IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 22 November 2023, accepted 19 December 2023, date of publication 22 December 2023,
date of current version 29 December 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3345889

APPLIED RESEARCH

Infrared Object Detection Method Based on DBD-YOLOv8

LINGYUN SHEN[1], BAIHE LANG[2], AND ZHENGXUN SONG[2,3]

[1]Department of Electronic Engineering, Taiyuan Institute of Technology, Taiyuan 030008, China
[2]School of Electronics and Information Engineering, Changchun University of Science and Technology, Changchun 130022, China
[3]Overseas Expertise Introduction Project for Discipline Innovation, Changchun University of Science and Technology, Changchun 130022, China

Corresponding authors: Lingyun Shen (shenshly@163.com) and Baihe Lang (langbh@gmail.com)

This work was supported in part by the Shanxi Province Talent Introduction Science and Technology Innovation Startup Fund, China.

ABSTRACT An innovative and improved method for infrared object detection, namely DBD-YOLOv8 (DCN-BiRA-DyHeads-YOLOv8), is presented. The inherent limitations of the YOLOv8 model in scenarios with a low signal-to-noise ratio and complex tasks are addressed, with a focus on improving the multi-scale feature representation within the YOLOv8 framework and effectively filtering out irrelevant regions. To achieve this, two crucial modules, D_C2f and D_SPPF, are integrated. Deformable convolutions (DCN) are utilized by these modules to dynamically adjust the visual receptive fields of the network. Furthermore, a Bi-level Routing Attention mechanism (BRA) and Dynamic Heads (DyHeads) are adapted within the feature fusion network, refining feature maps and enhancing semantic representation through attention mechanisms. Significant improvements are demonstrated by DBD-YOLOv8 when compared to the YOLOv8-n\s\m\l\x series models. Notably, improved average mAP@0.5 values on benchmark datasets, including FLIR, OTCBVS (Dataset 01), OTCBVS (Dataset 03), and VEDAI, are achieved by DBD-YOLOv8. The corresponding values are 84.8%, 96.3%, 99.7%, and 76.0%, respectively. These results represent increases of 7.9%, 1.5%, 0.1%, and 3.5%, respectively. Importantly, real-time requirements are met by the model's inference times, which measure 10.9ms, 32.0ms, 37.3ms, and 28.4ms accordingly for the previous datasets.

INDEX TERMS Infrared object detection, deformable convolution, Bi-level routing attention, dynamic head.

I. INTRODUCTION

Infrared radiation refers to electromagnetic waves emitted by objects at temperatures above absolute zero, with wavelengths ranging from 750 nm to 1 mm. Infrared imaging offers unique advantages such as all-weather capability, good concealment, smoke penetration, and blind-spot detection when compared to radar and visible light imaging [1]. Infrared imaging serves as an effective complement or alternative to active radar imaging and visible light imaging, finding applications in various military and civilian fields [2], [3].

Infrared object detection plays a crucial role in infrared search and tracking (IRST) applications, involving the detection, recognition, and labeling of object categories in infrared images. Precisely and efficiently locating objects in infrared images is essential for object detection, which serves as a foundation for subsequent tasks like image recognition, object segmentation, and object tracking.

Infrared images possess unique characteristics, including low signal-to-noise ratio and limited texture and detail information. These characteristics arise from factors affecting infrared detectors, such as atmospheric scattering and refraction, optical defocus, background noise, clutter interference, and detector noise. Furthermore, the demand for real-time object detection in applications like autonomous driving and guided tracking adds complexity to infrared object detection. Accurately detecting and recognizing various categories of objects, including multi-scale, occluded, and small objects (e.g., those with a contrast ratio lower than 15%, signal-to-noise ratio below 1.5, object size less than 0.15% of the entire image, or fewer than 80 pixels) with limited shape and texture

The associate editor coordinating the review of this manuscript and approving it for publication was Shuo Sun.

information, remains a prominent research area in infrared object detection [4].

Infrared object detection methods can be categorized into four types: filter-based, local information-based, data structure-based, and deep learning-based [5]. Deep learning-based methods, leveraging their effective feature representation capabilities, have become the predominant trend in infrared image object detection research [6]. Existing deep learning-based object detection methods for infrared images can be broadly categorized into two main approaches based on different detection strategies: those utilizing convolutional neural networks (CNNs) and those employing Transformers [7]. Additionally, there are methods that combine elements of both approaches.

Within the CNN-based methods, there are two distinct categories: region proposal-based detection methods (two-stage network algorithms) and regression-based object detection methods (one-stage detection algorithms).

Region proposal-based methods generate candidate box images containing potential object positions, which are then fed into the detection network for classification and localization. Combining shallow and deep feature maps and utilizing multi-scale features enhances the detection accuracy. Representative algorithms in this category include R-CNN [8], Faster R-CNN [9], and HyperNet [10]. Regression-based methods directly determine the object category and location in the network architecture, simplifying the algorithms' implementation and improving computational speed. Representative algorithms in this category include SSD [11] and the YOLO series [12], [13], [14].

In recent years, Transformer-based object detection methods have gained substantial traction due to the remarkable performance of self-attention mechanisms in capturing long-range information and enabling global modeling of object features. This has led to an increased exploration of the Transformer model's potential application in computer vision, including object detection [7]. Prominent examples of this exploration include MMViT [15] and RIFormer [16], which have leveraged Transformer-based architectures to facilitate the development of end-to-end approaches that strike a balance between accuracy and computational cost. These models have achieved significant breakthroughs in object detection by introducing improvements to the backbone network [15], [16] and novel detection heads [17], [18]. Notably, DETR (DEtection TRansformer) [18] has emerged as a pioneering visual model that applied the Transformer to object detection, paving the way for further research and advancements in the field.

The YOLO series has gained popularity in object detection research due to its excellent performance in terms of generalization, real-time processing, lightweight design, and scalability. Efforts have focused on combining methods such as global perception and multi-scale feature fusion to enhance feature extraction and classification capabilities, addressing the challenge of semantic feature extraction versus object scale to improve algorithm interpretability [19].

In this study, we propose enhancements to the YOLOv8 algorithm as the basis for our model. In the backbone network, we improve the extraction of multi-scale features for irregular and occluded objects by integrating a Deformable Convolution Network (DCN v2) [20], [21], [22]. To enhance the detection accuracy of occluded objects and small objects, we have replaced the convolutional units in CBS with DCNv2, thereby improving the second and third layers, C2f, of the backbone network, which correspond to the feature maps of layers P3 and P4. Additionally, we have improved the convolutional layers in SPPF by incorporating DCNv2. The introduction of DCNv2 serves the purpose of adaptively adjusting the receptive field to accommodate target deformations and thereby ensuring an enhanced detection accuracy for occluded objects and small objects. Additionally, we introduce a Bi-level Routing Attention (BRA) mechanism [23] that selectively filters out irrelevant regions in the feature map while retaining highly relevant regions. This approach enhances the model's ability to learn multi-scale features and improves computational efficiency. In the feature fusion network, we utilize DyHeads, which integrates attention mechanisms to enhance feature semantics. DyHeads employs a unified approach that incorporates scale awareness, spatial awareness, and task awareness, leading to improved focus and generalization performance of the model [24].

The DBD-YOLOv8 model demonstrates a significant improvement in the detection accuracy of infrared targets, particularly in scenarios with multiple objects, including occluded and small targets. Importantly, these improvements in detection accuracy are achieved without a significant increase in model inference time, ensuring the model remains suitable for real-time detection requirements.

II. RELATED WORK

In the realm of infrared object detection, the fundamental aspect of object detection algorithms revolves around three key components: feature extraction, feature fusion, and the integration of various feature processing techniques. These elements work in unison to enhance the capabilities of feature extraction and classification.

A. FEATURE EXTRACTION

In scenarios involving non-cooperative objects or sensitive applications with limited datasets, the common approach is to enhance algorithm generalization to address the problem of model overfitting caused by small sample sizes. The design of the backbone network aims to effectively extract and facilitate the subsequent fusion of multi-scale features. Network architectures like CSPDarkNet and Transformer have successfully improved detection performance. However, excessively complex structures can lead to increased model complexity.

DETR architecture consists of a backbone network, a Transformer encoder, and a Transformer decoder. The backbone network, commonly based on CNN [25], [26] or Transformer [15], [16] models, extracts image features. The

Transformer encoder integrates multi-scale information [27] and re-encodes image features to enhance relevant information while suppressing irrelevant details. The Transformer decoder utilizes query vectors to aggregate features related to the objects within the image [28], extracting relevant object information for detection. However, it is important to note that the Transformer encoder in DETR has certain limitations. From an encoding perspective, the Transformer encoder redundantly re-encodes highly encoded image features, resulting in functional repetition within the network. From a feature fusion standpoint, although the Transformer encoder successfully fuses multi-scale features, its multi-layered hierarchical structure and large parameter size significantly increase the complexity of network optimization. Consequently, the model convergence speed is slowed down, necessitating substantial computational resources. Thus, the effectiveness of the Transformer encoder does not align with its significant computational cost. To overcome these limitations, alternative approaches are being explored to improve the efficiency and effectiveness of feature encoding and fusion in the context of the Transformer model [29], [30].

Despite the exceptional detection accuracy of Transformer-based methods, they come with high computational costs and inadequate convergence speed. Additionally, the unique characteristics of infrared and remote sensing images, such as small object scales and diverse categories, limit the advantages of Transformer in the context of infrared object detection.

The YOLOv8 backbone network is based on the CSPDarkNet-53 network structure. To mitigate the issue of losing low-level details due to the overshadowing of shallow detail features by deep semantic features in complex background feature fusion, contextual aggregation is introduced between the backbone network and the feature extraction network. This integration enhances feature fusion and spatial interaction capabilities [31]. When detecting small objects, a combination of shallow position information and deep semantic information is utilized to strengthen feature fusion capabilities [32]. For the detection of multi-scale, dense, and occluded objects, several techniques have been proposed. These techniques include iterative feature extraction using the backbone network [33], fusion with attention mechanisms [34], optimization of attention feature maps using Long Short-Term Memory (LSTM) to enhance feature extraction capabilities [35], and the utilization of multi-scale context and enhanced channel attention to improve the representation of object features in complex backgrounds [36].

To avoid the loss of fine-grained details caused by deep convolution iterations, dilated convolutions can be employed to maintain higher resolution and larger receptive fields [37]. Deformable convolutions can be used to adaptively adjust the network's visual receptive fields [38].

B. FEATURE FUSION

Research on feature fusion primarily focuses on efficiently integrating multiple features and combining methods for multi-feature fusion to enhance feature extraction and classification capabilities. Attention mechanisms dynamically learn feature weights or attention distribution based on context and task requirements, enabling adaptive focus on key features.

There are various attention mechanisms and feature fusion methods, including spatial attention mechanisms, channel attention mechanisms, channel-spatial mixed attention mechanisms, and self-attention mechanisms. Among these, the channel-spatial mixed attention model combines the advantages of channel attention and spatial attention. It adaptively refines and maps important channels and spatial regions for feature extraction, attention weighting, and fusion, resulting in more reliable attention information and comprehensive feature representation.

While Convolutional Block Attention Module (CBAM) [39] and Bottleneck Attention (BAM) [40] utilize channel and spatial attention mechanisms to capture feature representations across different dimensions, enhancing network expressiveness and robustness, they suffer from high computational complexity, excessive focus on local features, and limitations in model generalization. On the other hand, Shuffle Attention Net (SA-Net) leverages the idea of feature separation and interaction to accurately comprehend relationships between image features. However, it needs to address the issue of information loss during information exchange and recombination. Efficient Pyramid Split Attention (ESPA) employs a spatial pyramid attention mechanism to handle multi-scale visual information and improve feature representation. However, it comes with high computational complexity and significant training and tuning costs.

Multi-Head Self-Attention (MHSA) improves model performance by increasing the number of attention heads [41], but this introduces scalability issues, particularly in terms of computational and spatial complexity.

In addition, feature pyramid networks are used to extract multi-scale features of different objects [42], and loss functions are optimized [43]. The anchor-free mechanism is employed to address the difficulties of imbalanced positive and negative samples and hyperparameter tuning [44]. Furthermore, improvements have been made to enhance position regression accuracy [45] and computational speed [46] in the Non-Maximum Suppression (NMS) algorithm.

III. IMPROVEMENT FOR DBD-YOLOv8 MODEL

Compared to previous versions of YOLO, YOLOv8 demonstrates significant improvements in both accuracy and speed. YOLOv8 provides different-sized models, namely n/s/m/l/x scales, based on different scaling factors to accommodate various scene requirements. The depth and width factors of the models vary, where larger models yield better detection performance but slower inference speed. The YOLOv8 network architecture consists of three main parts: the Backbone layer, Neck layer, and Head layer. This design allows each branch of the backbone network, neck, and head to focus on

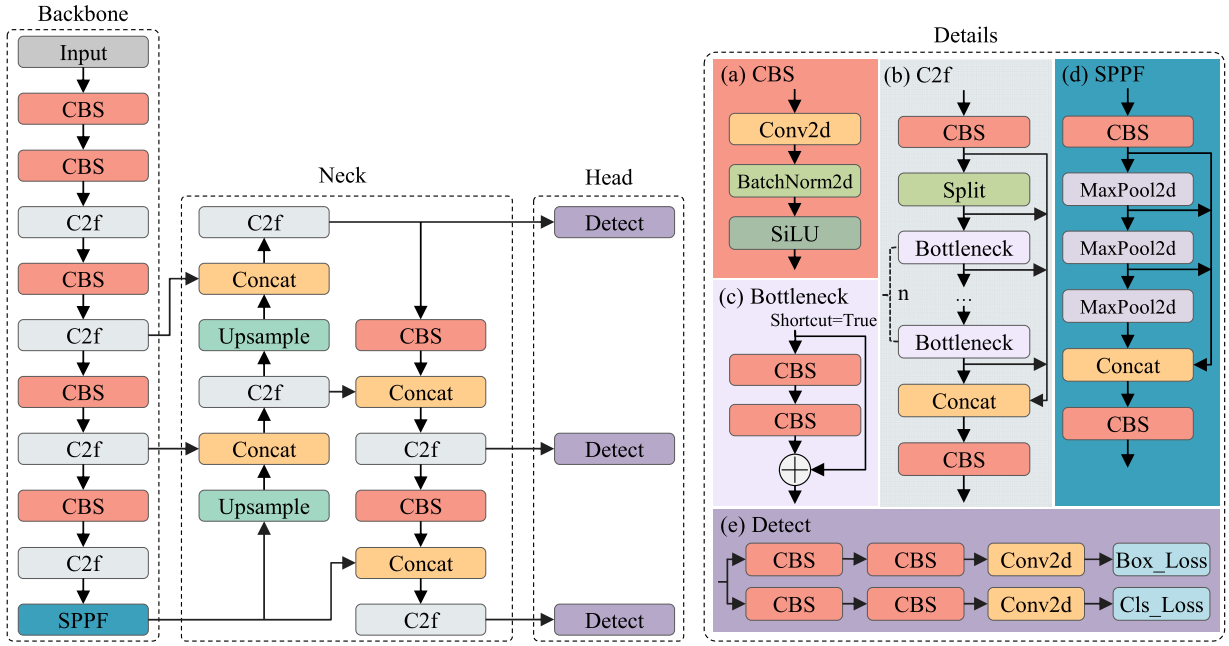


FIGURE 1. The network architecture diagram of YOLOv8 and the schematic illustration of its main constituent module details.

its respective task, thereby improving the overall accuracy of the model.

The YOLOv8 Backbone network comprises three modules: CBS, C2f, and SPPF. The CBS module consists of convolution (Conv), batch normalization (BN), and Sigmoid Linear Unit (SiLu) activation function. The C2f module adopts the Cross Stage Partial (CSP) structure, where the output of the first CBS module is split into two channel branches. One branch concatenates n residual blocks of BottleNeck, while the other branch performs additive concatenation with the first branch, resulting in feature extraction.

The Spatial Pyramid Pooling Fast (SPPF) employs three pooling layers with the same kernel size (5×5) to reduce computation while enhancing feature extraction efficiency. Through the fusion of local and global features, the SPPF enhances the network's receptive field and achieves multi-scale feature fusion through concatenation, contributing to improved object detection.

The Backbone is primarily used for feature extraction. YOLOv8 enhances the capability of feature representation through the lightweight c2f module. The channel numbers are altered through splitting and concatenation operations based on the scaling factors, reducing computational complexity and model capacity. The SPPF layer is added at the end to increase the receptive field and capture features at different levels in the scene.

The Neck is mainly responsible for feature fusion and adopts the Dual-Stream Feature Pyramid Network (FPN) structure, combining the Feature Pyramid Network (FPN) and Path Aggregation Network (PAN). It fuses the semantic and spatial information of multi-scale feature maps. As the image passes through deep networks, the receptive field and semantic features strengthen, but the resolution decreases, making it challenging to capture features of small objects. FPN utilizes a top-down (upsampling) approach to transmit deep semantic features to shallow layers and performs tensor concatenation with the same-sized feature maps from the backbone network, enhancing the detection performance of small objects. PAN utilizes a bottom-up (downsampling) approach to transmit shallow-level color, edge, contour, and spatial features to deep layers while performing tensor concatenation with the feature maps from FPN. This achieves comprehensive fusion of multi-scale features, compensating for feature loss caused by deepening the network and enriching feature granularity information.

The Head utilizes the Decoupled Head to separate localization and classification into two branches. It predicts objects of small, medium, and large scales based on the fused P3, P4, and P5 feature maps, respectively. YOLOv8 employs an Anchor-Free framework, directly predicting bounding boxes within the grid, which avoids the need for initializing anchor box sizes and the computational overhead of NMS, thereby improving object localization efficiency.

The network structure of YOLOv8 and its module details are depicted in Figure 1.

The objective of this study is to propose enhancements to the YOLOv8 algorithm to improve the detection accuracy of occluded and small objects in infrared target scenarios without compromising real-time detection requirements.

The proposed enhancements include integrating a Deformable Convolution Network (DCNv2) in the backbone network to extract multi-scale features for irregular and occluded objects. The convolutional units in CBS and SPPF are replaced with DCNv2 to improve the detection accuracy

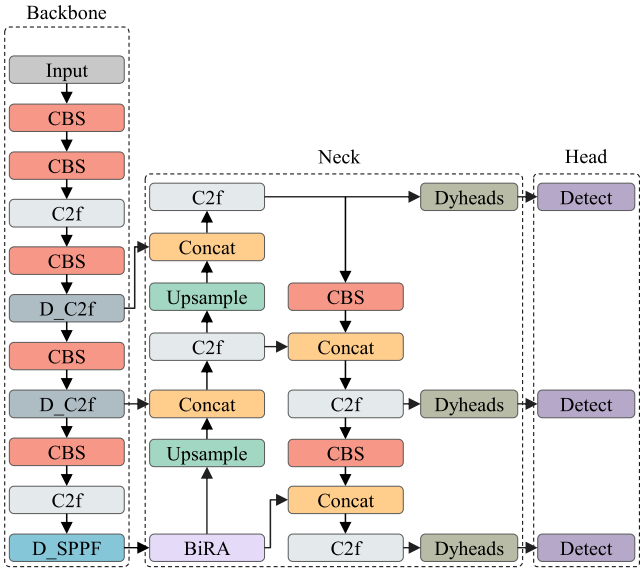
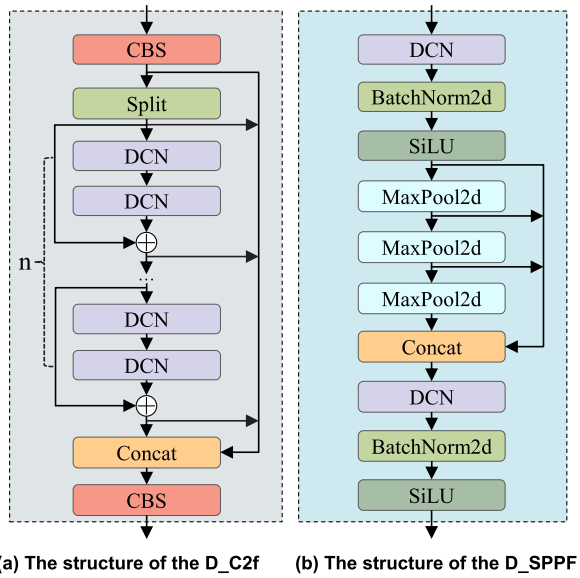


FIGURE 2. The network framework scheme of the DBD-YOLOv8 model.



(a) The structure of the D_C2f (b) The structure of the D_SPPF

FIGURE 3. The structure of the D_C2f and D_SPPF module.

of occluded and small objects. A BRA mechanism is introduced to filter out irrelevant regions and enhance the model's ability to learn multi-scale features. DyHeads is utilized in the feature fusion network to enhance feature semantics. The proposed enhancements effectively improve the detection accuracy of occluded and small objects in infrared target scenarios, making the DBD-YOLOv8 model suitable for real-time detection requirements.

The enhanced network architecture of the DBD-YOLOv8 model is illustrated in Figure 2, showcasing its refined structure.

A. D_C2f AND D_SPPF MODULE

Due to the diverse nature of infrared objects, which exhibit complexity in terms of size, shape, position, and orientation, standard convolution operations often struggle to accurately capture the precise location of objects or only capture partial information. This limitation is particularly evident in detecting occluded and small objects, which are susceptible to noise and interference.

In the backbone network of YOLOv8, the visual layer features encompass rich spatial information but may lack semantic information, which limits their effectiveness. Additionally, fixed standard convolution kernels lack flexibility, resulting in limited receptive fields. Consequently, the network is prone to missed detections or false detections when encountering multi-scale objects, occluded objects, or small objects. To overcome this issue, the introduction of deformable convolutional networks in the visual layer can enhance feature representation [47].

By introducing DCN v2 in the C2f and SPPF network structures, we have reconfigured the YOLOv8 algorithm to improve its detection capability. While the DCN itself does not significantly increase the number of parameters and FLOPs in the model, the stacking of multiple DCN layers can lead to increased inference time in practical applications.

Experimental results demonstrate that using DCNs generally enhances network performance. However, it is important to note that excessive use of DCN layers does not necessarily improve network performance; instead, it may reduce network speed and increase parameter tuning costs [48].

To address these challenges, we conducted experiments to evaluate detection accuracy and inference speed. Through this process, we found that replacing the second and third C2f modules in the backbone network with D_C2f modules, as well as replacing SPPF with D_SPPF, yields the optimal optimization results. Specifically, we employed deformable convolutions in the detection layers P3, P4, and SPPF. This modification enables the network to adaptively adjust its receptive field during the model sampling process, aligning better with the shapes and sizes of objects and enhancing its robustness. Furthermore, during the model's prediction and regression process, the algorithm can adjust the regression parameters of predicted bounding boxes, further improving the model's ability to express features related to multi-scale objects, occlusions, and small objects. Figure 3 provides a comprehensive visualization of the detailed module structures of D_C2f and D_SPPF.

During regression prediction, the D_C2f module adjusts the regression parameters of bounding boxes, thereby enhancing the model's feature representation capability for objects of different scales, occluded objects, and small objects. This improvement results in increased completeness of information for the objects under test. The D_C2f module achieves this enhancement by stacking multiple deformable convolution modules, which not only extract diversified and multi-scale feature information from objects but also expand the network's receptive field while reducing the number of parameters.

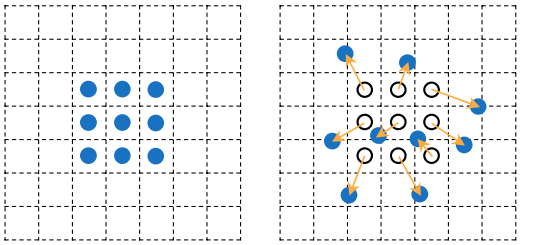


FIGURE 4. Sampling comparison between standard convolution and deformable convolution.

In the feature fusion stage, the D_SPPF module enhances the network model's geometric modeling capability, allowing it to focus on object areas and adapt to object deformations and scale changes. This capability enables the network to capture features at different levels in the scene, thereby increasing the receptive field.

The purpose of employing the DCN module is to enhance the model's ability to extract invariant features. DCN achieves this by learning an offset for each sampling point of the convolution kernel, enabling it to adapt to the geometric shape of the object. By learning different optimal convolution kernel structures based on diverse object data, DCN strengthens the feature extraction capability for infrared objects at various scales. For a visual comparison between standard convolution and deformable convolution sampling, refer to Figure 4.

Deformable convolution is an extension of standard convolution that introduces offset values to the sampling points. Let x represent the input feature map and y denote the output feature map. N and n represent the total number of sampling points and the enumeration of these points, respectively. After sampling the input feature map x, the center sampling point is denoted as p_0, and p_n represents the offset of p_0 within the convolution kernel range. This offset is a constant value that assists the model in learning biases, thereby enhancing accuracy and stability.

In deformable convolution, each position p_0 introduces an offset value $\{\Delta p_n | n = 1, 2, \cdots, N\}$. The feature matrix outputted by DCN v1 [20] is as follows:

$$y(p_0) = \sum_{p_0 \in R} w(p_n) \cdot x(p_0 + p_n + \Delta p_n) \tag{1}$$

In this context, $w(p_n)$ represents the weight assigned to the position p_n, while $x(p_n)$ denotes the pixel value of the input feature map at position p_n. The variable Δp_n represents the offset at position p_n. Since Δp_n is typically a decimal value, $x(p_0 + p_n + \Delta p_n)$ may not correspond to an existing point on the feature map, making it unsuitable for direct sampling. Therefore, DCN v1 utilizes bilinear interpolation to achieve the desired offset effect.

While DCN v1 effectively fits small objects by introducing the offset module, it also introduces excessive and irrelevant context information. This additional context may include irrelevant background information, which can interfere with feature extraction by the model. Building upon DCN v1, DCN v2 not only learns the offset of sampling points but also the modulation of each sampling point. It introduces a modulation scalar Δm_n to suppress irrelevant background information, enabling the model to reduce the interference from irrelevant regions during feature learning. The modulation scalar $\Delta m_n \in [0, 1]$ differentiates whether the introduced region is the region of interest. If a sampling point's region is not of interest, the weight can be learned as 0. The feature matrix outputted by DCN v2 [21] is as follows:

$$y(p_0) = \sum_{p_n \in R} w(p_n) \cdot x(p_0 + p_n + \Delta p_n) \cdot \Delta m_n \tag{2}$$

In DCN v1, the introduced offset aims to identify the region where relevant information is located. In DCN v2, the introduced modulation is used to assign weights to the identified positions. These two aspects ensure the accurate extraction of valuable information.

B. Bi-LEVEL ROUTING ATTENTION MECHANISM

The attention mechanism in this study aims to emulate the selective perception mechanism observed in the human visual system. To leverage the advantages of both channel attention and spatial attention, we propose the Channel and Spatial Mixed Attention (CSMA) mechanism. This approach adaptively selects critical channels and spatial regions, combines the channel and spatial attention weights, and generates a mixed feature vector. By considering the interactive relationship between input data in channel and spatial dimensions, our mechanism provides more comprehensive and reliable attention information.

To further enhance the model's performance, we introduce the Multi-Head Self-Attention (MHSA) technique, which increases the number of attention heads. However, this improvement introduces scalability concerns, particularly in terms of computational and spatial complexity. To address these issues, we employ the Bi-Level Routing Attention (BRA) approach. BRA facilitates the extraction of attention weights between feature maps of different scales, resulting in a hierarchical channel attention vector. Subsequently, these vectors are recalibrated and used to weight the corresponding feature maps. The output is a multi-scale feature map that contains richer and more detailed feature information, as illustrated in Figure 5.

First, the input feature map $X \in \mathbb{R}^{H \times W \times C}$ is partitioned into non-overlapping regions of size $s \times s$, which are then mapped to region-level feature vectors $X^r \in \mathbb{R}^{S^2 \times \frac{HW}{S^2} \times C}$. The query, key, and value tensor $Q, K, V \in \mathbb{R}^{S^2 \times \frac{HW}{S^2} \times C}$ are obtained through linear projections [23].

$$Q = X^r W^q, \ K = X^r W^k, \ V = X^r W^v \tag{3}$$

Here, the weights $W^q, \ W^k, \ W^v \in \mathbb{R}^{C \times C}$ correspond to the projection of the query, key, and value, respectively.

Next, an adjacency matrix is constructed using a directed graph to determine the relationships between different keys

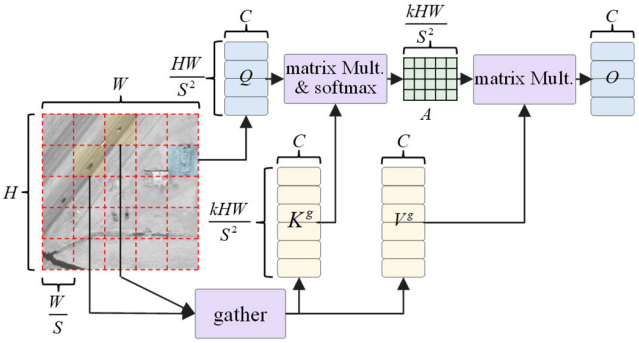


FIGURE 5. BRA Module process mechanism diagram.

and values. By performing region-level averaging on Q and K, Q^r, $K^r \in \mathbb{R}^{S^2 \times C}$ is obtained. The adjacency matrix of the region-to-region association graph is calculated by multiplying Q^r with the transpose of K^r.

$$A^r = Q^r (K^r)^T \in \mathbb{R}^{S^2 \times S^2} \tag{4}$$

To identify the most relevant regions, a top-k operation is performed row-wise, resulting in a routing index matrix.

$$I^r = topindex(A^r) \in \mathbb{R}^{S^2 \times k} \tag{5}$$

Consequently, the i-th row of I^r contains the indices of the k most relevant regions to the i-th region.

Finally, the key tensor and value tensor of the region index matrix are collected, as shown in Equation 6 [23].

$$\begin{cases} K^g = gather(K,\ I^r) \in \mathbb{R}^{S^2 \times \frac{kHW}{S^2} \times C} \\ V^g = gather(V,\ I^r) \in \mathbb{R}^{S^2 \times \frac{kHW}{S^2} \times C} \end{cases} \tag{6}$$

This collection facilitates the generation of fine-grained Token-to-Token attention.

$$O = Attention(Q,\ K^g,\ V^g) + LCE(V) \tag{7}$$

To enhance the local context, we employ the Local Context Enhancement (LCE) function [49], which utilizes deep convolution with a kernel size of 5.

The detailed structure of the BiRA module is illustrated in Figure 6.

C. DYHEADS MODULE

The YOLOv8 model has only three detection heads, which may lead to missed or false detections when detecting occluded or small objects. Since the output of the YOLOv8 backbone network is a three-dimensional tensor with dimensions of width × height × channels, we propose to incorporate a dynamic detection head called Dynamic Head (DyHeads). This additional head aims to unify scale-aware attention, spatial-aware attention, and task-aware attention, enabling the integration of attention mechanisms on specific dimensions of the feature tensor. By doing so, we can effectively enhance the model's performance in detecting occluded and small objects.

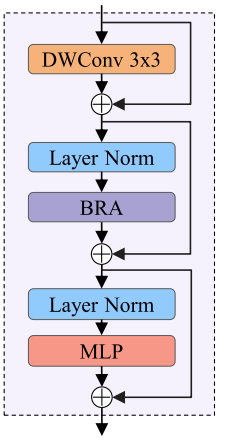


FIGURE 6. The structure of the BiRA module.

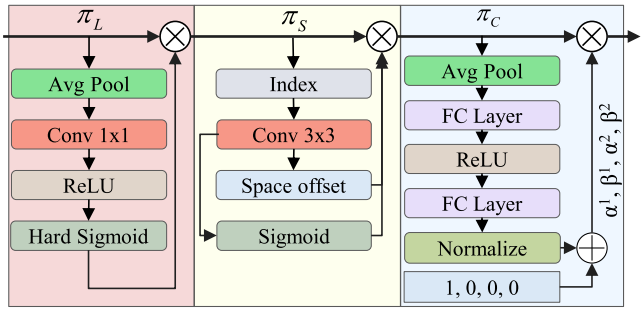


FIGURE 7. The dynamic head module.

Object detection heads aim to enhance the classification and localization of objects. To further improve the handling of high-resolution multi-scale semantic information generated by YOLOv8, we employ the Dynamic Head. The rescaled feature pyramid, which can be represented as a four-dimensional tensor, is denoted as $\mathcal{F} \in \mathbb{R}^{L \times H \times W \times C}$, where L represents the number of levels in the pyramid, and H, W, and C represent the height, width, and number of channels of intermediate-level features, respectively. The structure of the dynamic head is illustrated in Figure 7.

Furthermore, we define $S = H \times W$ and reshape the tensor into a three-dimensional tensor, $\mathcal{F} \in \mathbb{R}^{L \times S \times C}$. The dynamic head unifies object detection head methods from three dimensions: scale-awareness (Level-wise), spatial-awareness (Spatial-wise), and task-awareness (Channel-wise), by utilizing an attention mechanism.

Despite its name, the dynamic head plays a role similar to a neck, strengthening the semantic features. To achieve this, DyHeads employs a separated attention mechanism that sequentially applies three attention mechanisms, with each one focusing on a specific dimension [24]:

$$W(\mathcal{F}) = \pi_C(\pi_S(\pi_L(\mathcal{F}) \cdot \mathcal{F}) \cdot \mathcal{F}) \cdot \mathcal{F} \tag{8}$$

Among them, $\pi_L(\cdot)$, $\pi_S(\cdot)$, and $\pi_C(\cdot)$ are three separate attention functions applied to dimensions L, S, and C, respectively.

To dynamically fuse features from different scales based on their semantic importance, we introduce the scale-awareness attention mechanism, denoted as attention function $\pi_L(\cdot)$.

$$\pi_L(\mathcal{F}) \cdot \mathcal{F} = \sigma\left(f\left(\frac{1}{SC}\sum_{S,C}\mathcal{F}\right)\right) \cdot \mathcal{F} \tag{9}$$

where $\sigma(x) = max(0, min(1, \frac{x+1}{2}))$ represents the hard-sigmoid function, while $f(\cdot)$ corresponds to the linear transformation achieved through a 1×1 convolution.

The spatial-awareness attention mechanism initially incorporates sparsity in attention learning through the use of DCN v2. Subsequently, it aggregates features from different levels at the same spatial position.

$$\pi_S(\mathcal{F}) \cdot \mathcal{F} = \frac{1}{L}\sum_{l=1}^{L}\sum_{k=1}^{K} w_{l,k} \cdot \mathcal{F}(l; p_k + \Delta p_k; c) \cdot \Delta m_k \tag{10}$$

The task-awareness attention mechanism dynamically enables and disables feature channels to accommodate different tasks.

$$\pi_C(\mathcal{F}) \cdot \mathcal{F} = max(\alpha^1(\mathcal{F}) \cdot \mathcal{F}_c + \beta^1(\mathcal{F}), \alpha^2(\mathcal{F}) \cdot \mathcal{F}_c + \beta^2(\mathcal{F})) \tag{11}$$

where $\theta(\cdot) = [\alpha^1, \alpha^2, \beta^1, \beta^2]^T$ represents a hyperfunction utilized for learning the control activation threshold, while $\mathcal{F}_c$ denotes the feature slice corresponding to the c-th channel.

These three types of attention can be sequentially applied to detection heads to enhance the algorithm's ability to detect occlusions and small objects. As the number of stacking increases, the dynamic head achieves higher precision.

Inspired by the dynamic head, a DyHeads was designed to enhance the effectiveness of object detection by combining the infrared object detection characteristics of YOLOv8. Since $\pi_S(\cdot)$ bears resemblance to deformable convolution modules, spatial-awareness has already been improved in the backbone network enhancements through the use of deformable convolutions. To strike a balance between accuracy, complexity, and inference speed, while reducing model parameters and computational requirements, only scale-aware and task-aware attention mechanisms are employed to enhance the perception ability of the detection head. In our algorithm, we integrated four dynamic head blocks (DyHeads) with m=4. The schematic diagram of the DyHeads module is depicted in Figure 8.

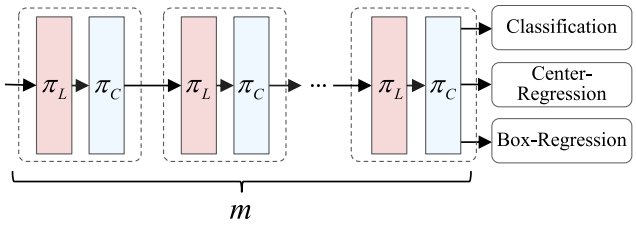


FIGURE 8. The structure of the DyHeads module.

IV. EXPERIMENTS AND RESULTS ANALYSIS

A. EXPERIMENTAL ENVIRONMENT

The experiments were conducted on a mobile service platform equipped with a GeForce RTX 4060 GPU and an Intel i9-13900HX CPU. The runtime environment utilized CUDA Toolkit 11.7, while the deep learning framework employed was PyTorch 1.13.1.

B. DATASETS

Experimental benchmarks for model training and testing were chosen from several infrared datasets, namely FLIR [50], OTCBVS (Dataset 01) [51], OTCBVS (Dataset 03) [51], and VEDAI [52].

The FLIR dataset comprises 14,452 labeled thermal infrared images captured in various scenarios, encompassing applications such as security surveillance, industrial inspection, drones, and autonomous driving. It consists of three classes: person, bicycle, and car. To conduct the experiments, a random sampling approach was employed, resulting in a training set of 8,862 images, a validation set of 976 images, and a test set of 390 images.

From the OTCBVS dataset, two subsets were chosen. Dataset 01 contains 284 infrared images with person objects, while Dataset 03 consists of 17,089 infrared images featuring both person and car objects. Each subset was divided into training, validation, and test sets using an 8:1:1 ratio.

The VEDAI dataset contains 1,250 pairs of co-registered visible and near-infrared aerial images with a resolution of 1024×1024. It encompasses nine object categories, including airplanes, ships, various vehicles, and bicycles. On average, each image contains 5.5 vehicle objects, which account for only 0.7% of the total image pixels. This dataset presents challenges such as small and multiple objects, variations in orientation, lighting conditions, shadows, specular reflections, and occlusions. For the experiments, a random sampling method was employed, resulting in a training set of 1,000 images, a validation set of 125 images, and a test set of 125 images.

C. EVALUATION METRICS

The model's performance was evaluated using precision (P), recall (R), and mean average precision (mAP) metrics.

1) PRECISION

Precision refers to the probability of correctly predicting positive samples among the predicted positive samples.

$$P = \frac{TP}{TP + FP} \tag{12}$$

where TP represents the number of true positive samples predicted as positive and FP represents the number of false positive samples predicted as positive.

TABLE 1. Model training hyperparameter settings.

Hyperparameter Options	Setting
Input Resolution	640x640
Initial Learning Rate 0 (lr0)	0.01
Learning Rate Float (lrf)	0.01
Batch_size	8
Epochs	200

2) RECALL

Recall represents the probability of correctly predicting positive samples among all the predicted samples.

$$R = \frac{TP}{TP + FN} \tag{13}$$

FN represents the number of false negative samples predicted as negative.

3) mAP

mAP (mean Average Precision) is the mean value of precision for all detection categories. It is calculated using the following formula:

$$mAP = \frac{\sum_0^N AP_n}{N} \tag{14}$$

where N represents the total number of classes. APn refers to the average precision of class n, which is calculated as the area under the Precision-Recall curve. The mean Average Precision (mAP) is denoted as mAP@0.5, representing the average accuracy value when the IoU parameter threshold is set to 0.5. Furthermore, mAP@0.5:0.95 is utilized to indicate a range of IoU parameter thresholds from 0.5 to 0.95, with a step size of 0.05.

D. HYPERPARAMETER SETTINGS

To optimize the model's performance during the training process, a learning rate decay method was employed. This method involved adjusting the model's parameters update speed using an initial learning rate (lr0). Additionally, a learning rate coefficient (lrf) was utilized to control the decay of the learning rate over the course of training. The final learning rate was determined by multiplying the initial learning rate with the coefficient. To ensure sufficient training steps, the iteration was set to 200. Throughout the training, the learning rate gradually decreased, promoting model stability, facilitating smooth convergence, and minimizing fluctuation to achieve the optimal solution. Table 1 illustrates the hyperparameter configurations employed during model training.

To assess the impact of target classes on detection outcomes, we conducted object detection and recognition tests on the YOLOv8-based improved model, DBD-YOLOv8n, using the FLIR validation dataset. Figure 9 presents the confusion matrix depicting the distribution of target classes in the DBD-YOLOv8n model. Additionally, Figure 10 displays the precision-recall (P-R) curve.

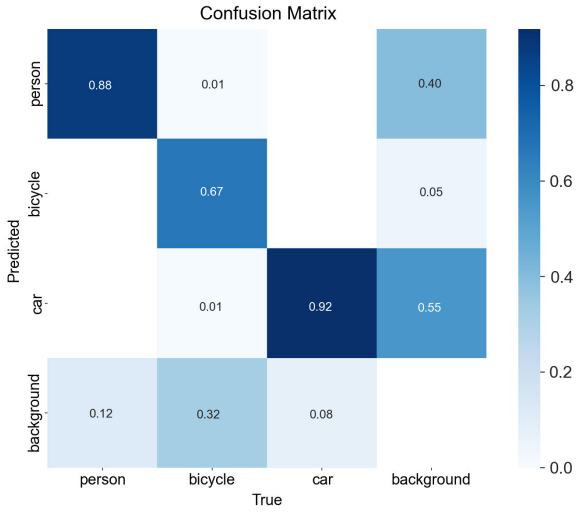


FIGURE 9. Confusion matrix of target class distribution (FLIR).

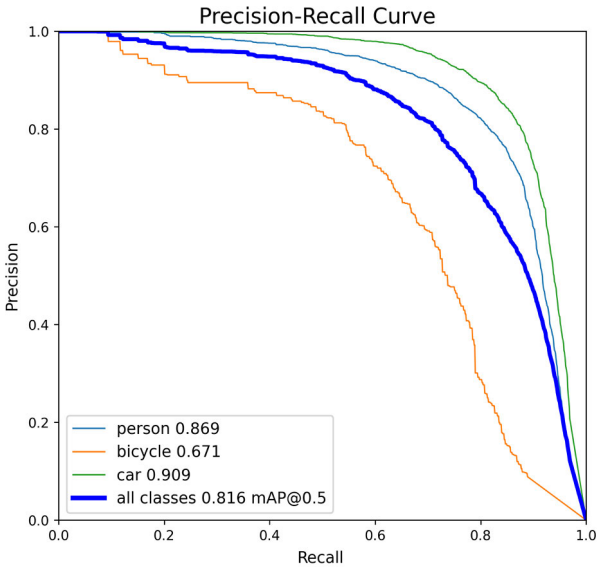
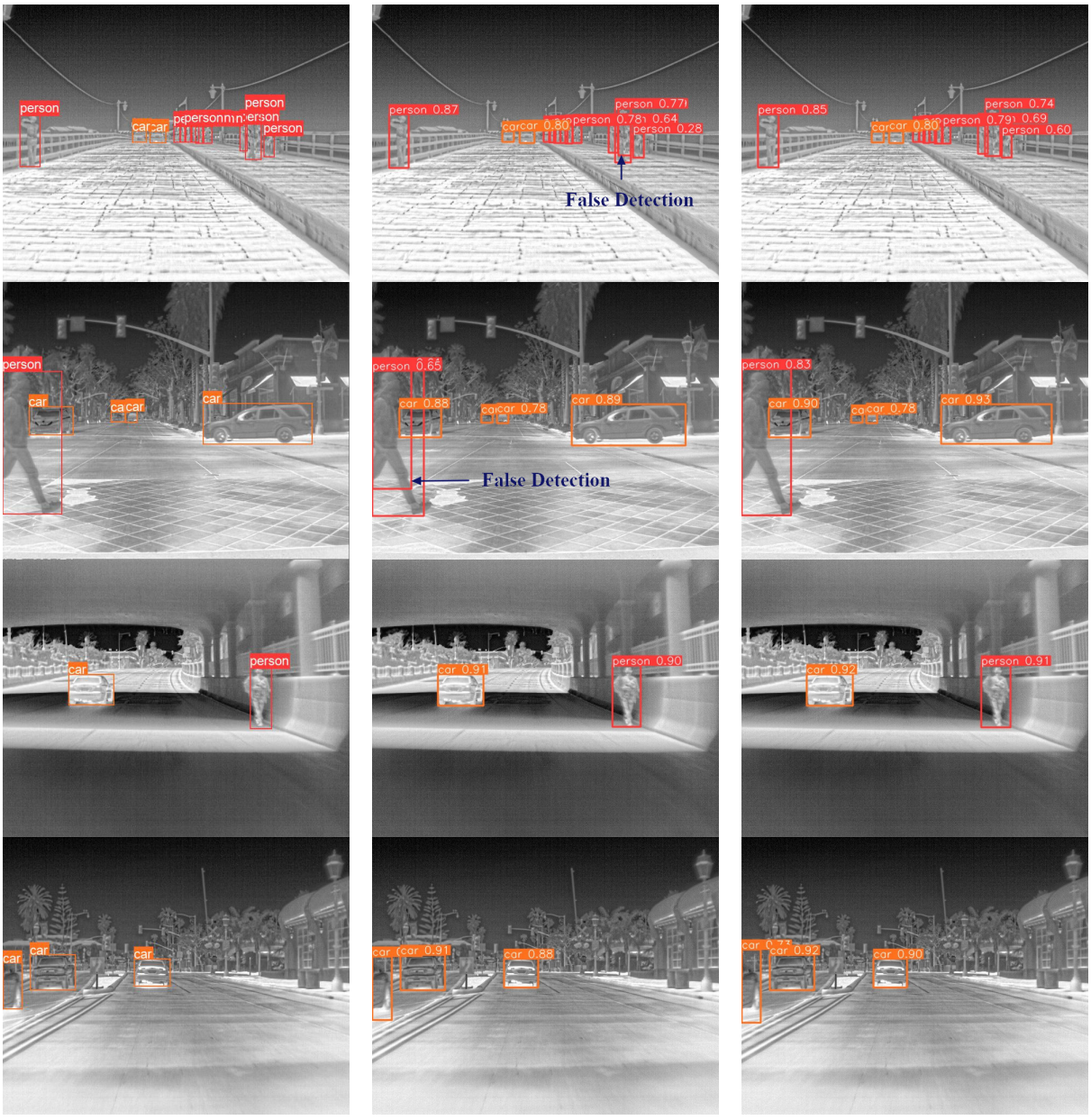


FIGURE 10. Precision-recall curve (FLIR).

E. ABLATION EXPERIMENTS

A series of ablative experiments were conducted on the FLIR, OTCBVS (Dataset 01), OTCBVS (Dataset 03), and VEDAI infrared datasets to evaluate the impact of the improved modules in the DBD-YOLOv8 model on the performance of infrared object detection. To compare the visual detection results between the original YOLOv8n model and the improved DBD-YOLOv8n model, the same set of images was used for object detection experiments.

Table 2 summarizes the ablative experiment results for FLIR, OTCBVS (Dataset 01 and 03), and VEDAI. Based on the results in Table 2, the YOLOv8n+DCN model shows improvements in precision (P), recall (R), and mean average precision mAP@0.5 compared to the YOLOv8n baseline model. Specifically, on the FLIR,



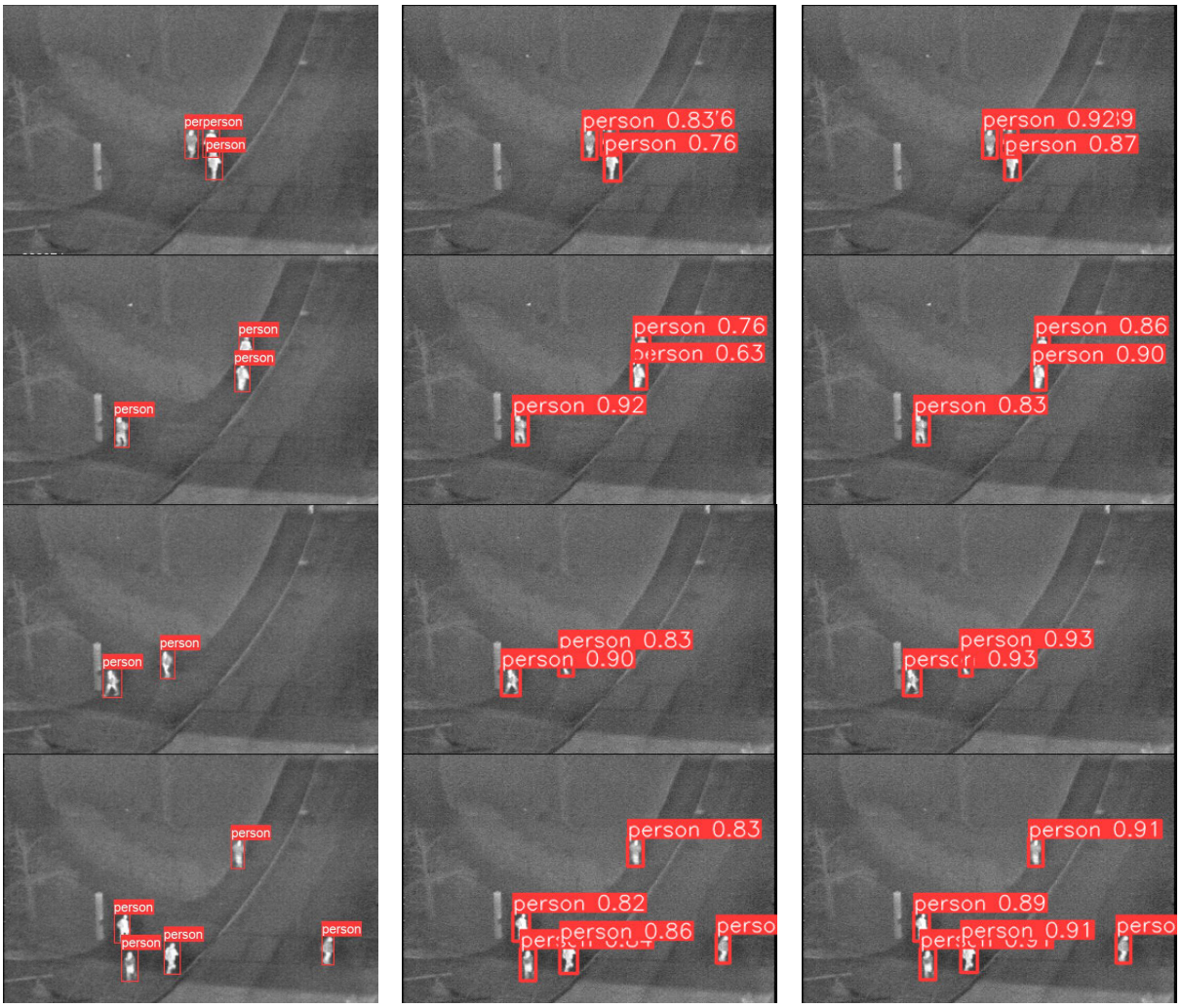
(a) Object label of input image. (b) Object detection result of YOLOv8n. (c) Object detection result of DBD-YOLOv8n.

FIGURE 11. Comparison of object detection results on FLIR test sets between YOLOv8n and DBD-YOLOv8n.

OTCBVS (Dataset 01), OTCBVS (Dataset 03), and VEDAI test sets, the YOLOv8n+DCN model achieves a P increase of 2.2%, 1.4%, 1.2%, and 1.7% respectively. The corresponding R increases are 1.5%, 0.7%, 0.6%, and 1.0%, while the mAP@0.5 increases are 4.8%, 0.6%, 0.3%, and 2.3%. These results suggest that integrating DCN into the backbone network allows for adaptive adjustment of the network's visual receptive field, better accommodating object shapes and sizes, reducing environmental noise effects, and effectively improving the representation capability of multi-scale, occluded, and small objects in infrared images, thereby enhancing detection accuracy.

The introduction of BiRA between the backbone network and the feature fusion network leads to significant improvements in P, R, and mAP@0.5 on the FLIR, OTCBVS (Dataset 01), OTCBVS (Dataset 03), and VEDAI test sets, despite a slight increase in parameter count. The mAP@0.5 reaches 73.5%, 91.3%, 96.8%, and 69.8%. These findings suggest that BiRA refines the feature maps by capturing multi-scale feature information and enhances the feature learning capability, thereby improving the overall detection accuracy of the network.

The addition of Dyheads to the feature fusion network in the YOLOv8n+Dyheads model further improves



(a) Object label of input image. (b) Object detection result of YOLOv8n. (c) Object detection result of DBD-YOLOv8n.

FIGURE 12. Comparison of object detection results on OTCBVS (Dataset 01) Test Sets between YOLOv8n and DBD-YOLOv8n.

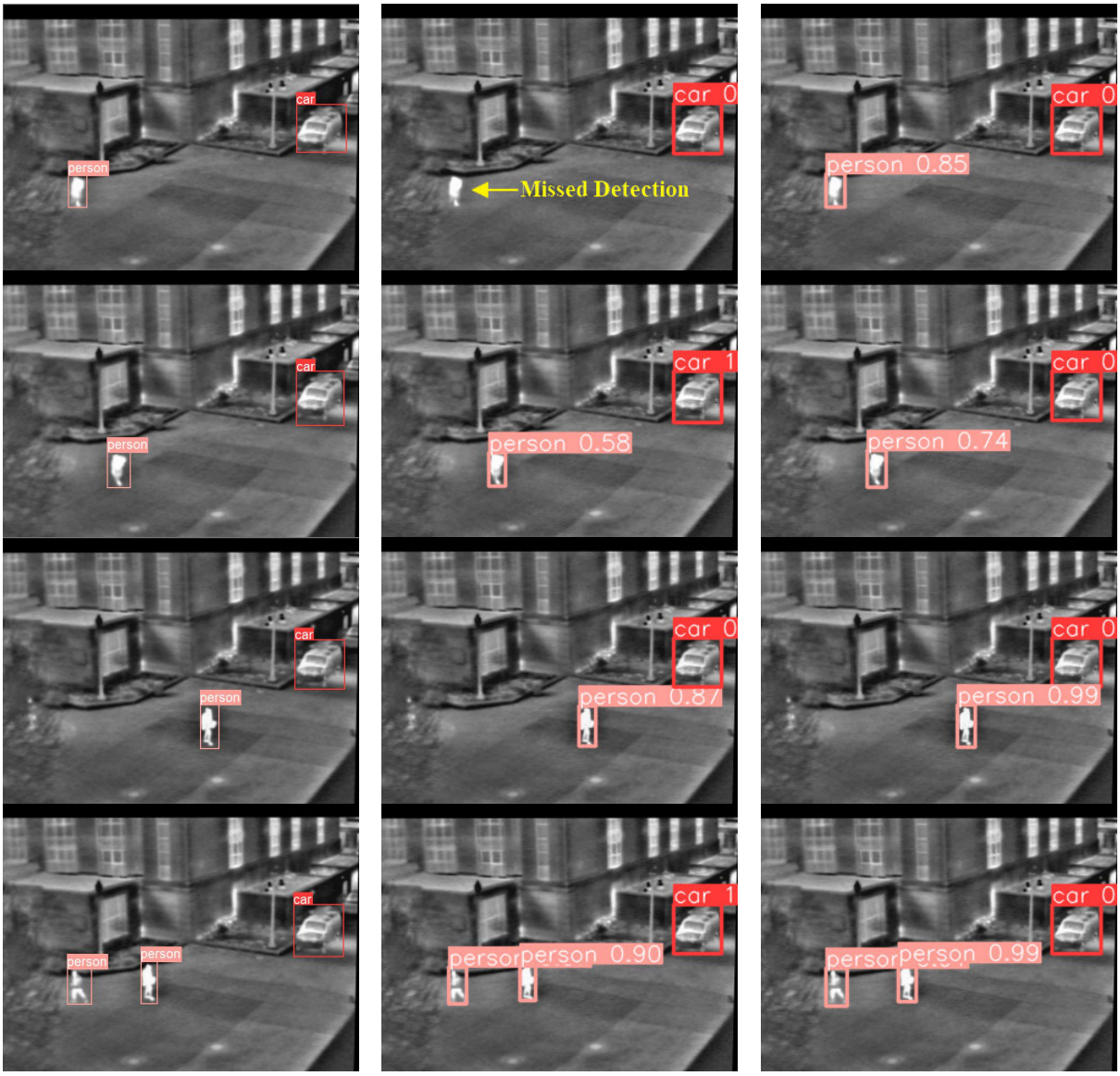
TABLE 2. Results of ablation experiment.

Model [1]				P (%) [2]				R (%) [2]				mAP@0.5 (%) [2]				Inference time (ms) [2]			
B	DCN	Bi	Dy	D1	D2	D3	D4	D1	D2	D3	D4	D1	D2	D3	D4	D1	D2	D3	D4
√				76.8	92.3	96.6	73.5	70.4	90.0	95.5	67.5	72.2	90.5	96.4	68.1	5.8	16.5	18.7	10.6
√	√			78.5	93.6	97.8	74.8	71.5	90.6	96.1	68.2	75.7	91.0	96.7	69.7	5.7	15.3	16.6	9.4
√		√		78.6	94.7	98.1	74.9	72.4	91.0	96.7	68.5	73.5	91.3	96.8	69.8	5.8	16.5	18.7	10.6
√			√	79.9	93.8	98.3	74.6	73.6	91.1	96.1	69.3	74.0	91.2	96.2	70.0	5.6	14.8	17.8	9.2
√	√	√		80.5	95.4	98.6	75.2	74.1	91.1	96.4	70.2	76.3	91.5	97.0	71.1	5.7	14.6	16.9	9.0
√	√	√	√	81.8	96.7	98.8	76.5	80.0	92.5	96.5	72.8	81.6	94.0	99.6	72.8	5.7	15.9	18.1	10.3

[1] B: Baseline (Yolov8n), DCN: D_C2f + D_SPPF, Bi: BiRA, Dy: DyHeads.
[2] D1: FLIR, D2: OTCBVS (Dataset 01), D3: OTCBVS (Dataset 03), D4: VEDAI.

the mAP@0.5 on the FLIR, OTCBVS (Dataset 01), OTCBVS (Dataset 03), and VEDAI test sets, reaching74.0%, 91.2%, 96.2%, and 70.0% respectively. This indicates that Dyheads, combined with attention mechanisms, strengthen the semantic representation of features, leading to improved training effectiveness and detection performance of the model.

Finally, the DCN-improved D_C2f, D_SPPF, and BiRA modules are combined, the DBD-YOLOv8n model achieves the highest mAP@0.5 performance on the FLIR, OTCBVS



(a) Object label of input image. (b) Object detection result of YOLOv8n. (c) Object detection result of DBD-YOLOv8n.

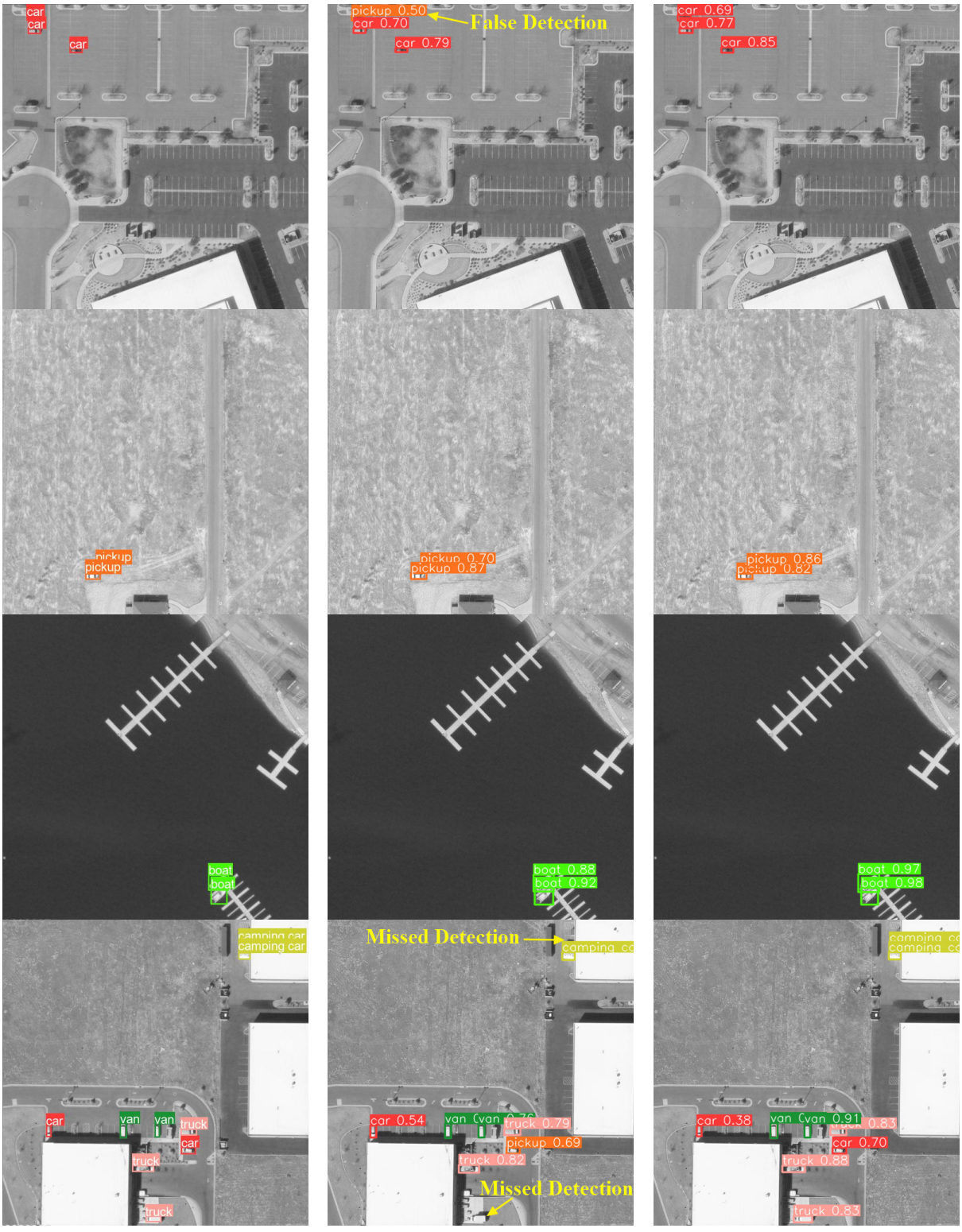
FIGURE 13. Comparison of object detection results on OTCBVS (Dataset 03) Test Sets between YOLOv8n and DBD-YOLOv8n.

(Dataset 01), OTCBVS (Dataset 03), and VEDAI test sets, with values of 81.6%, 94.0%, 99.6%, and 72.8% respectively. The inference times were recorded as 6.2ms, 16.5ms, 19.3ms, and 12.3ms respectively, demonstrating the real-time capability of the proposed object detection system to meet the requirements of dynamic detection scenarios. This indicates that each module contributes independently to the performance.

F. EXPERIMENTAL RESULTS COMPARISON WITH DIFFERENT MODELS

To quantitatively analyze the object detection performance of the DBD-YOLOv8 model, we conducted training and testing comparisons on publicly available datasets, namely FLIR, OTCBVS (Dataset 01), OTCBVS (Dataset 03), and VEDAI. We compared the DBD-YOLOv8 series models with popular object detection models, including Faster R-CNN, YOLOv3, and YOLOv5. The experimental results are presented in Table 3.

The object detection results on FLIR, OTCBVS (Dataset 01), OTCBVS (Dataset 03), and VEDAI Test Sets were compared between YOLOv8n and DBD-YOLOv8n, as presented in Figures 11 to 14. The results clearly demonstrate that the DBD-YOLOv8n model exhibits significant enhancements over the original model. Specifically, it effectively reduces false detections and missed detections, thereby improving the detection performance for multi-scale occluded objects and small objects.



(a) Object label of input image. (b) Object detection result of YOLOv8n. (c) Object detection result of DBD-YOLOv8n.

FIGURE 14. Comparison of object detection results on VEDAI test sets between YOLOv8n and DBD-YOLOv8n.

Compared to Faster R-CNN, YOLOv3, and YOLOv5n, the DBD-YOLOv8 model achieves improved average precision (mAP@0.5) for all object categories. The DBD-YOLOv8 model, which incorporates improvements such as DCN, the

TABLE 3. Experimental results of different object detection methods on the RSOD, NWPU VHR-10, DIOR, and VEDAI datasets.

Method	mAP@0.5 (%) [1]				mAP@0.5:0.95 (%) [1]				Inference time (ms) [1]			
	D1	D2	D3	D4	D1	D2	D3	D4	D1	D2	D3	D4
DETR	77.5	95.0	99.2	73.2	59.0	76.1	78.2	50.1	18.4	49.5	50.6	48.2
Deformable DETR [53]	79.2	96.5	99.5	75.0	60.2	77.0	79.3	51.6	10.5	35.2	37.5	30.4
SSD	68.5	80.3	82.4	59.4	50.1	49.5	58.8	43.2	26.3	35.8	38.5	45.6
Faster R-CNN	70.4	81.6	83.3	60.3	51.6	50.7	60.7	44.3	32.5	50.3	54.2	60.3
YOLOv3	71.8	85.1	88.6	63.1	53.8	58.6	66.4	42.8	25.8	36.8	41.2	45.8
YOLOv5n	75.5	90.2	95.2	67.5	56.4	66.5	74.2	46.6	6.7	17.7	20.1	27.5
YOLOv8n	75.6	93.3	99.5	70.8	57.4	73.6	75.5	47.5	5.8	16.5	18.7	10.6
YOLOv8s	77.1	94.1	99.5	72.4	58.8	75.3	76.0	48.3	6.3	20.7	21.8	13.7
YOLOv8m	78.7	94.9	99.6	73.1	60.6	76.7	76.1	49.2	10.5	33.5	34.0	26.9
YOLOv8l	79.6	95.6	99.6	74.5	62.4	76.2	76.7	50.1	12.6	41.6	42.5	26.1
YOLOv8x	81.8	96.2	99.6	76.2	64.3	77.1	77.5	50.9	17.8	48.3	65.7	48.4
DBD-YOLOv8n	81.6	94.0	99.6	72.8	58.2	75.5	76.0	48.8	5.7	15.9	18.1	10.3
DBD-YOLOv8s	83.0	95.3	99.6	74.0	60.0	76.1	77.8	50.0	6.0	19.5	20.5	13.5
DBD-YOLOv8m	84.5	96.5	99.7	75.8	61.5	77.9	79.0	51.7	9.1	28.6	26.7	18.5
DBD-YOLOv8l	86.1	97.1	99.7	77.9	63.1	78.8	80.5	53.0	11.5	36.3	38.8	21.3
DBD-YOLOv8x	88.6	98.4	99.7	79.4	64.9	80.0	81.8	54.1	16.5	37.9	40.7	35.0

[1] D1: FLIR, D2: OTCBVS (Dataset 01), D3: OTCBVS (Dataset 03), D4: VEDAI.

added BiRA module, and the Dyheads module, exhibits significant performance advancements compared to the YOLOv8 series models (YOLOv8-n\s\m\l\x). On the FLIR, OTCBVS (Dataset 01), OTCBVS (Dataset 03), and VEDAI test sets, the average mAP@0.5 increases by 7.9%, 1.5%, 0.1%, and 3.5%, respectively, reaching 84.8%, 96.3%, 99.7%, and 76.0%. The average inference times are 10.9ms, 32.0ms, 37.3ms, and 28.4ms (equivalent to 92fps, 31fps, 27fps, and 35fps). Although there is a slight increase in the number of parameters, resulting in a partial loss of inference time, the model's average inference time still meets the real-time detection requirements.

V. CONCLUSION

To address the challenges associated with low signal-to-noise ratio and insufficient texture details in infrared images, we propose an enhanced version of the YOLOv8 base series (YOLOv8-n\s\m\l\x) specifically designed for infrared object detection tasks. Our approach aims to improve the accuracy of detecting multi-scale, occluded, and small objects.

Our approach introduces two key modules: the DCN-based D_C2f and D_SPPF modules, which replace the C2f and SPPF modules of the backbone network. These modules utilize deformable convolutions to replace the feature maps of layers P3, P4, and SPPF. By dynamically adjusting the convolution kernel shape and adapting the network's receptive field based on the object's position, our model improves its ability to focus on object regions, capture object deformations, and handle scale variations. As a result, our model exhibits enhanced capabilities in representing multi-scale features.

Additionally, we introduce the BiRA module, which incorporates a dual-route attention mechanism between the backbone and neck networks. This module extracts attention weights for feature maps of different scales and hierarchical channel attention vectors, correcting and weighting the feature maps to generate richer and more refined multi-scale feature maps. This enhancement strengthens the model's feature learning capabilities.

To further enhance the semantic representation of features and improve object detection and localization accuracy, we introduce the Dyheads module within the neck network. This module integrates attention mechanisms that promote scale-awareness (Level-wise), spatial awareness (Spatial-wise), and task awareness (Channel-wise). By combining these different awareness aspects, our model achieves a more robust semantic representation of features.

Our proposed method for infrared object detection, called DBD-YOLOv8, significantly enhances the model's feature representation capabilities for detecting multi-scale, occluded, and small objects. It also improves the model's ability to learn and fuse multi-scale features. Although this enhancement slightly increases model complexity and incurs a minor sacrifice in inference time, it leads to a notable improvement in the average detection accuracy. Experimental results validate the effectiveness and applicability of the DBD-YOLOv8 model in infrared object detection.

In our forthcoming research, our primary objective is to explore more efficient methods for extracting and fusing object information. Our aim is to develop a robust detection model specifically tailored for identifying infrared objects within intricate backgrounds, leveraging their spatiotemporal

features. A particular focus of our investigation is the integration of cutting-edge Transformer models into the realm of object detection. We draw inspiration from the Transformer model's remarkable ability to capture long-range dependencies and dynamically aggregate spatial information. Our intention is to address inherent limitations associated with fixed receptive fields observed in YOLO, as well as potential overlooked dependencies between objects. To achieve this, we plan to incorporate both the Vision Transformer (ViT) and the bi-directional feature pyramid network (BiFPN) architecture into our research framework. This integration is anticipated to significantly enhance the model's proficiency in handling intricate backgrounds and augment object detection performance.

REFERENCES

[1] X. Shao, H. Fan, G. Lu, and J. Xu, "An improved infrared dim and small target detection algorithm based on the contrast mechanism of human visual system," *Infr. Phys. Technol.*, vol. 55, no. 5, pp. 403–408, Sep. 2012, doi: 10.1016/j.infrared.2012.06.001.

[2] Y. Dai and Y. Wu, "Reweighted infrared patch-tensor model with both nonlocal and local priors for single-frame small target detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 8, pp. 3752–3767, Aug. 2017, doi: 10.1109/JSTARS.2017.2700023.

[3] X. Zhang, W. Jin, P. Yuan, C. Qin, H. Wang, J. Chen, and X. Jia, "Research on passive wide-band uncooled infrared imaging detection technology for gas leakage," in *Proc. Int. Conf. Opt. Instrum. Technol., Opt. Syst. Modern Optoelectronic Instrum.*, Beijing, China, Mar. 2020, pp. 144–157, doi: 10.1117/12.2542906.

[4] P. B. Chapple, D. C. Bertilone, R. S. Caprari, S. Angeli, and G. N. Newsam, "Target detection in infrared and SAR terrain images using a non-Gaussian stochastic model," *Proc. SPIE*, vol. 3699, pp. 122–132, Jul. 1999, doi: 10.1117/12.352951.

[5] M. Zhao, W. Li, L. Li, J. Hu, P. Ma, and R. Tao, "Single-frame infrared small-target detection: A survey," *IEEE Geosci. Remote Sens. Mag.*, vol. 10, no. 2, pp. 87–119, Jun. 2022, doi: 10.1109/MGRS.2022.3145502.

[6] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, Jun. 2014, pp. 580–587, doi: 10.1109/CVPR.2014.81.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, and A. N. Gomez, "Attention is all you need," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, Long Beach, CA, USA: Curran Associates, 2017, pp. 1–11.

[8] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional networks for accurate object detection and segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 1, pp. 142–158, Jan. 2016, doi: 10.1109/TPAMI.2015.2437384.

[9] R. Gavrilescu, C. Zet, C. Fosalau, M. Skoczylas, and D. Cotovanu, "Faster R-CNN: An approach to real-time object detection," in *Proc. Int. Conf. Expo. Electr. Power Eng. (EPE)*, Oct. 2018, pp. 0165–0168, doi: 10.1109/icepe.2018.8559776.

[10] T. Kong, A. Yao, Y. Chen, and F. Sun, "HyperNet: Towards accurate region proposal generation and joint object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 845–853, doi: 10.1109/CVPR.2016.98.

[11] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg, "SSD: Single shot MultiBox detector," in *Computer Vision—ECCV* (Lecture Notes in Computer Science). Springer, 2016, pp. 21–37, doi: 10.1007/978-3-319-46448-0_2.

[12] 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. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 779–788, doi: 10.1109/CVPR.2016.91.

[13] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6517–6525, doi: 10.1109/CVPR.2017.690.

[14] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.

[15] Y. Liu, N. Ong, K. Peng, B. Xiong, Q. Wang, R. Hou, M. Khabsa, K. Yang, D. Liu, D. S. Williamson, and H. Yu, "MMViT: Multiscale multiview vision transformers," 2023, *arXiv:2305.00104*.

[16] N. Kitaev, L. Kaiser, and A. Levskaya, "Reformer: The efficient transformer," 2020, *arXiv:2001.04451*.

[17] J. Lin, X. Mao, Y. Chen, L. Xu, Y. He, and H. Xue, "D^{2}ETR: Decoder-only DETR with computationally efficient cross-scale attention," 2022, *arXiv:2203.00860*.

[18] C. Nicolas, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 213–229, doi: 10.1007/978-3-030-58452-8_13.

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017, doi: 10.1145/3065386.

[20] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 764–773, doi: 10.1109/ICCV.2017.89.

[21] X. Zhu, H. Hu, S. Lin, and J. Dai, "Deformable ConvNets v2: More deformable, better results," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9300–9308, doi: 10.1109/CVPR.2019.00953.

[22] W. Wang, J. Dai, Z. Chen, Z. Huang, Z. Li, X. Zhu, X. Hu, T. Lu, L. Lu, H. Li, X. Wang, and Y. Qiao, "InternImage: Exploring large-scale vision foundation models with deformable convolutions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 14408–14419, doi: 10.1109/CVPR52729.2023.01385.

[23] L. Zhu, X. Wang, Z. Ke, W. Zhang, and R. Lau, "BiFormer: Vision transformer with bi-level routing attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 10323–10333, doi: 10.1109/CVPR52729.2023.00995.

[24] X. Dai, Y. Chen, B. Xiao, D. Chen, M. Liu, L. Yuan, and L. Zhang, "Dynamic head: Unifying object detection heads with attentions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 7369–7378, doi: 10.1109/CVPR46437.2021.00729.

[25] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 11976–11986, doi: 10.1109/CVPR52688.2022.01167.

[26] X. Ding, X. Zhang, J. Han, and G. Ding, "Scaling up your kernels to 31×31: Revisiting large kernel design in CNNs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 11963–11975, doi: 10.1109/CVPR52688.2022.01166.

[27] Z. Gao, L. Wang, B. Han, and S. Guo, "AdaMixer: A fast-converging query-based object detector," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5354–5363, doi: 10.1109/CVPR52688.2022.00529.

[28] B. Roh, J. Shin, W. Shin, and S. Kim, "Sparse DETR: Efficient end-to-end object detection with learnable sparsity," 2021, *arXiv:2111.14330*.

[29] F. Gu, J. Lu, C. Cai, Q. Zhu, and Z. Ju, "EANTrack: An efficient attention network for visual tracking," *IEEE Trans. Autom. Sci. Eng.*, early access, Oct. 3, 2023, doi: 10.1109/TASE.2023.3319676.

[30] F. Gu, J. Lu, and C. Cai, "RPformer: A robust parallel transformer for visual tracking in complex scenes," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–14, 2022, doi: 10.1109/TIM.2022.3170972.

[31] Y. Liu, Y. Zhang, S. Liu, S. Coleman, Z. Wang, and F. Qiu, "Salient object detection by aggregating contextual information," *Pattern Recognit. Lett.*, vol. 153, pp. 190–199, Jan. 2022, doi: 10.1016/j.patrec.2021.12.011.

[32] L. Hou, K. Lu, J. Xue, and L. Hao, "Cascade detector with feature fusion for arbitrary-oriented objects in remote sensing images," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, London, U.K., Jul. 2020, pp. 1–6, doi: 10.1109/ICME46284.2020.9102807.

[33] D. Xu and Y. Wu, "FE-YOLO: A feature enhancement network for remote sensing target detection," *Remote Sens.*, vol. 13, no. 7, p. 1311, Mar. 2021, doi: 10.3390/rs13071311.

[34] L. Chen, W. Shi, and D. Deng, "Improved YOLOv3 based on attention mechanism for fast and accurate ship detection in optical remote sensing images," *Remote Sens.*, vol. 13, no. 4, p. 660, Feb. 2021, doi: 10.3390/rs13040660.

[35] X. Hua, X. Wang, T. Rui, H. Zhang, and D. Wang, "A fast self-attention cascaded network for object detection in large scene remote sensing images," *Appl. Soft Comput.*, vol. 94, Sep. 2020, Art. no. 106495, doi: 10.1016/j.asoc.2020.106495.

[36] Q. Ran, Q. Wang, B. Zhao, Y. Wu, S. Pu, and Z. Li, "Lightweight oriented object detection using multiscale context and enhanced channel attention in remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 5786–5795, 2021, doi: 10.1109/JSTARS.2021.3079968.
[37] W. Liu, L. Ma, J. Wang, and H. Chen, "Detection of multiclass objects in optical remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 5, pp. 791–795, May 2019, doi: 10.1109/LGRS.2018.2882778.
[38] L. Shen, B. Lang, and Z. Song, "DS-YOLOv8-based object detection method for remote sensing images," *IEEE Access*, vol. 11, pp. 125122–125137, 2023, doi: 10.1109/ACCESS.2023.3330844.
[39] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.* (Lecture Notes in Computer Science), 2018, pp. 3–19, doi: 10.1007/978-3-030-01234-2_1.
[40] J. Park, S. Woo, J.-Y. Lee, and I. So Kweon, "BAM: Bottleneck attention module," 2018, *arXiv:1807.06514*.
[41] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," 2020, *arXiv:2012.12877*.
[42] C. Li, B. Luo, H. Hong, X. Su, Y. Wang, J. Liu, C. Wang, J. Zhang, and L. Wei, "Object detection based on global-local saliency constraint in aerial images," *Remote Sens.*, vol. 12, no. 9, p. 1435, May 2020, doi: 10.3390/rs12091435.
[43] Z. Tong, Y. Chen, Z. Xu, and R. Yu, "Wise-IoU: Bounding box regression loss with dynamic focusing mechanism," 2023, *arXiv:2301.10051*.
[44] P. Das, A. Chakraborty, R. Sankar, O. K. Singh, H. Ray, and A. Ghosh, "Deep learning-based object detection algorithms on image and video," in *Proc. 3rd Int. Conf. Intell. Technol. (CONIT)*, Hubli, India, Jun. 2023, pp. 1–6, doi: 10.1109/CONIT59222.2023.10205601.
[45] S. Liu, D. Huang, and Y. Wang, "Adaptive NMS: Refining pedestrian detection in a crowd," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 6452–6461, doi: 10.1109/CVPR.2019.00662.
[46] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee, "YOLACT: Real-time instance segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9156–9165, doi: 10.1109/ICCV.2019.00925.
[47] L. Deng, Y. Gong, X. Lu, X. Yi, Z. Ma, and M. Xie, "Focus-enhanced scene text recognition with deformable convolutions," in *Proc. IEEE 5th Int. Conf. Comput. Commun. (ICCC)*, Dec. 2019, pp. 1685–1689, doi: 10.1109/ICCC47050.2019.9064428.
[48] W. Xi, L. Sun, and J. Sun, "Upgrade your network in-place with deformable convolution," in *Proc. 19th Int. Symp. Distrib. Comput. Appl. Bus. Eng. Sci. (DCABES)*, Oct. 2020, pp. 239–242, doi: 10.1109/dcabes50732.2020.00069.
[49] S. Ren, D. Zhou, S. He, J. Feng, and X. Wang, "Shunted self-attention via multi-scale token aggregation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 10843–10852, doi: 10.1109/cvpr52688.2022.01058.
[50] F. Teledyne. *FREE FLIR Thermal Dataset for Algorithm*. Accessed: Nov. 10, 2023. [Online]. Available: https://www.flir.com/oem/adas/adas-dataset-form/
[51] N. Lannan, L. Zhou, and G. Fan, "A multiview depth-based motion capture benchmark dataset for human motion denoising and enhancement research," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 427–436, doi: 10.1109/CVPRW56347.2022.00058.
[52] S. Razakarivony and F. Jurie, "Vehicle detection in aerial imagery : A small target detection benchmark," *J. Vis. Commun. Image Represent.*, vol. 34, pp. 187–203, Jan. 2016, doi: 10.1016/j.jvcir.2015.11.002.
[53] 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*.

LINGYUN SHEN received the B.S. and M.S. degrees from the Changchun University of Science and Technology, and the Ph.D. degree from the University of Chinese Academy of Sciences, China, in 2015. She is currently an Associate Professor with the Taiyuan Institute of Technology. Her current research interests include machine vision and intelligent processing of information.

BAIHE LANG received the B.S. and M.S. degrees from the Changchun Institute of Optics, Fine Mechanics and Physics, China, in 2000. He is currently an Associate Professor with the Changchun University of Science and Technology. His current research interests include deep learning and intelligent processing of information.

ZHENGXUN SONG received the B.S. degree from the Changchun Institute of Optics, Fine Mechanics and Physics, and the M.S. degree from the Jilin University of Technology, China. He is currently a Professor with the International Research Center for Nano Handling and Manufacturing of China, Changchun University of Science and Technology, and the Overseas Expertise Introduction Project for Discipline Innovation. His research interests include micro-nano manipulation technology, optical communication technology, and intelligent processing of information.