Differentiated Attention Guided Network Over Hierarchical and Aggregated Features for Intelligent UAV Surveillance

Houzhang Fang, *Member, IEEE*, Zikai Liao, Xuhua Wang, Yi Chang, *Member, IEEE*, and Luxin Yan, *Member, IEEE*

***Abstract*—Intelligent unmanned aerial vehicle (UAV) surveillance based on infrared imaging has wide applications in the anti-UAV system for protecting urban security and aerial safety. However, weak target features and complex background distraction pose great challenges for the accurate detection of UAVs. To address this issue, we propose a novel differentiated attention guided network to adaptively strengthen the discriminative features between UAV targets and complex background. First, a novel spatial-aware channel attention (SCA) is introduced into deep layers via preserving critical spatial features and leveraging channel interdependencies to focus on the large-scale targets. The channel-modulated deformable spatial attention is introduced into shallow layers via refining channel context and dynamically perceiving the spatial features for focusing on the small-scale targets. A combination of the above two attention mechanisms is employed in intermediate layers of the network for concentrating on the medium-scale targets. Then, we embed a feature aggregator at the detection branches to guide the information exchange of high-level feature maps and low-level feature maps with a bottom-up context modulation, and integrate an SCA at the end to further boost the distinctive feature representation for task-awareness. The above design can adaptively enhance multiscale UAV target features and suppress complex background interferences, leading to better detection performance, especially for small targets. Extensive experiments on real infrared UAV datasets reveal that the proposed method outperforms the baseline object detectors by a large margin, validating its feasibility in real-world infrared UAV detection. The source code can be found at https://github.com/KALEIDOSCOPEIP/DAGNet.**

***Index Terms*—Attention mechanism, infrared target detection, network transformation, real-time UAV surveillance, unmanned aerial vehicles (UAVs).**

Manuscript received 25 August 2022; revised 9 November 2022; accepted 25 December 2022. Date of publication 17 January 2023; date of current version 24 July 2023. This work was supported in part by the Open Research Fund of the National Key Laboratory of Science and Technology on Multispectral Information Processing under Grant 6142113220303, in part by the National Natural Science Foundation of China under Grant 41501371, Grant 61971460, and Grant 62101294, and in part by the Natural Science Foundation of Shaanxi Province under Grant 2018JM4036. Paper no. TII-22-3653. *(Houzhang Fang and Zikai Liao contributed equally to this work.) (Corresponding authors: Houzhang Fang; Xuhua Wang.)*

Houzhang Fang, Zikai Liao, and Xuhua Wang are with the School of Computer Science and Technology, Xidian University, Xi'an 710071, China (e-mail: houzhangfang@gmail.com; lzk773629528@163.com; daleiwxh@163.com).

Yi Chang and Luxin Yan are with the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430074, China (e-mail: owuchangyuo@gmail.com; yanluxin@hust.edu.cn).

This article has supplementary material provided by the authors and color versions of one or more figures available at https://doi.org/10.1109/TII.2022.3232777.

Digital Object Identifier 10.1109/TII.2022.3232777

I. Introduction

RECENTLY, unmanned aerial vehicles (UAVs) have been commonly used in civil and military fields for its high mobility, small size, and low cost, but their great threats toward aerial security as well as public safety raises serious concerns and calls for necessary containment measures. Due to the round-the-clock working capability and long monitoring distance, thermal infrared imaging has become one of the most appropriate and important sensing technologies for constant surveillance of UAVs [1], [2], [3], [4] in the anti-UAV system, as illustrated in Fig. 1. However, there exist the following challenges in the task of infrared UAV surveillance.

1) *Weak target features*: Compared to optical images, UAV targets usually occupy only a small portion of the whole image and lack conspicuous features, such as texture and color, making it difficult to identify UAV targets.
2) *Varying scale of UAV targets*: The dynamic changes of UAVs during the motion lead to multiscale characteristics in the observed UAV images. In contrast, the effective detection of small-scale UAV targets is of more significance, since long-range monitoring provides adequate time for early alarm and countermeasures.
3) *Complex background interference*: UAV targets can fly in backgrounds such as strong and bright clouds, trees and bushes, compounds, etc. The targets can be easily interfered or submerged by these backgrounds, leading to weak and dim UAV targets.

In this article, we aim at detecting multiscale infrared UAV targets in single frame under various complex background conditions.

In the past few years, many works have been dedicated to solving the problem of infrared target detection, which can be divided into two categories: 1) model-driven methods; 2) data-driven methods. The former depends largely on handcrafted features of infrared targets, such as partial sum of tensor

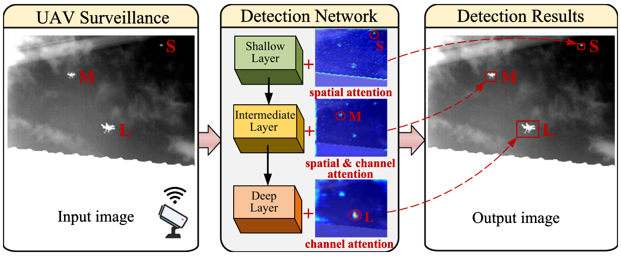


Fig. 1. Illustration of intelligent UAV surveillance. The infrared sensors capture images and transfer them to the proposed detection network. The detection results will be further transferred to interdict UAVs. S/M/L denote small/medium/large scale UAV targets, respectively. The heatmaps on the right side of the different network layers show DAGNet's close attention to infrared UAV targets of different scales.

nuclear norm (PSTNN) [5] and double-neighborhood gradient method (DNGM) [6]. These methods are comparatively easy to implement, but they usually yield poor detection results under complex backgrounds or when the target scales changes frequently. The latter, based on convolutional neural networks (CNNs), can adaptively learn the features from sample images. Some state-of-the-art object detection baseline methods (e.g., Faster region-convolutional neural network (R-CNN) [7], Cascade R-CNN [8], single shot detector (SSD) [9], you only look once v5 (YOLOv5) [10], RetinaNet [11], EfficientDet [12]) were proposed to deal with multiscale object detection issues. However, these methods are tailored for high resolution optical images with rich details, and it is difficult for them to achieve high detection performance for infrared images with low resolution and coarser details. Meanwhile, many recent studies attempted to design CNN-based models dedicated to infrared target detection. Hou et al. [13] proposed robust infrared small target detection network (RISTDNet) to detect infrared UAV targets with few obvious features. Fang et al. [2] proposed dilated residual U-Net (DRUNet) incorporating global and local properties using dilated residual networks. Dai et al. [14] proposed the asymmetric contextual modulation (ACM), utilizing context with attention to capture features of small infrared targets. These methods are more compatible and effective for infrared UAV target detection than baseline object detectors, but they still fail to take full advantage of infrared UAV target features.

To better tackle the difficulties within infrared UAV target detection mentioned above, we propose a differentiated attention guided network (DAGNet) to adaptively capture and enhance the discriminative features between multiscale UAV targets and the complex background. Concretely, our DAGNet possesses the following characteristics:

1) **Careful arrangement of attention mechanisms in backbone**: The attention mechanism is intrinsically capable of selecting the discriminative hierarchical features in backbone and automatically suppressing the background. However, the contribution of different hierarchical features to the detection of targets at different scales is different, so it is crucial to integrate suitable attention mechanisms into the network. In this study, we embed a novel channel-modulated deformable spatial attention (CDSA) in deep layers for focusing on the large-scale targets, spatial-aware channel attention (SCA) in low-level layers to focus on the small-scale targets, and dual-dimension combined attention (DCA) as a combination of the CDSA and SCA at stages in between for concentrating on the medium-scale targets.
2) **Spatial-aware Channel Attention**: Channel attention is used to adaptively weight features distributed on the channel and can be viewed as a task-oriented selection process for objects. Hu et al. [15] proposed the channel attention squeeze-and-excitation network (SENet) using global average pooling and Woo et al. [16] proposed an improved channel attention called convolutional block attention module (CBAM) with global average and max pooling operations. However, these attention mechanisms utilize only one global value to represent the whole spatial properties of one channel feature map, which severely weaken the feature representation ability of infrared UAV targets, and thus might not correctly represent the true significance of each channel [14]. To address this problem, we design the SCA. It adopts both $n \times n$ max pooling and $n \times n$ convolution to aggregate spatial information in order to better represent the spatial properties for one channel feature map. By this, features of infrared UAV targets can effectively reflect on channel weights and thus be correctly been strengthened, which leads to higher detection performance.
3) **Channel-modulated Deformable Spatial Attention**: Spatial attention is used to weight features distributed on the spatial dimension, where it adaptively highlights the target pixel regions and suppresses background contents. Woo et al. [16] proposed an spatial attention mechanism with two channel pooling operations, which, however, may also damage the valid representation of channel-wise features within one spatial element, and is unable to model geometric transformation. In this work, we propose a CDSA, where we adopt point-wise convolution for channel information exchange, and subsequent deformable convolution [17] to dynamically adjust the receptive field in light of UAV target scales and poses. Less compression of channel information and dynamic spatial information fulfills a more effective spatial information representation, which helps discriminate the real infrared UAV target especially in complex background conditions, while being more sensitive toward the location of the target.
4) **FA at the feature pyramid network**: In the original feature pyramid network (FPN) [18], the high-level feature map will be fused with the neighboring low-level feature map in a top-down manner for the purpose of multiscale detection. However, infrared UAV targets might be submerged by the background in deep layers, and high-level features from deep layers cannot provide accurate semantic information about the targets. Therefore, the feature aggregator (FA) is proposed.

   First, a channel attention modulation module is introduced to guide the refinement of the high-level features with spatial details of the neighboring low-level feature maps in a bottom-up manner. Then, we add one SCA at the end to increase task awareness. Compared to the original fusion strategy, our FA is capable of not only aggregating low-level and high-level features for a valid multiscale infrared UAV detection, but also once again highlighting the target as well as suppressing the backgrounds.

In summary, our main contributions are summarized as follows.

1) For infrared-imaged UAV detection tasks under complex background, we propose the DAGNet based on the differentiated attention feature enhancement mechanism (DAFEM). It arranges different types of attention mechanisms in the network according to the contribution of different hierarchical target features to the detection of targets at different scales. Such arrangement can adaptively strengthen the hierarchical multiscale UAV targets features as well as suppress the background content, thus boosting detection performance.
2) In order to preserve more multiscale (especially small-scale) infrared UAV target features within the network for finer feature representations, we propose the CDSA. For better UAV target identification against distracted targets as well as better target localization, we devise the SCA. The DCA is proposed to enhance intermediate features that contain abundant spatial and channel features. These attention mechanisms can adaptively characterize the distinctive features of multiscale infrared UAV targets and the dynamic scenes, making the detection model more robust to complex backgrounds.
3) In order to improve multiscale detection performance for infrared UAV targets, we propose the FA. The incorporation of appropriate attention mechanisms strengthens the network's target localization and classification awareness, by incorporating spatial details from low-level feature maps to high-level feature maps with a context modulation module, and adaptively selecting crucial task-related features. This can further improve the performance of detecting multiscale infrared UAV targets.

The rest of this article is organized as follows. Section II briefs the related works. Section III specifies the proposed network. Section IV shows the experimental results. Finally, Section V concludes this article.

II. Related Work

A. Anchor-Based Object Detection

Anchor-based object detection methods can generally be categorized into the two-stage and single-stage methods. The former ones (e.g., Faster R-CNN [7]) often exploit a region proposal network (RPN) to generate region proposals in the first stage, and then forwards them to the second stage for further classification and localization. The latter ones (e.g., SSD [9]) manage to straightly obtain class probabilities and object bounding boxes results based on global image classification and regression minus RPN. But neither can detect objects in an accurate and efficient manner. Methods such as RefineDet [19] were proposed to address this imbalance, but still limited in accuracy for their architectural design. In this article, we employ the attention mechanisms and network architecture transformation to build a robust and efficient network for infrared UAV detection.

B. Attention Mechanisms

Attention mechanisms, such as channel attention and spatial attention, are widely adopted to enhance critical features in the network. Hu et al. [15] introduced the SENet to obtain the global channel-wise correlative response and weight the feature map accordingly. Woo et al. [16] then proposed the CBAM, coupling channel-wise attention with space-wise attention to enhance features from two different aspects. Although previous studies [20], [21], [22] embedded attention mechanisms, their approaches lacked careful analysis of the feature representations of infrared targets on different output feature maps from different layers, thus still not capable enough of dealing with problems like the complex background in the UAV detection. Our method takes advantage of multiple different attention mechanisms that are important for multiscale infrared UAV target detection.

C. Infrared Small Target Detection Methods

Infrared small target detection methods can be categorized as model-driven traditional ones and data-driven deep-learning ones. The former relies on handcrafted feature extractor (FE) to discriminate infrared small target from different backgrounds (*e.g.*, PSTNN [5] and DNGM [6]). Despite some performance achievements, these methods are highly susceptible to varying detection circumstances like complex background (*e.g.*, strong clouds and trees), and suffer from strenuous hyper-parameters tuning, thus not compatible with real-scene infrared UAV detection tasks. The latter learns to detect with a network architecture and a large amount of data (*e.g.*, RISTDNet [13] and ACM [14]). Although they perform better than model-driven methods, they face challenges in accurately extracting and representing multiscale UAV target features under complex dynamic scenes, thus many false alarms or missed detections. Our method makes use of different attention mechanisms to enhance target features dynamically and adaptively, thus achieving better results than other state-of-the-art approaches (i.e., DNGM [6] and ACM [14]).

III. Methodology

In this section, we describe the details of the overall proposed network design, covering the transformable convolutional object detection network and the differentiated attention mechanisms.

A. Transformable Convolutional Object Detection Network

We design the transformable convolutional object detection network as illustrated in Fig. 2, in order to establish a single-path backbone network for boosting inference speed. It consists of the following three main parts: 1) the FE; 2) FA; 3) the detection head (DH).

1) Feature Extractor (FE): FE is the backbone network of the proposed DAGNet, which is used to effectively and efficiently extract features of infrared UAV targets, inspired by RepVGG [23]. We arrange five network stages in the FE, which

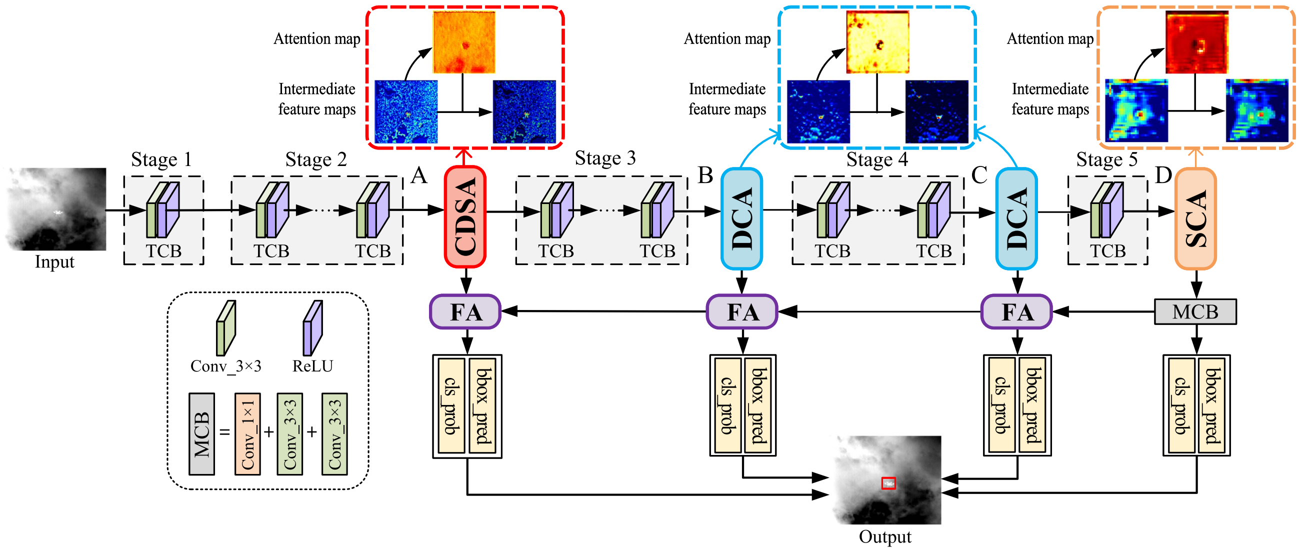


Fig. 2. Overview of the architecture of DAGNet. The first tier (composed of five network stages with TCBs) is the FE. The CDSA, DCA, and SCA are the proposed attention mechanisms, which stand for channel-modulated deformable spatial attention, dual-dimension combined attention, and spatial-aware channel attention, respectively. The second tier is the FAs. MCB is the multicolvolution block consisting of one 1 × 1 and two 3 × 3 consecutive convolutions. The third tier is the detection head (bbox_pred and cls_prob). The dashed boxes extended from CDSA, DCA, and SCA indicate the effect of the corresponding attention mechanism, where the red-orange images denote the attention maps, and the left/right bluish images are the intermediate feature maps before/after the influence of the attention.

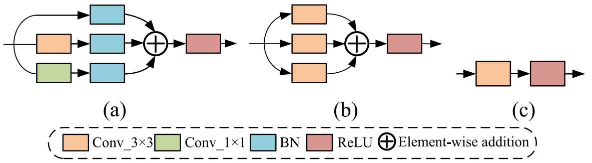


Fig. 3. Illustration of the TCB. (a) Original TCB in the training phase. (b) Intermediate form of the TCB. (c) Final TCB structure in the inference phase.

consists of 1, 4, 6, 16, and 1 transformable convolution blocks (TCBs), respectively. The specification of FE architecture can be found in the supplementary material.

The TCB is the main component of FE. The training-phase and the inference-phase TCB are given in Fig. 3(a) and (c). In the training phase, the input feature map will be fed into three convolution paths, i.e., 3 × 3 convolutional path, 1 × 1 convolutional path, and identity path. Then, the results from three paths are added together and go through an activation layer. In the inference phase, the TCB will transform from a multipath form to a single-path one in order to increase the inference speed, where it only contains one 3 × 3 convolution layer and an activation layer. This transformation of TCB can not only learn multireceptive field feature representations in the training phase, but also considerably reduce model complexity in the inference phase, which helps to achieve a better tradeoff between detection performance and efficiency [23].

2) Feature Aggregator: To realize multiscale infrared UAV target detection, we take the output feature maps from stage 2 to 5 from the backbone, and fuse them gradually in a top-down way. In the backbone network, low-level feature maps contain finer structural detail features conducive to localization and smaller targets detection; deep feature maps have richer semantic features beneficial to classification and larger targets detection. Small UAV targets in high-level features can be easily overwhelmed by the complex background, thus a direct top-down feature fusion is hard to help highlight the details of infrared UAV targets. The original FPN [shown in Fig. 4(a)] only considered a vanilla operation of fusing the shallow feature map with the neighboring upsampled deep feature map, which lacks a comprehensive inspection of the importance of both feature maps. While in ACM [shown in Fig. 4(b)], the two weighting branches exchange feature importance information with both shallow and deep feature maps, where semantic and structural information complement each other. But there is no relatively valid feature recalibration operation after feature

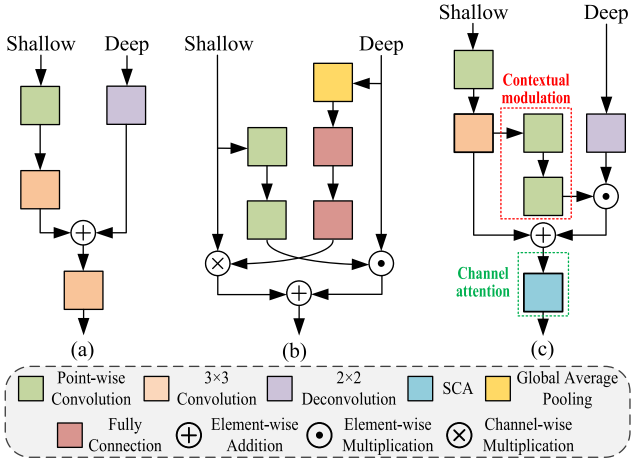


Fig. 4. Comparison of different feature fusion strategies. (a)–(c) are the original FPN [18], the ACM [14], and the proposed FA, respectively. Activation operations are omitted.

fusion, since features might misalign after being weighted. To deal with this issue, we propose to propagate low-level spatial details by multiplying an element-wise attention map from low-level feature maps into deep feature maps, and employ a proposed channel attention SCA after feature fusion. This constitutes a bottom-up contextual modulation path that is capable of complementing spatial properties in deep feature maps, and thus increases the network's awareness of the target location, which further facilitates the detection of infrared UAV targets.

The proposed FA is presented in Fig. 4(c). We firstly use one point-wise convolution and two 3 × 3 convolutions to calibrate features from low-level feature map as well as to fix the channel to 256. And we use deconvolution to upsample the deep feature map to its 2× spatial size. Then two consecutive point-wise convolutions are used on the low-level feature map to generate the channel attention map. The first point-wise convolution compresses the channel to 64, while the second one restores it to 256, which aggregates channel contextual information. After a sigmoid layer, this attention map is multiplied onto the deep layer, and then the low-level and the deep feature maps are added together. Finally, the fused feature map will go through an SCA to recalibrate the features as well as to increase the model's task awareness. Given the input low-level feature map S and the deep feature map D, the fusion operation can be formulated as follows:

$$\mathrm{S}' = \mathrm{Conv\_3 \times 3(Conv\_3 \times 3(PWConv(S)))} \tag{1}$$

$$\mathrm{S}'_{\mathrm{map}} = \sigma(\mathrm{PWConv(PWConv(S'))})) \tag{2}$$

$$\mathrm{D}' = \mathrm{S}'_{\mathrm{map}} \odot \mathrm{DeConv\_2 \times 2(S)} \tag{3}$$

$$\mathrm{O} = \mathrm{SCA(D' + S')} \tag{4}$$

where PWConv denotes the point-wise convolution, DeConv_2 × 2 represents a deconvolution with a kernel of 2 × 2, σ is the sigmoid function, and $\odot$ means element-wise multiplication. S′ and D′ are two feature maps before addition fusion, $\mathrm{S}'_{\mathrm{map}}$ is the spatial attention map, and O is the output fused feature map.

3) DH and Loss Function: The detection head classifies and locates potential targets, and finally yields the detection results. As shown in Fig. 2, we construct four detection branches for target classification and localization. We also apply an anchor calibration strategy that refines the prior boxes in the network, where these boxes are coarsely calibrated before feature aggregation according to the output feature representations, and are further fed to the final bounding boxes regression for finer localization. This strategy alleviates the one-shot-regression disadvantage introduced by the single-stage methods [19].

The loss function is divided into two parts: 1) the first part derives from the first-stage anchor adjustment, which coarsely decides anchors' object attribute (foreground or background) and their locations and sizes; 2) the second part performs multiclass categorization and precise localization assisted by aforementioned calibration results. The overall loss function is given as follows:

$$\mathfrak{L} = \frac{1}{N_{fs}}\left\{\sum_i \mathfrak{L}_o(p_i, [gt_i^l \geq 1]) + \sum_i [gt_i^l \geq 1]\mathfrak{L}_r(x_i, gt_i^b)\right\} + \frac{1}{N_{ss}}\left\{\sum_i \mathfrak{L}_m(c_i, [gt_i^l \geq 1]) + \sum_i [gt_i^l \geq 1]\mathfrak{L}_r(t_i, gt_i^b)\right\} \tag{5}$$

where subscripts fs and ss indicate the first stage and the second stage; $\mathfrak{L}_o$, $\mathfrak{L}_m$, and $\mathfrak{L}_r$ are the losses of objectness categorization, multiclass categorization, and bounding box regression; N_{fs} and N_{ss} denote the number of positive anchors in the first stage and the second stage, respectively; i defines the index of the anchor box, gt_i^l and gt_i^b are the class label and the bounding box from ground truths, p_i is the objectness score, and c_i is the predicted class score. x_i means the coordinates of the first-stage calibrated anchors, while t_i means the final predicted coordinates of the output bounding boxes. The bracket represents the signum function with the determining condition placed inside. More details can be found in [19].

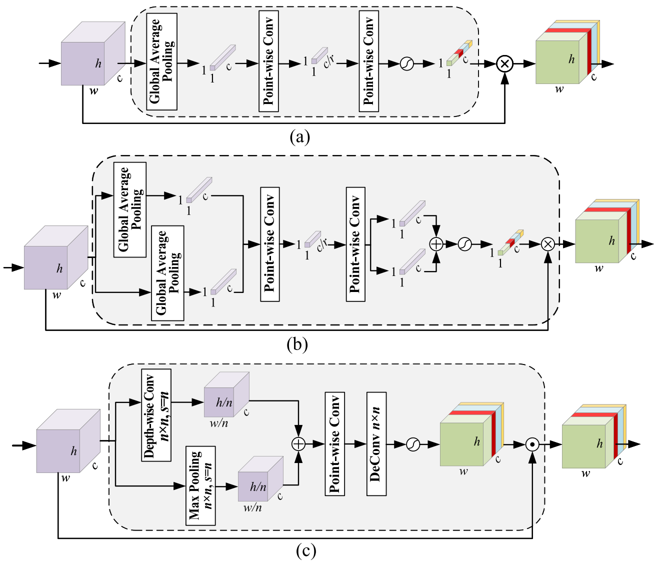
B. Differentiated Attention Feature Enhancement Mechanism

In order to highlight the features of infrared UAV targets according to the distribution of them from both spatial and channel dimensions, we propose the differentiated attention feature enhancement mechanism (DAFEM), which includes SCA, CDSA, DCA, and an arrangement of these attention mechanisms used in our backbone network according to the feature representations.

1) Spatial-Aware Channel Attention: We design SCA to sort out those UAV-related features by adaptively weighting the significance of channel-wise features. The overview of this attention is shown in Fig. 5(c), and the SENet channel attention and CBAM channel attention are shown in Fig. 5(a) and (b), respectively.

Assume the input feature map is $X \in \mathbb{R}^{C\times H\times W}$, where C, H, W denote the number of channels, height and width, respectively. Unlike SENet [15] and CBAM [16] that use global pooling operations to obtain global spatial responses for each channel, we instead use one max pooling operation with a kernel size of n × n and a convolution operation with the same kernel to do so. This results in two tensors $X_1 \in \mathbb{R}^{C\times \frac{H}{n}\times \frac{W}{n}}$ and $X_2 \in \mathbb{R}^{C\times \frac{H}{n}\times \frac{W}{n}}$, where each element in one channel represents a specific n × n region of the original input feature map. Considering the size of infrared UAV targets in our dataset is at least 4 × 4 pixels, we set $n = 4$ (Ablation studies can be found in the supplementary material). Compared to SENet and CBAM using global pooling operations, the spatial properties of infrared UAV targets are more likely to be retained after pooling, which is greatly conducive to infrared UAV target detection since their features can be easily submerged by the background in global pooling operations. Besides, the upcoming channel attention is able to more finely determine which channel is more closely related to the infrared UAV target.

After spatial shrinkage, X_1 and X_2 are added together as X'. We then use a point-wise convolution to compute the overall



(a)

(b)

(c)

Fig. 5. Comparison of different channel attention mechanisms. (a) Original SENet channel attention [15]. (b) CBAM channel attention [16]. (c) Proposed SCA. The main difference is that it is able to preserve more spatial properties for finer determination of which channel is more related to UAV targets, so as to increase UAV target classification performance against distracted objects.

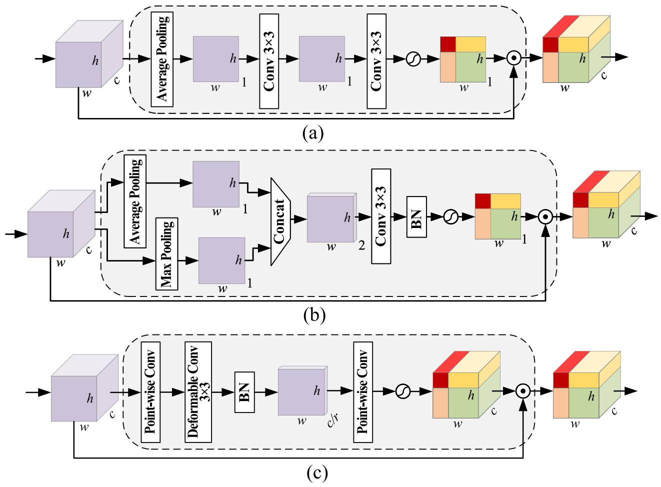
significance of each channel, and then a deconvolution operation to restore the spatial dimension from $(C, \frac{H}{n}, \frac{W}{n})$ to (C, H, W). After going through a sigmoid layer to convert statistics between 0 and 1, the final channel attention map $X_{\text{cmap}} \in \mathbb{R}^{C\times H\times W}$ is achieved, and it will be multiplied onto the input feature map element-wisely. This self-gated mechanism can be formulated as

$$X_1 = \text{MP\_}n \times n, X_2 = \text{Conv\_}n \times n_{(r)}(X) \tag{6}$$

$$Y_c(X) = X \odot \sigma(\text{DeConv\_}2 \times 2(\text{PWConv}(X_1 + X_2))) \tag{7}$$

where X is the input feature map, X_1 and X_2 are the shrunk intermediate feature maps using max pooling and convolution, MP means max pooling, σ denotes sigmoid function, and $Y_c(X)$ is the output channel-wisely augmented feature map.

By using SCA, the features of infrared UAV targets across channel dimension will be given scores close to 1, while other background-related channels will be given scores close to 0. Therefore, infrared UAV target features will be enhanced while backgrounds contents will be suppressed channel-wisely, thus improving the network's task awareness to discriminate between real targets and complex backgrounds. As for SENet channel attention [shown in Fig. 5(a)] and CBAM channel attention [shown in Fig. 5(b)], they both resort to compressing spatiality to only a 1 × 1 value, since they are effective for capturing more global properties of comparatively big objects. However, infrared UAV targets are typically small in size and dim in illuminance, with less global properties but more local properties. Our proposed SCA [shown in Fig. 5(a)] can capture more local properties and preserve more spatial details for finer channel importance determination, thus improving UAV target classification performance against distracted objects.



(a)

(b)

(c)

Fig. 6. Comparison of different spatial attention mechanisms. (a) SENet-like spatial attention [15]. (b) CBAM spatial attention [16]; (c) Proposed CDSA. The main difference is CDSA compresses less information, which results to a better network awareness toward the location of UAV targets in spatial dimension. This advantage can lead to better localization of targets, and can preserve more features in the network.

2) Channel-Modulated Deformable Spatial Attention:

Space-wise features are equally significant to channel-wise features, as they contain rich information of targets' locations. Thus, we design CDSA to augment spatial properties on spatial dimension to highlight the infrared UAV targets as well as repressing backgrounds, as is shown in Fig. 6(c). We also provide a comparison with the SENet-like spatial attention and the CBAM spatial attention in Fig. 6(a) and (b), respectively.

Given the same input feature map $X \in \mathbb{R}^{C\times H\times W}$, we first use one point-wise convolution to modulate channel contextual information as well as reduce channel dimensionality to $\frac{C}{r}$ (In our method we set $r = 4$, ablation study about this hyperparameter can be found in the supplementary material). Thereafter, we use a 3 × 3 deformable convolution and a batch normalization (BN) to compute spatial significance statistics. This deformable convolution can dynamically capture spatial features according to UAV target scales and poses to form a spatial attention map, which adaptively highlights target regions as well as suppressing background contents, and hence improves detection performance. We then use another point-wise convolution to restore channel dimensionality back to C, operate a sigmoid function and obtain the final spatial attention map X_{smap}. As for now, each spatial element on X_{smap} ranges from 0 to 1, where values close to 1 indicate the region is more likely to be target-related, while those close to 0 indicate otherwise. Then, X_{smap} will be element-wisely multiplied onto the input feature map. This spatial attention is formulated as

$$Y_s(X) = X \odot \sigma(\text{PWConv}(\text{BN}(\text{DeformConv}(\text{PWConv}(X))))) \tag{8}$$

where $Y_s(X)$ is the output feature map space-wisely augmented, DeformConv denotes the 3 × 3 deformable convolution, and BN means batch normalization.

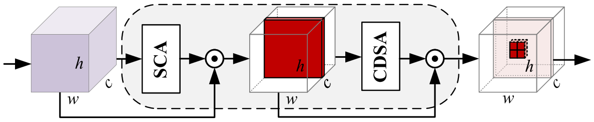


Fig. 7. Structure of the proposed DCA.

Similar to SCA, CDSA is able to assign higher scores to regions of targets and lower scores to background contents, which helps network to more precisely locate the target in the image. Furthermore, since the infrared UAV targets are comparatively small in size and dim in illuminance, igniting their spatial existence is of great priority. By using CDSA to enhance their spatial properties, features of infrared UAV target can be more effectively utilized in deep layers. In comparison with the SENet-like spatial attention [shown in Fig. 6(a)] and the CBAM spatial attention [shown in Fig. 6(b)], one evident difference is that our proposed CDSA [shown in Fig. 6(a)] does not compress channel dimensionality to 1, where severe information loss may result. In addition, the 3 × 3 deformable convolution operation is more robust to object scales and shapes. These lead to better localization of UAV targets.

3) Dual-Dimension Combined Attention: Feature maps from intermediate stages of network usually contain both spatial and semantic features. In this study, We couple SCA and CDSA together and form DCA, as shown in Fig. 7. In DCA, features are firstly augmented in channel dimension and then in spatial dimension using SCA and CDSA consecutively. Our initial thought for this structure is that we allege it is important to first find which dimensions contain rich features of infrared UAV targets, and then further search for the specific locations of the targets on those channel feature maps.

4) Arrangement of Attention Mechanisms: In SENet [15], there exists only the SE Block (channel attention) for each convolutional block throughout the backbone network; similarly, in CBAM [16], only the CBAM block (channel attention and spatial attention) is present for each convolutional block in the backbone network. However, in FPN [18], it is implied that low-level feature maps contain finer structural features conducive to localization and smaller targets detection; deep feature maps have richer semantic features beneficial to classification and larger targets detection. From this perspective, there are different hierarchical feature representations in the backbone network, suggesting that different attention mechanisms should be employed corresponding to the feature hierarchy. Since the proposed SCA is to enhance channel semantic features, CDSA to spatial structural features, and DCA to both, we follow the feature representations across our backbone network FE, and embed different attention at the end of the network stages to enhance the correspondingly rich features. We mark the four output feature maps from FE stages 2–5 as A, B, C, and D, respectively, and the arrangement of attention mechanisms is given as follows:

1) Feature map A has a relatively large spatial size which we perceive contains more fine spatial features that contribute to targets localization, so we opt to use spatial attention on it, as we embed one CDSA at the end of stage 2;
2) Feature map D has most channels that contain rich semantics. In this case, we impose channel attention on it by embedding one SCA at the end of stage 5;
3) Feature maps B and C, in our perspective, contain both relatively rich structural and semantic features. Therefore, we embed one DCA at the end of stage 3 and 4, respectively.

With DAFEM, low-level structural features can be effectively enhanced by CDSA, which ensures the features of infrared UAV targets to be passed down to deeper stages of the network. As a result, those features are more likely to exist in deep layers. With these retained features and the extra channel enhancement from SCA, infrared UAV targets can be more accurately classified and robust to the complex background, and thus have better detection performance.

IV. Experiments

In this section, we conduct experiments to validate the effectiveness of our method with other comparison methods. We first introduce our datasets and the experiment setups we use to train and test all methods, then present the quantitative and qualitative experiment results. Finally, we design some ablation experiments and report the results.

A. Datasets

We select nine infrared image sequences, each of which exemplifies some representative problems in anti-UAV scenarios, such as strong clouds background and motion blur. The total number of all image sequences is 75 666 and the sizes of the infrared UAV targets range from about 15 pixels to more than 200 pixels for a multiscale infrared detection task. We have meticulously labeled the targets in each sequence, and divided them into the training set and the testing set with no intersection. The division ratio is set to 9:1. Among our datasets are two open-access infrared UAV datasets (Seq. 8 [24] and Seq. 9 [25]), which we use to validate our method. Further details can be referred to the supplementary material.

B. Evaluation Metrics and Experimental Setups

We employ the following commonly used metrics to verify the performance of each method: 1) Detection precision (P); 2) recall (R); 3) F1-measure (F1). FPS, network parameters, and FLOPs are also included to evaluate the detection efficiency as well as the model complexity. Theoretically, a qualified detection method should be low in parameters and FLOPs, while high in P, R, F1, and FPS.

The proposed network is trained by an SGD optimizer with a total of 120 000 iterations, an initial learning rate of 0.001, a batch size of 24, a momentum of 0.9, and a weight decay of 0.0005. The learning rate decays at iterations 80 000 and 100 000 by a magnitude. The framework is implemented on a server with a NVIDIA GeForce RTX 3090 GPU and accelerated by CUDA 11.1. We use Python 3.9 and Pytorch 1.8.1 for software implementation.

TABLE I
QUANTITATIVE COMPARISONS OF THE PROPOSED METHOD AND OTHER METHODS

# Seq.	Metrics	Methods								
		Faster R-CNN	Cascade R-CNN	RetinaNet	EfficientDet	YOLOv5	RefineDet	DNGM	ACM	Ours
1	P	0.8103	0.8389	0.7815	0.9570	0.844	0.9271	0.4674	0.8033	**0.9844**
	R	0.7398	0.7974	0.6988	0.8849	0.826	0.9568	0.8217	0.7343	**0.9656**
	F1	0.7734	0.8176	0.7378	0.9195	0.8349	0.9417	0.5959	0.7673	**0.9749**
	FPS	19.33	13.61	29.40	21.18	**62.5**	35.59	12.91	38.12	38.64
2	P	0.9848	0.9914	0.9801	0.9710	0.988	0.9799	0.8105	0.9219	**1.0000**
	R	0.9888	0.9919	0.9696	0.9588	0.967	0.9553	0.8863	0.9571	**1.0000**
	F1	0.9868	0.9916	0.9748	0.9649	0.9774	0.9674	0.8467	0.9392	**1.0000**
	FPS	17.53	14.47	31.10	27.63	**57.14**	52.14	19.28	48.39	44.62
3	P	0.9813	0.9749	0.9110	0.8039	0.9459	0.9586	0.4199	0.8278	**0.9966**
	R	0.9255	0.9377	0.8994	0.8844	0.9772	0.9558	0.4392	0.7954	**1.0000**
	F1	0.9526	0.9559	0.9089	0.8975	0.9613	0.9572	0.4293	0.8113	**0.9983**
	FPS	18.59	14.52	31.24	27.56	**77.52**	54.68	12.84	40.14	43.99
4	P	0.9913	0.9900	0.9901	0.9842	0.9950	0.9880	0.8963	0.9258	**1.0000**
	R	0.9954	0.9967	0.9907	0.9743	**1.0000**	0.6568	0.8574	0.9338	**1.0000**
	F1	0.9933	0.9933	0.9904	0.9792	0.9975	0.7891	0.8764	0.9298	**1.0000**
	FPS	19.16	12.94	30.96	27.61	**76.34**	54.83	16.69	48.77	42.19
5	P	0.9514	0.9301	0.6917	0.8814	0.9563	0.9865	0.6528	0.8764	**0.9939**
	R	0.9073	0.9105	0.7963	0.8502	0.9488	0.9564	0.7347	0.9594	**0.9922**
	F1	0.9288	0.9202	0.7403	0.8655	0.9525	0.9758	0.6919	0.9160	**0.9925**
	FPS	16.66	12.18	30.93	26.61	**76.92**	52.78	13.10	45.82	43.97
6	P	0.8696	0.9002	0.8379	0.8140	0.9162	0.9798	0.3817	0.9297	**0.9931**
	R	0.8803	0.8834	0.7900	0.7953	0.8934	0.9008	0.5964	0.8831	**0.9628**
	F1	0.8749	0.8917	0.8132	0.8045	0.9047	0.9674	0.4655	0.9058	**0.9777**
	FPS	18.82	13.48	29.48	22.81	**77.52**	50.79	10.65	40.16	42.56
7	P	0.8998	0.9157	0.7398	0.8906	0.9459	0.9797	0.4127	0.9251	**0.9963**
	R	0.8691	0.9103	0.7581	0.7928	0.9445	0.9669	0.3557	0.7538	**0.9797**
	F1	0.8842	0.9129	0.7488	0.8389	0.9452	0.9733	0.3821	0.8037	**0.9879**
	FPS	16.48	12.51	29.15	26.88	**76.34**	49.13	16.11	39.88	40.74
8	P	0.8626	0.8298	0.7181	0.8501	0.9222	0.7999	0.4951	0.8345	**0.8979**
	R	0.7949	0.8103	0.6957	0.8293	0.8600	0.8054	0.3371	0.8120	**0.8854**
	F1	0.8274	0.8199	0.7067	0.8395	0.8900	0.8026	0.4011	0.8231	**0.8916**
	FPS	17.14	12.75	30.76	27.02	**76.34**	39.48	18.32	33.45	39.12
9	P	0.9997	0.9970	0.9998	0.9799	**1.0000**	0.8011	0.7674	0.9816	**1.0000**
	R	0.9982	**1.0000**	0.9992	0.9701	0.9970	0.7963	0.8511	0.9771	0.9985
	F1	0.9989	0.9985	0.9995	0.9750	0.9985	0.8008	0.7913	0.9793	**0.9992**
	FPS	19.27	15.94	31.25	33.34	**72.46**	41.11	19.98	38.89	39.12

The bold entities indicate that it is the largest in a row.

For comparison, we select Faster R-CNN [7] and Cascade R-CNN [8] as the two-stage baseline methods, and SSD [9], RefineDet [19], RetinaNet [11], EfficientDet [12], and YOLOv5 [10] as the one-stage baseline methods. We also choose infrared small target detection method for comparison, i.e., the model-driven method DNGM [6] and the data-driven method ACM [14]. Note that for the Faster R-CNN, Cascade R-CNN and RetinaNet, their bakcbone network is ResNet-50; for SSD and RefineDet, their backbone is VGG-16; for the EfficientDet, we choose its D1 architecture; For YOLOv5, we choose its yolov5l configuration. For infrared small target detection methods, we adopt their default settings. All these experiments are performed with the same hardware and software, with each method tuned to the optimal.

C. Quantitative Results

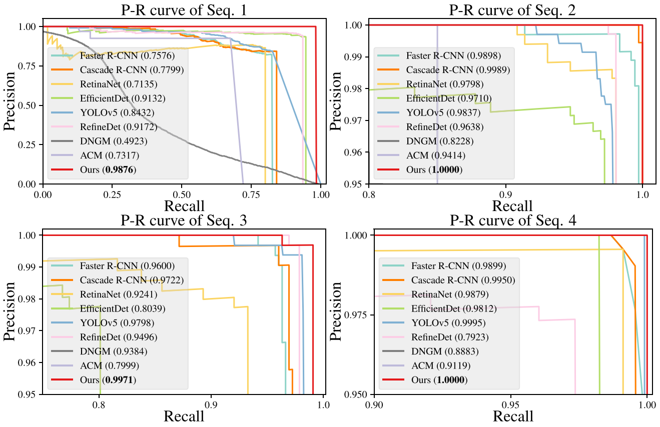


Fig. 8. P-R curves of Seq. 1, Seq. 2, Seq. 3, and Seq. 4. The area under the curve (AUC) values are placed after methods' names.

In this part, we present the quantitative comparisons for eight selected infrared image sequences of the proposed method along with eight of the aforementioned comparison methods, as shown in Table I. It can be seen that the proposed method surpasses the other ones. Especially for Seq. 2 and Seq. 4, our method has reached impressive 1.0 scores in P, R, and F1, which is a great demonstration of its ability to detect weak infrared UAV targets in complex background circumstances. Also our methods achieves a good balance between detection precision and efficiency, verifying that our method is capable of real-time UAV surveillance.

Furthermore, we present the P-R curves of eight methods for Seq. 1–Seq. 4 in Fig. 8. Generally, a larger area under the P-R curve represents a better detection performance. It can be seen

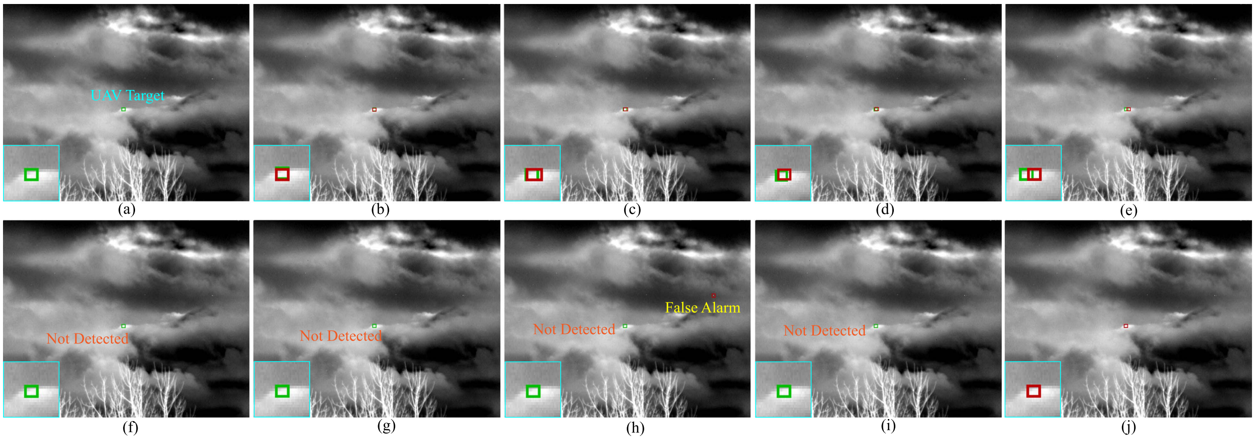


Fig. 9. Qualitative detection results of Frame No. 2403 in sequence 2. The green and red boxes indicate the ground truth and the detection box, respectively. The detection confidence scores for each method are: (b) 0.85; (c) 0.52; (d) 0.62; (e) 0.54; (f) Not detected and false alarm; (g) Not detected; (h) Not detected; (i) Not detected (j) **1.00**.

TABLE II
NETWORK PARAMETERS AND FLOPS OF EACH METHOD

Method	Parameters (M)	FLOPs (G)
Faster R-CNN	41.40	167.94
Cascade R-CNN	69.21	195.74
RetinaNet	36.33	163.29
EfficientDet	6.68	6.02
YOLOv5	46.11	107.8
RefineDet	33.91	97.12
ACM	0.52	8.38
Ours	64.20	123.98

TABLE III
EFFECTIVENESS OF SCA, CDSA, DCA, AND FA

SCA	CDSA	DCA	FA	Metrics			
				P	R	F1	FPS
-	-	-	-	0.9565	0.9327	0.9445	47.87
✓	-	-	-	0.9601	0.9497	0.9549	**53.38**
-	✓	-	-	0.9781	0.9641	0.9710	46.69
-	-	✓	-	0.9758	0.9572	0.9664	46.19
-	-	-	✓	0.9633	0.9497	0.9565	49.78
✓	✓	✓	✓	**0.9858**	**0.9686**	**0.9772**	45.59

The bold entities imply that it is the largest in a column.

that our method is consistently superior in accuracy to the other seven methods.

Moreover, we list the model parameters and the FLOPs of each method in Table II. For both model parameters and FLOPs, ACM is of the lowest and Cascade R-CNN is of the highest, while our method is relatively moderate in both metrics. This reveals another balance regarding model complexity between two-stage and single-stage methods, which is beneficial to detection tasks in industrial scenarios without having to utilize heavy computational resources.

D. Qualitative Results

Figs. 9–11 are the qualitative detection results of our method and other comparison methods for Seq. 2, Seq. 3, and Seq. 8, respectively. Due to space limit, qualitative results of Seq. 1, Seq. 4, Seq. 5, Seq. 6, Seq. 8, and Seq. 9 are placed in the supplementary material. The green bounding boxes indicate the actual locations of the targets, and the red boxes indicate the detection results. Close-ups are given at the left-bottom corner of each image for better visualization, and confidence scores are given in each caption. In ground truth images, we use cyan words "UAV Target" to indicate where the target is; in detection images, we use yellow words "False Alarm" to indicate an incorrect detection, and orange words "Not Detected" to imply a failure of detection.

Fig. 9 is a typical example of an infrared UAV target immersed in bright clouds barely exposed to be clearly seen. Fig. 10 shows the UAV target largely occluded by background trees. Fig. 11 is a common scene where the unexpected motion blur by the imaging device misting the target. The confidence scores by the comparison methods vary greatly in the experiments. Note that DNGM yields poor detection results in these three images, having missed detections and false alrms; ACM also obtains relatively poor detection results; RefineDet missed the target for Seq. 2; two-stage baseline methods perform consistently; one-stage methods, such as EfficientDet, RefineDet, and YOLOv5, fail to detect targets in Seq. 2 or Seq. 3. Inconsistency and miss detection may delay the following counter measures. In contrast, our method has successfully detected the UAV targets and consistently achieved satisfactory confidence scores and overlapped better than the others, showing the robustness of our method under different complex backgrounds.

E. Ablation Study

Several ablations are conducted to report the contributions of the proposed techniques and ideas, i.e., SCA, CDSA, DCA, and FA. More ablation studies can be found in the supplementary material.

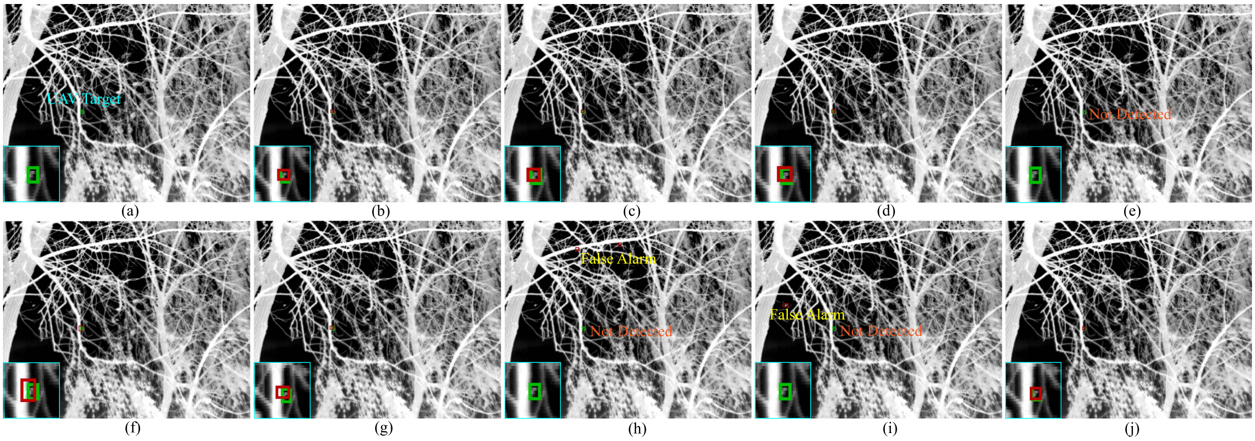


Fig. 10. Qualitative detection results of Frame No. 2715 in sequence 3. The green and red boxes indicate the ground truth and the detection box, respectively. The detection confidence scores for each method are: (b) 0.96; (c) 0.80; (d) 0.63; (e) Not detected; (f) 0.85; (g) 0.72; (h) Not detected and false alarm; (i) Not detected and false alarm; (j) **1.00**.

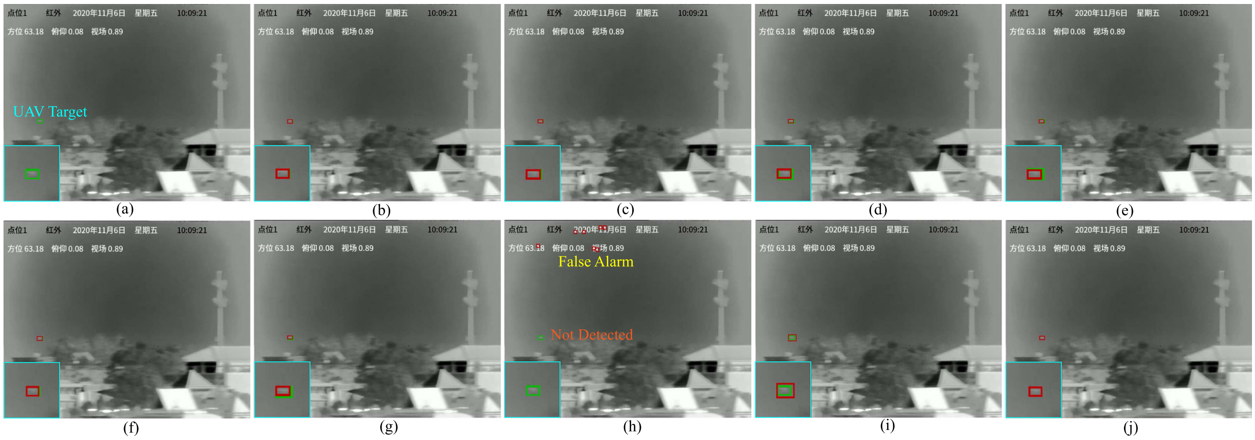

点位1 红外 2020年11月6日 星期五 10:09:21
方位 63.18 俯仰 0.08 视场 0.89
UAV Target
False Alarm
Not Detected
(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)
(i)
(j)

Fig. 11. Qualitative detection results of Frame No. 89 in sequence 8. The green and red boxes indicate the ground truth and the detection box, respectively. The detection confidence scores for each method are: (b) 0.97; (c) 0.94; (d) 0.85; (e) 0.60; (f) 0.95; (g) 0.64; (h) Not detected and false alarm; (i) 0.79; (j) **1.00**.

(1) *Integration of SCA, CDSA, DCA, and FA*. Table III shows the quantitative results of our detection method with/without the proposed SCA, CDSA, DCA, and FA. It is evident that individually they are able to improve the original detection performance, and the combination of them achieves the best. Compared to the first-row method, our method gains increments in four metrics, which validates our purpose of incorporating both in the network.

(2) *Arrangement scheme for attention mechanisms in FE*. Table IV lists the quantitative results of different arrangements in the backbone network of the proposed attention mechanisms, and the last row represents our proposed arrangement (i.e., CDSA at first stage, SCA at last stage, and DCA at intermediate stages). In order to comprehensively demonstrate the validity of our proposed arrangement, we experiment with different other arrangements of attention mechanisms and study their quantitative detection results, including embedding SCA before CDSA. It can be seen that the proposed arrangement achieves the best performance in Table IV. We can have the following extra observations.

1) From row 1 to row 5, as more SCA are gradually embedded in the shallower layers, the detection performance evidently decreases. While CDSAs are being embedded back to the network from deep layers, the performance of network again increases, specifically for metric P. This verifies that spatial property enhancement is critical to more effectively discriminate infrared UAV target against its background.
2) From row 1 and row 5–8 we can see that, with a gradual integration of SCA from the early stage, the overall detection performance decreases evidently; from row 1 to row 5 we can witness a rapid performance increase with more CDSA integration at shallow stages. We conjuncture that is due to the sufficient exploitation of target's finer features, and embedding spatial attention CDSA at early stage is relatively more capable of utilizing those features.

TABLE IV

EFFECTIVENESS OF DIFFERENT ARRANGEMENTS OF ATTENTION MECHANISMS INSIDE BACKBONE NETWORK

Arranging scheme				Metrics			
A	B	C	D	P	R	F1	FPS
CDSA	CDSA	CDSA	CDSA	0.9153	0.9206	0.9179	46.51
CDSA	CDSA	CDSA	SCA	0.9157	0.9362	0.9258	47.89
CDSA	CDSA	SCA	SCA	0.9042	0.9197	0.9119	49.91
CDSA	SCA	SCA	SCA	0.8976	0.9440	0.9202	50.18
SCA	SCA	SCA	SCA	0.8864	0.9127	0.8994	52.04
SCA	SCA	SCA	CDSA	0.8897	0.9016	0.8956	**53.98**
SCA	SCA	CDSA	CDSA	0.8976	0.9084	0.9030	50.89
SCA	CDSA	CDSA	CDSA	0.9106	0.9097	0.9104	51.38
CDSA	DCA	SCA	SCA	0.9498	0.9203	0.9348	48.40
CDSA	CDSA	DCA	SCA	0.9399	0.9416	0.9455	49.99
CDSA	DCA	DCA	SCA	**0.9858**	**0.9686**	**0.9772**	45.59

The bold entities imply that it is the largest in a column.

3) It can be observed that, from row 1 to row 5, a vanilla arrangement of only using one type of attention does not yield better results, while incorporating a mixture of attention mechanisms is more likely to achieve better results.
4) From row 9 and 10, methods incorporating DCA in intermediate stage yield better overall performance results than all previous arrangements, and row 10 is about 1% better than row 9. This not only means there is rich structural and semantic information in the intermediate stage and our DCA is able to exploit them effectively, but also reveals the significance of more exploitation of spatial properties.

In summary, we can conclude from Table IV that, if spatial features are more properly utilized, especially at low-level layers, infrared UAV targets can be detected more effectively. Furthermore, it can be seen that our arrangement, with two DCA embedded at layers in between, achieves the best performance of all, which supports our claim that layers between shallow layers and deep layers need to use combined attention mechanisms to both retain and enhance critical features for better detection performance.

V. CONCLUSION

This article proposes a novel DAGNet that is carefully designed based on the contribution of hierarchical features to different scale target detection, in order to improve the performance of detecting multiscale infrared UAV targets. We designed the channel-modulated deformable spatial attention, spatial-aware channel attention (SCA), and a combination of them for focusing on small-scale, large-scale, and medium-scale targets, respectively. We also integrated the feature aggregator into the detection branches to encode spatial details of shallow layers into deep layers as well as increase the task-awareness with SCA. Our method is verified to have good generality on real infrared UAV data, with better results than those by state-of-the-art baseline object detection methods (e.g., YOLOv5 [10]) and infrared small target detection methods (e.g., ACM [14]), and is potentially applicable in anti-UAV surveillance systems. For future research into infrared UAV target detection, we deem it is critical to focus on preserving more features of UAV targets, especially of small and dim ones, so as to prevent those features from vanishing in the network.

REFERENCES

[1] J. Xie, C. Gao, J. Wu, Z. Shi, and J. Chen, "Small low-contrast target detection: Data-driven spatiotemporal feature fusion and implementation," *IEEE Trans. Cybern.*, vol. 52, no. 11, pp. 11847–11858, Nov. 2022.

[2] H. Fang, M. Xia, G. Zhou, Y. Chang, and L. Yan, "Infrared small UAV target detection based on residual image prediction via global and local dilated residual networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[3] H. Fang, X. Wang, Z. Liao, Y. Chang, and L. Yan, "A real-time anti-distractor infrared UAV tracker with channel feature refinement module," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop*, 2021, pp. 1240–1248.

[4] H. Fang, L. Ding, L. Wang, Y. Chang, L. Yan, and J. Han, "Infrared small UAV target detection based on depthwise separable residual dense network and multiscale feature fusion," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–20, 2022.

[5] L. Zhang and Z. Peng, "Infrared small target detection based on partial sum of the tensor nuclear norm," *Remote Sens.*, vol. 11, no. 4, pp. 1–34, 2019.

[6] L. Wu, Y. Ma, F. Fan, M. Wu, and J. Huang, "A double-neighborhood gradient method for infrared small target detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 8, pp. 1476–1480, Aug. 2021.

[7] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[8] Z. Cai and N. Vasconcelos, "Cascade R-CNN: High quality object detection and instance segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 5, pp. 1483–1498, May 2021.

[9] W. Liu et al., "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.

[10] G. Jocher et al., "ultralytics/yolov5: V6.0 - YOLOv5n 'nano' models, roboflow integration, TensorFlow export, OpenCV DNN support," Oct. 2021. [Online]. Available: https://doi.org/10.5281/zenodo.5563715

[11] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, Feb. 2020.

[12] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: Scalable and efficient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10778–10787.

[13] Q. Hou, Z. Wang, F. Tan, Y. Zhao, H. Zheng, and W. Zhang, "RISTDnet: Robust infrared small target detection network," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2021.

[14] Y. Dai, Y. Wu, F. Zhou, and K. Barnard, "Asymmetric contextual modulation for infrared small target detection," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2021, pp. 950–959.

[15] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, “Squeeze-and-excitation networks,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Aug. 2020.

[16] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, “CBAM: Convolutional block attention module,” in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.

[17] J. Dai et al., “Deformable convolutional networks,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 764–773.

[18] 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.*, 2017, pp. 2117–2125.

[19] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li, “Single-shot refinement neural network for object detection,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4203–4212.

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

[21] X. Tong, B. Sun, J. Wei, Z. Zuo, and S. Su, “EAAU-Net: Enhanced asymmetric attention u-net for infrared small target detection,” *Remote Sens.*, vol. 13, no. 16, pp. 1–20, 2021.

[22] F. Chen et al., “Local patch network with global attention for infrared small target detection,” *IEEE Trans. Aerosp. Electron. Syst.*, vol. 58, no. 5, pp. 3979–3991, Oct. 2022.

[23] X. Ding, X. Zhang, N. Ma, J. Han, G. Ding, and J. Sun, “RepVGG: Making VGG-style convnets great again,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 13733–13742.

[24] F. Svanström, F. Alonso-Fernandez, and C. Englund, “A dataset for multisensor drone detection,” *Data Brief*, vol. 39, pp. 1–11, 2021.

[25] “2021 IEEE international conference on computer vision 2nd Anti-UAV workshop & challenge,” https://anti-uav.github.io/dataset, Accessed: Dec. 8, 2021.

Houzhang Fang (Member, IEEE) received the Ph.D. degree in control science and engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2014.

He is currently an Associate Professor with the School of Computer Science and Technology, Xidian University, Xi’an, China. His research interests include target recognition, target tracking, and image restoration.

Zikai Liao received the B.S. degree in software engineering from the Chongqing University of Posts and Telecommunications, Chongqing, China, in 2021. He is currently working toward the M.S. degree in software engineering with Xidian University, Xi’an, China.

His research interests include computer vision and deep learning.

Xuhua Wang received the Ph.D. degree in electronic science and technology from Air Force Engineering University, Xi’an, China, in 2013.

He is currently a Lecturer with the School of Computer Science and Technology, Xidian University, Xi’an, China. His research interests include application of UAV cluster operation and anti-UAV technology.

Yi Chang (Member, IEEE) received the B.S. degree in automation from the University of Electronic Science and Technology of China, Chengdu, China, in 2011, and the M.S. degree in pattern recognition and intelligent systems and the Ph.D. degree in control science and engineering from the Huazhong University of Science and Technology (HUST), Wuhan, China, in 2014 and 2019, respectively.

From 2014 to 2015, he was a Research Assistant with Peking University, Beijing, China. From 2019 to 2021, he was a Postdoctoral Researcher with Peng Cheng Laboratory, Shenzhen, China. He is currently an Assistant Professor with the School of Artificial Intelligence and Automation, HUST. His research interests include image processing, computer vision, and machine learning.

Luxin Yan (Member, IEEE) received the B.S. degree in electronic communication engineering and the Ph.D. degree in pattern recognition and intelligence system from the Huazhong University of Science and Technology (HUST), Wuhan, China, in 2001 and 2007, respectively.

He is currently a Professor with the School of Artificial Intelligence and Automation, HUST. His research interests include multispectral image processing, pattern recognition, and real-time embedded systems.