HCF-Net: Hierarchical Context Fusion Network for Infrared Small Object Detection

Shibiao Xu[1], ShuChen Zheng[1], Wenhao Xu[1], Rongtao Xu[3,4], Changwei Wang[2,3,5,*],
Jiguang Zhang[3,4], Xiaoqiang Teng[1] Ao Li[6,*], Li Guo[2]

[1]Artificial Intelligence, Beijing University of Posts and Telecommunications
[2]Key Laboratory of Computing Power Network and Information Security, Ministry of Education,
Shandong Computer Science Center, Qilu University of Technology (Shandong Academy of Sciences)
[3]State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences
[4]School of Artificial Intelligence, University of Chinese Academy of Sciences
[5] Shandong Provincial Key Laboratory of Computer Networks, Shandong Fundamental Research Center for Computer Science
[6]School of Computer Science and Technology, Harbin University of Science and Technology

Abstract—**Infrared small object detection is an important computer vision task involving the recognition and localization of tiny objects in infrared images, which usually contain only a few pixels. However, it encounters difficulties due to the diminutive size of the objects and the generally complex backgrounds in infrared images. In this paper, we propose a deep learning method, HCF-Net, that significantly improves infrared small object detection performance through multiple practical modules. Specifically, it includes the parallelized patch-aware attention (PPA) module, dimension-aware selective integration (DASI) module, and multi-dilated channel refiner (MDCR) module. The PPA module uses a multi-branch feature extraction strategy to capture feature information at different scales and levels. The DASI module enables adaptive channel selection and fusion. The MDCR module captures spatial features of different receptive field ranges through multiple depth-separable convolutional layers. Extensive experimental results on the SIRST infrared single-frame image dataset show that the proposed HCF-Net performs well, surpassing other traditional and deep learning models. Code is available at https://github.com/zhengshuchen/HCFNet.**

Index Terms—**Infrared small object detection, Deep learning, Multi-scale features.**

I. INTRODUCTION

Infrared small object detection is a crucial technology for identifying and detecting minute objects in infrared images. Due to the ability of infrared sensors to capture the infrared radiation emitted by objects, this technology enables precise detection and identification of small objects, even in dark or low-light environments. As a result, it holds significant application prospects and value in various fields, including military, security, maritime rescue, and fire monitoring.

However, Infrared small object detection is still challenging for the following reasons. First, deep learning currently serves

as the primary method for infrared small object detection. However, almost all existing networks adopt classic downsampling schemes. Infrared small objects, due to their small size, often come with weak thermal signals and unclear contours. There is a significant risk of information loss during multiple downsampling processes. Second, compared to visible light images, infrared images lack physical information and have lower contrast, making small objects easily submerged in complex backgrounds.

To tackle these challenges, We propose an infrared small object detection model named HCF-Net. This model aims for a more precise depiction of object shape and boundaries, enhancing the accuracy of object localization and segmentation by framing infrared small object detection as a semantic segmentation problem. As illustrated in Fig. 1, it incorporates three key modules: PPA, DASI, and MDCR, which address the challenges mentioned above on multiple levels.

Specifically, as a primary component of the encoder-decoder, PPA employs hierarchical feature fusion and attention mechanisms to maintain and enhance representations of small objects, ensuring crucial information is preserved through multiple downsampling steps. DASI enhances the skip connection in U-Net, focusing on the adaptive selection and delicate fusion of high and low-dimensional features to enhance the saliency of small objects. Positioned deep within the network, MDCR reinforces multi-scale feature extraction and channel information representation, capturing features across various receptive field ranges. It more finely models the differences between objects and backgrounds, enhancing its ability to locate small objects. The organic combination of these modules enables us to address the challenges of small object detection more effectively, improving detection performance and robustness.

In summary, our contributions in this paper can be summarized as follows:

- We model infrared small object detection as a semantic segmentation problem and propose HCF-Net, a layer-
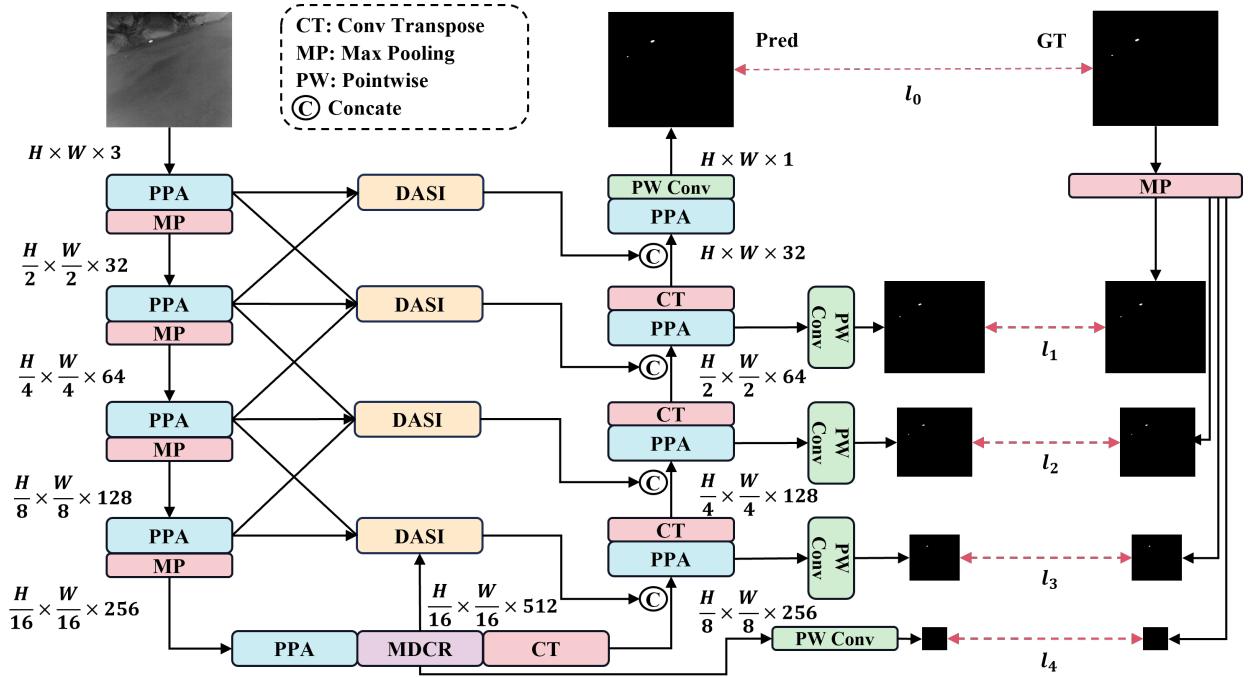
Fig. 1. Network Architecture. The encoder primarily comprises the parallelized patch-aware attention (PPA) module and max-pooling layers, while the decoder mainly consists of PPA and convolutional transpose (CT) layers. We incorporate the multi-dilated channel refiner (MDCR) module as an intermediary layer to bridge the encoder and decoder. Within the skip-connection component, we introduce the dimension-aware selective integration (DASI) module to enhance the fusion and propagation of features across different network layers.

wise context fusion network that can be trained from scratch.

- Three practical modules have been proposed: parallelized patch-aware attention (PPA) module, dimension-aware selective integration (DASI) module, and multi-dilated channel refiner (MDCR) module. These modules effectively alleviate the issues of small object loss and low background distinctiveness in infrared small object detection.
- We evaluate the proposed HCF-Net's detection performance on the publicly available single-frame infrared image dataset SRIST and demonstrate a significant advantage over several state-of-the-art detection methods.

II. RELATED WORK

A. Traditional Methods

In the early stages of infrared small object detection, the predominant approaches were model-based traditional methods, generally categorized into filter-based methods, methods based on the human visual system, and low-rank methods. Filter-based methods are typically limited to specific and uniform scenarios. For example, TopHat [1] estimates the scene background using various filters to separate the object from a complex background. Methods based on the human visual system are suitable for scenarios with large objects and strong background differentiation, such as LCM [2], which measures the contrast between the center point and its surrounding environment. Low-rank methods are suitable for fast-changing and complex backgrounds but lack real-time performance in practical applications, often requiring additional assistance

such as GPU acceleration. Examples of these methods include IPI [3], which combines low-rank background with sparsely shaped objects using low-rank decomposition, PSTNN [4] employing a non-convex method based on tensor nuclear norms, RIPT [5] that focuses on reweighted infrared patch tensors, and NIPPS [6], an advanced optimization approach that attempts to incorporate low-rank and prior constraints. While effective in specific scenarios, traditional methods are susceptible to interference from clutter and noise. In complex real-world scenarios, modeling objects is significantly affected by model hyperparameters, resulting in poor generalization performance.

B. Deep Learning Methods

In recent years, with the rapid development of neural networks, deep learning methods have significantly advanced the infrared small object detection task. Deep learning approaches [7]–[14] exhibit higher recognition accuracy than traditional methods without relying on specific scenes or devices, demonstrating increased robustness and significantly lower costs, gradually taking a dominant position in the field. Wang et al. [15] used the model trained by ImageNet Large Scale Visual Recognition Challenge (ILSVRC) data to complete the infrared small object detection task. Liangkui et al. [16] combined with the data generated from oversampling, a multi-layer network was proposed for small object detection. Zhao et al. [17] developed an encoder-decoder detection method (TBC-Net) combining semantic constraint information of infrared small objects. Wang et al. [18] employed a generator and discriminator to address two distinct tasks: miss detection and

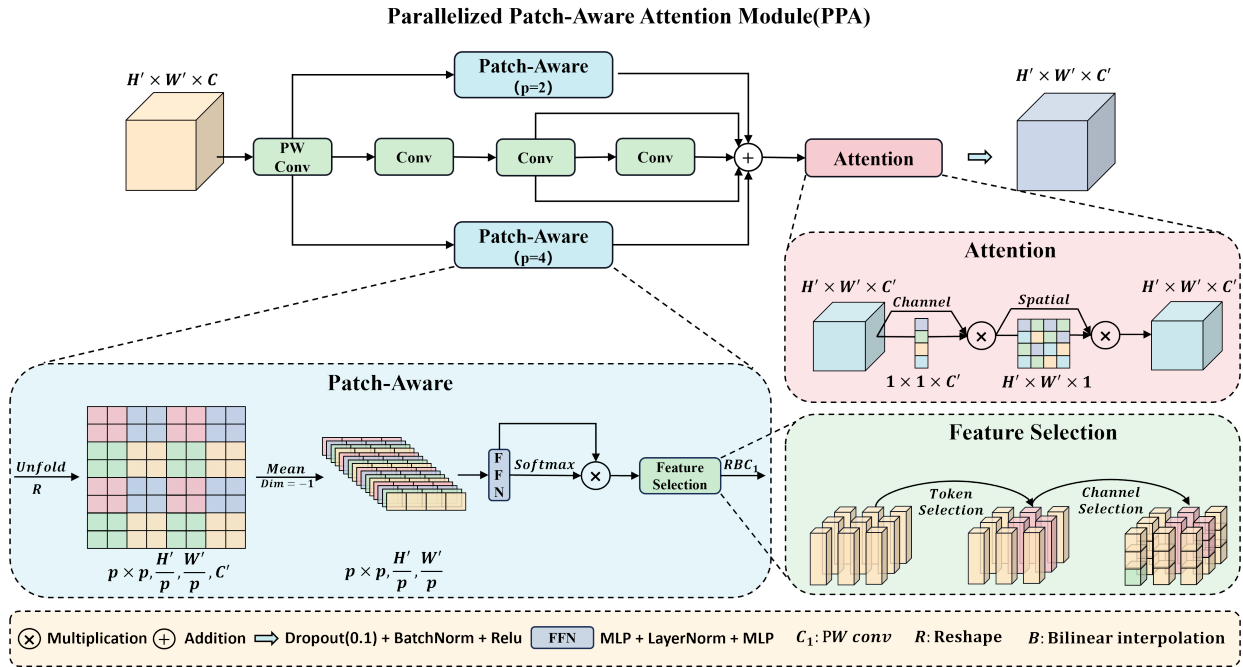**Parallelized Patch-Aware Attention Module(PPA)**

Fig. 2. Detailed structure of the parallelized patch-aware attention module. This module primarily consists of two components: multi-branch fusion and attention mechanisms. The multi-branch fusion component includes patch-aware and concatenated convolutions. The 'p' parameter in patch-aware is set to 2 and 4, representing local and global branches, respectively.

false alarm, achieving a balance between these aspects. Nasser et al. [19] proposed a deep convolutional neural network model for automatic object recognition (ATR). Zhang et al. proposed AGPCNet [20], which introduced attention-guided context modules. Dai et al. introduced the asymmetric context modulation ACM [21] and introduced the first real-world infrared small object dataset, SIRST. Wu et al. [22] proposed a "U-Net within U-Net" framework to achieve multi-level representation learning of goals.

III. METHOD

In this section, we will be discussing HCF-Net in detail. As shown in Fig. 1, HCF-Net is an upgraded U-Net architecture that consists of three crucial modules: PPA, DASI, and MDCR. These modules make our network more suitable for detecting small infrared objects and effectively tackle the challenges of small object loss and low background distinctiveness. Next, we will provide a brief introduction to PPA in Sec. III-A, followed by an overview of DASI in Sec. III-B, and finally, an introduction to MDCR in Sec. III-C.

A. Parallelized Patch-Aware Attention Module

In infrared small object detection tasks, small objects are prone to losing crucial information during multiple down-sampling operations. As depicted in Fig. 1, PPA substitutes traditional convolution operations in the encoder and decoder's fundamental components to better address this challenge.

1) Multi-branch feature extraction: The primary strength of PPA resides in its multi-branch feature extraction strategy. As depicted in Fig. 2, PPA employs a parallel multi-branch approach and Each branch is tasked with extracting features at

various scales and levels. This multi-branch strategy facilitates the capture of multi-scale features of the object, consequently improving the accuracy of small object detection. Specifically, this strategy involves three parallel branches: the local, global, and serial convolution branches. Given the input feature tensor $\mathbf{F} \in \mathbb{R}^{H' \times W' \times C}$, it is first adjusted through point-wise convolution to obtain $\mathbf{F}' \in \mathbb{R}^{H' \times W' \times C'}$. Then, through the three branches, you can calculate $\mathbf{F}_{local} \in \mathbb{R}^{H' \times W' \times C'}$, $\mathbf{F}_{global} \in \mathbb{R}^{H' \times W' \times C'}$, and $\mathbf{F}_{conv} \in \mathbb{R}^{H' \times W' \times C'}$ separately. Finally, these three results are summed to obtain $\tilde{\mathbf{F}} \in \mathbb{R}^{H' \times W' \times C'}$.

Specifically, the distinction between the local and global branches is established by controlling the patch size parameter p, which is realized through the aggregation and displacement of non-overlapping patches in spatial dimensions. Furthermore, we compute the attention matrix between non-overlapping patches to enable local and global feature extraction and interaction.

Initially, we employ computationally efficient operations, including Unfold and reshape, to partition $\mathbf{F}'$ into a set of spatially contiguous patches $(p \times p, H'/p, W'/p, C)$. Subsequently, we conduct channel-wise averaging to yield $(p \times p, H'/p, W'/p)$, followed by linear computations using FFN [23]. Subsequently, we apply the activation function to obtain the probability distribution in the spatial dimension for the linearly computed features and adjust their weights accordingly.

In the weighted outcomes, we employ feature selection [24] to choose pertinent features for the task from tokens and channels. To be specific, let $d = \frac{H' \times W'}{p \times p}$, and represent the weighted outcome as $(\mathbf{t}_i)_{i=1}^{C'}$, where $\mathbf{t}_i \in \mathbb{R}^d$ represents the i-th output token. Feature selection operates on each token,

yielding the output as $\hat{\mathbf{t}}_i = \mathbf{P} \cdot sim(\mathbf{t}_i, \xi) \cdot \mathbf{t}_i$, where $\xi \in \mathbb{R}^{C'}$ and $\mathbf{P} \in \mathbb{R}^{C' \times C'}$ are task-specific parameters, and $sim(\cdot, \cdot)$ is a cosine similarity function bounded within [0,1]. Here, ξ functions as the task embedding, specifying which tokens are relevant to the task. Each token $\mathbf{t}_i$ is reweighted based on its relevance to the task embedding (measured by cosine similarity), effectively simulating token selection. Subsequently, we apply a linear transformation of $\mathbf{P}$ for channel selection for each token, followed by reshape and interpolation operations, ultimately producing the features $\mathbf{F}_{local} \in \mathbb{R}^{H' \times W' \times C'}$ and $\mathbf{F}_{global} \in \mathbb{R}^{H' \times W' \times C'}$. Finally, we substitute the conventional 7x7, 5x5, and 3x3 convolution layers with a serial convolution consisting of three 3x3 convolution layers. This results in three distinct outputs: $\mathbf{F}_{conv1} \in \mathbb{R}^{H' \times W' \times C'}$, $\mathbf{F}_{conv2} \in \mathbb{R}^{H' \times W' \times C'}$, and $\mathbf{F}_{conv3} \in \mathbb{R}^{H' \times W' \times C'}$, which are then summed to obtain the serial convolution output $\mathbf{F}_{conv} \in \mathbb{R}^{H' \times W' \times C'}$.

2) Feature fusion and attention: Following feature extraction via the multi-branch feature extraction, we conduct adaptive feature enhancement using attention mechanisms. The attention module comprises a sequence of efficient channel attention [25] and spatial attention [26] components. In this context, $\tilde{\mathbf{F}} \in \mathbb{R}^{H \times W \times C'}$ is successively processed by a one-dimensional channel attention map $\mathbf{M}_c \in \mathbb{R}^{1 \times 1 \times C'}$ and a two-dimensional spatial attention map $\mathbf{M}_s \in \mathbb{R}^{H' \times W' \times 1}$. This process can be summarized as follows:

$$\mathbf{F}_c = \mathbf{M}_c(\tilde{\mathbf{F}}) \otimes \tilde{\mathbf{F}}, \quad \mathbf{F}_s = \mathbf{M}_s(\mathbf{F}_c) \otimes \mathbf{F}_c, \quad (1)$$

$$\mathbf{F}'' = \delta(\mathcal{B}(dropout(\mathbf{F}_s))), \quad (2)$$

where $\otimes$ denotes element-wise multiplication, $\mathbf{F}_c \in \mathbb{R}^{H \times W \times C'}$ and $\mathbf{F}_s \in \mathbb{R}^{H \times W \times C'}$ represent features after channel and spatial selection, $\delta(\cdot)$ and $\mathcal{B}(\cdot)$ represent Rectified Linear Unit (*ReLU*) and Batch Normalization (*BN*), respectively, and $\mathbf{F}'' \in \mathbb{R}^{H \times W \times C'}$ is the final output of PPA.

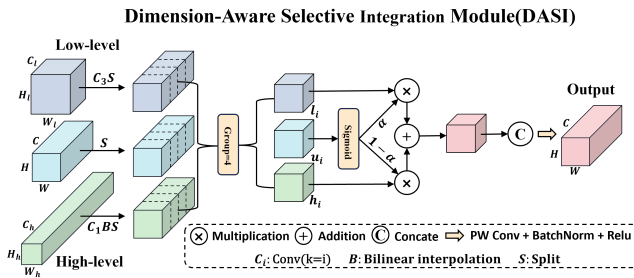### B. Dimension-Aware Selective Integration Module

Fig. 3. Detail structure of the dimension-aware selective integration module.

During the multiple downsampling stages in infrared small object detection, high-dimensional features may lose information about small objects, while low-dimensional features may fail to provide sufficient context. To address this, we propose a novel channel partition selection mechanism (depicted in Fig. 3), enabling DASI to adaptively select appropriate features for fusion based on the object's size and

characteristics. In particular, DASI initially aligns the high-dimensional features $\mathbf{F_h} \in \mathbb{R}^{H_h \times W_h \times C_h}$ and low-dimensional features $\mathbf{F_l} \in \mathbb{R}^{H_l \times W_l \times C_l}$ with the features of the current layer $\mathbf{F_u} \in \mathbb{R}^{H \times W \times C}$ through operations like convolution and interpolation. Subsequently, it divides them into four equal segments in the channel dimension, resulting in $(\mathbf{h}_i)_{i=1}^4 \in \mathbb{R}^{H \times W \times \frac{C}{4}}$, $(\mathbf{l}_i)_{i=1}^4 \in \mathbb{R}^{H \times W \times \frac{C}{4}}$, and $(\mathbf{u}_i)_{i=1}^4 \in \mathbb{R}^{H \times W \times \frac{C}{4}}$, where $\mathbf{h}_i$, $\mathbf{l}_i$, and $\mathbf{u}_i$ denote the i-th partitioned features of high-dimensional, low-dimensional, and current layer features, respectively. These partitions are computed according to the following formulas:

$$\alpha = sigmoid(\mathbf{u}_i), \quad \mathbf{u}_i' = \alpha \mathbf{l}_i + (1 - \alpha)\mathbf{h}_i, \quad (3)$$

$$\mathbf{F}_u' = [\mathbf{u}_1', \mathbf{u}_2', \mathbf{u}_3', \mathbf{u}_4'], \quad \hat{\mathbf{F_u}} = \delta(\mathcal{B}(Conv(\mathbf{F_u'}))), \quad (4)$$

where $\alpha \in \mathbb{R}^{H \times W \times \frac{C}{4}}$ represents the values obtained through the activation function applied to $\mathbf{u}_i$, $\mathbf{u}_i' \in \mathbb{R}^{H \times W \times \frac{C}{4}}$ represents the selectively aggregated results for each partition. After merging $(\mathbf{u}_i')_{i=1}^4$ in the channel dimension, we obtain $\mathbf{F_u'} \in \mathbb{R}^{H \times W \times C}$. The operations $Conv()$, $\mathcal{B}()$, and $\delta()$ denote convolution, batch normalization (*BN*), and rectified linear unit (*ReLU*), respectively, ultimately resulting in the output $\hat{\mathbf{F_u}} \in \mathbb{R}^{H \times W \times C}$.

If $\alpha > 0.5$, the model prioritizes fine-grained features, while if $\alpha < 0.5$, it emphasizes context features.

C. Multi-Dilated Channel Refiner Module

In the MDCR, we introduce multiple depth-wise separable convolution layers with varying dilation rates to capture spatial features across a range of receptive field sizes, which allows for more detailed modeling of the differences between objects and backgrounds, enhancing its ability to discriminate small objects.
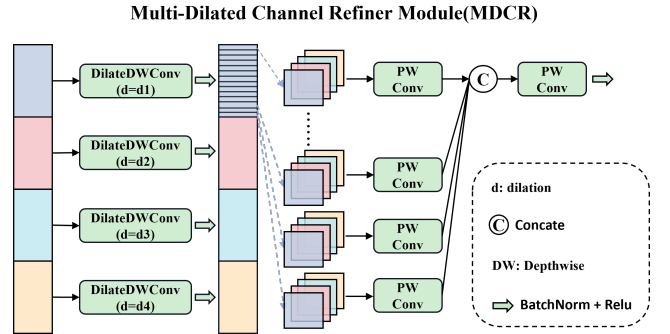

Fig. 4. Detail structure of multi-dilated channel refiner module.

Illustrated in Fig. 4, MDCR partitions the input features $\mathbf{F_a} \in \mathbb{R}^{H \times W \times C}$ into four distinct heads along the channel dimension, generating $(\mathbf{a}_i)_{i=1}^4 \in \mathbb{R}^{H \times W \times \frac{C}{4}}$. Each head then undergoes separate depth-wise separable dilated convolution with distinct dilation rates, yielding $(\mathbf{a}_i')_{i=1}^4 \in \mathbb{R}^{H \times W \times \frac{C}{4}}$. We designate the convolution dilation rates as $d1$, $d2$, $d3$, and $d4$.

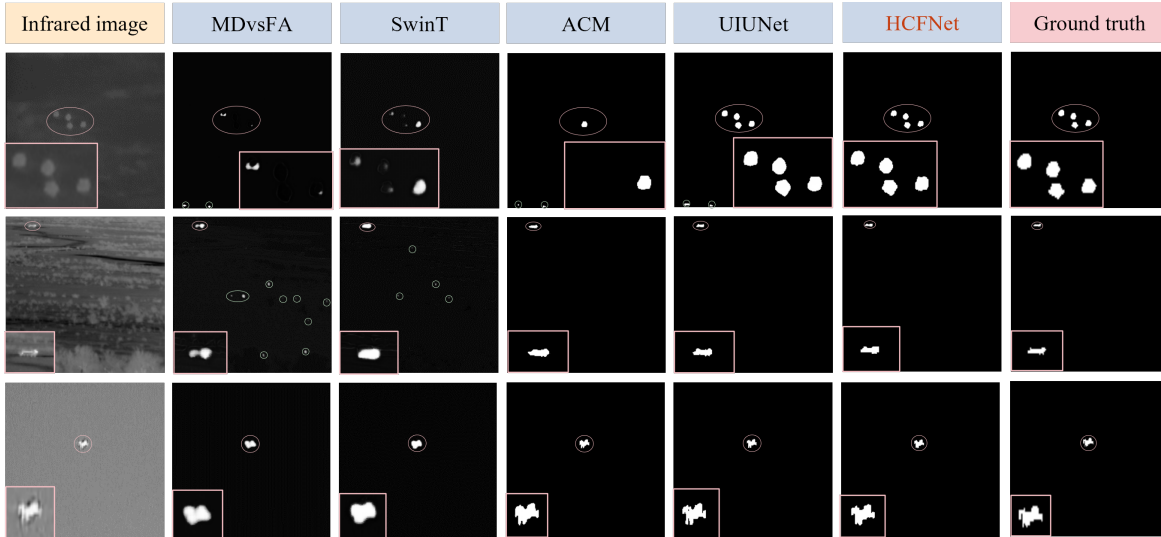$$\mathbf{a}_i' = DDWConv(\mathbf{a}_i), \quad (5)$$

Fig. 5. Visual examples of representative methods are provided. Pink and green circles represent true positive and false positive objects, respectively. The pink rectangle zooms in on true positive objects for a more apparent distinction of detection accuracy among different methods.

where $\mathbf{a}'_i$ denotes the features acquired by applying depth-wise separable dilated convolution to the i-th head. The operation $DDWConv()$ represents depth-wise separable dilated convolution, and i takes values in $1, 2, 3, 4$.

MDCR enhances the feature representation through channel segmentation and recombination. Specifically, we split $\mathbf{a}'_i$ into individual channels to obtain $(\mathbf{a}^j_i)^{\frac{C}{4}}_{j=1} \in \mathbb{R}^{H \times W \times 1}$ for each head. Following this, we interleave these channels across the heads to form $(\mathbf{h}_j)^{\frac{C}{4}}_{j=1} \in \mathbb{R}^{H \times W \times 4}$, thereby enhancing the diversity of multi-scale features. Subsequently, we perform inter-group and cross-group information fusion using pointwise convolution to obtain the output $\mathbf{F_o} \in \mathbb{R}^{H \times W \times C}$, achieving a lightweight and efficient aggregation effect.

$$\mathbf{h}_j = \mathbf{W}_{inner}([\mathbf{a}^j_1, \mathbf{a}^j_2, \mathbf{a}^j_3, \mathbf{a}^j_4]), \tag{6}$$

$$\mathbf{F_o} = \delta(\mathcal{B}(\mathbf{W}_{outer}([\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_j]))), \tag{7}$$

where $\mathbf{W}_{inner}$ and $\mathbf{W}_{outer}$ are the weight matrices used in pointwise convolution. Here, $\mathbf{a}^j_i$ represents the j-th channel of the i-th head, while $\mathbf{h}_j$ denotes the j-th group of features. We have $i \in 1, 2, 3, 4$ and $j \in 1, 2, ..., \frac{C}{4}$. The functions $\delta()$ and $\mathcal{B}()$ correspond to rectified linear units (*ReLU*) and batch normalization (*BN*), respectively.

D. Loss design

As depicted in Fig.1, we employed a deep supervision strategy to further resolve the issue of small objects being lost during downsampling. The loss at each scale comprises binary cross-entropy loss and Intersection over union loss and is defined as follows:

$$l_i = Bce(y, \hat{y}) + Iou(y, \hat{y}), \quad \mathcal{L} = \sum_{i=0}^{5} \lambda_i \cdot l_i, \tag{8}$$

where $(l_i)^5_{i=0}$ represents the losses at multiple scales, $\hat{y}$ is the ground truth mask, and y is the predicted mask. The loss

TABLE I
ABLATION STUDY ON THE SIRST DATASET IN IoU(%) AND nIoU(%). HERE ✓ MEANS THAT THIS COMPONENT IS APPLIED. NOTE THAT OUR BASELINE (BAS.).

Bas.	PPA	DASI	MDCR	SIRST IoU	nIoU
✓				71.2	74.4
✓	✓			75.3	76.9
✓	✓	✓		77.9	76.1
✓	✓	✓	✓	**80.1**	**78.3**

TABLE II
COMPARATIVE EVALUATION ON THE SIRST DATASET. WE REPORT METRIC IoU (%) AND nIoU (%).

Method	IoU	nIoU
Top-Hat [1]$_{Infrared\ Phys\ Techn'}$2006	5.86	25.42
LCM [2]$_{T\ Geosci\ Remote'}$2013	6.84	8.96
PSTNN [4]$_{Remote\ Sens-Basel'}$2019	39.44	47.72
IPI [3]$_{TIP'}$2013	40.48	50.95
RIPT [5]$_{J-STARS'}$2017	25.49	33.01
NIPPS [6]$_{Infrared\ Phys\ Techn'}$2016	33.16	40.91
MDvsFA [18]$_{ICCV'}$2019	56.17	59.84
SwinT [27] $_{ICCV'}$2021	70.53	69.89
ACM [21]$_{WACV'}$2021	72.45	72.15
UIUNet [22] $_{TIP'}$2022	78.25	75.15
HCFNet (Ours)	**80.09**	**78.31**

weights for each scale are defined as $[\lambda_0, \lambda_1, \lambda_2, \lambda_3, \lambda_4] = [1, 0.5, 0.25, 0.125, 0.0625]$.

IV. EXPERIMENTS

A. Datasets and Evaluation Metrics

Our methods are assessed using SIRST [21] in two standard metrics: Intersection over Union (IoU) and normalized Intersection over Union (nIoU) [21]. SIRST was partitioned into training and test sets in an 8:2 ratio during our experiments.

B. Implementation Details.

We perform experiments with HCF-Net on an NVIDIA GeForce GTX 3090 GPU. For input images of size 512×512 pixels and featuring three color channels, HCF-Net's computational cost is 93.16 GMac (Giga Multiply-Accumulate operations), comprising 15.29 million parameters. We employ the Adam optimizer for network optimization, employing a batch size of 4 and training the model 300 epochs.

C. Ablation and Comparison

This section introduces ablative experiments and comparative experiments conducted on the SIRST dataset. Firstly, as shown in Table I, we use U-Net as a baseline and systematically introduce different modules to demonstrate their effectiveness. Secondly, as indicated in Table II, our proposed method achieves outstanding performance on the SIRST dataset, with IoU and nIoU scores of 80.09% and 78.31%, respectively, significantly surpassing other methods. Finally, Fig. 5 presents visual results for various methods. In the first row, it can be observed that our method accurately detects more objects with a meager false-positive rate. The second row demonstrates that our method can still precisely locate objects in complex backgrounds. Finally, the last row indicates that our method provides a more detailed description of shape and texture features.

V. CONCLUSION

In this paper, we address two challenges in infrared small object detection: small object loss and background clutter. To tackle these challenges, we propose HCF-Net, which incorporates multiple practical modules that significantly enhance small object detection performance. Extensive experiments have demonstrated the superiority of HCF-Net, outperforming traditional segmentation and deep learning models. This model is poised to be crucial in infrared small object detection.

REFERENCES

[1] Ming Zeng, Jian xun Li, and Zhang xiao Peng, "The design of top-hat morphological filter and application to infrared target detection," *Infrared Physics & Technology*, vol. 48, pp. 67–76, 2006.

[2] CL 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*, vol. 52, no. 1, pp. 574–581, 2013.

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

[4] Landan Zhang and Zhenming Peng, "Infrared small target detection based on partial sum of the tensor nuclear norm," *Remote Sensing*, vol. 11, no. 4, pp. 382, 2019.

[5] 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*, vol. 10, no. 8, pp. 3752–3767, 2017.

[6] Yimian Dai, Yiquan Wu, and Yu Song, "Infrared small target and background separation via column-wise weighted robust principal component analysis," *Infrared Physics & Technology*, vol. 77, pp. 421–430, 2016.

[7] Rongtao Xu, Changwei Wang, Jiguang Zhang, Shibiao Xu, Weiliang Meng, and Xiaopeng Zhang, "Rssformer: Foreground saliency enhancement for remote sensing land-cover segmentation," *IEEE Transactions on Image Processing*, vol. 32, pp. 1052–1064, 2023.

[8] Rongtao Xu, Ye Li, Changwei Wang, Shibiao Xu, Weiliang Meng, and Xiaopeng Zhang, "Instance segmentation of biological images using graph convolutional network," *Engineering Applications of Artificial Intelligence*, vol. 110, pp. 104739, 2022.

[9] Changwei Wang, Rongtao Xu, Shibiao Xu, Weiliang Meng, and Xiaopeng Zhang, "Cndesc: Cross normalization for local descriptors learning," *IEEE Transactions on Multimedia*, 2022.

[10] Changwei Wang, Rongtao Xu, Shibiao Xu, Weiliang Meng, and Xiaopeng Zhang, "Da-net: Dual branch transformer and adaptive strip upsampling for retinal vessels segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2022, pp. 528–538.

[11] Changwei Wang, Rongtao Xu, Yuyang Zhang, Shibiao Xu, Weiliang Meng, Bin Fan, and Xiaopeng Zhang, "Mtldesc: Looking wider to describe better," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022, vol. 36, pp. 2388–2396.

[12] Changwei Wang, Rongtao Xu, Ke Lv, Shibiao Xu, Weiliang Meng, Yuyang Zhang, Bin Fan, and Xiaopeng Zhang, "Attention weighted local descriptors," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[13] Changwei Wang, Lele Xu, Rongtao Xu, Shibiao Xu, Weiliang Meng, Ruisheng Wang, and Xiaopeng Zhang, "Triple robustness augmentation local features for multi-source image registration," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 199, pp. 1–14, 2023.

[14] Rongtao Xu, Changwei Wang, Jiaxi Sun, Shibiao Xu, Weiliang Meng, and Xiaopeng Zhang, "Self correspondence distillation for end-to-end weakly-supervised semantic segmentation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023, vol. 37, pp. 3045–3053.

[15] WanTing Wang, Hanlin Qin, Wenxiong Cheng, Chunmei Wang, Hanbing Leng, and Huixin Zhou, "Small target detection in infrared image using convolutional neural networks," in *AOPC 2017: Optical Sensing and Imaging Technology and Applications*. SPIE, 2017, vol. 10462, pp. 1335–1340.

[16] LIN Liangkui, Wang Shaoyou, and Tang Zhongxing, "Using deep learning to detect small targets in infrared oversampling images," *Journal of Systems Engineering and Electronics*, vol. 29, no. 5, pp. 947–952, 2018.

[17] 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.

[18] Huan Wang, Luping Zhou, and Lei Wang, "Miss detection vs. false alarm: Adversarial learning for small object segmentation in infrared images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 8509–8518.

[19] Nasser M Nasrabadi, "Deeptarget: An automatic target recognition using deep convolutional neural networks," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 55, no. 6, pp. 2687–2697, 2019.

[20] Tianfang Zhang, Siying Cao, Tian Pu, and Zhenming Peng, "Agpcnet: Attention-guided pyramid context networks for infrared small target detection," *arXiv preprint arXiv:2111.03580*, 2021.

[21] Yimian Dai, Yiquan Wu, Fei Zhou, and Kobus Barnard, "Asymmetric contextual modulation for infrared small target detection," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 950–959.

[22] Xin Wu, Danfeng Hong, and Jocelyn Chanussot, "Uiu-net: U-net in u-net for infrared small object detection," *IEEE Transactions on Image Processing*, vol. 32, pp. 364–376, 2022.

[23] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[24] Baifeng Shi, Siyu Gai, Trevor Darrell, and Xin Wang, "Refocusing is key to transfer learning," *arXiv preprint arXiv:2305.15542*, 2023.

[25] Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu, "Eca-net: Efficient channel attention for deep convolutional neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11534–11542.

[26] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[27] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10012–10022.