SCINet: Spatial and Contrast Interactive Super-Resolution Assisted Infrared UAV Target Detection

Houzhang Fang, *Member, IEEE*, Lan Ding, Xiaolin Wang, Yi Chang, *Member, IEEE*, Luxin Yan, *Member, IEEE*, Li Liu, *Member, IEEE*, and Jinrui Fang

***Abstract*— Unmanned aerial vehicle (UAV) detection based on thermal infrared imaging has been one of the most important sensing technologies in the anti-UAV system. However, the technical limitations and long-range detection of thermal sensors often lead to acquiring low-resolution (LR) infrared images, thereby bringing great challenges for the subsequent target detection task. In this article, we propose a novel spatial and contrast interactive super-resolution network (SCINet) for assisting infrared UAV target detection. The network consists of two main subnetworks: a spatial enhancement branch (SEB) and a contrast enhancement branch (CEB). The SEB embeds the lightweight convolution module and attention mechanism to highlight the spatial structure detail features of infrared UAV targets. The proposed CEB incorporates the center-oriented contrast-aware module and multibranch collapsible module, which can provide local contrast priors to reconstruct the super-resolved UAV target image. The spatial features of the intermediate layers from the SEB are integrated into the CEB as a rich gradient prior. Besides, the output features of the CEB are aggregated into those of the SEB for further supplementing the contrast of spatial features in return. Dual-branch feature interaction (DBFI) of the SEB and CEB can further enhance the spatial details and target saliency of the targets. In addition, we also introduce an infrared UAV detection network via a new dual-dimensional feature calibration module (DFCM) for boosting the detection performance. Extensive experiments demonstrate that the SCINet outperforms the state-of-the-art (SOTA) SR methods on real infrared UAV sequences and improves the detection performance of infrared small UAV targets. The code is available at https://github.com/IVPLaboratory/SCINet.**

Received 5 July 2024; revised 11 September 2024; accepted 25 September 2024. Date of publication 1 October 2024; date of current version 10 October 2024. This work was supported in part by the Open Research Fund of the National Key Laboratory of Science and Technology on Multispectral Information Intelligent Processing under Grant 61421132301; in part by the National Natural Science Foundation of China under Grant 41501371, Grant 62101294, and Grant 61971460. *(Corresponding author: Houzhang Fang.)*

Houzhang Fang and Xiaolin Wang are with the School of Computer Science and Technology, Xidian University, Xi'an 710071, China (e-mail: houzhangfang@xidian.edu.cn; wxl@stu.xidian.edu.cn).

Lan Ding is with iFLYTEK Company Ltd., Hefei 230088, China (e-mail: 1635126912@qq.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: yichang@hust.edu.cn; yanluxin@hust.edu.cn).

Li Liu is with the School of Aerospace Science and Technology, Xidian University, Xi'an 710071, China (e-mail: dr.liliu@hotmail.com).

Jinrui Fang is with the Center for Data Science, New York University, New York City, NY 10011 USA (e-mail: jf4959@nyu.edu).

This article has supplementary downloadable material available at https://doi.org/10.1109/TGRS.2024.3471786, provided by the authors.

Digital Object Identifier 10.1109/TGRS.2024.3471786

***Index Terms*— Feature fusion, infrared small target detection, infrared super-resolution (SR), unmanned aerial vehicles (UAVs).**

I. Introduction

THE unmanned aerial vehicles (UAVs) have been widely applied in various fields, such as environmental surveillance, aerial photography, and precision agriculture [1], [2], [3]. However, the unauthorized use of UAVs raises great threats to public safety and aerial security. Many accidents involving UAVs threatening public areas and strikes on aircraft around airports have been reported [4], [5], [6]. Therefore, UAV detection is critically important in protecting public safety and urban security. UAV detection based on thermal infrared imaging has been one of the most important sensing technologies in the anti-UAV system for its strong reliability and long-distance monitoring ability [7], [8]. Consequently, accurate identification and localization of UAVs is an important part of guaranteeing the effectiveness of anti-UAV systems.

However, due to the limitations of infrared sensing techniques and the long imaging distance between the infrared sensor and UAVs, the UAV targets in the acquired low-quality infrared images are often small in size and have weak features (e.g., contour and shape) [9], [10], [11], making it difficult to extract discriminative features for the subsequent detection task. In addition, UAVs during motion are susceptible to interference from complex backgrounds, including dense clouds, prominent building edges, and strong illumination variations. Therefore, it is challenging to accurately detect infrared UAV targets directly on LR images, especially for small-scale ones. The qualitative comparison of our method on infrared images with different resolutions in two complex scenes is shown in Fig. 1. The results illustrate that detecting UAV targets directly on LR infrared images easily leads to false alarms and miss detection. In contrast, the proposed SR method can overcome the above problems and improve the detection performance in complicated scenarios.

Currently, methods for addressing infrared small target detection can be primarily categorized into two groups: model-driven methods [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22] and data-driven methods [23], [24], [25], [26]. The former relies heavily on hand-crafted features to construct

1558-0644 © 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: XIDIAN UNIVERSITY. Downloaded on October 11,2024 at 13:03:53 UTC from IEEE Xplore. Restrictions apply.

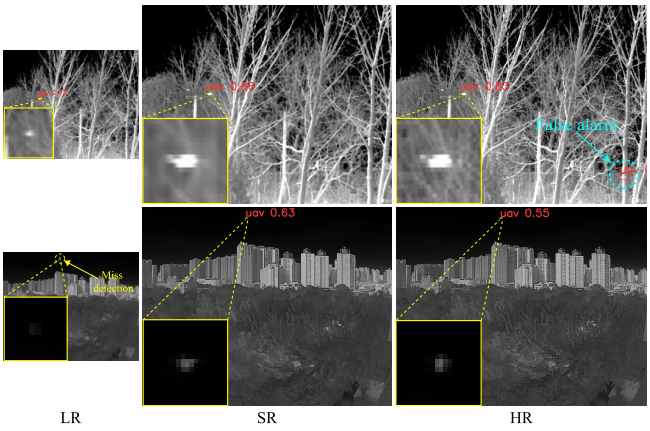


Fig. 1. Qualitative comparison of detection results of our method on real infrared images with different resolutions in two complex scenes. The first, second, and third columns present the detection results for the low-resolution (LR) images, super-resolution (SR) images with the proposed method, and the original high-resolution (HR) images, respectively. The yellow area in the lower left corner is a local close-up of the target region under different resolution images. The proposed SCINet cannot only improve the quality of LR images but also help overcome false alarms and miss detection in LR images. Furthermore, the detection performance of SR images even surpasses that of original HR images.

target detection models. These methods tend to yield satisfactory results for targets in simple scenarios, such as high local contrast and strong salient. They may generate false alarms in scenarios with complex backgrounds and a low signal-to-noise ratio. The latter exhibits significant advantages owing to the ability to autonomously learn intricate feature representations from data. Many general object detection methods, such as Cascade R-CNN [27], FCOS [28], have presented impressive performance, while these methods are tailored for HR visible images with rich target details. Recently, several studies have endeavored to design convolutional neural networks (CNNs)-based models for infrared target detection. Dai et al. [23] utilized an attention-based asymmetric context modulation (ACM) module to highlight infrared UAV target features. Zhang et al. [24] adopted Taylor finite difference and two-orientation attention aggregation block to extract edge and shape features of targets for improving the detection of infrared small targets. Fang et al. [25] meticulously constructed differentiated attention mechanisms in different layers to strengthen the detection ability of infrared UAV targets. However, due to the inherent limitations of infrared sensors and the long detection range, the observed infrared UAV images usually have low resolution and lack obvious shape and contrast features. Consequently, the detection performance of the above methods is still constrained by the resolution and intrinsic quality of the images.

To address the aforementioned challenges, we propose in this work to develop an infrared image SR algorithm to enhance the spatial and contrast features of the UAV targets for facilitating the downstream application, such as infrared weak UAV target detection under complicated backgrounds [29], [30], [31]. Existing single-frame image SR methods are mainly divided into two categories: traditional SR methods [32], [33], [34], [35], [36], [37] and deep learning-based methods [38], [39]. The traditional SR methods includes statistics-based [32], [33], prediction-based [34], [35], and sparse representation methods [36], [37]. These methods rely on handicraft features to reconstruct SR images, which are suboptimal in tackling with challenging scenarios. For the deep learning-based methods, Ma et al. [38] proposed a structure-preserving SR (SPSR) method to solve the issue of geometric distortions in the SR problem, thereby restoring the geometric structure of the SR image. Wei et al. [39] exploited the attention mechanism to perform adaptive learning to recover flat regions, edges, and corner regions separately. However, the above methods are mainly designed for visible light images, aiming to restore the texture structure features and improve the visual effect of the image.

In recent years, infrared images SR methods are also introduced for improving image resolution [30], [40], [41], [42], [43], [44]. Ren et al. [30] proposed a region-generative adversarial network (GAN) to perform SR reconstruction of infrared images. Chen et al. [43] employed an iterative error reconstruction mechanism to perform SR in a coarse-to-fine manner. Prajapati et al. [44] introduced a channel splitting-based CNN (ChaSNet) for thermal image SR eliminating the redundant features in the network. Although the above infrared SR methods can yield good results, they mainly reconstruct high-quality images from the perspective of improving spatial resolution or visual quality, rather than focusing on simultaneously boosting the spatial and contrast information of the target. In fact, the spatial structure and contrast features of targets are important for infrared UAV target detection. However, infrared UAV targets generally have weak spatial features and low contrast, making it difficult to directly extract sufficient features of the aforementioned two types from LR infrared UAV images. Therefore, designing an infrared image SR model is crucial to enhance the spatial details of infrared UAV targets while strengthening the target contrast, thereby facilitating the subsequent detection task.

In this article, we present a novel spatial and contrast interactive SR network (SCINet) for assisting infrared UAV target detection. As shown in Fig. 2, the SCINet decouples the SR task into two subtasks: spatial enhancement branch (SEB) and contrast enhancement branch (CEB), which interactively fuse the spatial and contrast features. To improve the efficiency of the SEB while enhancing spatial structure details, we adopt the lightweight convolution module [45] to reduce the number of parameters and exploit the attention mechanism to make the network focus on the high-frequency feature regions of interest, which strengthens the representation ability of the SEB. Generally, there is a relatively high contrast between UAV targets and the surrounding background. Moreover, since image gradient information can reveal the sharpness and contrast of the local region of an image, we leverage the gradient features of the infrared image to guide the SR reconstruction to focus more on the UAV target region, thus enhancing the feature information of the UAV target. In this work, we design the CEB composed of contrast enhancement residual dense blocks (CE-RDBs) for contrast feature extraction and enhancement. The CE-RDB is adopted to extract hierarchical contrast features at different levels of the network. Each CE-RDB consists of a center-oriented

Authorized licensed use limited to: XIDIAN UNIVERSITY. Downloaded on October 11,2024 at 13:03:53 UTC from IEEE Xplore. Restrictions apply.

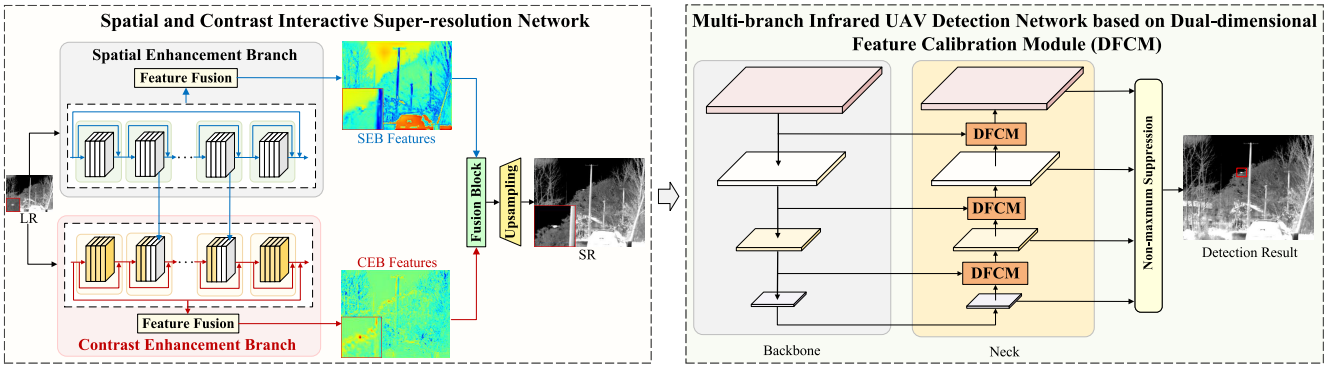


Fig. 2. Overview of the working diagram of the proposed framework for infrared small UAV target detection. The SCINet is adopted to enhance the spatial and contrast features of the SR images. The resulting SR images are sent to the proposed multibranch infrared UAV detection network based on a dual-dimensional feature calibration module (DFCM) to improve the detection performance of weak and small UAV targets.

contrast block (CCB) and multibranch collapsible blocks (MCBs). The CCB utilizes cross-center difference convolution (CDC) to accomplish learnable contrast feature extraction and enhancement of UAV targets. In addition, the MCB module combines multiple receptive fields of different scales to enhance the representation capability of a single convolutional layer. Furthermore, to reduce the computational complexity of the CEB, we incorporate the reparameterization technique into the CEB by transforming the training-time MCB into an inference-time single path. The densely connected CE-RDB in CEB not only generates a larger receptive field to promote the utilization of contrast features but also preserves information on small targets.

Besides, SCINet improves the SR performance of infrared UAV images through the dual branches feature interaction of the SEB and CEB. On the one hand, the CEB incorporates feature information from several intermediate levels of the SEB. This is because a well-designed SEB can carry rich gradient information, which is crucial for feature extraction and enhancement of the contrast branch. Therefore, exploiting these intermediate features of the CEB as strong prior can improve the performance of the CEB. On the other hand, the contrast features generated by the CEB are also further integrated into the SEB to provide additional local contrast priors for reconstructing reliable and detailed SR images. The dual-branch feature interaction (DBFI) can further prominently highlight the feature discrimination between the UAV targets and background, thereby strengthening the target saliency of SR images. In conclusion, in addition to resolution enhancement, the SCINet accomplishes contrast enhancement for the UAV target region, thus providing finer and sufficient target information for accurate detection and minimizing the occurrence of miss detections and false alarms.

To further improve the detection performance, we construct a multibranch infrared UAV detection network based on a new spatial and channel DFCM. In the DFCM, we adopt the attention mechanism to recalibrate and enhance the concatenated features in the two dimensions of sequential spatial and channel. With DFCM, the network fully integrates accurate localization details in low-level and fine semantic information at a high level in a dynamically weighted manner for more discriminative after fusion, thus improving the target detection performance. In addition, since the infrared UAV targets are weak in features and small in size, the spatial positioning information of the UAV target is easily lost in the deep layers of the network. Therefore, we also incorporate a new detection branch by fusing the features in the shallower layer into the high-level layer to improve the detection performance of small-scale UAV targets.

Overall, the main contributions of this work are summarized as follows.

1) We present a novel framework for tackling infrared small UAV target detection in complex backgrounds by integrating an SR network and a detection network to collaboratively improve the detection capabilities of infrared small-scale UAV targets.
2) We propose a novel SCINet that integrates the SEB and CEB. The SEB improves the spatial details and resolution of LR UAV images and the CEB performs the contrast enhancement of the infrared UAV targets. Furthermore, DBFI can facilitate the model to learn the correlations and complementarities between spatial and contrast features, as well as enhance spatial and contrast features simultaneously, thus significantly boosting the infrared UAV image SR performance.
3) We introduce a new multibranch infrared UAV detection network based on the DFCM. The DFCM can recalibrate and enhance the aggregated features of different hierarchical levels. We additionally incorporate a shallower branch into the deep layer to improve the detection performance of small-scale UAV targets.
4) We conduct a comprehensive comparison of various SR and detection methods on synthetic and real-world images. Extensive experiments demonstrate that the proposed SR and detection networks outperform the state-of-the-art (SOTA) methods.

The article is organized as follows. Section II reviews the related works of image SR and target detection. Section III introduces the design motivation and details of our model. Section IV presents the experiment results and analyses. Finally, we conclude the whole article in Section V.

Authorized licensed use limited to: XIDIAN UNIVERSITY. Downloaded on October 11,2024 at 13:03:53 UTC from IEEE Xplore. Restrictions apply.

II. Related Work

A. Efficient Single Image SR

Image SR is an inherently ill-posed optimization problem. Many conventional image SR algorithms [32], [33], [34], [35], [36], [37] including those based on statistics or sparse representations have encountered difficulties in accurately formulating the mapping from LR to HR images. Consequently, these challenges have led to suboptimal SR performance. However, with the advancements in deep learning techniques, the methods based on CNNs have demonstrated impressive capabilities in the realm of SR. These CNN-based methods have significantly improved the overall SR performance, effectively overcoming the limitations associated with traditional methods. The VDSR [46] proposed by Kim et al. introduces the idea of residual to reconstruct the SR image by enabling the network to learn the high-frequency residual partial between SR and LR images. Kim et al. [47] proposed the DRCN network by employing the recursive neural networks in the SR task for the first time, further improving the SR efficiency and image quality. Tong et al. [48] proposed SRDenseNet, which applied the dense block for the SR reconstruction and achieved superior performance. Wang et al. [49] exploited the GAN network named ESRGAN to enable SR to achieve better visual quality, especially in terms of realism and texture details. Ma et al. [38] proposed the SPSR to alleviate the issue of geometric distortions and boost SR performance by utilizing the gradient guidance. Although the aforementioned methods have achieved superior SR performance, the computational cost is too high. As a result, lightweight SR networks have attracted increasing attention to address the inefficient issue. Liu [50] et al. presented the RFDN that achieves a better tradeoff in terms of performance and model complexity by exploiting residual feature distillation mechanisms. Li et al. [45] proposed the BSRN that achieves SOTA performance while being highly efficient by reducing redundant computations in convolution operations.

Although the SR algorithm mentioned above achieves superior performance, it is mainly designed for visible light images to reconstruct smooth textures and detailed structures. Since infrared images lack apparent texture and structural details, it is difficult to obtain high-quality SR infrared images by the above methods. With the widespread application of HR infrared images, the research on infrared images SR has also seen significant growth. Chen et al. [43] utilized an iterative error reconstruction mechanism to enhance the generation of SR images enabling the model to recover more textures and edge details. Prajapati et al. [44] introduced ChaSNet, which utilizes channel splitting-based CNN to reduce redundancy in the network, thereby maintaining high-frequency details of the SR images. Meanwhile, Huang et al. [51] proposed a progressive SR GAN and adopted a multistage transfer learning strategy to improve SR performance, especially in scenarios with limited data samples. Although these methods are proposed for infrared images, they mainly focus on producing high-fidelity infrared SR images. They failed to fully take into account the distinctive imaging mechanism and military applications of infrared imaging systems, as well as the potential advantages and assistance of infrared image SR to downstream tasks.

B. Infrared Small Target Detection

The methods for small target detection are primarily categorized into model-driven [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22] and data-driven [27], [52], [53], [54], [55], [56], [57]. The model-driven methods largely rely on hand-crafted features and prior knowledge of the targets. However, these methods may encounter limitations and challenges in circumstances like complex and dynamic backgrounds, thereby affecting the performance of small target detection. The data-driven methods exploit CNNs to detect infrared targets due to their adaptive feature learning ability. The Cascade R-CNN [27] as a classic two-stage network significantly enhances target detection performance through its multistage refinement process but also increases the computational burden. Many one-stage target detection algorithms have demonstrated satisfactory performance in detecting small targets. For instance, RetinaNet [53] significantly improves the detection capability of less noticeable small targets by introducing focal loss. The YOLO v5 [55] effectively improves the ability of cross-scale feature fusion by optimizing the feature extraction network, thus exhibiting great ability and potential in the detection tasks of multiscale targets. However, the one-stage detection methods still face challenges when dealing with small targets in complex backgrounds and dynamic environments. To enhance the accuracy of small-scale target detection while maintaining efficiency, many research efforts have gradually integrated one-stage and two-stage detectors. For example, methods like RefineDet [57] and DAGNet [25] combine the efficiency of the one-stage with the accuracy of the two-stage network to improve the overall detection performance and efficiency. However, such methods tend to be sensitive to target occlusion and complex backgrounds. In addition, other techniques also have been introduced to improve the detection performance of small-scale targets, such as anchor-free design [58], [59], decoupling of classification and detection tasks [59], [60], and high-density detection strategies. Although these methods improve the detection performance of small-scale targets to some extent, they are still limited in some complex backgrounds. In this article, we integrate the attention mechanisms and a shallower detection branch into the detection network to boost the detection performance of infrared small UAV targets.

C. Infrared Small Target Detection With SR

Some existing works suggest that SR preprocessing can effectively improve target detection performance. Xiaolin et al. [61] proposed a target detection model S2ANET-SR based on the SR preprocessing step to improve the detection accuracy of small targets in remote sensing images. Ying et al. [62] first proposed an SR network called MoCoPNet to super-resolve LR infrared images and further investigate its influence on the performance of infrared target

Authorized licensed use limited to: XIDIAN UNIVERSITY. Downloaded on October 11,2024 at 13:03:53 UTC from IEEE Xplore. Restrictions apply.

detection. Li and Shen [31] initially utilized SRResNet to super-resolve the LR images, followed by processing these SR images through their newly proposed YOLOSR-IST network, thus improving the detection of small targets in infrared remote sensing images. Zhang et al. [63] presented SuperYOLO, which utilizes auxiliary multiresolution learning for HR detection of multiscale targets. The above methods mainly leverage the existing SR or detection networks to investigate the impact of SR techniques on target detection performance but do not provide a framework for integrating SR and detection technology. In this work, we introduce a novel framework for tackling infrared small UAV target detection under a complicated background. The framework comprehensively enhances the spatial resolution and contrast features of LR infrared UAV images by integrating SR into the detection network, thus achieving a better performance against other methods.

III. Proposed Method

In this section, we first provide a detailed description of the proposed UAV target detection framework assisted by the SR network in Section III-A. Then, we introduce the SCINet in Section III-B. Finally, we describe the proposed multibranch infrared UAV detection network based on dual-dimensional feature calibration in Section III-C.

A. Overall Architecture

Fig. 2 is a working diagram of the proposed infrared UAV target detection framework assisted by the SR network. The entire framework comprises two key components: the SCINet and the multibranch infrared UAV detection network based on dual-dimensional feature calibration. The SCINet decouples the infrared UAV image SR task into two subtasks: spatial enhancement and contrast enhancement, and achieves SR image reconstruction with simultaneous enhancement of spatial resolution and target contrast features through two-branch feature interaction. The improved detection network processes the enhanced SR images. To reduce the occurrence of false alarms and miss detections, we introduce the DFCM module into the feature fusion network and add a shallower detection branch to achieve greater detection capability of the model.

B. Spatial and Contrast Interactive SR Network

The detailed architecture of the SCINet is depicted in Fig. 3. The SCINet is composed of two branches, namely, the SEB and the CEB. The SEB is implemented by stacking the lightweight convolution modules and attention mechanism of the SR network BSRN [45] to refine the extraction of spatial structural features of the image. Considering that infrared UAV images have a local contrast prior to the spatial dimension, we introduce the CEB to achieve center-oriented contrast-aware feature extraction. The CEB can effectively enhance the contrast feature of the UAV target, thus providing contrast priors for SR reconstruction. Furthermore, the CEB incorporates several intermediate-layer features from the SEB. Such a scheme is inspired by the idea that a well-designed SEB can carry rich gradient information, which facilitates the contrast extraction and enhancement of the CEB. Consequently, we exploit the features in the SEB as a gradient prior to promoting the performance of the CEB. Finally, the output contrast features from the CEB are returned to be fused with the spatial features of the SEB to reconstruct the final SR features and generate the SR image.

Let us denote the input LR image and output SR image as I_{LR} and I_{SR}, respectively. We first replicate the input LR image n times along the channel dimension and then send it to BSConv proposed in BSRN for shallow feature extraction. The operation can be formulated as

$$F_{\mathrm{SF}} = O_{\mathrm{SF}}\left(\left[I^1_{\mathrm{LR}}, I^2_{\mathrm{LR}}, \cdots I^n_{\mathrm{LR}}\right]\right) \tag{1}$$

where O_{SF} denotes the module of shallow feature extraction achieved by BSConv, which consists of a 1×1 convolution and a depth-wise convolution. $[I^1_{\mathrm{LR}}, I^2_{\mathrm{LR}}, \ldots, I^n_{\mathrm{LR}}]$implies the concatenation of the LR image n times.

Next, the shallow feature F_{SF} will be fed into the SEB and the CEB as input, respectively. In the SEB, we gradually refine the extracted features by a stack of ESDBs in the BSRN. The stacking of ESDBs in the SEB not only gradually enhances the overall spatial details of the image but also ensures the retention of key structural information, further improving the quality of SR images. The process can be written as

$$F^k_{\mathrm{SEB}} = O_{\mathrm{ESDB}}\left(F^{k-1}_{\mathrm{SE}B}\right), k = 1, 2, \cdots n. \tag{2}$$

The proposed CE-RDBs are integrated into the CEB to mine and encode rich contrast features of the infrared images. In the CEB, the hierarchical contrast features are reused by utilizing dense connections, which not only strengthen the contrast features of the UAV target but also maintain the background information. As shown in Fig. 3, the CEB fuses the feature representations of the second, fourth, and sixth intermediate-layer features from the SEB. The process can be defined as

$$F^k_{\mathrm{CEB}} = \begin{cases} O_{\text{CE-RDB}}\left(F^{k-1}_{\mathrm{CEB}}\right), & k = 2, 4, 6, \\ O_{\text{CE-RDB}}\left(F^{k-1}_{\mathrm{CEB}}, F^{k-1}_{\mathrm{SEB}}\right), & k = 1, 3, 5. \end{cases} \tag{3}$$

To make full use of the multilevel local features, we concatenate the hierarchical features from all depths, respectively, in the two branches. Then, the concatenated features are optimized by a 1×1 convolutional layer, which effectively reduces the number of channels of the fused features and strengthens the representation capability of the model. The residual learning is involved in the CEB. The long residual connection establishes an effective feature transfer and compensation path, which not only promotes contrast information transfer from low-level to high-level but also effectively alleviates the issue of gradient vanishing. The multilevel feature fusion is formulated as

$$F^{\mathrm{fused}}_{\mathrm{SEB}} = \mathrm{Conv}\left(\left[F^1_{\mathrm{SEB}}, F^2_{\mathrm{SEB}}, \cdots, F^k_{\mathrm{SEB}}\right]\right) \tag{4}$$

$$F^{\mathrm{fused}}_{\mathrm{CEB}} = \mathrm{Conv}\left(\left[F^1_{\mathrm{CEB}}, F^2_{\mathrm{CEB}}, \cdots, F^k_{\mathrm{CEB}}\right]\right) + F_{\mathrm{SF}} \tag{5}$$

where Conv denotes the 1×1 convolution.

We incorporate the output contrast information and spatial detail features by a fusion block. The fusion block first

Authorized licensed use limited to: XIDIAN UNIVERSITY. Downloaded on October 11,2024 at 13:03:53 UTC from IEEE Xplore. Restrictions apply.

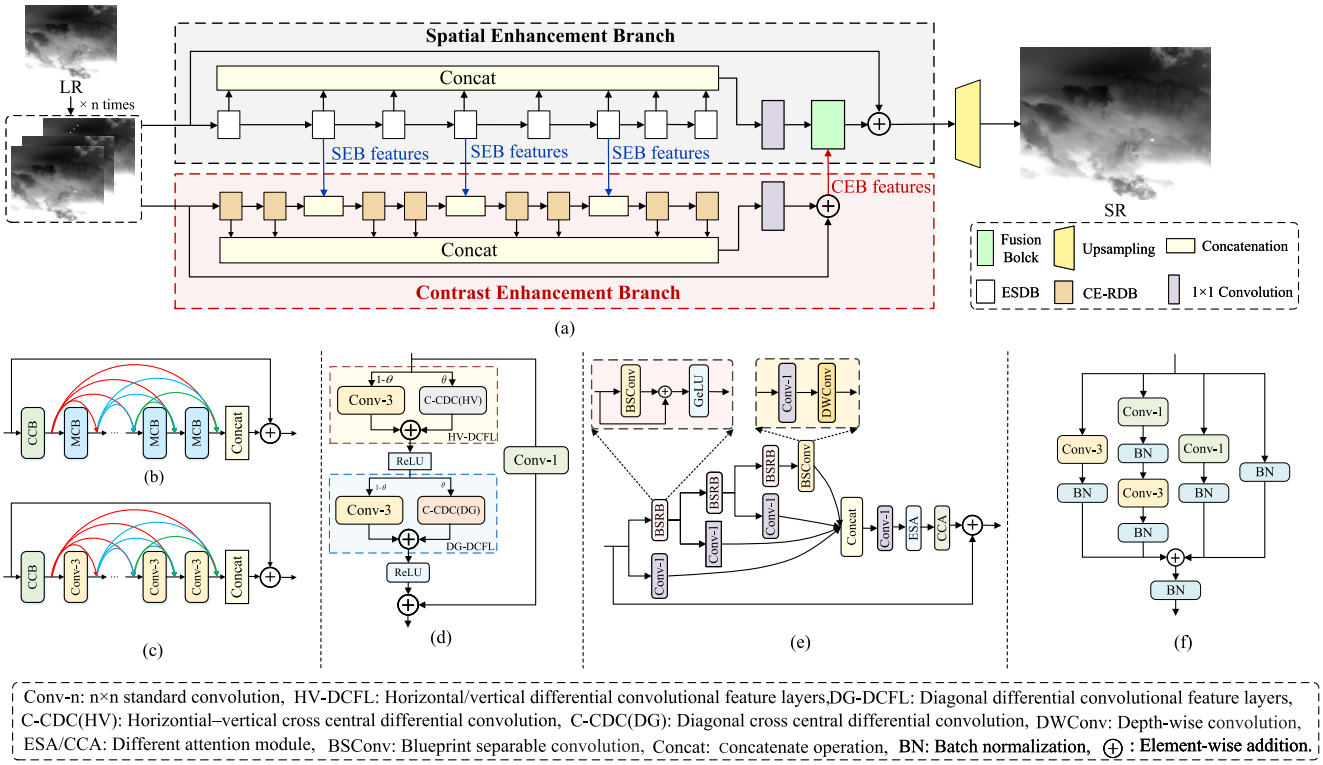


Fig. 3. Overview of the proposed SCINet framework. The gray box represents the SEB, and the red box represents the CEB. The CEB aims to improve the contrast of UAV targets while maintaining the background information. It incorporates multilevel representations from the SEB to guide the SR process by a fusion block in return. (a) Overview of the proposed SCINet framework. (b) Structure of CE-RDB during training. (c) Structure of CE-RDB during inference. (d) CCB. (e) Efficient separable distillation block (ESDB). (f) MCB.

concatenates the features from two branches together, and then the GELU activation and the other BSConv are utilized to further aggregate features. The advantage of such a fusion strategy is that it effectively integrates contrast features with spatial structural features before the up-sampling operation, thus ensuring a significant improvement in terms of target enhancement of the SR image. The process can be represented as

$$F_{\text{fused}} = O_{\text{DF}}\left(\text{GELU}\left(O_{\text{FB}}\left[F_{\text{SEB}}^{\text{fused}}, F_{\text{CEB}}^{\text{fused}}\right]\right)\right) \tag{6}$$

where O_{FB} represents the fusion block which concatenates three RRDB blocks in ESRGAN [49], and O_{DF} denotes the operation achieved by BSConv.

After the dual-branch feature fusion, we incorporate residual connections to facilitate the information flow from the initial layer to the deeper layer of the network. This is because the shallow features contain critical details from the LR image, and maintaining these features is crucial for the network to capture finer details, especially in reconstructing high-frequency areas such as edges and structural details. Finally, we utilize the pixel shuffle operation to reconstruct the SR results. The reconstruction stage is formulated as

$$I_{\text{SR}} = O_{\text{rec}}(F_{\text{fused}} + F_{\text{SF}}) \tag{7}$$

where O_{rec} represents the pixelshuffle operation.

We exploit the pixel-wise L_1 loss between the SR images and corresponding HR images to optimize the model. The loss function is

$$\begin{aligned}\mathcal{L}(\theta) &= \frac{1}{k}\sum_{i=1}^{k}\left\|F\left(I_{\text{LR}}^{k}:\theta\right) - I_{\text{HR}}^{k}\right\|_1 \\ &= \frac{1}{k}\sum_{i=1}^{k}\left\|I_{\text{SR}}^{k} - I_{\text{HR}}^{k}\right\|_1\end{aligned} \tag{8}$$

where θ is the network parameter to be learned. The $\{(I_{\text{HR}}^{k}, I_{\text{SR}}^{k})\}_{k=1}^{K}$ denotes K HR image and generated SR image pairs.

In the CEB, we propose the CE-RDB to enhance the contrast of UAV targets. The structure of CE-RDB during training is shown in Fig. 3(b). During training, each CE-RDB consists of one CCB and $k-1$ MCBs to achieve contrast feature representation ($k = 6$ in our module). The CE-RDB uses dense connections to alleviate the gradient vanishing and strengthen the propagation and exploitation of hierarchical contrast features. In particular, reusing these hierarchical features not only improves the performance of SR but also effectively enhances the contrast features of UAV targets.

The CCB is shown in Fig. 3(d). Each CCB is composed of a horizontal/vertical differential convolutional feature layer (HV-DCFL) and a diagonal differential convolutional feature layer (DG-DCFL). The differential convolutional feature layer (DCFL) contains differential convolutions in different directions and vanilla 3×3 convolution. The DCFL can be

Authorized licensed use limited to: XIDIAN UNIVERSITY. Downloaded on October 11,2024 at 13:03:53 UTC from IEEE Xplore. Restrictions apply.

formulated as

$$o(p_0) = \theta \cdot \underbrace{\sum_{p_n \in \mathcal{R}} w(p_n) \cdot (i(p_0 + p_n) - i(p_0))}_{\text{Difference convolution}} + (1-\theta) \cdot \underbrace{\sum_{p_n \in \mathcal{R}} w(p_n) \cdot i(p_0 + p_n)}_{\text{Vanilla convolution}} \quad (9)$$

where o, w, i, and p_0 denote the output feature map, the weights of the convolution kernel, the pixel value in the input feature map, and the current location on both input and output, respectively. $p_0 = (0, 0)$, p_n represent the neighboring locations around p_0 and $p_n \in \mathcal{R}$. Specifically, $\mathcal{R} = \{(-1, 0), (0, -1), (0, 0), (0, 1), (1, 0)\}$ in horizontal/vertical direction and $\mathcal{R} = \{(-1, -1), (-1, 1), (0, 0), (1, -1), (1, 1)\}$ in diagonal direction. Vanilla convolution extracts intensity-level information, while differential convolution extracts gradient-level information. The parameter $\theta \in [0, 1]$ is a hyper-parameter used to control the importance of gradient-level information and intensity-level features. The larger θ is the greater the contribution of the differential convolution. In our experiments, θ is set to 0.8. In addition, the residual connection is leveraged in CCB to retain the feedforward contrast feature and promote the propagation of information flow, where 1×1 convolution is used to reduce the number of channels from 64 to 32.

Fig. 4 presents different contrast measure modules, where the handcrafted dilated local contrast measure (DLCM) can only reserve the contrast information in one direction. The learnable contrast measure is achieved by the central difference convolution (CDC), which can improve the contrast of small UAV targets. However, CDC aggregates the central gradient features from four directions of the entire local adjacent region, which may cause feature redundancy and make training difficult to optimize. The cross-central difference convolution (C-CDC) decouples the original differential convolution into two symmetrically crossed suboperators, horizontal/vertical C-CDC [C-CDC(HV)] and diagonal C-CDC [C-CDC(DG)]. It has fewer parameters than the CDC and this decoupling method is also conducive to model convergence and features robust extraction. However, C-CDC does not retain and transfer the feedforward contrast features, which may cause the issue of shallow contrast features being lost layer by layer. The proposed CCB conducts decoupling in two directions in a serial manner, regulating the contribution of vanilla convolution and differential convolution for feature extraction through the hyper-parameter θ, and introduces local residual connections to preserve contrast flow. Experimental results indicate that serial decoupling outperforms parallel decoupling, so the proposed CCB is more in line with the task of infrared UAV image SR, aiming to recover reliable and detailed SR images with high-contrast UAV targets.

The MCB is shown in Fig. 3(f). Each MCB contains four branches, i.e., 3×3 convolutional branch, 1×1 followed by 3×3 convolutional branch, 1×1 convolutional branch, and identity branch. These branches with convolutions of different scales, so the different sizes of receptive fields can

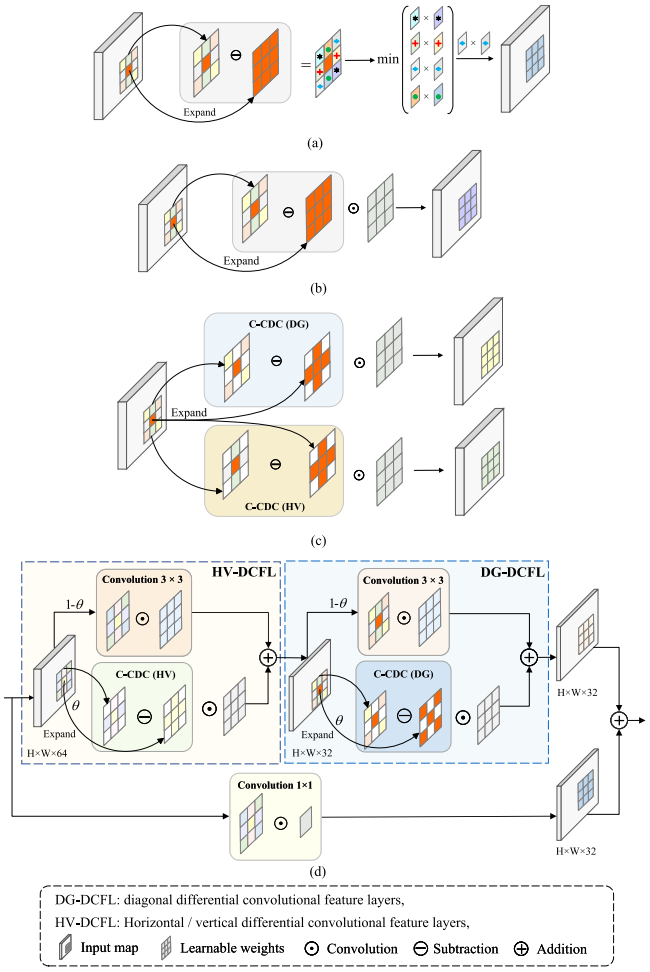


Fig. 4. Structure of different contrast extraction modules. (a) DLCM. (b) CDC. (c) C-CDC. (d) CCB.

be obtained. Therefore, the MCB enhances the representation ability of a single convolutional layer by combining multiple branches of different scale convolution, thereby generating a richer feature space. Many works have proved that the identity branch can alleviate the gradient disappearance problem and play a complementary role in the characteristics of small-scale targets, so we also introduce the identity branch in the MCB.

After the training is completed, we collapse the MCB into a single branch which only consists of the common 3×3 convolution and activation layer by reparameterization. The process of converting the MCB into a single-path one is shown in Fig. 5. We first convert the BN layer in the original MCB into a normal 3×3 convolution, and thus fuse all the convolutional layers with the BN layer to get the intermediate form (b); then we perform padding to all the 1×1 convolutions in the picture (b) to turns them into a 3×3 convolution. Due to the additivity of convolutions, we transform two 3×3 convolutions in series into one 3×3 convolution to get form (c). Finally, we add the weights and offsets of the 3×3 convolutions of the four branches in (c) to get the collapsed single-branch form (d) and complete the collapse from multibranch to single-branch. Through the above reparameterization, the structure of CE-RDB during the inference phase is shown in Fig. 3(c). It can be found

Authorized licensed use limited to: XIDIAN UNIVERSITY. Downloaded on October 11,2024 at 13:03:53 UTC from IEEE Xplore. Restrictions apply.

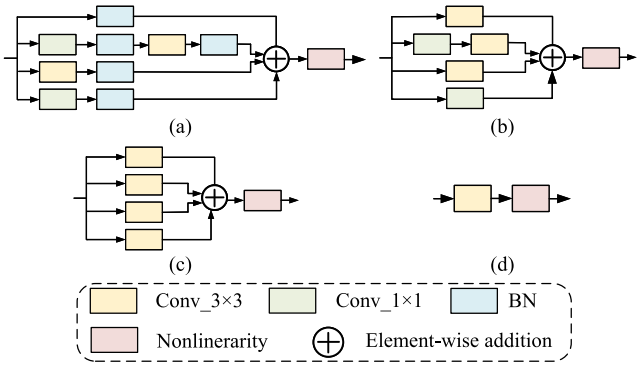


Fig. 5. Flowchart of collapsing the MCB during training to a single branch during inference. (a) Original MCB in the training phase. (b) and (c) Intermediate form of the MCB. (d) Final collapsed single-branch form. Conv_3×3 and Conv_1×1 denote the 3×3 convolution and 1×1 convolution, respectively. BN and nonlinearity represent the batch normalization and GELU function.

that the MCB realizes the decoupling of the model structure during training and inference. This collapse of the MCB allows the model not only to learn multireceptive field feature representations during the training stage but also significantly reduce the model complexity during the inference phase. This design achieves a tradeoff between performance and efficiency, which is more conducive to deployment in practice.

C. Multibranch Infrared UAV Detection Network Based on Dual-Dimensional Feature Calibration

1) Shallower Detection Branch Customized for Small-Scale UAV Targets: To enhance the network's ability to capture semantic information, lots of detectors increase the network's depth in the backbone and perform subsequent classification and location on deeper layers of the network. This design leads to the loss of crucial location information about small-scale targets as the down-sampling rate increases, resulting in a significant drop in the performance of small-scale target detection. Taking YOLO as an example, it retains feature maps at three scales with downsampling rates of 32, 16, and 8 in the backbone, and then these feature maps are sent into the feature pyramid networks (FPNs) and path aggregation network (PAN) for multiscale feature fusion, and finally transmitted to the different scale prediction heads for detection. However, when dealing with small-scale UAV targets that exhibit subtle and weak features, deep networks struggle to preserve shallow-level information, bringing great challenges and difficulties for accurate detection of small-scale UAV targets.

In the realm of small-scale target detection, the preservation of shallow feature information is very crucial and vital. To this end, we deliberately retain the feature map from the shallower layer of the network, which corresponds to the feature map with a downsampling factor of 4. This shallower layer contains abundant spatial information, which is well-suited for recognizing and localizing weak small UAV targets in infrared images. The introduction of the shallower layer detection branch enables the network to transfer the spatial location information of the low-level to the high-level of the network, thus improving the detection performance of small UAV targets. In conclusion, the proposed multibranch infrared UAV detection network comprises four specialized detection branches that are different from the original YOLO v5 with three prediction heads. Fig. 6 is the architecture of our detection network.

2) Improvement of Low-High Level Feature Fusion Based on DFCM: The low-level features contain rich spatial localization information, especially of the spatial properties of small-scale UAV targets. The high-level features usually contain finer semantic representations, which help the network improve its understanding of the context of the complex scene. However, the simple feature fusion operations (i.e., concatenation and addition) will cause feature misalignment, so fully integrating features from different levels of the network is crucial for small-scale UAV target detection. Consequently, we propose a spatial-channel DFCM based on the attention mechanism. The motivation of the DFCM is to embed low-level fine spatial location details of UAV targets into high-level global feature information, which dynamically calibrates and modulates aggregated features in an element-wise manner under the guidance of channel and spatial dual dimensions. The proposed DFCM is presented in Fig. 7.

Let C_i and P_i (the resolution is $1/2^i$ of the input images) represent the same scale feature maps of the corresponding output by the backbone and the FPN at level i, respectively, and assume that their dimensions of $C_i, P_i \in \mathbb{R}^{C\times H\times W}$. We first concatenate the C_i and P_i in the channel dimension and then utilize plain 3×3 convolution to reduce the channel from $2C$ to C to perform shallow feature fusion. We denote the fused feature as $F_i \in \mathbb{R}^{C\times H\times W}$. The output of F_i can be formulated as

$$F_i = \mathrm{Conv}_1([C_i, P_i]) \tag{10}$$

where $i \in \{l_{\min}, \cdots, l_{\max}\}$ and we set $l_{\min} = 2$ and $l_{\max} = 4$ in our experiments. The $[C_i, P_i]$ implies the concatenation operation, and Conv_1 denotes the convolution layer with a kernel size of $C \times 2C \times 3 \times 3$.

Next, we adopt depth-wise convolution, average pooling, and max pooling to obtain the attention response in the spatial dimension, and then add the three obtained feature maps in element-wise to capture cross-scale correlation

$$F_i^{\mathrm{depth}} = \mathrm{DWConv}(F_i) \tag{11}$$

$$F_i^{\mathrm{MP}} = \mathrm{M\ P}(F_i) \tag{12}$$

$$F_i^{\mathrm{AP}} = \mathrm{A\ P}(F_i) \tag{13}$$

$$F_i^{\mathrm{spatial}} = F_i^{\mathrm{depth}} + F_i^{\mathrm{MP}} + F_i^{\mathrm{AP}} \tag{14}$$

where DWConv, MP, and AP denote the depth-wise convolution, average pooling, and max pooling, respectively, whose kernel sizes are $n \times n$, and n is set to 4 in our DFCM.

After spatial shrinkage, we then leverage the point-wise convolution to scale and restore the channel to calculate the significance of each channel. Furthermore, the scaling operation before the nonlinearity function can make the training process stable

$$F_i^{\mathrm{channel}} = (\mathrm{PWConv}_2\left(\mathrm{PWConv}_1\left(F_i^{\mathrm{spatial}}\right)\right) \tag{15}$$

Authorized licensed use limited to: XIDIAN UNIVERSITY. Downloaded on October 11,2024 at 13:03:53 UTC from IEEE Xplore. Restrictions apply.

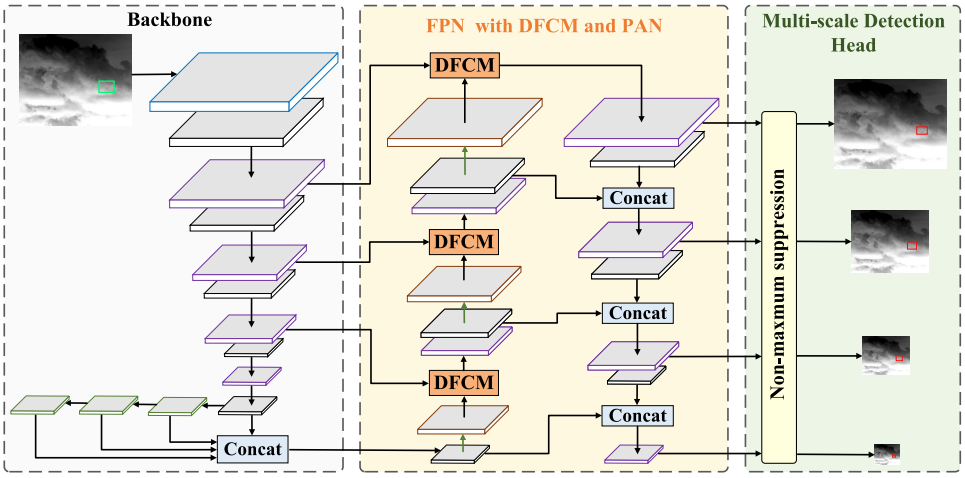


Fig. 6. Architecture of the multibranch infrared UAV detection network based on the DFCM. The blue, purple, black, and green cubes represent the 6×6 convolution, C3 module, the convolutional layer, and max-pooling in the original YOLO v5, respectively. The C3 module in the backbone can sequentially downsample the feature map by 2 times. The green arrow represents the upsampling operation and the Concat represents the concatenate operation along the channel dimension.

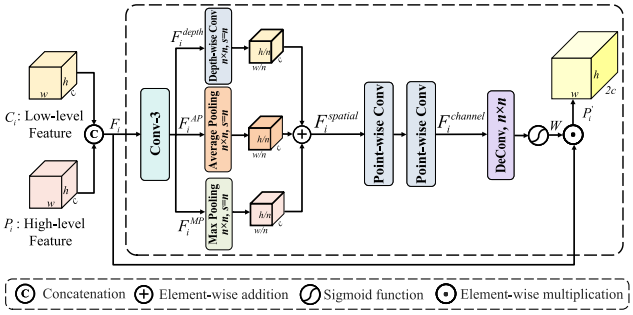


Fig. 7. Architecture of DFCM.

where PWConv_1 and PWConv_2 represent the point-wise convolution for which the kernel size is $(C/r) \times C \times 3 \times 3$ and $C \times (C/r) \times 3 \times 3$, respectively, and r is set to 2 in our DFCM.

Then, a deconvolution is used to rescale the spatial and channel dimension from $C \times (H/n) \times (W/n)$ to $2C \times H \times W$. Finally, the dynamic element-wise attention weight $W \in \mathbb{R}^{2C \times H \times W}$ is achieved by using the sigmoid function, and it will be multiplied by the input feature map element-wisely. The final output of the DFCM is calculated by

$$W = \sigma\left(\mathrm{DeConv}\left(F_i^{\mathrm{channed}}\right)\right) \tag{16}$$

$$P_i' = W \otimes ([C_i, P_i]) \tag{17}$$

where DeConv represents a deconvolution with a kernel of 4×4, σ is the sigmoid function, and $\otimes$ denotes the element-wise multiplication.

The proposed DFCM can calibrate the aggregated features by utilizing the dual-dimension attention mechanisms, thus fully integrating low-level patterns and high-level semantic representations, which helps the network to more precisely identify and locate the UAV target in the image. In addition, infrared UAV targets are usually smaller in size and weaker in features, with fewer global features but more local features. Therefore, we do not scale them to only 1×1 or 1value, whether in space or channel, but scale them to a certain scale when extracting attention weights. This operation is conducive to capturing more local features of small-scale UAV targets and preventing features from being submerged in complex backgrounds due to excessive pooling, thereby further improving the detection performance of infrared small-scale UAV targets.

The overall loss function of the detection network is given as follows:

$$\mathcal{L} = \lambda_1 \sum_i \mathcal{L}_{\mathrm{cls}}^{\mathrm{pos}}\left(\mathrm{gt}_i^c, c_i\right) + \lambda_2 \sum_i \mathcal{L}_{\mathrm{obj}}^{\mathrm{pos+ne\ g}}(\mathrm{gt}_i^b, b_i) + \lambda_3 \sum_i \mathcal{L}_{\mathrm{loc}}^{\mathrm{pos}}\left(\mathrm{gt}_i^b, b_i\right) \tag{18}$$

where $\mathcal{L}_{\mathrm{cls}}^{\mathrm{pos}}$, $\mathcal{L}_{\mathrm{obj}}^{\mathrm{pos+neg}}$, and $\mathcal{L}_{\mathrm{loc}}^{\mathrm{pos}}$ indicate the losses of categorization of positive samples, objectness categorization, and bounding box regression loss of positive samples, respectively. λ_1, λ_2, and λ_3 represent the balance coefficient of the above three losses, respectively. In addition, both classification loss and objectness loss use binary cross-entropy (BCE) loss, and location loss utilizes CIOU loss. i denotes the index of the anchor box, gt_i^c and gt_i^b are the class label and bounding box of the ground truths (GTs). c_i and b_i mean the class label and bounding box of the final prediction results. In our work, $\lambda_1 = 0.5$, $\lambda_2 = 1$, and $\lambda_3 = 0.05$, which are consistent with the default hyperparameter settings of YOLOv5.

IV. Experimental Results and Discussion

In this section, we analyze and verify the superiority and effectiveness of our method through a series of experiments. First, we introduce our experiment settings, including datasets, SOTA SR and detection baseline models, evaluation metrics for SR and detection networks, and training details. Next, we conduct a comprehensive ablation study from the perspective of SR and detection to explore the effectiveness of each proposed module. Finally, the SOTA baselines are compared with the proposed method.

Authorized licensed use limited to: XIDIAN UNIVERSITY. Downloaded on October 11,2024 at 13:03:53 UTC from IEEE Xplore. Restrictions apply.

A. Experiments Settings

1) Datasets: Our datasets contain two types, the first is a total of 16 685 infrared UAV images collected by ourselves using DJI Phantom 4, Pro V2.0, and DJIMini2 as the UAV targets, and the other is a total of 3983 open-access datasets from the second anti-UAV workshop and challenge [64]. We divide the whole dataset into nine image sequences according to the scene characteristics. Each sequence represents the scene that may be encountered in the anti-UAV mission, such as complex background, interference target, and target occlusion. Meanwhile, we divide the training set, validating set, and testing set in each sequence according to the ratio of 8 : 1 : 1 without intersection. All images are in the size of 512×640, where the infrared UAV targets range in size from around 10 pixels to over 250 pixels. In the SR task, the LR images are acquired by bicubic interpolation with a scaling factor of 4 from the HR images (512×640). Therefore, the LR images have a size of 128×160, where the infrared UAV targets range in size from around 4 pixels to over 70 pixels. The specific information of the dataset is given in the supplementary material.

2) Implementation Details: The SCINet consists of eight ESDBs and eight CE-RDBs, and the CEB receives the features generated from the second, fourth, and sixth ESDBs of the SEB. At the beginning of the SCINet, we replicated LR images in the channel dimension 4 times and randomly cropped them into 48×48 patches as the input. The batch size is 8 and the model is optimized by an Adam optimizer with a total of 3 000 000 iterations. The initial learning rate is set to 1×10^{-4} and is adjusted at 2 000 000 iterations using the CosineAnnealingRestartLR learning rate adjustment algorithm, which decreases the learning rate from the initial value to a minimum value of 1×10^{-7} in each training cycle, and then restores it to the initial value until the total number of iterations is reached. The proposed multibranch infrared UAV detection network is trained by an SGD optimizer with a momentum of 0.937 and a weight decay of 0.0005. The learning rate is set to 1×10^{-2} initially and the batch size is 8. The entire training process involves 300 epochs. Our experiments are conducted on a server with an NVIDIA GeForce RTX 3090 GPU and accelerated by CUDA 11.1. We also use PyTorch 1.12.1 and Python 3.8 for soft implementation.

3) Baseline Methods for SR and Detection: In the SR task, we compare our SCINet with several SR baselines methods: Bicubic, SRDensetNet [48], BSRN [45], SPSR [38], MoCoPNet [62], ChaSNet [44], IERN [43], PSRGAN [51], DASR [65], and WaveMixSR [66]. Bicubic interpolation is a traditional image SR method that uses a cubic interpolation algorithm to interpolate pixels in both horizontal and vertical directions. The main purpose of SRDenseNet, SPSR, IERN, and WaveMixSR is to restore natural and realistic high-quality images. The MoCoPNet, ChaSNet, PSRGAN, and DASR are SR algorithms designed for infrared images. BSRN is a lightweight SR network. For a fair comparison, we retrain all the compared methods on our infrared UAV dataset. The above methods share the same platform and environment as the proposed method.

In the detection task, we selected nine typical methods with superior performance, namely, Cascade R-CNN [27], AutoAssign [67], RetinaNet [53], DDOD [60], YOLOX [59], RefineDet [57], YOLO v5 [55], ATSS [56], and DAGNet [25], as a comparison to perform detection on SR infrared UAV images. Cascade R-CNN is a classic two-stage object detection algorithm, while DAGNet and RefineDet integrate concepts from both one-stage and two-stage detection approaches. The remaining six algorithms, including RetinaNet, DDOD, YOLOX, ATSS, YOLO v5, and AutoAssign, are categorized as one-stage baseline methods. Among these, some employ an anchor-free design, while others achieve decoupling of localization and classification tasks. Note that for the Cascade R-CNN, AutoAssign, ATSS, DDOD, and RetinaNet, their backbone network is ResNet-50; for RefineDet, its backbone is VGG-16; for YOLO v5 and YOLOX, we choose its YOLO v5m configuration. The above methods also share the same platform and environment as ours.

4) Evaluation Metrics: We adopt the peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM) to quantitatively evaluate the performance of different SR methods. Furthermore, we utilize the contrast gain (CG) to evaluate the contrast enhancement ability of a small target relative to the background. In order to assess the target detection capability of different methods, we utilize the precision (P), recall (R), $F1$-score ($F1$), and mean average precision (mAP) to evaluate the detection methods quantitatively. The calculation details of all the above metrics can be found in the supplementary materials.

In particular, network parameters, floating-point operations per second (FLOPs), and frame per second (FPS) are also taken into account to measure the model complexity and computation cost.

B. Ablation Study

In this section, we further verify and analyze the effectiveness of different modules in the proposed method, and conduct ablation experiments of two subnetworks, the SCINet and the detection network. The details of the ablation study of the two networks are presented as follows.

1) Ablation Experiments of the SCINet: Several experiments are conducted to explore the contributions of different components in the proposed SCINet, i.e., the CEB, CCB, and MCB. Table I shows the quantitative experimental results. We select the BSRN network as our baseline method, as indicated in the third row of Table I. The CEB in the fourth row of Table I represents the ordinary CEB. We replace all the CE-RDB with the RDB and retrain the network from scratch. It is evident that the incorporation of CEB can improve SR performance (i.e., 0.024/0.022 PSNR/SSIM gain on average). This demonstrates that the goal of enhancing the contrast of the target is achieved, regardless of the type of CEB used. However, with the introduction of CCB, in comparison to regular CEB, which utilizes the plain convolution and CDC, CCB not only reduces the number of network parameters through decoupling but also introduces 0.913/0.015 gains on

Authorized licensed use limited to: XIDIAN UNIVERSITY. Downloaded on October 11,2024 at 13:03:53 UTC from IEEE Xplore. Restrictions apply.

TABLE I
EFFECTIVENESS OF CEB, CCB, MCB, AND DBFI

CEB	CCB	MCB	DBFI	Metrics		
				PSNR	SSIM	Params.(MB)
-	-	-	-	33.3605	0.9034	**0.336**
✓	-	-	-	34.1645	0.9095	2.531
✓	✓	-	-	34.2734	0.9183	2.525
✓	✓	✓	-	35.3652	0.9223	2.525
✓	✓	✓	✓	**36.0998**	**0.9250**	3.376

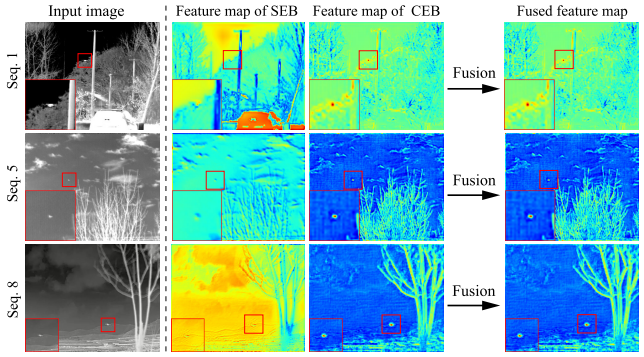


Fig. 8. Visualization of response maps from three representative sequences before and after applying SEB. The red box indicates the UAV target. The first column shows the input images. The second and third columns display the response maps obtained by SEB and CEB. The fourth column represents the response maps obtained by dual-branch fusion. The features of the target region are significantly enhanced after leveraging CEB.

PSNR/SSIM. This further demonstrates that the proposed CCB is capable of effectively restoring SR images with high-contrast targets. Following the integration of the MCB module, the network's capacity to extract multiscale features is further augmented. Consequently, both PSNR and SSIM obtain additional improvements. It is worth noting that MCB employs reparameterization, which improves the performance of the SR network without introducing additional computational costs in the inference stage. Furthermore, through DBFI, the CEB receives the intermediate-layer features from the SEB and is subsequently integrated into the SEB to provide additional local contrast for the SR process. The introduction of this scheme makes the model aggregates more gradient information, directing the network more focus toward high-frequency target edge features thus resulting in enhanced target contrast while improving the overall SR performance.

Fig. 8 exhibits the feature maps generated by the SEB, the CEB, and the dual branches, respectively. It becomes clear that relying solely on the SEB branch for SR tasks primarily extracts global features, yet lacks sufficient attention to target regions. In contrast, the CEB leverages the proposed CCB and MCB substantially enhancing the contrast of target regions. Beyond this, the CEB not only enhances target areas but also improves the high-frequency edges and intricate background details, such as the enhanced clarity of tree outlines, further confirming that the CEB can gather gradient information in different directions to make the network pay more attention to high-frequency features. However, the use of CEB alone still lacks a comprehensive understanding of the global context of the image, which is crucial for achieving high-quality SR tasks. Therefore, the proposed SCINet network integrates the merits of both SEB and CEB to promote the fusion of global context information and local high-frequency features, thereby simultaneously enhancing target contrast and SR performance.

TABLE II
EFFECTIVENESS OF THE SR, SHALLOWER DETECTION BRANCH, AND DFCM

SCINet	Shallower detection branch	DFCM	Metrics		
			P	R	F1
-	-	-	0.764	0.467	0.578
✓	-	-	0.963	0.935	0.948
✓	✓	-	0.964	0.952	0.957
✓	✓	✓	**0.976**	**0.964**	**0.970**

2) Ablation Experiments of the Multibranch Infrared UAV Detection Network: Several experiments are conducted to explore the contributions of different components, i.e., the shallower detection branch and the DFCM in the proposed detection network. We also test the effectiveness of the SR technique as a preprocessing step for the detection network. Table II shows the experimental results. It can be seen from Table II that compared with the detection results of LR images, the SR preprocessing by utilizing the SCINet greatly improves the detection performance, introducing 0.196/0.468/0.370 gains on P, R, and $F1$ on average. It is worth noting that the SR technique contributes the most to the metric R. The detection results of the SR images prove that the SR can effectively address some of the miss detection, which further explains from the aspect of the detection network that the proposed SCINet can effectively enhance the contrast between the target and the background, thereby greatly improving the detection performance of LR infrared UAV images.

The above results indicate that the SR technique can significantly improve the detection performance of LR infrared UAV images. However, due to the fact that the UAV target is small in size and easily overwhelmed by the complex background under long-range imaging, relying only on high-quality SR image reconstruction cannot fully address all the challenges in detection. Most networks, including YOLO v5, struggle to maintain the spatial features of weak and small-scale UAV targets, and thus often encounter difficulties in detecting UAV targets in complex backgrounds. To address these challenges, we specifically design a shallower detection branch for small-scale UAV targets and propose the DFCM to address the feature misalignment of deep and shallow layers. The shallower detection branch preserves the spatial attributes of small-scale UAV targets to the greatest extent. The DFCM calibrates and enhances the aggregated features of the different depth networks through the dual-dimension attention mechanism, which allows the model to focus on the target area and ignore irrelevant background information. The ablation experiment results in Table II suggest that the above two modules can gradually improve the detection performance, further validating the effectiveness of the proposed detection network.

C. Hyperparameters and Network Structure Analysis

In this section, we will study the design of the network structures in our model and examine the impact of different SR ratios on detection performance.

Authorized licensed use limited to: XIDIAN UNIVERSITY. Downloaded on October 11,2024 at 13:03:53 UTC from IEEE Xplore. Restrictions apply.

TABLE III
SR PERFORMANCE COMPARISON OF UPSAMPLING AND DUAL-BRANCH FUSION ORDER

Dual-branch fusion order	PSNR	SSIM	CG
PFAU	35.7962	0.9217	5.3319
PFBU	**36.0998**	**0.9250**	**5.3820**

1) Impact of Order of Upsampling and Dual-Branch Fusion on SR Performance: In SCINet architecture, the SEB is utilized to enhance the spatial detail features of images, while the CEB provides strong contrast prior knowledge to the SR task. This section investigates the impact of the order of upsampling and CEB and SEB dual-branch fusion on SR performance. Table III presents the experimental results. The second row in Table III denotes the postfusion after upsampling (PFAU), that is, we first upsample the features of the SEB branch and then fuse the contrast features of the CEB branch with it. The third row represents prefusion before upsampling (PFBU) which involves aggregating the features of the two branches first and then upsampling the fused features.

The results from Table III demonstrate that PFBU achieves superior SR performance. Compared to PFAU, PFBU comprehensively aggregates the spatial detail and contrast features before the SR reconstruction. Therefore, the SCINet can achieve a more accurate and refined reconstruction of the fused features, ensuring that the network can focus more on the feature information of the target area before upsampling. Therefore, in this work, we choose the PFBU to aggregate the feature guidance of both branches.

2) Impact of Different Scale Factors of SR on Detection Performance: The SR technique can enhance pixel-level details of LR images, which facilitates more accurate detection of infrared small-scale UAVs, and thus can improve the detection capability of small-scale UAV targets in LR images. In this section, we examine the impact of varying scale factors of SR on infrared small target detection performance to find the optimal scale factor that maximizes the performance of the detection network. In our experiments, we utilize the detection result of the LR images as the baseline result, equivalent to $1\times$ SR. We employ the SCINet to assess the impact of the factors of $2\times$, $4\times$, $8\times$, and $16\times$ SR on detection performance. The P, R, and $F1$ are exploited as metrics to measure the detection performance of the model comprehensively. It is well known that as the scale factor increases, the computational complexity of the model also increases, which could make the SR and detection models unsuitable for real-time applications. Therefore, FPS is also considered to assess the influence of the SR scale factor on detection performance, aiming to comprehensively select the ideal SR scale factor. Table IV presents the detection metrics and FPS at different scale factors on the validation set, while Fig. 9 shows the evolution curves of the $F1$ score and FPS under various scale factors of SR.

From Table IV and Fig. 9, it is evident that with the increase in scale factor of SR, there is a general improvement in the P, R, and $F1$. Compared to the LR image, the image with $2\times$ SR achieved significant gains on R/$F1$ (i.e., 0.298/0.214 on average). This demonstrates that the SCINet can effectively mitigate the problem of miss detections and false alarms of small UAV targets by recovering local salient features in the image and enhancing the contrast between the target and the background. Compared to the LR and $2\times$ SR, the $4\times$ SR images achieved over 0.950 in the P, R, and $F1$, indicating that the model performance has further effectively improved, and the FPS of the model is 53, satisfying the requirements for real-time detection. When the LR images are super-resolved by the scale factor of $8\times$ and $16\times$, the metrics tend to stabilize or decrease. This is because the proposed SCINet has well extracted and enhanced the contrast features of the UAV targets. Excessive scale factors can no longer bring more feature information about the UAV targets, and might even enhance background noise and clutter, negatively impacting detection. Moreover, when the scale factor of SR reached $8\times$ or higher, the FPS of the detection network declined below 30, failing to meet real-time detection requirements. Therefore, we focused on $4\times$ SR for further research to achieve a tradeoff between accuracy and efficiency.

TABLE IV
P, R, $F1$, AND FPS OF THE PROPOSED SCINET FOR DIFFERENT SCALE FACTORS ON LR IMAGES

Scale factor of SR	P	R	F1	FPS
LR images	0.764	0.467	0.578	84
$2\times$ SR	0.823	0.764	0.792	68
$4\times$ SR	0.976	0.964	0.970	53
$8\times$ SR	0.978	0.965	0.971	21
$16\times$ SR	0.977	0.965	0.970	5

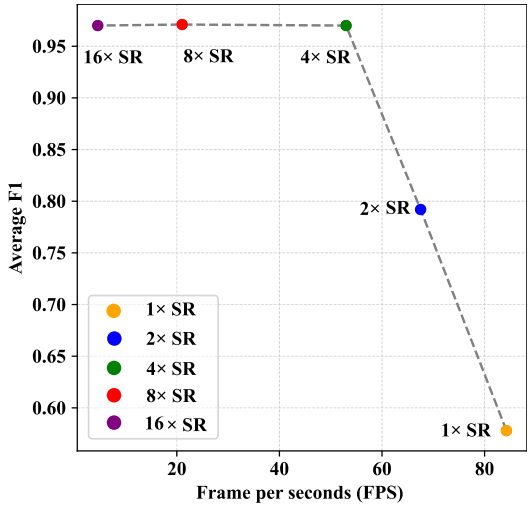


Fig. 9. Comparison of $F1$ and FPS of the different scale factors of SR on the validation set. A tradeoff between performance and efficiency is achieved with a factor is 4.

D. Comparison With SOTA Methods

In this section, extensive experiments of our method and the aforementioned comparative methods on nine real infrared sequences are performed to verify the superior performance of the proposed SR network and detection network.

Authorized licensed use limited to: XIDIAN UNIVERSITY. Downloaded on October 11,2024 at 13:03:53 UTC from IEEE Xplore. Restrictions apply.

TABLE V

QUANTITATIVE COMPARISONS OF THE PROPOSED METHOD AND OTHER SR METHODS. THE BEST PERFORMANCE IS HIGHLIGHTED IN RED, WHILE THE SECOND-BEST AND THIRD-BEST RESULTS ARE IN GREEN AND BLUE COLORS, RESPECTIVELY

# Seq.	Metrics	Bicubic	SRDenseNet	BSRN	SPSR	MoCoPNet	ChaSNet	IERN	PSRGAN	DASR	WaveMixSR	SCINet
1	PSNR	22.8812	25.7641	25.0865	22.4789	**26.4215**	**26.3989**	26.2177	21.2094	24.3761	20.8154	**26.2647**
	SSIM	0.6380	**0.8134**	0.7645	0.6676	0.8103	**0.8140**	0.8041	0.6095	0.8041	0.7415	**0.8143**
	CG	**1.8001**	1.5593	**1.8024**	1.6882	1.7886	1.7734	1.7792	1.7344	1.6500	1.2100	**1.8154**
	FPS	4	5	**19**	5	1	7	10	**19**	**12**	7	**16**
2	PSNR	31.2776	33.6139	34.2484	32.2245	**34.8030**	**34.9014**	34.7551	29.5835	30.1565	25.3147	**35.2375**
	SSIM	0.8604	0.9086	0.9150	0.8691	0.9159	**0.9254**	**0.9239**	0.8243	0.8065	0.7574	**0.9295**
	CG	**5.2411**	4.9798	5.2235	5.0202	5.2094	5.0031	5.0757	**5.4690**	5.0149	4.2087	**5.2432**
	FPS	5	8	**19**	5.102	1	6	11	**18**	14	12	**17**
3	PSNR	28.3981	31.4343	31.0152	28.6298	31.5413	**31.5667**	**31.6867**	26.5713	31.2541	28.5217	**31.5684**
	SSIM	0.7931	0.8797	0.8729	0.7932	0.8785	**0.8864**	**0.8853**	0.7405	0.8715	0.7942	**0.8870**
	CG	10.2011	10.3046	10.4893	10.5962	**10.7483**	10.4922	10.5894	**10.6678**	10.1123	10.1046	**10.8115**
	FPS	4	6	**18**	5	1	6	**11**	**18**	12	8	**17**
4	PSNR	34.2829	36.9093	36.8811	34.2089	37.4515b	**37.5865**	**37.5222**	31.6758	35.1764	31.1865	**37.5869**
	SSIM	0.8594	0.8845	**0.8848**	0.7935	0.8820	**0.8930**	0.8916	0.7886	0.8874	0.7165	**0.8862**
	CG	17.6475	17.1441	17.6633	16.7034	**17.7749**	**17.7024**	17.5993	17.5875	17.4907	16.4135	**17.9569**
	FPS	5	6	**18**	5	1	6	10	**19**	13	9	**17**
5	PSNR	25.1268	27.9562	26.4394	23.1806	27.4492	**28.4868**	**28.4399**	22.6235	24.4589	23.4891	**28.4873**
	SSIM	0.7745	0.8551	0.8318	0.7555	0.8532	**0.8690**	**0.8706**	0.7047	0.8561	0.7914	**0.8751**
	CG	4.3657	4.0822	**4.5823**	3.9085	4.5298	**4.5412**	4.4201	3.9759	4.0527	3.0614	**4.7325**
	FPS	5	6	**18**	5	1	7	11	**18**	**12**	8	**17**
6	PSNR	31.4227	34.2315	32.8480	30.3120	33.9690	**34.7538**	**34.7951**	29.0758	33.6498	29.9147	**34.9353**
	SSIM	0.8648	0.9163	0.8955	0.8403	0.9159	**0.9241**	**0.9259**	0.8128	0.9176	0.8719	**0.9259**
	CG	0.8925	0.8714	**0.8929**	0.8671	**0.9926**	0.8644	0.8713	0.8273	0.8639	0.8036	**1.0716**
	FPS	5	5	**18**	5	1	6	11	**18**	**12**	9	**17**
7	PSNR	35.7478	41.1208	39.6749	38.4344	39.7185	**41.4871**	**42.3269**	30.0087	38.6141	37.4427	**42.3791**
	SSIM	0.9695	**0.9812**	0.9747	0.9682	0.9801	0.9811	**0.9822**	0.9335	0.9475	0.8642	**0.9826**
	CG	2.6304	2.4863	2.6872	2.3466	**2.6941**	2.5876	2.5767	**2.6627**	2.6345	2.1201	**2.7567**
	FPS	4	5	**18**	5	1	6	10	**21**	13	8	**17**
8	PSNR	32.2404	33.0919	32.8663	30.5969	33.0238	**33.2276**	**33.4702**	28.1249	32.1914	30.9185	**33.5073**
	SSIM	0.9116	0.9278	0.9198	0.8940	**0.9285**	0.9259	**0.9283**	0.8661	0.8681	0.8192	**0.9294**
	CG	2.8353	2.3277	2.6451	**2.7948**	**2.7473**	2.6937	2.7215	3.2950	2.5237	2.2548	**2.8421**
	FPS	4	5	**18**	5	1	6	11	**21**	13	8	**17**
9	PSNR	29.4297	**36.2056**	32.1153	31.5009	34.7273	35.9188	**36.0015**	26.1996	35.9198	31.8918	**36.5086**
	SSIM	0.9905	0.9897	0.9899	0.9875	**0.9946**	**0.9915**	0.9912	0.9839	0.9418	0.8914	**0.9919**
	CG	1.0122	0.8489	0.9195	**0.9624**	**1.0071**	0.9129	0.9140	0.9029	0.9046	0.8447	**1.0914**
	FPS	4	6	**18**	5	1	7	11	**22**	12	9	**17**

TABLE VI

NETWORK PARAMETERS AND FLOPS OF EACH SR METHOD

Methods	Parameters (M)	FLOPs (G)
Bicubic	-	-
SRDenseNet	2.660	548.253
BSRN	0.344	13.657
SPSR	24.833	2272.256
MoCoPNet	4.829	8612.864
ChaSNet	14.479	590.959
IERN	1.747	178.161
PSRGAN	0.341	64.969
DASR	6.848	16.765
WaveMixSR	1.704	6622.776
SCINet	3.447	140.996

1) Quantitative Evaluation of SR Methods: Table V presents the quantitative comparison results of different SR methods. In addition to the commonly used SR performance evaluation metrics PSNR and SSIM, we introduce the tailored metric CG. The CG effectively measures the contrast difference between the target region and the surrounding background in the SR images and LR images, thus demonstrating the capability of different methods in target enhancement. In addition, the FPS is also utilized to evaluate the efficiency of a model. It can be observed that the SCINet outperforms most of the compared methods on PSNR and SSIM, and achieves the highest CG scores in most of the test sequences except Seq. 2, which demonstrates the superiority of our method. These scores indicate that our SCINet not only accurately generates high-quality SR images but also enhances the contrast between the UAV target and the background. This is attributed to the custom-designed CEB, which utilizes the CE-RDB to aggregate center-orient contrast. Furthermore, the CEB employs local and global residual learning for the transfer and compensation of local and global contrast features to further strengthen the intensity and contour details of the UAV targets. Notably, the interaction of features between the SEB and the CEB enables the SCINet can effectively improve the spatial details and resolution of LR images while enhancing the contrast of target regions. It is noteworthy that while the FPS of our method does not surpass that of PSRGAN or the BSRN, SCINet achieves a better SR performance with a reasonable increase in runtime, maintaining efficiency within an acceptable range.

Table VI displays the model parameters and FLOPs for the aforementioned comparison methods. Our method is relatively

Authorized licensed use limited to: XIDIAN UNIVERSITY. Downloaded on October 11,2024 at 13:03:53 UTC from IEEE Xplore. Restrictions apply.

TABLE VII

QUANTITATIVE COMPARISONS OF THE PROPOSED METHOD AND OTHER DETECTION METHODS. THE BEST THREE RESULTS ARE SHOWN IN RED, GREEN, AND BLUE FONTS, RESPECTIVELY

# Seq.	Metrics	Cascade R-CNN	AutoAssign	RetinaNet	DDOD	YOLOX	RefineDet	YOLO v5	ATSS	DAGNet	Ours
1	P	0.899	0.974	0.954	0.983	0.905	**0.998**	**0.994**	0.993	0.991	**1.000**
	R	**0.997**	0.978	0.990	0.985	0.992	**0.996**	0.996	**0.997**	0.962	**1.000**
	mAP@0.5	**0.996**	0.967	0.983	0.976	0.984	**0.990**	**0.995**	**0.995**	0.961	**0.996**
	FPS	22	34	33	33	**44**	30	**66**	33	27	**60**
2	P	0.653	0.561	0.749	0.634	0.755	**0.751**	**0.776**	0.719	0.525	**0.797**
	R	0.771	0.577	**0.905**	0.614	0.844	0.639	0.746	**0.904**	0.594	**0.897**
	mAP@0.5	0.634	0.374	**0.832**	0.489	0.811	0.501	0.767	**0.853**	0.209	**0.836**
	FPS	21	34	33	33	**46**	**38**	**42**	33	28	35
3	P	0.853	0.931	0.816	0.967	0.883	**0.983**	**0.994**	0.864	0.973	**1.000**
	R	0.818	0.914	0.919	0.907	0.913	0.818	**0.951**	**0.968**	0.731	**0.981**
	mAP@0.5	0.820	**0.914**	0.905	0.906	0.91	0.812	**0.977**	**0.977**	0.716	**0.994**
	FPS	23	35	34	34	**45**	35	**63**	36	28	**53**
4	P	0.824	0.953	0.848	0.950	0.960	0.984	**0.999**	**0.987**	0.942	**1.000**
	R	0.946	0.953	**0.992**	0.950	0.978	0.953	**0.997**	0.989	0.822	**1.000**
	mAP@0.5	0.925	0.919	0.988	0.920	0.976	0.941	**0.995**	**0.989**	0.782	**0.996**
	FPS	22	24	32	35	**44**	22	**67**	33	27	**59**
5	P	0.467	0.748	0.731	0.688	**0.924**	0.741	**0.905**	0.898	0.893	**0.913**
	R	0.455	0.568	0.847	0.331	**0.915**	0.662	0.812	**0.895**	0.833	**0.899**
	mAP@0.5	0.275	0.465	0.791	0.252	**0.920**	0.528	**0.884**	0.859	0.867	**0.919**
	FPS	25	32	33	33	**42**	36	**68**	34	28	**52**
6	P	0.500	0.989	0.800	0.977	**0.996**	0.977	**0.998**	0.967	0.951	**1.000**
	R	0.992	0.989	0.994	0.977	**0.997**	0.943	**0.998**	0.989	0.863	**1.000**
	mAP@0.5	0.759	0.978	0.984	0.973	**0.996**	0.926	**0.995**	0.975	0.827	**0.997**
	FPS	22	33	35	31	**39**	21	**53**	32	26	**41**
7	P	**0.984**	0.997	0.954	**0.997**	0.987	0.993	**1.000**	0.949	0.989	**1.000**
	R	0.985	0.997	0.987	**0.998**	**0.999**	0.9492	**1.000**	0.996	0.967	**1.000**
	mAP@0.5	0.982	0.994	0.969	**0.996**	0.992	0.987	**0.995**	0.984	0.859	**0.998**
	FPS	25	30	**31**	30	**41**	22	**41**	**31**	30	**32**
8	P	0.972	0.941	0.965	0.951	0.795	0.998	**0.999**	0.897	**0.997**	**1.000**
	R	0.917	0.911	**0.972**	**0.944**	0.922	0.895	0.918	0.919	0.868	**0.923**
	mAP@0.5	0.946	0.927	**0.964**	0.924	0.956	0.894	**0.959**	0.952	0.868	**0.960**
	FPS	22	33	32	33	**36**	22	**45**	32	25	**37**
9	P	0.987	0.980	0.938	0.986	0.997	0.995	**0.999**	0.915	**0.998**	**1.000**
	R	0.987	0.980	**0.997**	0.987	0.994	0.993	0.996	**0.995**	0.992	**1.000**
	mAP@0.5	0.986	0.973	0.989	0.981	**0.994**	0.992	**0.995**	0.988	0.992	**0.998**
	FPS	27	34	32	30	**42**	24	**52**	32	37	**45**

moderate in terms of parameters and FLOPs. This balance is particularly notable when upscaling images with a resolution of 128 × 160 by a scale of 4, where our method ranks third in FLOPs. Consequently, our SCINet attains superior SR performance without an excessive reliance on computational resources. This reveals that our method is highly cost-effective, achieving an optimal balance between precision and efficiency.

2) Quantitative Evaluation of Detection Methods: In this section, we present the quantitative evaluation for nine 4× SR image sequences of the proposed detection method along with nine typical target detection algorithms, as shown in Table VII. It can be observed from Table VII that the proposed detection network generally outperforms consistently in P, R, and mAP, especially in Seq. 1 and Seq. 4, where the target scale is relatively small, and our method reaches impressive 1.0 scores in P and R. This further proves that the introduction of the shallower layer detection branch significantly enhances the model capability of detecting weak small UAV targets in complex backgrounds. Our method consistently delivered outstanding detection performance on nine real complicated scenarios, which verified the robustness and adaptability of our method. This can be attributed to the proposed DFCM, which uses both spatial and semantic information to enhance the model's feature fusion capability, thereby effectively minimizing false alarms and miss detections of challenging targets. Furthermore, the FPS of our method consistently exceeds 30, which satisfies the real-time UAV surveillance.

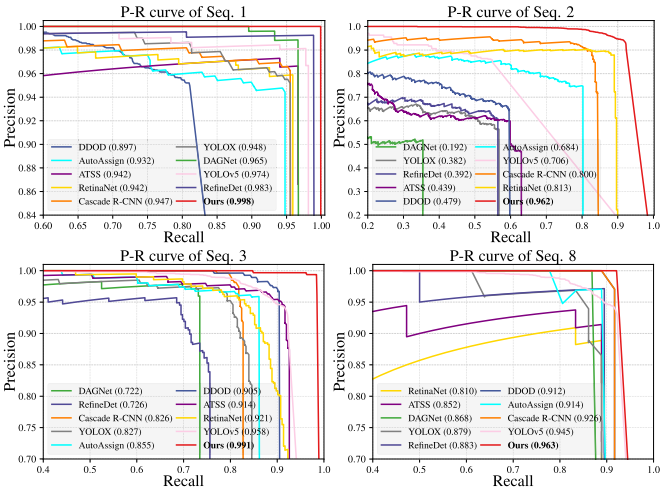


Fig. 10. P-R curves of Seq. 1, Seq. 2, Seq. 3, and Seq. 8. The area under the curve (AUC) values are placed after the methods' names.

Fig. 10 shows the P-R curves obtained by the proposed method and other methods for Seq. 1, Seq. 2, Seq. 3, and Seq. 8. It is widely acknowledged that the PR curve

Authorized licensed use limited to: XIDIAN UNIVERSITY. Downloaded on October 11,2024 at 13:03:53 UTC from IEEE Xplore. Restrictions apply.

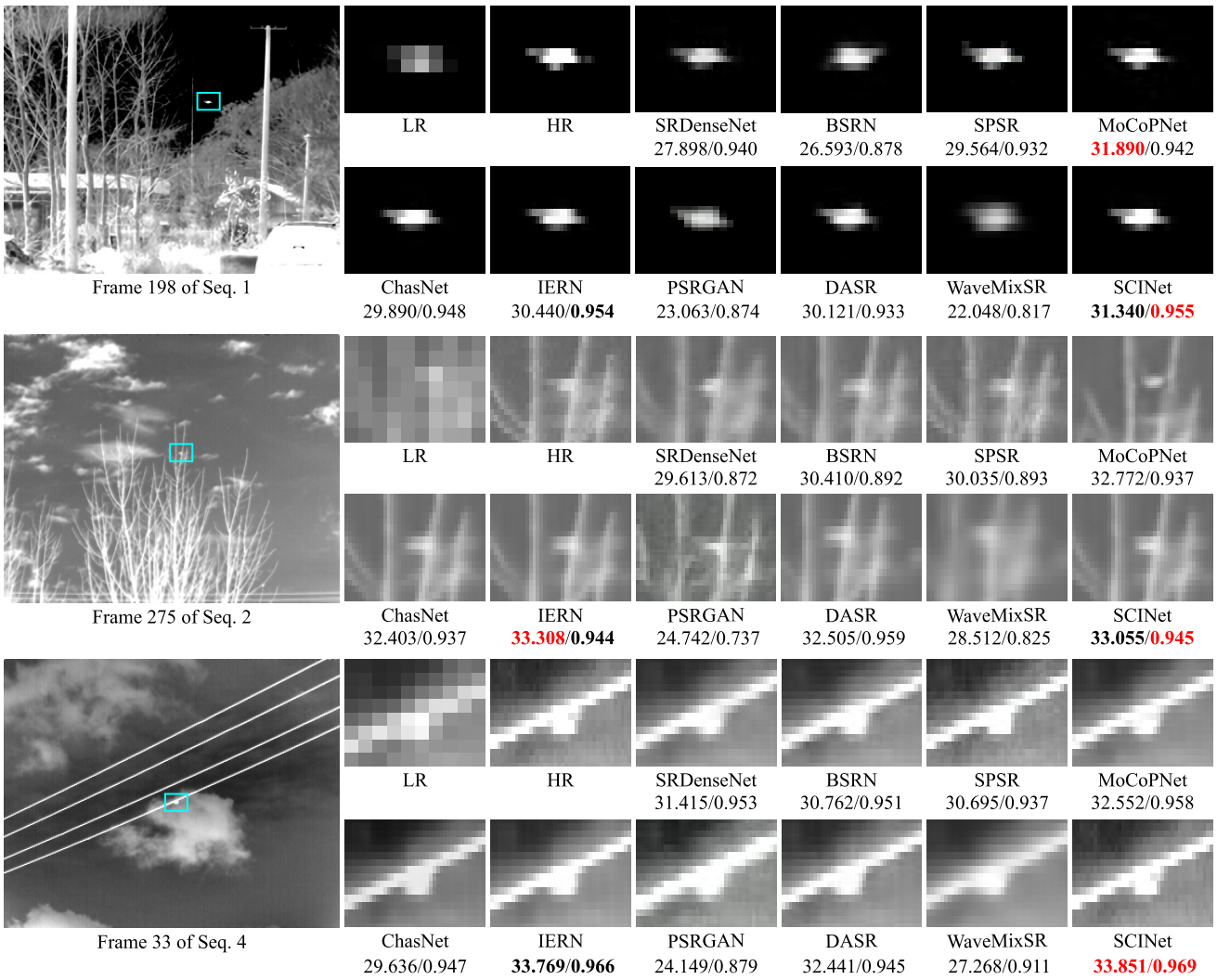


Fig. 11. Qualitative and quantitative comparison of our SCINet with SOTA SR methods on four LR images for 4× SR. In addition to the whole SR images with our SCINet (leftmost), we also present the close-up views of target areas in the LR image and the SR images restored using different methods. Below the SR images, the PSNR and SSIM of each close-up area are also displayed. The best and second-best results are highlighted with bold red and black text.

assesses precision at various recall levels, and comprehensively reflects the detection performance of a model. The area under the PR curve is known as the AP. A higher AP value indicates a superior detection performance. As illustrated in Fig. 10, our method achieves the highest AP values across four representative test sequences, further highlighting the superiority and effectiveness of our method.

Moreover, the model parameters and the FLOPs of each detection method are also listed in Table VIII. It can be seen from Tables VII and VIII that our method does not significantly increase the parameters and FLOPs, and the proposed method achieves superior performance than the two-stage method with fewer parameters. The proposed method shows a favorable accuracy-efficiency tradeoff compared with the SOTA models.

3) Impact of SR on Infrared Small Target Detection: To verify the impact of the SR on target detection performance, Table IX lists the detection results of three different detection algorithms on four representative sequences at different resolutions. The SR images are obtained using the proposed SCINet for 4× SR. The best and second-best results are represented in red and green fonts, respectively. It is evident that SR images, regardless of the detection algorithm, exhibit a significant improvement in detection performance compared to LR images. This improvement is especially noticeable in the challenging Seq. 5, characterized by extremely small

TABLE VIII
NETWORK PARAMETERS AND FLOPS OF EACH DETECTION METHOD

Methods	Parameters (M)	FLOPs (G)
Cascade R-CNN	68.93	153.39
AutoAssign	35.97	103.25
RetinaNet	36.33	155.35
DDOD	32.15	98.18
YOLOX	25.33	69.50
RefineDet	33.91	97.12
YOLO v5	21.32	103.21
ATSS	32.07	85.66
DAGNet	64.20	123.98
Ours	23.17	107.80

Authorized licensed use limited to: XIDIAN UNIVERSITY. Downloaded on October 11,2024 at 13:03:53 UTC from IEEE Xplore. Restrictions apply.

TABLE IX

QUANTITATIVE COMPARISON OF OUR METHOD AND OTHER SOTA DETECTION METHODS FOR IMAGES WITH DIFFERENT RESOLUTIONS ON FOUR SEQUENCES. THE BEST TWO RESULTS ARE SHOWN IN RED AND GREEN, RESPECTIVELY. THE UNDERLINE INDICATES THAT THE DETECTION PERFORMANCE RESULT OF SR IMAGES EXCEEDS THE ORIGINAL HR IMAGES

Methods	Resolution	Seq. 2			Seq. 3			Seq. 5			Seq. 8		
		P	R	mAP@0.5	P	R	mAP@0.5	P	R	mAP@0.5	P	R	mAP@0.5
YOLO v5	LR	0.554	0.047	0.325	0.861	0.0247	0.457	0.000	0.000	0.000	0.911	0.579	0.767
	SR	0.776	0.746	0.767	0.994	0.951	0.977	0.905	0.812	0.884	0.999	0.918	0.959
	HR	0.763	0.724	0.743	0.997	0.991	0.993	0.912	0.899	0.919	1.000	0.921	0.964
DAGNet	LR	0.131	0.124	0.119	0.250	0.003	0.001	0.000	0.000	0.000	0.413	0.052	0.035
	SR	0.525	0.592	0.209	0.973	0.731	0.716	0.893	0.833	0.867	1.000	0.868	0.868
	HR	0.526	0.599	0.254	0.989	0.745	0.725	0.864	0.847	0.882	1.000	0.763	0.763
Ours	LR	0.682	0.144	0.357	0.889	0.356	0.470	0.483	0.408	0.390	0.957	0.816	0.876
	SR	**0.797**	**0.897**	**0.836**	**1.000**	**0.981**	**0.994**	**0.913**	**0.899**	**0.919**	**1.000**	**0.923**	**0.960**
	HR	**0.848**	**0.837**	**0.857**	**1.000**	**0.996**	**0.955**	**0.995**	**0.872**	**0.934**	**1.000**	**0.931**	**0.978**

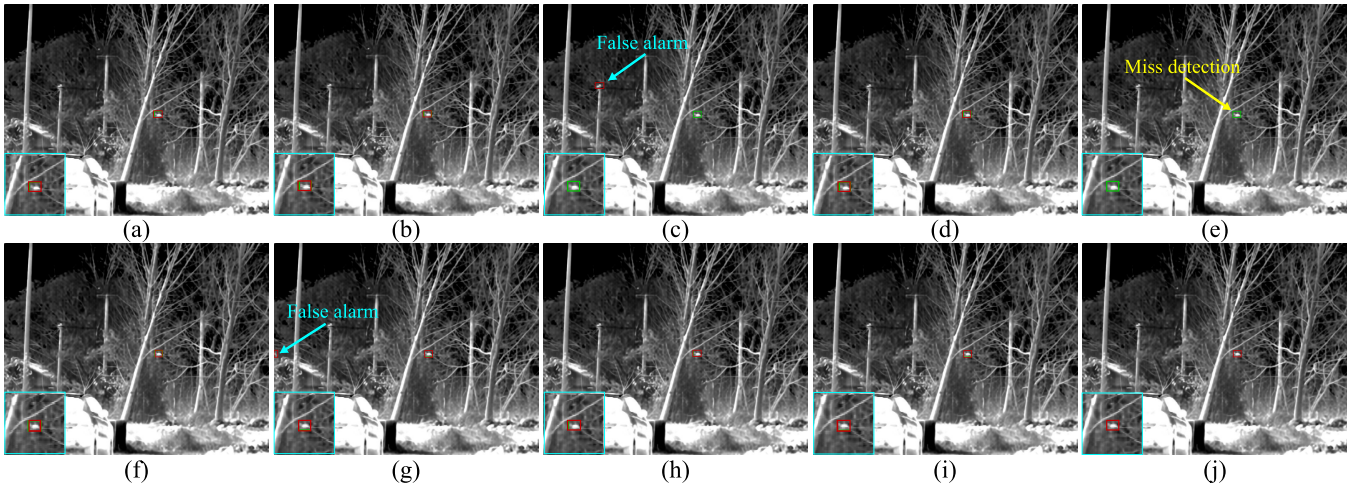


Fig. 12. Qualitative detection results of Frame No. 120 in Seq. 1 of 4× SR images generated by SCINet. The green and red boxes indicate the GT and the detection box, respectively. The detection method and confidence scores for each method are: (a) Cascade R-CNN, 0.94; (b) AutoAssign, 0.87; (c) RetinaNet, not detected and false alarm; (d) DDOD, 0.76; (e) YOLOX, not detected; (f) RefineDet, 0.88; (g) YOLO v5, 0.93 and false alarm; (h) ATSS, 0.66; (i) DAGNet, 0.90; and (j) Ours, 0.99.

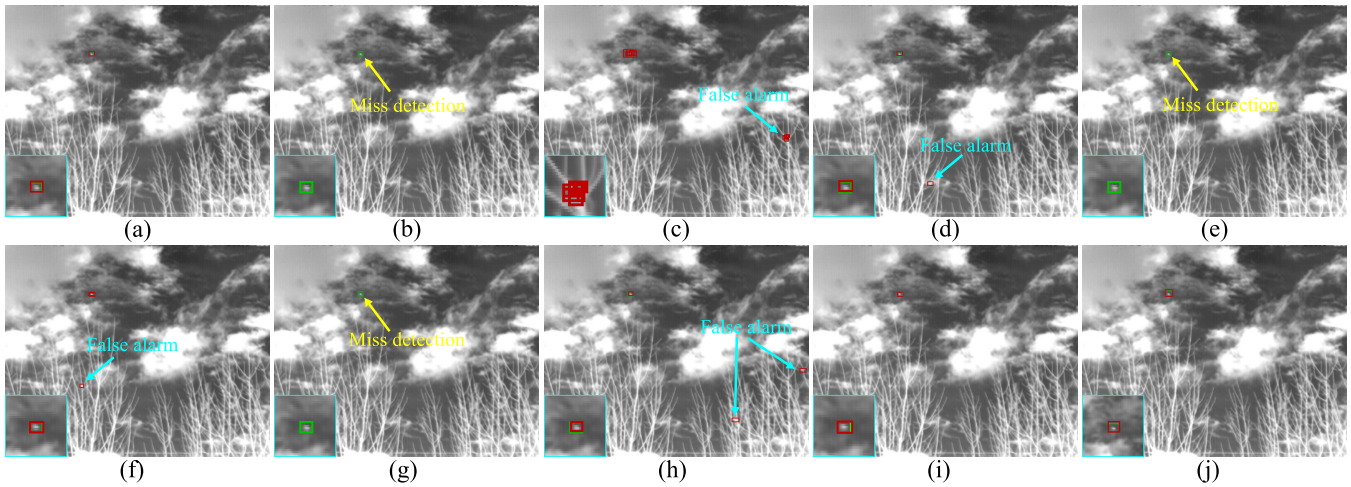


Fig. 13. Qualitative detection results of Frame No. 18 in Seq. 2 of 4× SR images generated by SCINet. The green and red boxes indicate the GT and the detection box, respectively. The detection method and confidence scores for each method are: (a) Cascade R-CNN, 0.90; (b) AutoAssign, not detected; (c) RetinaNet, false alarms; (d) DDOD, 0.72 and false alarm; (e) YOLOX, Not detected; (f) RefineDet, 0.84 and false alarm; (g) YOLO v5, not detected; (h) ATSS, 0.68 and false alarms; (i) DAGNet, 0.87; and (j) Ours, 1.00.

target scales. For Seq. 5, both the YOLO v5 and DAGNet achieve zero values for three metrics on LR images, whereas our SR network significantly improves image quality and contrast of the target area, alleviating the challenges in UAV detection in this sequence. This demonstrates that the proposed SR algorithm effectively enhances the contrast between the

Authorized licensed use limited to: XIDIAN UNIVERSITY. Downloaded on October 11,2024 at 13:03:53 UTC from IEEE Xplore. Restrictions apply.

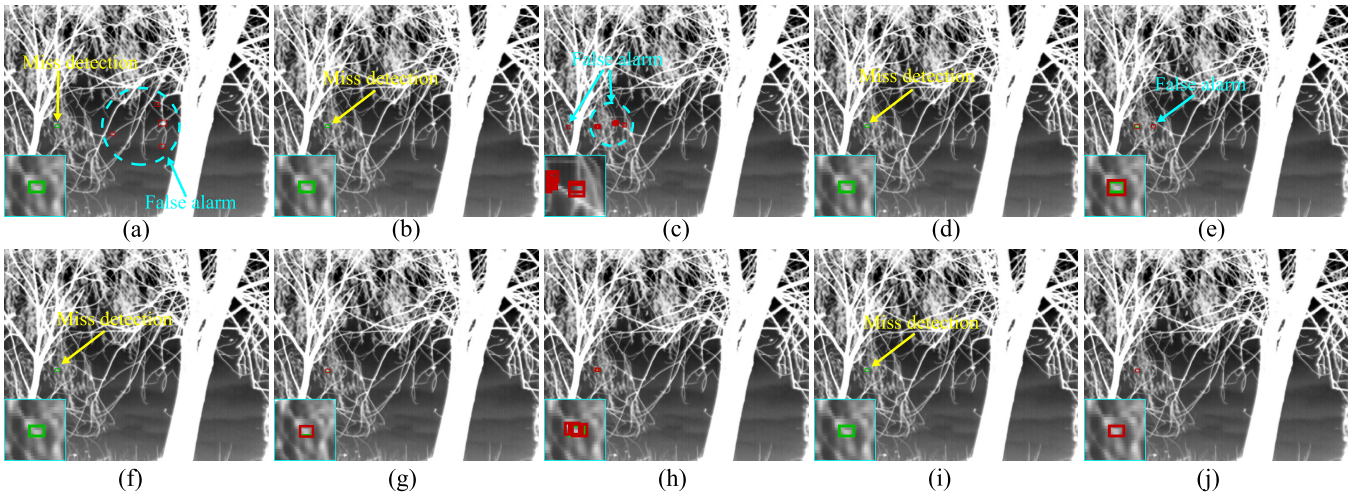


Fig. 14. Qualitative detection results of Frame No. 275 in Seq. 3 of 4× SR images generated by SCINet. The green and red boxes indicate the GT and the detection box, respectively. The detection method and confidence scores for each method are: (a) Cascade R-CNN, not detected and false alarms; (b) AutoAssign, not detected; (c) RetinaNet, 0.58 and false alarms; (d) DDOD, not detected; (e) YOLOX, 0.70 and false alarm; (f) RefineDet, not detected; (g) YOLO v5, 0.73; (h) ATSS, 0.68 and false alarms; (i) DAGNet, not detected; and (j) Ours, 0.93.

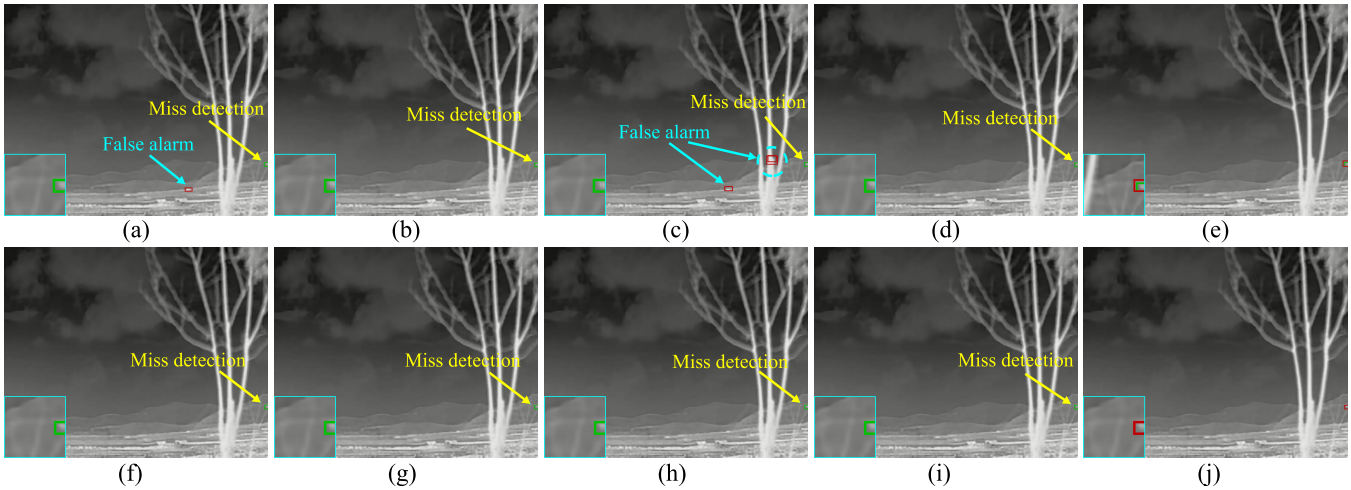


Fig. 15. Qualitative detection results of Frame No. 7 in Seq. 8 of 4× SR images generated by SCINet. The green and red boxes indicate the GT and the detection box, respectively. The detection method and confidence scores for each method are: (a) Cascade R-CNN, not detected and false alarm; (b) AutoAssign, not detected; (c) RetinaNet, not detected and false alarms; (d) DDOD, not detected; (e) YOLOX, 0.59; (f) RefineDet, not detected; (g) YOLO v5, not detected; (h) ATSS, not detected; (i) DAGNet, not detected; and (j) Ours, 0.87.

target and the background, thereby promoting the UAV target detection performance. Obviously, in Table IX, some detection results of SR images even surpass those of HR images. This suggests that the proposed SR method enhances discriminative features between the target and background, which again validates that the SR can improve detection performance.

4) Qualitative Evaluation of SR Methods on Synthetic LR Images: Fig. 11 presents the SR visual comparisons of our SCINet with SOTA SR methods for the representative frames of Seq. 1, Seq. 2, Seq. 4, and Seq. 8. The results for the remaining five sequences are included in the supplementary material. To better focus on the enhancement of target regions, we provide close-up views of the UAV targets for each SR result. Below each SR image, we also give the PSNR and SSIM of the target local area.

In terms of SR performance, our method generates SR images that are closer to the HR images, achieving the best or second-best results in PSNR and SSIM across all sequences. For instance, in Seq. 4, our method precisely recovers the sharp edges of targets and utility poles. Regarding target enhancement, SR images generated by our method exhibit significantly brighter target areas, indicating an effective improvement in the contrast of UAV targets. Take Seq. 1 as an example, where the SCINet notably enhances the contrast in the target area. Unlike the primary goal of other SR tasks to recover highly realistic SR images, we focus more on discriminative feature enhancement of the UAV target area, which can guide the detection network to pay more attention to the target rather than the background, thereby enhancing detection performance. In addition, our well-designed CEB aims to boost high-frequency features in images, so it may not perform as well as other methods in some low-frequency areas. However, this property also reduces the false alarms.

5) Qualitative Evaluation of Detection Methods: Figs. 12–15 are the qualitative detection results of our method and other comparison methods for Seq. 1, Seq. 2, Seq. 3, and

Authorized licensed use limited to: XIDIAN UNIVERSITY. Downloaded on October 11,2024 at 13:03:53 UTC from IEEE Xplore. Restrictions apply.

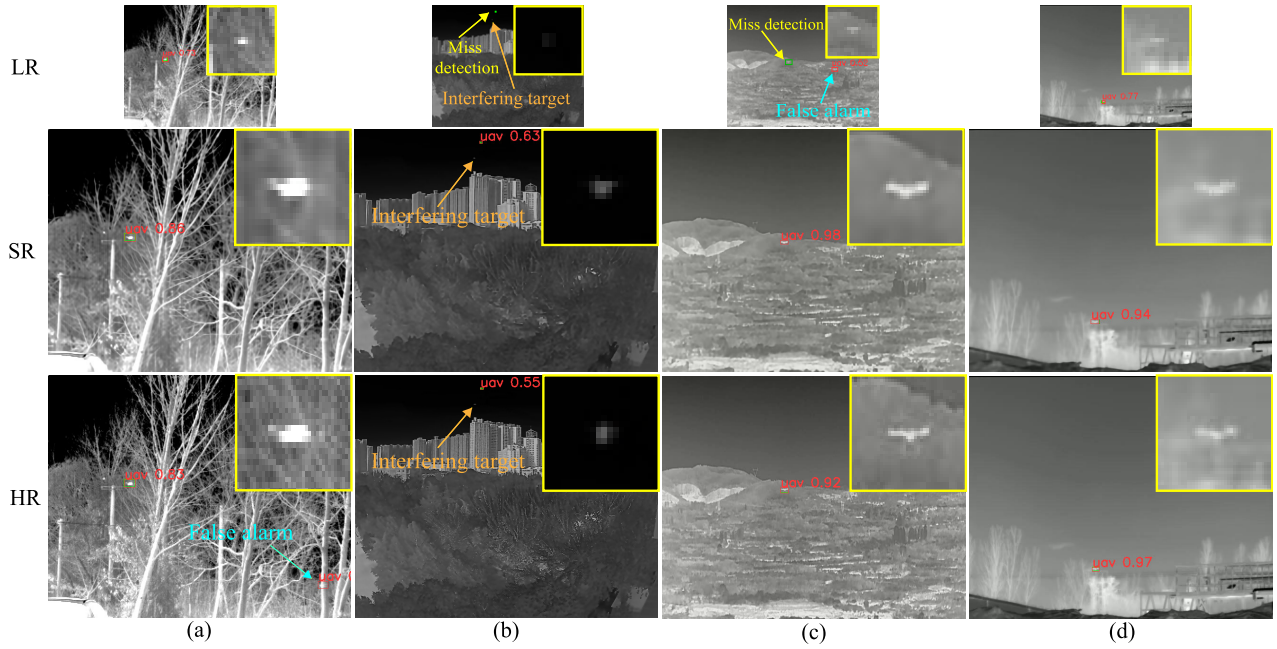


Fig. 16. Qualitative comparison of detection results for images of different resolutions in four sequence scenarios. The first, second, and third rows represent the detection results of LR, SR, and HR images, respectively. The SR and detection results are obtained by the proposed SCINet and detection network in this work, respectively. (a)–(d) 73rd, 54th, 68th, and 28th frames of Seqs. 1, 5, 6, and 7, respectively. The yellow area in the upper right corner is a local close-up of the target region under different resolution images.

TABLE X

QUANTITATIVE COMPARISON OF OUR SCINET AND OTHER SR METHODS ON REAL INFRARED IMAGES USING NO-REFERENCE EVALUATION METRICS NIQE, PIQE, AND BRISQUE. THE BEST AND SECOND BEST ARE BOLDED AND UNDERLINED, RESPECTIVELY. LOWER NIQE, PIQE, AND BRISQUE PROVE BETTER RESULTS

Metrics	Bicubic	SRDenseNet	BSRN	SPSR	MoCoPNet	ChaSNet	IERN	PSRGAN	SCINet
NIQE	16.2217	14.7711	14.6046	**12.6342**	14.2407	15.1596	14.8522	13.1375	<u>13.0932</u>
PIQE	68.7502	18.8547	29.7315	<u>13.1374</u>	32.4851	22.5272	25.1555	15.4645	**13.0970**
BRISQUE	66.5008	37.9603	49.0231	**35.4356**	45.2089	48.8657	49.7410	38.3210	<u>36.6796</u>

Seq. 8, respectively. Due to space limits, qualitative results of Seq. 4, Seq. 5, Seq. 6, Seq. 7, and Seq. 9 are placed in the supplementary material. In these figures, green bounding boxes represent the actual target locations of the UAV targets, while red boxes denote the detection results of different methods. In addition, to improve visibility, we provide a close-up view of the target region in the lower left corner of each image, and the target confidence score for each method is presented in the legend below. In the detection image, the term "False alarm" is highlighted in cyan, indicating a false detection, while "Miss detection" is marked in yellow, indicating the case where the target was not detected.

Figs. 12–15 illustrate exemplary cases of small UAV target detection in four distinct complex scenes. In Fig. 12 of Seq. 1, although most methods correctly identify the UAV target, there is an obvious gap in their confidence scores. Evidently, the RetinaNet and YOLO v5 show false alarms, and the YOLOX also exhibits miss detection. This is because the background of Seq. 1 is relatively complicated, making it challenging for one-stage networks to effectively eliminate the background interference. In contrast, our method not only detects UAV targets accurately but also achieves the highest confidence score, a fact that further verifies the effectiveness of the introduced DFCM. The DFCM can generate attention through both spatial and channel dimensions and then weight and calibrate the aggregated features, thus significantly enhancing the network's ability to distinguish UAV targets from the background and allowing the model to locate small-scale UAV targets in complex backgrounds. In Fig. 13, the UAV target is submerged in bright clouds and characterized by small and weak features. Most of the one-stage methods exhibit miss detection or false alarms. Although Cascade R-CNN and DAGNet can detect the targets, our method demonstrates a higher confidence score. This reveals the effectiveness of integrating the shallower layer detection branch. Fig. 14 presents a scenario where the UAV target is partially obscured by trees and shares similar features with the foliage, making it challenging for most methods to extract distinctive characteristics, leading to numerous false alarms and miss detections. The RefineDet and DAGNet did not yield satisfactory results, and even the two-stage algorithm Cascade R-CNN also presents miss detection. By contrast, our method successfully detects the target and consistently maintains a high confidence score. In Fig. 15, the UAV target is the edge

Authorized licensed use limited to: XIDIAN UNIVERSITY. Downloaded on October 11,2024 at 13:03:53 UTC from IEEE Xplore. Restrictions apply.

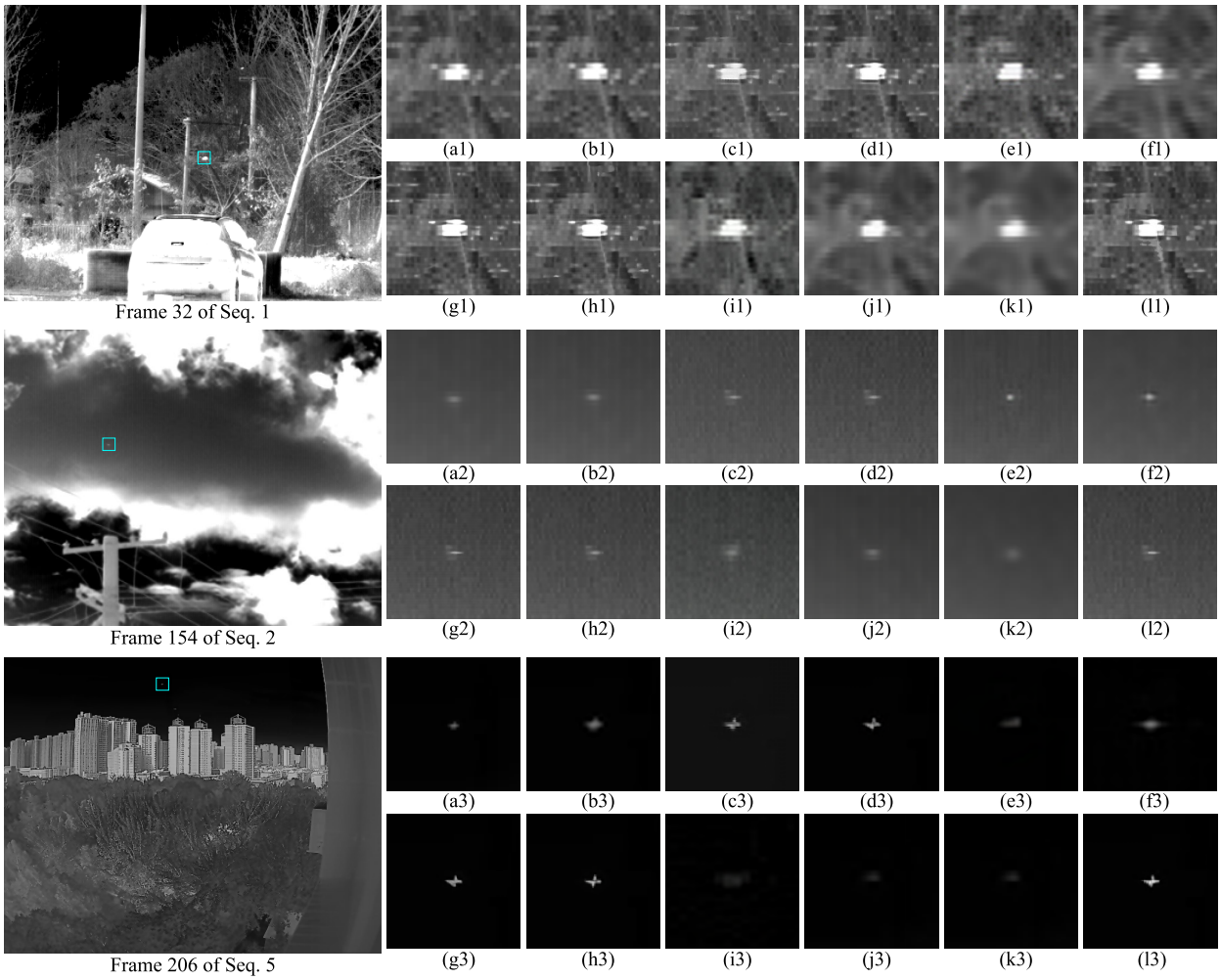


Fig. 17. Close-up visual comparison for 4× SR from different methods on captured infrared UAV images. (a)–(l) Original HR UAV image and 4× SR images of Bicubic, SRDenseNet, BSRN, SPSR, MoCoPNet, ChaSNet, IERN, PSRGASN, DASR, WaveMixSR, and Ours, respectively.

of the field of view. At this time, the UAV target is almost invisible, making it difficult for most methods to focus on such limited feature information. However, our method still effectively detects the UAV target and demonstrates robustness in various complex backgrounds.

6) Qualitative Evaluation of the Impact of SR on Infrared Small Target Detection: Fig. 16 presents a qualitative comparison of detection results for images of different resolutions in four representative sequences. The first, second, and third rows in Fig. 16 show the detection results for LR images, 4× SR images, and HR images, respectively. In the experiments of this section, both the SR and detection networks employ the proposed methods in this article. It is observed from Fig. 16 that the size of UAV targets in LR images is extremely tiny, which frequently results in miss detections. However, the UAV targets after the SR are accurately detected, and the confidence scores are also significantly improved compared to those of the LR images. From Fig. 16(a) to (c), it is evident that the SR images even achieve better detection performance than the HR images, effectively eliminating false alarms and enhancing confidence scores. This improvement is attributed to two key factors. First, the SR images obtained through the SCINet enhance the high-frequency features in the target area, thereby increasing the contrast difference between the target and the background, which is conducive to target detection. Second, the detection network proposed in our study dynamically calibrates and fuses features from different depths of the network, ensuring that the shallower layer branch retains abundant feature information for infrared small UAV targets, thereby enhancing detection performance.

7) Additional Experiments of the SR Network SCINet on Real Infrared UAV Images: To validate the effectiveness of the proposed SR method, we conduct experiments on infrared images collected from real-world scenarios. Due to the absence of corresponding high-quality GT images for captured thermal infrared images, it is infeasible to employ GT-based evaluation metrics, such as the PSNR and SSIM. Consequently, we adopt the natural image quality evaluator (NIQE) [68], perception-based image quality evaluator (PIQE) [69], and no-reference image spatial quality evaluator (BRISQUE) [70], which are widely accepted no-reference image quality assessment metrics in the field of image quality evaluation. The quantitative assessment results are presented in Table X. The lower the NIQE, PIQE, and BRISQUE, the better the SR performance. It can be seen from Table X that the proposed SCINet demonstrates commendable performance in NIQE, PIQE, and BRISQUE metrics, compared to other methods.

Authorized licensed use limited to: XIDIAN UNIVERSITY. Downloaded on October 11,2024 at 13:03:53 UTC from IEEE Xplore. Restrictions apply.

Of these metrics, PIQE scores are the most favorable, while the performance in NIQE and BRISQUE is second only to SPSR, ranking second among test datasets under real-world degradation. This suggests that the proposed SR method cannot only solve the SR task of synthetic LR infrared images but also achieve satisfactory results on real infrared images. Qualitative comparison results of SR reconstruction of real infrared UAV images are shown in Fig. 17. It can be observed that SCINet can recover more refined and realistic target areas, including the wings of UAV targets and the high-frequency edges of tree trunks. In addition, with the help of the CEB, our method can significantly improve the contrast of the UAV target area and increase the feature difference from the background.

V. Conclusion

In this article, we introduce a novel SR network for assisting infrared small UAV target detection under complex backgrounds in LR images. It incorporates the proposed SR model SCINet into the detection network to enhance the UAV image resolution and quality. The SCINet introduces a customized CEB to provide a contrast prior to reconstructing high-quality SR images, and enhances spatial and contrast features through the feature interaction of the CEB and SEB, effectively improving SR performance and visual quality. Furthermore, we propose a multibranch infrared UAV target detection network with the DFCM module. The DFCM leverages spatial and channel dual-dimensional attention mechanisms to recalibrate and enhance the aggregated features between different layers, enabling the model to better learn the discriminative features of the target regions. Extensive experimental results demonstrate that the proposed SCINet and the detection network outperform SOTA SR and detection methods. The research in this work highlights the effectiveness of SR in infrared small UAV target detection tasks and provides an effective way to improve the performance of infrared UAV target detection. In the future, we will further improve the efficiency of our SR method to achieve real-time performance for practical applications. In addition, we also will explore image SR and target detection in an end-to-end manner within a unified framework.

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] J. Xie, J. Yu, J. Wu, Z. Shi, and J. Chen, "Adaptive switching spatial–temporal fusion detection for remote flying drones," *IEEE Trans. Veh. Technol.*, vol. 69, no. 7, pp. 6964–6976, Jul. 2020.

[3] N. P. Bertrand, J. Lee, K. F. Prussing, S. Shapero, and C. J. Rozell, "Infrared search and track with unbalanced optimal transport dynamics regularization," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 12, pp. 2072–2076, Dec. 2021.

[4] E. Vattapparamban, I. Güvenç, A. I. Yurekli, K. Akkaya, and S. Uluağaç, "Drones for smart cities: Issues in cybersecurity, privacy, and public safety," in *Proc. Int. Wireless Commun. Mobile Comput. Conf. (IWCMC)*, 2016, pp. 216–221.

[5] X. Shi, C. Yang, W. Xie, C. Liang, Z. Shi, and J. Chen, "Anti-drone system with multiple surveillance technologies: Architecture, implementation, and challenges," *IEEE Commun. Mag.*, vol. 56, no. 4, pp. 68–74, Apr. 2018.

[6] Y. Zhao and Y. Su, "Cyclostationary phase analysis on micro-Doppler parameters for radar-based small UAVs detection," *IEEE Trans. Instrum. Meas.*, vol. 67, no. 9, pp. 2048–2057, Sep. 2018.

[7] R. K. Willardson, E. R. Weber, D. D. Skatrud, and P. W. Kruse, *Uncooled Infrared Imaging Arrays and Systems*. New York, NY, USA: Academic, 1997.

[8] S. Deane et al., "Comparison of cooled and uncooled IR sensors by means of signal-to-noise ratio for NDT diagnostics (don't short) of aerospace grade composites," *Sensors*, vol. 20, no. 12, p. 3381, Jun. 2020.

[9] N. J. W. Morris, S. Avidan, W. Matusik, and H. Pfister, "Statistics of infrared images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–7.

[10] Z. Zhao, S. Xu, C. Zhang, J. Liu, and J. Zhang, "Bayesian fusion for infrared and visible images," *Signal Process.*, vol. 177, Dec. 2020, Art. no. 107734.

[11] J. Chen, K. Wu, Z. Cheng, and L. Luo, "A saliency-based multiscale approach for infrared and visible image fusion," *Signal Process.*, vol. 182, May 2021, Art. no. 107936.

[12] S. D. Deshpande, M. H. Er, R. Venkateswarlu, and P. Chan, "Max-mean and max-median filters for detection of small targets," *Proc. SPIE*, vol. 3809, pp. 74–83, Jul. 1999.

[13] M. Zeng, J. Li, and Z. Peng, "The design of top-hat morphological filter and application to infrared target detection," *Infr. Phys. Technol.*, vol. 48, no. 1, pp. 67–76, Apr. 2006.

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

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

[16] Y. Dai and Y. Wu, "Reweighted infrared patch-tensor model with both nonlocal and local priors for single-frame small target detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 8, pp. 3752–3767, Aug. 2017.

[17] Y. Shi, Y. Wei, H. Yao, D. Pan, and G. Xiao, "High-boost-based multiscale local contrast measure for infrared small target detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 1, pp. 33–37, Jan. 2018.

[18] S. Yao, Y. Chang, and X. Qin, "A coarse-to-fine method for infrared small target detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 2, pp. 256–260, Feb. 2019.

[19] L. Zhang and Z. Peng, "Infrared small target detection based on partial sum of the tensor nuclear norm," *Remote Sens.*, vol. 11, no. 4, p. 382, Feb. 2019.

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

[21] P. Zhang, L. Zhang, X. Wang, F. Shen, T. Pu, and C. Fei, "Edge and corner awareness-based spatial–temporal tensor model for infrared small-target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 12, pp. 10708–10724, Dec. 2021.

[22] Y. Sun, J. Yang, and W. An, "Infrared dim and small target detection via multiple subspace learning and spatial–temporal patch-tensor model," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 3737–3752, May 2021.

[23] 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. (WACV)*, Jan. 2021, pp. 949–958.

[24] M. Zhang, R. Zhang, Y. Yang, H. Bai, J. Zhang, and J. Guo, "ISNet: Shape matters for infrared small target detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 867–876.

[25] H. Fang, Z. Liao, X. Wang, Y. Chang, and L. Yan, "Differentiated attention guided network over hierarchical and aggregated features for intelligent UAV surveillance," *IEEE Trans. Ind. Informat.*, vol. 19, no. 9, pp. 9909–9920, Sep. 2023.

[26] H. Fang et al., "DANet: Multi-scale UAV target detection with dynamic feature perception and scale-aware knowledge distillation," in *Proc. 31st ACM Int. Conf. Multimedia*, vol. 19, Oct. 2023, pp. 2121–2130.

[27] 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.

Authorized licensed use limited to: XIDIAN UNIVERSITY. Downloaded on October 11,2024 at 13:03:53 UTC from IEEE Xplore. Restrictions apply.

[28] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: A simple and strong anchor-free object detector," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 4, pp. 1922–1933, Apr. 2020.

[29] G. Suryanarayana and R. Dhuli, "Simultaneous edge preserving and noise mitigating image super-resolution algorithm," *AEU-Int. J. Electron. Commun.*, vol. 70, no. 4, pp. 409–415, Apr. 2016.

[30] K. Ren, Y. Gao, M. Wan, G. Gu, and Q. Chen, "Infrared small target detection via region super resolution generative adversarial network," *Appl. Intell.*, vol. 52, no. 10, pp. 11725–11737, Aug. 2022.

[31] R. Li and Y. Shen, "YOLOSR-IST: A deep learning method for small target detection in infrared remote sensing images based on super-resolution and Yolo," *Signal Process.*, vol. 208, Jul. 2023, Art. no. 108962.

[32] K. I. Kim and Y. Kwon, "Single-image super-resolution using sparse regression and natural image prior," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 6, pp. 1127–1133, Jun. 2010.

[33] Z. Xiong, X. Sun, and F. Wu, "Robust web image/video super-resolution," *IEEE Trans. Image Process.*, vol. 19, no. 8, pp. 2017–2028, Aug. 2010.

[34] R. Keys, "Cubic convolution interpolation for digital image processing," *IEEE Trans. Acoust. Speech Signal Process.*, vol. ASSP-29, no. 6, pp. 1153–1160, Dec. 1981.

[35] M. Irani and S. Peleg, "Improving resolution by image registration," *CVGIP, Graph. Models Image Process.*, vol. 53, no. 3, pp. 231–239, May 1991.

[36] J. Yang, J. Wright, T. Huang, and Y. Ma, "Image super-resolution as sparse representation of raw image patches," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[37] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Trans. Image Process.*, vol. 19, no. 11, pp. 2861–2873, Nov. 2010.

[38] C. Ma, Y. Rao, J. Lu, and J. Zhou, "Structure-preserving image super-resolution," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 7898–7911, Nov. 2022.

[39] P. Wei et al., "Component divide-and-conquer for real-world image super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 101–117.

[40] X. Zhang, C. Li, Q. Meng, S. Liu, Y. Zhang, and J. Wang, "Infrared image super resolution by combining compressive sensing and deep learning," *Sensors*, vol. 18, no. 8, p. 2587, Aug. 2018.

[41] T. Y. Han, Y. J. Kim, and B. C. Song, "Convolutional neural network-based infrared image super resolution under low light environment," in *Proc. 25th Eur. Signal Process. Conf. (EUSIPCO)*, Aug. 2017, pp. 803–807.

[42] Z. He, S. Tang, J. Yang, Y. Cao, M. Y. Yang, and Y. Cao, "Cascaded deep networks with multiple receptive fields for infrared image super-resolution," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 8, pp. 2310–2322, Aug. 2019.

[43] L. Chen, R. Tang, M. Anisetti, and X. Yang, "A lightweight iterative error reconstruction network for infrared image super-resolution in smart grid," *Sustain. Cities Soc.*, vol. 66, Mar. 2021, Art. no. 102520.

[44] K. Prajapati et al., "Channel split convolutional neural network (ChaSNet) for thermal image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 4363–4372.

[45] Z. Li et al., "Blueprint separable residual network for efficient image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 833–843.

[46] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1646–1654.

[47] J. Kim, J. K. Lee, and K. M. Lee, "Deeply-recursive convolutional network for image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1637–1645.

[48] T. Tong, G. Li, X. Liu, and Q. Gao, "Image super-resolution using dense skip connections," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4809–4817.

[49] X. Wang et al., "ESRGAN: Enhanced super-resolution generative adversarial networks," in *Proc. Eur. Conf. Comput. Vis. Workshops (ECCVW)*, Sep. 2018, pp. 63–79.

[50] J. Liu, J. Tang, and G. Wu, "Residual feature distillation network for lightweight image super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, Jan. 2020, pp. 41–55.

[51] Y. Huang, Z. Jiang, R. Lan, S. Zhang, and K. Pi, "Infrared image super-resolution via transfer learning and PSRGAN," *IEEE Signal Process. Lett.*, vol. 28, pp. 982–986, 2021.

[52] W. Liu et al., "SSD: Single shot MultiBox detector," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, Oct. 2016, pp. 21–37.

[53] T. 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.

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

[55] G. Jocher, Oct. 2021, "Ultralytics/YOLOv5: V6.0—YOLOv5n 'nano' models, roboflow integration, TensorFlow export, OpenCV DNN support," doi: 10.5281/zenodo.5563715.

[56] S. Zhang, C. Chi, Y. Yao, Z. Lei, and S. Z. Li, "Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9759–9768.

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

[58] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9626–9635.

[59] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "YOLOX: Exceeding YOLO series in 2021," 2021, *arXiv:2107.08430*.

[60] Z. Chen, C. Yang, Q. Li, F. Zhao, Z.-J. Zha, and F. Wu, "Disentangle your dense object detector," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 4939–4948.

[61] F. Xiaolin et al., "Small object detection in remote sensing images based on super-resolution," *Pattern Recognit. Lett.*, vol. 153, pp. 107–112, Jan. 2022.

[62] X. Ying et al., "Local motion and contrast priors driven deep network for infrared small target superresolution," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 5480–5495, 2022.

[63] J. Zhang, J. Lei, W. Xie, Z. Fang, Y. Li, and Q. Du, "SuperYOLO: Super resolution assisted object detection in multimodal remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, pp. 1–15.

[64] J. Zhao et al., "The 2nd anti-UAV workshop & challenge: Methods and results," 2021, *arXiv:2108.09909*.

[65] S. Liang, K. Song, W. Zhao, S. Li, and Y. Yan, "DASR: Dual-attention transformer for infrared image super-resolution," *Infr. Phys. Technol.*, vol. 133, Sep. 2023, Art. no. 104837.

[66] P. Jeevan, A. Srinidhi, P. Prathiba, and A. Sethi, "WaveMixSR: Resource-efficient neural network for image super-resolution," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, vol. 13, Jan. 2024, pp. 5872–5880.

[67] B. Zhu et al., "AutoAssign: Differentiable label assignment for dense object detection," 2020, *arXiv:2007.03496*.

[68] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a 'completely blind' image quality analyzer," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, Mar. 2013.

[69] N. Venkatanath, D. Venkatanath, M. Chandrasekhar, S. S. Channappayya, and S. S. Medasani, "Blind image quality evaluation using perception based features," in *Proc. 21st Nat. Conf. Commun. (NCC)*, Feb. 2015, pp. 1–6.

[70] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4695–4708, Dec. 2012.



Houzhang Fang (Member, IEEE) received the Ph.D. degree in control science and engineering from 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.

Authorized licensed use limited to: XIDIAN UNIVERSITY. Downloaded on October 11,2024 at 13:03:53 UTC from IEEE Xplore. Restrictions apply.



Lan Ding received the B.S. degree in software engineering from Ningxia University, Yinchuan, China, in 2021, and the M.S. degree in electronic information (software engineering direction) from Xidian University, Xi'an, China, in 2024.

She is currently an Engineer with iFLYTEK Company Ltd., Hefei, China. Her research interests include computer vision and deep learning.



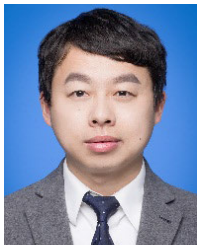
Xiaolin Wang received the bachelor's and master's degrees from the School of Computer Science and Technology, Xidian University, Xi'an, China, in 2020 and 2023, respectively, where he is currently pursuing the Ph.D. degree.

His research interests include object tracking, object detection, and infrared image processing.



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 Huazhong University of Science and Technology (HUST), Wuhan, China, in 2014 and 2019, respectively.

He is currently an Associate 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 systems from 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.



Li Liu (Member, IEEE) received the Ph.D. degree in control science and technology from Huazhong University of Science and Technology, Wuhan, China, in 2016.

He has been an Assistant Professor with Xidian University, Xi'an, China, since 2016. Since 2021, he has been a Visiting Scholar with the University of Helsinki, Helsinki, Finland, where he was a Post-Doctoral Researcher from 2021 to 2023. His main research interests include infrared image correction, image deblurring, image enhancement, object detection, and recognition.



Jinrui Fang received the B.S. degree in informatics: data science from the University of Washington, Seattle, WA, USA, in 2023. He is currently pursuing the M.S. degree in data science with New York University, New York City, NY, USA.

His research interests include deep learning and natural language processing.

Authorized licensed use limited to: XIDIAN UNIVERSITY. Downloaded on October 11,2024 at 13:03:53 UTC from IEEE Xplore. Restrictions apply.