Object Detection by Channel and Spatial Exchange for Multimodal Remote Sensing Imagery

Guozheng Nan, Yue Zhao, Liyong Fu, and Qiaolin Ye, *Member, IEEE*

***Abstract*—Smart satellites and unmanned aerial vehicles (UAVs) are typically equipped with visible light and infrared (IR) spectrum sensors. However, achieving real-time object detection utilizing these multimodal data on such resource-limited devices is a challenging task. This article proposes HyperYOLO, a real-time lightweight object detection framework for multimodal remote sensing images. First, we propose a lightweight multimodal fusion module named channel and spatial exchange (CSE) to effectively extract complementary information from different modalities. The CSE module consists of two stages: channel exchange and spatial exchange. Channel exchange achieves global fusion by learning global weights to better utilize cross-channel information correlation, while spatial exchange captures details by considering spatial relationships to calibrate local fusion. Second, we propose an effective auxiliary branch module based on the feature pyramid network for super resolution (FPNSR) to enhance the framework's responsiveness to small objects by learning high-quality feature representations. Moreover, we embed a coordinate attention mechanism to assist our network in precisely localizing and attending to the objects of interest. The experimental results show that on the VEDAI remote sensing dataset, HyperYOLO achieves a 76.72% mAP_{50}, surpassing the SOTA SuperYOLO by 1.63%. Meanwhile, the parameter size and GFLOPs of HyperYOLO are about 1.34 million (28%) and 3.97 (22%) less than SuperYOLO, respectively. In addition, HyperYOLO has a file size of only 7.3 MB after the removal of the auxiliary FPNSR branch, which makes it easier to deploy on these resource-constrained devices.**

***Index Terms*—Multimodal feature fusion, remote sensing image (RSI), RGB-infrared object detection, super resolution (SR).**

I. Introduction

OBJECT detection is an important task in the field of remote sensing image (RSI) processing, which not only contributes to applications in monitoring natural disasters and military reconnaissance but also has far-reaching impacts on urban planning and forest management. Traditional image feature extraction [1], [2] is important in computer vision, but its performance in object detection is limited by complex visual patterns and manual engineering. On the contrary, deep learning significantly improves detection performance by automatically learning discriminative features. In recent years, due to the rapid development of deep learning technology, many excellent algorithms [3], [4], [5], [6] have emerged in the field of object detection.

However, compared to general object detection tasks, RSIs have various characteristics such as complex backgrounds, small and densely arranged objects, and shadow occlusion. Therefore, it is necessary to adjust and optimize the model structure according to the characteristics of RSIs. Traditional object detection algorithms [7], [8], [9] are typically designed based on a single modality, primarily utilizing the visual information from images for detection, lacking the assistance of other modalities' information. This may result in limited feature representation capability and difficulty capturing the diversity and contextual information of objects in complex scenes. In addition, different sensors may encounter various defects and noise when acquiring target detection data [10]. These sensors can be influenced by factors such as weather conditions, terrain, obstructions, and shadows, leading to a decrease in image quality or incomplete target information. Therefore, relying solely on a single modality for target detection may have certain limitations. Furthermore, certain targets may be difficult to discern in specific modalities but may become more apparent in other modalities. For instance, in RGB images, some targets may blend with the background color, making them challenging to distinguish. However, in IR images, these targets may exhibit distinct thermal features compared to the surrounding environment, making them easier to identify. By fusing information from both RGB and IR modalities, it is possible to enhance the visibility and distinctiveness of targets under different spectra. Manish et al. [11] improved the performance of detection through the introduction of a fusion strategy for multimodal data. To discover potential correlations between different modalities, many researchers employ complex fusion modules, such as transformer [12] and illumination-aware [13], which lead to increased computational complexity. Similarly, widely adopted fusion methods, encompassing feature-level and decision-level fusion [14], [15], [16] may lead to redundant computations among different modality branches or the introduction of additional backbone networks, thereby restricting the deployment of the model. The recent development direction of remote sensing object detection algorithms [9], [17],

Manuscript received 9 March 2024; accepted 7 April 2024. Date of publication 12 April 2024; date of current version 24 April 2024. This work was supported in part by the National Key Research and Development Program under Grant 2022YFD2201005-03, in part by the National Natural Science Foundation of China under Grant 62072246 and Grant 32371877, in part by the Technology Winter Olympics Special Project under Grant 201001D, and in part by the Forest Fire Comprehensive System Construction-Unmanned Aerial Patrol Monitoring System of Chongli under Grant DA2-20001. *(Corresponding author: Qiaolin Ye.)*

Guozheng Nan, Yue Zhao, and Qiaolin Ye are with the College of Information Science and Technology, College of Artificial Intelligence, Nanjing Forestry University, Nanjing 210037, China (e-mail: gzn@njfu.edu.cn; zyue0109@163.com; yqlcom@njfu.edu.cn).

Liyong Fu is with the Institute of Forest Resource Information Techniques, Chinese Academy of Forestry, Beijing 100091, China, and also with the College of Forestry, Hebei Agricultural University, Baoding 071000, China.

Digital Object Identifier 10.1109/JSTARS.2024.3388013

[18] involves the use of super resolution (SR) technology to learn the mapping relationship from low resolution (LR) images to high resolution (HR) images, enabling the reconstruction and detection of LR images. Although this approach can improve the performance of small target detection by increasing the detailed information in the image, the benefit comes at the cost of increased model complexity, introducing a certain level of intricacy and time overhead. Zhang et al. [19] introduced an auxiliary SR branch to guide the detector in learning high-quality HR representations, facilitating the distinction of small objects from the LR input background. It is worth noting that many upsampling methods, like bilinear and nearest-neighbor interpolation, estimate new pixel values from nearby pixels. While these increase image size, they may lose texture details, and thus, hinder small object reconstruction in LR images. In addition, many methods fail to fully utilize multiscale information when increasing the size of images, which limits their performance.

In recent years, the you only look once (YOLO) series of algorithms [7], [20], [21], [22], [23] has emerged as a representative in the field of object detection due to its rapid, accurate, and proven engineering capabilities. To further enhance real-time performance and achieve efficient object detection in computationally constrained environments, several lightweight and real-time improvement algorithms based on YOLO have emerged. Zi et al. [24] proposed TP-YOLO, integrating self-attention mechanisms and omnidimensional dynamic convolution (ODConv) [25] into YOLOv8 [23] to improve small target detection while reducing parameters and computational complexity. However, ODConv may not be suitable for sparse image processing since it requires computing unique convolution kernels for each position and channel, making it difficult to achieve stable and reliable convolution operations with sparse input data. In addition, Nvidia's acceleration library, TensorRT, is not very friendly toward ODConv operations. Zhang et al. [26] introduced FFCA-YOLO, an improved version of YOLOv5 [22], to address the issue of insufficient feature representation in small object detection, achieving performance enhancement through feature fusion and spatial context-aware modules while minimizing complexity. However, FFCA-YOLO is an algorithm designed based on a single modality, and its detection performance may not meet expectations when encountering extreme weather conditions or severe target occlusion.

In light of the aforementioned challenges, we propose a real-time object detection framework for multimodal RSIs that excels not only in having fewer parameters and lower computational complexity but also in delivering exceptional detection accuracy. First, we choose the YOLOv7tiny [7] architecture as our detection baseline. It has a smaller network structure and fewer parameters, making it easier to deploy and run with limited computational resources. Second, we propose a novel lightweight fusion module employing channel and spatial exchange (CSE) to ensure that each modality retains its unique features while effectively integrating the prominent features of other modalities. Moreover, an efficaciously assisted branch module based on the feature pyramid network [8] for super resolution (FPNSR), which uses the PixelShuffle (PS) upsampling [27] method and multiscale feature information, is designed to preserve the features of small targets in LR input within the backbone network. Meanwhile, it is necessary to remove the FPNSR branch to avoid additional computational overhead in the inference and deployment stages. Finally, we introduced a coordinate attention (CA) [28] mechanism to suppress interference from complex backgrounds on small objects, while focusing on the regions of interest to enhance boundary accuracy and preserve fine details. In summary, the main contributions of our article are as follows.

1) We propose a computationally efficient lightweight multimodal fusion method that symmetrically and compactly combines internal information in a bidirectional manner, utilizing spatial and channel relationships to ensure both upper and lower branches focus on each other's complementary information.
2) We designed the FPNSR auxiliary branch to address inconspicuous features in small targets within LR inputs, utilizing SR techniques to enhance the model's ability to learn HR feature representations and improve the identification of small objects in cluttered backgrounds.
3) The FPNSR branch is a flexible component, guiding the network to enhance the detector's responsiveness to small targets during training, and it can be removed during deployment to reduce computational demands while maintaining target detection accuracy.
4) We introduce the CA mechanism to enable our framework to focus on specific regions within the target image, thereby more effectively capturing crucial information and achieving precise localization and recognition of the target.

II. Related Work

A. Remote Sensing Object Detection

HR satellite RSIs usually contain rich information about ground objects, making the high-precision extraction of these objects through the application of remote sensing object detection a current hot research direction [29], [30]. Object detection algorithms often fall into two categories: two-stage methods based on anchor boxes, such as Faster R-CNN [31] and Mask R-CNN [32], and single-stage detectors like SSD [33] and the YOLO series [22], [34], [35], among others. Recently, some anchor-free detectors, such as the fully convolutional FCOS [36], and transformer-based algorithms like deformable DETR [37] and Swin Transformer [38], have also achieved remarkable results. The two-stage detection method consists of two core components: the feature extraction network and the candidate region generation network. Initially, the feature extraction network identifies the regions of interest in the image. Subsequently, the candidate region generation network further processes the features to generate potential target candidate regions. While this two-stage approach achieves high accuracy, it comes at the cost of slower detection speed. On the contrary, the single-stage detection method directly predicts the category and location information of the object without generating candidate regions to improve the detection speed. Given the significant

differences between general objects and remote sensing objects, researchers have proposed various solutions to address the issue of small object detection in remote sensing imagery for the one-stage detector YOLO series networks effectively. Zakria et al. [39] proposed an improvement for the YOLOv4 [21] network by introducing a nonmaximum suppression threshold classification setting and an anchor box assignment scheme. Lin et al. [30] introduced a decoupled detection head and terminal attention mechanism to enhance the target localization performance of the YOLOv5 framework. Yi et al. [40] used a visual transform for feature extraction and designed an attention-guided bidirectional FPN to improve the performance of YOLOv8 for small target detection in RSIs.

Although the aforementioned improvement methods have made progress in certain aspects, they only utilize unimodal data and fail to fully exploit the inherent value of multimodal data. The vibrant advancement of imaging technology offers greater opportunities for collecting multimodal data in RSIs presenting new possibilities to improve the accuracy of RSI analysis and detection.

B. SR in Object Detection

Existing methods for small object detection in RSIs mainly focus on two aspects: context information and multiscale processing [41], [42], [43]. However, these methods overlook a crucial issue, which is the severe loss of feature information for small objects after multiple downsampling operations in RSIs, as well as the inadequate preservation of HR contextual information. As a result, the current models exhibit poor detection accuracy for RSIs. The SR technology is a highly regarded research direction in the fields of computer vision and image processing [44], [45]. In recent years, it has made significant advancements and found widespread application, particularly in the domain of RSI processing [18]. Ji et al. [9] introduced a two-branch network for simultaneous SR and target detection, where the SR branch generates HR feature maps for use in the detection branch, and jointly optimizes the SR and target detection losses to train the network. Courtrai et al. [46] improved small object detection in RSIs by using generative adversarial networks (GANs) and the EDSR [47] network to generate HR images for input into a detector. Although the above methods address the challenge of small target detection to some extent, they are not suitable for practical deployment of models in real-time application scenarios due to the introduction of a large number of additional computations, including the complexity associated with SR techniques and the fact that HR features increase the complexity of the detection model.

Recently, Wang et al. [48] and Zhang et al. [19] proposed an SR module, respectively, which can maintain an HR representation even with LR inputs, while reducing model computation in segmentation and detection tasks. However, a limitation of these SR modules is their underutilization of multiscale features. In SR networks, multiscale features play a crucial role in enhancing reconstruction quality and preserving fine details. By integrating information from different scales, the network can better capture subtle changes and texture details in the image. Building upon these structures, we propose a multiscale SR module, namely, the FPNSR module, that achieves high-quality LR reconstruction to preserve HR representation by integrating features from multiple scales.

C. CA Mechanism

The CA mechanism is a computational unit used to enhance the feature representation capability of convolutional neural networks (CNNs). Its design purpose is to assist the model in focusing on important locations and content while addressing the potential issue of position information loss in the squeeze-and-excitation (SE) [49] attention module. To counteract spatial information loss caused by 2-D global pooling layers, the CA mechanism utilizes two 1-D networks to generate X and Y 1-D features, producing corresponding attention features aligned with the spatial characteristics of the image. Specifically, as illustrated in Fig. 2, the CA mechanism utilizes two 1-D global pooling layers to extract directional features along the vertical and horizontal directions from image features. Then, these directional feature maps are concatenated and subjected to dimension reduction using a 1×1 convolution, followed by nonlinear activation operations, generating a new feature map. Subsequently, the feature map is split along the spatial dimension, resulting in two split features. Each split feature is further subjected to dimension expansion using a 1×1 convolution and finally combined with a sigmoid activation function to obtain the final attention vector feature. This operation effectively captures long-term dependencies in image features along both directions, preserving spatial information. The combination of these attention vector features with the original image is achieved through elementwise multiplication, resulting in image features weighted by attention scores which indicate the degree of emphasis on the regions of interest within the image features.

The CA mechanism not only captures crucial features across channel dimensions but also possesses the ability to perceive and extract spatial coordinate features in different directions, effectively highlighting objects of interest in the input features. In addition, with a low computational cost and complexity, the CA mechanism can be efficiently utilized in object detection models, achieving powerful enhancement of features.

III. Proposed Method

A. Baseline Architecture

Designed for edge computing devices, YOLOv7tiny is a model in the YOLOv7 series that boasts a smaller model size and faster inference speed. As shown in Fig. 1, the YOLOv7tiny network architecture consists of three main components: the Backbone, Neck, and Head. The Backbone section is composed of several Convolution-BatchNorm-LeakyReLU (CBL) modules, UP modules, MP modules, and efficient layer aggregation network (ELAN) [50] modules. The UP is built using CBL modules and upsampling operations. The MP performs max pooling operations. The ELAN, composed of multiple stacked CBL modules and featuring a two-branch structure, contributes to the reduction of gradient propagation delay and information

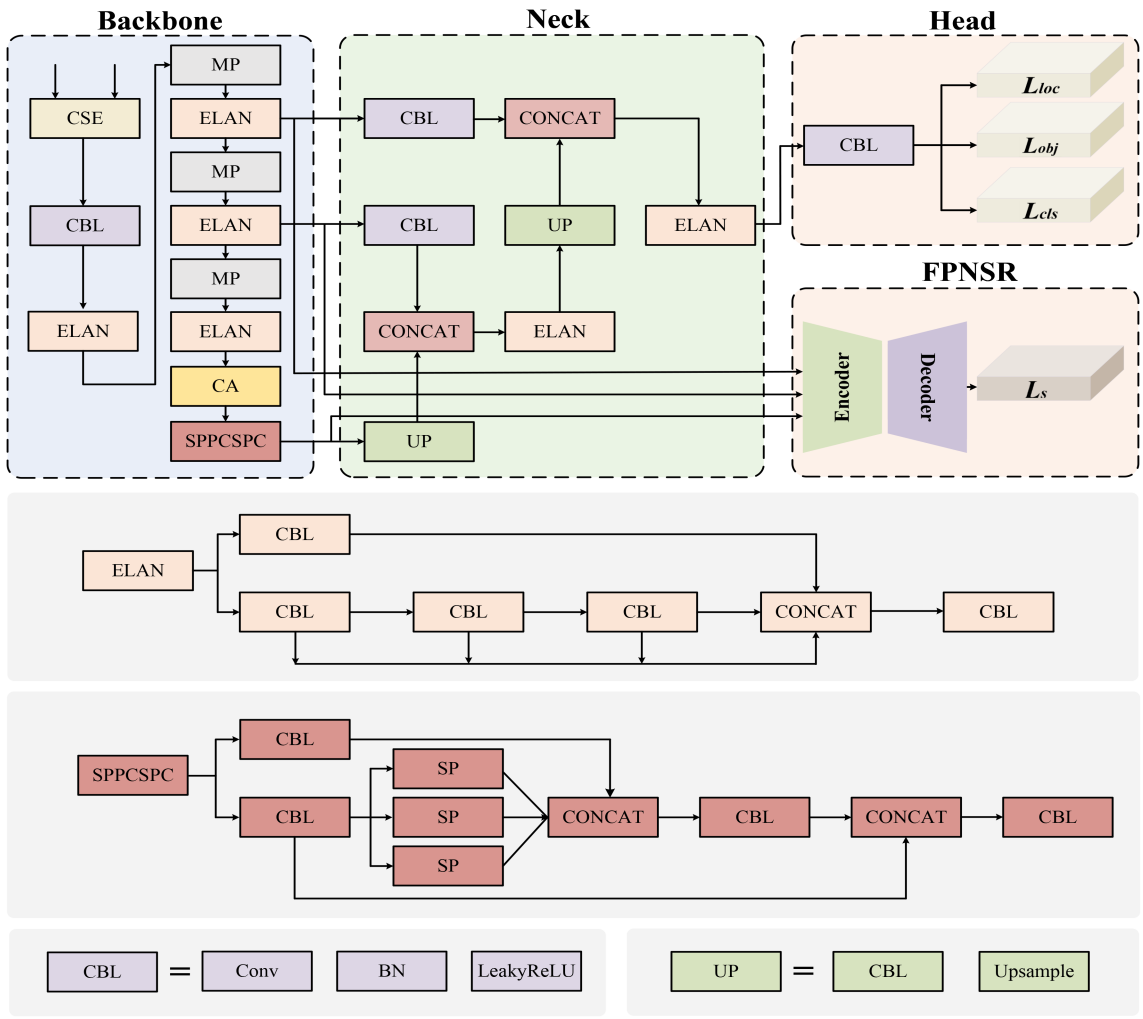

Fig. 1. Overview of the proposed HyperYOLO framework. Our new contributions include 1) the CSE fusion module, 2) the assisted branch module based on the FPNSR, 3) the only detector designed to enhance small-scale objects, and 4) the CA mechanism added before the SPPCSPC module.

loss. The spatial pooling pyramid cross stage partial convolution (SPPCSPC) modules, based on the spatial pyramid pooling (SPP) [51] modules, add a "residual path" and stack it with the output of parallel max pooling layers with different kernel sizes, capturing information at multiple scales. The Neck section employs the path aggregation network (PANet) [52] structure, introducing both top–down and bottom–up path aggregation to merge low-level details with high-level semantic information. The Head section adopts CBL modules for channel adjustment to predict bounding box positions, class information, and confidence scores. To enhance the model's ability for target localization and scale perception, we have introduced a CA attention module in the preceding layer of the SPPCSPC. This module can adaptively adjust the weights of different channels in the feature map, ensuring that channels crucial for the current task receive more attention, while less important channels receive less attention. In addition, we removed two detectors from the PANet structure, leaving only the one that enhances small-scale objects (small-scale detector). The purpose of this modification is to expedite the convergence speed of our network in the task of detecting small objects in RSIs, while simultaneously meeting the requirements of model iteration more quickly, without sacrificing accuracy.

B. CSE Multimodal Fusion

The CSE fusion module can learn shared and complementary features between different modalities by exchanging channel and spatial information, thereby achieving better feature extraction and interaction. The architecture of the CSE module is depicted in Fig. 3. Due to the use of the auxiliary training branch, the first step is to downsample both input modalities, $I_{\mathrm{RGB}}, I_{\mathrm{IR}} \in \mathbb{R}^{C\times H\times W}$ to $F_{\mathrm{RGB}}, F_{\mathrm{IR}} \in \mathbb{R}^{C\times \frac{H}{n}\times \frac{W}{n}}$, where n represents the downsampling factor. The enhanced features, denoted as $F_{e\mathrm{RGB}}$ and $F_{e\mathrm{IR}}$, are obtained by applying a 1×1 convolution to F_{RGB} and F_{IR}, followed by elementwise multiplication with themselves, which are formulated as

$$F_{e\mathrm{RGB}} = (\mathrm{Conv}_{1\times1}(F_{\mathrm{RGB}})) \otimes F_{\mathrm{RGB}} \tag{1}$$

$$F_{e\mathrm{IR}} = (\mathrm{Conv}_{1\times1}(F_{\mathrm{IR}})) \otimes F_{\mathrm{IR}} \tag{2}$$

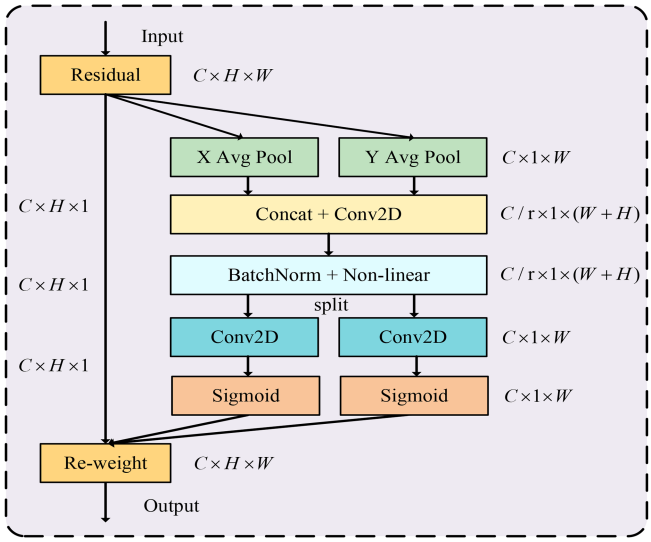

Fig. 2. Network architecture of the CA module. C represents the number of channels in the feature map. H and W denote the height and width of the feature map, respectively. r is expressed as the reduction factor.

where $\otimes$ represents elementwise matrix multiplication. Apply average pooling and max pooling separately to the features obtained after concatenating two enhanced features, and then concatenate the results again to acquire F

$$F_e = \text{Concat}(F_{e\text{RGB}}, F_{e\text{IR}}) \tag{3}$$

$$F = \text{Concat}\left(AP\left(F_e\right), MP\left(F_e\right)\right) \tag{4}$$

where $\text{Concat}(\cdot)$ indicates the concatenation operation along the channel axis. The average pooling and max pooling are used to preserve more feature information. F represents the global contextual features in the I_{RGB} and I_{IR} channel dimensions. After applying the MLP, which consists of two linear layers and two activation functions (ReLU and Sigmoid) to F, it is divided along the channel axis to obtain the channel attention vectors $W_{\text{RGB}}^{C} \in \mathbb{R}^{3C}$ and $W_{\text{IR}}^{C} \in \mathbb{R}^{C}$

$$W_{\text{RGB}}^{C}, W_{\text{IR}}^{C} = f_{\text{split}}\left(f_{\text{mlp}}(F)\right) \tag{5}$$

where W_{RGB}^{C} and W_{IR}^{C}, respectively, represent the attention weights on the RGB input feature and IR input feature channels. The introduction of these weights allows the model to better utilize the cross-channel information correlation, thereby enhancing the processing capability for different channel features. Through channel exchange, the two input modal features can suppress irrelevant backgrounds while enhancing the representation capability of object features. Channel exchange focuses on learning global weights for global fusion, and further introduces spatial exchange to calibrate local fusion. F_e is first fed with two 1×1 convolution layers assembled with ReLU and Sigmoid functions to obtain the feature map S, which is then divided into two spatial weight maps $W_{\text{RGB}}^{S} \in \mathbb{R}^{\frac{H}{n}\times\frac{W}{n}}$ and $W_{\text{IR}}^{S} \in \mathbb{R}^{\frac{H}{n}\times\frac{W}{n}}$

$$S = \text{Sigmoid}\left(\text{Con}v_{1\times1}\left(\text{ReLU}\left(\text{Con}v_{1\times1}\left(F_e\right)\right)\right)\right) \tag{6}$$

$$W_{\text{RGB}}^{S}, W_{\text{IR}}^{S} = f_{\text{split}}(S). \tag{7}$$

The bimodal inputs integrate global feature information through embedding into the channel and spatial vector weights to, respectively, attain RGB_{out} and IR_{out}

$$\text{RGB}_{\text{out}} = F_{e\text{RGB}} \otimes W_{\text{RGB}}^{C} + F_{e\text{RGB}} \otimes W_{\text{RGB}}^{S} \tag{8}$$

$$\text{IR}_{\text{out}} = F_{e\text{IR}} \otimes W_{\text{IR}}^{C} + F_{e\text{IR}} \otimes W_{\text{IR}}^{S}. \tag{9}$$

This weighted fusion mechanism allows the model to fully leverage the correlations between features from different channels and spatial positions, thereby enabling the output features to better reflect the structural details and semantic information of the image. RGB_{out} and IR_{out} need to be added to the original modality features and fed into the 1×1 convolution separately to obtain more comprehensive feature representations

$$F_{\text{out1}} = \text{Conv}_{1\times1}\left(F_{\text{RGB}} + \text{RGB}_{\text{out}}\right) \tag{10}$$

$$F_{\text{out2}} = \text{Conv}_{1\times1}\left(F_{\text{IR}} + \text{IR}_{\text{out}}\right). \tag{11}$$

The final fused feature of the Backbone's input, achieved by dynamically weighting the complete features of different modalities using an attention mechanism, is represented as

$$F_o = \text{CA}\left(\text{Concat}\left(F_{\text{out1}}, F_{\text{out2}}\right)\right) \tag{12}$$

where $\text{CA}(\cdot)$ refers to the CA attention network, which facilitates the fusion by dynamically assigning attention to spatial positions.

C. Feature Pyramid Network for Super Resolution

In deep neural networks, shallow features typically have higher resolution and rich geometric information but smaller receptive fields and lack semantic information. Conversely, deep features have larger receptive fields and rich semantic information but relatively lower resolution and less geometric information. To preserve the features of small targets as much as possible in the backbone network, we adopt a feature pyramid structure to construct the FPNSR auxiliary branch. By introducing lateral connections and upsampling operations, we fuse feature maps from different scales to generate features with high-level semantic information, providing clearer texture details for the SR network. The flexible FPNSR branch is composed of a simple encoder–decoder structure, where the encoder captures texture details at various scales to generate a high-semantic lowest level feature, and the decoder is responsible for upsampling the lowest level feature. Its role during training is to facilitate the backbone in constructing HR feature representations, enhancing the detection model's responsiveness to small objects.

For the FPNSR branch, we select the results from the fifth, seventh, and 11th modules of the Backbone network as the low-level, midlevel, and high-level features, respectively. Shallow layers in neural networks typically excel at capturing low-level features such as edges and textures, while as the network depth increases, later layers gradually capture more abstract high-level features such as object shapes and parts. These selected features are utilized as inputs for FPNSR due to their provision of diverse and valuable feature representations at various levels. As depicted in Fig. 4, the high-level feature is upsampled and concatenated with the midlevel feature processed by the

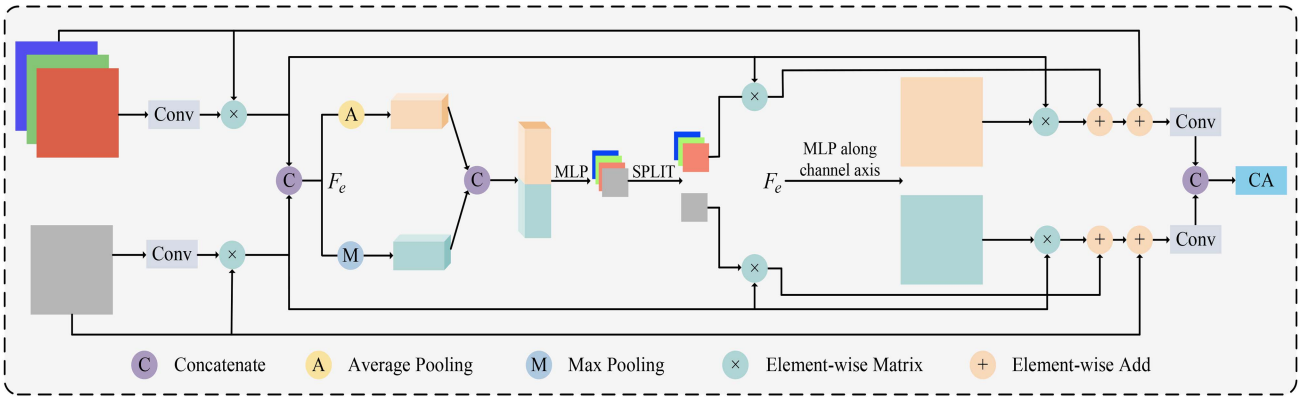

Fig. 3. Architecture of the CSE fusion module.

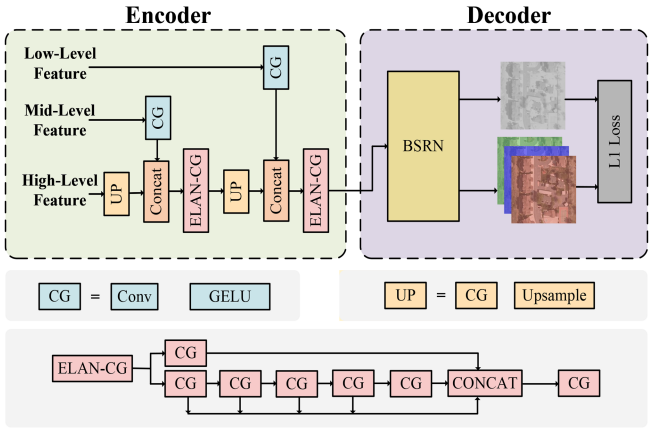

Fig. 4. FPNSR structure of HyperYOLO. The FPNSR can be considered as a simple encoder–decoder model.

convolution-GELU (CG) module (1×1 convolution) before being fed to the ELAN-CG feature enhancement module, which is constructed with the ELAN structure consisting of seven stacked CG modules. The CG module includes a convolution and GELU [53] activation function. The ELAN-CG module utilizes a two-branch structure, with the first branch employing a CG module (1×1 convolution) for channel adjustment, and the second branch initiating with a CG module (1×1 convolution) followed by four CG modules (3×3 convolution) for feature extraction. Ultimately, the features from both branches are concatenated to yield the final result of feature extraction, enhancing the expressive capacity of features without modifying the width and height of the input features through the regulation of the shortest and longest gradient paths. After upsampling the output of the first ELAN-CG module and concatenating it with the low-level feature, the concatenated result is fed to another ELAN-CG module to generate the output LR feature of the encoder.

In the decoder, the LR feature is upsampled into the HR feature using the BSRN [54] network, which employs blueprint separable convolution (BSConv) [55] to effectively reduce redundant computations and uses PS upsampling method that combines channel information to fill pixels. The use of the GELU activation function ensures consistency with the BSRN network in terms of representation capacity and nonlinear characteristics. The BSRN network performs exceptionally well in SR tasks, particularly in reconstructing high-quality and visually pleasing HR images from LR inputs. In addition, the BSConv technique offers advantages in situations where computational resources or time constraints are important. In our FPNSR structure, refraining from using batch normalization is more beneficial for preserving the original contrast and feature information of the image. The output size of the FPNSR module is twice that of the downsampled input image.

D. Loss Function

The total loss function of our network, composed of detection loss L_o and FPNSR loss L_s, is defined by the following formula:

$$L_{\text{total}} = \alpha L_o + \beta L_s \tag{13}$$

where α and β act as balance coefficients, enabling flexible control and adjustment of the training task. Since the network's input is RGB and IR multimodal data, we need to split the results of FPNSR into three channels for S_{rgb} and one channel for S_{ir}. L_s is obtained by adding the L1 loss between the input images I_{rgb} and I_{ir} with S_{rgb} and S_{ir}, respectively, which is expressed as

$$L_s = \| I_{\text{rgb}} - S_{\text{rgb}} \|_1 + \| I_{\text{ir}} - S_{\text{ir}} \|_1 \tag{14}$$

The loss function L_o encompasses three distinct components: the object detection loss L_{obj} (reflecting the confidence score of object presence), the classification loss L_{cls}, and the bounding box regression loss L_{loc}, which is expressed as follows:

$$L_o = \lambda_1 L_{\text{cls}} + \lambda_2 L_{\text{obj}} + \lambda_3 L_{\text{loc}} \tag{15}$$

where λ_1, λ_2, and λ_3 are weight coefficients used to adjust the influence of each loss term on the overall loss function.

IV. Experiments and Results

A. Datasets

The experiment uses two publicly available datasets: the vehicle detection in aerial imagery (VEDAI) [56] and the Bayberry

Tree (BT) [57]. The VEDAI dataset is designed for the study of aerial visual object recognition, which includes two image sizes, 512×512 and 1024×1024, as well as two modal types, RGB images, and IR images. Each image in the dataset contains an average of 5.5 targets, occupying approximately 0.7% of the total pixel count per image. The dataset consists of a total of 1210 images, which focus on different backgrounds and include various scenes such as grasslands, mountains, deserts, rural areas, and urban areas. Within these images, the targets are categorized into nine classes, including car, truck, van, camping, pickup, boat, and more.

The BT dataset was created by aerial photography using the DJI Phantom 4 drone. It was captured in the experimental zone of Dayangshan Forest Park, Yongjia County, Zhejiang Province, China, on January 23–24, 2019, and consists of 284 HR RSIs, each with a resolution of 1024×682. Owing to the lack of IR modality images in the dataset, an infrared generative adversarial network (InfraGAN) [59] was used to generate the texture features of IR modality. The training dataset for this InfraGAN was derived from the VEDAI dataset. When generating images in the IR modality for the BT dataset, we initially employed the letterbox technique to resize the RGB images to 512×512. Subsequently, these resized images are passed through the InfraGAN network to generate IR images.

B. Evaluation Metrics

We utilize two types of mean average precision (mAP) as the primary evaluation metrics, namely, mAP_{50} and $mAP_{50:95}$. The calculation of mAP is based on recall and precision. Recall represents the ratio of correctly detected targets to the actual number of targets, while precision represents the ratio of correctly detected targets to the total number of detected targets. Recall and precision are calculated as follows:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{16}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{17}$$

where TP is the number of correctly predicted positive samples, FN is the number of false negatives (actual positives incorrectly predicted as negatives), and FP is the number of negative samples wrongly predicted as positive by the model. The mAP is a comprehensive metric obtained by calculating the area under the precision-recall curve. It measures the average precision of the model at different levels of recall and provides an overall performance evaluation. The mAP is calculated by the integration method as

$$\text{mAP} = \frac{1}{N} \sum_{i=1}^{N} \int_0^1 P_i(R) dR \tag{18}$$

where P represents precision, R represents recall, and N denotes the number of categories. In addition, we employ two commonly used metrics, which are the model Parameters (Params) and giga floating-point operations per second (GFLOPs) to evaluate the performance and efficiency of the model.

TABLE I
COMPARISON RESULTS OF COMPLEXITY AND ACCURACY IN DIFFERENT BASELINE YOLO FRAMEWORKS ON THE FIRST FOLD OF THE VEDAI VALIDATION SET USING THE CONCAT FUSION METHOD

Methods	Layers	Params(M)	GFLOPs	mAP_{50}(%)
YOLOv3 [20]	270	61.5	52.8	62.6
YOLOv3SPP [20]	270	62.6	49.8	67.4
YOLOrs [11]	241	20.2	46.4	55.8
YOLOv4 [21]	393	52.5	38.2	65.7
YOLOR [58]	410	52.5	38.2	67.3
YOLOv5s [22]	224	7.10	5.32	62.2
YOLOv5m [22]	308	21.1	16.1	64.5
YOLOv5l [22]	397	46.6	36.7	63.9
YOLOv5x [22]	476	87.3	69.7	64.0
YOLOv7tiny [7]	**209**	**6.03**	**4.21**	62.3
YOLOv7 [7]	315	36.5	33.1	**70.0**
YOLOv7x [7]	363	70.8	60.3	66.8

C. Implementation Details

All experiments were performed on a workstation equipped with an NVIDIA 3060 GPU, using the PyTorch framework. Images from the BT dataset were randomly divided into training and testing sets at an 8:2 ratio. Following [19], the VEDAI dataset is designed for tenfold cross-validation. Each fold has 1089 training images and 121 validation images. Ablation experiments are conducted on the first fold of data. When comparing with state-of-the-art (SOTA) methods, we employ the average of the results from tenfolds. The network is trained using the standard stochastic gradient descent (SGD), with a momentum of 0.937 and weight decay of 0.0005 employed for the Nesterov accelerated gradients and a batch size of 4. The initial learning rate is set to 0.01, and cosine annealing is used for learning rate decay. To achieve the assisted FPNSR branch, the input images of the network are downsampled from 1024×1024 to 512×512 during the training process, while the image size is set to 512×512 during the testing process.

D. Ablation Studies

Our ablation study aims to validate the effectiveness of the proposed module, and for this purpose, we conducted a series of experiments in the first fold of the VEDAI dataset. In addition, we will validate the effectiveness of our model on both the VEDAI and BT datasets by comparing it with SOTA methods. In our experiment, we compared the performance across different metrics for various models and highlighted the best performance of the model metrics by using bold font to emphasize it.

1) Validation of the Baseline Framework: We evaluated the performance of different base frameworks in terms of accuracy and inference capability by considering the model's layer count, parameter size, GFLOPs, and mAP_{50}. All networks were trained and tested using an image resolution of 512×512, and a multimodal fusion method along channels was employed for all. As shown in Table I, although YOLOv7 achieves the best detection performance, it has 106 more layers than YOLOv7tiny (315 versus 209), its parameter size is approximately six times

TABLE II
COMPARISON RESULTS OF DIFFERENT FUSION METHODS ON THE FIRST FOLD OF THE VEDAI DATASET

Methods	$mAP_{50}(\%)$	$mAP_{50:95}(\%)$	Params(M)	GFLOPs
CONCAT	78.25	48.16	**3.4366**	**13.31**
MF	$78.94_{\uparrow 0.69}$	$49.31_{\uparrow 1.15}$	3.4732	13.98
CSE	$\mathbf{80.12}_{\uparrow 1.87}$	$\mathbf{51.08}_{\uparrow 2.92}$	3.4745	13.99

TABLE III
ABLATION EXPERIMENT RESULTS ABOUT CA AND FPNSR ON THE FIRST FOLD OF THE VEDAI VALIDATION SET

Method	CA	FPNSR	$mAP_{50}(\%)$	$mAP_{50:95}(\%)$	Params(M)	GFLOPs
YOLOv7tiny [7] Small-scale Detector	✗	✗	80.12	51.08	**3.4745**	**13.99**
	✗	✓	$80.99_{\uparrow 0.87}$	$51.84_{\uparrow 0.76}$	**3.4745**	**13.99**
	✓	✗	$80.35_{\uparrow 0.23}$	$51.82_{\uparrow 0.74}$	3.5002	14.01
	✓	✓	$\mathbf{82.58}_{\uparrow 2.46}$	$\mathbf{52.42}_{\uparrow 1.34}$	3.5002	14.01

TABLE IV
VALIDATION RESULTS OF FPNSR BRANCH FOR DIFFERENT BASELINES ON THE FIRST FOLD OF THE VEDAI VALIDATION SET

Methods	Layers	Params(M)	GFLOPs	$mAP_{50}(\%)$
YOLOv3SPP [20]	270	61.5	52.8	73.1
YOLOv3SPP+FPNSR	270	61.5	52.8	$\mathbf{75.6}_{\uparrow 2.5}$
YOLOR [58]	410	52.5	38.2	78.9
YOLOR+FPNSR	410	52.5	38.2	$\mathbf{80.7}_{\uparrow 1.8}$
YOLOv7 [7]	315	36.5	33.1	78.7
YOLOv7+FPNSR	315	36.5	33.1	$\mathbf{79.9}_{\uparrow 1.2}$
YOLOv7x [7]	363	70.8	60.3	78.1
YOLOv7x+FPNSR	363	70.8	60.3	$\mathbf{79.0}_{\uparrow 0.9}$

larger than that of YOLOv7tiny (36.5 versus 6.03 M), and its GFLOPs is approximately 7.9 times higher than that of YOLOv7tiny (33.1 versus 4.21). Considering the practical deployment and real-time performance of the model, we have chosen YOLOv7tiny as our most suitable baseline network. Despite its slightly lower mAP_{50} compared to YOLOv5l and YOLOv3, YOLOv7tiny has significantly fewer layers, smaller parameter sizes, and lower GFLOPs than other models. This makes it more efficient and well-suited for deployment in resource-constrained environments. The aforementioned experimental results validate the rationale behind choosing YOLOv7tiny as the baseline detection framework.

2) Comparisons of Different Fusion Methods: To evaluate the proposed fusion method, we also employed two other methods: feature concatenation (CONCAT) along the channel axis and multimodal fusion (MF) [19]. The results in Table II indicate that while there is not a significant difference in the Params and GFLOPs among the three methods, the utilization of the CSE module resulted in the best performance, achieving 80.12% mAP_{50} and 51.08% $mAP_{50:95}$, surpassing the CONCAT method by 1.87% and 2.92%, respectively. It shows that the CSE module provides effective feature representation for detection. The CSE module significantly improves accuracy through multiple mechanisms. First, it leverages the channel relationships within feature maps to capture fine-grained details that are crucial for object detection. Second, by recalibrating the spatial relationships in feature responses, it allows the fusion branches to focus on complementary information. In addition, the CSE module emphasizes important features in multimodal data while suppressing irrelevant ones, thereby enhancing the discriminative ability of the fused features and ultimately improving object detection performance.

3) Effectiveness Analysis of FPNSR Auxiliary Branch: To analyze the effectiveness of the FPNSR auxiliary branch, we selected four additional baseline networks: YOLOv3SPP, YOLOR, YOLOv7, and YOLOv7x, which differ in terms of network structure and training strategies. Due to the use of the FPNSR auxiliary branch, it is necessary to downsample the input image size from 1024×1024 to 512×512. To eliminate the impact of this downsampling on experimental accuracy, we ensured that all models performed this operation. As shown in Table IV, compared to the bare baseline, the baseline with the addition of the FPNSR auxiliary branch demonstrates favorable performance: YOLOv3SPP+FPNSR showcases a 2.5% improvement in mAP_{50} over YOLOv3, YOLOR+FPNSR exhibits a 1.8% increase in mAP_{50} compared to YOLOR, YOLOv7+FPNSR demonstrates a 1.2% improvement in mAP_{50} relative to YOLOv7, and YOLOv7x+FPNSR registers a 0.9% uptick in mAP_{50} when compared to YOLOv7x. The experimental results indicate that the FPNSR auxiliary branch has a positive impact on object detection tasks across different baseline models, without introducing additional parameters or computational costs.

4) Impacts of FPNSR Auxiliary Branch and CA: To explore the relationship between the CA attention module and the FPNSR auxiliary branch, a series of ablation experiments were conducted on our selected baseline network. As shown in Table III, compared to the baseline network, training solely with FPNSR resulted in small margin improvements of 0.87% and 0.76% in mAP_{50} and $mAP_{50:95}$, respectively. This suggests that the FPNSR module contributes to enhancing the performance of the baseline network. However, adding CA in the layer preceding the SPPCSPC modules, mAP_{50} and $mAP_{50:95}$ remarkably increased by 2.46% and 1.34%, respectively. The CA attention incorporates positional information into the attention mechanism to provide the texture structure of the region of interest for the FPNSR auxiliary module. These findings highlight the significance of incorporating both the CA module and the FPNSR auxiliary branch in the network architecture, as they synergistically contribute to achieving better object detection performance.

5) Comparisons with Single Modality Model: Table V summarizes the performance of different modalities on the first fold of the VEDAI validation set for models YOLOv7tiny, YOLOv7, YOLOv7x, SuperYOLO, and our proposed HyperYOLO, where the first three models employ the CONCAT method for multimodal fusion. It is clear that the mAP score in multimodal (multi) mode is higher than in unimodal (IR or RGB) model. Despite a slight increase in the Params and GFLOPs of the multimodal model, it is highly worthwhile considering the improvement it brings to mAP. In a single modality, HyperYOLO achieves

TABLE V
PERFORMANCE COMPARISON OF DIFFERENT MODALITIES FOR VARIOUS MODELS ON THE FIRST FOLD OF THE VEDAI DATASET

Methods		$mAP_{50}(\%)$	$mAP_{50:95}(\%)$	Params(M)	GFLOPs
YOLOv7tiny [7]	IR	60.34	34.23	**6.03**	**4.19**
	RGB	66.30	37.18	**6.03**	**4.19**
	Multi	**70.02**	**41.03**	6.03	4.21
YOLOv7 [7]	IR	64.84	39.89	**36.52**	**33.05**
	RGB	74.78	44.27	**36.52**	**33.05**
	Multi	**78.71**	**46.14**	36.52	33.12
YOLOv7x [7]	IR	66.60	40.91	**70.83**	**60.20**
	RGB	76.79	47.16	**70.83**	**60.20**
	Multi	**78.08**	**48.76**	70.83	60.30
SuperYOLO [19]	IR	67.53	40.80	**4.83**	**16.61**
	RGB	77.55	48.69	**4.83**	**16.61**
	Multi	**81.31**	**49.89**	4.85	17.98
HyperYOLO	IR	75.88	46.30	**3.48**	**12.62**
	RGB	79.61	49.77	**3.48**	**12.62**
	Multi	**82.58**	**52.42**	3.50	14.01

TABLE VI
PERFORMANCE COMPARISON OF DIFFERENT MODALITIES FOR VARIOUS MODELS ON THE FIRST FOLD OF THE BT DATASET

Methods		$mAP_{50}(\%)$	$mAP_{50:95}(\%)$	Params(M)	GFLOPs
YOLOv7tiny [7]	IR	91.06	51.46	**6.03**	**4.19**
	RGB	94.00	56.82	**6.03**	**4.19**
	Multi	**95.12**	**60.07**	6.03	4.21
YOLOv7 [7]	IR	92.13	51.24	**36.52**	**33.05**
	RGB	94.96	56.70	**36.52**	**33.05**
	Multi	**94.98**	**60.71**	36.52	33.12
YOLOv7x [7]	IR	92.30	53.67	**70.83**	**60.20**
	RGB	94.64	58.38	**70.83**	**60.20**
	Multi	**95.19**	**61.02**	70.83	60.30
SuperYOLO [19]	IR	92.26	51.10	**7.05**	**20.22**
	RGB	95.55	60.41	**7.05**	**20.22**
	Multi	**96.05**	**61.94**	7.07	21.59
HyperYOLO	IR	93.14	53.27	**6.03**	**16.77**
	RGB	95.61	63.20	**6.03**	**16.77**
	Multi	**96.58**	**64.32**	6.05	18.17

significantly higher mAP scores than other frameworks, especially in the case of the IR modality, with mAP_{50} and $mAP_{50:95}$ surpassing the YOLOv7 by 11.04% and 6.41%, respectively. The Params and GFLOPs of HyperYOLO are about 10× and 2.6× less than YOLOv7. We not only validated the effectiveness of HyperYOLO on the VEDAI dataset with smaller targets but also extended our validation to the BT dataset with larger targets. We solely consider these models' fusion detection capability for the BT dataset, thus, both HyperYOLO and SuperYOLO cancel auxiliary training branches and use the complete PANet structure with three detectors. Similarly, in Table VI, we present experimental results for three modalities of different models on the first fold of the BT validation set. We enhance the detection accuracy of the model by leveraging IR images generated using the InfraGAN network through multimodal fusion. This indicates that the IR images generated by this network have a positive impact on our object detection framework. We observed that the multimodal mode in the HyperYOLO framework demonstrated excellent performance in terms of the $mAP_{50:95}$ metric, achieving a score of 64.32%. Compared to other models, it exhibited remarkable advantages in precise target localization [high intersection over union (IOU)] and tolerance to variations in object positions (low IOU). Our HyperYOLO single-modal models (IR and RGB) achieved mAP_{50} scores of 93.14% and 95.61% respectively, surpassing other models. This demonstrates the effectiveness of the CA mechanism we introduced in our model, which enhances the weights of crucial feature channels and improves the model's perception and discrimination of objects. Meanwhile, Fig. 5 presents the confusion matrix of the prediction results obtained by using the YOLOv7 and the proposed HyperYOLO multimodal models on the VEDAI dataset. It's evident that the confusion matrix plot for the predictions of our HyperYOLO model has larger values on the diagonal, indicating a higher number of correctly detected samples. Furthermore, our model's confusion matrix exhibits fewer nondiagonal elements compared to YOLOv7's, indicating a reduced occurrence of misclassifications or missed detections. Fig. 6 shows the effect of visualizing the prediction results for the two datasets using HyperYOLO's models with different modalities. It can be observed that the CSE fusion method utilizes complementary information to accurately detect objects that might be undetected or incorrectly predicted when using a single modality model.

6) Comparisons with State-of-the-Art Methods: To bolster the credibility of our results and ensure the resilience and relevance of our model, we implemented a tenfold cross-validation on both the VEDAI and BT datasets. This strategic approach not only minimized experimental errors but also enabled a thorough evaluation of the model's stability and its capacity to generalize effectively across diverse data samples. SuperYOLO has achieved excellent tradeoffs between speed and accuracy on the VEDAI dataset, so it is necessary for us to compare our method with SuperYOLO to assess our performance in this aspect. Moreover, it is worth noting that SuperYOLO outperforms numerous models within the YOLOv3, YOLOv4, and YOLOv5 series in terms of overall performance. Table VII presents a comparative analysis of model performance, encompassing YOLOv7tiny, YOLOv7, YOLOv7x, YOLOR, YOLORx, R50-CSP, X50-CSP, L-FFCA-YOLO, TP-YOLO, SuperYOLO, and HyperYOLO, evaluated on the VEDAI and BT datasets. The backbone networks of the R50-CSP and X50-CSP detection models, respectively, adopt improved ResNet and ResNeXt structures [60], both of which incorporate the cross stage partial (CSP) [61] mechanism to reduce information loss and enhance network performance. L-FFCA-YOLO is a streamlined version of FFCA-YOLO, which reconstructs the backbone and neck using partial convolution (PConv) [62]. Table VII clearly demonstrates that our HyperYOLO model excels on the VEDAI dataset, achieving a remarkable 76.72% mAP_{50} and 47.98% $mAP_{50:95}$. This performance surpasses SuperYOLO by 1.63% and 1.89%, respectively, and outperforms YOLOv7x by an impressive margin of 5.01% and 3.48%. Moreover, HyperYOLO achieves these results with significantly fewer parameters and GFLOPs compared to both SuperYOLO and YOLOv7x, with approximately

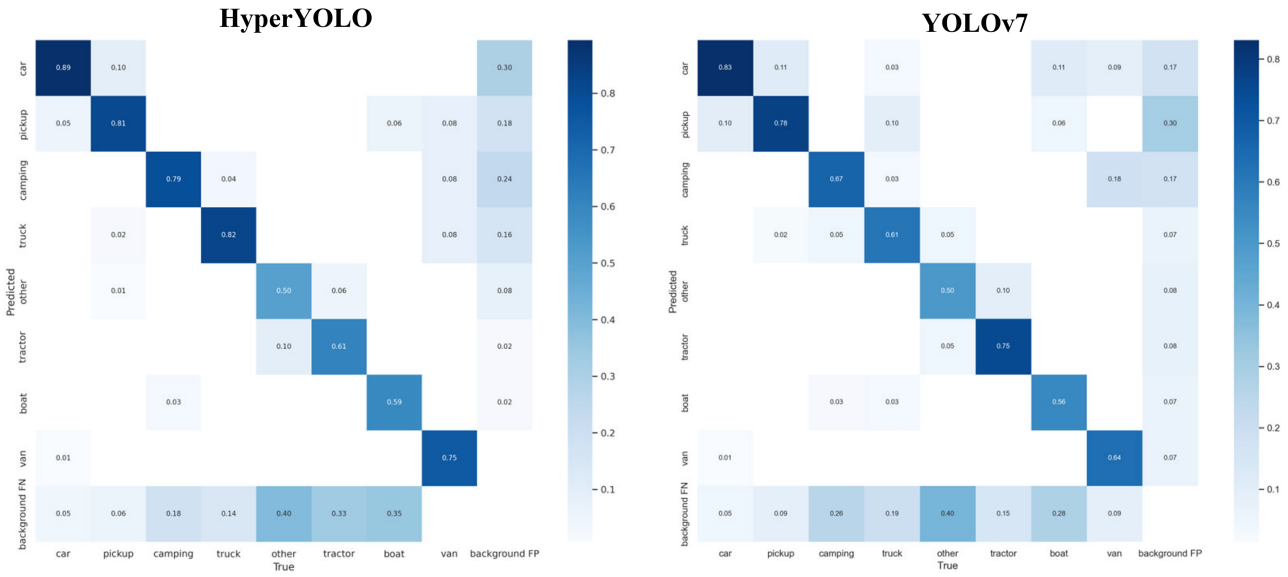

Fig. 5. Confusion matrix plot of the predictions for both HyperYOLO and YOLOv7 illustrates the correspondence between predicted results and true labels in matrix form.

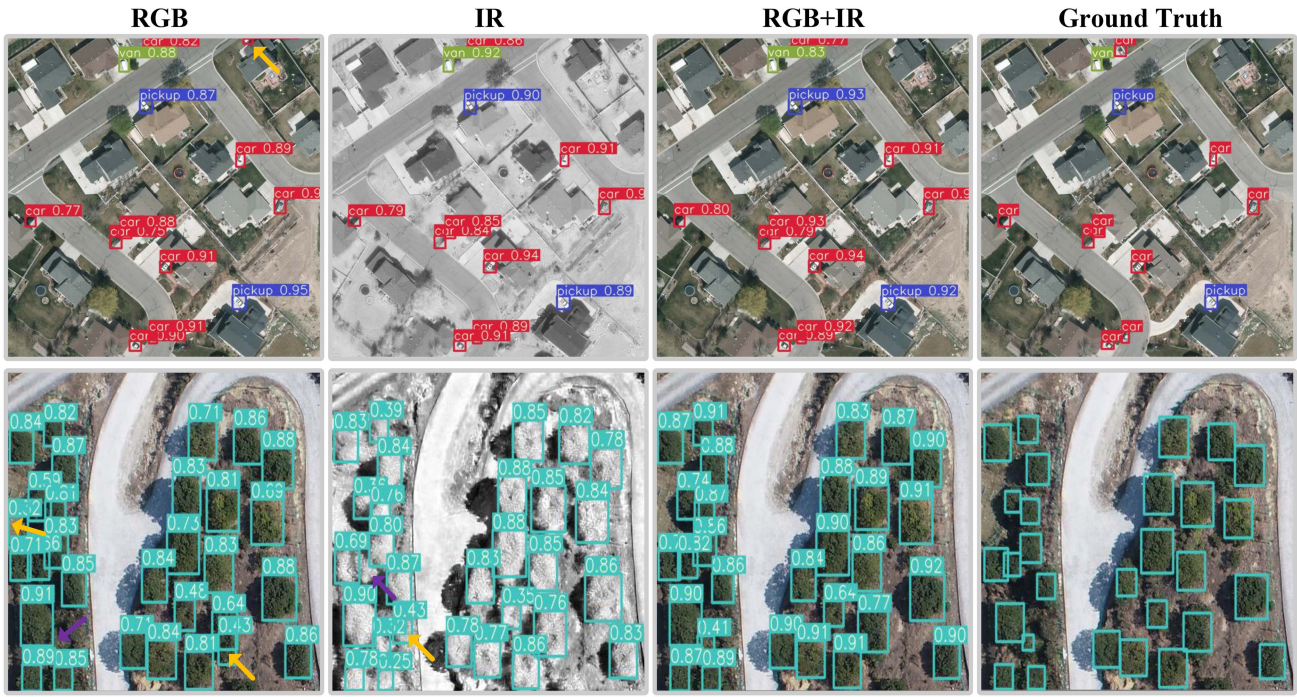

Fig. 6. Visual results for different modalities of the VEDAI and BT datasets using the HyperYOLO. The yellow arrow indicates the false positive, and the purple indicates the false negative. The model of the IR modal found targets that the RGB modal model failed to detect, while the RGB modal model corrected erroneous targets in the IR modal detection. The multimodal model compensates for the limitations of the unimodal model.

1.34 million fewer parameters and 3.97 less GFLOPs than SuperYOLO, and 20× fewer parameters and 4.3× less GFLOPs than YOLOv7x. Compared to the lightweight models TP-YOLO (4.29 M Params and 2.93 GFLOPs) and L-FFCA-YOLO (5.06 M Params and 11.89 GFLOPs), our HyperYOLO (3.50 M Params and 14.01 GFLOPs) still exhibits significant advantages in terms of performance. Although TP-YOLO has the lowest GFLOPs, its performance in terms of Params and mAP is far inferior to that of HyperYOLO. Our HyperYOLO model not only has the lowest parameter count but also outperforms other models in terms of mAP. This fact further validates the effectiveness of our proposed CSE and FPNSR modules. On the BT dataset, HyperYOLO still achieved the best results with 95.59% mAP_{50} and 64.62% $mAP_{50:95}$, surpassing other models while maintaining lower Params and GFLOPs. Although the parameter count of HyperYOLO is slightly higher than YOLOv7tiny, its

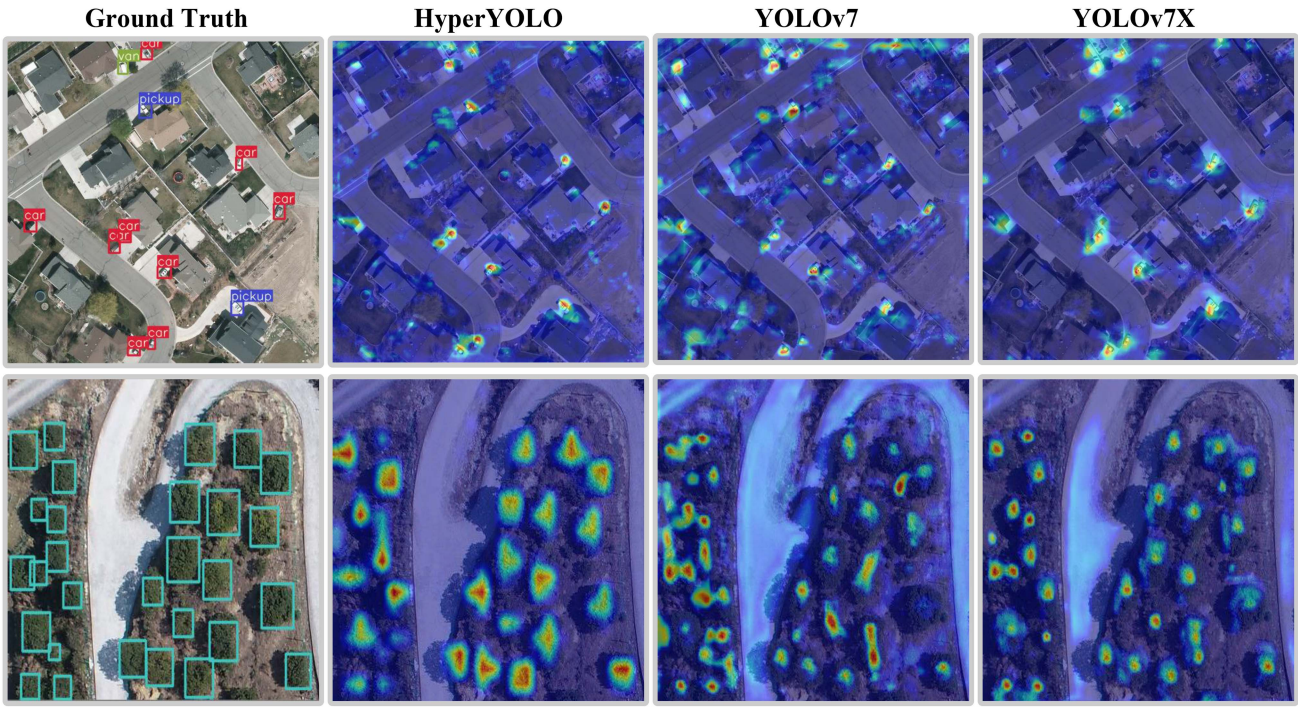

Fig. 7. Heat map comes from the head layers of our proposed HyperYOLO and YOLOv7 and YOLOV7x, respectively. The heat map is typically used to visualize the activation values or attention distribution of a neural network, helping to understand the network's response to input data.

TABLE VII
PERFORMANCE COMPARISON OF VARIOUS MULTIMODAL MODELS ON BOTH THE VEDAI AND BT DATASETS

Datasets	Methods	$mAP_{50}(\%)$	$mAP_{50:95}(\%)$	Params(M)	GFLOPs
VEDAI	YOLOv7tiny [7]	65.93	38.37	6.03	4.21
	YOLOv7 [7]	69.77	42.91	36.52	33.12
	YOLOv7x [7]	71.71	44.50	70.83	60.30
	YOLOR [58]	72.03	44.70	52.50	38.22
	YOLORx [58]	71.18	43.29	96.41	72.09
	R50-CSP [60] [61]	70.17	43.29	33.88	20.49
	X50-CSP [60] [61]	70.46	42.78	31.87	18.99
	L-FFCA-YOLO [24]	73.3	44.7	5.06	11.89
	TP-YOLO [24]	70.45	42.26	4.29	**2.93**
	SuperYOLO [19]	75.09	46.09	4.85	17.98
	HyperYOLO	**76.72**	**47.98**	**3.50**	14.01
BT	YOLOv7tiny [7]	94.60	61.49	6.03	4.21
	YOLOv7 [7]	94.78	63.46	36.52	33.12
	YOLOv7x [7]	95.46	63.26	70.83	60.30
	YOLOR [58]	94.29	59.85	52.50	38.22
	YOLORx [58]	93.82	59.66	96.41	72.09
	R50-CSP [60] [61]	94.02	59.29	33.88	20.49
	X50-CSP [60] [61]	94.11	58.37	31.87	18.99
	L-FFCA-YOLO [24]	95.02	62.31	5.06	11.89
	TP-YOLO [24]	94.27	61.29	**4.29**	**2.93**
	SuperYOLO [19]	95.19	59.65	7.07	21.59
	HyperYOLO	**95.59**	**64.62**	6.05	18.17

performance exceeds that of YOLOv7tiny by 0.99% in mAP_{50} and 3.13% in $mAP_{50:95}$. Furthermore, in Fig. 7, we visualized the heatmaps of the Head layers in HyperYOLO, YOLOv7, and YOLOv7x multimodal models using the XGrad-CAM [63] technique, showcasing the attention distribution of CNNs. The visualization results of the HyperYOLO model for the VEDAI dataset clearly show a more distinct representation of the target structure, entirely focusing on the target without additional attention to other areas. Regarding the BT dataset, the visualization results of the HyperYOLO model indicate that its attention coverage on targets is significantly superior to the results obtained with YOLOv7 and YOLOv7x. This phenomenon suggests that when dealing with certain RSIs containing larger objects, HyperYOLO can still identify points of interest more accurately and effectively. The presented data underscores our model's ability to strike an advantageous equilibrium between speed and accuracy, showcasing its robust and reliable performance across a spectrum of remote sensing datasets. This proficient balance not only ensures efficient processing but also highlights the model's resilience in handling diverse and complex information inherent in remote sensing scenarios.

V. CONCLUSION

In this article, we have proposed a real-time lightweight object detection framework tailored for multimodal RSIs, featuring faster inference speed and lower computational resource consumption. This provides an innovative and feasible solution for deploying real-time object detection in resource-limited environments. To address the high computational complexity of multimodal fusion methods, we have proposed an innovative CSE fusion module that effectively captures information from

different modalities. Furthermore, we have introduced an auxiliary FPNSR branch to enhance the framework's capability in recognizing small objects, and this is removed in the inference and deployment stages to avoid additional computational overhead. The experiments demonstrate that both the CSE fusion module and FPNSR module have significantly positive effects on our baseline network.

The performance and inference capabilities of the framework proposed in this article underscore the value of the CSE fusion module and FPNSR auxiliary training branch in RSIs object detection tasks, offering a feasible solution for future research in multimodal object detection. In the future, we will explore advanced techniques such as knowledge distillation, pruning, and quantization to enhance the performance of our model and adapt it to a broader range of resource-constrained devices, such as vehicle-mounted cameras and handheld medical scanners.

REFERENCES

[1] L. Fu et al., "Learning robust discriminant subspace based on joint $L_{2,p}$- and $L_{2,s}$-norm distance metrics," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 1, pp. 130–144, Jan. 2022.

[2] Q. Ye, J. Yang, H. Zheng, and L. Fu, "Comments and corrections convergence analysis on trace ratio linear discriminant analysis algorithms," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published, doi: 10.1109/TNNLS.2024.3355422.

[3] L. Fu, D. Zhang, and Q. Ye, "Recurrent thrifty attention network for remote sensing scene recognition," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 10, pp. 8257–8268, Oct. 2021.

[4] 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.*, 2020, pp. 213–229.

[5] D. Zhang, Y. Lin, J. Tang, and K.-T. Cheng, "CAE-GReaT: Convolutional-auxiliary efficient graph reasoning transformer for dense image predictions," *Int. J. Comput. Vis.*, pp. 1–19, 2023.

[6] Y. Shen, D. Zhang, Z. Song, X. Jiang, and Q. Ye, "Learning to reduce information bottleneck for object detection in aerial images," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, 2023, Art. no. 3001705.

[7] 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.*, 2023, pp. 7464–7475.

[8] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2117–2125.

[9] H. Ji, Z. Gao, T. Mei, and B. Ramesh, "Vehicle detection in remote sensing images leveraging on simultaneous super-resolution," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 4, pp. 676–680, Apr. 2020.

[10] J. Wang, K. Song, Y. Bao, L. Huang, and Y. Yan, "CGFNet: Cross-guided fusion network for RGB-T salient object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 5, pp. 2949–2961, May 2022.

[11] M. Sharma et al., "Yolors: Object detection in multimodal remote sensing imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 1497–1508, 2021.

[12] Y. Zhu, X. Sun, M. Wang, and H. Huang, "Multi-modal feature pyramid transformer for RGB-infrared object detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 9, pp. 9984–9995, Sep. 2023.

[13] L. Tang, J. Yuan, H. Zhang, X. Jiang, and J. Ma, "PIAFusion: A progressive infrared and visible image fusion network based on illumination aware," *Inf. Fusion*, vol. 83, pp. 79–92, 2022.

[14] L. Gómez-Chova, D. Tuia, G. Moser, and G. Camps-Valls, "Multimodal classification of remote sensing images: A review and future directions," *Proc. IEEE*, vol. 103, no. 9, pp. 1560–1584, Sep. 2015.

[15] D. Guan, Y. Cao, J. Yang, Y. Cao, and M. Y. Yang, "Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection," *Inf. Fusion*, vol. 50, pp. 148–157, 2019.

[16] H. Hermessi, O. Mourali, and E. Zagrouba, "Multimodal medical image fusion review: Theoretical background and recent advances," *Signal Process.*, vol. 183, 2021, Art. no. 108036.

[17] J. Yang, K. Fu, Y. Wu, W. Diao, W. Dai, and X. Sun, "Mutual-feed learning for super-resolution and object detection in degraded aerial imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5628016.

[18] Y. Wang et al., "Remote sensing image super-resolution and object detection: Benchmark and state of the art," *Expert Syst. Appl.*, vol. 197, 2022, Art. no. 116793.

[19] 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, Art. no. 5605415.

[20] R. Joseph et al., "Yolov3: An incremental improvement," 2018, *arXiv:1804.02767*.

[21] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.

[22] G. Jocher et al., "ultralytics/yolov5: V6.0," 2021. [Online]. Available: https://github.com/ultralytics/yolov5

[23] Ultralytics, "ultralytics/yolov8," 2023. [Online]. Available: https://github.com/ultralytics/ultralytics

[24] Y. Di, S. L. Phung, J. Van Den Berg, J. Clissold, and A. Bouzerdoum, "Tp-yolo: A lightweight attention-based architecture for tiny pest detection," in *Proc. IEEE Int. Conf. Image Process.*, 2023, pp. 3394–3398.

[25] C. Li, A. Zhou, and A. Yao, "Omni-dimensional dynamic convolution," 2022, *arXiv:2209.07947*.

[26] Y. Zhang, M. Ye, G. Zhu, Y. Liu, P. Guo, and J. Yan, "FFCA-YOLO for small object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5611215.

[27] W. Shi et al., "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1874–1883.

[28] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 13713–13722.

[29] Q. Ming, L. Miao, Z. Zhou, and Y. Dong, "CFC-Net: A critical feature capturing network for arbitrary-oriented object detection in remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5605814.

[30] J. Lin, Y. Zhao, S. Wang, and Y. Tang, "YOLO-DA: An efficient YOLO-based detector for remote sensing object detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, 2023, Art. no. 6008705.

[31] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1440–1448.

[32] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2961–2969.

[33] W. Liu et al., "SSD: Single shot multibox detector," in *Proc. 14th Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.

[34] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 779–788.

[35] J. Redmon and A. Farhadi, "Yolo9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 7263–7271.

[36] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 9627–9636.

[37] 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*.

[38] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 10012–10022.

[39] Z. Zakria, J. Deng, R. Kumar, M. S. Khokhar, J. Cai, and J. Kumar, "Multiscale and direction target detecting in remote sensing images via modified YOLO-v4," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 1039–1048, 2022.

[40] H. Yi, B. Liu, B. Zhao, and E. Liu, "Small object detection algorithm based on improved yolov8 for remote sensing," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 1734–1747, 2024.

[41] Z. Deng, H. Sun, S. Zhou, J. Zhao, L. Lei, and H. Zou, "Multi-scale object detection in remote sensing imagery with convolutional neural networks," *ISPRS J. Photogrammetry Remote Sens.*, vol. 145, pp. 3–22, 2018.

[42] B. Cheng, Z. Li, B. Xu, C. Dang, and J. Deng, "Target detection in remote sensing image based on object-and-scene context constrained CNN," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 8013705.

[43] S. Zhang, G. He, H.-B. Chen, N. Jing, and Q. Wang, "Scale adaptive proposal network for object detection in remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 6, pp. 864–868, Jun. 2019.

[44] J. Noh, W. Bae, W. Lee, J. Seo, and G. Kim, "Better to follow, follow to be better: Towards precise supervision of feature super-resolution for small object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 9725–9734.
[45] M. Haris, G. Shakhnarovich, and N. Ukita, "Task-driven super resolution: Object detection in low-resolution images," in *Proc. 28th Int. Conf. Neural Inf. Process.*, 2021, pp. 387–395.
[46] L. Courtrai, M.-T. Pham, and S. Lefèvre, "Small object detection in remote sensing images based on super-resolution with auxiliary generative adversarial networks," *Remote Sens.*, vol. 12, no. 19, 2020, Art. no. 3152.
[47] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, "Enhanced deep residual networks for single image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2017, pp. 136–144.
[48] L. Wang, D. Li, Y. Zhu, L. Tian, and Y. Shan, "Dual super-resolution learning for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 3774–3783.
[49] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.
[50] C.-Y. Wang, H.-Y. M. Liao, and I.-H. Yeh, "Designing network design strategies through gradient path analysis," 2022, *arXiv:2211.04800*.
[51] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015.
[52] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8759–8768.
[53] D. Hendrycks and K. Gimpel, "Gaussian error linear units (Gelus)," 2016, *arXiv:1606.08415*.
[54] Z. Li et al., "Blueprint separable residual network for efficient image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 833–843.
[55] D. Haase and M. Amthor, "Rethinking depthwise separable convolutions: How intra-kernel correlations lead to improved mobilenets," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 14600–14609.
[56] S. Razakarivony and F. Jurie, "Vehicle detection in aerial imagery: A small target detection benchmark," *J. Vis. Commun. Image Representation*, vol. 34, pp. 187–203, 2016.
[57] D. Wang and W. Luo, "Bayberry tree recognition dataset based on the aerial photos and deep learning model," *J Glob. Change Data Discover*, vol. 3, no. 3, pp. 290–296, 2019.
[58] C.-Y. Wang, I.-H. Yeh, and H.-Y. M. Liao, "You only learn one representation: Unified network for multiple tasks," 2021, *arXiv:2105.04206*.
[59] M. A. Özkanoğlu and S. Ozer, "InfraGAN: A GAN architecture to transfer visible images to infrared domain," *Pattern Recognit. Lett.*, vol. 155, pp. 69–76, 2022.
[60] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[61] C.-Y. Wang, H.-Y. M. Liao, Y.-H. Wu, P.-Y. Chen, J.-W. Hsieh, and I.-H. Yeh, "CSPNet: A new backbone that can enhance learning capability of CNN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. workshops*, 2020, pp. 390–391.
[62] G. Liu, F. A. Reda, K. J. Shih, T.-C. Wang, A. Tao, and B. Catanzaro, "Image in painting for irregular holes using partial convolutions," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 85–100.
[63] R. Fu, Q. Hu, X. Dong, Y. Guo, Y. Gao, and B. Li, "Axiom-based grad-cam: Towards accurate visualization and explanation of CNNs," 2020, *arXiv:2008.02312*.

Guozheng Nan received the B.S. degree in computer science and technology from Henan Normal University, Xinxiang, China, in 2022. He is currently working toward the M.S. degree in computer technology with the College of Information Science and Technology, College of Artificial Intelligence, Nanjing Forestry University, Nanjing, China.

His current research interests include computer vision and remote sensing.

Yue Zhao received the B.S. degree in network engineering from the Tongda College of Nanjing University of Posts and Telecommunications, Yangzhou, China, in 2022. She is currently working toward the M.S. degree in computer technology with the College of Information Science and Technology, College of Artificial Intelligence, Nanjing Forestry University, Nanjing, China.

Her current research interests include computer vision and remote sensing.

Liyong Fu received the B.S. degree in forestry from Shanxi Agricultural University, Jinzhong, China, in 2007, the M.S. degree in forest biometrics from Nanjing Forestry University, Nanjing, China, in 2009, and the Ph.D. degree in forest biometrics from the Chinese Academy of Forestry, Beijing, China, in 2012.

He is currently a Full Professor with the Department of Forest Management and Statistics, Chinese Academy of Forestry, and the College of Forestry, Hebei Agricultural University, Baoding, China.

Qiaolin Ye (Member, IEEE) received the B.S. degree in computer science from the Nanjing Institute of Technology, Nanjing, China, in 2007, the M.S. degree in computer science and technology from Nanjing Forestry University, Nanjing, in 2009, and the Ph.D. degree in pattern recognition and intelligence systems from the Nanjing University of Science and Technology, Nanjing, in 2013.

He is currently a Full Professor with the Department of Computer Science, Nanjing Forestry University, and the Key Laboratory of Intelligent Information Processing, Nanjing Xiaozhuang University, Nanjing. His research interests include machine learning, data mining, and pattern recognition.