LAST-Net: Local Adaptivity Spatial Transformer Network for Multiobject Detection in UAV Remote Sensing Thermal Infrared Imagery

Min Li, Jinhui Lan, Ying Zhang, and Kun Huang

***Abstract*— Object detection in drone infrared remote sensing is a critical technology in the fields of intelligent transportation and computer vision. However, this technology faces numerous challenges due to its inherent characteristics, such as complex background interference that hinders semantic scene understanding, low-contrast images that lead to the loss and blurring of feature information, and disparities in object scale and data distribution that result in mismatches in the feature map. Existing studies have focused on improving individual issues using convolutional neural networks (CNNs) or transformer architectures, but they have not achieved satisfactory results in UAV infrared scenarios. This article proposes the local adaptivity spatial transformer network (LAST-Net) to address these challenges: 1) to address feature extraction in complex remote sensing scenes, the convolutional spatial transformer block (CST-Block) is proposed, which effectively integrates convolutional operations and transformer architecture through parallel computation to extract both local and global features, while suppressing background interference; 2) to address feature fusion for the low-contrast issue in infrared images, the dual-modality adaptive attention (DMAA) mechanism is proposed, which leverages the interaction and feedback between content and spatial matrices to enable coarse-to-fine queries, thereby effectively aggregating object information; and 3) to handle object scale variation and sample imbalance, the coordinate crossing loss function (CCLF) is proposed, which dynamically adjusts weight factors to match different scales and multiscene objects. Experiments on the DroneVehicle and HIT-UAV datasets show that LAST-Net achieves mean average precisions (mAPs) of 86.3% and 90.9%, surpassing the state-of-the-art algorithms.**

***Index Terms*— Attention mechanism, drone, infrared remote sensing imagery, small object detection, transformer.**

I. Introduction

DRONE-BASED remote sensing object detection is essential in military applications [1], intelligent transportation [2], and natural disaster early warning [3]. In contrast to visible light cameras, thermal infrared cameras generate images by detecting temperature variations among objects rather than depending on external light sources. This capability enables effective operation in low-light conditions and through haze. However, the inherent characteristics of infrared images and remote sensing images still present numerous challenges for UAV-based infrared object detection [4], [5], [6]. (Challenge 1) Blurred semantic information and spatial details due to complex backgrounds: drone images are captured from a bird's-eye view, often contain dense and disordered objects that may be obscured by vegetation or buildings [Fig. 1(a) and (b)], leading to blurred semantic information and spatial detail. (Challenge 2) Low contrast issues: in infrared images, when the object's temperature closely matches that of the background, its edges become blurred, leading to reduced contrast [Fig. 1(c) and (d)]. (Challenge 3) Multiscale objects and imbalanced data sample issues: current UAV-based infrared datasets are relatively scarce, and variations in flight altitude result in multiscale changes in object size within images [Fig. 1(e) and (f)], and the varying numbers of road objects lead to differences in label quantities across the existing datasets [Fig. 1(g) and (h)].

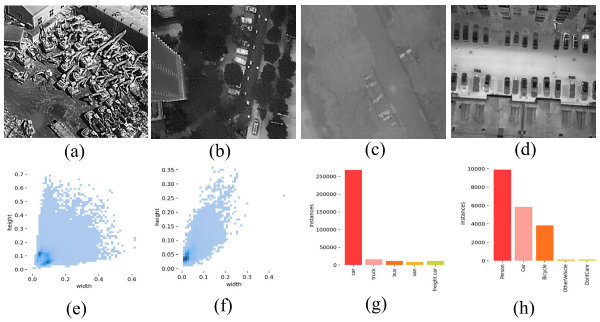

Fig. 1. UAV-based infrared images under different conditions. (a) Dense object scene in complex background environment. (b) Occluded object scene. (c) Low-contrast image. (d) Different feature representations of objects within the same scene. (e) Aspect ratios of object sizes relative to the overall image in the DroneVehicle dataset. (f) Aspect ratios of object sizes relative to the overall image in the HIT-UAV dataset. (g) Annotation counts for each object category in the DroneVehicle dataset. (h) Annotation counts for each object category in the HIT-UAV dataset.

Over the past decade, deep learning technologies have significantly advanced object detection [7]. Convolutional neural networks (CNNs) [8] have demonstrated exceptional performance across various fields due to their robust

Received 3 February 2025; revised 19 March 2025; accepted 11 April 2025. Date of publication 21 April 2025; date of current version 8 May 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 62476024, in part by the 14th Five-Year Plan Funding of China under Grant 50916040401, and in part by the Fundamental Research Program under Grant 514010503-201. *(Min Li and Jinhui Lan are co-first authors.) (Corresponding author: Jinhui Lan.)*

The authors are with Beijing Engineering Research Center of Industrial Spectrum Imaging, School of Automation and Electrical Engineering, University of Science and Technology Beijing, Beijing 100083, China (e-mail: ustb923@163.com).

Digital Object Identifier 10.1109/TGRS.2025.3562966

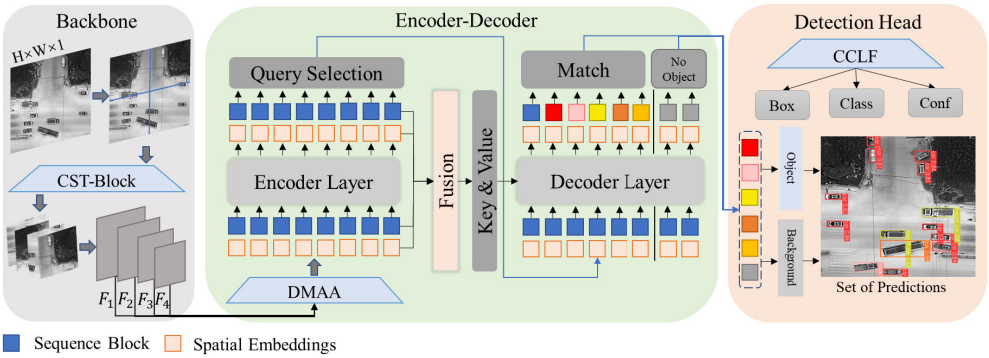

Fig. 2. Architecture of the LAST-Net, including backbone, encoder, decoder, and detection head. The input to the encoder–decoder consists of all the features extracted by the backbone network, where the blue module represents the sequence block and the yellow module represents the position embedding.

feature extraction capabilities [9]. However, convolutional networks still exhibit certain limitations. Due to the restricted receptive field of convolutions, CNNs are inadequate in extracting global features, which results in degraded performance when dealing with complex backgrounds. Additionally, while deeper convolutional layers can capture more abstract features, they may also lead to the loss of detailed features, making them less effective in addressing low-contrast issues.

In recent years, transformer models have significantly improved performance in object detection tasks due to their robust spatial modeling capabilities [10], [11], [12]. Unlike traditional convolution-based methods, transformers utilize an attention mechanism [13], [14], [15], [16] to process information from all positions, effectively capturing input sequence dependencies and acquiring global information from the first layer. This is particularly beneficial for remote sensing images with rich semantic content. This presents a new research direction for developing more efficient transformer-based detection algorithms. However, when the dataset size is small or when multiscale object detection is required, its performance declines significantly.

Therefore, we propose the transformer architecture into the UAV-based infrared remote sensing object detection task and combine it with the feature extraction strengths of CNNs to propose the local adaptivity spatial transformer network (LAST-Net). Fig. 2 illustrates the network's structure, which comprises three parts: the backbone, encoder–decoder, and detection head. Specifically, the backbone network combines the transformer framework and convolution operation, proposing the convolutional spatial transformer block (CST-Block). This module effectively suppresses background interference, enhances object features, and highlights important regions, achieving effective feature extraction in complex backgrounds (addressing the challenge 1). In the encoder–decoder structure, we propose the dual-modality adaptive attention (DMAA) mechanism, where the content matrix first performs coarse-grained querying, followed by fine-grained querying by the spatial matrix. By enabling effective feedback between these two components, the DMAA mechanism accurately locates weak objects in low contrast images (addressing the challenge 2). Finally, the detection head proposes the coordinate crossing loss function (CCLF), which assigns higher weights to samples with fewer labels to optimize training. Combined with the attention mechanism, it facilitates precise one-to-one matching between predicted boxes and objects, resolving the multiscale feature mapping issue (addressing the challenge 3). The main contributions of this article are as follows.

1) We propose the CST-block, which leverages convolution operations to capture local pixel information while utilizing the transformer's global modeling capability to extract semantic information. Through feature prealignment and parallel computation, the CST-block ensures consistency in feature representation and effectively fuses local and global information. This effective combination enhances the feature extraction capabilities.
2) We design the DMAA mechanism, where the query matrix assigns weights to object information, while the spatial matrix correlates and aggregates similar features across distant regions. By establishing deep interactions and feedback mechanisms between the two matrices, DMAA enables efficient feature fusion, enhancing the contrast between objects and the background.
3) We introduce the CCLF, which adjusts parameters to accurately match predicted bounding boxes to objects, facilitating effective identification of objects across various categories and scales while achieving multiscale feature mapping. Additionally, CCLF addresses dataset imbalance by assigning higher weights to small samples, enhancing the model's generalization ability in small sample conditions.

The remainder of this article is organized as follows. Section II reviews related research; Section III details the proposed model architecture and implementation; Section IV presents the experimental setup and results of ablation studies; Section V discusses our algorithm's differences and applicability compared to the existing algorithms; Section VI concludes this article and outlines future research directions.

II. Related Work

A. Infrared Object Detection

Infrared object detection algorithms can be classified into three categories: local prior-based, nonlocal prior-based, and learning-based methods [17], [18].

1) Local Prior-Based Infrared Object Detection: Traditional algorithms include max-mean and max-median filters [19] and morphological top-hat filters [20]. These methods focus on the relationship between objects and backgrounds by adjusting filter parameters to identify objects in infrared images. However, their effectiveness is limited by inherent noise. To enhance detection accuracy, Chen et al. [21] improved object signals and suppressed background clutter using local contrast measures. Yu et al. [22] developed local contrast learning to generate feature maps during training, further enhancing detection performance. Additionally, many methods have applied these approaches at multiple scales to tackle a broader range of scenarios [23], [24].

2) Nonlocal Prior-Based Infrared Object Detection: These algorithms approach detection by measuring nonlocal similarity in the background and analyzing object sparsity. Gao et al. [25] changed infrared small object detection as a low-rank and sparse matrix recovery problem, a typical nonlocal prior detection approach. However, this method faces challenges in environments with highly variable backgrounds. Yao et al. [26] proposed a strategy that combines nonlocal self-similarity for initial detection with local contrast for refinement, helping differentiate between noise and clutter. Zhang and Peng [27] introduced a tensor nuclear norm and a robust infrared patch model to effectively suppress edge artifacts, demonstrating strong robustness. Zhu et al. [28] utilized nonlocal low-rank methods to construct a low-rank model for saliency learning across diverse scenes.

3) Learning-Based Infrared Object Detection: Advancements in deep learning have significantly enhanced detection accuracy and robustness in complex scenarios through improved feature extraction, model training, and detection techniques. Cao et al. [29] introduced a dual-channel fusion module based on deep neural networks, excelling in autonomous driving systems, especially in low-light conditions. Devaguptapu et al. [30] proposed MMTOD, a pseudo-multimodal framework that improves object detection in thermal imagery. Hou et al. [31] combined handcrafted features with a CNN to address challenges posed by infrared small objects with low contrast and low signal-to-noise ratios (SNRs) in complex backgrounds. Li et al. [32] proposed Yolo-firi, which expands the shallow cross-stage partial (CSP) connection module by introducing an enhanced attention mechanism that focuses on objects while suppressing background noise.

In summary, local and nonlocal prior methods adapt well to varying environmental conditions but may face challenges with small temperature differences between objects and backgrounds or significant environmental changes [33], [34]. In contrast, learning-based methods, through training on large-scale datasets, can automatically learn richer and more abstract feature representations, leading to higher detection accuracy and improved robustness across diverse backgrounds and lighting conditions.

B. UAV-Based Remote Sensing Object Detection

In recent years, UAV-based remote sensing object detection technology has rapidly advanced through deep learning techniques. These methods can be categorized into two main types: two-stage detectors and single-stage detectors, based on whether they generate region proposals [35], [36]. A typical two-stage detector is faster R-CNN [37], widely used in remote sensing object detection [38], [39]. Ye et al. [40] designed a rotated region of interest (RRoI) learner to detect densely packed objects. Avola et al. [41] proposed a filtering and fusion network inspired by faster R-CNN, designed specifically for remote sensing images. The MS-faster R-CNN [42] combines with the deep tracking algorithm for effective detection and tracking of UAV images. Tang et al. [43] introduced keypoint R-CNN, which enhances RoI prediction with a nonconvolutional region proposal network branch, improving small object detection.

While two-stage detectors have significantly improved detection accuracy, their slower processing speed has led to a preference for lightweight networks in UAV applications. Consequently, many researchers are turning to single-stage algorithms. R3Det [44] implements a fast and accurate detection mechanism through an incremental regression single-stage detector. Studies [45], [46] have optimized YOLO-based models for multiscene object detection in infrared remote sensing. Sun et al. [47] integrated a recursive feature pyramid into a single-stage framework for rotated objects, balancing speed and accuracy. Li et al. [48] employed channel substitution and pixel-level weighted fusion for accuracy and speed. Tan et al. [49] enhanced feature extraction through dilated convolution, which improved detection accuracy. Wang et al. [50] introduced a global attention module (GAM) to effectively handle long-range dependencies and enhance the capability of capturing global information. Fu et al. [51] proposed a lighter model based on minimum point distance boundary regression.

In summary, current UAV-based remote sensing object detection methods primarily focus on designing and improving convolutional networks, enhancing detection capabilities for small objects and complex backgrounds by introducing new network architectures or optimizing existing model components.

C. Transformer and Attention Mechanism

The transformer model [52] has gained significant focus for its robust spatial modeling capabilities [54], [55], [56] and is widely used in object detection tasks [57], [58], [59], [60]. Its self-attention mechanism captures relationships between different regions in an image, enabling global information processing. Recently, researchers have applied transformers to remote sensing object detection. For instance, SCRDet [61] introduced a multilayer feature fusion mechanism to enhance anchors and improve small object detection accuracy. Ye et al. [62] developed a self-attention mechanism based on efficient convolution transformer blocks, improving the detection of occluded objects through the extraction of contextual information. Lu et al. [63] utilized a cross-shaped window transformer as the backbone, combined with a feature pyramid structure, significantly improving small object detection. Zhang et al. [64] proposed an adaptive multigranularity routing mechanism to promote token sparsity within

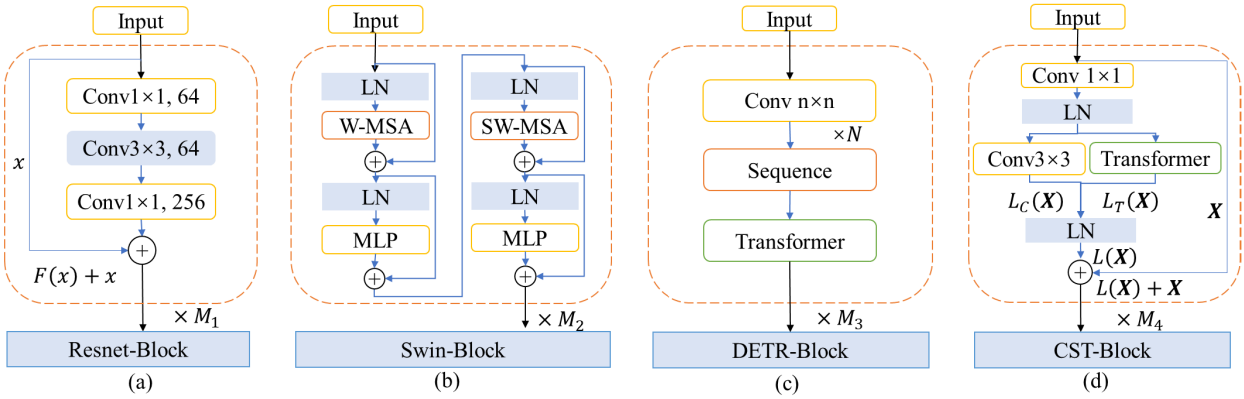

Fig. 3. Four backbone network layers. (a) Convolutional network block (Resnet-Block). (b) Transformer block (Swin-Block). (c) Hybrid convolutional and transformer block (DETR-Block). (d) CST-Block. The input is the raw infrared image. LN indicates layer normalization and MLP indicates multilayer perceptron. M_1, M_2, M_3, and M_4 indicate the number of layers in the backbone networks of each block.

the transformer, reducing computational costs without sacrificing accuracy, and indicating vision transformers as viable alternatives. Ni et al. [65] proposed a novel baseline for detecting and counting in remote sensing images to generate caption information. Yuan and Wei [66] designed an intercross attention (ICA) to learn cross-attention between RGB and infrared modalities, enhancing small object detection accuracy.

Some researchers have improved the attention mechanism to address various downstream tasks [67]. Fu et al. [60] proposed the dual attention network (DANet), which uses attention across both spatial and channel dimensions to model the semantic dependencies and address the scene segmentation task. Woo et al. [68] proposed CBAM, which uses channel and spatial attention to sequentially infer attention maps, enhancing the capability of object classification tasks. Li et al. [69] added attention mechanisms to CNNs to increase their depth and width, improving detection robustness. Ma et al. [70] incorporated attention and a dual-adaptive NMS network into the feature pyramid, enhancing the accuracy of object detection.

In summary, the transformer architecture and attention mechanism have been widely applied to various downstream tasks, showcasing significant advantages in remote sensing image processing. However, research specifically focused on object detection in UAV-based infrared remote sensing images remains quite limited.

III. Proposed Method

This section provides a detailed overview of the LAST-Net, starting with an introduction to the overall network architecture, followed by an in-depth discussion of the CST-Block, DMAA, and CCLF modules.

LAST-Net is an end-to-end architecture that converts input images into detection results (as illustrated in Fig. 2). The backbone network (gray area) segments the grayscale image into nonoverlapping feature patches, incorporating spatial embeddings to extract local pixel features and global positional information, thereby ensuring effective information preservation at each layer. Subsequently, the features extracted by the backbone network are fed in parallel into the DMAA module (green area), which employs a learnable strategy for the content and position of its feature layers through an encoding–decoding mechanism. The DMAA combines features from multiple encoder subsets, producing keys and values initialized as queries for the decoder. Attention is applied to update features layer by layer, integrating contextual information to identify objects. Finally, the detection head (yellow area) combines with the CCLF for multiscale feature mapping. By dynamically adjusting anchor box sizes and positions to align with object locations, the model generates bounding boxes, classes, and confidence scores for each object, enabling accurate class recognition and precise localization in multiscale object detection.

A. Convolutional Spatial Transformer Block

Motivation: Currently, the main design approaches for backbone networks are based on three primary methods. One approach is based on residual convolutional networks [as shown in Fig. 3(a)], which are widely used due to their feature extraction capabilities. However, convolutional networks are limited by the need to stack a large number of convolutional layers; for instance, ResNet50 [9] comprises 50 convolutional operations. The other approach is the transformer architecture [as shown in Fig. 3(b)], which has gained significant enhancement in recent years due to its excellent spatial information extraction capabilities. However, this architecture is less effective for multiscale object detection; for example, the swin transformer [55] demonstrates low accuracy when detecting small objects. In addition, some researchers flatten the feature maps of CNNs into sequences for processing by transformers, which are then reshaped back into feature maps after the transformer output, such as in the DETR network [as shown in Fig. 3(c)]. However, during the process of flattening and reshaping features, the misalignment between CNN and transformer features can lead to information loss. Moreover, this method of feature transformation in high-resolution remote sensing images introduces additional computational load, reducing the model's efficiency.

After analyzing the advantages and disadvantages of CNNs and transformers, we found that CNNs are proficient in leveraging local receptive fields to extract features from local

pixel information, while transformers are effective in global modeling, enabling the extraction of contextual information and the detection of semantic features using global information. Therefore, this article combines the strengths of both to design the CST-Block, as shown in Fig. 3(d). First, in the CST-Block, a 1×1 convolution is used to project the CNN's output features to a dimension and scale consistent with the transformer features. Then, layer normalization (LN) is applied to normalize both types of features, ensuring consistency in their distributions. This approach preemptively addresses the inconsistency in the representation of the two feature types during fusion. The CNN processes the input image through convolution and pooling operations to extract local spatial features (such as edges and shapes). Meanwhile, the transformer captures long-range contextual information, providing global semantic representations. Subsequently, the outputs of the convolutional and transformer layers are fused through sufficient combination and parallel computation, fully integrating local and global feature information to generate the final feature map. Additionally, inspired by residual learning, shallow features are preserved and transmitted to deeper layers through skip connections, effectively reducing the degradation problem in deep networks. This design significantly enhances the effectiveness of information extraction, minimizes detail loss, and reduces computational overhead, thereby enabling more comprehensive feature extraction from complex remote sensing images while preserving all valuable information.

Specifically, let the input features be denoted as $\boldsymbol{X}$. The CST-Block extracts convolutional features, denoted as $L_C(\boldsymbol{X})$, and extracts transformer features, denoted as $L_T(\boldsymbol{X})$

$$L_C(\boldsymbol{X}) = \text{Pool}(\text{Conv}(\boldsymbol{X})) \tag{1}$$

$$L_T(\boldsymbol{X}) = \text{softmax}(\text{Encoder}(\boldsymbol{X}), \text{Decoder}(\boldsymbol{X})). \tag{2}$$

These features are then interacted to obtain the intermediate features, denoted as $L(\boldsymbol{X})$

$$L(\boldsymbol{X}) = \text{LayerNorm}(\text{Inter}(L_C(\boldsymbol{X}), L_T(\boldsymbol{X}))). \tag{3}$$

Next, $L(\boldsymbol{X})$ interacts with the original input features $\boldsymbol{X}$, while residual connections are maintained to ensure that the model's learning outcomes are at least as effective as those derived from the original features, resulting in the final output of the CST-Block

$$\text{CST}(\boldsymbol{X}) = L(\boldsymbol{X}) + \boldsymbol{X}. \tag{4}$$

The CST-Block is applied to the backbone framework, as shown in Fig. 4. The input $x \in \mathbb{R}^{H \times W \times 1}$, $H \times W$ denotes the pixel resolution and the infrared image is single channel. First, the input image is divided into nonoverlapping patches using patch partition. Next, linear embedding is utilized to linearly transform the image features into a sequence of vectors, followed by projection and activation for a single-layer feature map. LN is applied between the CST-Block and multilayer perceptron (MLP) modules at each stage to mitigate the gradient vanishing issue. The feature map dimensions are progressively reduced through patch merging. The feature map output at each stage has dimensions $(H/m) \times (W/m) \times nC$. In the first stage, F_1, the feature map size is reduced

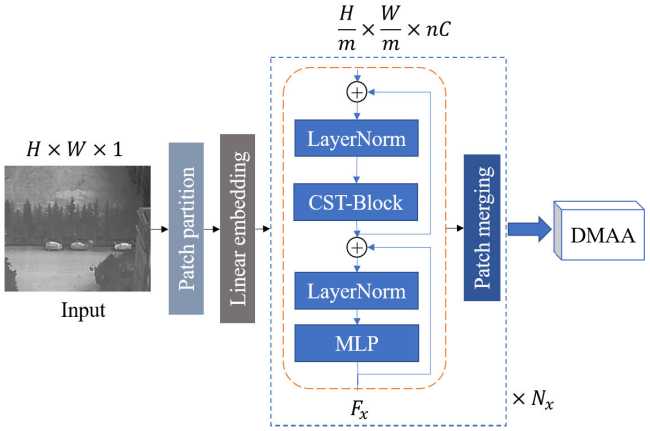

Fig. 4. Overall framework of the backbone.

to $(H/4) \times (W/4)$, and the number of channels increases to C, with the number of modules being N_1. Similarly, in the second stage, F_2, $m = 8$ and $n = 2$, and the number of modules is N_2. In the third stage, F_3, $m = 16$ and $n = 4$, and the number of modules is N_3. In the fourth stage, F_4, $m = 32$ and $n = 8$, and the number of modules is N_4. The total number of layers in the backbone network is $N = N_1 + N_2 + N_3 + N_4$. The impact of different network depths on detection results is validated in the ablation experiments.

B. Dual-Modality Adaptive Attention

Motivation: 1) In infrared remote sensing scenarios, infrared images typically exhibit low contrast, and the objects often occupy only a small portion of the pixels. However, current attention mechanisms lack a differentiated query strategy, making it challenging to effectively distinguish foreground and background information in low-contrast infrared images; 2) previous studies have achieved some performance improvements in semantic segmentation and object classification tasks by introducing dual attention mechanisms [60], [68] and adding spatial attention [69], [70] in RGB object detection networks. However, directly applying these attention mechanisms to object detection tasks involves processing channel and spatial information independently through serial or parallel operations. This approach not only lacks interaction between the two types of information but also introduces additional computational overhead; and 3) most attention mechanisms are integrated directly into CNNs or transformers, without fully exploring the potential advantages of hybrid CNN-transformer frameworks.

We propose a DMAA for feature fusion in object detection to address the aforementioned three issues. As shown in Fig. 5, the DMAA comprises two core components: content attention and spatial attention: 1) content attention applies coarse-grained queries to each feature layer, performing pixel-level weighting on local features to extract and assign weights to object information; 2) spatial attention combines the boundary information of each feature layer to associate and aggregate similar features that are spatially distant. By capturing positional relationships through nonlocal operations, it enhances the scene understanding capability; and 3) deep

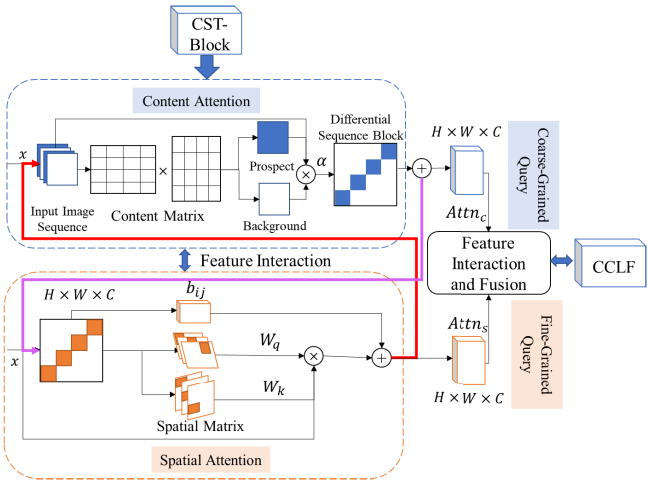

Fig. 5. Framework of dual-modality adaptive attention.

interaction and feedback between the two attention modules, highlighted by the red and purple lines in Fig. 5, represent our key innovation and improvement over traditional serial and parallel structures. The output of content attention serves as the input for spatial attention, and the result of spatial attention further provides feedback to the content attention module, forming a more robust interaction pattern. Each iteration updates the feature representation based on previous attention results, achieving progressive refinement.

Specifically, the content attention processes the feature blocks extracted by the backbone network, independently calculating attention for each block. If a block contains foreground information, it is assigned a higher weight, while a lower weight is given to blocks that do not contain foreground information, with adaptive updates made layer by layer. We define C_{ij} represent the influence of the i channel of the content matrix on the j channel, expressed as follows:

$$C_{ij} = \frac{\exp(x_i \cdot x_j)}{\sum_{i=1}^{N}\sum_{j=1}^{M}\exp(x_i \cdot x_j)}. \tag{5}$$

The output of the content attention is denoted as Attn_c, representing the sum of all channel matrices, expressed as follows:

$$\text{Attn}_c = \alpha \sum_{i=1}^{N}\sum_{j=1}^{M}(C_{ij}x_i) + x_j \tag{6}$$

where α represents the weighting operator for the content attention features.

Subsequently, the spatial attention learns the differentiated sequences obtained from content attention based on the weight values. Spatial attention captures the broad dependencies between blocks according to their weights, consolidating the same object within these blocks. The output of the spatial attention is denoted as Attn_s, and the expression is

$$\text{Attn}_s = x_i W_q W_k^T x_j^T + b_{ij} \tag{7}$$

where $W_q = \text{MLP}(\text{SE}(x_{i,j}))$ represents the spatial encoding that queries data shared across all layers using an MLP. $W_k = \text{Con}(C_{ij}, \text{SE}(x_{i,j}))$ represents the concatenation (Con) of all original features with the positional encoded feature information. Additionally, b_{ij} represents the positional bias.

Finally, the two attention modules facilitate deep interaction and feedback between content and spatial attention. Each iteration updates the current feature representation based on the previous attention result, gradually refining the features. The fused attention is denoted as Attn and can be expressed as

$$\text{Attn} = F(\text{Attn}_c, \text{Attn}_s) \tag{8}$$

where the function $F(\cdot)$ denotes the interaction and fusion of feature layer information.

C. Coordinate Crossing Loss Function

The UAV infrared dataset faces significant variations in object scales and challenges related to sample imbalance. To address this, we propose the CCLF in the detection head. The CCLF integrates bounding box loss, classification loss, and confidence loss during the training process, assigning greater emphasis to minority class samples to balance contributions across different categories. Furthermore, it enables one-to-one matching between objects and categories, facilitating accurate multiscale object detection.

The bounding box loss directly measures the degree of overlap between the predicted and ground truth boxes by evaluating their relative positional offsets in the detection head. In multiscale scenarios, anchor boxes are dynamically adjusted to ensure that the regression loss is more precise across different scales. Let the input feature representation of a feature map be $(x, y, h, w) \in R$. In the CCLF, (x, y) determines the center of the predicted bounding box, while (h, w) dynamically adjusts the attention object's position map to align with the ground truth and bounding box. The relative position predicted by the detection head is denoted as $(\Delta x, \Delta y, \Delta h, \Delta w) \in R$. We adjust the size and position of the anchor boxes to modulate the position attention, thereby providing preset positional priors for subsequent pooling operations. The PE represents the positional embedding, expressed as follows:

$$\text{PE}(\text{Attn}_{q,k,v}) = \begin{pmatrix} x_{q,k,v} + \Delta x, y_{q,k,v} + \Delta y, \\ w_{q,k,v} + \Delta h, h_{q,k,v} + \Delta w \end{pmatrix}. \tag{9}$$

By performing elementwise multiplication between the attention and the positional embedding, we achieve a decoupled representation. This results in a vector representation, which can be expressed as follows:

$$\text{CLN}(A_{q,k,v}) = \text{Dec}(\text{PE}(A_{q,k,v}) \cdot A_{q,k,v}). \tag{10}$$

When applied to multiple scales, this becomes

$$\text{M-CLN}(A_{q,k,v}) = \text{Dec}\left(\sum_{i=1}^{M}(\text{PE}(A_{q,k,v}) \cdot A_{q,k,v})\right). \tag{11}$$

Additionally, to address the imbalance in the dataset due to varying object category frequencies, we assign higher weights to underrepresented categories during training. The classification loss is defined as $C(\hat{u})$, expressed as follows:

$$C(\hat{u}) = -\delta_t(1-\vartheta_t)^{\kappa}\log(\vartheta_t) \tag{12}$$

TABLE I
DRONEVEHICLE AND HIT-UAV DATASET PARAMETERS

Indicators	DroneVehicle dataset	HIT-UAV dataset
Drone platform	DJI M200	DJI Matrice M210 V2
Thermal infrared camera	Zenmuse XT 2	DJI Zenmuse XT2
Resolution	640x512	640x512
Lens focal length	8mm	25mm
Flying height	80 m, 100 m, and 120 m	60-130m
Shooting angles	15°, 30°, and 45°	30°-90°
FOV	57.12°×42.44°	25°×20°
Wavelength	7.5 ~ 13.5 μm	7.5 - 13.5 μm
Scene	Urban roads, residential areas, parking lots, highways, etc.	Schools, parking lots, roads, playgrounds, etc.
Categories	Car, truck, bus, van, and freight	Person, car, bicycle, othervehicle, and dontcare
Year	2020	2023

where δ_t is a hyperparameter used to balance positive and negative samples during training. ϑ_t is a hyperparameter designed to balance easy and hard samples. κ is a weighting factor that addresses the imbalance in the quantity of easy and hard samples. In remote sensing images, small objects constitute a large proportion. The weighting factor adjustment strategy is employed to enhance small object detection and further optimize performance. Higher weights are assigned to small objects to ensure they are not dominated by larger objects during training.

The CCLF performs bounding box regression, class prediction, and confidence score prediction. As training progresses, the difficulty of objects at different scales may vary. To address this, an adaptive loss weighting adjustment strategy can be employed, allowing the network to automatically adjust the contribution of each layer's loss. This facilitates better model convergence while ensuring effective detection across multiple scales. Finally, the confidence loss quantifies the probability that a predicted box contains the object by distinguishing between foreground and background and provides a confidence threshold. The loss function is expressed as

$$\mathcal{F} = \mathcal{L}_{\text{box}}\left(\text{M-CLN}\left(A_{q,k,v}\right), b\right) + \mathcal{L}_{\text{class}}\left(C\left(\hat{u}\right), c\right) + \mathcal{L}_{\text{conf}}(f) \tag{13}$$

where b represents the bounding box information, c represents the class information, and f indicates the confidence information.

IV. EXPERIMENTS AND DISCUSSION

In this section, we first introduce the datasets, evaluation metrics, and experimental platform used. We then present both quantitative and qualitative comparisons between our proposed algorithm and SOTA methods across two datasets. Finally, we conduct ablation studies to quantitatively evaluate the effectiveness of the three proposed modules.

A. Datasets

1) DroneVehicle Dataset [71]: This dataset is a large UAV vehicle detection dataset specifically designed for vehicle detection. It includes a diverse range of lighting, occlusion, and scale variations. We utilized the original dataset's division of 17 990 images as the training set and 1469 images as the validation set. The original image size was 840 × 712 pixels, which was cropped to 640 × 512 pixels after removing the white borders.

2) HIT-UAV Dataset [72]: This dataset is designed for object detection in UAV infrared imagery. It contains images captured under various flight altitudes, angles, dates, and lighting conditions. The dataset includes 2898 thermal infrared images, with 2335 images used for training and 563 images for validation.

Both datasets offer annotated infrared images to validate the network's performance. They encompass different urban scenes and object types, with notable differences in the number of object annotations, facilitating a comprehensive evaluation of the algorithms. Detailed parameters are presented in Table I.

TABLE II
EXPERIMENTAL PLATFORM AND PARAMETERS

Hardware	CPU	Intel Xeon Silver 4114
	GPU	NVIDIA GTX3060
Software	System	Ubuntu20.04
	Deep learning framework	Pytorch 2.1, python 3.10, pycharm 2023
Training parameters	Batch size	16
	IoU threshold	0.5
	Learning rate	0.0001
	Weight decay	0.05

B. Implementation Details

The information regarding the hardware, software, and training parameters used in the experiments is presented in Table II. Additionally, the input image size is set to 640 × 512 pixels based on the dataset standards, and the output size is consistent with the input size.

C. Comparison to SOTA

In this section, we evaluate our LAST-Net on two UAV-based infrared remote sensing datasets—the DroneVehicle and

TABLE III
COMPARATIVE EXPERIMENTAL RESULTS OF OUR ALGORITHM WITH VARIOUS SOTA ALGORITHMS ON THE DRONEVEHICLE DATASET, FOCUSING PRIMARILY ON THE PRECISION AT AN IOU THRESHOLD OF 0.5. BOLD INDICATES THE BEST RESULT, AND UNDERLINE INDICATES THE SECOND-BEST RESULT

Model	Backbone	AP%					mAP %	Params (M)	FLOPs (G)	Test time (FPS)
		Car	Truck	Bus	Van	Freight				
Faster-RCNN [37]	ResNet50	89.9	48.9	86.9	45.8	48.3	64.0	41.1	107.2	12
DETR [53]	ResNet50	91.4	61.2	89.9	57.2	58.0	71.5	41.0	86.0	28
Deformable DETR [54]	ResNet50	92.5	65.3	90.1	58.2	60.7	73.4	40.0	173.0	19
RT-DETR [57]	ResNet50	96.0	79.7	95.3	66.6	74.9	82.5	42.7	136.0	108
LF-MDet [6]	ResNet50+LET	82.2	73.6	86.6	57.0	59.6	71.8	38.7	77.7	67
C^2Former [66]	ResNet50+FPN	90.2	68.3	89.8	58.5	64.4	74.2	100.8	89.9	21
Yolov7s [73]	CSPDarkNet53+ E-ELAN+MPConv	96.7	73.4	95.0	64.4	70.0	79.9	36.9	104.7	51
Yolov8s [74]	CSPDarkNet53+C2f	98.3	78.8	96.9	66.2	73.8	82.8	**11.1**	**28.7**	**125**
I2MDet [75]	CSPNet+Kaleidoscope	96.3	73.4	93.2	58.6	65.0	77.3	31.9	48.9	-
Dual-YOLO [76]	CSPDarknet	98.1	65.7	95.8	46.6	52.9	71.8	175.1	-	62
DTNet-B [60]	CSPBlock	90.2	78.1	89.2	65.7	67.9	78.2	27.7	57.1	-
ViT-B+RVSA [77]	ViT-B+RVSA	89.7	52.3	88.0	44.4	51.0	65.1	113.1	252	9
Swin Transformer [55]	Swin-T	93.4	67.5	91.6	61.2	71.3	77.0	48.0	267.0	15
LAST-Net (ours)	CST	**98.9**	**85.6**	**97.8**	**70.0**	**79.0**	**86.3**	39.2	88.6	101

the HIT-UAV dataset—and compare its performance against SOTA methods. We focus on metrics, such as average precision (AP), mean AP (mAP), model parameter (Params), floating-point operations per second (FLOPs), and frames per second (FPS), to quantitatively validate the algorithm's effectiveness. Additionally, we present the visual results for a qualitative analysis.

1) Comparison to SOTA in the DroneVehicle Dataset: Table III presents a comparison of vehicle detection performance between LAST-Net and other SOTA algorithms on the DroneVehicle dataset. The results indicate that LAST-Net achieves the best performance in terms of AP for individual classes (including car, truck, bus, van, and freight) as well as mAP across all classes. Specifically, LAST-Net outperforms the second-ranked YOLOv8 by improving AP by 0.6% and 0.9% points for the car and bus classes, respectively, and increasing mAP by 3.5% points. In addition, LAST-Net shows significant improvements over RT-DETR in AP for the truck, van, and freight classes, with increases of 5.9%, 3.4%, and 4.1% points, respectively. Notably, compared to the original transformer architectures, LAST-Net not only reduces model parameters and FLOPs, significantly improving detection speed, but also demonstrates strong competitiveness when compared to CNN architectures.

Based on the results in Table III, we visualized the output of different backbone networks and detection performance results, including RT-DETR, C^2Former, YOLOv8, Swin Transformer, and our algorithm LAST-Net. The visualization results are shown in Fig. 6, where red circles indicate missed detections and blue circles denote false detections. All algorithms generally perform well in detecting objects across most scenes. However, in scenarios with extremely low contrast or complex backgrounds, other algorithms often result in missed or false detections. In contrast, our network significantly reduces both missed detections and false positives, effectively identifying objects and providing qualitative evidence of the approach's effectiveness.

2) Comparison to SOTA in the HIT-UAV Dataset: In Table IV, we present the performance comparison results of LAST-Net with several SOTA algorithms on the HIT-UAV dataset. The results show that our network achieves the best AP for the person, car, othervehicle, and dontcare classes, as well as the overall mAP. Additionally, LAST-Net ranks second in AP for the bicycle class. Specifically, compared to the second-ranked YOLOv8, LAST-Net improves AP by 0.8%, 2%, and 1.7% points for the car, othervehicle, and dontcare classes, respectively. When compared to the second-ranked RT-DETR, LAST-Net shows a 0.2% point increase in AP for the person class and a 2.1% point improvement in mAP. Moreover, as shown in Table IV, although the parameters and FLOPs are not the most optimal, they are still significantly reduced compared to most algorithms. Additionally, the test time (FPS) as the second best among the evaluated methods.

Based on the results in Table IV, we visualized the output of different backbone network algorithms, including faster-RCNN, RT-DETR, YOLOv8, Swin Transformer, and our LAST-Net algorithm. The visualization results are shown in Fig. 7, where red circles indicate missed detections and blue circles denote false detections. The figure reveals that in the top–down view, other algorithms often miss detections for smaller objects like persons and bicycles due to limited feature representation. Additionally, in complex scenes, there is a tendency to misidentify signs and wheels as persons and bicycles. In contrast, LAST-Net not only specializes in detecting vehicles with distinct features but also performs well in identifying less pronounced objects like persons and bicycles.

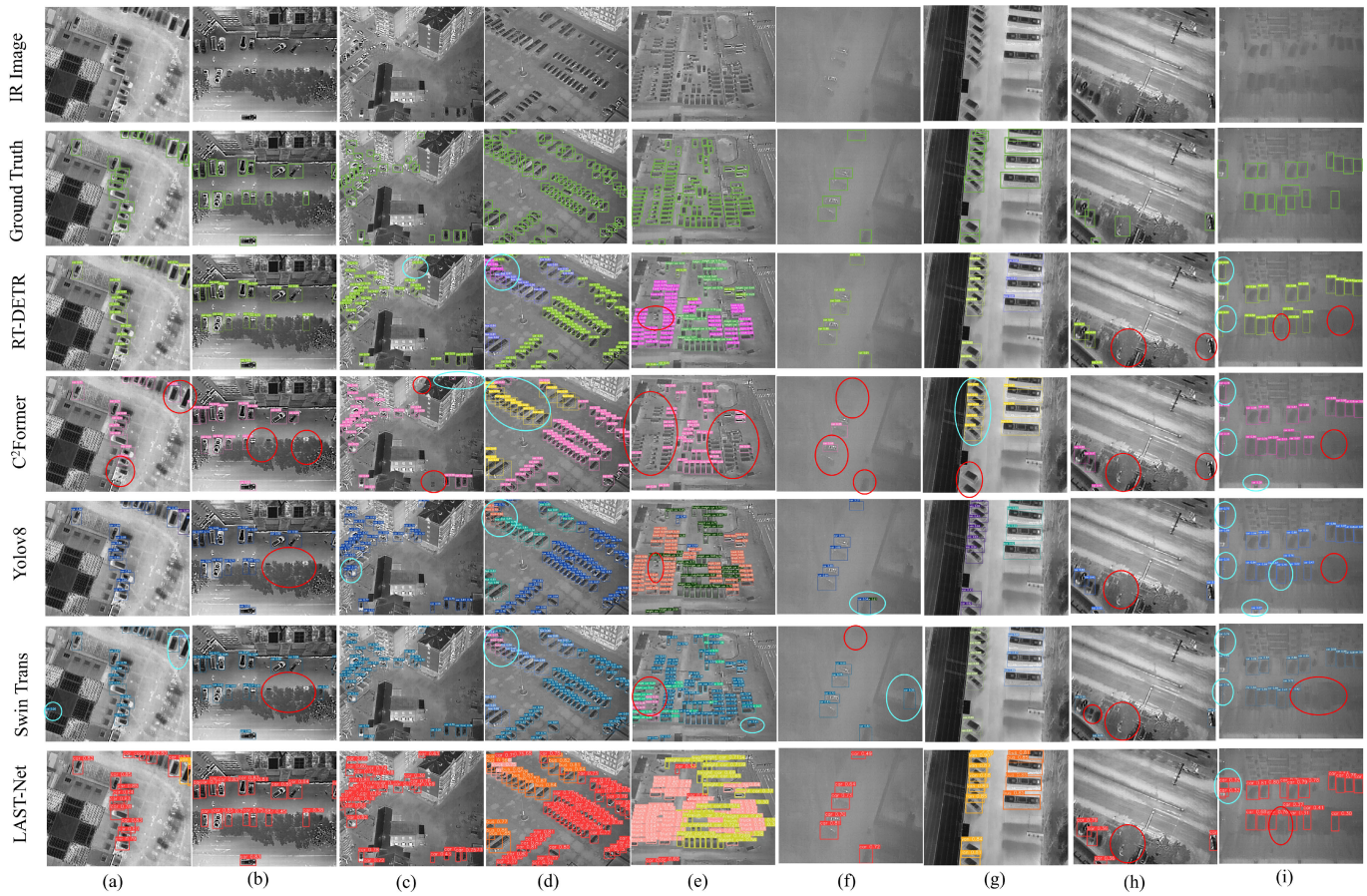

Fig. 6. Visualization examples of various methods in different scenarios on the DroneVehicle dataset. (a) Low contrast between object and background. (b) Occlusion scenario. (c) Complex urban environment. (d) Multiple objects. (e) Oblique viewing angle. (f) Low-resolution imagery. (g) Multiscale objects. (h) Extreme environments. (i) Complex architectural scenes.

TABLE IV

COMPARATIVE EXPERIMENTAL RESULTS OF OUR ALGORITHM WITH VARIOUS SOTA ALGORITHMS ON THE HIT-UAV DATASET, FOCUSING PRIMARILY ON THE PRECISION AT AN IOU THRESHOLD OF 0.5. BOLD INDICATES THE BEST RESULT, AND UNDERLINE INDICATES THE SECOND-BEST RESULT

Model	Backbone	AP%					mAP %	Params (M)	FLOPs (G)	Test time (FPS)
		Person	Car	Bicycle	Other Vehicle	Dont Care				
Faster-RCNN [37]	ResNet50+FPN	84.4	95.1	74.6	65.7	50.5	74.1	41.1	107.2	12
DETR [53]	ResNet50	88.1	96.1	83.8	63.1	69.6	80.1	41.0	86.0	28
Deformable DETR [54]	ResNet50	92.1	96.2	84.0	65.2	70.1	81.5	40.0	173.0	19
RT-DETR [57]	ResNet50	94.5	97.0	**92.4**	83.1	77.2	88.8	42.7	136.0	108
Yolov7s [73]	CSPDarkNet53+ E-ELAN+MPConv	92.1	96.0	82.1	65.6	76.2	82.4	36.9	104.7	51
Yolov8s [74]	CSPDarkNet53+C2f	92.4	97.6	87.6	84.3	81.8	88.7	11.1	28.7	**125**
PHSI-RTDETR [78]	RPConv-Block	94.4	97.2	90.8	65.5	65.1	82.6	13.9	47.5	127
MFFNet [79]	ShuffleNetv2	84.0	94.6	85.7	66.7	76.7	81.5	**4.2**	**16.0**	-
Swin Transformer [55]	Swin-T	91.1	96.8	89.9	70.2	71.0	83.8	48.0	267.0	15
LAST-Net (ours)	CST	**94.7**	**98.4**	91.6	**86.3**	**83.5**	**90.9**	39.2	88.6	101

D. Ablation Study

In this section, we validate the effectiveness of the proposed modules through ablation experiments. We first conduct a combination test of the three main components: our proposed backbone network CSTNet, the attention mechanism DMAA, and the loss function CCLF for multiscale object detection, and then, we evaluate the impact of their respective parameters.

1) Component Combination Experiment: We selected the well-known and high-precision RT-DETR model as the baseline model. In the ablation study, we individually tested the

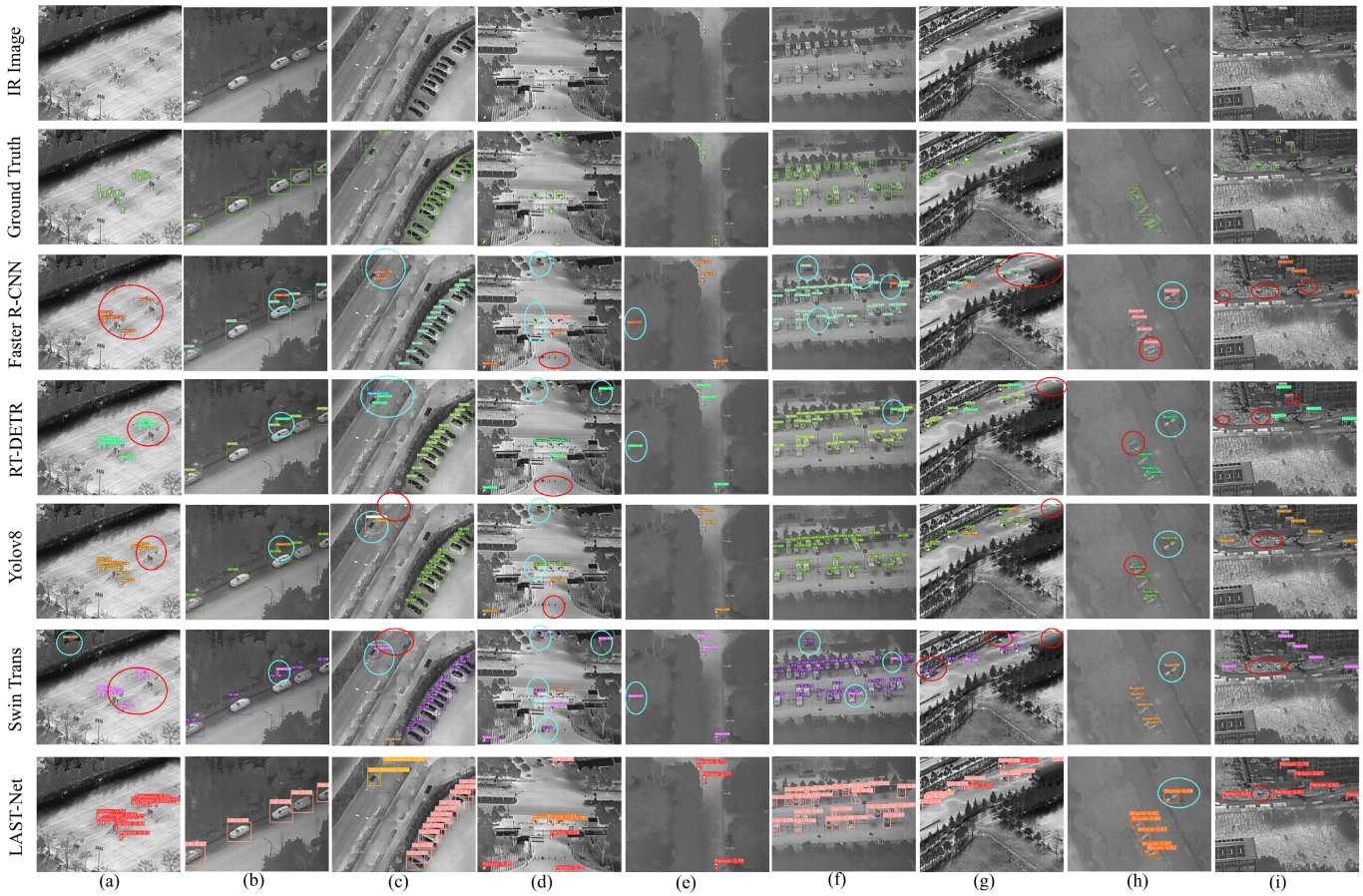

Fig. 7. Visualization examples of various methods in different scenarios on the HIT-UAV dataset. (a) Small object scale. (b) Large object scale. (c) Multiobject scenario. (d) Complex background. (e) Low image contrast. (f) Top–down view. (g) Oblique view. (h) Low contrast images. (i) Complex scene.

TABLE V
COMPONENT COMBINATION EXPERIMENT ON THE DRONEVEHICLE AND HIT-UAV DATASET. BOLD INDICATES THE BEST RESULT

Model	CST	DMAA	CCLF	mAP%		Params(M)	FLOPs(G)
				Drone Vehicle dataset	HIT-Drone dataset		
Baseline	-	-	-	82.5	88.8	42.7	136.0
LAST-Net	√	×	×	84.0(+1.5)	89.7(+0.9)	**38.8**	**78.1**
	×	√	×	83.8(+1.3)	89.6(+0.8)	40.1	91.3
	×	×	√	84.9(+2.4)	89.9(+1.1)	43.2	136.6
	√	√	×	85.7(+3.2)	90.1(+1.3)	39.0	86.7
	√	×	√	86.0(+3.5)	90.5(+1.7)	39.6	87.7
	×	√	√	85.8(+3.3)	90.2(+1.4)	41.3	89.6
	√	√	√	**86.3**(+3.8)	**90.9**(+2.1)	39.2	88.6

contributions of the CSTNet, DMAA, and CCLF components, as well as their combined effects. The experimental results are shown in Table V. Our findings indicate that each module enhances performance independently. Combined use has greater improvements. Notably, integrating all three components produces the most significant overall effect. Specifically, the combination of the three modules, compared to the baseline, resulted in a 3.8% increase in mAP on the DroneVehicle dataset, a 2.1% increase on the HIT-UAV dataset, and a reduction of 3.5 M parameters. Furthermore, compared to the baseline, our network not only achieves higher precision but also demonstrates impressive performance in terms of parameter and FLOPs.

2) Structure of CSTNet: Building on prior research, we designed four CSTNet architectures: CSTNet-B, CSTNet-A, CSTNet-L, and CSTNet-H for backbone networks, as illustrated in Table VI. Among them, CSTNet-B consists of 12 network layers with a minimum parameter count of 32.2 M, while CSTNet-H has 37 layers and achieves a maximum mAP of 87.3%. Experimental results show that increasing network depth not only significantly enhances model precision but also increases the parameter count. To balance speed and precision, we selected CSTNet-A for all subsequent experiments.

3) Analysis of CCLF: We examined the impact of the adjustable factor κ in the loss function on detection precision. Specifically, a low κ may fail to adequately emphasize

TABLE VI

IMPACT OF THE NUMBER OF FEATURE LAYERS IN BACKBONE NETWORKS ON THE DRONEVEHICLE DATASET. BOLD INDICATES THE BEST RESULTS

Mode	$[N_1, N_2, N_3, N_4]$	mAP%	Params(M)
Base	[2, 2, 6, 2]	84.2	**32.2**
Advance	[2, 2, 15, 2]	86.3	39.2
Large	[2, 2, 18, 2]	86.7	42.0
High	[5, 5, 22, 5]	**87.3**	69.1

TABLE VII

EFFECT OF THE ADJUSTABLE FACTOR κ IN THE LOSS FUNCTION ON THE PRECISION AT AN IOU THRESHOLD OF 0.5 IN THE DRONEVEHICLE DATASET. THE BEST RESULTS ARE INDICATED IN BOLD

κ	AP%					mAP%
	Car	Truck	Bus	Van	Freight	
0	97.8	81.6	95.0	67.4	75.0	83.4
0.5	98.3	82.9	94.9	68.1	75.2	83.9
1	98.7	84.6	96.1	68.9	77.8	85.2
2	**98.9**	**85.6**	**97.8**	**70.0**	**79.0**	**86.3**
5	98.0	82.4	95.7	67.2	73.4	82.3

TABLE VIII

EFFECT OF THE HYPERPARAMETER δ_t IN THE LOSS FUNCTION ON THE PRECISION AT AN IOU THRESHOLD OF 0.5 IN THE DRONEVEHICLE DATASET. THE BEST RESULTS ARE INDICATED IN BOLD

δ_t	AP%					mAP%
	Car	Truck	Bus	Van	Freight	
0.10	97.4	84.0	95.9	67.5	77.3	84.4
0.25	98.0	84.7	96.7	68.9	78.0	85.3
0.50	98.2	85.1	97.3	69.5	78.6	85.7
0.75	**98.9**	**85.6**	**97.8**	**70.0**	**79.0**	**86.3**
0.90	98.3	84.9	96.6	69.1	78.5	85.5
0.99	98.1	84.1	96.1	68.0	77.9	84.8

difficult-to-classify samples, leading to decreased overall precision, while a high κ can result in overfitting or training instability. Experimental results, as shown in Table VII, indicate that different κ settings significantly influence model performance. As the adjustable factor κ increases, the mAP gradually increases. However, after reaching its peak at $\kappa = 2$, the mAP begins to decrease. Optimal precision was achieved at $\kappa = 2$ particularly in categories with fewer labels (such as truck, van, and freight), indicating that κ effectively addresses the issue of sample imbalance. Thus, we adopted $\kappa = 2$ for all experiments and subsequent studies.

We also evaluated the impact of the hyperparameter $\delta_t (\delta_t \epsilon (0, 1))$ on the model's precision in balancing positive and negative samples. As shown in Table VIII, the model achieves the highest precision when $\delta_t = 0.75$, with the highest precision for the car category. This shows that categories with more labels, having more positive samples, are easier to classify correctly, while categories with fewer labels, having insufficient positive samples, experience significantly higher classification and detection difficulty. Moreover, adjusting the value of δ_t reveals that this parameter has a greater impact

TABLE IX

EFFECT OF THE HYPERPARAMETER ϑ_t IN THE LOSS FUNCTION ON THE PRECISION IN THE DRONEVEHICLE DATASET. THE BEST RESULTS ARE INDICATED IN BOLD

κ	Iou	ϑ_t	AP_S%	AP_M%	AP_L%
2	0.5	0.25	77.3	86.1	96.9
		0.5	**81.2**	**87.7**	**98.3**
		0.75	79.2	87.0	96.7
		0.9	76.9	85.8	95.5
	0.75	0.25	51.3	67.9	81.3
		0.5	**53.4**	**69.1**	**82.7**
		0.75	51.9	68.4	81.8
		0.9	50.7	67.8	81.0

on categories with fewer labels (such as truck, bus, van, and freight) than on categories with more labels (such as car), with the most notable improvement in the van category. These results show that our loss function effectively addresses the imbalance between positive and negative samples.

It can be observed from Tables VII and VIII that the optimal parameter values are $\kappa = 2$ and $\delta_t = 0.75$, which are used as benchmarks. The experiment uses IoU thresholds of 0.5 and 0.7 to analyze the impact of varying ϑ_t on the objects of different sizes (large, medium, and small). According to the experimental results in Table IX, when $\vartheta_t = 0.5$, the model achieves the highest precision. The detection precision is 81.2% for small objects, 87.7% for medium-sized objects, and 98.3% for large objects. Moreover, the results show that the model significantly improves detection precision for small objects, indicating that adjusting ϑ_t helps address the classification challenge of hard-to-separate samples, ensuring more accurate classification and detection of these samples.

4) Comprehensive Performance Comparison: We employed the precision–recall (P–R) curve and $F1$ curve to evaluate our model's overall performance. The P–R curve measures precision and recall at various thresholds, with a curve closer to the upper right corner indicating better performance. The $F1$ curve shows $F1$ score variations at different thresholds, with a score closer to 1 reflecting superior performance. As shown in Fig. 8, our network is closer to the top-right corner on the P–R curve and closer to the top on the $F1$ curve, indicating that our network overall outperforms the baseline in terms of performance. Notably, even in challenging categories, our network consistently achieves high precision and recall, validating its robustness and effectiveness in complex scenarios.

V. DISCUSSION

The experimental results indicate that our network performs well on both the DroneVehicle and HIT-UAV datasets. In this section, we will conduct an in-depth discussion and analysis based on the experimental findings.

1) In the DroneVehicle dataset, cars, trucks, buses, vans, and freight cars exhibit similar shapes and sizes. Some algorithms merge these categories to enhance detection accuracy. However, distinguishing between vehicle types is essential for practical applications. As shown

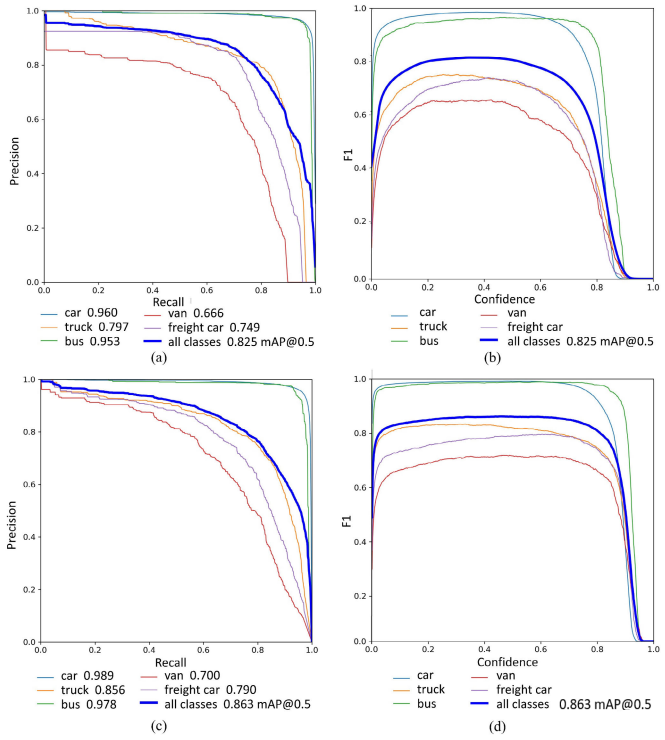

Fig. 8. Comparison of the P–R curves and $F1$ score curves between the baseline model and our proposed LAST-Net. (a) P–R curve of the baseline model. (b) $F1$ curve of the baseline model. (c) P–R curve of our LAST-Net. (d) $F1$ curve of our LAST-Net.

in Fig. 6, our network effectively differentiates these details, avoiding false positives and missed detections, even for objects with similar features.

2) In the HIT-UAV dataset, due to the limited sample features of pedestrians and bicycles in the top–down view, our network effectively avoids missed detections compared to other networks, as shown in Fig. 7.
3) The DroneVehicle dataset contains nearly ten times more samples than the HIT-UAV dataset, as indicated in Table I. By analyzing the results of ablation and comparative experiments across the two datasets, Tables III–V demonstrate that precision improvements are significantly more pronounced for the DroneVehicle dataset, indicating that larger datasets have a greater impact on enhancing transformer model performance.
4) In a series of comparative experiments with advanced algorithms, our method demonstrates greater robustness and higher precision, as shown in Tables III and IV and Figs. 6 and 7. This is particularly evident in scenarios with multiple objects and complex backgrounds. Compared to infrared small object detection algorithms, our approach achieves high detection precision even in low-contrast and low-resolution images. In contrast to remote sensing object detection algorithms, our method consistently identifies and classifies various objects under challenging conditions, such as oblique angles, occlusion, and multiple objects, highlighting its superior adaptability and generalization capabilities.
5) Overall, as shown in Tables III and IV, transformer-based models typically have more parameters than those based on CNNs-based. However, our network significantly reduces parameter count while maintaining high performance compared to other transformer-based frameworks. To further optimize the model and enhance its practicality, future work will focus on minimizing parameters and improving algorithm efficiency.

VI. Conclusion

To tackle the challenges of complex backgrounds, low contrast, multiscale objects, and imbalanced data samples in UAV-based infrared remote sensing images, we propose the LAST-Net algorithm. First, we propose the CST-Block, which integrates the powerful modeling capabilities of transformers with residual convolutional networks in the backbone. This combination enables the extraction of features from local to global scales, enhancing the network's versatility in addressing complex backgrounds. Second, the DMAA mechanism within the encoder–decoder structure achieves pixel-level weighting by fusing content and position matrices, effectively improving the low contrast of infrared images. Finally, in the detection head, the CCLF adjusts weight factors to help mitigate data imbalance issues, enhancing detection precision for classes with fewer labels. Additionally, by combining the DMAA mechanisms, one-to-one matching for multiscale and dense objects is achieved in both classification and localization, facilitating multiobject detection.

We conducted extensive experiments using the DroneVehicle and HIT-UAV datasets. The results indicate that our network achieves optimal detection precision compared to SOTA algorithms. Additionally, we performed a comprehensive series of ablation studies to demonstrate the network's effectiveness from both quantitative and qualitative perspectives. However, considering the limited payload capacity of UAVs, we will continue to optimize the algorithm to reduce the number of parameters in the future, making it better suited for practical applications.

References

[1] H. Yao, R. Qin, and X. Chen, "Unmanned aerial vehicle for remote sensing applications—A review," *Remote Sens.*, vol. 11, no. 12, p. 1443, Jun. 2019.
[2] D. Feng et al., "Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 3, pp. 1341–1360, Mar. 2021.
[3] S. Yuan, K. Ota, M. Dong, and J. Zhao, "A path planning method with perception optimization based on sky scanning for UAVs," *Sensors*, vol. 22, no. 3, p. 891, 2022.
[4] H. Wang et al., "Cross-modal oriented object detection of UAV aerial images based on image feature," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5403021.
[5] X. Kong, C. Yang, S. Cao, C. Li, and Z. Peng, "Infrared small target detection via nonconvex tensor fibered rank approximation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Apr. 2022, Art. no. 5000321.
[6] X. Sun, Y. Yu, and Q. Cheng, "Low-rank multimodal remote sensing object detection with frequency filtering experts," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5637114.
[7] D. Wang and J. Lan, "PPDet: A novel infrared pedestrian detection network in a per-pixel prediction fashion," *Infr. Phys. Technol.*, vol. 119, Dec. 2021, Art. no. 103965.
[8] D.-X. Zhou, "Deep distributed convolutional neural networks: Universality," *Anal. Appl.*, vol. 16, no. 6, pp. 895–919, Nov. 2018.
[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[10] H. Bao, L. Dong, S. Piao, and F. Wei, "BEiT: BERT pre-training of image transformers," 2021, *arXiv:2106.08254*.

[11] W. Wang et al., "CrossFormer++: A versatile vision transformer hinging on cross-scale attention," 2023, *arXiv:2303.06908*.

[12] R. Pope et al., "Efficiently scaling transformer inference," in *Proc. Mach. Learn. Syst. Nov.*, Jan. 2022, pp. 606–624.

[13] M. Guo et al., "Attention mechanisms in computer vision: A survey," *Comput. Vis. Media*, vol. 8, no. 3, pp. 331–368, 2022.

[14] X. Chu et al., "Twins: Revisiting the design of spatial attention in vision transformers," in *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS), Sep.*, Jan. 2021, pp. 9355–9366.

[15] Y. Tian, Z. Zhou, Z. Cui, and Z. Cao, "Scene adaptive SAR incremental target detection via context-aware attention and Gaussian-box similarity metric," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, 2025, Art. no. 5205217.

[16] X. Tao, M. E. Paoletti, Z. Wu, J. M. Haut, P. Ren, and A. Plaza, "An abundance-guided attention network for hyperspectral unmixing," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5205217.

[17] C. Zhao, J. Wang, N. Su, Y. Yan, and X. Xing, "Low contrast infrared target detection method based on residual thermal backbone network and weighting loss function," *Remote Sens.*, vol. 14, no. 1, p. 177, Jan. 2022.

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

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

[20] Y. Zhang, Y. Cao, and X. Xiang, "Ir small target detection based on morphological top-hat filter," *Comput. Meas. Control.*, vol. 19, no. 6, pp. 1269–1272, Jun. 2011.

[21] C. L. P. Chen, H. Li, Y. Wei, T. Xia, and Y. Y. Tang, "A local contrast method for small infrared target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 1, pp. 574–581, Jan. 2014.

[22] C. Yu et al., "Infrared small target detection based on multiscale local contrast learning networks," *Infr. Phys. Technol.*, vol. 123, Jun. 2022, Art. no. 104107.

[23] C. Yang, J. Ma, M. Zhang, S. Zheng, and X. Tian, "Multiscale facet model for infrared small target detection," *Infr. Phys. Technol.*, vol. 67, pp. 202–209, Nov. 2014.

[24] H. Deng, X. Sun, M. Liu, C. Ye, and X. Zhou, "Infrared small-target detection using multiscale gray difference weighted image entropy," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 52, no. 1, pp. 60–72, Feb. 2016.

[25] 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.

[26] 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.

[27] 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.

[28] H. Zhu, H. Ni, S. Liu, G. Xu, and L. Deng, "TNLRS: Target-aware non-local low-rank modeling with saliency filtering regularization for infrared small target detection," *IEEE Trans. Image Process.*, vol. 29, pp. 9546–9558, 2020.

[29] Y. Cao, T. Zhou, X. Zhu, and Y. Su, "Every feature counts: An improved one-stage detector in thermal imagery," in *Proc. IEEE 5th Int. Conf. Comput. Commun. (ICCC)*, Dec. 2019, pp. 1965–1969.

[30] C. Devaguptapu, N. Akolekar, M. M. Sharma, and V. N. Balasubramanian, "Borrow from anywhere: Pseudo multi-modal object detection in thermal imagery," 2019, *arXiv:1905.08789*.

[31] Q. Hou, Z. Wang, F. Tan, Y. Zhao, H. Zheng, and W. Zhang, "RISTDnet: Robust infrared small target detection network," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[32] S. Li, Y. Li, Y. Li, M. Li, and X. Xu, "YOLO-FIRI: Improved YOLOv5 for infrared image object detection," *IEEE Access*, vol. 9, pp. 141861–141875, 2021.

[33] X. Wu, D. Hong, and J. Chanussot, "UIU-Net: U-Net in U-Net for infrared small object detection," *IEEE Trans. Image Process.*, vol. 32, pp. 364–376, 2023.

[34] W. Li, M. Zhao, X. Deng, L. Li, L. Li, and W. Zhang, "Infrared small target detection using local and nonlocal spatial information," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 9, pp. 3677–3689, Sep. 2019.

[35] C. Zhang, K.-M. Lam, and Q. Wang, "CoF-Net: A progressive coarse-to-fine framework for object detection in remote-sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5600617.

[36] M. Li, J. Lan, L. Wang, Y. Zhang, and K. Huang, "Infrared multiobject contrast enhancement and detection based on layered visual transformer network for autonomous driving," *IEEE Sensors J.*, vol. 24, no. 22, pp. 38244–38255, Nov. 2024.

[37] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2016.

[38] R. Qin, Q. Liu, G. Gao, D. Huang, and Y. Wang, "MRDet: A multihead network for accurate rotated object detection in aerial images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5608412.

[39] J. Ding, N. Xue, Y. Long, G.-S. Xia, and Q. Lu, "Learning RoI transformer for oriented object detection in aerial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2844–2853.

[40] X. Ye, F. Xiong, J. Lu, J. Zhou, and Y. Qian, "F^3-Net: Feature fusion and filtration network for object detection in optical remote sensing images," *Remote Sens.*, vol. 12, no. 24, p. 4027, Dec. 2020.

[41] D. Avola et al., "MS-faster R-CNN: Multi-stream backbone for improved faster R-CNN object detection and aerial tracking from UAV images," *Remote Sens.*, vol. 13, no. 9, p. 1670, Apr. 2021.

[42] J. Butler and H. Leung, "A novel keypoint supplemented R-CNN for UAV object detection," *IEEE Sensors J.*, vol. 23, no. 24, pp. 30883–30892, Dec. 2023.

[43] G. Tang, J. Ni, Y. Zhao, Y. Gu, and W. Cao, "A survey of object detection for UAVs based on deep learning," *Remote Sens.*, vol. 16, no. 1, p. 149, Dec. 2023.

[44] X. Yang, J. Yan, Z. Feng, and T. He, "R3Det: Refined single-stage detector with feature refinement for rotating object," in *Proc. AAAI Conf. Artif. Intell., Dec.*, vol. 35, May 2021, pp. 3163–3171.

[45] M. Liu, X. Wang, A. Zhou, X. Fu, Y. Ma, and C. Piao, "UAV-YOLO: Small object detection on unmanned aerial vehicle perspective," *Sensors*, vol. 20, no. 8, p. 2238, Apr. 2020.

[46] C. Jiang et al., "Object detection from UAV thermal infrared images and videos using YOLO models," *Int. J. Appl. Earth Observ. Geoinformation*, vol. 112, Aug. 2022, Art. no. 102912.

[47] P. Sun, Y. Zheng, Z. Zhou, W. Xu, and Q. Ren, "R4Det: Refined single-stage detector with feature recursion and refinement for rotating object detection in aerial images," *Image Vis. Comput.*, vol. 103, pp. 1–12, Nov. 2020.

[48] M. Li, X. Zhao, J. Li, and L. Nan, "ComNet: Combinational neural network for object detection in UAV-borne thermal images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 8, pp. 6662–6673, Aug. 2021.

[49] L. Tan, X. Lv, X. Lian, and G. Wang, "YOLOv4_Drone: UAV image target detection based on an improved YOLOv4 algorithm," *Comput. Electr. Eng.*, vol. 93, Jul. 2021, Art. no. 107261.

[50] F. Wang, H. Wang, Z. Qin, and J. Tang, "UAV target detection algorithm based on improved YOLOv8," *IEEE Access*, vol. 11, pp. 116534–116544, 2023.

[51] Q. Fu, Q. Zheng, and F. Yu, "LMANet: A lighter and more accurate multiobject detection network for UAV remote sensing imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, pp. 1–5, 2024.

[52] A. Vaswani et al., "Attention is all you need," 2017, *arXiv:1706.03762*.

[53] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Aug. 2020, pp. 213–229.

[54] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," 2020, *arXiv:2010.04159*.

[55] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.

[56] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[57] Y. Zhao et al., "DETRs beat YOLOs on real-time object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 16965–16974.

[58] J. Yang et al., "Focal self-attention for local–global interactions in vision transformers," 2021, *arXiv:2107.00641*.

[59] D. Shi, "TransNeXt: Robust foveal visual perception for vision transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 17773–17783.

[60] J. Fu et al., "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3141–3149.

[61] X. Yang et al., "SCRDet: Towards more robust detection for small, cluttered and rotated objects," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8231–8240.

[62] T. Ye, W. Qin, Z. Zhao, X. Gao, X. Deng, and Y. Ouyang, "Real-time object detection network in UAV-vision based on CNN and transformer," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–13, 2023.

[63] W. Lu et al., "A CNN-transformer hybrid model based on CSWin transformer for UAV image object detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 1211–1231, 2023.

[64] C. Zhang, J. Su, Y. Ju, K.-M. Lam, and Q. Wang, "Efficient inductive vision transformer for oriented object detection in remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5616320.

[65] Z. Ni, Z. Zong, and P. Ren, "Incorporating object counts into remote sensing image captioning," *Int. J. Digit. Earth*, vol. 17, no. 1, pp. 1–16, Aug. 2024.

[66] M. Yuan and X. Wei, "C^2former: Calibrated and complementary transformer for RGB-infrared object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5403712.

[67] R. Cheng, H. Zeng, B. Zhang, X. Wang, and T. Zhao, "FFA-net: Fast feature aggregation network for 3D point cloud segmentation," *Mach. Vis. Appl.*, vol. 34, no. 5, pp. 1–14, Jul. 2023.

[68] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 3–19.

[69] W. Li, K. Liu, L. Zhang, and F. Cheng, "Object detection based on an adaptive attention mechanism," *Sci. Rep.*, vol. 10, no. 1, pp. 1–13, Jul. 2020.

[70] W. Ma, T. Zhou, J. Qin, Q. Zhou, and Z. Cai, "Joint-attention feature fusion network and dual-adaptive NMS for object detection," *Knowl.-Based Syst.*, vol. 241, Apr. 2022, Art. no. 108213.

[71] Y. Sun, B. Cao, P. Zhu, and Q. Hu, "Drone-based RGB-infrared cross-modality vehicle detection via uncertainty-aware learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 10, pp. 6700–6713, Oct. 2022.

[72] J. Suo, T. Wang, X. Zhang, H. Chen, W. Zhou, and W. Shi, "HIT-UAV: A high-altitude infrared thermal dataset for unmanned aerial vehicle-based object detection," 2022, *arXiv:2204.03245*.

[73] C.-Y. Wang, A. Bochkovskiy, and H.-Y.-M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 7464–7475.

[74] D. Reis, J. Kupec, J. Hong, and A. Daoudi, "Real-time flying object detection with YOLOv8," 2023, *arXiv:2305.09972*.

[75] N. Zhang, Y. Liu, H. Liu, T. Tian, and J. Tian, "Oriented infrared vehicle detection in aerial images via mining frequency and semantic information," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5002315.

[76] C. Bao, J. Cao, Q. Hao, Y. Cheng, Y. Ning, and T. Zhao, "Dual-YOLO architecture from infrared and visible images for object detection," *Sensors*, vol. 23, no. 6, p. 2934, Mar. 2023.

[77] D. Wang et al., "Advancing plain vision transformer toward remote sensing foundation model," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5607315.

[78] S. Wang et al., "PHSI-RTDETR: A lightweight infrared small target detection algorithm based on UAV aerial photography," *Drones*, vol. 8, no. 6, p. 240, Jun. 2024.

[79] B. Wan et al., "MFFNet: Multi-modal feature fusion network for V-D-T salient object detection," *IEEE Trans. Multimedia*, vol. 26, pp. 2069–2081, 2024.