Attentional Local Contrast Networks for Infrared Small Target Detection

Yimian Dai, *Student Member, IEEE*, Yiquan Wu, Fei Zhou, and Kobus Barnard, *Member, IEEE*

***Abstract*— To mitigate the issue of minimal intrinsic features for pure data-driven methods, in this article, we propose a novel model-driven deep network for infrared small target detection, which combines discriminative networks and conventional model-driven methods to make use of both labeled data and the domain knowledge. By designing a feature map cyclic shift scheme, we modularize a conventional local contrast measure method as a depthwise parameterless nonlinear feature refinement layer in an end-to-end network, which encodes relatively long-range contextual interactions with clear physical interpretability. To highlight and preserve the small target features, we also exploit a bottom-up attentional modulation integrating the smaller scale subtle details of low-level features into high-level features of deeper layers. We conduct detailed ablation studies with varying network depths to empirically verify the effectiveness and efficiency of the design of each component in our network architecture. We also compare the performance of our network against other model-driven methods and deep networks on the open SIRST data set as well. The results suggest that our network yields a performance boost over its competitors.**

***Index Terms*— Attention mechanism, deep learning, feature fusion, infrared small target, local contrast.**

I. Introduction

INFRARED small target detection plays an important role in applications such as early-warning systems and maritime surveillance systems because of the ability of infrared imaging that can offer a clear image without illumination or penetrate obstructions, such as fog, smoke, and other atmospheric conditions [1]. Due to the long imaging distance, the infrared target generally ends up only occupying a few pixels in the image, lacking texture or shape characteristics [2]. Facing the lack of intrinsic features, the traditional methods utilize the spatiotemporal continuity of the target on the image sequence for detection by assuming a static background or consistent targets in adjacent frames [3]. Recently, the need for early warning has made researchers pay more and more attention to the single-frame detection task, particularly with the advance in hypersonic aircraft in which the fast-changing backgrounds and inconsistent target motion traces caused by the rapid relative movement between sensor platforms and targets can make the performance of sequential detection methods degrade significantly [4]. Therefore, detecting small targets in a single image is of great importance for such applications.

Conventional model-driven approaches model the infrared small targets as an outlier popping out from the slowly transitional background where nearby pixels are highly correlated [5]. Detecting infrared small targets is thus a form of blob detection, which is a problem with a long history in the image processing literature [6]. However, in real-world scenarios, the problem is more complex; there are more distractors that also stand out as outliers in the background [7]. Hence, model-driven methods must make strong prior assumptions about the small target, e.g., the most sparse [4] or salient [8], about the background, e.g., smooth [9] or nonlocal correlated [10] or both [11]. Traditional image processing formulations of the problem usually utilize only grayscale values as features in the spatial domain [12], lacking a semantic discriminability between the real targets and distractors; the resulting methods can typically handle only very salient targets of high local contrasts and not the dim ones buried in complex background. Furthermore, algorithms exploiting such small target priors are sensitive to the hyperparameters relevant to the image content, e.g., the sparsity control hyperparameter λ in low-rank methods [13] and the preset target size in local contrast methods [14], which fails easily in highly variable scenes with fast-changing backgrounds.

Infrared small target detection can also be modeled as a supervised machine-learning problem, but it has long been stuck with insufficient training data due to the difficulty of collecting infrared small target images. To this end, we recently contribute an open and high-quality data set specially created for single-frame small infrared target detection termed "SIRST," which enables the training of deep networks.[1] However, with the labeled data, accurate infrared small target detection or segmentation can still be challenging for many off-the-shelf baseline networks [15], [16]. The object appearance centered feature representation [17], in which

Manuscript received July 28, 2020; revised October 29, 2020; accepted November 19, 2020. Date of publication January 5, 2021; date of current version October 26, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 61573183, in part by the Open Project Program of the National Laboratory of Pattern Recognition (NLPR) under Grant 201900029, in part by the Nanjing University of Aeronautics and Astronautics PhD Short-Term Visiting Scholar Project under Grant 180104DF03, and in part by the Excellent Chinese and Foreign Youth Exchange Program of China Association for Science and Technology, China Scholarship Council, under Grant 201806830039. *(Corresponding author: Yimian Dai.)*

Yimian Dai, Yiquan Wu, and Fei Zhou are with the College of Electronic and Information Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China (e-mail: yimian.dai@gmail.com; nuaaimage@gmail.com; hellozf1990@163.com).

Kobus Barnard is with the Department of Computer Science, The University of Arizona, Tucson, AZ 85721 USA (e-mail: kobus@cs.arizona.edu).

Digital Object Identifier 10.1109/TGRS.2020.3044958

[1]https://github.com/YimianDai/sirst

many deep networks rely on, may fail due to the scarcity of target intrinsic characteristics as well as the presence of background distractors. Also, most convolutional networks learn high-level semantic features by gradually attenuating the feature map size, making small targets easily being overwhelmed by surrounding background features in deep layers. This means that a specialized "network surgery" is essential for reliable detection performance.

In this article, we advocate the idea of combining the feature learning capacity of deep networks and the physical mechanisms of model-driven methods into an end-to-end network to handle the intrinsic feature scarcity issue for detecting infrared small targets in a single image. In this scheme, the backbone network extracts high-level semantic features of the input image, and then, a certain model-inspired module encodes them into local contrast measures. Unlike purely data-driven methods [18] and purely model-driven methods [19], our approach fully makes use of both labeled data and domain knowledge. As a result, it solves the inaccurate modeling and hyperparameter sensitivity problems of model-driven methods since the network learns discriminative features automatically. It also mitigates the minimal intrinsic feature issue of data-driven approaches by incorporating the domain knowledge of local contrast prior to deep networks.

More specifically, we propose attentional local contrast network (ALCNet), a new model-driven deep network for single-frame infrared small target detection. This advance stems from two key improvements over previous work. First, by designing an acceleration scheme based on feature map cyclic shift, we modularize a local contrast measure method by Wei *et al.* [20] as a depthwise parameterless nonlinear feature refinement layer with clear physical interpretability, which explicitly breaks the limited receptive field imposed by the local nature of convolutional kernels and encodes relatively long-range contextual interactions. Second, to highlight and preserve the small target features, besides adjusting the network downsampling scheme, we also exploit a bottom-up attentional modulation (BLAM) module that encodes the smaller scale subtle details of low-level features into high-level features of deeper layers. Finally, the cross-layer fused feature maps are used for the segmentation task.

To verify the effectiveness of the proposed ALCNet architecture, we conduct extensive ablation studies to investigate the importance of encoding local contrast prior, multiscale local contrast (MLC) measure, and BLAM. We also compare it with other state-of-the-art model-driven methods and data-driven methods on the public SIRST benchmark. The experimental results indicate that the proposed ALCNet achieves the best performance. Our code, trained models, and results are available online.[2]

II. Related Works

Human visual system inspired methods [8] build the local contrast property into models, which is a major characteristic of these featureless infrared small targets. These methods depend heavily on the local contrast measure used to enhance the target and suppress the background clutters [9]. For this purpose, different variants of local contrast measures have been proposed to deal with pixel pulse noise [21] and edges [20]. However, all these methods assume a strict match between the target size and the preset nested cell or patch structure [8], as they take the mean or maximum grayscale or entropy of each patch on the original image as handcrafted features [14]. Such oversimplified raw features are also the root of the inaccurate modeling issue of these model-driven methods. Although this work borrows the multiscale patch-based contrast measure (MPCM) [20] as guidance, unlike the above approaches, our network does not rely on any predefined feature representation nor do we assume that the small target has higher local contrast than any background component. Our network automatically learns the target feature from the labeled data, which makes it more robust in detecting dim targets in complex backgrounds. To the best of our knowledge, this article is the first to modularize a conventional local contrast method and embed it into a convolutional network for infrared small target detection.

Small object detection is a key challenge in generic computer vision as well because there is little signal on the object to exploit [22]. Besides carefully designing the anchor matching strategy [23] or training the network in a scale-aware scheme [24], encoding context that offers more evidence beyond the object extent is also highly explored as a solution to mitigate the problems arising from small objects [25], [26], which is generally done by extracting and concatenating the features of an enlarged window around the objects [22]. Unlike these implicit context encoding approaches, our network explicitly encodes the interaction of each element on the feature map with its distant neighbors, namely the local contrast according to some well-defined physical mechanisms [20]. In addition, the scarcity of intrinsic characteristics is different between generic small objects and the infrared small target studied in this article. Small objects in generic vision tasks generally occupy around 1% of image area [27], while the pixels of an infrared small target may only take 0.1% or less of the image area.

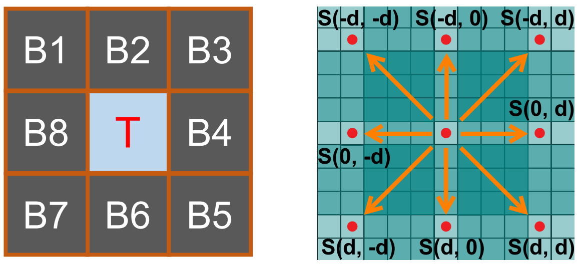
Cross-layer feature fusion has proved to be an effective approach to alleviate the scale variation issue in computer vision, which is often implemented via linear combinations such as addition or concatenation [16], [28] or nonlinear top-down modulation [29], [30]. In Table I, we provide a brief summary of cross-layer feature integration schemes, where PWConv denotes the pointwise convolution, and Concat is the abbreviation of concatenation. $\mathbf{X}$ and $\mathbf{Y}$ are the feature maps of different scales. By default, $\mathbf{Y}$ is the one with a larger receptive field. $\mathbf{G}$ denotes the global channel attention module [31], and $\mathbf{L}$ is the local channel attention module we adopt. However, these approaches are not designed for the task of infrared small target detection, which may fail without considering the special characteristics of this task. Unlike the generic computer vision task, the bottleneck in infrared small target detection is how to preserve and highlight the features of dim and small targets in deep layers, rather than lacking high-level semantics in shallow layers. To this end, our work employs a reverse

[2]https://github.com/YimianDai/open-alcnet

TABLE I

BRIEF OVERVIEW OF CROSS-LAYER FEATURE INTEGRATION SCHEMES. THE PROPOSED ALCNET DIFFERS IN THE MODULATION PATHWAY DIRECTION AND ATTENTION MODULE

Manner	Formulation	Reference
Addition	$\mathbf{X}+\mathbf{Y}$	[28]
Concatenation	$\mathrm{PWConv}(\mathrm{Concat}(\mathbf{X},\mathbf{Y}))$	[16, 32]
Refinement	$\mathbf{X}+\mathbf{G}(\mathbf{Y})\otimes\mathbf{Y}$	[31, 33]
Top-down Modulation	$\mathbf{G}(\mathbf{Y})\otimes\mathbf{X}+\mathbf{Y}$	[29]
Bottom-up Modulation	$\mathbf{X}+\mathbf{L}(\mathbf{X})\otimes\mathbf{Y}$	*ours*



(a) (b)

Fig. 1. Transition from patch to dilation. (a) Nested structure of patch-based contrast measure [20]. The patch size is equal to the filter size. Patches are strictly nonoverlapped. (b) Structure of our DLC measure, in which each point shares the overlapped receptive fields.

bottom-up pointwise attentional modulation pathway, which we think is vital for infrared small target detection.

III. MODULARIZING THE LOCAL CONTRAST PRIOR

In this section, we describe how to convert a conventional patch-based local contrast measure [20] to a nonlinear feature refinement layer, which can be inserted into a network as a plug-in module. Specially, we will deal with the challenges: 1) how to get rid of the patch-based paradigm of local contrast measure [8] which cannot fit into an end-to-end network and 2) how to measure the local contrast on feature maps fast.

A. Dilated Local Contrast Measure

In traditional local contrast measure methods [8], [20], the patch size is both the scale of feature extraction and the scale of local contrast measure, imposing an explicit equality constraint on these two, as shown in Fig. 1(a). To measure the local contrast in an end-to-end network, we abandon the concept of "patch" and its patch mean feature representation in MPCM [20]. As a replacement, we borrow the dilation rate from DeepLab series [34] as a hyperparameter controlling the scale of local contrast measure, as shown in Fig. 1(b). As a result, the feature scale and local contrast measure scale are decoupled in the network, which enables us to measure MLC on multiscale feature maps.

Given an intermediate feature $\mathbf{F} \in \mathbb{R}^{C\times H\times W}$, a specific position (c, i, j), and a dilation rate d, one can write the directional local contrast in a scalar form as

$$\mathbf{D}_{[c,i,j]}^{(x,y)} = \left(\mathbf{F}_{[c,i,j]} - \mathbf{F}_{[c,i-x,j-y]}\right) \cdot \left(\mathbf{F}_{[c,i,j]} - \mathbf{F}_{[c,i+x,j+y]}\right) \quad (1)$$

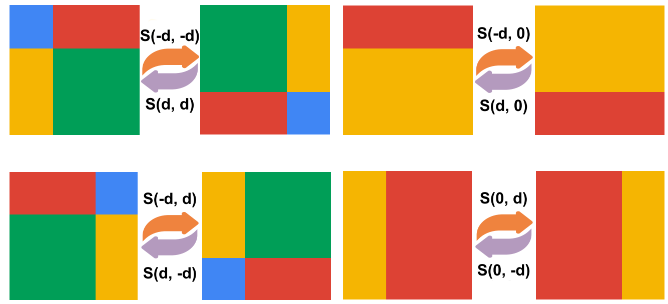


Fig. 2. Illustration of the cyclic shift scheme to obtain eight neighborhood feature maps. The map at the end of the arrow denotes the original feature map, and the map at the arrowhead denotes the shifted feature map in a particular direction.

where $(x, y) \in \Omega = \{(-d,-d), (-d,0), (-d,d), (0,-d)\}$ is the direction index. Then, the scalar local contrast under dilation rate d can be obtained as

$$\mathbf{C}_{[c,i,j]}^{d} = \min_{(x,y)\in\Omega}\left\{\mathbf{D}_{[c,i,j]}^{(x,y)}\right\}. \quad (2)$$

B. Cyclic Shift Accelerating Scheme

To fast calculate the subtraction between the center point and its neighborhoods, MPCM filters the mean image with eight preset kernels [20]. It should be noted that if we utilize this accelerating scheme in a depthwise way in the network, it requires $8(3d)^2\,HW$ multiplications and $8(3d)^2\,HW$ additions for each feature map. The computational cost of the local contrast measure will grow rapidly as the dilation rate d increases, which can be a bottleneck in inference speed.

In this work, we offer another accelerating trick by hypothesizing that the feature map margins are smooth and similar. Actually, this is a reasonable hypothesis. On the one hand, the original infrared image has both strong local and nonlocal correlations. On the other hand, most background components have been suppressed by previous convolutional layers. Therefore, we can formulate 1 into a tensor form and compute it as a whole as

$$\mathbf{D}^{(x,y)} = \left(\mathbf{F} - \mathbf{S}_{(x,y)}\right) \otimes \left(\mathbf{F} - \mathbf{S}_{(-x,-y)}\right) \quad (3)$$

where $\otimes$ is the elementwise multiplication and $\mathbf{S}_{(x,y)}$ denotes the cyclically shifted $\mathbf{F}$ in the direction (x, y) with a stride d. The depthwise cyclic shift scheme is shown in Fig. 2. Each arrow inside denotes the transformation from the original feature map to the shifted feature map. The shifted feature map in the opposite direction can be obtained via reversing the cyclic shift. With this cyclic shift trick, we can reduce the computational cost to only $8HW$ subtractions for local difference measure on each feature map. For instance, with this trick, MPCM can be computed around 15% faster, from 2.67 to 3.07 FPS. Then, the local contrast feature map under dilation rate d [dilated local contrast (DLC)] can be obtained via elementwise maximum operation as

$$\mathrm{DLC}(\mathbf{F}, d) = \max_{(x,y)\in\Omega}\left\{\mathbf{D}^{(x,y)}\right\}. \quad (4)$$

Discussion: To some extent, 3 is a special formulation of the attention mechanism with clear physical interpretability, in which the features that match the local contrast prior are

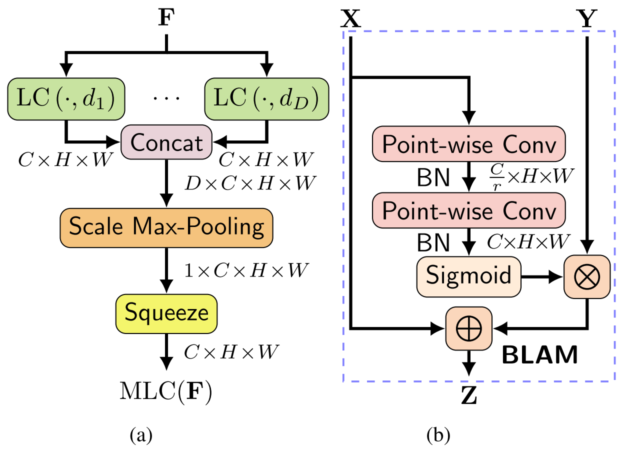


Fig. 3. Illustration of the proposed modules. (a) Same-layer MLC module, which embeds the local contrast prior into networks. (b) Cross-layer BLAM module, which embeds smaller scale details into high-level coarse feature maps.

emphasized, while the rest features are suppressed. Under a large dilation rate d, the local contrast measure explicitly breaks the limited effective receptive field [35] and encodes relatively long-range contextual interactions on feature maps.

C. Multiscale Local Contrast Measure

A central issue in infrared small target detection is the scale variation of targets. Both traditional local contrast methods and state-of-the-art convolutional network architectures share the same motivation that models should have different patch sizes or receptive fields for targets of different scales. To remedy this issue, an intuitive way is to leverage an MLC measure on feature maps. Given an intermediate feature map $\mathbf{F}$, we measure MLC in the same layer $\text{MLC}(\mathbf{F}) \in \mathbb{R}^{C\times H\times W}$ by applying the DLC module DLC with various dilation rates $\{d_1, d_2, \ldots, d_D\}$ as

$$\text{MLC}(\mathbf{F}) = \text{Squeeze}(\text{SMP}(\text{Concat}\\ (\text{DLC}(\mathbf{F}, d_1), \ldots, \text{DLC}(\mathbf{F}, d_D)))) \quad (5)$$

where SMP denotes the scale max-pooling operation that max pools the concatenated feature tensor along the scale axis, and the Squeeze operation removes single-dimensional entries from the shape of the feature map. The flowchart of 5 is shown in Fig. 3(a). It is noteworthy that in traditional local contrast methods, it is a risky operation of taking the maximum value among different scales as the target contrast that may cause false alarms since some background distractors are possible to have a higher contrast measure than the real targets under a preset oversimplified feature representation. However, this problem is largely mitigated in our ALCNet because the network can learn to adjust the feature representation dynamically in an end-to-end way according to the labeled data and loss function, thereby making sure that the real target has maximum local contrast.

IV. Attentional Local Contrast Network

In this section, we describe the overall architecture and final optimization formulation of the proposed method. Specially, we will deal with the challenges: 1) how to highlight the features of infrared small targets in high-level layers of a coarse representation of the input image and 2) how to handle the class imbalance issue between small targets and the background.

A. Bottom-Up Local Attentional Modulation

A deeper network can provide better semantic features and an understanding of the context of the scene, which helps to resolve the ambiguity between the target and the background distractors. However, as the network deepens, the risk of losing the target spatial details is also increasing. To tackle this issue, our ALCNet resorts to a cross-layer bottom-up local attentional modulation (BLAM) module to embed low-level information into high-level coarse feature maps, as shown in Fig. 3(b). The local channel attention mechanism $\mathbf{L}$ locally aggregates channel feature context for each spatial position individually as

$$\mathbf{L}(\mathbf{X}) = \sigma(\mathcal{B}(\text{PWConv}_2(\delta(\mathcal{B}(\text{PWConv}_1(\mathbf{X})))))) \quad (6)$$

where PWConv, σ, δ, and $\mathcal{B}$ denote the point-wise convolution [36], Sigmoid function, rectified linear unit (ReLU) [37], and batch normalization (BN) [38], respectively. The kernel sizes of PWConv_1 and PWConv_2 are $(C/4) \times C \times 1 \times 1$ and $C \times (C/4) \times 1 \times 1$, respectively, consisting of a bottleneck structure. It is noteworthy that the attentional weight map $\mathbf{L}(\mathbf{X}) \in \mathbb{R}^{C\times H\times W}$ has the same shape as the input feature maps and can thus be used to highlight the subtle details in an element-wise manner—spatially and across channels. Therefore, the BLAM module is able to be dynamically aware of the subtle details of small infrared targets.

The motivation of the BLAM module is to embed smaller scale details into high-level coarse feature maps, which is achieved as a dynamically weighted modulation of the high-level features under the guidance of low-level features. Given $\mathbf{X}$ as the low-level feature and $\mathbf{Y}$ as the high-level feature, the cross-layer fused feature $\mathbf{Z}' \in \mathbb{R}^{C\times H\times W}$ can be obtained via the bottom-up local attentional module as

$$\mathbf{Z} = \mathbf{X} + \mathbf{L}(\mathbf{X}) \otimes \mathbf{Y} \quad (7)$$

where $\otimes$ denotes the elementwise multiplication. In the context of MLC measure, by replacing $\mathbf{X}$ and $\mathbf{Y}$ with $\text{MLC}(\mathbf{X})$ and $\text{MLC}(\mathbf{Y})$, we can obtain the fused MLC feature map $\mathbf{Z}' \in \mathbb{R}^{C\times H\times W}$ via

$$\begin{aligned}\mathbf{Z}' &= \text{MLC}(\mathbf{X}) \uplus \text{MLC}(\mathbf{Y}) \\ &= \text{MLC}(\mathbf{X}) + \mathbf{L}(\text{MLC}(\mathbf{X})) \otimes \text{MLC}(\mathbf{Y}) \quad (8)\end{aligned}$$

where $\uplus$ denotes the cross-layer feature fusion via BLAM.

B. Network Architecture

A high-resolution prediction map is vital for detecting infrared small targets due to their small sizes. To preserve the small target, we use a modified ResNet-20 [39] as the backbone network to extract the feature maps. To unlike conventional approaches that downsample the image 32 times, the backbone in our ALCNet only subsamples the image

TABLE II

BACKBONE ARCHITECTURE. WE SCALE THE MODEL BY DEPTH (THE BLOCK NUMBER b IN EACH STAGE) TO STUDY THE RELATIONSHIP BETWEEN THE PERFORMANCE AND THE NETWORK DEPTH. WHEN $b = 3$, IT IS THE STANDARD RESNET-20 BACKBONE [39]

Stage	Output	ResBlock
Conv-1	480 × 480	3×3 conv, 16
Stage-1	480 × 480	$\begin{bmatrix} 3 \times 3 \text{ conv}, 16 \\ 3 \times 3 \text{ conv}, 16 \end{bmatrix} \times b$
Stage-2	240 × 240	$\begin{bmatrix} 3 \times 3 \text{ conv}, 32 \\ 3 \times 3 \text{ conv}, 32 \end{bmatrix} \times b$
Stage-3	120 × 120	$\begin{bmatrix} 3 \times 3 \text{ conv}, 64 \\ 3 \times 3 \text{ conv}, 64 \end{bmatrix} \times b$

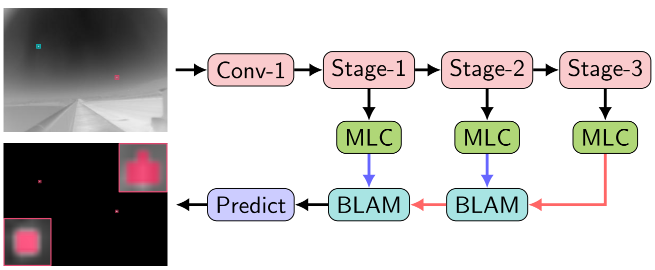


Fig. 4. Architectures of the proposed ALCNet, which incorporates same-layer MLC modules and cross-layer BLAM modules into an FPN. The blue line and red line represent the channel number transformation and the upsampling operator, respectively.

twice at stage2_1 and stage3_1 with a stride of 2, as shown in Table II. To stack images of different sizes into a batch, each image is resized to 512×512 and randomly cropped to 480×480 during training.

With the BLAM module, the MLC feature maps of different stages can be further fused in a smart way, which recovers the full spatial resolution at the network output by iteratively fusing coarse, high-layer feature maps with fine, low-layer feature maps as

$$\mathrm{M^2LC}(f) = \uplus\big(\mathrm{MLC}(\mathbf{F}^{(1)}), \uplus(\cdots, \uplus\big(\mathrm{MLC}(\mathbf{F}^{(L-1)}), \mathrm{MLC}(\mathbf{F}^{(L)})\big)\big)\big) \quad (9)$$

where $\mathrm{M^2LC}(f) \in \mathbb{R}^{C\times H\times W}$ is the final local contrast feature maps given an infrared image f and $\uplus$ denotes the BLAM module. It should be noted that for the sake of simplicity, 9 omits 1×1 convolutions that are used to adjust the number of filters in the proposed ALCNet. Finally, the two-stage MLC feature maps $\mathrm{M^2LC}(f)$ are used to predict small infrared targets. The whole proposed network that conducts two-stage MLC measurement is proposed, as shown in Fig. 4.

C. Problem Formulation and Optimization

To handle the class imbalance issue between infrared small target and the background, we adopt the Soft-IoU loss function [40] for this highly unbalanced segmentation task, which is defined as follows:

$$\ell_{\text{soft-IoU}}(p, y) = \frac{\sum_{i,j} p_{i,j} \cdot y_{i,j}}{\sum_{i,j} p_{i,j} + y_{i,j} - p_{i,j} \cdot y_{i,j}} \quad (10)$$

where $p = \sigma(\mathrm{M^2LC}(f, \Theta)) \in \mathbb{R}^{H\times W}$ is the prediction score map and $y \in \mathbb{R}^{H\times W}$ is the labeled mask, given an infrared image f. Θ denotes the weights of the proposed ALCNet. Given N training samples, Θ will be learned during training by minimizing the total loss as

$$\Theta = \arg\min_{\Theta} \sum_{n=1}^{N} \ell_{\text{soft-IoU}}(\sigma(\mathrm{M^2LC}(f, \Theta))_n, y_n). \quad (11)$$

For optimization, we adopt AdaGrad [41] as an optimizer with a learning rate of 0.1 and the strategy described by He *et al.* [42] for weight initialization, a total of 400 epochs, a weight decay of 10^{-4}, and a batch size of 10.

V. EXPERIMENTS

To analyze the potential of the proposed ALCNet, we compare it with state-of-the-art baselines on the public SIRST data set. We also conduct a comprehensive ablation study to investigate the effectiveness of the design of the ALCNet and its behavior under different parameter and computational budgets. In particular, the following questions will be investigated in our experimental evaluation.

1) *Q1:* Our key insight is to embed the local contrast measure into convolution networks as a parameterless nonlinear feature extraction layer. Given the same parameter budget, we investigate the question of how the proposed DLC module helps learn more discriminative features for infrared small targets (see Section V-B1).
2) *Q2:* From another perspective, the proposed ALCNet can be viewed as a learnable and dynamic MLC feature measure. We will examine the impact of MLC feature fusion in the same layer and across layers, respectively, as well as the importance of BLAM (see Section V-B2).
3) *Q3:* Generally, multiscale feature fusion is done by integrating high-level semantic information into low-level features in a top-down manner, but our ALCNet utilizes a reverse BLAM instead. In our study (see Section V-B3), we investigate the question of how important the bottom-up modulation and the local feature context are for infrared small targets.
4) *Q4:* Finally, we will analyze how the proposed ALCNet compares to other state-of-the-art model- or data-driven methods (see Section V-C).

A. Experimental Settings

Before investigating the questions Q1–Q4, we first introduce our experimental settings in detail.

1) Data set: For experimental evaluation, we resort to our public SIRST data set, which is the largest open data set for single-frame infrared small target detection to the best of our knowledge. It contains 427 representative images and 480 instances of different scenarios from hundreds of real-world videos and is roughly split into approximately 50%

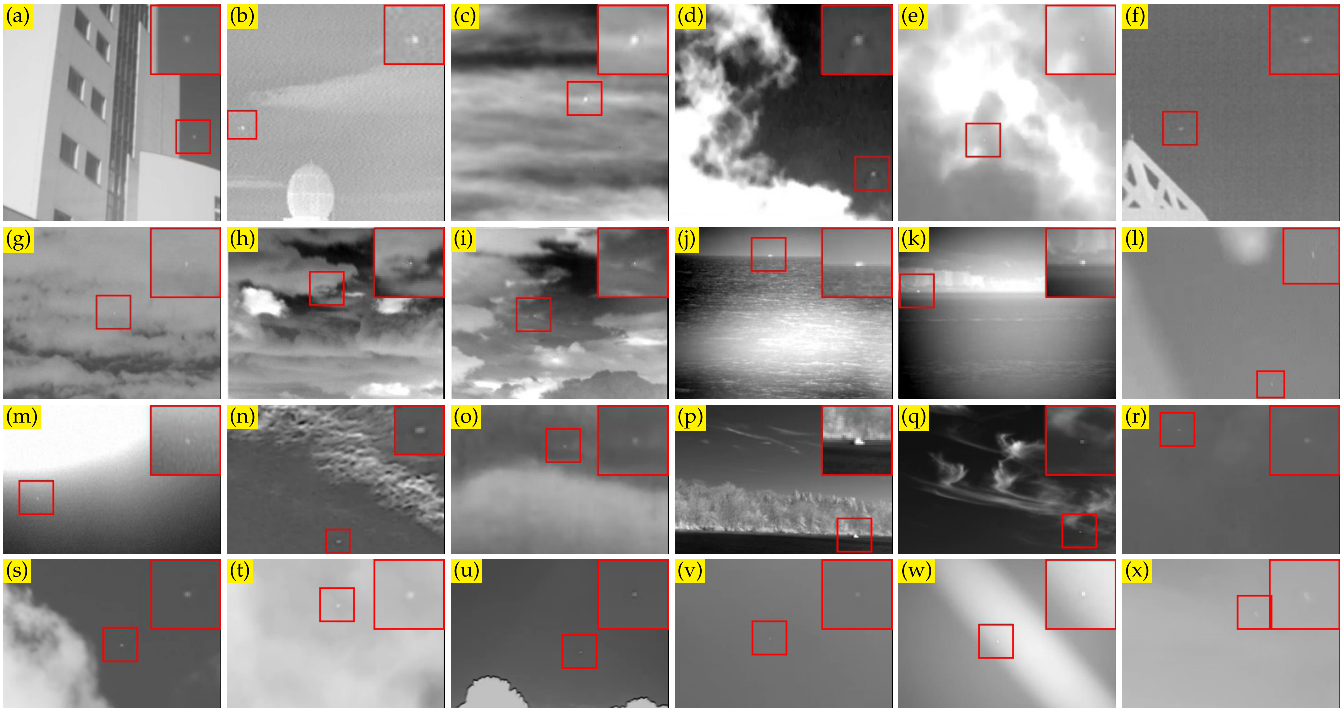


Fig. 5. Representative infrared images from the SIRST data set with various backgrounds, which excludes many trivial cases. For better visualization, the demarcated area is enlarged, which is better to be seen by zooming on a computer screen. (a)–(x) Collected infrared small target images.

train, 20% validation, and 30% test. Fig. 5 shows some representative images from the SIRST data set, from which we can see that many infrared small targets are extremely dim and buried in complex backgrounds with heavy clutter. In addition, only 35% targets in the data set contain the brightest pixel in the image. Therefore, the methods purely based on the target saliency assumption or merely thresholding on the raw image would not work well.

2) Implementation Details: For data-driven methods, feature pyramid networks (FPNs) [28], selective kernel networks [43] style FPN (SK-FPN), FPN with global attention upsampling (GAU-FPN) [29], and TBC-Net [18] are selected for comparison. These methods share the same loss function, optimizer, and other hyperparameters as the proposed ALCNet. For nonlearning model-driven methods, we choose stable multisubspace learning (SMSL) [19], facet kernel and random walker (FKRW) [44], MPCM [20], infrared patch-image model (IPI) [4], nonnegative IPI model via partial sum minimization of singular values (NIPPS) [10], and reweighted infrared patch-tensor model (RIPT) [11] for comparison. Their detailed hyperparameter settings are listed in Table III, which are determined by an exhaustive search on the trainval set of the SIRST data set.

3) Evaluation Metrics: Besides the intersection over union (IoU) metric and receiver operating characteristic (ROC) curve, we also choose the normalized IoU (nIoU) to evaluate the proposed ALCNet. nIoU is specially designed for the SIRST data set as a more balanced metric between model- and data-driven methods, which is defined as

$$\mathrm{nIoU} = \frac{1}{N}\sum_{i}^{N}\frac{\mathrm{TP}[i]}{\mathrm{T}[i]+\mathrm{P}[i]-\mathrm{TP}[i]} \tag{12}$$

where TP, T, and P denote the true positive, true, and positive, respectively. Unlike conventional filtering or target-background separation approaches, the proposed ALCNet outputs binary decisions. Therefore, traditional evaluation metrics for background suppression, including local signal-to-noise ratio gain, background suppression factor, and signal-to-clutter ratio gain, are not suitable here.

B. Ablation Study

We start by investigating the questions Q1–Q3 raised above. To better understand the proposed ALCNet, we consider several competitors constructed by removing or replacing specific parts of ALCNet. In Table IV, we illustrate these ablation study architectures according to their first-stage same-layer and second-stage cross-layer feature extraction schemes. "Plain" means that no local contrast module is used. The bottom-up global attention module (BGAM) and top-down local attention module (TLAM) are shown in Fig. 6(a) and (b).

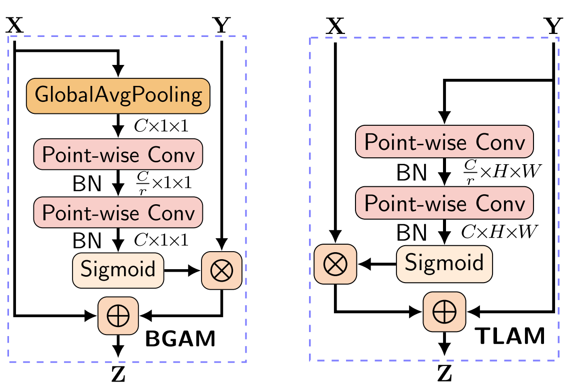
1) Impact of Local Contrast Prior (Q1): We start by comparing the FPN and DLC-FPN that embeds a DLC module before cross-layer fusion. Note that, in FPN, we add a 3×3 convolution as postprocessing to improve its performance, whereas DLC-FPN does not have such postprocessing, and the DLC module does not induce additional parameters. Therefore, the FPN has a bit more parameters than DLC-FPN when b is the same. Fig. 7 presents their comparison on IoU and nIoU given a gradually increased network depth. The dilation rate of DLC-FPN is 13. It can be seen that compared with FPN, the performance of the DLC-FPN is consistently and significantly better. Especially, the performance of DLC-FPN when $b = 3$ is approximately the same as FPN ($b = 4$). This result suggests that incorporating the local contrast prior

TABLE III
DETAILED HYPERPARAMETER SETTINGS OF MODEL-DRIVEN METHODS FOR COMPARISON

Methods	Hyper-parameter settings
MPCM [20]	$N = 1, 3, ..., 9$
FKRW [44]	$K = 4$, $p = 6$, $\beta = 200$, window size:11×11
SMSL [19]	Patch size: 50×50, $\lambda = \frac{2\times L}{\sqrt{\min(m,n)}}$, $L = 2.0$, threshold factor: $k = 1$
IPI [4]	Patch size: 50×50, stride: 10, $\lambda = L/\min(\mathrm{m}, \mathrm{n})^{1/2}$,$L = 4.5$, threshold factor: $k = 10$, $\varepsilon = 10^{-7}$
NIPPS [10]	Patch size: 50×50, stride: 10, $\lambda = \frac{L}{\sqrt{\min(m,n)}}$, $L = 2.0$, energy constraint ratio: $r = 0.11$, threshold factor: $k = 10$
RIPT [11]	Patch size: 50×50, stride: 10, $\lambda = \frac{L}{\sqrt{\min(I,J,P)}}$, $L = 0.001$, $h = 0.1$, ϵ=0.01, $\varepsilon = 10^{-7}$, threshold factor:$k = 10$

TABLE IV
ILLUSTRATION OF ARCHITECTURES OF THEIR DIFFERENT SAME-LAYER MLC FEATURE EXTRACTION AND CROSS-LAYER FEATURE FUSION SCHEMES IN THE ABLATION STUDY

Same-Layer	Cross-Layer	Architecture
Plain	None	PlainFCN
Plain	Skip Connection via Addition	FPN
DLC	Skip Connection via Addition	DLC-FPN
MLC	Skip Connection via Addition	MLC-FPN
MLC	Skip Connection via Maximum	Max-FPN
MLC	Top-down Local Attention	TLA-FPN
MLC	Bottom-up Global Attention	BGA-FPN
MLC	Bottom-up Local Attention	ALCNet (ours)



(a) (b)

Fig. 6. Architectures for the ablation study. (a) BGAM. (b) TLAM.

into deep networks helps alleviate the minimal intrinsic feature issue.

This performance gain can be explained from two perspectives. On the one hand, the local contrast module transforms the network from appearance-based recognition to local contrast-based recognition, which helps suppress more background clutter with the domain knowledge. On the other hand, it can also be viewed as a particular spatial attention module with clear physical interpretability, which encodes relatively long-range contextual interactions in a depthwise manner.

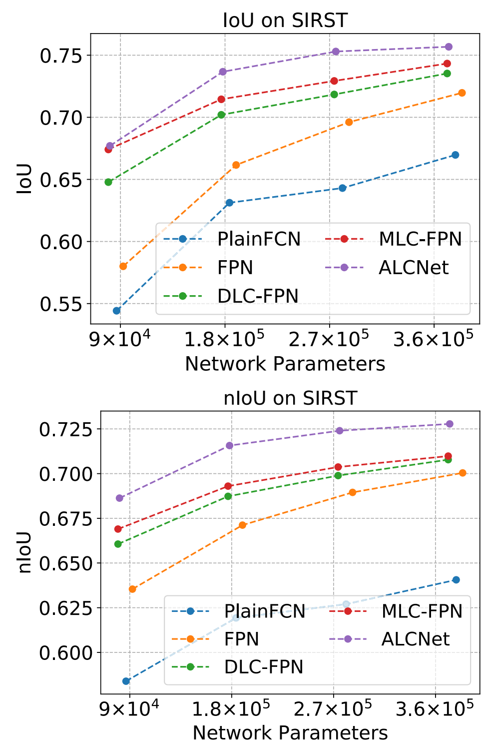


Fig. 7. Performance comparison of ablation architectures. The comparison among FPN, DLC-FPN, and MLC-FPN suggests that incorporating local contrast prior helps and a multiscale measure in the same layer can further boost the performance. The results comparing PlainFCN and FPN suggest that cross-layer feature fusion is of vital importance. Furthermore, the results comparing MLC-FPN and ALC-Net suggest that one should pay more attention to the cross-layer feature fusion, and a more sophisticated scheme yields consistent performance gains.

2) Impact of MLC Integration (Q2): Next, we investigate the importance of the MLC measure in our ALCNet. In Fig. 8, we provide the performance of DLC-FPN with various dilation rates and network depth and comparison with FPN ($b = 4$). It can be seen that, just like its nonlearning counterparts, our DLC network is sensitive to the dilation rate, which is a hyperparameter. Generally, DLC-FPN performs better than FPN. However, with a lousy dilation rate, DLC-FPN is possible to perform worse than FPN, especially on the nIoU metric.

To solve this issue, we adopt the MLC measure on the same layer feature maps. Fig. 7 presents the comparison between single-scale DLC-FPN and multiscale MLC-FPN given a gradually increased network depth. The dilation rates of MLC-FPN are 13 and 17. It can be seen that by covering multiple dilation rates, the performance of MLC-FPN is consistently better than DLC-FPN. The MLC measure enables the network to aggregate multiscale spatial information in the same layer, which greatly improves the robustness against the target scale variation. It should be noted that the gap between DLC-FPN and MLC-FPN is not so significant because we select the best dilation rate for DLC-FPN.

The importance of the cross-layer local contrast feature fusion is also shown in Fig. 7. PlainFCN that did not fuse the feature maps of different layers via skip connections is significantly worse. The results suggest that cross-layer feature integration is of vital importance for infrared small targets. Furthermore, ALCNet that replaces the simple elementwise addition with our specially designed BLAM module performs consistently better than the rest of competitors. Especially, compared with the second-best MLC-FPN ($b = 4$), our ALCNet can perform similarly on the IoU metric and even better on the nIoU metric with only around 50% parameters. The results suggest that one should pay attention to the cross-layer feature fusion for infrared small targets and that more sophisticated fusion mechanisms hold the potential to consistently yield better results. We believe that the reason behind the performance gain brought by the cross-layer feature fusion is that the low-level features offer detailed information for the accurate localization of small targets, and the high-level features help solve the ambiguous cases with their semantic information.

3) Impact of Cross-Layer Fusion Manners (Q3): Next, we also investigate and compare our BLAM with two other ablation modules. The first one is the BGAM, which aggregates the global contextual information by adding a global averaging pooling at the beginning of the local channel attention module, as shown in Fig. 6(a). The second one that reversed the ALCNet's modulation direction from bottom-up to top-down is the TLAM module shown in Fig. 6(b). Table V shows the results, from which it can be seen that the following conditions hold.

1) The performance of BGAM-FPN is not as good as TLAM-FPN and ALCNet, which suggests that for infrared small targets, given the same additional budget for parameters and computation costs, one should aggregate the fine local information as guidance instead of the global contextual information.
2) The difference between TLAM-FPN and ALCNet is that TLAM-FPN embeds the high-level semantic information into low-level features in a top-down manner. In contrast, the proposed ALCNet reverses this modulation direction by using the low-level feature as guidance to refine the high-level feature maps.

The results are strong support for designing BLAM pathways for infrared small targets. It is because unlike the semantic segmentation task in generic vision data sets, the localization error occupies most of the overall error. Therefore, compared with top-down semantic guidance, bottom-up detailed information is more helpful for accurate segmentation.

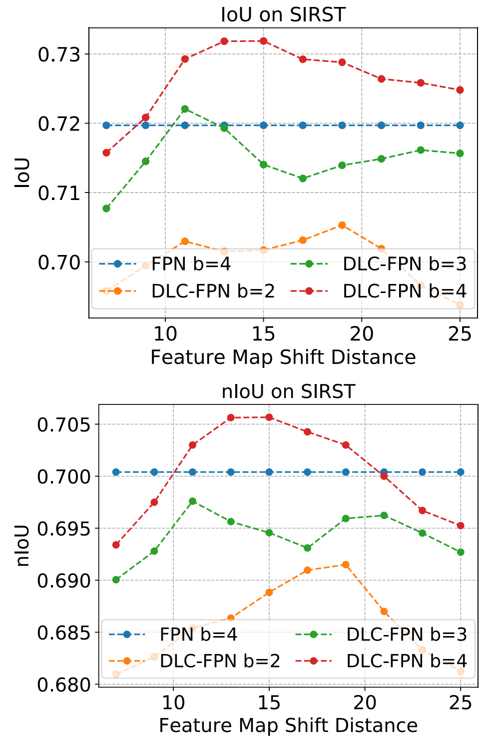


Fig. 8. IoU and nIoU performance of DLC-FPN with varying dilation rates and network depths, which suggests that a single-scale local contrast measure is sensitive to the dilated value.

C. Comparison to State-of-the-Art Approaches

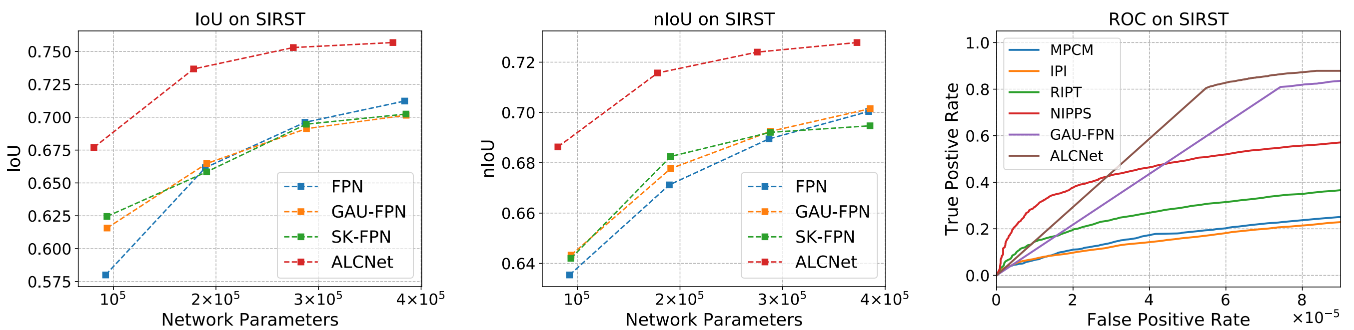
Finally, we address question Q4 by comparing our ALCNet with several model-driven methods and other state-of-the-art network competitors. First, we compare the proposed ALCNet with other deep convolutional networks, namely, FPN [28], SK-FPN [43], and GAU-FPN [29], on the SIRST data set, given a gradual increase of the depths of the networks. The results are shown in Fig. 9(a) and (b). It can be seen that the following holds. First, the proposed ALCNet achieves a significantly better performance for all experimental settings, which demonstrates its effectiveness compared with other baselines. These results reaffirm that one can obtain better infrared small detection performance by incorporating local contrast domain knowledge and by dynamically modulating the high-level features with guidance from low-level feature maps. Second, even with a half parameter number ($b = 2$), ALCNet still performs better than these baseline networks with $b = 4$. The results suggest that by utilizing the local contrast prior and paying attention to the cross-layer feature fusion, one can obtain a more efficient convolutional network that yields better performance with fewer layers or parameters per network.

Next, we compare our proposed ALCNet with the baseline networks and other state-of-the-art nonlearning model-driven methods on the IoU and nIoU metrics in Table VI as well as the computational time. We also validate our ALCNet on the

TABLE V

COMPARISON OF DIFFERENT CROSS-LAYER FEATURE FUSION SCHEMES ON THE SIRST DATA SET. THE RESULTS COMPARING BGA-FPN AND ALCNET SUGGEST THAT THE LOCAL CONTEXTUAL AGGREGATION IS VITAL FOR INFRARED SMALL TARGETS. THE COMPARISON BETWEEN TLA-FPN AND ALCNET SUGGESTS THAT GIVEN THE SAME COMPUTATIONAL AND PARAMETER BUDGET, THE BOTTOM-UP MODULATION PERFORMS BETTER THAN THE TOP-DOWN ONE

Contextual Scale	Modulation Direction	Formulation	Architecture	IoU				nIoU			
				$b=1$	$b=2$	$b=3$	$b=4$	$b=1$	$b=2$	$b=3$	$b=4$
None	None	$\mathbf{X}+\mathbf{Y}$	FPN	0.674	0.713	0.729	0.744	0.669	0.691	0.702	0.710
		$\max(\mathbf{X},\mathbf{Y})$	Max-FPN	0.665	0.713	0.722	0.734	0.674	0.698	0.706	0.712
Global	Bottom-Up	$\mathbf{X}+\mathbf{G}(\mathbf{X})\otimes\mathbf{Y}$	BGA-FPN	0.676	0.714	0.731	0.736	0.679	0.698	0.704	0.711
Local	Top-Down	$\mathbf{L}(\mathbf{X})\otimes\mathbf{X}+\mathbf{Y}$	TLA-FPN	**0.688**	0.729	0.750	0.753	**0.688**	0.708	0.722	0.718
	Bottom-Up	$\mathbf{X}+\mathbf{L}(\mathbf{X})\otimes\mathbf{Y}$	ALCNet	0.677	**0.737**	**0.753**	**0.757**	0.686	**0.716**	**0.724**	**0.728**



(a) (b) (c)

Fig. 9. Predictive performance comparison with other state-of-the-art methods. (a) and (b) Comparison with other state-of-the-art baseline networks on IoU and nIoU with a gradual increase of network depth. (c) Comparison with both data- and model-driven methods on ROC. The proposed ALCNet consistently yields the best performance.

ROC metric. The results are shown in Fig. 9(c). It can be seen that the following conditions hold.

1) All convolutional networks perform better than the nonlearning model-driven methods, which shows that learning from the data offers a promising way leading to better performance for infrared small detection.
2) Our ALCNet achieves the best among learning and nonlearning approaches, showing the effectiveness of the proposed architecture. Compared with GAU-FPN, the proposed ALCNet achieves a higher detection rate and a lower false-alarm rate simultaneously.
3) It should be noted that to keep the comparison fair, the average inference time per sample of every method, including both model-driven methods and deep networks, is evaluated as each sample at a time with the same CPU.

It can be seen that once finishing training, the inference of deep networks is much faster than conventional saliency detection or low-rank plus sparse decomposition methods. Although our ALCNet is around ten times slower than the purely data-driven FPN due to the proposed local contrast measure layer, it is still among the fastest methods compared with model-driven methods. However, considering the performance boost brought by the proposed method, we think it is a good tradeoff between the performance and inference time. In addition, our network can be further accelerated with GPUs.

D. Error Diagnosis and Limitations

In this section, we analyze the reasons for false positives and false negatives as well as their impact on the detection performance. All the prediction results of the test set are available online.[3] The images on which the proposed ALCNet cannot perform very well are shown in Fig. 10. Actually, the overall performance of the proposed ALCNet is quite good, only having two miss detections, namely, Fig. 10(a) and (b). It can be seen that the majority of segmentation errors stems from the target boundary, either exceeding the labeled mask by a few pixels or segmenting the target incompletely. It should be noted that the human label is ambiguous or inaccurate in terms of one or two pixels' shift, which has a large impact on our final IoU and nIoU metrics. For example, for a minimum target of 2×2 pixels, even one pixel's shift leads the label to 3×3, which induces about 50% errors for this target. Actually, such boundary error also exists in generic vision tasks. However, due to the large object size, the impact is not as prominent as our infrared small target detection task. However, as long as the segmented pixels are located in the same connected domain, these boundary errors will not cause true false alarms or missed detections. Furthermore, the false positives caused by incomplete detection like Fig. 10(c), (d), and (f), in which one target is predicted as two close but

[3] https://github.com/YimianDai/open-alcnet/tree/master/results/pred

TABLE VI
COMPARISON WITH OTHER STATE-OF-THE-ART METHODS ON IoU AND nIoU

Metric	SMSL	FKRW	MPCM	IPI	NIPPS	RIPT	FPN	SK-FPN	GAU-FPN	TBC-Net	ALCNet
IoU	0.081	0.268	0.357	0.466	0.473	0.146	0.720	0.702	0.701	0.734	**0.757**
nIoU	0.279	0.339	0.445	0.607	0.602	0.245	0.700	0.695	0.702	0.713	**0.728**
Time on CPU/s	0.595	0.399	0.347	11.699	5.707	6.398	0.031	0.035	0.033	0.049	0.378

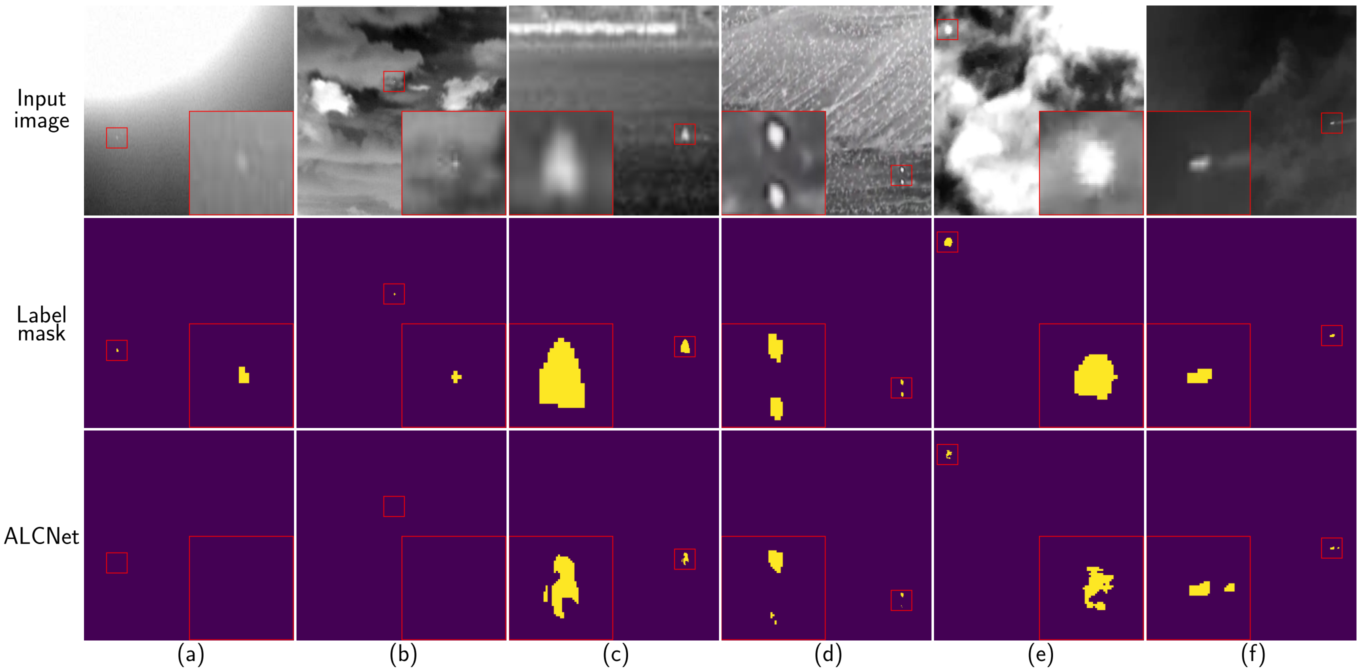


Fig. 10. Illustrations of the images on which the proposed ALCNet cannot perform very well. (a)–(f) Corresponding input image, label mask, and failed results produced by ALCNet.

separated regions, can be alleviated by simple morphological dilation operations under the target sparsity prior. The real issue is the miss detections. From Fig. 10(a) and (b), it can be seen that the main reason is that the infrared small target is too dim. Besides, the small size of the target also causes its weight to be very small in the loss function, which is easily overwhelmed by the boundary error of larger targets during training.

It should be noted that in this work, we follow the convention of this field and model infrared small target detection as a segmentation problem. Actually, we start our work as a bounding box regression problem just like the conventional object detection task in generic computer vision. However, due to the small size, the IoU threshold has to be much lower than the default value in generic vision data sets, which leads to a severe duplicate detection problem as another form of the localization error. Therefore, it remains an open question that how to represent the infrared small target so that the evaluation metric can reveal the true detection performance and not be affected by the boundary error in any means.

VI. CONCLUSION

We have presented our ALCNet for infrared small target detection. In particular, we find that convolutional networks that integrate the local contrast prior, which is generally modeled in nonlearning methods, are very promising and worthy of further research. By breaking the conventional nonoverlapped patch constraint, we extract and fuse the local contrast feature maps in two stages, namely in the same layer and across layers, to better transplant the domain knowledge into networks. Besides, to highlight and preserve the small target in high-level coarse features, we utilize a BLAM module that embeds subtle low-level details into high-level layers. We conduct extensive ablation studies and comparison with other state-of-the-art methods. The proposed ALCNet significantly outperforms the compared purely model-driven methods and purely data-driven networks on the open SIRST data set, which suggests that one should pay attention to combining deep networks with domain knowledge to detect small infrared targets and a target-preserving cross-layer fusion scheme holds the potential to yield better results.

REFERENCES

[1] Z. Wu, N. Fuller, D. Theriault, and M. Betke, "A thermal infrared video benchmark for visual analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2014, pp. 201–208.
[2] W. Zhang, M. Cong, and L. Wang, "Algorithms for optical weak small targets detection and tracking: Review," in *Proc. Int. Conf. Neural Netw. Signal Process.*, vol. 1, Dec. 2003, pp. 643–647.

[3] I. S. Reed, R. M. Gagliardi, and L. B. Stotts, "Optical moving target detection with 3-D matched filtering," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 24, no. 4, pp. 327–336, Jul. 1988.

[4] C. Gao, D. Meng, Y. Yang, Y. Wang, X. Zhou, and A. G. Hauptmann, "Infrared patch-image model for small target detection in a single image," *IEEE Trans. Image Process.*, vol. 22, no. 12, pp. 4996–5009, Dec. 2013.

[5] G. Arce and M. McLoughlin, "Theoretical analysis of the max/median filter," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-35, no. 1, pp. 60–69, Jan. 1987.

[6] K. Mikolajczyk and C. Schmid, "Scale & affine invariant interest point detectors," *Int. J. Comput. Vis.*, vol. 60, no. 1, pp. 63–86, 2004.

[7] D. Liu, L. Cao, Z. Li, T. Liu, and P. Che, "Infrared small target detection based on flux density and direction diversity in gradient vector field," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 7, pp. 2528–2554, Jul. 2018.

[8] C. L. Philip Chen, H. Li, Y. Wei, T. Xia, and Y. Yan Tang, "A local contrast method for small infrared target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 1, pp. 574–581, Jan. 2014.

[9] H. Deng, X. Sun, M. Liu, C. Ye, and X. Zhou, "Small infrared target detection based on weighted local difference measure," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 7, pp. 4204–4214, Jul. 2016.

[10] Y. Dai, Y. Wu, Y. Song, and J. Guo, "Non-negative infrared patch-image model: Robust target-background separation via partial sum minimization of singular values," *Infr. Phys. Technol.*, vol. 81, pp. 182–194, Mar. 2017.

[11] Y. Dai and Y. Wu, "Reweighted infrared patch-tensor model with both nonlocal 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.

[12] Y. Zhao, H. Pan, C. Du, Y. Peng, and Y. Zheng, "Bilateral two-dimensional least mean square filter for infrared small target detection," *Infr. Phys. Technol.*, vol. 65, pp. 17–23, Jul. 2014.

[13] H. Zhu, S. Liu, L. Deng, Y. Li, and F. Xiao, "Infrared small target detection via low-rank tensor completion with top-hat regularization," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 2, pp. 1004–1016, Feb. 2020.

[14] X. Bai and Y. Bi, "Derivative entropy-based contrast measure for infrared small-target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 4, pp. 2452–2466, Apr. 2018.

[15] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 3431–3440.

[16] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent (MICCAI)*, Munich, Germany, 2015, pp. 234–241.

[17] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, Aug. 2013.

[18] 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*. [Online]. Available: http://arxiv.org/abs/2001.05852

[19] X. Wang, Z. Peng, D. Kong, and Y. He, "Infrared dim and small target detection based on stable multisubspace learning in heterogeneous scene," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 10, pp. 5481–5493, Oct. 2017.

[20] Y. Wei, X. You, and H. Li, "Multiscale patch-based contrast measure for small infrared target detection," *Pattern Recognit.*, vol. 58, pp. 216–226, Oct. 2016.

[21] J. Han, Y. Ma, B. Zhou, F. Fan, K. Liang, and Y. Fang, "A robust infrared small target detection algorithm based on human visual system," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 12, pp. 2168–2172, Dec. 2014.

[22] P. Hu and D. Ramanan, "Finding tiny faces," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1522–1530.

[23] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Z. Li, "S3FD: Single shot scale-invariant face detector," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 192–201.

[24] B. Singh, M. Najibi, and L. S. Davis, "SNIPER: Efficient multi-scale training," in *Proc. Annu. Conf. Neural Inf. Process. Syst. (NeurIPS)*, Montreal, QC, Canada, 2018, pp. 9333–9343.

[25] H. Zhang *et al.*, "Context encoding for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7151–7160.

[26] A. Shrivastava and A. Gupta, "Contextual priming and feedback for faster R-CNN," in *Proc. 14th Eur. Conf. Comput. Vis. (ECCV)*, Amsterdam, The Netherlands, 2016, pp. 330–348.

[27] B. Singh and L. S. Davis, "An analysis of scale invariance in object detection—SNIP," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, Jun. 2018, pp. 3578–3587.

[28] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 936–944.

[29] H. Li, P. Xiong, J. An, and L. Wang, "Pyramid attention network for semantic segmentation," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, Newcastle, U.K., 2018, pp. 1–13.

[30] W. Yuan, S. Wang, X. Li, M. Unoki, and W. Wang, "A skip attention mechanism for monaural singing voice separation," *IEEE Signal Process. Lett.*, vol. 26, no. 10, pp. 1481–1485, Oct. 2019.

[31] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[32] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[33] J. Hu, L. Shen, S. Albanie, G. Sun, and A. Vedaldi, "Gather-excite: Exploiting feature context in convolutional neural networks," in *Proc. Annu. Conf. Neural Inf. Process. Syst. (NeurIPS)*, Montreal, QC, Canada, 2018, pp. 9423–9433.

[34] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *Proc. 4th Int. Conf. Learn. Represent. (ICLR)*, San Juan, Puerto Rico, 2016, pp. 1–10.

[35] W. Luo, Y. Li, R. Urtasun, and R. S. Zemel, "Understanding the effective receptive field in deep convolutional neural networks," in *Proc. Annu. Conf. Neural Inf. Process. Syst. (NeurIPS)*, Barcelona, Spain, 2016, pp. 4898–4906.

[36] M. Lin, Q. Chen, and S. Yan, "Network in network," in *Proc. 2nd Int. Conf. Learn. Represent. (ICLR)*, Banff, AB, Canada, 2014, pp. 1–10.

[37] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. 27th Int. Conf. Mach. Learn. (ICML)*, Haifa, Israel, 2010, pp. 807–814.

[38] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. 32nd Int. Conf. Mach. Learn. (ICML)*, Lille, France, 2015, pp. 448–456.

[39] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Proc. 14th Eur. Conf. Comput. Vis. (ECCV)*, Amsterdam, The Netherlands, 2016, pp. 630–645.

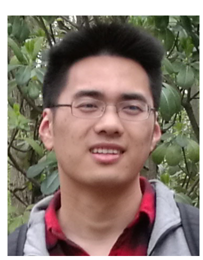
[40] M. A. Rahman and Y. Wang, "Optimizing intersection-over-union in deep neural networks for image segmentation," in *Proc. 12th Int. Symp. Vis. Comput. (ISVC)*, Las Vegas, NV, USA, 2016, pp. 234–244.

[41] J. C. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," in *Proc. Conf. Learn. Theory (COLT)*, 2010, pp. 257–269.

[42] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1026–1034.

[43] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 510–519.

[44] Y. Qin, L. Bruzzone, C. Gao, and B. Li, "Infrared small target detection based on facet kernel and random walker," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 7104–7118, Sep. 2019.

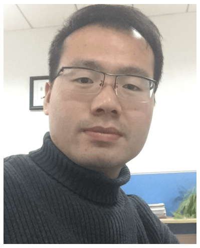


Yimian Dai (Student Member, IEEE) received the B.S. degree from the Nanjing University of Aeronautics and Astronautics (NUAA), Nanjing, China, in 2013, where he is pursuing the Ph.D. degree.

As a Visiting Student, he spent a half year at the Department of Computer Science, University of Copenhagen, Copenhagen, Denmark, and later two years at the Department of Computer Science, The University of Arizona, Tucson, AZ, USA. His research interests include sparse representation, deep learning, and their applications in image classification, target detection, and image understanding.

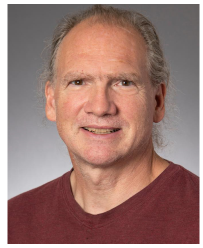
Yiquan Wu received the M.S. and Ph.D. degrees from the Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 1987 and 1998, respectively.

He is a Professor and a Ph.D. Supervisor with the Department of Information and Communication Engineering, Nanjing University of Aeronautics and Astronautics, where he is involved in teaching and research in the areas of image processing and recognition, target detection and tracking, and intelligent information processing.



Fei Zhou received the M.S. degree in electronic engineering from Xinjiang University, Ürümqi, China, in 2017. He is pursuing the Ph.D. degree with the College of Electronic and Information Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing, China.

His research interests include signal processing, target detection, and image processing.



Kobus Barnard (Member, IEEE) received the Ph.D. degree in computer science with specialization in computational color constancy from Simon Fraser University, Burnaby, BC, Canada, in 2000.

He then spent two years at the University of California at Berkeley, Berkeley, CA, USA, as a Post-Doctoral Researcher, working on modeling the joint statistics of images and associated text. He is a Professor with the Department of Computer Science, The University of Arizona, Tucson, AZ, USA. He also holds appointments in statistics, cognitive science, electrical and computer engineering, and the BIO5 Institute. He leads the Interdisciplinary Visual Intelligence Laboratory. His research interests include addressing interdisciplinary computational intelligence by developing top-down statistical models that are predictive, semantic, and explanatory. Application domains include computer vision, multimedia data, biological structure and processes, astronomy, and human social interaction. His work has been funded by multiple grants from the National Science Foundation, including a CAREER Award, and awards from the Defense Advanced Research Projects Agency, the Office of Naval Research, the Arizona Biomedical Commission, and the University of Arizona BIO5 Institute.

Dr. Barnard received the Governor General Gold Medal across all disciplines for his Ph.D. dissertation.