Asymmetric Contextual Modulation for Infrared Small Target Detection

Yimian Dai[1] Yiquan Wu[1] Fei Zhou[1] Kobus Barnard[2]

[1]College of Electronic and Information Engineering, Nanjing University of Aeronautics and Astronautics
[2]Department of Computer Science, University of Arizona

Abstract

Single-frame infrared small target detection remains a challenge not only due to the scarcity of intrinsic target characteristics but also because of lacking a public dataset. In this paper, we first contribute an open dataset with high-quality annotations to advance the research in this field. We also propose an asymmetric contextual modulation module specially designed for detecting infrared small targets. To better highlight small targets, besides a top-down global contextual feedback, we supplement a bottom-up modulation pathway based on point-wise channel attention for exchanging high-level semantics and subtle low-level details. We report ablation studies and comparisons to state-of-the-art methods, where we find that our approach performs significantly better. Our dataset and code are available online[1].

1. Introduction

Infrared small target detection is the key technique for applications including early warning systems, precision-guided weapons, and maritime surveillance systems. In many cases, the traditional assumptions of static backgrounds do not apply [17]. Therefore, researchers have started to pay more attention to the single-frame detection problem recently [10].

The prevalent idea from the signal processing community is to directly build models that measure the contrast between the infrared small target and its neighborhood context [2, 10]. By applying a threshold on the final saliency map, the potential targets are then segmented out. Despite being learning-free and computationally friendly, these *model-driven methods* suffer from the following shortcomings:

1. The target hypotheses of having global unique saliency, sparisty, or high contrast do not hold in real-world images. Real dim targets can be inconspicuous and low-contrast, whereas many background distractors satisfy these hypotheses, resulting in many false alarms.
2. Many hyper-parameters, such as λ in [10] and h in [4], are sensitive and highly relevant with the image content, which is not robust enough for highly variable scenes.

[1]https://github.com/YimianDai/open-acm

In short, these methods are handicapped because they lack a high-level understanding of the holistic scene, making them incapable to detect the extreme dim ones and remove salient distractors. Hence, it is necessary to embed high-level contextual semantics into models for better detection.

1.1. Motivation

It is well known that deep networks can provide high-level semantic features [12], and attention modules can further boost the representation power of CNNs by capturing long-range contextual interactions [9]. However, despite the great success of convolutional neural networks in object detection and segmentation [36], *very few deep learning approaches have been studied in the field of infrared small target detection*. We suggest the principal reasons are as follows:

1. **Lack of a public dataset so far**. Deep learning is data-hungry. However, until now, there is no public infrared small target dataset with high-quality annotations for the single-frame detection scenario, on which various new approaches can be trained, tested, and compared.
2. **Minimal intrinsic information**. SPIE defines the infrared small target as having a total spatial extent of less than 80 pixels (9×9) of a 256×256 image [34]. The lack of texture or shape characteristics makes purely target-centered representations inadequate for reliable detection. Especially, in deep networks, small targets can be easily overwhelmed by complex surroundings.
3. **Contradiction between resolution and semantics**. Infrared small targets are often submerged in complicated backgrounds with low signal-to-clutter ratios. For networks, detecting these dim targets with low false alarms needs both a high-level semantic understanding of the whole infrared image and a fine-resolution prediction map, which is an endogenous contradiction of deep networks since they learn more semantic representations by gradually attenuating the feature size [14].

In addition, these state-of-the-art networks are designed for generic image datasets [15, 19]. *Directly using them for infrared small target detection can fail catastrophically due to the large difference in the data distribution*. It requires a re-customization of the network in multiple aspects including

1. *re-customizing the down-sampling scheme:* Many studies emphasize that when designing CNNs, the receptive fields of predictors should match the object scale range [29, 20]. Without a re-customization of the down-sampling scheme, the feature of infrared small targets can hardly be preserved as the network goes deeper.
2. *re-customizing the attention module:* Existing attention modules tend to aggregate global or long-range contexts [15, 9]. The underlying assumption is that objects are relatively large and distribute more globally, which is consistent with objects in ImageNet [30]. However, this is not the case for infrared small targets, and a global attention module would weaken their features. This gives rise to the question of what kind of attention module is suitable for highlighting infrared small targets.
3. *re-customizing the feature fusion approach*: Recent works fuse cross-layer features in a one-directional, top-down manner [18, 32], aiming to select the right low-level features based on high-level semantics. However, since small targets may have already been overwhelmed by the background in deep layers, a pure top-down modulation may not work, even harmful.

Therefore, besides an annotated dataset and a re-adjustment on spatial down-sampling, it also needs a re-design of the attention module and feature fusion approach.

1.2. Contributions

To support *data-driven methods*, we first contribute an open dataset to advance the research of Single-frame InfraRed Small Target detection dubbed *SIRST*. Representative frames are selected from hundreds of infrared small target sequences and are manually labeled into five annotation forms, which enables the training of various machine learning approaches. To the best of our knowledge, SIRST is not only the first such public of this kind but also the largest ($4\times$ larger) compared with other private datasets [31]. Moreover, a new evaluation metric is also proposed to better balance the data-driven methods and traditional model-driven methods.

In this paper, we advocate the idea of mutually exchanging high-level semantics and low-level fine details for all level features as a solution for the issues arising from the scale mismatch between infrared small targets and objects in generic datasets. To this end, we propose an *asymmetric contextual modulation* (ACM) mechanism, a plug-in module that can be integrated into multiple host networks. Our approach supplements the state-of-the-art top-down high-level semantic feedback pathway with a reverse bottom-up contextual modulation pathway to encodes the smaller scale visual details into deeper layers, which we think is a key ingredient to achieve better performance for infrared small targets.

Moreover, this mutual modulation between high-level and low-level features is implemented in an asymmetric way, in which the top-down modulation is achieved by a conventional *global channel attention modulation* (GCAM) [18] to propagate high-level large scale semantic information down to shallow layers, whereas the bottom-up modulation is achieved by a *pixel-wise channel attention modulation* (PCAM) to preserve and highlight infrared small targets in high-level features. Our idea behind the proposed PCAM is that scale is not exclusive to spatial attention, and channel attention can also be achieved in multiple scales by varying the spatial pooling size. For infrared small targets, the proposed PCAM is a perfect fit for its small size.

By replacing the existing cross-layer feature fusion operations with the proposed ACM module, we can construct new networks that perform significantly better than the original host networks with only a modest number of additional parameters. Ablation studies on the impact of different modulation schemes show the effectiveness of the proposed ACM module. Experiments on the proposed SIRST dataset demonstrate that compared to other state-of-the-art methods, the networks based on the proposed ACM module achieves the best detection performance of infrared small targets.

2. Related Work

2.1. Single-Frame Infrared Small Target Detection

Due to the lack of a public dataset, most state-of-the-art methods in this field are still non-learning and heuristic methods highly dependent on target/background assumptions. Generally, most researchers model the single-frame detection problem as outlier detection under various assumptions, e.g., a salient outlier [3, 8], a sparse outlier in a low-rank background [5, 40], a pop-out outlier in smooth background [33, 7]. Then an outlierness map can be obtained via saliency detection, sparse and low-rank matrix/tensor decomposition, or local contrast measurements. Finally, the infrared small target is segmented out given a certain threshold. Although being computationally friendly and learning-free, these approaches suffer from the insufficient discriminability and hyper-parameter sensitivity to scene changing.

We notice that there are few deep learning-based infrared small target detection approaches [31, 39]. Our work differs in two important aspects: 1) We propose the ACM module for cross-layer feature fusion which is specially customized for infrared small targets. 2) We aim to build a benchmark for infrared small target detection, in which we not only offer a public dataset with high-quality annotations, but also a toolkit with implementations of state-of-the-art methods, customized evaluation metrics, and data augmentation tricks.

2.2. Cross-Layer Feature Fusion in Deep Networks

For accurate object localization and segmentation, state-of-the-art networks follow a coarse-to-fine strategy to hierarchically combine subtle features from lower layers and coarse semantic features from higher layers, e.g., U-Net [27]

and Feature Pyramid Networks (FPN) [22]. However, most works focus on constructing sophisticated pathways to bridge features across layers [12]. The feature fusion approach itself is generally achieved by simple linear approaches, either summation or concatenation, which can not provide networks with the ability to dynamically select the relevant features from lower layers. Recently, a few methods [18, 35] have been proposed to use high-level features as guidance to modulate the low-level features via the global channel attention module [15] in long skip connections.

Please note that the proposed ACM module follows the idea of cross-layer modulation, but differs in two important aspects: 1) Instead of a one-directional top-down pathway, our ACM module exchanges high-level semantics and fine details in two-directional top-down and bottom-up modulation pathways. 2) A point-wise channel attention module for the bottom-up modulation pathway is utilized to preserve and highlight the subtle details of infrared small targets.

2.3. Datasets for Infrared Small Targets

Unlike the computer vision tasks based on optical image datasets [28, 23], infrared small target detection is trapped by data scarcity for a long time due to many complicated reasons. Most algorithms are evaluated on private datasets consisting of very limited images [31], which is easy to make the performance comparison unfair and inaccurate. Some machine learning approaches utilize the sequence datasets like OSU Thermal Pedestrian [6] for training and test. However, objects in these datasets are not small targets, which not only do not meet the SPIE definition [34], but also are not in line with typical application scenarios of infrared small target detection. Besides, the sequential dataset is not appropriate for single-frame detection task, since the test set should not overlap with the training and validation sets.

In contrast, our proposed SIRST dataset is the first to explicitly build an open single-frame dataset by only selecting one representative image from a sequence. Moreover, these images are annotated with five different forms to support to model the detection task in different formulations. Limited by the difficulties in infrared data acquisition (mid-wavelength or short-wavelength), to the best of our knowledge, SIRST is not only the first public but also the largest compared to other private datasets [31].

3. SIRST: From Model-Driven to Data-Driven

Our motivation for contributing SIRST is to bridge the recent advance in data-driven deep learning and the field of infrared small target detection that is dominant by model-driven methods [40]. To this end, we present SIRST not only as a dataset but also as a toolkit of implementations of state-of-the-art methods and customized evaluation metrics.

3.1. Image Collection and Annotation

The proposed SIRST dataset contains 427 images including 480 instances, which is roughly split into 50% train, 20% validation, and 30% test. To avoid the overlap among training, validation, and test sets, we only select one representative image form each infrared sequence. Due to the scarcity of infrared sequences, besides short-wavelength and mid-wavelength infrared images, SIRST also includes infrared images of 950 nm wavelength. Fig. 1 shows some representative images, from which we can see that many targets are extremely dim and buried in complex backgrounds with heavy clutter. Even for humans, detecting them is not an easy task, which requires a high-level semantic understanding of the holistic scene and a concentrated search.

Unlike object detection in generic datasets, infrared small target detection is an outlier detection problem, which is a binary decision. Since the target is too small and lacks intrinsic characteristics, all of them are classified into one category without further distinguishing their specific classes. We provide the images with five kinds of annotations to support image classification, instance segmentation, bounding box regression, semantic segmentation, and instance spotting. The annotation pipeline is outlined in Fig. 2. Each target is confirmed by observing its moving in a sequence to make sure it is a real target, not pixel-wise pulse noise.

3.2. Dataset Statistics

The distribution of the target number per image is shown in Fig. 3(a). It shows that about 90% of images only contain a single target. This fact supports many model-driven methods to convert the detection task into finding the most sparse or salient target [10, 33]. However, it should be noted that around 10% of images still contain additional targets that would be ignored under such global unique assumptions.

The distribution of the target size proportion is given in Fig. 3(b), where about 55% targets only occupy 0.02% of the image area. Given an image of 300×300, the target is merely 3×3 pixels. Generally, detecting smaller objects requires more contextual information, and infrared small targets push this difficulty to an extreme degree due to the low contrast and background clutters. Therefore, when designing CNNs, the primary priority should be preserving and highlighting features of infrared small targets in deep layers.

The target brightness distribution in percentile rank is given in Fig. 3(c). Note that only 35% targets are the brightest in images. Hence, picking the brightest pixels in the image is not a good idea, resulting in a detection rate of 0.35 with a false alarm rate of 65%. As a comparison, the proposed method in this paper can achieve a detection rate of 0.84 with a false alarm rate of 0.0065% which is 10,000 times smaller. Considering that 65% of targets have a very similar brightness with the background or even darker, we should think twice about the target saliency assumption.

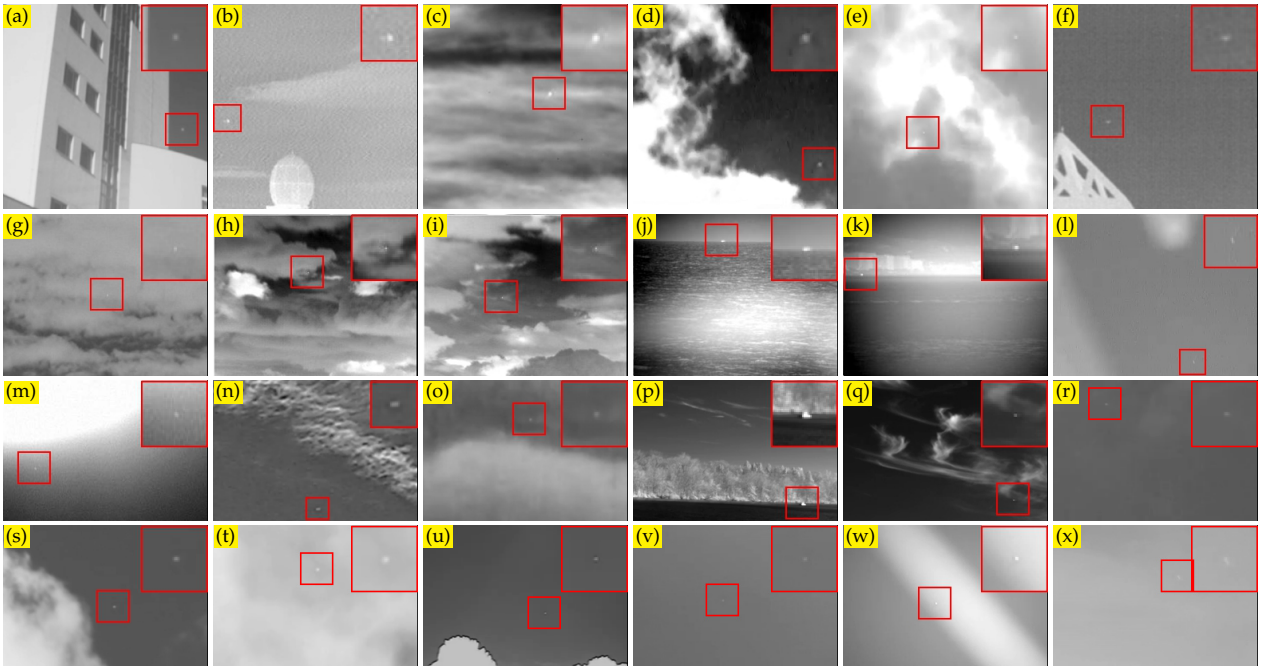
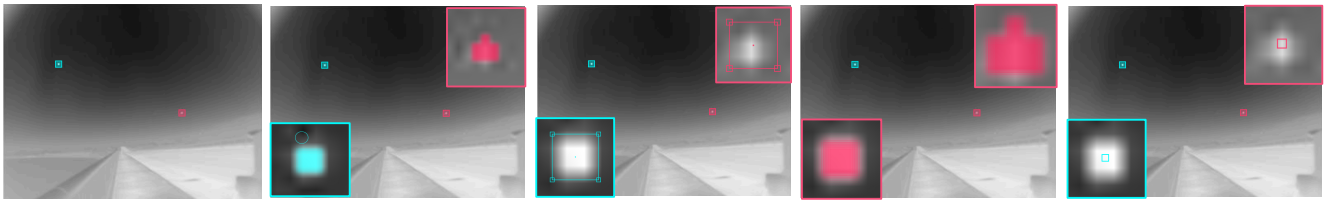


Figure 1: The representative infrared images from the SIRST dataset with various backgrounds.



(a) Category Labeling (b) Instance Segmentation (c) Bounding Box (d) Semantic Segmentation (e) Instance Spotting

Figure 2: Illustration of different kinds of annotations in the proposed SIRST dataset.

3.3. Normalized Intersection over Union

The evaluation metric is also an issue when bridging deep learning with infrared small targets. On the one hand, traditional metrics like background suppression factor or signal to clutter ratio gain are developed for filtering methods to measure the background residual around targets. However, the deep networks output a binary mask, where the values of these metrics would be infinity in most cases. On the other hand, traditional methods tend to model the infrared small target detection as a segmentation process [10], but sacrifice the integrity of the segmented targets for higher detection rate [4], which is very disadvantaged when compared with deep networks designed for semantic segmentation.

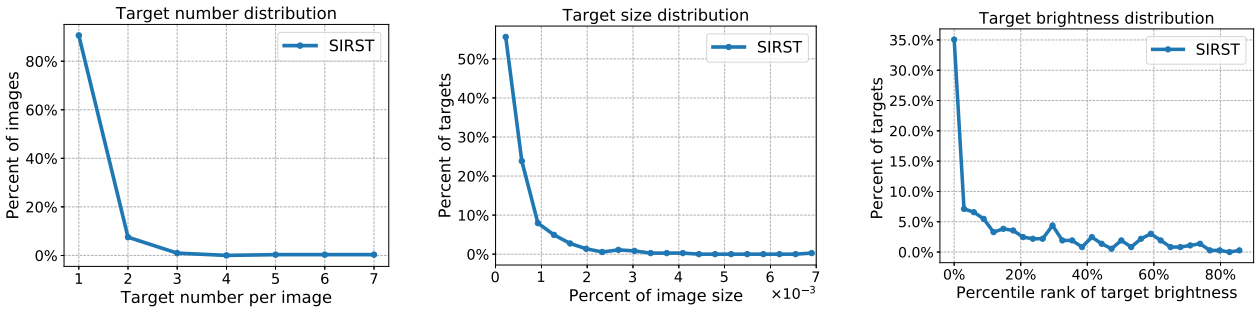
To better balance the model-driven and data-driven methods, we propose the normalized Intersection over Union (nIoU) as a replacement of the IoU, which is defined as

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

where N is the total sample number. With nIoU, we can observe an improvement of model-driven methods and a slight drop of data-driven methods compared to their IoU values. Please note that both IoU and nIoU can not replace the receiver operating characteristic (ROC) curve, since they reflect the segmentation effect under a fixed threshold, while ROC reflects the overall effect under a sliding threshold.

3.4. SIRST Toolkit and Leaderboard

To promote reproducible research, besides an annotated dataset, SIRST is also an open toolkit that provides data processing utilities, common model components, loss functions, and evaluation metrics that are specially designed for infrared small target detection. Building upon those modular APIs, SIRST provides implementations of state-of-the-art networks with trained models. For model-driven methods, the models with the best hyper-parameter settings are also presented with accelerating schemes that do not harm the final performance [4]. Based on this open toolkit, we construct a leaderboard for the selected methods as a place for a fair comparison. Through it, we hope to explore the right evolvement direction for infrared small target detection.



(a) Target Number Distribution (b) Size Ratio Distribution (c) Brightness Distribution

Figure 3: Illustration of SIRST dataset statistics.

4. Asymmetric Contextual Modulation

We now propose the ACM module and the corresponding networks to deal with the challenges: 1) how to construct a deep model to detect infrared small targets lacking intrinsic information; 2) how to encode the high-level contextual information without overwhelming finer details of targets.

4.1. Rethinking Top-Down Attentional Modulation

Given a low-level feature $\mathbf{X}$ and a high-level feature $\mathbf{Y}$ both with C channels and feature maps of size $H \times W$, the top-down attentional modulation [18] can be formulated as

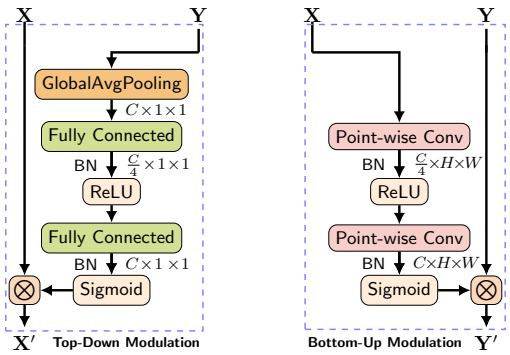
$$\mathbf{X}' = \mathbf{G}(\mathbf{Y}) \otimes \mathbf{X} = \sigma\left(\mathcal{B}\left(\mathbf{W}_2 \delta\left(\mathcal{B}\left(\mathbf{W}_1 \mathbf{y}\right)\right)\right)\right) \otimes \mathbf{X}, \quad (2)$$

where $\mathbf{y}$ is the global feature context obtained by global average pooling $\mathbf{y} = \frac{1}{H \times W} \sum_{i=1,j=1}^{H,W} \mathbf{Y}[:, i, j]$. δ, $\mathcal{B}$, σ, $\otimes$ denote the Rectified Linear Unit (ReLU) [24], Batch Normalization (BN) [16], Sigmoid function, and element-wise multiplication, respectively. $\mathbf{W}_1 \in \mathbb{R}^{\frac{C}{r} \times C}$ and $\mathbf{W}_2 \in \mathbb{R}^{C \times \frac{C}{r}}$ are two fully connected layers. r is the channel reduction ratio.

This top-down modulation shown in Fig. 4(a) implies two assumptions: 1) high-level features provide more accurate semantic information about the target; 2) the global channel context is a competent modulation signal. However, these two assumptions are not necessarily true for infrared small targets as the network goes deeper because, in high-level features, small targets can be easily submerged by the background, and their features are also weakened in the global average pooling. Although the semantic information embedded via the top-down modulation can help handle ambiguity, the prerequisite is that small targets are still preserved.

4.2. Bottom-Up Point-wise Attentional Modulation

To highlight the subtle details of infrared small targets in deep layers, we propose a point-wise channel attention modulation module, in which the channel feature context of each spatial position is aggregated individually. Contrary to the top-down modulation, this modulation pathway propagates the context information in a bottom-up manner to enrich the high-level features with spatial details of low-level feature



(a) (b)

Figure 4: Illustration for one-directional modulation modules. (a) Top-down global attentional modulation [18], (b) The proposed bottom-up point-wise attentional modulation.

maps as illustrated in Fig. 4(b). The contextual modulation weights $\mathbf{L}(\mathbf{X}) \in \mathbb{R}^{C \times H \times W}$ can be computed as

$$\mathbf{L}(\mathbf{X}) = \sigma\left(\mathcal{B}\left(\mathrm{PWConv}_2\left(\delta\left(\mathcal{B}\left(\mathrm{PWConv}_1(\mathbf{X})\right)\right)\right)\right)\right) \quad (3)$$

where PWConv denotes the point-wise convolution [21]. The kernel sizes of PWConv_1 and PWConv_2 are $\frac{C}{4} \times C \times 1 \times 1$ and $C \times \frac{C}{4} \times 1 \times 1$, respectively. It is noteworthy that $\mathbf{L}(\mathbf{X})$ has the same shape as $\mathbf{Y}$, which can highlight the infrared small target in an element-wise way. Then the modulated high-level feature $\mathbf{Y}' \in \mathbb{R}^{C \times H \times W}$ can be obtained via

$$\mathbf{Y}' = \mathbf{L}(\mathbf{X}) \otimes \mathbf{Y}. \quad (4)$$

4.3. Asymmetric Contextual Modulation Module

Our goal is to simultaneously leverage top-down global attentional modulation and bottom-up local attentional modulation to exchange multi-scale context for a richer encoding of both semantic information and spatial details. To this end, the proposed asymmetric contextual modulation for the cross-layer feature fusion is achieved via

$$\mathbf{Z} = \mathbf{G}(\mathbf{Y}) \otimes \mathbf{X} + \mathbf{L}(\mathbf{X}) \otimes \mathbf{Y}, \quad (5)$$

where $\mathbf{Z} \in \mathbb{R}^{C \times H \times W}$ is the fused feature, which is illustrated in Fig. 5, where ReLU and BN are omitted for simplicity.

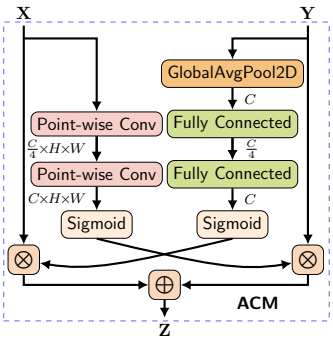


Figure 5: The proposed asymmetric contextual modulation.

4.4. Examples: FPN and U-Net

Following the main practices in this field [10, 33], we model infrared small target detection as a semantic segmentation problem. To show the universality and modularity of the proposed ACM module, we choose FPN [22] and U-Net [27] as host networks. By replacing the original cross-layer feature fusion operations, e.g., the addition in FPN or concatenation in U-Net with the proposed ACM module, we can construct new networks, namely ACM-FPN and ACM-U-Net for the task of infrared small target detection, as shown in Fig. 6. We use the ResNet-20 [14] as the backbone architecture as shown in Table 1, in which we scale the model by depth (the block number b in each stage) to study the relationship between the performance and network depth. Only when $b = 3$, it is the standard backbone of ResNet-20. It should be noted that to preserve the small targets, we adjust the down-sampling scheme specially for this task. In Table 1, the sub-sampling is only performed by at the first convolution layer of Stage-2 and Stage-3.

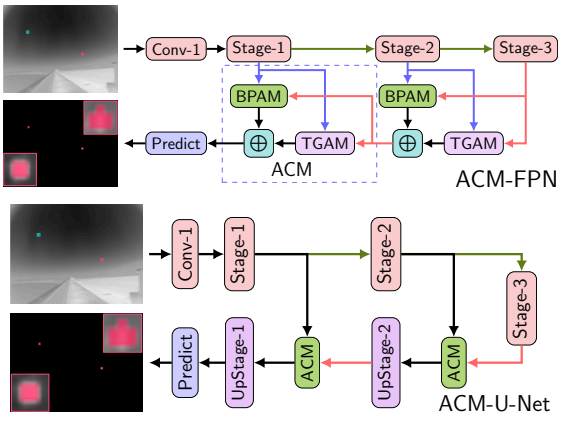


Figure 6: The proposed ACM-FPN and ACM-U-Net.

5. Experiments

We conduct ablation studies and comparison to state-of-the-art methods to verify the effectiveness of the proposed ACM module and the networks. In particular, the following questions will be studied in our experimental evaluation:

Table 1: Backbone for ACM-FPN and ACM-U-Net

Stage	Output	Backbone
Conv-1	480 × 480	3×3 conv, 16
Stage-1 / UpStage-1	480 × 480	$\begin{bmatrix} 3 \times 3 \text{ conv}, 16 \\ 3 \times 3 \text{ conv}, 16 \end{bmatrix} \times b$
Stage-2 / UpStage-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	$\begin{bmatrix} 3 \times 3 \text{ conv}, 64 \\ 3 \times 3 \text{ conv}, 64 \end{bmatrix} \times b$

1. Q1: We will investigate the impact of adjusting the down-sampling scheme for the networks to show that preserving small targets in deep layers is a priority when designing networks for infrared small target detection.
2. Q2: One main contribution of this paper is the supplement of the bottom-up modulation pathway which enables the network to exchange low-level and high-level information in a bi-directional way. We will investigate that given the same parameter budget and point-wise channel attention, whether the bi-directional modulation can outperform the top-down modulation scheme.
3. Q3: Our another contribution is the asymmetric modulation, in which the top-down and bottom-up modulations are achieved via global channel attention and point-wise channel attention, respectively. It raises a question that how important this asymmetric modulation is? Will it outperform other symmetric schemes?
4. Q4: Finally, we will analyze how the networks based on the proposed ACM module compare to other model-driven methods and baseline networks, see Section 5.3.

5.1. Experimental Settings

We model the infrared small target detection as a semantic segmentation task and resort to the proposed SIRST dataset for experimental evaluation. FPN [22] and U-Net [27] are chosen as host networks, where ResNet-20 is the backbone for both. The ROC curve, IoU, and the proposed nIoU are chosen as the evaluation metrics. Since most of the experimental networks cannot take advantage of pre-trained weights, every architecture instantiation is trained from scratch for fairness. The strategy described by He *et al.* [13] is used for weight initialization. We choose the Soft-IoU [26] as the loss function and the Nesterov Accelerated Gradient method as the optimizer. We use a learning rate of 0.05, a batch size of 8, and a total number of 300 epochs.

For data-driven methods, we choose FPN [22], U-Net [27], selective kernel (SK) networks [19] style FPN and U-Net (SK-FPN/SK-U-Net), global attention upsampling (GAU) [18] based GAU-FPN/GAU-U-Net for comparison. For model-driven methods, we choose eleven methods including top-hat filter [1], local contrast method (LCM) [2],

Table 2: Ablation study on the impact of the **down-sampling** scheme and **modulation** scheme.

Modulation Scheme	FPN as Host Network								U-Net as Host Network							
	IoU				nIoU				IoU				nIoU			
	$b=1$	$b=2$	$b=3$	$b=4$	$b=1$	$b=2$	$b=3$	$b=4$	$b=1$	$b=2$	$b=3$	$b=4$	$b=1$	$b=2$	$b=3$	$b=4$
TopDownLocal	0.595	0.648	0.693	0.713	0.635	0.662	0.688	0.703	0.648	0.710	0.713	0.718	0.673	0.692	0.694	0.697
BiGlobal	0.599	0.660	0.685	0.693	0.645	0.674	0.696	0.684	0.682	0.716	0.723	0.730	0.688	0.708	0.707	0.719
BiLocal	0.591	0.662	0.713	0.722	0.657	0.694	0.709	0.714	0.670	0.715	0.718	0.742	0.680	0.710	0.713	0.720
Regular-ACM	**0.683**	**0.703**	0.711	0.711	0.661	0.671	0.680	0.675	0.684	0.700	0.692	0.692	0.637	0.650	0.646	0.643
ACM	0.645	0.700	**0.714**	**0.731**	**0.684**	**0.702**	**0.713**	**0.721**	**0.707**	**0.732**	**0.741**	**0.743**	**0.709**	**0.720**	**0.726**	**0.731**

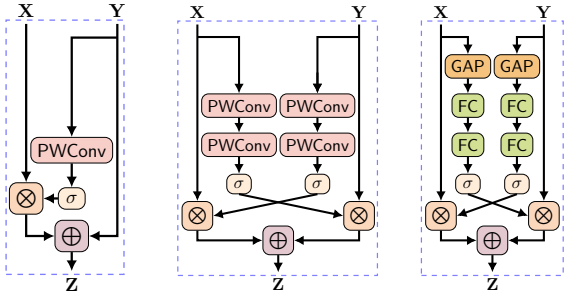
improved LCM (ILCM) [11], local saliency method (LSM) [2], facet kernel and random walker (FKRW) [25], multi-scale patch-based contrast measure (MPCM) [33], infrared patch-image model (IPI) [10], non-negative IPI model based on partial sum of singular values (NIPPS) [5], reweighted infrared patch-tensor model (RIPT) [4], partial sum of the tensor nuclear norm (PSTNN) [38], and non-convex rank approximation minimization (NRAM) [37].

5.2. Ablation Study

Impact of Down-Sampling Scheme: First, we investigate the impact of the down-sampling scheme by comparing the adjusted scheme in Table 1 and the *regular* scheme in [14] that the feature maps are down-sampled four times more. The comparison results are shown in Table 2. It can be seen that ACM based networks outperform significantly better than the Regular-ACM based networks, especially as the network goes deeper. The results show that it is necessary to customize the network down-sampling scheme for infrared small target detection. Otherwise, excessive down-sampling will cause the loss of small target features in deep layers.

Impact of Bi-directional Attentional Modulation: In this part, we compare the one-directional top-down modulation module, i.e., TopDownLocal as shown in Fig. 7(a), with the two-directional modulation module, i.e., BiLocal as shown in Fig. 7(b). To keep the comparison fair, we keep the parameter budget of the point-wise channel attention the same for both, namely C^2. From Table 2, it can be seen that BiLocal always performs better than the TopDownLocal, which shows that it is better to use bi-directional attentional modulation, instead of top-down modulation only. We believe this performance gain comes from the low-level fine details that are embedded in high-level features via the proposed bottom-up modulation pathway, which helps to preserve small targets in deep layers.

Impact of Asymmetric Attentional Modulation: Table 2 presents a comparison among BiLocal, BiGlobal (Fig. 7(c)), and the proposed ACM to verify the effectiveness of the proposed asymmetric attentional modulation, in which we can see that compared to the modulation scheme whose channel attention scales are both local (BiLocal) or global



(a) TopDownLocal (b) BiLocal (c) BiGlobal

Figure 7: Architectures for the ablation study on modulation scheme. (a) Top-down modulation with point-wise channel attention module (TopDownLocal); (b) Bi-directional modulation with point-wise channel attention module (BiLocal); (c) Bi-directional modulation with global channel attention module (BiGlobal). All these architectures share the same number of learning parameters C^2.

(BiGlobal), the proposed ACM module which utilizes global channel attention in the top-down pathway and point-wisely local channel attention in the bottom-up pathway, performs the best in all settings. The results verify our hypothesis of the proposed asymmetric modulation, that is, top-down modulation needs a global channel attention module for the high-level semantic information of the whole image, while the bottom-up modulation requires a point-wise channel attention mechanism for the low-level finer details.

5.3. Comparison to State-of-the-Art Methods

In this subsection, we first compare the proposed ACM-FPN and ACM-U-Net with other state-of-the-art networks as the network depth grows in Fig. 8. It can be seen that 1) The proposed networks achieve best in all kinds of settings, even with fewer layers. Moreover, this performance advantage will not subside as the network goes deeper. It demonstrates the goal of this paper that *with the proposed ACM module, host networks can gain a significant performance boost, even with fewer layers or parameters per network*. 2) As the network depth grows, the advantage of merely top-down global attentional modulation subsides gradually. For example, when $b=4$, the baseline FPN and U-Net perform even

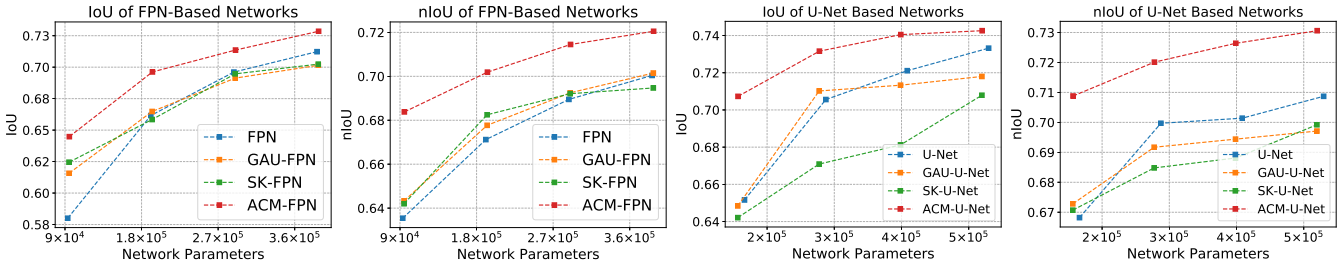


Figure 8: The IoU/nIoU comparison with other cross-layer modulation schemes given FPN and U-Net as host networks.

Table 3: IoU and nIoU comparison among 19 methods.

Methods	Model-Driven											Data-Driven							
	Local Contrast Measurement						Local Rank + Sparse Decomposition					FPN Based				U-Net Based			
	Tophat	LCM	ILCM	LSM	FKRW	MPCM	IPI	NIPPS	RIPT	PSTNN	NRAM	FPN	SK	GAU	ACM	U-Net	SK	GAU	ACM
IoU	0.220	0.193	0.104	0.1864	0.268	0.357	0.466	0.473	0.146	0.605	0.294	0.720	0.702	0.701	**0.731**	0.733	0.708	0.718	**0.743**
nIoU	0.352	0.207	0.123	0.2598	0.339	0.445	0.607	0.602	0.245	0.504	0.424	0.700	0.695	0.701	**0.721**	0.709	0.699	0.697	**0.731**

better than SK-FPN/SK-U-Net and GAU-FPN/GAU-U-Net, which shows that there is a high risk for the high-level semantic features to overwhelm the features of small targets in top-down modulation. This also proves the necessity of the proposed bottom-up attentional modulation pathway.

Next, we compare the performance of the proposed networks with other state-of-the-art model-driven methods as well as the data-driven networks. Table 3 shows the IoU and nIoU comparison results of a total of 19 methods. It can be seen that 1) The proposed networks achieve the best in both IoU and nIoU evaluation, showing the effectiveness of the proposed asymmetric attentional modulation; 2) The data-driven methods all perform better than the model-driven methods, which shows that with the proposed SIRST dataset, we should pay more attention to data-driven methods to obtain state-of-the-art performance. 3) For model-driven methods, their nIoU numbers are generally higher than IoU numbers, while the data-driven methods are the opposite. It validates our argument that the networks tend to improve performance on larger targets to minimize the loss function and pay less attention to the smaller ones. It is fair to conclude that nIoU is a better metric than IoU in evaluating the performance of infrared small target detection.

Finally, we compare the ROC curves among seven selected methods in Fig. 9. It can be seen that the proposed ACM-FPN and ACM-U-Net achieve the best, showing the effectiveness of the proposed ACM module. Another interesting point is that although RIPT performs worse than both MPCM and IPI in nIoU and IoU in Table 3, it performs better than them in terms of ROC in Fig. 9. To our understanding, the reason behind this is that IoU and nIoU reflect the segmentation effect under a fixed threshold, while ROC reflects the overall effect under a sliding threshold. It shows that RIPT trades off the detection ability with the target integrity.

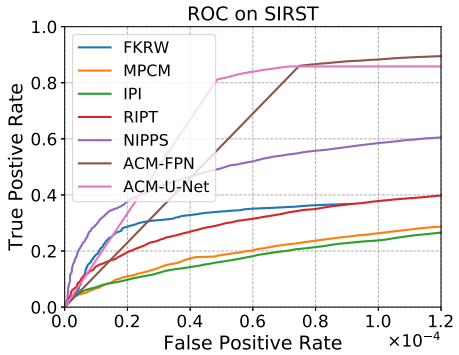


Figure 9: ROC comparison of selected methods

6. Conclusion

In this paper, we first contribute an open dataset for detecting and segmenting infrared small targets in single-frame scenarios. Further, we propose the asymmetric contextual modulation which is specially designed for infrared small targets. The innovation is two-fold. First, the supplement of the bottom-up modulation pathway enables the networks to embed low-level contexts of fine details into high-level features. Second, the point-wise channel attention module highlights the features of infrared small targets, instead of being overwhelmed by their background neighborhoods. Extensive ablation experiments demonstrate the effectiveness of the proposed architecture. Compared with other state-of-the-art approaches, our networks can achieve better performance with fewer parameters and layers.

Acknowledgement

Supported by NSFC No. 61573183, Open Project of NLPR No. 201900029, NUAA PhD visiting scholar project No. 180104DF03, and CSC No. 201806830039.

References

[1] Xiangzhi Bai and Fugen Zhou. Analysis of New Top-Hat Transformation and the Application for Infrared Dim Small Target Detection. *Pattern Recognition*, 43(6):2145–2156, Jun 2010.

[2] C. L. Philip Chen, Hong Li, Yantao Wei, Tian Xia, and Yuan Yan Tang. A local contrast method for small infrared target detection. *IEEE Transactions on Geoscience and Remote Sensing*, 52(1):574–581, 2014.

[3] Yuwen Chen and Yunhong Xin. An efficient infrared small target detection method based on visual contrast mechanism. *IEEE Geoscience and Remote Sensing Letters*, 13(7):962–966, 2016.

[4] Yimian Dai and Yiquan Wu. Reweighted infrared patch-tensor model with both nonlocal and local priors for single-frame small target detection. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 10(8):3752–3767, 2017.

[5] Yimian Dai, Yiquan Wu, Yu Song, and Jun Guo. Non-negative infrared patch-image model: Robust target-background separation via partial sum minimization of singular values. *Infrared Physics & Technology*, 81:182–194, 2017.

[6] James W. Davis and Mark A. Keck. A two-stage template approach to person detection in thermal imagery. In *7th IEEE Workshop on Applications of Computer Vision / IEEE Workshop on Motion and Video Computing (WACV/MOTION 2005), 5-7 January 2005, Breckenridge, CO, USA*, pages 364–369. IEEE Computer Society, 2005.

[7] H. Deng, X. Sun, M. Liu, C. Ye, and X. Zhou. Small infrared target detection based on weighted local difference measure. *IEEE Transactions on Geoscience and Remote Sensing*, 54(7):4204–4214, 2016.

[8] Lili Dong, Bin Wang, Ming Zhao, and Wenhai Xu. Robust infrared maritime target detection based on visual attention and spatiotemporal filtering. *IEEE Transactions on Geoscience and Remote Sensing*, 55(5):3037–3050, 2017.

[9] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3146–3154, 2019.

[10] Chenqiang Gao, Deyu Meng, Yi Yang, Yongtao Wang, Xiaofang Zhou, and Alexander G. Hauptmann. Infrared patch-image model for small target detection in a single image. *IEEE Transactions on Image Processing*, 22(12):4996–5009, 2013.

[11] Jinhui Han, Yong Ma, Bo Zhou, Fan Fan, Kun Liang, and Yu Fang. A Robust Infrared Small Target Detection Algorithm Based on Human Visual System. *IEEE Geoscience and Remote Sensing Letters*, 11(12):2168–2172, May 2014.

[12] Bharath Hariharan, Pablo Andrés Arbeláez, Ross B. Girshick, and Jitendra Malik. Hypercolumns for object segmentation and fine-grained localization. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 447–456. IEEE Computer Society, 2015.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *IEEE International Conference on Computer Vision (ICCV)*, ICCV '15, pages 1026–1034, Washington, DC, USA, 2015.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[15] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7132–7141, 2018.

[16] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning (ICML)*, pages 448–456, 2015.

[17] Sungho Kim. High-speed incoming infrared target detection by fusion of spatial and temporal detectors. *Sensors*, 15(4):7267–7293, 2015.

[18] Hanchao Li, Pengfei Xiong, Jie An, and Lingxue Wang. Pyramid attention network for semantic segmentation. In *British Machine Vision Conference (BMVC)*, pages 1–13, 2018.

[19] Xiang Li, Wenhai Wang, Xiaolin Hu, and Jian Yang. Selective kernel networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 510–519, 2019.

[20] Yanghao Li, Yuntao Chen, Naiyan Wang, and Zhao-Xiang Zhang. Scale-aware trident networks for object detection. In *IEEE International Conference on Computer Vision (ICCV)*, pages 6053–6062, 2019.

[21] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. In *International Conference on Learning Representations (ICLR)*, pages 1–10, 2014.

[22] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. Feature pyramid networks for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 936–944, 2017.

[23] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pages 740–755, Cham, 2014.

[24] Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines. In *International Conference on Machine Learning (ICML)*, ICML'10, pages 807–814, USA, 2010.

[25] Yao Qin, Lorenzo Bruzzone, Chengqiang Gao, and Biao Li. Infrared small target detection based on facet kernel and random walker. *IEEE Transactions on Geoscience and Remote Sensing*, 57(9):7104–7118, 2019.

[26] Md Atiqur Rahman and Yang Wang. Optimizing intersection-over-union in deep neural networks for image segmentation. In *International Symposium on Visual Computing*, pages 234–244, 2016.

[27] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234–241, 2015.

[28] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[29] Bharat Singh and Larry S. Davis. An analysis of scale invariance in object detection - SNIP. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3578–3587, June 2018.

[30] Bharat Singh, Mahyar Najibi, and Larry S. Davis. SNIPER: efficient multi-scale training. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 9333–9343, 2018.

[31] Huan Wang, Luping Zhou, and Lei Wang. Miss detection vs. false alarm: Adversarial learning for small object segmentation in infrared images. In *IEEE International Conference on Computer Vision (ICCV)*, pages 8508–8517, 2019.

[32] Wenguan Wang, Shuyang Zhao, Jianbing Shen, Steven C. H. Hoi, and Ali Borji. Salient object detection with pyramid attention and salient edges. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 1448–1457. Computer Vision Foundation / IEEE, 2019.

[33] Yantao Wei, Xinge You, and Hong Li. Multiscale patch-based contrast measure for small infrared target detection. *Pattern Recognition*, 58:216–226, 2016.

[34] Wei Zhang, Mingyu Cong, and Liping Wang. Algorithms for optical weak small targets detection and tracking: review. In *International Conference on Neural Networks and Signal Processing*, volume 1, pages 643–647, 2003.

[35] Weitao Yuan, Shengbei Wang, Xiangrui Li, Masashi Unoki, and Wenwu Wang. A skip attention mechanism for monaural singing voice separation. *IEEE Signal Processing Letters*, 26(10):1481–1485, 2019.

[36] Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Zhi Zhang, Haibin Lin, Yue Sun, Tong He, Jonas Mueller, R. Manmatha, Mu Li, and Alexander Smola. ResNeSt: Split-Attention Networks. *arXiv e-prints*, page arXiv:2004.08955, Apr. 2020.

[37] Landan Zhang, Lingbing Peng, Tianfang Zhang, Siying Cao, and Zhenming Peng. Infrared Small Target Detection via Non-Convex Rank Approximation Minimization Joint l2,1 Norm. *Remote Sensing*, 10(11):1821, Nov 2018.

[38] Landan Zhang and Zhenming Peng. Infrared Small Target Detection Based on Partial Sum of the Tensor Nuclear Norm. *Remote Sensing*, 11(4):382, Jan 2019.

[39] Mingxin Zhao, Li Cheng, Xu Yang, Peng Feng, Liyuan Liu, and Nanjian Wu. Tbc-net: A real-time detector for infrared small target detection using semantic constraint. *arXiv preprint arXiv:2001.05852*, 2019.

[40] Hu Zhu, Shiming Liu, Lizhen Deng, Yansheng Li, and Fu Xiao. Infrared small target detection via low-rank tensor completion with top-hat regularization. *IEEE Transactions on Geoscience and Remote Sensing*, 58(2):1004–1016, 2020.