DANet: Multi-scale UAV Target Detection with Dynamic Feature Perception and Scale-aware Knowledge Distillation

Houzhang Fang*†
Zikai Liao*
School of Computer Science and Technology
Xidian University
Xi'an, China

Lu Wang†
Qingshan Li
School of Computer Science and Technology
Xidian University
Xi'an, China

Yi Chang
Luxin Yan
School of Artificial Intelligence and Automation
Huazhong University of Science and Technology
Wuhan, China

Xuhua Wang
School of Computer Science and Technology
Xidian University
Xi'an, China

ABSTRACT

Multi-scale infrared unmanned aerial vehicle (UAV) targets (IRUTs) detection under dynamic scenarios remains a challenging task due to weak target features, varying shapes and poses, and complex background interference. Current detection methods find it difficult to address the above issues accurately and efficiently. In this paper, we design a dynamic attentive network (DANet) incorporating a scale-adaptive feature enhancement mechanism (SaFEM) and an attention-guided cross-weighting feature aggregator (ACFA). The SaFEM adaptively adjusts the network's receptive fields at hierarchical network levels leveraging separable deformable convolution (SDC), which enhances the network's multi-scale IRUT awareness. The ACFA, modulated by two crossing attention mechanisms, strengthens structural and semantic properties on neighboring levels for the accurate representation of multi-scale IRUT features from different levels. A plug-and-play anti-distractor contrastive regularization (ADCR) is also imposed on our DANet, which enforces similarity on features of targets and distractors from a new uncompressed feature projector (UFP) to increase the network's anti-distractor ability in complex backgrounds. To further increase the multi-scale UAV detection performance of DANet while maintaining its efficiency superiority, we propose a novel scale-specific knowledge distiller (SSKD) based on a divide-and-conquer strategy. For the "divide" stage, we intendedly construct three task-oriented teachers to learn tailored knowledge for small-, medium-, and large-scale IRUTs. For the "conquer" stage, we propose a novel element-wise attentive distillation module (EADM), where we employ a pixel-wise attention mechanism to highlight teacher and student IRUT features, and incorporate IRUT-associated prior knowledge for the collaborative transfer of refined multi-scale IRUT features to our DANet. Extensive experiments on real infrared UAV datasets demonstrate that our DANet is able to detect multi-scale UAVs with a satisfactory balance between accuracy and efficiency.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision tasks**; **Object detection**; **Object recognition**.

KEYWORDS

Unmanned aerial vehicle, multi-scale infrared target detection, attention mechanism, contrastive learning, knowledge distillation

ACM Reference Format:
Houzhang Fang, Zikai Liao, Lu Wang, Qingshan Li, Yi Chang, Luxin Yan, and Xuhua Wang. 2023. DANet: Multi-scale UAV Target Detection with Dynamic Feature Perception and Scale-aware Knowledge Distillation. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29-November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3581783.3612146

1 INTRODUCTION

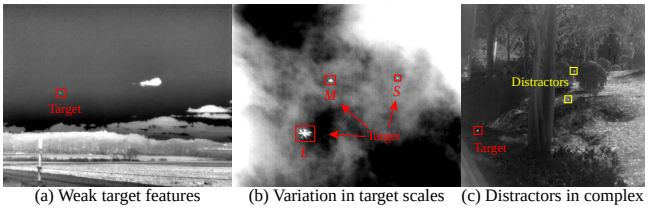
Unmanned aerial vehicles (UAVs), popularly employed in commercial and industrial applications such as aerial photography and environmental monitoring [9, 13], are becoming a great threat to aerial safety and public security. With the capability of monitoring UAVs at a long range in both day and night scenarios, the infrared thermal imaging-based UAV surveillance measure emerged as a key perception technology for UAV surveillance in anti-UAV systems. However, infrared UAV target (IRUT) detection still remains a very challenging task for weak target features (*i.e.*, dim in illuminance and small in size, thus easy to be submerged), variation in target scales (dynamic flying creates a varying IRUT scale range), and distractors in complex backgrounds (*e.g.*, birds and leaves similar to IRUTs), as shown in Fig. 1. These factors usually lead to missed detections and false alarms [9].

Many works have been dedicated to addressing the problems of infrared target detection [1, 2, 6, 7, 9, 10, 13, 14, 23, 34, 38, 41, 46–49], which can be categorized into two groups: traditional methods and

*Houzhang Fang and Zikai Liao contributed equally to this work and are co-first authors.
†Houzhang Fang and Lu Wang are the corresponding authors (houzhangfang@xidian.edu.cn, wanglu@xidian.edu.cn).

MM '23, October 29-November 3, 2023, Ottawa, ON, Canada

ACM ISBN 979-8-4007-0108-5/23/10…$15.00
https://doi.org/10.1145/3581783.3612146



(a) Weak target features (b) Variation in target scales (c) Distractors in complex backgrounds

Figure 1: Illustrations of three challenges in IRUT detection tasks. The 'L', 'M', and 'S' denote large-, medium-, and small-scale IRUTs.

deep learning (DL) -based methods. The former ones (*e.g.*, PSTNN [46], DNGM [40], WSLCM [14]), based on hand-crafted IRUT features, are comparatively light-weight but less robust to complex backgrounds [48]. The latter ones are more powerful since they can adaptively learn sophisticated feature representations of IRUTs driven by a large amount of data. Many object detection methods, such as FCOS [33] and ATSS [50], have shown satisfactory performance on general tasks of visible images, but are less competent in IRUT detection since IRUTs have relatively fewer feature details [9]. Despite some works (*e.g.*, MDvsFA [34], ACM [6], ALCNet [7], DNA-Net [23], ISNet [49], DAGNet [9]) attempted to design DL-based methods for infrared target detection, some issues remain unsolved: first, these methods mostly focus on small-scale infrared targets, with less attention on medium- and large-scale ones; second, they still suffer from poor discrimination toward distractors in complex backgrounds, which leads to false alarms; third, these methods usually have an imbalance between detection accuracy and efficiency, which is not compatible with real-time applications.

The feature weakness and varying scales of IRUTs bring great challenges for IRUT detection tasks. Recently, attention mechanisms [11, 19, 20, 35, 39] have drawn much attention for their ability to enhance target features and suppress the backgrounds, but their pooling operations compress valuable target features, which is harmful to IRUT detection [6]. Some works [6, 7, 9] adopt less feature compression, but their structures with fixed-size convolutions limit the perception of multi-scale IRUTs. To dynamically perceive targets with varying scales, studies have employed deformable convolution [5, 43, 52] in multi-scale object detection. Despite great achievements, this approach also increases model complexity and causes inference time overhead. Therefore, it is crucial to design a network mechanism for dynamic multi-scale IRUT feature perception and enhancement.

Distractors in complex backgrounds, with similar features to real IRUTs (particularly for small-scale ones), will affect detection performance as well, so learning the discriminative features of IRUTs and distractors is of great significance. Contrastive learning, enforcing similarity on positive/negative samples for discrimination, has been proven effective for distinguishing foreground/background objects [4, 18, 32, 37]. But these works did not fully consider that distractors have similar sparse distributions to those of IRUTs in the image, which are different from random backgrounds and should be specifically sampled. In addition, many works employ multi-layer perceptron (MLP) as feature projectors, which compresses features into vectors for similarity computation. Heavy feature compression may lead to losses of IRUT features (especially small-scale ones), and thus cause performance degradation.

For real-world applications, the detection model should perform both accurately and efficiently. Recently, detectors based on knowledge distillation (KD) [16] training strategy have shown impressive effectiveness, where sophisticated knowledge from a teacher model is transferred to a student model to compensate for the detection performance [24, 27, 29, 30, 42, 44, 45, 51]. There are also KD schemes that adopt multiple teachers to further improve the student's detection performance [12, 22, 25, 26, 28]. However, these methods fail to effectively transfer tailored multi-scale IRUT knowledge, and might lead to limited performance improvement. We stress that each teacher should discretely transfer knowledge to the student to increase multi-scale awareness.

To address the multi-scale IRUT detection task, in this paper we present a novel dynamic attentive network (DANet). For the backbone, we construct a four-stage multi-branch network to fully excavate feature representations of multi-scale IRUTs, and we integrate a scale-adaptive feature enhancement mechanism (SaFEM) at the end of each stage to adaptively enhance multi-scale IRUT features. A separable deformable convolution (SDC), divided into depth-wise and point-wise parts to reduce computation, is adopted to generate dynamic attention weight in the SaFEM, which alters the receptive fields according to the scale variation of IRUTs by adding a learnable weighted offset to the convolutional kernel. A grouping strategy is also used to learn richer feature representations from different subspaces in one convolution. Our SaFEM not only can effectively enhance IRUTs and suppress backgrounds, but also can dynamically perceive IRUTs of various scales without much model complexity increment. For the bottleneck, we embed an attention-guided cross-weighting feature aggregator (ACFA) in the feature fusion process. It employs two uncompressed attention branches to exchange critical fine spatial and coarse semantic properties for the accurate representation of multi-scale IRUT features from different levels.

We propose an anti-distractor contrastive regularization (ADCR), which extracts positive/negative samples from regional features of the IRUTs/distractors, and further projects them through a novel uncompressed feature projector (UFP) for the final similarity computation. ADCR is only used on our DANet during training, with no computation overhead during inference. With ADCR, the network is able to finely discriminate between IRUTs and distractors, and thus significantly reduce false alarms.

To further increase the multi-scale IRUT detection performance of our DANet without sacrificing detection speed, we present a new divide-and-conquer strategy-based scale-specific knowledge distiller (SSKD). At the "divide" stage, we use three task-oriented teacher models, *i.e.*, teachers for detecting small-, medium-, and large-scale IRUTs, to learn precise IRUT features corresponding to their own assigned scales and yield highly accurate detection results. At the "conquer" stage, we collaboratively transfer the multi-scale IRUT knowledge from the above three teachers to our DANet with a novel element-wise attentive distillation module (EADM). The EADM makes sure our DANet concentrates on learning refined feature knowledge from the teachers adopting an uncompressed attention mechanism, where it highlights the IRUT features in an element-wise way and suppresses the background.

Our main contributions are as follows.

1) A detector model DANet for multi-scale IRUT detection task is presented. It leverages SaFEM to highlight multi-scale IRUTs in complex backgrounds with SDC, which allows our model to have an effective dynamic perception of IRUTs with varying scales without heavy model complexity. We also introduce an ACFA to the feature fusion process in our DANet, which complements critical feature properties between two neighboring feature levels with spatial and semantic information. Furthermore, to address the distractor issue, we design an ADCR that utilizes UFP to enforce similarity calculation on IRUTs and distractors, improving the discrimination ability towards distractors of our DANet.

2) To further boost the detection accuracy on each scale of IRUTs for our DANet, we introduce a new SSKD based on a divide-and-conquer strategy. It employs three task-oriented teachers responsible for small-, medium-, and large-scale IRUT detection. To transfer this refined multi-scale knowledge to our DANet, we adopt EADM in our SSKD that utilizes various adaptive attention mechanisms to enhance and integrate critical features for effective learning.

3) We evaluate the proposed DANet on a composed multi-scale IRUT dataset. Extensive experiments demonstrate that our method achieves superior detection performance against other SOTA methods, and realizes real-time detection.

2 RELATED WORKS

Infrared Small Target Detection. Methods for this task can be categorized into traditional and DL-based ones. The former ones (*e.g.*, PSTNN [46], DNGM [40], WSLCM [14]) struggle to identify the targets, especially when targets are weak and the backgrounds are complex. This is because they mostly use the local contrast of the infrared images, thus yielding poor results. In contrast, DL-based methods (*e.g.*, MDvsFA [34], ACM [6], ALCNet [7], DNA-Net [23], ISNet [49], DAGNet [9]) are more effective and robust with a greater model capacity. Unlike detectors for visible images like FCOS [33] and ATSS [50], these methods are more effective for IRUT detection with architectures meticulously designed for infrared small target detection tasks. However, current DL-based methods mainly focus on small-scale infrared targets while neglecting real-world scenarios with multi-scale ones, and they suffer from insufficient usage of critical target features (*e.g.*, heavy feature compression operations, and lack of discrimination of distractors), which leads to performance degradation. In our work, we exploit the attention mechanism by integrating an SDC to comprehensively enhance our DANet's ability to perceive multi-scale IRUTs. We also employ contrastive learning to our DANet for discriminating distractors, which samples features directly from IRUTs and distractors, and projects them with no information loss for an accurate feature distinction.

Attention Mechanism. The purpose of the attention mechanism in vision tasks is to adaptively enhance target features while suppressing backgrounds. Classic attention mechanisms, such as SE-Net [19], CBAM [39], and ECA-Net [35], contain global pooling operations for obtaining attention weights, which is useful for visible images with rich features but not for infrared images with weak features. Although some works on infrared small targets (*e.g.*, ACM [6] and DAGNet [9]) altered attention mechanisms to preserve finer features for accurate feature enhancement, they only use stacks of convolutions with fixed receptive fields to perceive targets, which limits their capability of effectively enhancing IRUTs of varying scales. The proposed attention mechanism SaFEM combines a separable deformable convolution for the dynamic perception of IRUTs with no feature compression, and thus is able to enhance multi-scale IRUTs more precisely. We also propose ACFA utilizing an uncompressed weighting measure to realize a comprehensive fusion of different features from neighboring levels.

Knowledge Distillation. KD in vision tasks is to transfer powerful teacher knowledge to a student model so that the student model can achieve a better performance while maintaining its superior efficiency. Current KD methods for object detection mostly focus on general detection tasks for visible images with richer details [9], contrary to IRUT detection with weak target features and heavy background interference, thus misleading the student by incorrect knowledge transfer. Some works have adopted KD in infrared target detection (*e.g.*, CMD [29], AID [22]), but their teachers are not tailored for multi-scale IRUT detection, which limits the performance improvement. In this work, we develop SSKD with three teachers each responsible for IRUT detection of one single scale, and use EADM to effectively transfer refined knowledge to our DANet.

3 METHODOLOGY

3.1 DANet

In this subsection, we present the overview of our DANet, as displayed in Fig. 2. To achieve a subtle tradeoff between detection accuracy and efficiency, we compose the backbone of DANet of four network stages, which consist of 4, 6, 16, and 1 multi-branch blocks (MBBs). Inspired by RepVGG [8], each MBB, consisting of three branches (*i.e.*, a Conv3×3 branch, a Conv1×1 branch, and a shortcut branch), can extract and integrate rich multi-receptive field contextual properties of IRUTs, which enables our DANet to detect IRUTs more accurately. Additionally, in the inference phase, MBB will transform into a single-branch structure, reducing inference time. At the end of each stage, a SaFEM is embedded in order to dynamically enhance IRUT feature representations with uncompressed attention based on an SDC. The proposed SaFEM can further adjust the receptive fields of our DANet, and is also able to adaptively highlight critical features of IRUTs while suppressing backgrounds. To increase discrimination of distractors in the complex backgrounds, the SaFEM-enhanced feature maps will be fed into the proposed ADCR, which extracts features from regions of IRUTs/distractors and calculate similarities on features obtained by UFP. To ensure the network is able to fully exploit multi-level IRUT features, enhanced feature maps from each stage will be sent to ACFA for a comprehensive feature fusion by a cross-weighting mechanism. By ACFA, our DANet can simultaneously make full use of multi-level contexts of IRUTs for better multi-scale detection performance. In a nutshell, the use of SaFEM, ACFA, and ADCR jointly facilitates the accurate detection of our DANet.

3.1.1 Scale-adaptive Feature Enhancement Mechanism (SaFEM). Multi-scale IRUT detection tasks consist of targets with very dynamically heterogeneous characteristics (such as shapes, poses, contrast, *etc.*), where typical attention mechanisms with stacks of static convolutions will result in a limited perception of different IRUTs. Furthermore, with global pooling operations, many existing attention mechanisms fail to preserve critical spatial properties for IRUTs, leading to inaccurate feature enhancement. Our SaFEM leverages the SDC to extract rich details of multi-scale IRUTs with

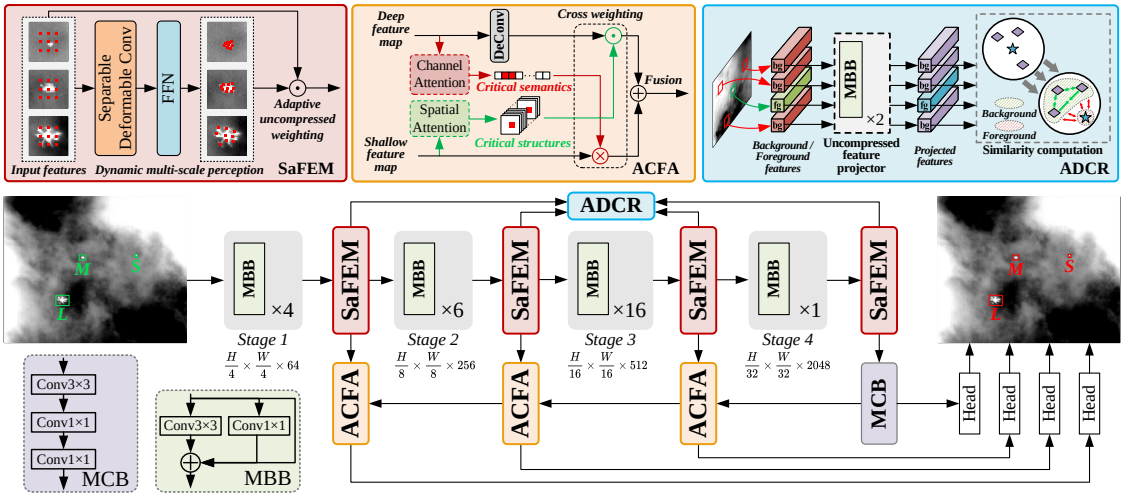


Figure 2: Overview of the proposed DANet. SaFEM, ACFA, and ADCR are introduced in Sections 3.1.1, 3.1.2, and 3.1.3, respectively.

dynamic receptive fields, while maintaining the entire dimensionality of feature maps for precise feature enhancement.

The structure of SaFEM is shown in Fig. 2, where we adopt an SDC and an FFN to achieve attention weight. To alleviate the complexity of deformable convolution that affects detection speed, inspired by [17, 36, 52], we detach the original deformable convolution into the dynamic spatial aggregation by a separable depth-wise deformable convolution, and the channel projection among sampling points by an FFN. The former is responsible for excavating multi-scale spatial features of IRUT by enforcing a location offset Δp on each sampling point of the original convolutional kernel. With the adaptive offsets representing scale variations, the multi-scale features of IRUT can be more effectively and efficiently captured and exploited. We also use a learnable modulation scalar m to modulate the significance of each location offset, so as to ensure the effectiveness of receptive field adjustment for multi-scale IRUTs. Moreover, to learn richer information from various representation subspaces at different locations, inspired by the multi-head self-attention mechanism, we split the SDC into four individual groups, where each group on this convolution will have different patterns of spatial aggregation, leading to more robust representations of multi-scale IRUT features. The SDC is formulated as follows:

$$\mathbf{y}(p_0) = \sum_{g=1}^{G}\sum_{k=1}^{K} \mathbf{w}_g \mathbf{m}_{gk} \mathbf{x}_g(p_0 + p_k + \Delta p_{gk}), \tag{1}$$

where K represents the total number of sampling points, and p_k is the k-th location of the predefined grid sampling points {(-1,-1), (-1,0), ..., (0,+1), ..., (+1,+1)}. G denotes the number of convolution groups, which we set to 4. For the g-th group, $\mathbf{w}_g \in \mathbb{R}^{C\times C'}$ is the location-unrelated projection weights of the group, and $C' = \frac{C}{G}$ is the group dimension. $\mathbf{m}_{gk}$ means the Softmax-normalized modulation scalar of the k-th point within the g-th group. $\mathbf{x}_g \in \mathbb{R}^{C'\times H\times W}$ represents the sliced input feature map, and Δp_{gk} is the dynamic offset related to the sampling location p_k in the g-th group. The overall process can be denoted as:

$$\mathbf{F}_e = \text{Softmax}(\text{FFN}(\text{SDC}(\mathbf{F}_{input}))) \odot \mathbf{F}_{input}, \tag{2}$$

where $\mathbf{F}_{input}$ and $\mathbf{F}_e$ denote the input and output feature map, $\odot$ means the element-wise multiplication. The FFN are two linear layers that project the channel dimensionality 2× then restore it back. Note that there is no spatial compression throughout, which is favorable for preserving fine details of multi-scale IRUTs. With SaFEM, the IRUT features will be adaptively and dynamically enhanced regardless of scale variations, thus yielding more robust detection results.

3.1.2 Attention-guided Cross-weighting Feature Aggregator (ACFA). Many works have rectified the structure of FPN for infrared small target detection [6, 7, 9] However, these works only focus on small-scale IRUT detection while overlooking other scales. Thus, we propose ACFA which employs dynamic channel and spatial attention mechanisms to extract critical semantics as well as structures, and then cross-weights on the neighboring feature map to complement feature details from different levels.

As shown in Fig. 2, assume the shallow and the deep feature maps are $\mathbf{F}_s \in \mathbb{R}^{C\times H\times W}$ and $\mathbf{F}_d \in \mathbb{R}^{C\times \frac{H}{2}\times \frac{W}{2}}$, respectively. $\mathbf{F}_s$ will first be fed into a spatial attention module to calculate the spatial weights of critical structures $\mathbf{w}_s \in \mathbb{R}^{C\times H\times W}$. We use SaFEM as the spatial attention module to highlight multi-scale features for a comprehensive perception, especially for small-scale IRUTs. Meanwhile, $\mathbf{F}_d$ will go through a channel attention module and yield the channel weights of critical semantics $\mathbf{w}_d \in \mathbb{R}^{C\times 1\times 1}$, which is instantiated by an SDC, a global average pooling, and an FFN. The global average pooling is to obtain the overall spatial response of each channel from $\mathbf{F}_d$, the SDC is to assist the pooling operation to yield more IRUT-representative responses of each channel for the channel weights, and the point-wise convolutions are for dimension projection. Overall, this channel attention is able to underline low-level critical semantics conducive to medium- and large-scale IRUT detection. Hence, $\mathbf{w}_s$ will be element-wisely multiplied onto $\mathbf{F}'_d \in \mathbb{R}^{C\times H\times W}$ up-sampled by a deconvolution, while $\mathbf{w}_d$ will be channel-wisely multiplied onto $\mathbf{F}_s$, thus constructing a cross-weighting process with the supplement of critical information of semantics from deep layers as well as structures from shallow layers. The two enhanced feature maps will then be summed together to form the final fused feature map. The ACFA is achieved via:

$$\mathbf{w}_s = \text{SaFEM}(\mathbf{F}_s), \tag{3}$$

$$\mathbf{w}_d = \text{Sigmoid}(\text{FFN}(\text{GAP}(\text{SDC}(\mathbf{F}_d)))), \tag{4}$$

$$\mathbf{F}_o = \mathbf{w}_s \odot \text{DeConv}(\mathbf{F}_d) + \mathbf{w}_d \otimes \mathbf{F}_s, \tag{5}$$

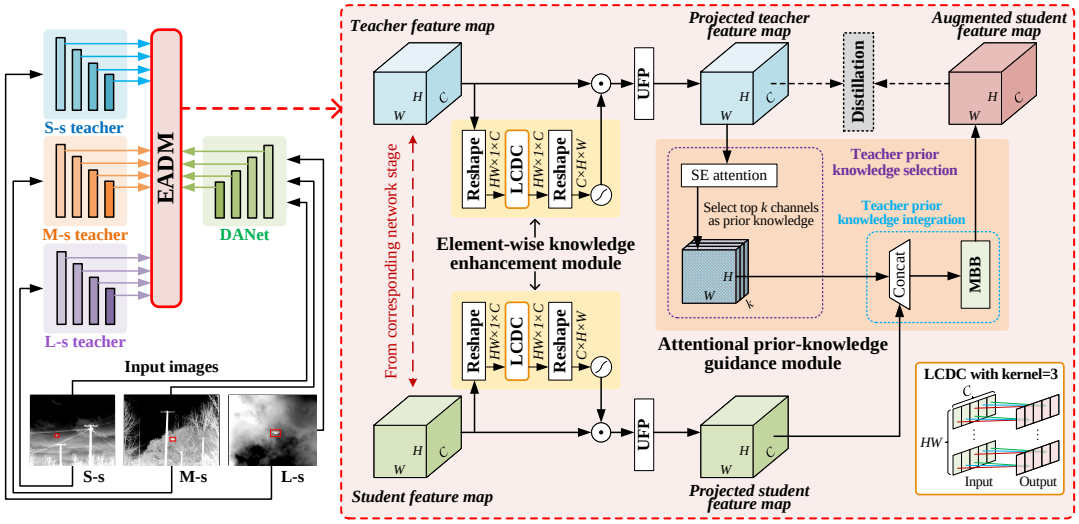


Figure 3: Overview of our scale-specific knowledge distiller (SSKD). "S-s", "M-s", and "L-s" denote "Small-scale", "Medium-scale", and "Large-scale", respectively. The output feature maps from each matching stage will all be computed for feature-based distillation with EADM. For simplicity, we only display the distillation between our DANet and one teacher, and show LCDC with a kernel of 3.

where GAP denotes global average pooling; $\odot$ and $\otimes$ are element-wise and channel-wise multiplication; $\mathbf{F}_o$ is the fused output feature map; DeConv means deconvolution.

3.1.3 Anti-Distractor Contrastive Regularization (ADCR). The backgrounds where multi-scale IRUTs fly in usually contain interferences and distracting objects, which might cause false alarms. To address this issue, we devise ADCR derived from contrastive learning, as shown in Fig. 2. Each output feature map from SaFEM will be sent into ADCR for a comprehensive comparison of IRUTs and distractors. Note that ADCR is only employed in the training phase.

The key to contrastive learning is to decide its positive and negative samples. Unlike existing works for classification [3, 15, 21] that determine samples by global features, we stress that the IRUT in each image is the positive sample, and the distractors having similar features to the IRUTs in the background are labeled as the negative samples. The ratio of positive and negative samples is set to 1:3, where we will randomly extract regions the same size as the IRUT in the image as the negative samples if there are no evident distractors. After obtaining the positive and negative samples, we further feed them into the uncompressed feature projector (UFP) to project features for the final similarity computation. Our UFP is comprised of two MBBs, which are able to extract high-level and multi-scale discriminative feature representations of IRUTS as well as distractors. Particularly, compared to many existing works adopting multi-layer perceptron that collapses features into vectors, our UFP maintains the dimensionality of each sample, crucial for preserving distinct features for the accurate discrimination of IRUTs and distractors. Then the projected features are used to construct the contrastive loss:

$$\mathcal{L}_{ADCR} = \sum_{a=1}^{N_p} \frac{-1}{N_p} \sum_{p=1}^{N_p} \log \frac{\Phi(s_a, s_p)}{\sum_{p=1}^{N_p} \Phi(s_a, s_p) + \sum_{n=1}^{N_n} \Phi(s_a, s_n)}, \quad (6)$$

$$\Phi(x, y) = \exp\left((x \cdot y)/\mu\right), \quad (7)$$

where s_a, s_p, and s_n denote the anchor, the positive, and the negative sample, respectively. N_p and N_n are the numbers of positive and negative samples. μ is a temperature parameter. "$\cdot$" represents the cosine similarity.

3.2 Scale-Specific Knowledge Distiller (SSKD)

Based on a divide-and-conquer strategy, we present the SSKD to further improve the multi-scale IRUT detection performance by discretely transferring refined single-scale IRUT knowledge from each responsible teacher model to our DANet using the proposed EADM, as shown in Fig. 3.

3.2.1 Divide: Task-oriented Multiple Teachers. To further increase the multi-scale IRUT detection performance, we conjecture that it is also important to improve the detection performance on each specific scale. We found that using three networks trained on three datasets of different scales (*i.e.*, small-, medium-, and large-scale datasets) will yield respectively better detection results than those by only one network trained on a hybrid multi-scale dataset. This suggests that multi-scale IRUT knowledge from these networks is more precise and robust, and it is worth further learning. To this end, we meticulously design three different networks for small, medium-, and large-scale IRUT detection, respectively, where each network can offer high detection performance on the IRUT dataset of their corresponding scales and provide sophisticated IRUT feature knowledge. We pre-train these networks to their optimum based on their corresponding datasets, and then use them to further train our DANet for further improving the multi-scale awareness. More details can be found in the supplementary material.

3.2.2 Conquer: Element-wise Attentive Distillation Module (EADM). IRUTs usually have weak features (especially for small-scale ones), and complicated background interferences might affect detection performance. In addition, the high-level knowledge from multiple teachers may be difficult for our DANet to learn. Accordingly, it is crucial to transfer the most valuable IRUT knowledge from three teachers to our DANet correctly. To this end, we devise EADM to tackle this problem, which consists of an element-wise knowledge enhancement module (EKEM) to highlight critical IRUT features as well as suppress background clutters in the intermediate feature

maps from the teachers for effective distillation, and a novel attentional prior-knowledge guidance module (APGM) to narrow the learning difficulty by adaptively aggregating teacher crucial prior knowledge to our DANet.

As shown in Fig. 3, we use three teacher models (*i.e.*, small-, medium-, and large-scale teachers) to further improve the multi-scale IRUT detection performance of our DANet. Note that we only assign images to the teachers corresponding to the scales of IRUTs within. For example, if an image in a training mini-batch contains one small-scale IRUT, this image will only be assigned to the small-scale teacher, while the other two teachers will not receive any inputs. Each teacher will have four stages in their architectures matching our DANet, and the output feature maps from the corresponding stages will be used for feature-based distillation. Take the student feature map $\mathbf{F}_i^s \in \mathbb{R}^{C\times H\times W}$ (i means input, C, H, and W means channel, height, and width) and a teacher feature map $\mathbf{F}_i^t \in \mathbb{R}^{C\times H\times W}$ as an example, these two feature maps will go through an EKEM for feature enhancement before distillation. First, they will be flattened to $HW \times 1 \times C$, and then sent to a local cross-channel depth-wise convolution (LCDC) operation to compute the significance of each spatial position. Unlike SENet [19] and ECANet [35], a depth-wise convolution with kernel size $(1, k)$ is used to aggregate the local channel response of all spatial positions from the input feature map, where the parameter k is decided by $k = \lfloor \frac{\log(c)+1}{\tau} \rfloor_{odd}$. The c denotes the number of input channels, τ is a temperature parameter set to 2, and $\lfloor s \rfloor_{odd}$ means the floor odd number of s. With channel information focally organized, each spatial position can represent its features more distinctively, which facilitates the accurate generation of the attention map. Following the reshape and the sigmoid layers, the attention map will be element-wisely multiplied on the input feature maps $\mathbf{F}_i^s$ and $\mathbf{F}_i^t$, thus enhancing critical IRUT features for improving the performance of knowledge transfer. The procedure of EKEM is formulated via $\mathbf{F}_e = \sigma(R(\text{LCDC}(R(\mathbf{F}_i)))) \odot \mathbf{F}_i$, where σ and R denote the sigmoid and reshape operations, respectively.

After EKEM, the enhanced features will be projected as $\mathbf{F}_p^s \in \mathbb{R}^{C\times H\times W}$ and $\mathbf{F}_p^t \in \mathbb{R}^{C\times H\times W}$, and fed to the proposed APGM. Our APGM can adaptively select important IRUT features from teachers by the teacher prior knowledge selection, and aggregates them with features from DANet by the teacher prior knowledge integration to construct the augmented student feature map $\mathbf{F}_a^s \in \mathbb{R}^{C\times H\times W}$. At last, the feature-based distillation will perform between $\mathbf{F}_p^t$ and $\mathbf{F}_a^s$. At TPKS, an SE channel attention will adaptively sort the significance of each channel from $\mathbf{F}_p^t$, and select the top k (we set $k = \frac{C}{8}$) most IRUT-related channels to concatenate with $\mathbf{F}_p^s$. The selected knowledge represents the most important features that our DANet needs to acquire in advance before distillation. Henceforth, different knowledge will be fused together for a comprehensive representation of multi-scale IRUT by an MBB to form the final $\mathbf{F}_a^s$ for distillation.

The process of EADM can be expressed as follows.

$$\mathbf{F}_p = \text{UFP}(\mathbf{F}_e), \mathbf{F}_a^s = \text{TPKI}(\mathbf{F}_p^s, \text{TPKS}(\mathbf{F}_p^t)), \tag{8}$$

$$\mathcal{L}_{SSKD} = \mathcal{D}(\mathbf{F}_p^t, \mathbf{F}_a^s), \tag{9}$$

where the distillation loss $\mathcal{D}$ is defined as a l_2-norm loss, and TPKI/TPKS denote the teacher prior knowledge selection/integration.

3.3 Overall Loss

The overall loss consists of the detection loss of DANet (a cross-entropy classification loss and a smooth-l_1 regression loss), ADCR loss, and SSKD loss. Therefore, the overall loss is written as

$$\mathcal{L} = \mathcal{L}_{cls} + \mathcal{L}_{reg} + \lambda_1 \mathcal{L}_{ADCR} + \lambda_2 \mathcal{L}_{SSKD}, \tag{10}$$

where we set the hyper-parameters λ_1 to 0.1, and λ_2 to 0.05.

4 EXPERIMENTS

4.1 Dataset and Evaluation Metrics

We conduct experiments on the real IRUT dataset consisting of small-, medium-, and large-scale datasets. We adopt our own datasets of 31,329 multi-scale IRUT images, where the numbers of the small, medium-, and large-scale images are 13,489, 11,039, and 6,801, respectively. We also add 10,712 multi-scale IRUT images from online public datasets [6, 31, 49] to the overall dataset to construct an overall dataset of 42,041 images. The ratio of the training set and the test set is set to 9:1. More details of the datasets (*e.g.*, scale separation criteria, dataset organization) can be found in the supplementary material. For the evaluation metrics, we use precision (P), recall (R), F1 measure ($F1$), and frames per second (FPS) to evaluate the model performance.

4.2 Implementation Setups

We first train our DANet on our composed multi-scale dataset. We adopt SGD optimizer with a learning rate of 10^{-3} and a weight decay 10^{-4} to train our DANet for a total of 476 epochs using a batch size 8. For SSKD, we pre-train the three teacher models to their optimum on their scale-corresponding datasets to ensure their effectiveness, and then freeze them to train our DANet using the whole dataset with batch size 4. While training DANet with SSKD, we will automatically decide which teacher model each image sample to feed according to the IRUT scale; in the testing phase, we feed the entire multi-scale test set to our DANet. All experiments are conducted using Nvidia RTX 3090, with PyTorch 1.9 and CUDA 10.2. For comparisons, we select DNGM [40] as the traditional infrared small target detection method; MDvsFA [34], ACM [6], ALCNet [7], DNA-Net [23], ISNet [49], DAGNet [9] as the DL-based infrared small target detection methods; and FCOS [33] and ATSS [50] as the baseline detectors.

4.3 Quantitative Results

As shown in Table 1, it can be evidently seen that the P, R, and F1 metrics of the proposed DANet are higher than those of the rest SOTA methods. For example, compared with the DAGNet which achieves the overall second-best quantitative results, our DANet has improved P, R, and F1 by 2.93%, 2.42%, and 2.67%, respectively. Meanwhile, our DANet maintains a decent detection efficiency of 31.57 FPS, meaning that DANet keeps a subtle balance between detection accuracy and efficiency. The traditional infrared small target detection method DNGM yields poor detection results due to limitations brought by its handcrafted feature filter design, thus not compatible with varying and complex detection conditions. The DL-based infrared small target detection methods obtain much better results than the DNGM but are still inferior to our DANet. We think it is due to their insufficient ability to perceive multi-scale IRUTs within complex backgrounds, where their architectures lack enough tailored structures to learn sophisticated and refined knowledge of IRUTs of different scales. The baseline detectors, however, achieve

Table 1: Quantitative results of our DANet and other SOTA methods on the composed multi-scale datasets.

Method	Evaluation metrics			
	P	R	F1	FPS
DNGM [40]	37.17	40.02	38.54	14.99
MDvsFA [34]	85.69	88.18	86.92	28.84
ACM [6]	89.14	88.68	88.91	**36.18**
ALCNet [7]	88.58	90.29	89.40	34.51
DNA-Net [23]	89.54	87.06	88.28	27.59
ISNet [49]	91.23	90.61	90.92	33.16
DAGNet [9]	92.75	90.09	91.40	34.95
FCOS [33]	86.88	82.97	84.88	26.10
ATSS [50]	87.63	90.46	89.03	29.77
DANet	**95.68**	**92.51**	**94.07**	31.57

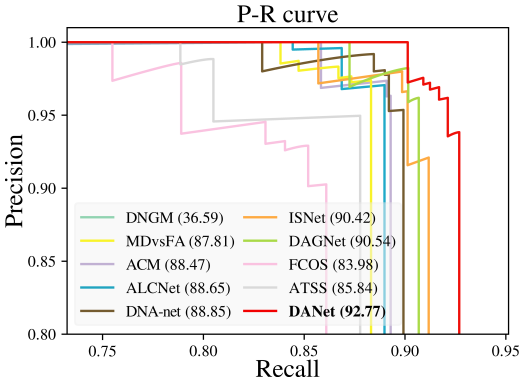


Figure 4: P-R Curve of our DANet with other SOTA methods. DNGM does not appear due to low detection accuracy.

mediocre results with less satisfying efficiency, which is because their structures are not meticulously designed for IRUT detection.

We also plot the P-R curves for each comparison method shown in Fig. 4. The p-R curve is a balanced graphical representation to evaluate an object detection algorithm, and the area under the curve indicates the average precision of the method, where the higher is the value the better. It can be seen that our method achieves the largest area under the curve with the best average precision of all detection methods.

4.4 Qualitative Results

We present six representative detection results in Fig. 5 to demonstrate the effectiveness of the proposed DANet. The methods we choose to compare are the six overall best-performed methods reported in Table 1, including our DANet. From Fig. 5 rows (a)-(c), the detection of three images containing large-, medium-, and small-scale IRUTs are displayed, whose target sizes are 52×42, 12×11, and 5×3, respectively. Rows (c), (e) and (f) are the results of the six methods on the public datasets "A dataset for multi-sensor drone detection" [31], IRSTD-1k [49] and NUAA-SIRST [6], respectively. It can be seen that our DANet is able to yield robust detection results regardless the IRUT scale variations. In comparison, the DAGNet fails to detect the large- and medium-scale IRUTs, and the ISNet fails to detect the small-scale IRUT. Besides, for medium- and small-scale IRUTs in rows (b) and (c), most methods mistake similar objects in the sky as the IRUTs, thus causing false alarms. As for rows (d) and (e), they are two typical scenes of complex backgrounds containing distractors. For row (d), the target is flying

Table 2: Ablation studies of the effectiveness of SaFEM, ACFA, ADCR, and SSKD for our DANet.

Model				Metrics			
SaFEM	ACFA	ADCR	SSKD	P	R	F1	FPS
-	-	-	-	90.07	87.93	88.98	**38.86**
✓	-	-	-	90.88	89.95	90.41	31.59
-	✓	-	-	90.39	88.57	89.47	34.83
-	-	✓	-	92.10	89.02	90.53	34.91
-	-	-	✓	91.93	90.44	91.18	35.85
✓	✓	-	-	93.85	91.76	92.79	35.85
✓	✓	✓	-	94.11	91.59	92.83	30.29
✓	✓	-	✓	93.87	92.19	93.02	31.58
✓	✓	✓	✓	**95.68**	**92.51**	**94.07**	31.57

in front of the buildings afar; for row (e), the target is with trees and bushes. We can see our DANet is the only method that accurately detects the IRUTs from the complex backgrounds, while the other methods all yield incorrect results affected by the distractors: in row (d), all methods mistake a section of the building in the trees as the IRUT, as ALCNet further misidentify an object in the bushes; in row (e), all the comparison methods mistake the tree leaves as the IRUTs. Row (f) demonstrates a scenario with strong building interference, where our DANet is the only one that accurately detects the IRUT while all the others fail. These results verify the effectiveness of our DANet, which adopts various adaptive attention mechanisms to improve the model's multi-scale IRUT awareness and uses contrastive learning to improve distractor discrimination.

4.5 Ablation Study

In this paper, we present the ablation studies on our SaFEM, ACFA, ADCR, and SSKD to demonstrate the effectiveness of our method. We use our DANet minus the proposed SaFEM, ACFA, ADCR, and SSKD as the comparison baseline for our ablation studies. More ablations and details (*e.g.*, different structures of the proposed techniques, SSKD with different teachers, LCDC effectiveness in EKEM, *etc.*) can be found in the supplementary material.

The impact of SaFEM, ACFA, and ADCR. As shown in Table 2, we notice that our DANet yields better results while utilizing any of SaFEM, ACFA, and ADCR, compared with row 1 where none of these techniques are used. It is worth noticing that the overall performance improvement peaks when our DANet is equipped with ADCR in contrast to SaFEM and ACFA, with an improvement of 2.03% in P and 1.55% in F1. This indicates that each of these techniques is able to effectively strengthen the multi-scale IRUT awareness of our DANet, as ADCR is more effective *w.r.t.* discriminating distractors in complex backgrounds. Our DANet performs even better while having SaFEM, ACFA, and ADCR all together, suggesting that these techniques offer to be a collaborative contribution to the effectiveness of multi-scale IRUT detection. The SaFEM can dynamically enhance multi-scale feature representations for ADCR and ACFA to more precisely learn high-level discriminative information for better detection performance.

The impact of SSKD. Reported in rows 1 and 7 from Table 2, it can be seen that there is an improvement of 1.06% in P, 2.51% in R, and 2.20% in F1 with the proposed SSKD. If our DANet is employed with SaFEM, ACFA, and ADCR, the performance is further improved by 1.57% in P, 0.92% in R, and 1.24% in F1, according to rows 6 and 8. This is because our designed task-oriented teachers

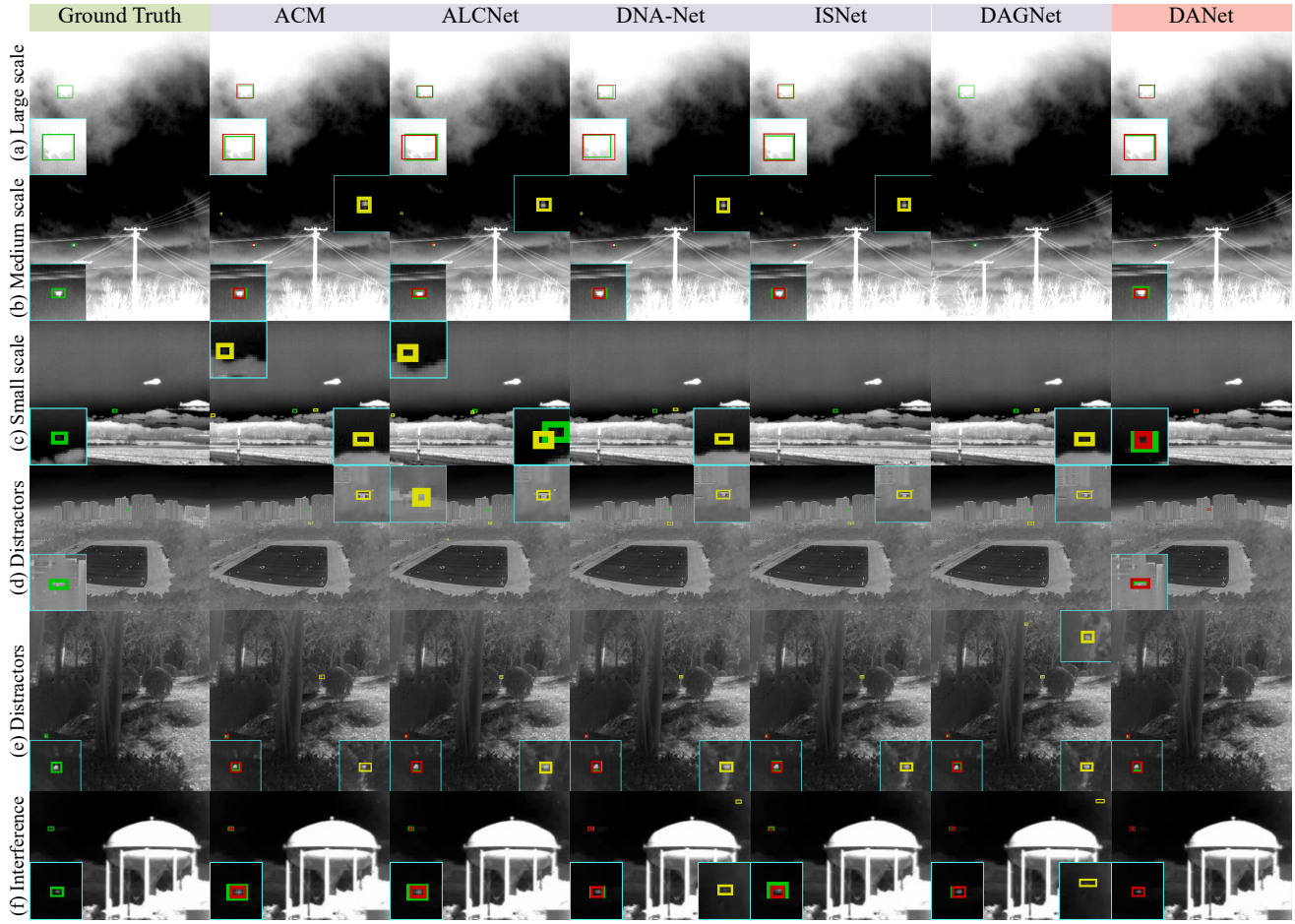


Figure 5: Qualitative results of our DANet and other SOTA methods. The comparison involves the six overall best-performed methods reported in Table 1. The target sizes in these images are 51×42, 12×11, 5×3, 11×8, 8×8, and 8×5. The green, red, and yellow boxes denote the ground truths, the correct detections, and the false alarms, respectively. We display close-ups in each image, where the ground truths and the correct detection results are magnified at the left-bottom corner, and the false alarms are put at other corners.

are able to extract high-level refined multi-scale IRUT knowledge, and transfer them to our DANet effectively via EADM, verifying the effectiveness of our SSKD. Additionally, attributed to the dynamic perception ability, our DANet is also capable of acquiring sophisticated multi-scale IRUT knowledge from the three teachers of SSKD, which cooperatively strengthens the knowledge learning process so as to improve the detection performance.

5 CONCLUSION

In this paper, we propose a DANet to address the multi-scale IRUT detection task in various complex backgrounds. To enable our DANet to have dynamic perceptions of IRUTs with various scales, we embed the SaFEM at the end of each stage of our DANet, which exploits the SDC with its dynamic offsets for convolutional kernels to realize accurate multi-scale IRUT feature highlighting as well as background suppressing. We also design an ACFA to fuse features from different levels by complementing high-level semantics with dynamic channel attention and low-level structures with dynamic spatial attention. An ADCR is devised to tackle the discrimination between IRUTs and distractors in complex backgrounds, which is by enforcing similarity computation on distinguishing features projected by a UFP. To further improve the multi-scale IRUT detection performance, we introduce the SSKD based on a divide-and-conquer strategy. At the "divide" stage, we utilize three different teacher models to solve IRUT detection on small-, medium-, and large-scale IRUT datasets, respectively. At the "conquer" stage, we adopt the EADM, which leverages attention mechanisms to collaboratively transfer tailored knowledge from the teachers to our DANet. Extensive experiments have verified the validity of our method, which obtains high detection accuracy and realizes real-time detection.

ACKNOWLEDGMENTS

This work was supported in part by the Open Research Fund of the National Key Laboratory of Multispectral Information Intelligent Processing Technology under Grant 6142113220303, the National Natural Science Foundation of China under Grants 61971460 and 62101294, the Fundamental Research Funds for the Central Universities under Grant ZYTS23201, and the JCJQ Program under Grant 2021-JCJQ-JJ-0060.

REFERENCES

[1] Cancan Chen, Runqiu Xia, Yang Liu, and Yue Liu. 2023. A Simplified Dual-Weighted Three-layer Window Local Contrast Method for Infrared Small Target Detection. *IEEE Geoscience and Remote Sensing Letters* 20 (2023), 1–5.

[2] Fang Chen, Chenqiang Gao, Fangcen Liu, Yue Zhao, Yuxi Zhou, Deyu Meng, and Wangmeng Zuo. 2022. Local Patch Network with Global Attention for Infrared Small Target Detection. *IEEE Transactions on Aerospace and Electronic Systems* 58, 5 (2022), 3979–3991.

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A Simple Framework for Contrastive Learning of Visual Representations. In *International Conference on Machine Learning*. PMLR, 1597–1607.

[4] Yi-Wen Chen, Xiaojie Jin, Xiaohui Shen, and Ming-Hsuan Yang. 2022. Video Salient Object Detection via Contrastive Features and Attention Modules. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 1320–1329.

[5] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. 2017. Deformable Convolutional Networks. In *Proceedings of the IEEE International Conference on Computer Vision*. 764–773.

[6] Yimian Dai, Yiquan Wu, Fei Zhou, and Kobus Barnard. 2021. Asymmetric Contextual Modulation for Infrared Small Target Detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 950–959.

[7] Yimian Dai, Yiquan Wu, Fei Zhou, and Kobus Barnard. 2021. Attentional Local Contrast Networks for Infrared Small Target Detection. *IEEE Transactions on Geoscience and Remote Sensing* 59, 11 (2021), 9813–9824.

[8] Xiaohan Ding, Xiangyu Zhang, Ningning Ma, Jungong Han, Guiguang Ding, and Jian Sun. 2021. RepVGG: Making VGG-style Convnets Great Again. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13733–13742.

[9] Houzhang Fang, Zikai Liao, Xuhua Wang, Yi Chang, and Luxin Yan. 2023. Differentiated Attention Guided Network Over Hierarchical and Aggregated Features for Intelligent UAV Surveillance. *IEEE Transactions on Industrial Informatics* 19, 9 (2023), 9909–9920.

[10] Houzhang Fang, Mingjiang Xia, Gang Zhou, Yi Chang, and Luxin Yan. 2021. Infrared Small UAV Target Detection based on Residual Image Prediction via Global and Local Dilated Residual Networks. *IEEE Geoscience and Remote Sensing Letters* 19 (2021), 1–5.

[11] Zilin Gao, Jiangtao Xie, Qilong Wang, and Peihua Li. 2019. Global Second-order Pooling Convolutional Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3024–3033.

[12] Jianping Gou, Liyuan Sun, Baosheng Yu, Lan Du, Kotagiri Ramamohanarao, and Dacheng Tao. 2022. Collaborative Knowledge Distillation via Multiknowledge Transfer. *IEEE Transactions on Neural Networks and Learning Systems* (2022), 1–13.

[13] Ismail Guvenc, Farshad Koohifar, Simran Singh, Mihail L. Sichitiu, and David Matolak. 2018. Detection, Tracking, and Interdiction for Amateur Drones. *IEEE Communications Magazine* 56, 4 (2018), 75–81.

[14] Jinhui Han, Saed Moradi, Iman Faramarzi, Honghui Zhang, Qian Zhao, Xiaojian Zhang, and Nan Li. 2020. Infrared Small Target Detection based on the Weighted Strengthened Local Contrast Measure. *IEEE Geoscience and Remote Sensing Letters* 18, 9 (2020), 1670–1674.

[15] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum Contrast for Unsupervised Visual Representation Learning. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*. 9729–9738.

[16] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the Knowledge in a Neural Network. *arXiv preprint arXiv:1503.02531* (2015).

[17] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. 2017. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv preprint arXiv:1704.04861* (2017).

[18] Hanzhe Hu, Jinshi Cui, and Liwei Wang. 2021. Region-Aware Contrastive Learning for Semantic Segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 16291–16301.

[19] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu. 2020. Squeeze-and-Excitation Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42, 8 (2020), 2011–2023.

[20] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. 2019. CCNet: Criss-Cross Attention for Semantic Segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 603–612.

[21] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised Contrastive Learning. *Advances in Neural Information Processing Systems* 33 (2020), 18661–18673.

[22] Qizhen Lan and Qing Tian. 2022. Instance, Scale, and Teacher Adaptive Knowledge Distillation for Visual Detection in Autonomous Driving. *IEEE Transactions on Intelligent Vehicles* 8, 3 (2022), 2358–2370.

[23] Boyang Li, Chao Xiao, Longguang Wang, Yingqian Wang, Zaiping Lin, Miao Li, Wei An, and Yulan Guo. 2022. Dense nested attention network for infrared small Target Detection. *IEEE Transactions on Image Processing* 32 (2022), 1745–1758.

[24] Sihao Lin, Hongwei Xie, Bing Wang, Kaicheng Yu, Xiaojun Chang, Xiaodan Liang, and Gang Wang. 2022. Knowledge Distillation via the Target-aware Transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10915–10924.

[25] Yuang Liu, Wei Zhang, and Jun Wang. 2020. Adaptive Multi-Teacher Multi-level Knowledge Distillation. *Neurocomputing* 415 (2020), 106–113.

[26] Jianyuan Ni, Anne HH Ngu, and Yan Yan. 2022. Progressive Cross-modal Knowledge Distillation for Human Action Recognition. In *Proceedings of the 30th ACM International Conference on Multimedia*. 5903–5912.

[27] Changyong Shu, Yifan Liu, Jianfei Gao, Zheng Yan, and Chunhua Shen. 2021. Channel-wise Knowledge Distillation for Dense Prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5311–5320.

[28] Wonchul Son, Jaemin Na, Junyong Choi, and Wonjun Hwang. 2021. Densely guided knowledge distillation using multiple teacher assistants. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9395–9404.

[29] Jingxian Sun, Lichao Zhang, Yufei Zha, Abel Gonzalez-Garcia, Peng Zhang, Wei Huang, and Yanning Zhang. 2021. Unsupervised Cross-Modal Distillation for Thermal Infrared Tracking. In *Proceedings of the 29th ACM International Conference on Multimedia*. 2262–2270.

[30] Zhicheng Sun and Yadong Mu. 2022. Patch-based Knowledge Distillation for Lifelong Person Re-Identification. In *Proceedings of the 30th ACM International Conference on Multimedia*. 696–707.

[31] Fredrik Svanström, Fernando Alonso-Fernandez, and Cristofer Englund. 2021. A dataset for Multi-sensor drone detection. *Data in Brief* 39 (2021), 107521.

[32] Maofeng Tang, Konstantinos Georgiou, Hairong Qi, Cody Champion, and Marc Bosch. 2023. Semantic Segmentation in Aerial Imagery Using Multi-Level Contrastive Learning With Local Consistency. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 3798–3807.

[33] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. 2020. FCOS: A Simple and Strong Anchor-Free Object Detector. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 4 (2020), 1922–1933.

[34] Huan Wang, Luping Zhou, and Lei Wang. 2019. 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*. 8509–8518.

[35] Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu. 2020. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11534–11542.

[36] Wenhai Wang, Jifeng Dai, Zhe Chen, Zhenhang Huang, Zhiqi Li, Xizhou Zhu, Xiaowei Hu, Tong Lu, Lewei Lu, Hongsheng Li, et al. 2022. InternImage: Exploring Large-Scale Vision Foundation Models with Deformable Convolutions. *arXiv preprint arXiv:2211.05778* (2022).

[37] Xuehui Wang, Kai Zhao, Ruixin Zhang, Shouhong Ding, Yan Wang, and Wei Shen. 2022. ContrastMask: Contrastive Learning to Segment Every Thing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11604–11613.

[38] Yantao Wei, Xinge You, and Hong Li. 2016. Multiscale Patch-based Contrast Measure for Small Infrared Target Detection. *Pattern Recognition* 58 (2016), 216–226.

[39] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. 2018. CBAM: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision*. 3–19.

[40] Lang Wu, Yong Ma, Fan Fan, Minghui Wu, and Jun Huang. 2020. A Double-Neighborhood Gradient Method for Infrared Small Target Detection. *IEEE Geoscience and Remote Sensing Letters* 18, 8 (2020), 1476–1480.

[41] Yunkai Xu, Minjie Wan, Xiaojie Zhang, Jian Wu, Yili Chen, Qian Chen, and Guohua Gu. 2023. Infrared Small Target Detection Based on Local Contrast-Weighted Multidirectional Derivative. *IEEE Transactions on Geoscience and Remote Sensing* 61 (2023), 1–16.

[42] Zhendong Yang, Zhe Li, Xiaohu Jiang, Yuan Gong, Zehuan Yuan, Danpei Zhao, and Chun Yuan. 2022. Focal and Global Knowledge Distillation for Detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4643–4652.

[43] Ze Yang, Shaohui Liu, Han Hu, Liwei Wang, and Stephen Lin. 2019. Reppoints: Point Set Representation for Object Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9657–9666.

[44] Lewei Yao, Renjie Pi, Hang Xu, Wei Zhang, Zhenguo Li, and Tong Zhang. 2021. G-DetKD: Towards General Distillation Framework for Object Detectors via Contrastive and Semantic-guided Feature Imitation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3591–3600.

[45] Linfeng Zhang and Kaisheng Ma. 2021. Improve Object Detection with Feature-based Knowledge Distillation: Towards Accurate and Efficient Detectors. In *International Conference on Learning Representations*.

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

[47] Mingjin Zhang, Haichen Bai, Jing Zhang, Rui Zhang, Chaoyue Wang, Jie Guo, and Xinbo Gao. 2022. RKformer: Runge-Kutta Transformer with Random-Connection Attention for Infrared Small Target Detection. In *Proceedings of the 30th ACM*

International Conference on Multimedia. 1730–1738.

[48] Mingjin Zhang, Ke Yue, Jing Zhang, Yunsong Li, and Xinbo Gao. 2022. Exploring Feature Compensation and Cross-level Correlation for Infrared Small Target Detection. In *Proceedings of the 30th ACM International Conference on Multimedia*. 1857–1865.

[49] Mingjin Zhang, Rui Zhang, Yuxiang Yang, Haichen Bai, Jing Zhang, and Jie Guo. 2022. ISNet: Shape Matters for Infrared Small Target Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 877–886.

[50] Shifeng Zhang, Cheng Chi, Yongqiang Yao, Zhen Lei, and Stan Z Li. 2020. Bridging the Gap Between Anchor-based and Anchor-free Detection via Adaptive Training Sample Selection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9759–9768.

[51] Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. 2022. Decoupled Knowledge Distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11953–11962.

[52] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. 2019. Deformable Convnets v2: More Deformable, Better Results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9308–9316.