Real-Time Infrared Small Target Detection With Nonlocal Spatial-Temporal Feature Fusion

Hai Xu, Sheng Zhong, Tianxu Zhang, and Xu Zou, *Member, IEEE*

***Abstract*—Infrared small target detection is a challenging task in which many researchers have made lots of achievements. While the performance of single-frame (SF) detection is still limited due to the lack of usage of multiframe (MF) continuous information, many spatial-temporal detection methods have been developed. However, most algorithms need to register the image or feed a set of images as the input. Inputting a batch of group images usually leads to large computations, which heavily affects their real-time processing capability in resource-limited machines. To tackle the problem, we propose a nonlocal multiframe network (NLMF-Net) with only a few additional computations (no more than 0.01 GFLOPs) compared to the SF baseline while achieving significant performance improvements. The NLMF-Net correlates features from grid cells with high confidence between current and past frames. While most background grid cells are removed after the SF processing, the MF feature fusion only focuses on a few potential target grid cells, resulting in high computation efficiency. The proposed vector length similarity module enlarges the difference between different grid cells and the non max similarity suppression further suppresses the backgrounds during the fusion, promoting the MF performance. The NLMF-Net can be readily deployed on Jetson Nano at a speed of 20 FPS for 288 × 384 image processing or Mi Pad 2 with a speed over 35 FPS for 128 × 128 part image processing. Extensive experiments show that our proposed method achieves state-of-the-art performance on three datasets while maintaining high efficiency in a real-time processing manner.**

***Index Terms*—Convolution neural network (CNN), infrared small target (IST) detection, real-time processing, spatial-temporal (ST) fusion.**

I. Introduction

INFRARED small target (IST) detection is an attractive task since it is widely applied in many areas, such as airport bird observation, unauthorized drone flight surveillance in urban areas, and remote sensing of spacecraft reentry. It is important for some scientific activities and urban safety. Over the past few decades, many methods have been proposed to tackle the issue. Due to the dim and shapeless characteristics of the ISTs, picking out ISTs from various complex backgrounds is a continuing challenge [1], [2].

For decades, many researchers have published remarkable works in IST detection [3], [4]. While single-frame (SF) detection may cause many false alarms in low signal-to-clutter ratio (SCR) or complex background images, multiframe (MF) methods [5], [6], [7], [8] present a more stable and robust performance by further leveraging spatial-temporal (ST) information. These traditional algorithms are designed with mathematical models and corresponding hyperparameters based on the prior assumptions. That is, once hyperparameters are selected, they can only work well on samples that satisfy the adopted prior assumptions. If they are implemented in some other conditions or evaluated based on datasets consisting of a large number of images in various spots, these classical methods may fail. And these ST tensors are typically sliced from multidimensional data matrices with large sizes, which significantly increases the computations for each processing.

Recently, with the development of deep learning methods applied in general object detection [9], [10] and segmentation [11], many convolution neural network (CNN)-based IST detection methods [1], [12], [13], [14] have been proposed. Some of them dominate the community of IST detection due to their strong learning ability and parameter self-adjustment during the training procedure. Most MF networks for IST detection [15], [16] also try to dig out more sequential features to achieve a more stable and better performance. They try to fuse ST features for IST detection and some remarkable performance has been achieved. However, a set of problems for practical applications still exist, which are given as follows.

1) While SF methods may be unable to distinguish some unstable background clusters, a group of registered or relatively static sequence images should be fed into the MF algorithms for both traditional and CNN-based methods. Direct feature extraction and fusion on batch images make processing time consuming.
2) Static images are too rigorous for practical cases. Thus, the registration process by scale-invariant features [17] or some other methods is inevitable and costs some additional time.
3) To balance the tradeoff between effectiveness and efficiency, for existing methods, only a few frames are utilized for each detection. That is, limited ST information is adopted for fusion and these complex models cannot be implemented on some source-limited machines.

To tackle the problem, we propose a nonlocal multiframe network (NLMF-Net) for real-time IST detection in sequence with ST information fusion in the feature field. The main

Manuscript received 25 January 2024; revised 2 March 2024; accepted 23 March 2024. Date of publication 27 March 2024; date of current version 12 April 2024. This work was supported by the National Natural Science Foundation of China (NSFC) under Grant 62176100. The computation is completed in the HPC Platform of Huazhong University of Science and Technology *(Corresponding author: Xu Zou.)*

The authors are with the National Key Laboratory of Multispectral Information Intelligent Processing Technology/ School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430074, China (e-mail: seahifly@hust.edu.cn; zoux@hust.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2024.3382389

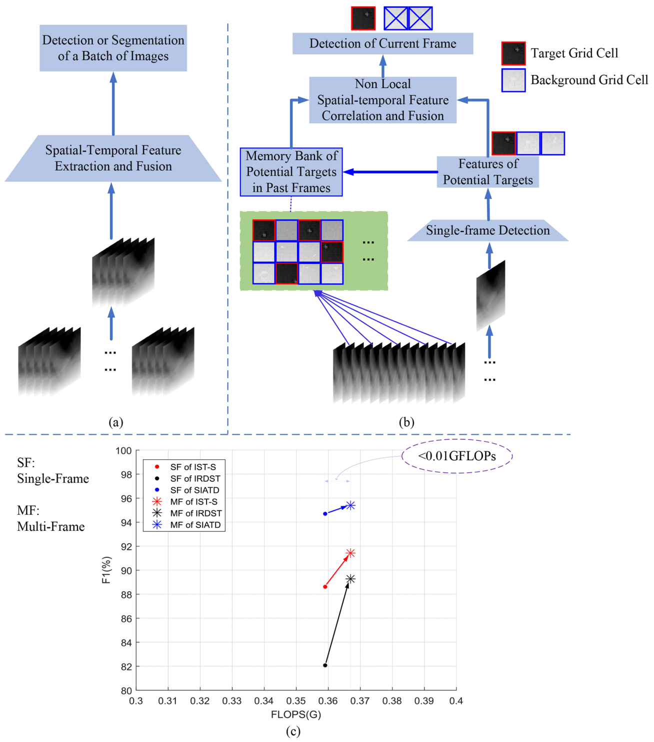


Fig. 1. (a) and (b) Advantage of our proposed NLMF-Net for sequential IST detection with high computational efficiency compared to current ones. For current methods, a batch of group images (usually should be registered or relatively static) would be fed into the model. However, such a large input data inevitably harms real-time properties. In contrast, our model processes the images one by one in real time. The ST fusion is applied in the feature field and focuses on high-confidence grid cells with potential targets. While the SF detection may fail to distinguish some backgrounds well, the MF fusion between the current and past frames in the memory bank can further suppress the target-similar background clusters. The movement of targets or camera platforms would not cause noticeable effects. (c) Our proposed NLMF-Net equipped with the MF strategy can consistently and effectively improve the performance of the SF baseline on three datasets while increasing no more than 0.01 GFLOPs computations. (a) Current Methods: Batch Image Input. (b) NLMF-Net: Feature Correlation and Fusion in Feature Field. (c) Promotion of NLMF-Net based on F1 evaluation of three datasets.

difference between our model and other methods is presented in Fig. 1. Since segmentation-based methods [1], [14] are too time consuming and cannot be applied on CPU-only or some source-limited machines for real-time processing, we build the NLMF-Net on the object level in a regression manner [2], [9], [12]. Besides, ISTs are sparse on the spatial dimension. Only features of a few grid cells with potential targets will be considered during the ST fusion. Our model takes a single real-time image as input and fuses ST features in a long-range quickly. Moreover, thanks to the vector length similarity module (VLSM) and non max similarity suppression (NMSS), unstable target-similar clusters can be further suppressed, and the performance increases compared to SF detection.

Contributions of this work can be summarized as follows.

1) Although there are some public datasets for IST detection studies, most datasets are for SF detection research or images in the datasets are synthetic or synthetic. Potential gaps between real and synthetic images may hurt the performance in real-world applications [18]. To alleviate the data dilemma and verify our research, we construct a training IST dataset and a test IST dataset, respectively. They are not overlapped.
2) We propose a lightweight nonlocal (NL) ST feature fusion framework, NLMF-Net. Input image registration is unessential since the feature correlation and fusion are only applied on high-confidence grid cells or image patches picked out from the SF detection. The framework promotes the performance on three datasets at the expense of no more than extra 0.01 GFlops.
3) To avoid inaccurate or insufficient correlation between the current and past frames, we propose the VLSM and NMSS to preserve correct correlation and fusion during the NL feature fusion process.
4) Experimental results based on our own and other public datasets show that the proposed framework improves the performance of SF detection obviously and outperforms other methods with a much faster speed. Our model can be readily deployed on CPU-only or resource-limited machines (e.g., Nvidia Jetson Nano and Mi Pad2).

II. Related Works

While the SF detection methods [4], [14], [19] fail to make use of the temporal information among contiguous images and cannot suppress some unstable clusters well, we mainly focus on multiframe detection for ISTs in this article and all the methods can be roughly divided into the traditional and CNN-based methods.

A. *Classic MF IST Detection*

Just like some SF methods [19], local contrast (LC) and NL self-correlation are also the most popular among MF methods for IST detection. At each processing, a group of images should be fed to them. As for LC methods, they usually split the task to solve the spatial maps and the temporal ones, respectively. Then, the last results can be obtained by fusing them just like ST local contrast filter (STLCF) [5] and ST local difference measure (STLDM) [20] algorithms. When it comes to NL methods, the MF input tensors are transformed from the small batch of sequence images. Although many optimization works have been made to split the targets well from the backgrounds for MFs, the low-rank and matrix recovery is usually implemented directly on these MF tensors. The improved multimode weighted tensor nuclear norm joint local weighted entropy contrast (IMNNL-WEC) [6] and the sparse regularization-based spatial-temporal twist tensor (SRSTT) [21] are the recent typical works that perform better than some SF methods [22]. However, it takes lots of time for the methods above and some others [23] to process the matrix recovery especially when a group of images are fed. Besides, traditional methods are based on prior assumptions and can only work well on some prior-satisfied conditions. When they are applied to evaluate a dataset with a large number of real images in various scenes, they cannot perform as well as some SF methods for both aforementioned and some others [24], [25] sometimes.

B. Lightweight Frameworks for Detection

Due to the limitations of classic methods, many researchers began to solve the problem based on CNNs or transformers. The authors in [1], [2], [13], [14], [26], and [27] performed much better than classic methods even on a dataset that is not overlapped with the training subset [2]. Although the deep-learning-based methods showed powerful abilities in object detection, it was a certain waste of computing resources for smaller target detection by simply adopting general frameworks. Excessively complex models will be detrimental to deployments on terminals, such as applications in small drones, cars, robots, or other similar scenes [28], [29]. For lightweight model building, besides efficient structures, such as depthwise separable or group convolution in some mobile networks [30], [31], it was common to clip the unnecessary branches [12], [32], [33] or reuse some backbones or modules [34] to make the models efficient. Especially for smaller objects or ISTs, many researchers also found the importance of shallow features [2], [35], [36]. For example, Sun et al. [35] designed a simple multireceptive field extraction (MRFE) module to further extract features with different receptive convolutions. They thought the shallow layers were suitable enough and the MRFE was light with only 12 layers. Different from MRFE, some other researchers [2], [36] extracted features from low to deep layers with a few convolutions. Widely adopted residual direct skip and dimension concatenation made the convolution combinations abundant, leading to features with adequate receptive fields and properties while the computations were controlled to small amounts. Although some SF networks perform well in real time on GPU platforms, most of them usually focus on smaller general object detection but not extremely small ISTs. The information of only an SF is still limited, which means more progress can be made if MF cues can be utilized.

C. CNN-Based MF IST Detection

As for MF IST detection, most researchers designed specific modules to further utilize temporal and spatial information. Lin et al. [37] proposed a video IST detection network. The 3-D cross-scale and fusion module connects features from different scales during the encoding and decoding. Such contextual information interaction leads to a better performance than some traditional methods. Yan et al. [15] connected a temporal multiscale feature extractor and a spatial multiscale feature refiner in series. Their spatio-temporal differential multiscale attention network (STDMANet) is trained with a mask-weighted heatmap loss and works better than some SF CNN-based methods [1], [26]. Different from the aforementioned networks that extract features directly on a group of sequence images, Du et al. [38] enhanced the targets and suppressed the strong spatially nonstationary clutter by an interframe energy accumulation enhancement mechanism. Then, the feature extraction module extracts the ST information and the last detection network predicts target positions among thousands of regions of interest boxes. Because these methods are designed to extract and enhance features based on a group input image, it is usually time consuming for unit processing just like MF traditional methods. Besides, the images should be registered sometimes, which also costs some additional time. To balance the performance and speed, they usually limit the frames to make sure the models can be implemented on powerful machines in real time.

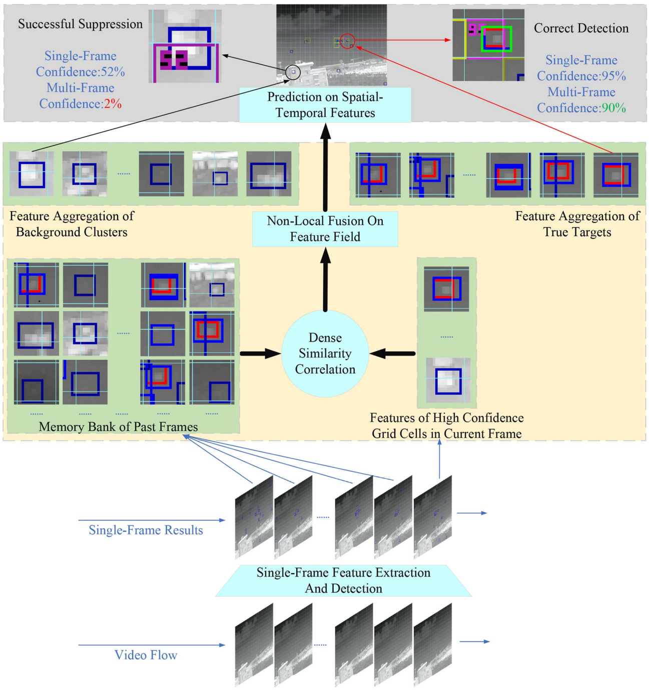

Successful Suppression
Single-Frame Confidence:52%
Multi-Frame Confidence:2%
Prediction on Spatial-Temporal Features
Correct Detection
Single-Frame Confidence:95%
Multi-Frame Confidence:90%
Feature Aggregation of Background Clusters
Non-Local Fusion On Feature Field
Feature Aggregation of True Targets
Dense Similarity Correlation
Memory Bank of Past Frames
Features of High Confidence Grid Cells in Current Frame
Single-Frame Results
Single-Frame Feature Extraction And Detection
Video Flow

Fig. 2. Framework of the MF IST detection based on NL ST feature fusion. All patterns with light blue backgrounds denote modules in NLMF-Net, while the others are data flow. Our NLMF-Net can further suppress the clusters that are not continuous and maintain the confidence of true targets well by fusing the ST information in the feature field. All the small image patches denote features of the picked out grid cells in each SF. In the images, all the red boxes, blue boxes, and green boxes denote the true target, SF results, and MF results, respectively. The light blue lines are the boundaries of the grid cells.

III. Methods

At present, most CNN-based MF IST detection methods focus on how to extract and fuse features from the batch input image. They are effective methods, but too time consuming and not competent for operation in real time under some resource constraints. We choose to capture the long range dependencies by spacetime NL operation [39] among MFs. It is a transformer style module and its various optimized versions can be applied in classification [40], superresolution [41], segmentation [42], [43], and so on. We can build space-time memory networks by using the query, key, and value concept. The keys or features of the current query frame associate and correlate with the ones of past frames in the memory bank. Values of space-time fusion can be read out by the matrix multiplication. Extracted features from the same backbone can be associated with each other. Features in the past frames are stored and fused with the current frame instead of extracting them with limited batch frames like [38].

The ST fusion of the NLMF-Net is based on the features of high-confidence grid cells. As presented in Fig. 2 from the bottom to the top, the SF network processes the current frame at first. Since the SF detection results can be obtained, features of high-confidence grid cells in the current frame are picked out. As the process goes on, more features in a few past frames are picked out and we can make a memory bank of past frames. Then, high-confidence grid cells of the current frame will be correlated with the ones in the memory bank. ST features can be aggregated and fused after the weights of features in the memory bank are obtained just as the feature aggregation examples presented at the top of the light yellow box. Finally, the MF prediction will be conducted on the ST features. High-confidence grid cells in the current frame will be stored in the memory bank for the following process.

Because the ST features are correlated, aggregated, and fused based on the features of grid cells or small image patches, our method does not need to register sequence images and fuse features on a local part. The movement of camera platforms makes few influences once the potential grid cells are picked out during the SF processing.

A. *SF Feature Extraction and Detection*

SF detection is also competent in most cases. The most difficult problem for SF detection is that it is nearly impossible to distinguish the target-similar clusters by only one still image. If the "false positive" metric is neglected, the SF network has the potential ability to pick most targets out. So, we adopt the multiscale and multilevel residual feature fusion network (MMRFF-Net) [2] for the SF detection and simplify it to be quicker. First, the inference speed is not only influenced by the computations of the model. More processing steps usually mean more time spent in some other computer operation such as data access from some memory hardware. Besides, ISTs are small in size. By drawing on the experience of infrared image observers, we only need to consider the surrounding background area to distinguish ISTs. Too large receptive fields may not be necessary, either. Therefore, we remove a convolution layer of the block in [2]. Both the global features and local features decrease by 1/4. Second, we replace the grid resample operations (GROs) with max-pooling layers and a single 2 × 2 convolution layer. One key point in [2] is that making full use of all scale features is beneficial for performance. We choose to continue to adhere to this rule and reduce some processing steps. We adjust all features to the 1/8 scale by the max-pooling layers directly and concatenate them on channel dimension. Then, the concatenated features are adjusted to 1/16 scale by a 2 × 2 convolution layer with the step set as 2, which can be fed to the decoupled head for the last prediction in SF processing.

Because layers in each block decrease and no GROs are used to further adjust features to a larger amount, the channels of the last fused features decrease. Some previous experiments [2] demonstrate that enough features are necessary for accurate predictions. So, we increase the channels of the third block from 4 to 8 to preserve the amount of features. The default channels of all blocks in our SF network are 2, 4, 8, and 8. Then, the channels of output features are 6, 12, 18, and 18. The total channels of the features before the 2 × 2 layers are 66. We set the output channel of the last 2 × 2 convolution layer as 192 before the SF prediction head. Finally, the decoupled head [2], [9] predicts targets at 1/16 scale by features with only 192 channels in dimension for each grid cell.

B. *Feature Correlation*

ISTs are sparse in images. There are only several targets in an image, which means only a few grid cells in which there are some potential targets. If features of all the grid cells are considered in the MF processing, the computations will be too large and redundant. After the SF feature extraction and detection, the confidence of all the grid cells can be obtained. We choose to pick out the top N (the default value is 10 in our experiments) grid cells with the highest confidence during the training and testing. The following feature correlation and aggregation in the ST fusion are conducted on these high-confidence grid cells. Compared to other ST fusion for IST detection, our model can correlate and fuse features in sequence for a large range without introducing too many computations while others usually extract and utilize a few frames for the balance between performance and speed. Many background grid cells are neglected means the computations can be controlled to a very small amount. If the image resolution is $H \times W$, the grid cell size is G (16 in this article), the frame number in the memory bank is T, and the feature channel size is C, the computations can be decreased from $O(T(\frac{HW}{G^2})^2C)$ to $O(TN^2C)$. To be more specific, over 99.94% of computations can be removed if our model is applied on 288 × 384 image processing. Besides, it is an NL processing and there is no need to register the images. The image registration is too time consuming. It is economical for our model to make correlation and aggregation in the feature field, which ensures the real-time processing of the network. The overall process can be formulated as follows:

$$s_{i,j} = \frac{q_j \cdot k_i}{|q_j| \cdot |k_i|} \cdot \left(1 - \frac{|q_j - k_i|}{|q_j| + |k_i|}\right), s_{i,j} \in S_{TN \times N} \quad (1)$$

where $s_{i,j}$ denotes the element in the correlation matrix S. It describes the similarity value between the jth picked out grid cell of the current frame and the ith one in the memory bank. q_j and k_i represent the encoded features of corresponding grid cells. More specifically, $q_j = \theta(F_c)$ and $q_i = \theta(F_{mb})$. F_c and F_{mb} denote features of the current frame and the ones in the memory bank, respectively. $s_{i,j}$ will be close to one if the grid cells are similar or otherwise close to zero.

The feature correlation module is also presented in Fig. 3. After SF processing, all the grid cells with high confidence in the past frames will be stored in a memory bank. All the extracted features will be fed to the same 1 × 1 convolution layer to be encoded as queries of the current frame or the keys of past frames. We avoid adopting the conventional QKV concept directly and encoding queries and keys by the same convolution operation. In other words, the query can be stored in the memory bank as keys directly after completing the processing of the current frame. During the testing, keys can be obtained from the memory bank directly without additional computations, which also contributes to reducing some computations.

After getting queries and keys, the similarity between each query and each key should be obtained before the fusion. The

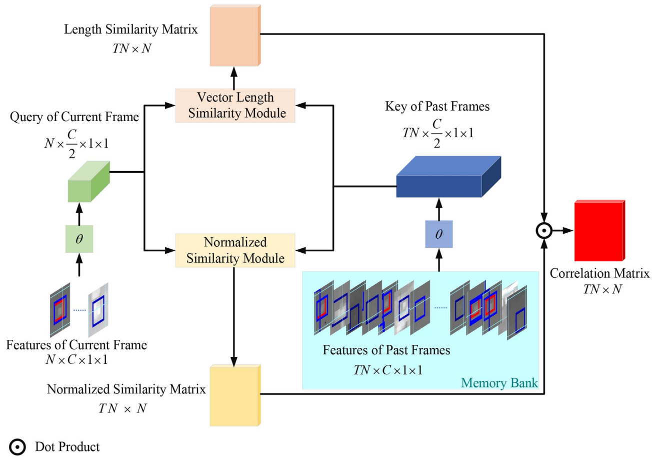

Length Similarity Matrix
$TN \times N$
Vector Length Similarity Module
Query of Current Frame
$N \times \frac{C}{2} \times 1 \times 1$
Key of Past Frames
$TN \times \frac{C}{2} \times 1 \times 1$
θ
θ
Normalized Similarity Module
Correlation Matrix
$TN \times N$
Features of Current Frame
$N \times C \times 1 \times 1$
Features of Past Frames
$TN \times C \times 1 \times 1$
Memory Bank
Normalized Similarity Matrix
$TN \times N$
Dot Product

Fig. 3. Structure of the feature correlation module. N denotes the constant number of the grid cells with the highest confidence in each frame during the training. T denotes the number of past frames in the memory bank. C denotes the channel size of the features. θ denotes the 1×1 convolution layer, which encodes all the grid cell features to the same implicit space.

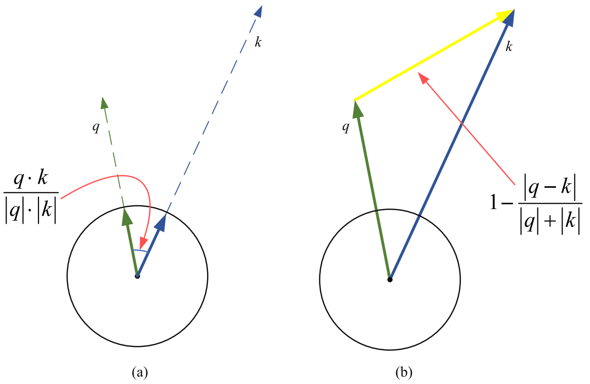

$\frac{q \cdot k}{|q| \cdot |k|}$
q
k
(a)
q
k
$1 - \frac{|q-k|}{|q|+|k|}$
(b)

Fig. 4. Definition of normalized similarity and the vector length similarity. q and k denote the encoded feature vectors of grid cells in the current frame and past frames, respectively. (a) Normalized Similarity. (b) Vector Length Similarity.

normalized similarity is the cosine function that has been widely applied in [2], [44], [45], [46]. Because we continue to adopt the self-contrast loss [2] and only the top N grid cells with the highest confidence are kept for ST correlation after SF detection, the normalized similarity values between these grid cells are usually high (over 95% or even more). Then, the fused ST features may consist of too many background components. On the contrary, if we choose to calculate the nonnormalized similarity, the difference will be too large. Only one grid cell in the memory bank will be in response during the following aggregation and fusion. To mitigate the problem, we propose a simple VLSM, which is presented in Fig. 4. It takes both the length and direction characteristics of the feature vectors into consideration and will not be influenced definitively by the length characteristics such as nonnormalized similarity. VLSM enlarges the difference among grid cells properly and ensures a more accurate aggregation during the ST feature fusion for those grid cells in which there are potential ISTs.

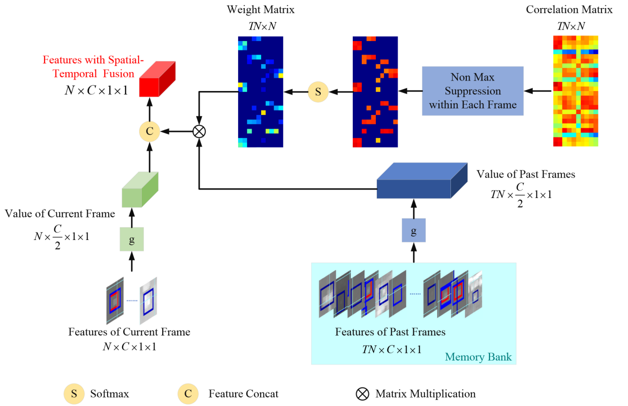

Weight Matrix
$TN \times N$
Correlation Matrix
$TN \times N$
Features with Spatial-Temporal Fusion
$N \times C \times 1 \times 1$
S
Non Max Suppression within Each Frame
C
Value of Past Frames
$TN \times \frac{C}{2} \times 1 \times 1$
Value of Current Frame
$N \times \frac{C}{2} \times 1 \times 1$
g
g
Features of Current Frame
$N \times C \times 1 \times 1$
Features of Past Frames
$TN \times C \times 1 \times 1$
Memory Bank
S Softmax
C Feature Concat
⊗ Matrix Multiplication

Fig. 5. ST feature fusion. g is the same 1×1 convolution that encodes features to values. During the fusion, the influence of the background clusters will be suppressed with NMSS within each frame. Only the features of continual targets with the highest similarities in the frames will be aggregated and fused, which promotes the performance.

C. ST Feature Fusion

In other works [39], [41], [42], the last aggregated features are directly obtained by matrix production. It is not practicable in our model for IST detection since the difference description between the target grid cells and the background ones is enlarged but not suppressed to zeros by VLSM. The aggregated features may consist of quite a few parts of backgrounds if we follow the conventional way [39], [41], [42]. To further suppress the backgrounds, we propose an NMSS within each frame. It is stable and consecutive for an IST to correlate one or two grid cells. So for, a target grid cell in the current frame, there will be no more than two correct grid cells to correlate and fuse in most cases. We suppress the lower ones to zero directly and only the highest and the second highest grid cells are kept left. The whole process is presented in Fig. 5, which can be formulated as follows:

$$F_{\mathrm{st}} = \mathrm{cat}\left(g\left(F_c\right), g\left(F_{\mathrm{mb}}\right) \otimes s\left(\mathrm{NMSS}\left(M_{TN \times N}\right)\right)\right) \tag{2}$$

where cat denotes the concatenation operation on channel dimension; g denotes the 1×1 convolution layer that converts features to the same value space; $\otimes$ denotes matrix production; NMSS denotes NMSS within each frame; F_{st}, F_c, F_{mb}, and M denote the ST features, features of the current frame, features in the memory bank, and correlation matrix, respectively; the size of the correlation matrix is $TN \times N$; T is the frame number of the images stored in the memory bank; N is the number of grid cells with the highest confidence in each SF; and s denotes softmax operation with temperature hyperparameter [42], [43] on column dimension as follows:

$$w_{i \to j} = \frac{e^{s_{i,j}/\gamma}}{\sum\limits_{i} e^{s_{i,j}/\gamma}} \tag{3}$$

where $s_{i,j}$ is obtained by (1) but not suppressed to zero by NMSS; $w_{i \to j}$ is the weight of the ith grid cell in the memory bank for the jth fused feature in the current frame; and γ is the temperature hyperparameter. The dimension of the fused ST features is the same as the one in SF detection. So, we adopt the same decoupled head for the MF prediction. Features with ST fusion will be fed to the MF head directly.

To present the computation influence more clearly, we compute and summarize computations of the correlation and fusion as follows:

$$\text{FLOPs}_{\text{ST}} = 2NC^2 + \frac{3}{2}NC + 4TN^2C + 13TN^2 \quad (4)$$

where both the feature correlation and fusion are considered; N, T, and C denote the number of high-confidence grid cells in the SF process, the number of frames stored in the memory bank, and the feature channel, respectively. Although more stored frames can increase the computations, the order of T is only 1 and only two terms contain it. The main computation contributions are from N and C. If the last prediction of the fused ST features is considered, the influence of N and C will be even enlarged. The additional MF computations will only rise up from 0.008 GFLOPs to 0.014 GFLOPs if we increase T from 20 to 100 based on our default settings. Relatively, if N or C are doubled, the additional computations will increase more since all terms contain them and some orders are two. Besides, the last prediction based on the fused ST features is also relevant to N and C. For building a lightweight MF model, it is important to constrain N and C while maintaining the performance. Default settings and corresponding analysis have been presented in previous subsections.

D. Loss Function

The SF network and MF fusion network are independent. They can be trained independently. In our experiments in Section V, the MF process contains the SF and the NLMF part. The overall loss function is as follows:

$$\mathcal{L} = \lambda_1 \cdot \mathcal{L}_{\text{SF}} + \lambda_2 \cdot \mathcal{L}_{\text{MF}} \quad (5)$$

where λ_1 and λ_2 denote the weight of SF and MF parts, respectively; $\mathcal{L}_{\text{SF}}$ is the SF detection loss, which is the same as [2]; and $\mathcal{L}_{\text{MF}}$ is the MF detection loss. It consists of only confidence loss and regression loss, which can be simply described as follows:

$$\begin{aligned}\mathcal{L}_{\text{MF}} = {} & \alpha_T \cdot \mathcal{M}\left(P_T \cdot \text{BCE}\left(C, \hat{C}\right)\right) \\ & + \alpha_B \cdot \mathcal{M}\left(P_B \cdot \text{BCE}\left(C, \hat{C}\right)\right) \\ & + \alpha_R \cdot \mathcal{M}\left(P_T \cdot \text{CIOU}\left(R, \hat{R}\right)\right) \quad (6)\end{aligned}$$

where BCE and CIOU denote binary cross entropy and complete-IOU [47] loss functions, respectively; and P_T and P_B represent whether there are targets in the grid cells or not. If there is a target in the picked out grid cell, P_T is 1 and P_B is 0. If there is no target, P_T is 0 and P_B is 1. C and $\hat{C}$ denote the confidence groundtruth and the prediction for grid cells that are picked out from the SF detection. R and $\hat{R}$ denote the box groundtruth and the prediction bounding boxes. $\mathcal{M}$ denotes the mean operation. α_T, α_B, and α_R denote the weight of each component in the loss function.

IV. Dataset and Evaluation

A. Dataset Description

In this article, we mainly focus on ISTs in the sky. Although synthetic images may also promote performance, we worry about the gap between real and synthetic images [18]. So, we only take real images for the first experiments (see Section V-B). As mentioned before, real images in sequence are lacking. We construct a training set and a test set from different cameras and cases. There are 249 sequences in our training dataset (161 358 images and 196 303 ISTs). The image sizes in eight sequences are 480 × 640, while others are 288 × 384. We name this dataset IST sequence training set, simplified as IST-ST. There are aircraft, planes, birds, and helicopter ITSs in IST-ST. The model trained on such a large number of images can meet the real-world application better. The test dataset consists of 26 sequences (16 404 images and 17 003 ISTs). All the images share the same size as 288 × 384 and they are captured by a static camera. In this test dataset, all the planes, birds, and aircraft are taken as targets. It is a static IST dataset and we name it IST-S. In our work, we broadly define targets smaller than 16 × 16 as ISTs. And over 99% of ISTs in our datasets are smaller than 6 × 6.

Besides, Sun et al. [35] shared a massive infrared small target dataset (IRDST) with 85 real image sequences and some synthetic images. To the best of our knowledge, it is the only open dataset in which there are many real sequences and ISTs satisfy our definition. Although IRDST is constructed by another team and images are captured by other cameras, it is still consistent with our datasets to some extent. It is more convincing to conduct experiments on nonoverlapped datasets to evaluate both the performance and robustness of all the models. So, we take the 85 real image sequences (40 656 images and 41 801 ISTs) as the second test dataset. In IRDST, a few images are 742 × 992, while all others are 480 × 720. More details can be referred to [35].

IST detection is a challenging task. Not only will the weak SCR of ISTs increase the difficulties, but the various shapes, sizes, and backgrounds will also do. Some large models may just overfit the training subset and perform well on the specific cases. So, in the first experiment comparison of this article, we train all the CNN methods on the same and only training dataset. All methods including classic methods are evaluated on IST-S and IRDST. All the datasets are not overlapped.

Besides, we conduct another experiment based on the open SIATD [48]. It is a semisynthetic dataset consisting of 175 sequences for both the training and test subsets. It has been widely adopted for comparison and analysis in [15], [26], and [27]. The backgrounds are captured by a camera with 512 × 640 resolution and all ISTs are synthetic. Targets in SIATD also satisfy the definition in our work. Indeed, all synthetic targets are smaller than 3 × 3. While potential gaps exist between real and synthetic samples, CNN models in this experiment are all retrained. In the official SIATD, both the training and test subsets contain 175 sequences. After checking all the images, we find that the 61st sequence in the test subset is broken. So we remove it and there are only 174 sequences in the evaluation

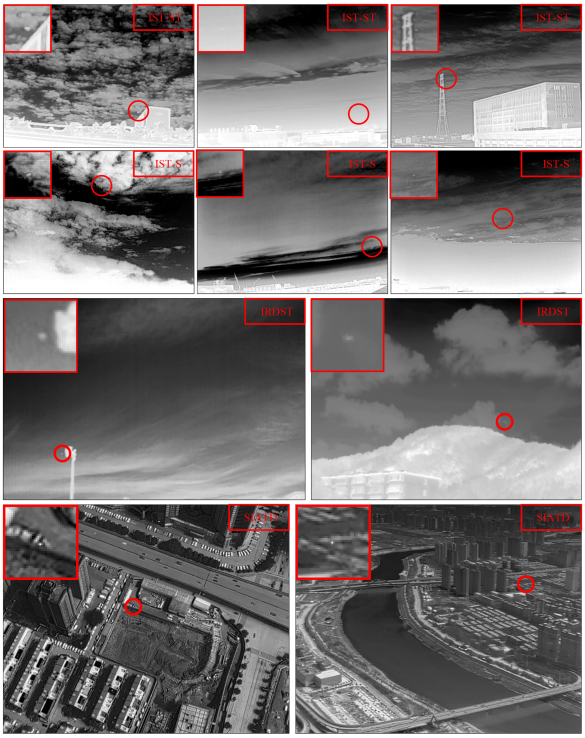


Fig. 6. Some examples from different datasets. Enlarged views of ISTs are presented at the left-top corner. There are various targets and backgrounds in the datasets for training and evaluation.

TABLE I
STATISTIC OF TARGET NUMBER FOR ALL DATASETS

IST num	0	1	2	≥3	frames	R/S
IST-ST	1530	132406	19805	7617	161358	R
IST-S	0	15805	599	0	16404	R
IRDST	0	39511	1145	0	40656	R
SIATD-train	612	39846	20352	14703	75513	S
SIATD-test	833	38957	19316	14333	73439	S

There are only one or two targets in most images. “R/S” denotes “real / Synthetic” images.

of the NLMF-Net. They contain rich scenarios such as rivers, buildings, vegetation, and cloud backgrounds. More details can be referred to [48].

Some examples of all the datasets are presented in Fig. 6. The statistics of the ISTs are presented in Table I and Fig. 7. Because there is only position information in the labels of SIATD, we set the target size in SIATD as 1 × 1 in the statistics. Validating performances of models on several large and diverse datasets is reliable and convincing.

B. Evaluation Methods

In this article, average precision (AP) [9] and receiver operating characteristic (ROC) [2] are adopted for the evaluation. Following the evaluation in [2], we take AP3p25 in [2] as the AP metric and name the corresponding ROC metric ROC3p25. When evaluating ROC, if a bounding box is not a correct detection, all the pixels in the box will be taken as false alarms for the regression-based methods.

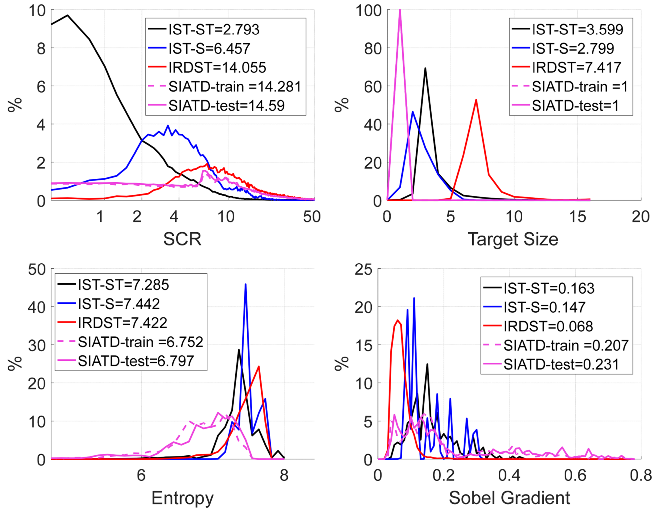


Fig. 7. Distribution of target size, SCR, image entropy, and local Sobel gradients for all datasets. The average values are presented in the left top corner or right top corner. All definitions can be referred to [2]. The training and test subset of SIATD are consistent in distributions, while other datasets vary with each other in the distributions. They are complex and vary in IST size and SCR distributions.

Besides, we relax AP3p25 to AP3pOr50. For AP3pOr50, a predicted bounding box will be taken as a correct detection when: the Euclidean distance between the center of groundtruth and the predicted bounding box is not larger than 3 and the center of groundtruth is in the predicted bounding box, or the intersection over union (IOU) between the predicted bounding box and groundtruth is not smaller than 50%. When evaluating methods based on AP3pOr50, predictions with lower IOU are not taken as false detection when ISTs are very small. If ISTs are larger, a successful detection only needs to satisfy the second rule. The corresponding ROC metric is named ROC3pOr50. Relatively, AP3p25 is a stricter metric demanding accuracy for both center distance and bounding box size prediction while AP3pOr50 may be more friendly for IST detection evaluation.

We also adopt the F1 score [15], [26], [27] to evaluate the balance performance of the precision and recall. Because AP evaluation can reveal the performance on the whole, we only present the best F1 scores for all methods in this article. All metric evaluation results will be presented in Section V.

V. Experiments

A. Network Training and Settings

During the training of NLMF-Net, we train the SF part for 10 000 iterations before the MF training. A well pretrained model can avoid incorrect feature correlation during the MF training. During the MF training, the images are sampled from 249 sequences randomly in IST-ST. To increase the sample combinations, images are selected from the original sequences randomly with a random interval from 1 to 3. The short training sample sequence varies the length from 2 to 15 randomly and the sequence order may be reversed before each iteration. We consistently crop the images to 240 × 320 randomly and make sure that ISTs are in the cropped images during the training.

TABLE II
EXPERIMENTAL RESULTS OF ALL METHODS

Category		Method	FPS			Params	FLOPs	IST-S						IRDST [35]					
								3p25 (%)			3pOr50 (%)			3p25 (%)			3pOr50 (%)		
			CPU	GPU	JN	(M)	(G)	AP	ROC	F1	AP	ROC	F1	AP	ROC	F1	AP	ROC	F1
SF	LC	ASPCM [3]	3.8	-	-	-	-	0.03	2.41	1.24	0.33	8.21	4.19	0.00	0.05	0.01	1.48	29.46	6.33
	NL	PSTNN [4]	4.0	-	-	-	-	4.78	35.43	17.45	6.66	39.40	21.90	0.72	10.12	5.57	3.40	23.97	12.63
	Seg	MDvsFA [13]	0.3	5.8	-	3.77	834	69.00	**86.27**	79.60	75.54	**92.93**	81.73	41.31	76.57	55.95	68.78	88.07	71.25
		ALCNet [14]	24	29	2.8	0.38	12.7	68.84	84.92	82.00	71.51	86.63	83.51	52.33	86.38	69.40	57.27	90.25	73.37
		DNANet [1]	1.7	19	1.3	4.70	48.1	45.93	60.93	65.60	50.37	66.25	68.52	34.44	77.37	54.27	35.52	78.82	55.18
	Reg	YOLOX-nano [9]	13.7	47	8.0	0.90	0.67	48.60	34.79	61.90	67.78	51.50	75.14	82.08	74.15	85.14	82.22	74.27	85.18
		YOLOX-x [9]	1.7	23	1.5	99.1	76.0	63.50	58.80	72.20	72.87	67.06	78.57	81.80	72.69	85.95	82.43	73.11	86.07
		MMRFF-Net [2]	40	89	27	0.44	0.48	81.24	80.55	80.97	87.15	85.13	84.38	81.52	82.92	79.59	85.86	81.16	81.89
		ours-SF	**56**	**180**	**30**	**0.34**	**0.36**	83.55	83.79	85.10	90.36	88.77	88.62	84.31	85.38	81.20	86.05	86.73	82.10
MF	LC	STLCF [5]	1.9	-	-	-	-	2.52	24.21	10.74	4.83	33.02	16.66	0.23	9.07	2.70	4.08	32.24	10.49
		STLDM [20]	0.3	-	-	-	-	5.42	22.51	16.54	6.17	23.78	17.80	0.72	9.85	5.35	1.30	15.79	7.87
	NL	IMNNLWEC [6]	0.1	-	-	-	-	27.73	62.32	47.93	32.05	67.70	51.92	0.35	8.39	3.19	3.78	34.51	12.30
		SRSTT [21]	0.04	-	-	-	-	8.76	33.19	29.15	16.07	44.83	39.37	1.63	17.58	9.06	3.11	24.66	12.66
	Reg	ours-MF	47	100	20	0.66	0.37	**84.79**	85.55	**87.49**	**91.47**	90.66	**91.43**	**87.95**	**89.26**	**86.68**	**92.17**	**92.40**	**89.27**

The best value is bold and the second best is underlined. In this table, “SF” denotes single-frame detection, “MF” denotes multiframe detection, “LC” denotes local-contrast methods, “NL” denotes nonlocal matrix recovering methods, “seg” denotes segmentation-based methods, and “reg” denotes regression-based mMethods. No acceleration like NVIDIA TensorTR or Open Neural Network Exchange (ONNX) is applied during speed evaluation on all platforms.

In a short training sample sequence, the crop positions of the images are not the same. In other words, the backgrounds are not aligned during the training no matter whether the original sequences are captured in a static way or not. Other random data augmentation such as image flip and gray stretch are the same for a small sequence in each iteration. We train our NLMF-Net for another 200 000 iterations by training the SF part and the whole network alternately. The setting of hyperparameters is usually flexible but follows some rules. For MF training, the loss function contains both the SF and MF parts. Since the loss of SF network backpropagates independently during the alternating training, weight parameter λ_1 in (5) is set lower than λ_2 during the MF alternate training. As for the weights in (6), it is not difficult for regression-based models to locate targets for the potential target grid cells. The main difficulty of IST detection is to discriminate whether there is a target in the grid cell. Besides, ISTs are sparse in images. Negative samples are much more than positive samples even in the high-confidence grid cells. To suppress backgrounds better, α_B is usually set higher. In our experiments, λ_1 and λ_2 are set as 0.1 and 1.0. α_T, α_B, and α_R are set as 0.1, 1.0, and 0.1, respectively. For the temperature hyperparameter in (3), because the similarities are normalized values to some extent, they are not larger than 1. To enlarge the weight of correct target grid cells, γ should be assigned a smaller value and we set it as 0.02.

TABLE III
MEMORY REQUIREMENT WHEN MODELS ARE APPLIED ON CPU OR GPU

Method	CPU (MB)	GPU (MB)
MDvsFA [13]	28	468
DNANet [1]	48	160
YOLOX-x [9]	787	859
YOLOX-nano [9]	33	15
MMRFF-Net [2]	21	18
ours-SF	14	16
ours-MF	15	16

B. Comparison of IST-S and IRDST

In this subsection, we compare our model with several state-of-the-art methods including adaptive scale patch-based contrast measure (ASPCM) [3], the partial sum of the tensor nuclear norm (PSTNN) [4], STLCF [5], STLDM [20], IMNNLWEC [6], SRSTT [21], MDvsFA [13], ALCNet [14], DNANet [1], MMRFF-Net [2], and YOLOX [9]. For fair comparisons, all classic and CNN methods are evaluated on an i7-10850H (2.3 GHz) with a NVIDIA RTX2060 platform. Quantitative evaluation values are presented in Table II and typical detection results are presented in Fig. 8. AP and ROC curves based on 3p25 and 3pOr50 metrics are presented in Fig. 9. The memory requirements for models when they are implemented on CPU or GPU are presented in Table III.

Classic methods are designed based on specific priors. They may work well on the prior-satisfied cases. However, too many false alarms come out when they are implemented on datasets with various cases for both LC and NL matrix recovery methods. Although some methods [5], [6], [20], [21] conduct the processing on MFs, it is still too difficult for them to suppress strong edges and segment the targets accurately even based on the AP3pOr50 metric.

For CNN-based methods, ALCNet [14] performs best among the segmentation-based methods due to the successful model-driven framework. Simple wide channels in [13] or dense skip feature fusion [1] cannot increase the performance and may lead to bad influences on the generation. Although ALCNet achieves high scores in the ROC metric, it still performs worse than regression-based methods because of pixel-level false alarms. These pixel-level false alarms are still very common just like the circled clusters in the fourth example, which limits the performance of AP and F1 evaluation. When compared with YOLOX-series methods (regression-based), the NLMF-Net outperforms a lot in IST-S and IRDST. Referred to Fig. 7, ISTs in IST-S are smaller than the ones in IRDST. The first and the second examples presented in Fig. 8 reveal that YOLOX cannot distinguish small targets as well as NLMF-Net. It is also demonstrated by the sharp drop of AP curves and gradual increase of ROC when recall and false alarms ratio are small in Fig. 9 on both datasets.

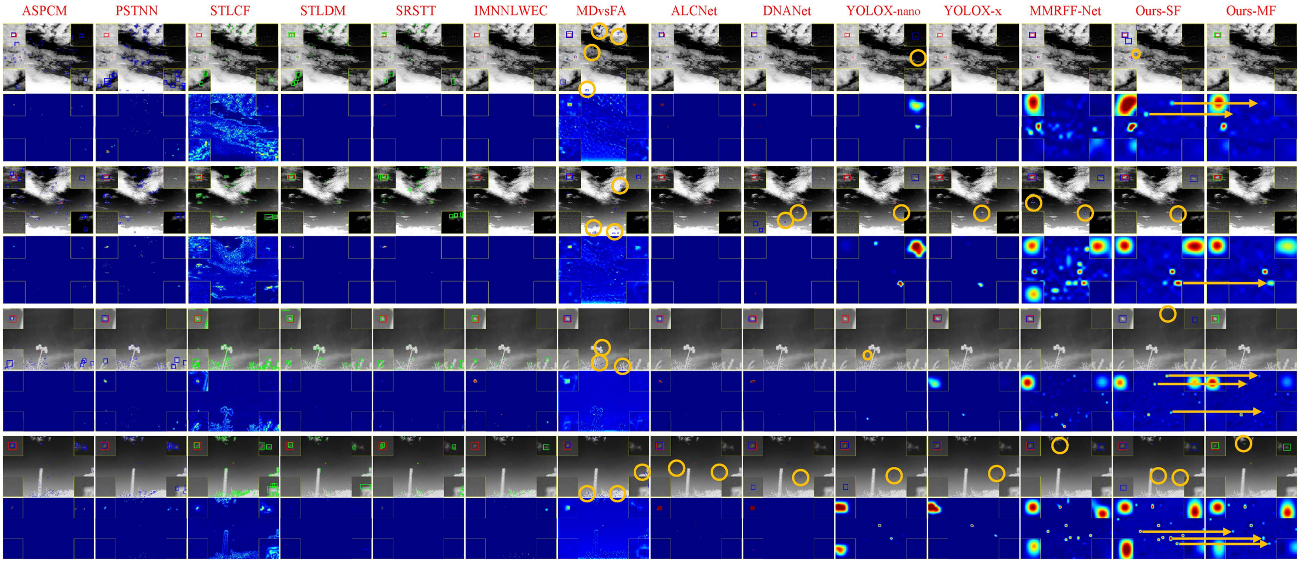
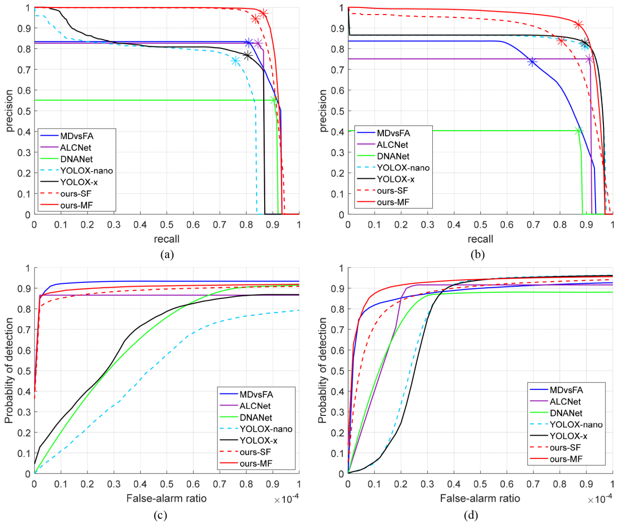


Fig. 8. Detection results from different methods. The first and second examples are from IST-S, while others are from IRDST. All red boxes, blue boxes, and green boxes denote groundtruth boxes, SF detection results, and MF detection results. “Ours-SF” and “Ours-MF” denote the simplified MMRFF backbone introduced in Section III-A and the NLMF-Net, respectively. All blue and green boxes without any red boxes overlapped are false alarms for the corresponding methods. Enlarged views around the true targets are shown at the left top corner, while typical false alarms are presented at other corners for all examples. Some but not all false alarms of CNN methods are labeled by orange circles. Our NLMF detection method can pick most targets out and further suppress the backgrounds by fusing ST features based on SF detection.



(a) (b) (c) (d)

Fig. 9. AP and ROC curve of CNN methods. Classic methods perform much worse than CNN methods. To show the performance of CNN methods clearly, corresponding curves of classic methods are not presented. The positions of best F1 scores are stared in AP curves. The NLMF-Net improves the performance of SF detection and performs better than other methods. (a) AP curve of IST-S. (b) AP curve of IRDST. (c) ROC of IST-S. (d) ROC of IRDST.

The NLMF-Net performs more balanced on both two datasets. The simplified MMRFF [2] backbone extraction fuses features from all scales and preserves the detection for ISTs with various sizes in the SF detection. Although the SF detection process fails to distinguish some nonstationary clusters, most of them are further suppressed when ST features are fused. Some typical examples are labeled by the orange arrows in Fig. 8. In these examples, some clusters possess very high confidence in SF detection. With only SF features, it is too difficult to distinguish them. Anyway, these similar cluster pixels may be just true targets in other cases. After correlating and fusing features among high-confidence grid cells, the MF head outputs a more accurate prediction for true targets and target-similar clusters. All quantitative values in Table II and curves in Fig. 9 demonstrate the validity of our NLMF-Net for increasing performance. More importantly, our NLMF framework is implemented only on high-confidence grid cells, which decreases the computations a lot. Compared to the only SF detection, NLMF-Net only increases computations by no more than 0.01 GFLOPs. It ensures that our model can be applied for 288 × 384 image processing at a speed of 47 FPS on CPU and 20 FPS on Jetson Nano. Besides the platforms mentioned previously, NLMF-Net can be also deployed on Xiaomi Mi Pad 2 (Intel Atom x5-Z8500 CPU) at about 8 FPS for 288 × 384 images and over 35 FPS for 128 × 128 part image processing when open neural network exchange (ONNX) is adopted. In Table III, we also present the memory requirements for CNN models when they are deployed on CPU or GPU. The model initialization, pretrained model loading, and inference process are considered during the memory statistics. More specifically, the maximum memory difference for CPU or GPU values between the beginning of the model initialization and the end of the inference are recorded. The NLMF-Net only needs a small memory for inference on both CPU and GPU.

C. Comparison of SIATD

In this experiment, we follow the official splitting of the dataset and the evaluation metric [48]. Because NLMF-Net will output the prediction bounding box size, we set the size of the label as 1 × 1 for all ISTs during the training. All the targets are smaller than 3 × 3 and most of them are just 1 × 1 by our rough statistic. During the evaluation, only the center positions of all

TABLE IV
EXPERIMENT RESULTS OF SIATD

Method	Precision (%)	Recall (%)	F1 (%)	FPS	Platform
MDvsFA [13]	98.01	56.23	71.46	8.7	2060 TI
ALCNet [14]	42.85	83.97	56.74	30	2060 TI
IAANet [26]	98.70	68.14	80.62	4.4	2060 TI
DNANet [1]	6.72	6.82	6.77	31	2060 TI
BPR-Net [27]	96.81	91.52	94.09	11	2060 TI
STDMANet [15]	97.55	**97.33**	**97.44**	27	3090 TI
MMRFF-Net [2]	98.71	89.39	93.82	<u>124</u>	2060 TI
ours-SF	<u>98.75</u>	90.98	94.70	**207**	2060 TI
ours-MF	**99.03**	<u>91.99</u>	<u>95.38</u>	120	2060 TI

The performance values are referred from [15] and [27]. In this table, the platform Nvidia GeForce RTX 2060 Ti or 3090 Ti is abbreviated as 2060 Ti or 3090 Ti directly. The F1 evaluation is based on the official method like [27]. The FPS values of all methods except STDMANet [15] are evaluated based on the open source codes or the provided codes from the authors on the same platform.

predictions are fed to the official evaluation method. Results are presented in Table IV.

When comparing with other state-of-the-art methods on SIATD, only STDMANet performs better than us, while NLMF-Net outperforms all other methods in F1 score. With NL ST feature fusion, the performance of NLMF-Net is promoted. In Table IV, only STDMANet and our NLMF-Net are MF detection networks. STDMANet is a network that takes a batch image as input. In its temporal multiscale feature extraction, the output features are aggregated from three paths. Especially, the input images should be aligned before the extraction by scale-invariant feature transform [17], which also costs some additional time. And the following dense convolutions lead to large computations. Its success in the highest F1 scores owes to the differential information, which can obtain the moving cues of targets especially when ISTs are weak and ambiguous with the backgrounds. By contrast, our NLMF-Net is an MF network based on NL patches. Thus, our NLMF-Net may fail to pick out some ambiguous targets but is much faster than STDMANet (95.38F1 at 120 FPS with 2060TI versus 97.44F1 at 27FPS with 3090TI).

Anyway, most other methods mainly try to detect more targets at the expense of many computations. NLMF-Net increases the accuracy by adopting a few computations and achieves the second-highest recall with the highest accuracy. Features from all scales and levels with a few channels are enough for IST detection. And NL ST feature fusion further increases the accuracy by focusing on the grid cells with high confidence. In Table IV, NL ST feature fusion improves the F1 score from 94.70% to 95.38% based on the official evaluation metric [48]. If we evaluate based on our metric, NLMF-Net will improve the F1 score from 94.20% to 95.00%. Too deep layers or wide channels are redundant for IST detection especially when targets are very small just like the ones in SIATD. Even our SF part network can easily beat other CNN methods except STDMANet with the quickest speed. Based on SF detection results, our NLMF-Net runs quicker than most other methods with the second-highest F1 score on the open SIATD. Because source codes of some methods [15], [27] are not open now and the other baselines are trained by previous researchers, we only show some examples of our NLMF-Net to present the superiority of NL ST fusion in Fig. 10.

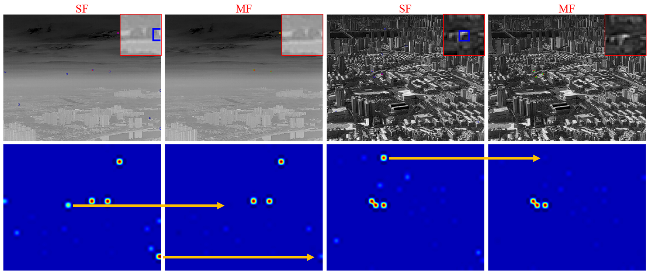


Fig. 10. Two detection examples of NLMF-Net on SIATD. NLMF-Net suppresses the background better than the SF network by fusing ST features. A typical background cluster is shown at the right-top corner. Some cluster suppression examples are labeled by orange arrows. NLMF-Net further suppresses some target-similar clusters and increases the accuracy of IST detection.

TABLE V
PERFORMANCE ON DIFFERENT DATASETS WHEN THE CONFIDENCE THRESHOLD IS SET AS 0.9 FOR BOTH SF BASELINE AND NLMF-NET

Datasets	IST-S		IRDST		SIATD	
	SF	MF	SF	MF	SF	MF
Targets	15134	12851	39879	32724	112434	108283
Backgrounds	6977	53	102422	1578	20151	631
False Alarm Ratio	0.316	0.004	0.720	0.046	0.152	0.005
Confidence Difference	0.754	0.871	0.403	0.859	0.711	0.927

In this table, "targets" denotes the successful detection number of true targets, "backgrounds" denotes the number of false alarm clusters, and the "confidence difference" denotes the average Euclidean distance in confidence dimension between backgrounds and targets.

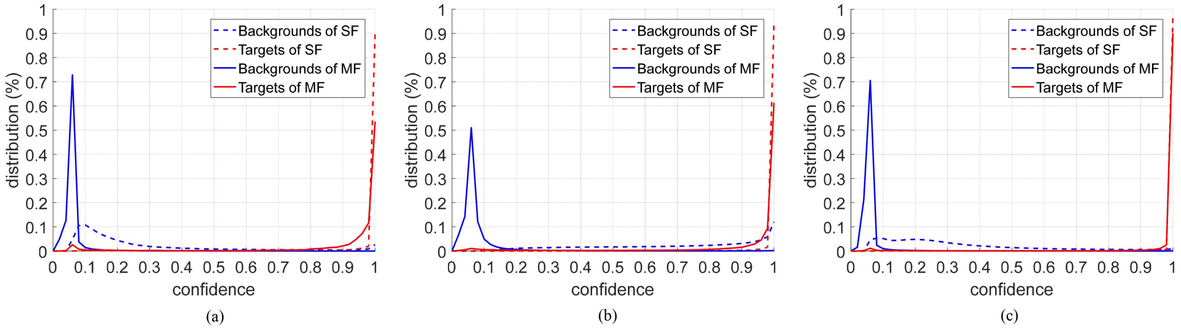
D. False Alarm Suppression of NLMF-Net

In previous subsections, we have presented some examples of background clutter suppression. The quantitative results also show that NLMF-Net can continuously increase F1 scores on three datasets. In this subsection, more concrete results are presented in Table V and Fig. 11 to show the powerful ability of NLMF-Net to suppress background clusters.

In Table V and Fig. 11, only the top N (Default value is 10) grid cells with the highest confidence are considered. Compared to the SF baseline, NLMF-Net enlarges the confidence difference by 15.5%, 113%, and 30.4% between targets and backgrounds with high confidence. Although the average confidence values for both targets and backgrounds are decreased, the background ones are suppressed more obviously. By correlating, aggregating, and fusing ST features among MFs, NLMF-Net can make a more accurate judgment for those target-similar clusters and decrease the false alarms by 98.7%, 93.6%, and 96.7% on the three datasets. On the whole, NLMF-Net promotes performance a lot by suppressing false alarms at the expense of a small recall decrease.

E. Ablation of NL Fusion

In this subsection, an ablation study of NL fusion is conducted. Quantitative evaluation results are presented in Table VI. Models with only the NL fusion strategy perform better than the SF model. In particular, F1 values are promoted continually by adding modules one by one. Although the AP and ROC may not



(a)

(b)

(c)

Fig. 11. Confidence distribution of target and background clusters in several datasets. NLMF-Net decreases the average confidence of background clusters by 0.217, 0.615, and 0.270, while the average target confidences are decreased by only 0.078, 0.097, and 0.027, respectively. NLMF-Net increases the accuracy by suppressing backgrounds. (a) Confidence distribution of IST-S. (b) Confidence distribution of IRDST. (c) Confidence distribution of SIATD.

TABLE VI
Ablation of Non-Local Fusion. In this Table, VLSM and NMSS Contribute a Lot to the Performance Promotion of NLMF-Net.

NL	VLSM	NMSS	SF/MF	IST-S						IRDST [35]					
				3p25			3pOr50			3p25			3pOr50		
				AP	ROC	F1	AP	ROC	F1	AP	ROC	F1	AP	ROC	F1
			SF	83.55	83.79	85.10	90.36	88.77	88.62	84.31	85.38	81.20	86.05	86.73	82.10
✓			MF	83.00	83.77	85.93	89.94	89.00	89.94	85.22	85.07	83.89	89.54	88.62	86.83
✓	✓		MF	82.85	83.86	86.19	89.78	88.98	90.03	**89.57**	89.22	86.50	91.12	90.53	87.50
✓	✓	✓	MF	**84.79**	**85.55**	**87.49**	**91.47**	**90.66**	**91.43**	87.95	**89.26**	**86.68**	**92.17**	**92.40**	**89.27**

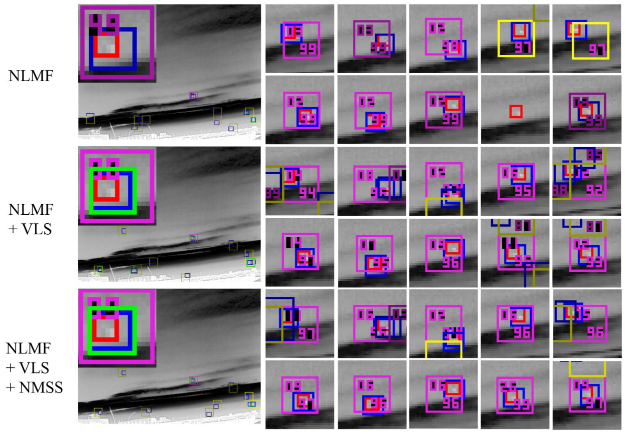


Fig. 12. How VLSM and NMSS promote the performance. The detection results are presented in the left and enlarged views of the true targets in the past ten frames are presented in the right. All red-purple and yellow boxes denote the grid cells of the targets or clusters. If the contribution proportions of the ST features are over 1%, the corresponding grid cells will be labeled by red-purple boxes. In each grid cell, the ST feature contribution proportion is printed at the left-top corner and the similarity is printed at the right-bottom corner. All red, blue, and green boxes denote groundtruth, SF, and MF detection results, respectively. VLSM and NMSS improve the feature correlation and fusion, which contributes to the performance promotion.

be promoted in a few cases, the scores may be just influenced by detection abilities when accuracy is lower. Anyway, a higher F1 means a better balance between accuracy and recall.

TABLE VII
Quantitative Statistics of VLSM Effect

Dataset	VLSM	SIM with IST	SIM with BG	Ratio of SIM
IST-S		20.15	168.45	0.120
	✓	18.47	34.23	0.540
IRDST		20.17	167.32	0.121
	✓	17.44	33.77	0.516

"SIM with IST" denotes the average similarity between the current frame target and all the target grid cells in the memory bank. "SIM with BG" denotes the ones between the current frame target and background grid cells. "Ratio of SIM" denotes the ratio of "SIM with IST" to "SIM with BG."

In our experiments, VLSM can enlarge the difference between true ISTs and target-similar clusters properly. With only normalized similarity, the proportion of true IST weight in the last ST features is usually low. A typical example is presented in Fig. 12. Without VLSM and NMSS, the similarity values between true targets and some target-similar clusters are usually high. The weight of the feature contribution from the true targets is only 17%. Such a low value means the NLMF-Net cannot collect enough and correct ST features from the consistent IST, which makes the model trained not well. If the VLSM is applied, the difference between different grid cells is enlarged obviously. And the weight increases to 65%, too. To further suppress the clusters in each frame, the NMSS helps increase the weight to 86%. Although some target-similar clusters still contribute to the last fusion, enough correct features are collected to ensure a more accurate and robust judgment for each high-confidence grid cell. More quantitative and comprehensive statistics are presented in Tables VII and VIII. By comparing the "Ratio of SIM," VLMS improves the feature correlation from 0.120 to 0.540 on IST-S and from 0.121 to 0.516 on IRDST. By comparing the "Ratio of Weight," NMSS further suppresses the backgrounds and improves the feature fusion from 6.226 to 10.11 on IST-S and from 3.344 to 4.960 on IRDST. On the whole, both the VLSM and NMSS help improve the performance of NLMF-Net.

TABLE VIII
QUANTITATIVE STATISTICS OF NMSS EFFECT

Dataset	NMSS	Weight of IST	Weight of BG	Ratio of Weight
IST-S		0.797	0.128	6.226
	✓	0.839	0.083	10.11
IRDST		0.689	0.206	3.344
	✓	0.739	0.149	4.960

"Weight of IST" denotes the contribution of fused features from target features while "weight of BG" denotes the contribution from background ones. "Ratio of weight" denotes the ratio of "weight of IST" to "weight of BG."

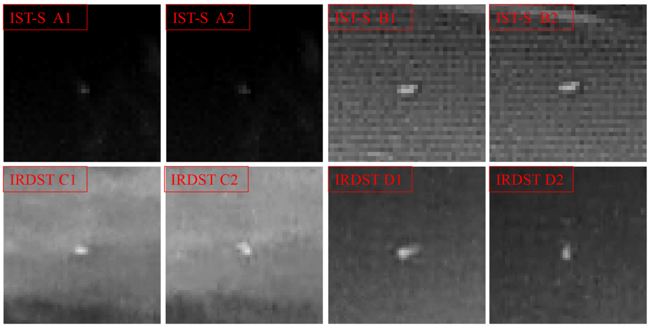


Fig. 13. Four enlarged views of noisy examples. The frame interval of examples is small. A certain degree of noise is widespread. The main noises in IST-S are fixed, while the ones in IRDST are random in most cases.

F. Noisy Robustness Analysis

Although great progress has been made in infrared imaging technology in recent decades, a certain degree of nonuniformity noises still widely exist. Some noisy examples are presented in Fig. 13.

In this subsection, we synthesize noisy images based on IST-S by adding random Gaussian noises directly. The additional noisy standard deviation (std.) values are set as 0.01 and 0.02 in the experiments. Only ALCNet and YOLOX-x are adopted for comparison. Quantitative results are presented in Table IX and corresponding curves are presented in Fig. 14. All the models are adopted from Section V-B directly and not retrained by adding heavier noises.

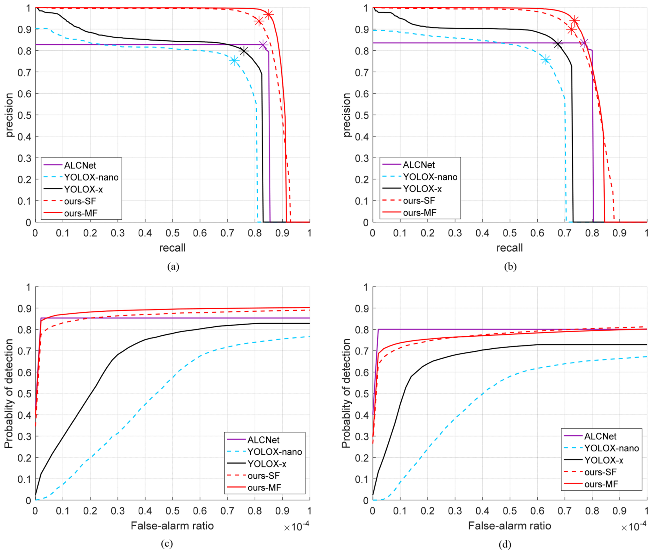


Fig. 14. AP curve and ROC on noisy datasets. (a) AP curve of IST-S (noisy std = 0.01). (b) AP curve of IST-S (noisy std = 0.02). (c) ROC of IST-S (noisy std = 0.01). (d) ROC of IST-S (noisy std = 0.02).

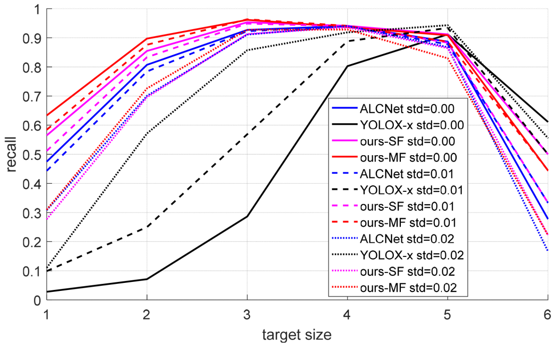


Fig. 15. Recall curves when demanding accuracy is 85% based on IST-S. ALCNet cannot achieve such a demanding accuracy and we present the recall curve when the F1 scores are the highest. NLMF-Net is robust to a certain degree of noise and performs well among ISTs with various sizes.

Deeper networks such as YOLOX-x are not competent for smaller IST detection. YOLOX-x predicts targets by only deep features, which limit their ability to detect smaller targets. Recall curves presented in Fig. 15 reveal that it can perform better when targets are larger. In Fig. 15, it seems that YOLOX-x achieves higher recall values on smaller targets when noises are heavier. It does not mean YOLOX-x performs better than itself in heavier noisy cases. Noise will bury some background clusters and targets at the same time. All methods drop their abilities to pick more targets out by comparing AP curves between Figs. 9 and 14. The synthetic noises just bury the target-similar clusters, and YOLOX-x achieves a relatively higher accuracy due to fewer false alarms in some cases. It increases the accuracy of YOLOX-x when recall is not high by referring to Figs. 9 and 14. Then, more successful detection results are counted during the noisy recall statistics. Actually, the performance of YOLOX-x also drops and YOLOX-x cannot distinguish smaller targets and background clusters especially when they are not larger than 4 × 4 for both original IST-S or synthetic noisy datasets. The other network ALCNet is a model-driven network that combines deep learning and traditional priors. It preserves the balance for detecting ISTs with various sizes but not as well as NLMF-Net. Although ALCNet achieves high scores in ROC metric, pixel-level false alarms limit its accuracy on object-level detection.

By contrast, NLMF-Net performs better on the whole. Although the performance is influenced by the noises, NLMF-Net still preserves a good recall on smaller targets. Features from all the scales help the model pick out true targets with various sizes. And ST feature fusion makes the model more robust to the background clusters especially when these clusters are not stable. The attention should be drawn that the performance of NLMF-Net tends to decrease more if noises are too heavy. MMRFF [2] backbone fuses both shallow and deep features. When the network cannot detect some small targets by taking full use of only deep features, shallow features may help the model to work it out. However, shallow features are not robust to noises. When heavier noises are involved, not only will the

TABLE IX
PERFORMANCE ON NOISY DATASETS

Noisy std		0.01						0.02					
Dataset	Method	3p25			3pOr50			3p25			3pOr50		
		AP	ROC	F1	AP	ROC	F1	AP	ROC	F1	AP	ROC	F1
IST-S	ALCNet	67.91	83.62	81.32	70.60	85.36	82.82	64.29	**78.46**	78.70	66.81	**80.06**	80.16
	YOLOX-nano	49.11	36.09	62.44	65.36	49.59	73.94	48.40	39.56	61.19	58.64	47.53	68.77
	YOLOX-x	63.28	60.95	72.36	71.43	68.25	77.87	60.39	61.04	70.68	65.74	65.91	74.51
	ours-SF	82.68	82.46	84.08	88.53	86.72	87.17	76.70	74.52	78.30	80.86	77.29	80.26
	ours-MF	**84.17**	**84.60**	**86.91**	**89.98**	**89.02**	**90.52**	**78.11**	75.08	**80.63**	**81.16**	77.26	**82.50**

Only the best values are bold.

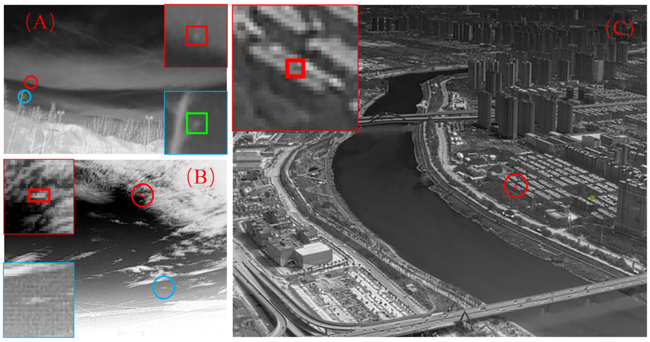


Fig. 16. Some ambiguous examples. The failures are labeled by red circles, while some target-similar clusters are labeled by light blue circles. It is difficult for SF methods or modules to distinguish such ambiguous targets and some stable clusters.

smaller targets become weaker, but also the shallow features may be also heavily influenced. It may even drop the performance on larger targets. That is, NLMF-Net can tolerate a certain degree of noise, which has been presented in Fig. 13. If more noises are added, performance for both smaller (not larger than 4 × 4) and large targets will drop down for our NLMF-Net.

G. Limitation Analysis

Though the performance of NLMF-Net is promoted by the ST feature fusion, there are still many other cases that the NLMF-Net cannot deal with. First, the performance of NLMF-Net highly relies on SF detection. NLMF-Net cannot deal with images when targets are ambiguous with backgrounds and researchers even cannot make a quick and correct judgment. The NL feature fusion is based on the grid cells with high confidence that are picked out by the SF network. In other words, if the extremely weak targets cannot be found by the SF network, they will be neglected in the following MF detection. Some examples are presented in Fig. 16. The other case has been also illustrated in Fig. 15 that our NLMF-Net cannot tolerate too heavy noises such as deeper networks [9], [14]. Heavier noises will harm the shallow features for SF prediction, which is not beneficial to the MF feature correlation and fusion.

Although the performance of NLMF-Net is limited by the SF detection and it cannot tolerate too heavy noise, NL ST feature fusion increases the accuracy obviously. And current imaging technology is sufficient to ensure a good image quality without too much noise. At least, NLMF-Net performs best in the original IST-S and IRDST with a quick speed compared to other methods. On the whole, NLMF-Net is practical and runs quickly to deal with IST detection and suppress the unstable clusters well.

VI. CONCLUSION

In this article, we propose an NLMF-Net for IST detection in sequence. The NLMF-Net fuses MF features by focusing on the high-confidence grid cells, while over 99.94% of redundant and ineffective features are neglected. The MF processing only introduces no more than 0.01 GFLOPs, which ensures real-time processing on resource-limited machines. Besides, the VLSM enlarges the difference between targets and background clusters. The NMSS further suppresses the weight of background clusters. They improve the feature correlation and fusion among grid cells between current and past frames. Based on the three datasets, NLMF-Net enlarges the confidence difference between targets and backgrounds by 15.5%, 113%, and 30.4%, which increase the accuracy and improve the performance by 2.81, 7.17, and 0.68 for F1 evaluation when compared to the SF baseline. Extensive experimental results demonstrate the balanced performance of the proposed network compared to classic and other CNN methods for both accuracy and speed in practical scenes.

REFERENCES

[1] B. Li et al., “Dense nested attention network for infrared small target detection,” *IEEE Trans. Image Process.*, vol. 32, pp. 1745–1758, Aug. 2022.

[2] H. Xu, S. Zhong, T. Zhang, and X. Zou, “Multiscale multilevel residual feature fusion for real-time infrared small target detection,” *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Apr. 2023, Art. no. 5002116.

[3] Z. Qiu, Y. Ma, F. Fan, J. Huang, and M. Wu, “Adaptive scale patch-based contrast measure for dim and small infrared target detection,” *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Dec. 2020, Art. no. 7000305.

[4] L. Zhang and Z. Peng, “Infrared small target detection based on partial sum of the tensor nuclear norm,” *Remote Sens.*, vol. 11, no. 4, 2019, Art. no. 382.

[5] L. Deng, H. Zhu, C. Tao, and Y. Wei, “Infrared moving point target detection based on spatial–temporal local contrast filter,” *Infrared Phys. Technol.*, vol. 76, pp. 168–173, 2016.

[6] Y. Luo, X. Li, S. Chen, C. Xia, and L. Zhao, “IMNN-LWEC: A. novel infrared small target detection based on spatial–temporal tensor model,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Dec. 2022, Art. no. 5004022.

[7] X. Zhao, K. Liu, K. Gao, and W. Li, “Hyperspectral time-series target detection based on spectral perception and spatial–temporal tensor decomposition,” *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Aug. 2023, Art. no. 5520812.

[8] X. Zhao, W. Li, C. Zhao, and R. Tao, “Hyperspectral target detection based on weighted cauchy distance graph and local adaptive collaborative representation,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Apr. 2022, Art. no. 5527313.

[9] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, “Yolox: Exceeding Yolo series in 2021,” 2021, *arXiv:2107.08430*.

[10] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, “Yolov4: Optimal speed and accuracy of object detection,” 2020, *arXiv:2004.10934*.

[11] H. Chen, K. Sun, Z. Tian, C. Shen, Y. Huang, and Y. Yan, “BlendMask: Top-down meets bottom-up for instance segmentation,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recogn.*, 2020, pp. 8573–8581.

[12] M. Ju, J. Luo, G. Liu, and H. Luo, “ISTDet: An efficient end-to-end neural network for infrared small target detection,” *Infrared Phys. Technol.*, vol. 114, 2021, Art. no. 103659.

[13] H. Wang, L. Zhou, and L. Wang, “Miss detection vs false alarm: Adversarial learning for small object segmentation in infrared images,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 8509–8518.

[14] Y. Dai, Y. Wu, F. Zhou, and K. Barnard, “Attentional local contrast networks for infrared small target detection,” *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 11, pp. 9813–9824, Nov. 2021.

[15] P. Yan, R. Hou, X. Duan, C. Yue, X. Wang, and X. Cao, “STDMANet: Spatio-temporal differential multiscale attention network for small moving infrared target detection,” *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Feb. 2023, Art. no. 5602516.

[16] M. Gupta et al., “Infrared small target detection enhancement using a lightweight convolutional neural network,” *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Sep. 2022, Art. no. 3513405.

[17] D. Lowe, “Object recognition from local scale-invariant features,” in *Proc. 7th IEEE Int. Conf. Comput. Vis.*, 1999, vol. 2, pp. 1150–1157.

[18] E. Bayraktar, C. B. Yigit, and P. Boyraz, “A hybrid image dataset toward bridging the gap between real and simulation environments for robotics: Annotated desktop objects real and synthetic images dataset: Adoreset,” *Mach. Vis. Appl.*, vol. 30, no. 1, pp. 23–40, 2019.

[19] S. Yao, Y. Chang, and X. Qin, “A coarse-to-fine method for infrared small target detection,” *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 2, pp. 256–260, Feb. 2019.

[20] P. Du and A. Hamdulla, “Infrared moving small-target detection using spatial–temporal local difference measure,” *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 10, pp. 1817–1821, Oct. 2020.

[21] J. Li, P. Zhang, L. Zhang, and Z. Zhang, “Sparse regularization-based spatial-temporal twist tensor model for infrared small target detection,” *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Jan. 2023, Art. no. 5000417.

[22] Y. Dai and Y. Wu, “Reweighted infrared patch-tensor model with both non-local and local priors for single-frame small target detection,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 8, pp. 3752–3767, Aug. 2017.

[23] F. Yan, G. Xu, J. Wang, Q. Wu, and Z. Wang, “Infrared small target detection via Schatten capped pNORM-based non-convex tensor low-rank approximation,” *IEEE Geosci. Remote Sens. Lett.*, vol. 20, Dec. 2022, Art. no. 6002105.

[24] M. Wu et al., “Infrared moving small target detection based on consistency of sparse trajectory,” *IEEE Geosci. Remote Sens. Lett.*, vol. 20, Mar. 2023, Art. no. 6003605.

[25] Z. Gao, H. Chang, X. Cheng, W. Liu, X. Wang, and W. Ying, “A fast detection method for infrared small targets in complex sea and sky background,” in *Proc. Asia Conf. Algorithms, Comput. Mach. Learn.*, 2022, pp. 44–49.

[26] K. Wang, S. Du, C. Liu, and Z. Cao, “Interior attention-aware network for infrared small target detection,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Mar. 2022, Art. no. 5002013.

[27] 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.

[28] E. Bayraktar, M. E. Basarkan, and N. Celebi, “A low-cost UAV framework towards ornamental plant detection and counting in the wild,” *ISPRS J. Photogrammetry Remote Sens.*, vol. 167, pp. 1–11, 2020.

[29] E. Bayraktar, B. N. Korkmaz, A. U. Erarslan, and N. Celebi, “Traffic congestion-aware graph-based vehicle rerouting framework from aerial imagery,” *Eng. Appl. Artif. Intell.*, vol. 119, 2023, Art. no. 105769.

[30] A. Howard et al., “Searching for MobileNetV3,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 1314–1324.

[31] X. Zhang, X. Zhou, M. Lin, and J. Sun, “ShuffleNet: An extremely efficient convolutional neural network for mobile devices,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6848–6856.

[32] H. Liu, F. Sun, J. Gu, and L. Deng, “SF-Yolov5: A lightweight small object detection algorithm based on improved feature fusion mode,” *Sensors*, vol. 22, no. 15, 2022, Art. no. 5817. [Online]. Available: https://www.mdpi.com/1424-8220/22/15/5817

[33] D. Biswas, M. M. Rahman, Z. Zong, and J. Tešić, “Improving the energy efficiency of real-time DNN object detection via compression, transfer learning, and scale prediction,” in *Proc. IEEE Int. Conf. Netw., Archit. Storage*, 2022, pp. 1–8.

[34] E. Bayraktar, Y. Wang, and A. DelBue, “Fast re-OBJ: Real-time object re-identification in rigid scenes,” *Mach. Vis. Appl.*, vol. 33, no. 6, 2022, Art. no. 97.

[35] H. Sun, J. Bai, F. Yang, and X. Bai, “Receptive-field and direction induced attention network for infrared dim small target detection with a large-scale dataset IRDST,” *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Jan. 2023, Art. no. 5000513.

[36] A. Amudhan and A. Sudheer, “Lightweight and computationally faster hypermetropic convolutional neural network for small size object detection,” *Image Vis. Comput.*, vol. 119, 2022, Art. no. 104396.

[37] Z. Lin, Y. Luo, Q. Ling, T. Wu, C. Xiao, and B. Li, “Video infrared small target detection combining hybrid attention with cross-scale feature fusion,” in *Proc. 2nd Int. Conf. Front. Electron., Inf. Comput. Technol.*, 2022, pp. 512–518.

[38] J. Du et al., “A spatial-temporal feature-based detection framework for infrared dim small target,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Oct. 2021, Art. no. 3000412.

[39] X. Wang, R. Girshick, A. Gupta, and K. He, “Non-local neural networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7794–7803.

[40] H. Hu, Z. Zhang, Z. Xie, and S. Lin, “Local relation networks for image recognition,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 3464–3473.

[41] J. Yu, J. Liu, L. Bo, and T. Mei, “Memory-augmented non-local attention for video super-resolution,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 17834–17843.

[42] H. Seong, J. Hyun, and E. Kim, “Kernelized memory network for video object segmentation,” in *Proc. 16th Eur. Conf. Comput. Vis.*, 2020, pp. 629–645.

[43] M. Li, L. Hu, Z. Xiong, B. Zhang, P. Pan, and D. Liu, “Recurrent dynamic embedding for video object segmentation,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 1332–1341.

[44] S. Hinterstoisser et al., “Gradient response maps for real-time detection of textureless objects,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 5, pp. 876–888, May 2012.

[45] P. Sun et al., “DanceTrack: Multi-object tracking in uniform appearance and diverse motion,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 20993–21002.

[46] S. Liu, Z. Li, and J. Sun, “Self-EMD: Self-supervised object detection without imagenet,” 2020, *arXiv:2011.13677*.

[47] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, “Distance-IOU loss: Faster and better learning for bounding box regression,” in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, pp. 12993–13000.

[48] X. Sun et al., “A dataset for small infrared moving target detection under clutter background,” 2022. [Online]. Available: https://www.scidb.cn/en/detail?dataSetId=808025946870251520

Hai Xu was born in 1994. He received the bachelor’s degree in control science and engineering from the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan, China, in 2016. He is currently working toward the Ph.D. degree in HUST.

His research interests include image recovery, pattern recognition, and object tracking, especially for infrared images.

Sheng Zhong was born in 1972, China. He received the Ph.D. degree in control science and engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2005.

He is currently a Professor with the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology. His research interests include computer vision, pattern recognition, and intelligent system.

Tianxu Zhang was born in Chongqing, China, in 1947. He received the B.S. degree in electrical engineering from the University of Science and Technology of China, Hefei, China, in 1970, the M.S. degree in computer science from the Huazhong University of Science and Technology, Wuhan, China, in 1981, and the Ph.D. degree in optical engineering from Zhejiang University, Hangzhou, China, in 1989. He is a Professor with the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan, China.

His research interests include image processing, computer vision, pattern recognition, and medical imaging.

Xu Zou (Member, IEEE) was born in 1991, China. He received the Ph.D. degree in biomedical engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2020.

He is currently an Assistant Professor with the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology. His research interests include image processing, computer vision, and intelligent system.