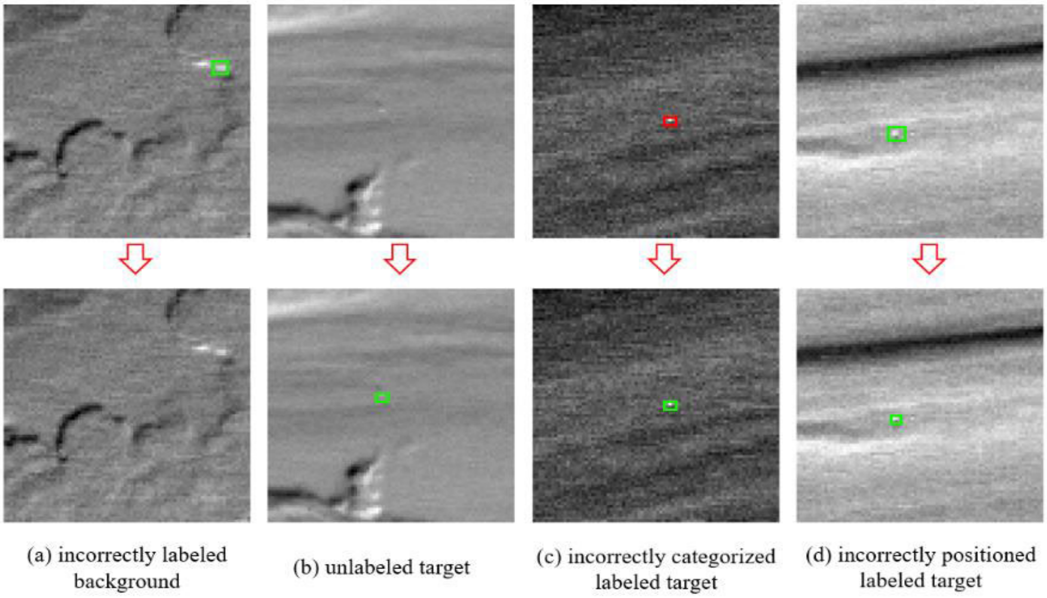
(a) incorrectly labeled background
(b) unlabeled target
(c) incorrectly categorized labeled target
(d) incorrectly positioned labeled target

FIGURE 6. Incorrect label and refined label.

With the deepening of network training and the enhancement of network target detection capability, the inaccurate label does not match the prediction result of the network, resulting in a high loss. Therefore, the loss is used as the judgment threshold, and labels that exceed the loss threshold are manually judged and refined. The modified dataset is retrained to improve the training effect.

C. ITERATIVE TRAINING BASED ON THE UPGRADE DATASET

In deep learning, diverse and rich datasets are crucial for the robustness of the network. However, it is always hard to collect a sufficient amount of data in the initial stage. At the same time, labeling all images requires too much manual operation due to the target-to-background proportion and massive original image sequence. So, it is difficult to establish a comprehensive and accurate dataset in the initial stage.

Infrared targets have motion characteristics in sequential images, and the target position information between consecutive frames is correlated. In consecutive frames, the movement direction of the same target should be relatively consistent, and the distance between positions should not be too large. By utilizing the motion characteristics of small targets, the detection results can be analyzed. In sequential images, targets that are detected multiple times and form trajectories are regarded as correct targets, whereas targets that appear only once and are not detected in the preceding and subsequent frames are regarded as interference targets. This method can be used to label the data and supplement the dataset when the network training is not sufficient. Fig. 7 illustrates the detection and miss detection of targets in the sequential images.

Supplementing all the new data into the dataset, not only lacks consideration for the diversity of the dataset but also increases training resources and time. For network learning, false positive and false negative data (false alarm and miss

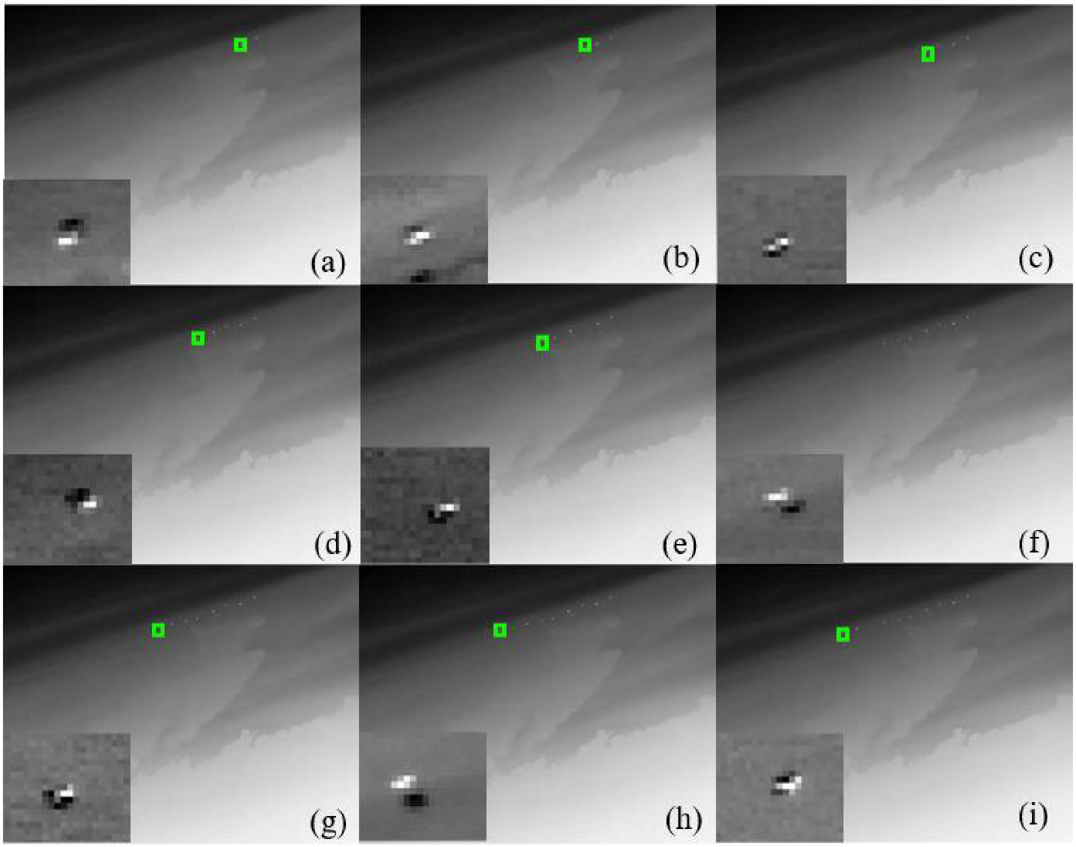

FIGURE 7. Targets (a -e, g – i) and missed target (f) in sequential images.

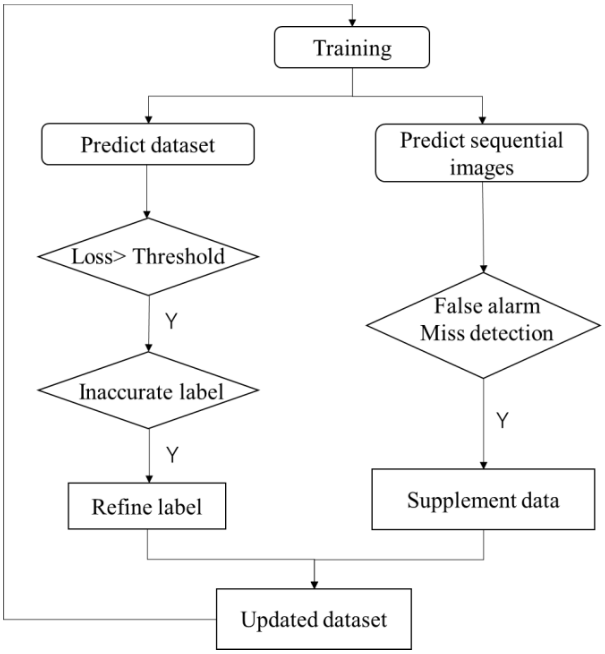

FIGURE 8. Flow chart of iterative training.

detection) can help to improve the training performance of the network. In infrared dim and small target detection, these data are more difficult for network learning, and cannot be effectively learned in the initial stage. Therefore, learning should be intensified for false positive and false negative data in the subsequent training process. New false positive and false negative data are constantly mined and added to the network training in the subsequent process, thus the network can focus more on learning these difficult points.

An initial network is obtained by training on the newly established dataset. The existing network is used to predict new original image sequences. When the detection result forms a trajectory in the preceding and following frames, but no target is detected in the current frame, it is considered as miss detection. If the target appears only once in the sequential images and is not detected in both the preceding and following frames, it is considered a false alarm. False alarm data and miss detection data are automatically identified through the above method, and then added to the dataset after manual double-check.

Iterative training is used to complete the above prediction, judgment, and data supplement operations. The trained network is used to predict sequential images, the false alarm and miss detection data are supplemented to the current dataset, and then the iterative training is continued. In each iteration, the dataset is first refined and supplemented, and

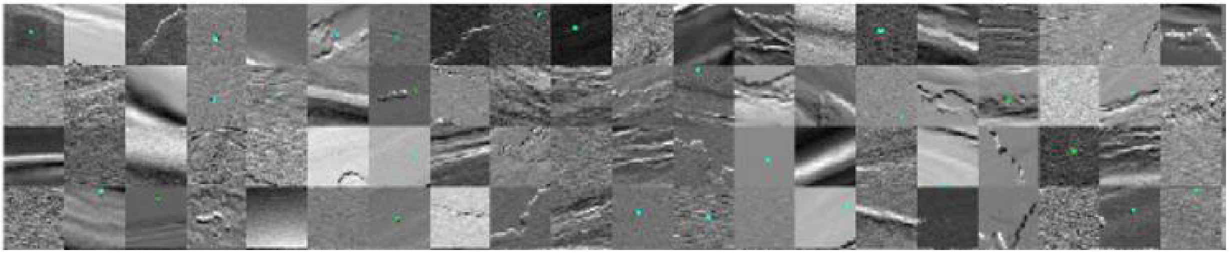

FIGURE 9. Examples of dataset.

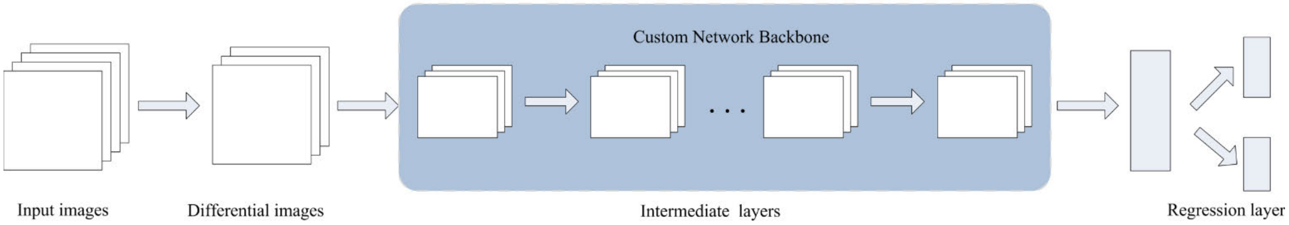

FIGURE 10. Network model.

the updated dataset is retrained. The flow chart is shown in Fig. 8.

IV. EXPERIMENT AND VERIFICATION

A. EXPERIMENTAL PREPARATION

The experimental dataset comes from infrared sequential images captured by the infrared focal plane detector under the sky background, with the original size being 640 × 512 × 4. The dataset is self-built dataset and consists of 89510 images, including 54489 pure background images and 35021 target images. Randomly allocate all datasets into two parts: training set and validation set, with 85010 training sets and 4500 validation sets. Fig. 9 shows an example of a partial dataset.

The CPU used is Intel(R) Xeon(R) Gold 6248R, and the GPU is GV100-32GB in this experiment. The initial learning rate for training is 1.0e-6. The main evaluation metrics used in this experiment are precision, recall rate, and F1 score.

Precision [20] measures the ratio of correctly predicted target number over all predicted target number, as shown in the Eq. (5). The TP is the number of correctly predicted target, and FP is the number of false predicted target.

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{5}$$

Recall [21] measures the ratio of correctly predicted target number over all real target number, as shown in the Eq. (6). The FN is the number of miss detection target.

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{6}$$

F1 score [22] is harmonic mean of Precision and recall, providing a better reflection of the model's performance as shown in the Eq. (7).

$$\text{F1} = \frac{2 \cdot \text{recall} \cdot \text{precision}}{\text{recall} + \text{precision}} \tag{7}$$

Different backbone networks are used to verify the usefulness of the above method. The input and output layers are unified for better comparison of backbone networks. A convolutional layer has been added to the input section to obtain differential images for extracting target moving information, and a unified regression layer has been used in the output section. The intermediate layers are the different backbone networks and the number of network layers is adjusted to obtain the same feature map size output, as shown in Fig. 10.

B. SMALL-SIZED IMAGE TRAINING

Due to hardware limitations, the network model size cannot be too large in engineering applications. Therefore, this article is trained on images of various sizes based on six widely used lightweight backbone networks (DarkNet-19 [23], ResNet-18 [24], GooLeNet [25], SuqeezeNet [26], MoblieNet [27], and EfficientNet [28]). The results are as follows in Table 2.

The experiment shows that as the size of the training image reduces, the resources required for network training reduce, the amount of operation required decreases, and the training time reduces. Compared to the size of 80 × 80, when training with the size of 64 × 64, there is a slight quiver in recall and precision, but the training time is significantly reduced. When using the size of 32 × 32, although the training time further decreases, the results show a moderate reduction of the network performance When the size of the training image

TABLE 2. Training results for images of different sizes.

Image Size	Network Backbone	Epoch	Training Time/min	Precision/%	Recall/%	F1/%
16×16×4	Darknet19	100	356	**41.81**	**79.03**	**54.69**
	Resnet-18	100	359	1.44	66.67	2.82
	Goolenet	100	360	32.89	36.45	34.58
	Suqeezenet	100	356	19.17	13.33	15.73
	Efficientnet	100	380	0.86	22.15	1.66
	Moblienet	100	361	1.81	67.74	3.53
32×32×4	Darknet19	100	398	**93.64**	**89.89**	**91.73**
	Resnet-18	100	391	69.58	93.23	79.69
	Goolenet	100	369	92.88	79.78	85.83
	Suqeezenet	100	400	75.52	82.80	78.99
	Efficientnet	100	440	34.54	76.99	47.69
	Moblienet	100	398	56.96	88.28	69.24
64×64×4	Darknet19	100	517	**95.11**	**94.73**	**94.92**
	Resnet-18	100	467	94.37	92.26	93.30
	Goolenet	100	505	95.22	92.37	93.77
	Suqeezenet	100	460	89.67	82.58	85.98
	Efficientnet	100	722	91.71	87.10	89.35
	Moblienet	100	532	92.16	87.74	89.90
80×80×4	Darknet19	100	546	**96.50**	**94.19**	**95.33**
	Resnet-18	100	516	94.66	91.94	93.28
	Goolenet	100	544	94.09	85.05	89.34
	Suqeezenet	100	506	88.85	75.59	81.69
	Efficientnet	100	963	92.60	85.59	88.96
	Moblienet	100	590	93.44	86.67	89.93

TABLE 3. Training results for images of iterative algorithm.

Iterations	Epoch	Number of images	Precision/%	Recall/%	F1/%
1	100	74215	95.40	54.47	69.35
2	50	77307	95.26	76.00	84.55
3	50	80528	96.60	86.45	91.24
4	50	83884	97.18	89.57	93.22
5	50	86879	96.37	92.23	94.25
6	50	89510	96.90	94.94	95.91

is reduced to 16 × 16, the performance of most networks is very poor.

This is because images of different sizes provide different background information. For infrared dim and small targets, images that are too small in size cannot fully express the complete background information. For example, images with size 16 × 16 are difficult to fully represent the details of cloud edges, resulting in poor learning for background and a large number of false alarms and miss detections. In addition, the detection ability of networks for edge positions is usually poor. The smaller the image size, the larger the proportion of edge positions, which also has an impact on the training effect.

When the size gradually increases, the image contains more background information. However, when the information contained in the image is sufficient to provide the complete background, it has a little effect on the improvement of network performance by continuing to increase the background information. Full-sized images provide redundant background information, thus most of the time is spent on

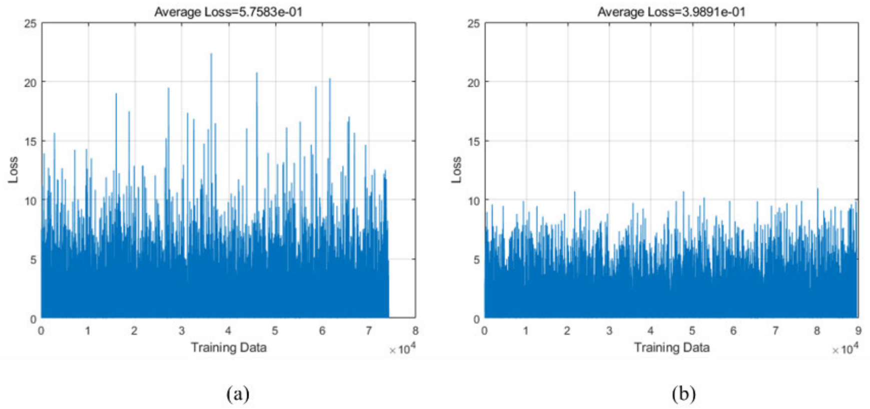

FIGURE 11. Distribution of loss first iteration (a) and last iteration (b).

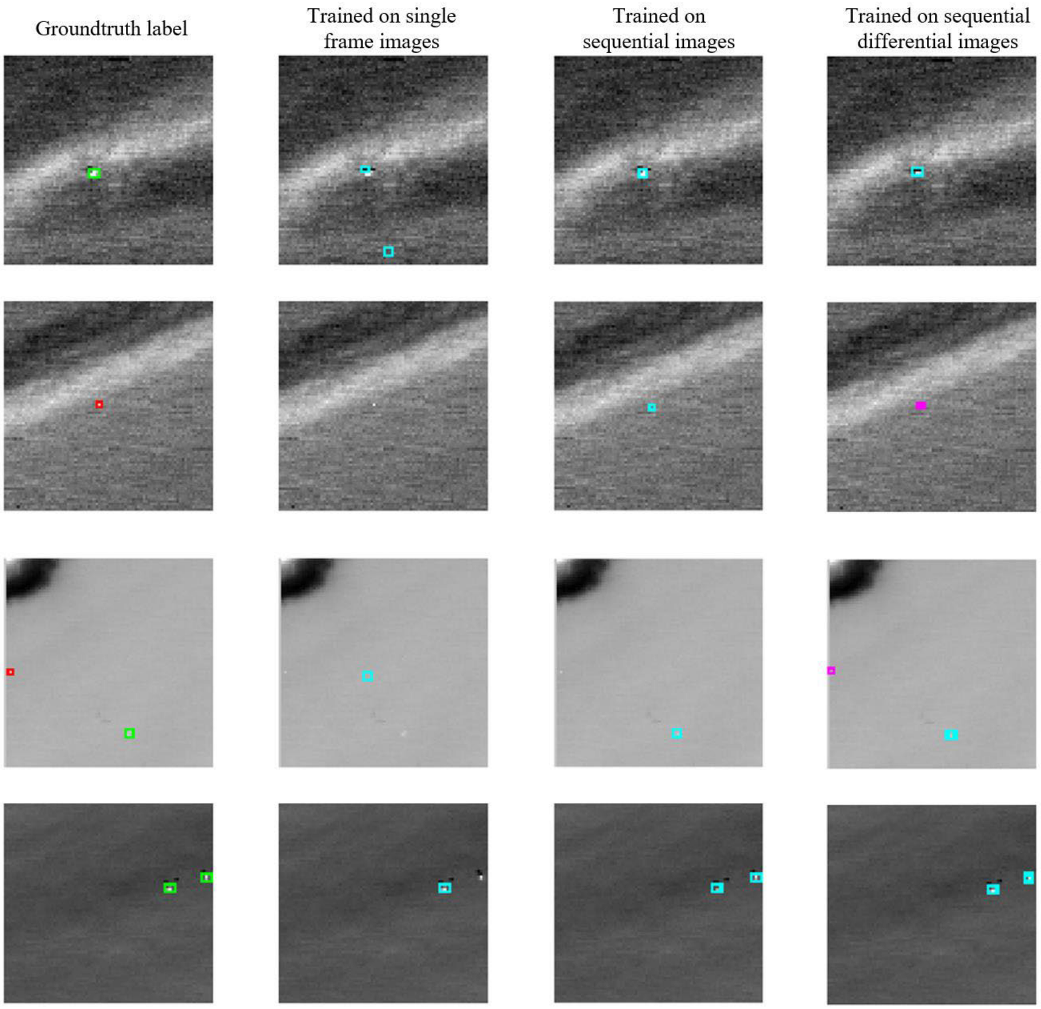
Groundtruth label
Trained on single frame images
Trained on sequential images
Trained on sequential differential images

FIGURE 12. Comparison of training results with different inputs.

calculating the background during network training, which ultimately leads to an increase in training time and a decrease in training efficiency.

Overall, the training effect of infrared dim and small target networks is influenced by the size of the images. The excessive image size with redundant background affects the learning of target features, resulting in long training time and poor training efficiency; The too small image size with incomplete background results in poor performance. Therefore, selecting appropriate small-sized images for training can effectively reduce training time and improve training efficiency. From the comparison of various networks, DarkNet19 has the best detection performance, with the highest precision rate of 96.50% and the highest recall rate of 94.19%.

C. ITERATIVE TRAINING

To validate the iterative training method, this experiment selected images with a size of 80 × 80 and the backbone network of DarkNet19 for iterative training to validate the iterative training method. The dataset is updated between each iteration. Between the iterations, we manually judge whether the label with high loss is inaccurate and refine the inaccurate label. Furthermore, new sequential images are predicted, and the image is added to the dataset when there is miss detection or false alarm data in the image. In the iterative training, the dataset is constantly revised and supplemented. The results of iterative training are shown in Table 3. The distribution of loss first iteration and last iteration is shown in the following Fig. 11.

As shown in Table 3, with the increasing number of iterations, the precision and recall of the network continue to improve, with the highest precision rate of 96.90% and the highest recall rate of 94.94%. From Fig. 11, it can be seen that label refinement based on loss evaluation metrics reduces the overall loss of the dataset, the average loss decreased from 0.5758 to 0.3989, decreased by 30.94%. The method avoids the sustained impact of inaccurate labels with high loss evaluation on training.

The results have demonstrated that network performance and training efficiency can be effectively improved by using this strategy to continuously supplement new false alarm and miss detection data, refine inaccurate labels, and carry out iterative training.

D. ITERATIVE TRAINING

To verify the impact of sequential differential images on training, we conducted comparative experiments using Darknet19 in the final dataset. We trained on the original single frame images, original sequential images, and sequential differential images, respectively. Fig. 12 shows the comparison of their prediction results.

From the experimental results, it can be seen that sequence differential images can effectively improve network performance compared to the original images.

V. CONCLUSION

Based on analyzing the impact of datasets on infrared dim and small target detection, this article proposes a strategy for training dim and small infrared target detection networks under sequential cloud background images. This strategy uses small-sized image datasets to reduce computing resources and training time. Through multiple iterations of training, it continuously refines inaccurate labels, updates and supplements the dataset, improves the quality of the dataset, and enhances the effectiveness of network training. The experiments are carried out and the results show that the above method can effectively shorten training time, improve training efficiency and performance of infrared dim and small target detection networks.

At present, the establishment of large-scale and high-quality datasets is still an important basic task of dim and small target detection. And the imaging results of the target are inevitably affected by the point diffusion function of the atmosphere and optical system, which seriously affects the accuracy of the annotation. Our method adopts semi-supervised learning method to improve the training efficiency and performance, but still requires manual operation and high cost to establish the dataset. In the future, innovation method such as auto evaluation metrics based on abnormal loss monitoring may help to enable unsupervised learning, reduce manual operations and costs.

REFERENCES

[1] Q. Liu, X. Lu, Z. He, C. Zhang, and W.-S. Chen, "Deep convolutional neural networks for thermal infrared object tracking," *Knowl.-Based Syst.*, vol. 134, pp. 189–198, Oct. 2017.

[2] M. Zhao, L. Cheng, X. Yang, P. Feng, L. Liu, and N. Wu, "TBC-Net: A real-time detector for infrared small target detection using semantic constraint," 2019, *arXiv:2001.05852*.

[3] W. Cai, P. Xu, Z. Yang, X. Jiang, and B. Jiang, "Dim-small targets detection of infrared images in complex background," *J. Appl. Opt.*, vol. 42, no. 4, pp. 643–650, Jul. 2021.

[4] J. Redmon and F. Ali, "YOLOv3: An incremental improvement," 2018, *arXiv: 804.02767*.

[5] L. Zhang, W. Lin, Z. Shen, D. Zhang, B. Xu, K. Wang, and J. Chen, "CA-U2-Net: Contour detection and attention in U2-Net for infrared dim and small target detection," *IEEE Access*, vol. 11, pp. 88245–88257, 2023.

[6] C. Bao, J. Cao, Y. Ning, T. Zhao, Z. Li, Z. Wang, L. Zhang, and Q. Hao, "Improved dense nested attention network based on transformer for infrared small target detection," in *Proc. Comput. Vis. Pattern Recognit.*, Nov. 2023, pp. 1745–1758.

[7] J.-H. Li, P. Zhang, X.-W. Wang, and S. Z. Huang, "Infrared small-target detection algorithms: A survey," *J. Image Graph.*, vol. 25, no. 9, pp. 1739–1753, Jan. 2020.

[8] Y. Liu, H.-J. Sun, and Y.-X. Zhao, "Infrared dim-small target detection under complex background based on attention mechanism," *Chin. J. Liquid Crystals Displays*, vol. 38, pp. 1455–1467, 2023.

[9] Y. Dai, X. Li, F. Zhou, Y. Qian, Y. Chen, and J. Yang, "One-stage cascade refinement networks for infrared small target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Feb. 2023, Art. no. 5000917.

[10] O. Gupta and R. Raskar, "Distributed learning of deep neural network over multiple agents," *J. Netw. Comput. Appl.*, vol. 116, pp. 1–8, Aug. 2018.

[11] Z.-F. Si and H.-G. Qi, "Survey on knowledge distillation and its application," *J. Image Graph.*, vol. 28, no. 9, pp. 2817–2832, Sep. 2023.

[12] S. K. Sun, J. Fan, Z. Sun, J. H. Qu, and T. T. Dai, "Survey of image data augmentation techniques for deep learning," *Comput. Sci.*, vol. 51, no. 1, pp. 1–23, Dec. 2023.

[13] W. Cao, M. Liu, M. Lu, X. Shao, Q. Liu, and J. Wang, "Influence of hyperparameters on performance of optical neural network training algorithms," *Laser Optoelelectron. Prog.*, vol. 60, no. 22, pp. 300–307, May 2023.

[14] T. DeVries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," 2017, *arXiv:1708.04552*.

[15] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "Mixup: Beyond empirical risk minimization," 2017, *arXiv:1710.09412*.

[16] P. Chen, S. Liu, H. Zhao, X. Wang, and J. Jia, "GridMask data augmentation," 2020, *arXiv:2001.04086*.

[17] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.

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

[19] Y. Liu, L. Geng, W. Zhang, Y. Gong, and Z. Xu, "Survey of video based small target detection," *J. Image Graph.*, vol. 9, no. 4, pp. 122–134, 2021.

[20] S. Du, K. Wang, and Z. Cao, "BPR-Net: Balancing precision and recall for infrared small target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Jul. 2023, Art. no. 5003515.

[21] J. Ma, H. Guo, S. Rong, J. Feng, and B. He, "Infrared dim and small target detection based on background prediction," *Remote Sens.*, vol. 15, no. 15, p. 3749, Jul. 2023.

[22] J. Lin, S. Li, L. Zhang, X. Yang, B. Yan, and Z. Meng, "IR-TransDet: Infrared dim and small target detection with IR-transformer," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Oct. 2023, Art. no. 5004813.

[23] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6517–6525.

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

[25] 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. (CVPR)*, Jun. 2015, pp. 1–9.

[26] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50× fewer parameters and < 0.5 MB model size," 2016, *arixv:1602.07360*.

[27] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.

[28] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.

WENXIN ZHAO received the bachelor's degree from Nanjing University of Science and Technology (NUST), in 2017. She is currently pursuing the master's degree with the Institute of Optics and Electronics (IOE), Chinese Academy of Sciences (CAS), China. Her research interests include image processing, deep learning, and target detection.

XUEFENG LAI received the Ph.D. degree from Shanghai Institute of Technical Physics of the CAS. He is currently an Associate Research Fellow and a Supervisor with the Institute of Optics and Electronics of the CAS. His research interests include infrared imaging, infrared photoelectric signal processing, and deep learning.

XULONG ZHAO is currently an Assistant Research Fellow with the Institute of Optics and Electronics (IOE), Chinese Academy of Sciences (CAS), China. His research interests include optical system optimization, infrared detection, and infrared radiation measurement.

YUCHENG XIA is currently an Assistant Research Fellow with the Institute of Optics and Electronics (IOE), Chinese Academy of Sciences (CAS), China. His research interests include image processing, deep learning, and infrared detection.

JINMEI ZHOU is currently a Research Fellow with the Institute of Optics and Electronics (IOE), Chinese Academy of Sciences (CAS), China. Her research interests include infrared radiation measurement, infrared imaging detection, and infrared system design.