Advanced Engineering Informatics 74 (2026) 104639

Contents lists available at ScienceDirect

Advanced Engineering Informatics

journal homepage: www.elsevier.com/locate/aei

Full length article

IPeDet: An end-to-end fine-grained feature aggregation network for UAV infrared pedestrian detection

Yi Li [a,b,1], Huiying Xu [a,b,*,1], Xinzhong Zhu [a,b,c], Hongbo Li [c,*], Yiming Sun [d], Ruidong Wang [a,b], Lingling Xu [e]

[a] Zhejiang Key Laboratory of Intelligent Education Technology and Application, Zhejiang Normal University, Jinhua, Zhejiang, 321004, China
[b] School of Computer Science and Technology of Zhejiang Normal University, Jinhua, Zhejiang, 321004, China
[c] Beijing Geekplus Technology Co., Ltd, Beijing, 100101, China
[d] School of Automation, Southeast University, Nanjing, Jiangsu, 210096, China
[e] School of Computer Science of Hangzhou Dianzi University, Hangzhou, Zhejiang, 310018, China

ARTICLE INFO

Keywords:
Feature fusion
UAV small object detection
Lightweight networks
Convolutional re-paramerization

ABSTRACT

Traditional visible light-based pedestrian detection methods face significant limitations in complex environments, often suffering from high miss rates and false alarms, which severely compromise perception reliability in real-world applications. To address these challenges, we present IPeDet, a robust infrared pedestrian detector designed for Unmanned Aerial Vehicle (UAV) scenarios, featuring enhanced fine-grained feature aggregation. Specifically, we introduce HGNetv2-CS, a channel-slimmed variant of HGNetv2, as a lightweight backbone specifically optimized for small-target detection. Furthermore, we incorporate Large Separable Kernel Attention (LSKA) at the early multi-scale fusion stage to capture explicit spatial dependencies. To enhance efficiency, Online Convolutional Re-Parameterization (OCRP) is adopted to stabilize the training of deep layers and reduce the computational overhead of feature fusion. Finally, we propose Multi-Path Coordinate Attention (MPCA), a novel mechanism that mitigates feature redundancy and fusion-induced aliasing through adaptive weight refinement. Extensive experiments demonstrate that IPeDet significantly outperforms both the baseline and mainstream object detectors. Concretely, IPeDet achieves a detection accuracy improvement of 5.88% in $mAP@50{:}95$ on the HIT-UAV dataset and a 6.23% gain on the M^3FD dataset, establishing a new state-of-the-art in accuracy and efficiency for infrared pedestrian detection.

1. Introduction

Pedestrian detection is a pivotal task in computer vision, focusing on the precise recognition and localization of individuals within images or video sequences. With the rapid evolution of intelligent transportation systems and smart surveillance, this technology has become an indispensable component of modern infrastructure [1,2]. Moreover, it serves as a foundational pillar for emerging research domains, such as human behavior analysis and motion understanding. Consequently, the pursuit of more efficient pedestrian detection methodologies possesses both significant practical utility and profound academic value [3–5]. Prior to the advent of deep learning methodologies, pedestrian detection predominantly depended on manually engineered features for target characterization [6–9]. While these manually crafted features are straightforward and user-friendly, they exhibit significant limitations, such as a propensity for misdetection and reliance primarily on superficial information. In recent years, the remarkable progress in deep learning has sparked growing interest in its application to pedestrian detection systems, driven by the superior performance of these advanced techniques.

While visible-light imagery delivers robust performance in daytime environments, its efficacy is severely compromised during nighttime or under suboptimal lighting conditions [10]. In contrast, infrared imaging has emerged as a vital complementary modality in surveillance systems, particularly for drone-based applications, by leveraging its distinct imaging capabilities [11]. Consequently, object detection utilizing infrared imagery from UAVs demonstrates significant potential across diverse fields, offering advantages such as cost-effectiveness, operational flexibility, and superior performance [12].

* Corresponding authors.
E-mail addresses: leeyee@zjnu.edu.cn (Y. Li), xhy@zjnu.edu.cn (H. Xu), zxz@zjnu.edu.cn (X. Zhu), Jason.li@geekplus.com (H. Li), sunyiming@seu.edu.cn (Y. Sun), iswangrd@zjnu.edu.cn (R. Wang), linglingxu@hdu.edu.cn (L. Xu).
[1] The authors contributed equally to this work.

https://doi.org/10.1016/j.aei.2026.104639
Received 11 July 2025; Received in revised form 26 March 2026; Accepted 26 March 2026

Nevertheless, deploying pedestrian detection on small UAV equipped with infrared sensors faces unique challenges: **Feature Degradation and Noise**: Detection accuracy is critically impeded by the intrinsic quality of infrared images, which often suffer from low contrast, blurred textures, and significant noise interference. These factors degrade the feature representation capability of the network, presenting a technical barrier that demands innovative signal processing and enhancement solutions [13,14]. **Small-Scale Target Identification:** Pedestrian targets in aerial infrared imagery are often characterized by their extremely small scale and weak thermal signatures, especially in critical UAV missions (such as nighttime search-and-rescue or security surveillance). The combination of low resolution and the lack of rich semantic information (such as color or texture) leave it difficult for conventional detectors to distinguish targets from the background. Consequently, missed detections caused by these weak thermal signatures and noise can directly lead to mission failure [15]. **Computational Constraints:** The intensive computational requirements of state-of-the-art models conflict with the limited hardware resources of UAVs. High processing latency hinders the ability to perform real-time, onboard inference, which is essential for dynamic missions. Therefore, ensuring the robustness of infrared detection systems in dynamic and complex aerial settings has become a central focus of the field [16,17]. While existing literature predominantly focuses on detection under favorable conditions, a substantial gap remains regarding reliability in complex, unstructured environments. Furthermore, effectively reconciling the inherent trade-off between detection accuracy and inference latency is increasingly paramount for practical UAV deployment.

Solving these interconnected problems is crucial for the advancement of autonomous UAV operations. In critical scenarios such as nighttime emergency rescue and border surveillance, the inability to distinguish a small, weak thermal signature from background noise directly translates to unacceptable miss rates [18]. Furthermore, if a model is too computationally heavy to run onboard the UAV in real-time, it cannot support dynamic decision-making or obstacle avoidance, rendering the system practically useless. Despite these pressing needs, existing literature predominantly focuses on either maximizing accuracy at the cost of massive computational overhead, or deploying overly simplified models that fail in unstructured, noisy environments [19]. A substantial gap remains in developing a specialized architecture that directly combats infrared feature degradation and small-scale target loss while strictly adhering to the severe power and memory constraints of edge devices.

To address these crucial gaps, we propose IPeDet, a lightweight end-to-end infrared pedestrian detector tailored for UAV scenarios. The motivation behind our design is to explicitly map robust feature enhancement mechanisms to the specific constraints of UAV infrared imagery. Our approach leverages a channel-slimmed HGNetv2 backbone to optimize feature representation while maintaining computational efficiency. We introduce Large Separable Kernel Attention (LSKA) to capture extensive spatial features from shallow layers, thereby boosting small-scale feature aggregation, and we employ online convolutional re-parameterization to ensure training convergence in deeper stages. Furthermore, a novel Multi-Path Coordinate Attention (MPCA) mechanism is devised to reduce feature redundancy and enhance discriminability. Extensive experiments demonstrate that IPeDet outperforms most mainstream detectors, effectively adhering to the lightweight paradigm. The main contributions of this work are summarized as follows:

- We propose IPeDet, an end-to-end lightweight infrared pedestrian detector tailored for UAV scenarios. By employing a channel-slimmed HGNetv2-CS backbone and a dedicated detection head, this framework effectively resolves the trade-off between feature representation capability and model complexity, ensuring high efficiency on resource-constrained devices.
- To address the limitation of insufficient spatial context in shallow layers, we introduce Large Separable Kernel Attention (LSKA) at the early multi-scale fusion stage. This mechanism expands the receptive field to capture explicit spatial dependencies, thereby enhancing the model's ability to locate small-scale targets.
- We devise a joint optimization strategy for robust feature fusion. Specifically, Online Convolutional Re-Parameterization (OCRP) is adopted to stabilize deep-layer training and reduce computational overhead, while our novel designed Multi-Path Coordinate Attention (MPCA) mitigates feature aliasing and redundancy, ensuring more discriminative feature representation.

2. Related works

2.1. General efficient object detector

The rapid evolution of deep learning has revolutionized object detection, with Convolutional Neural Networks (CNNs) serving as the cornerstone of modern algorithms. These methods are generally categorized into two-stage and one-stage architectures based on their inference workflow. Two-stage detectors operate by first generating a sparse set of candidate region proposals from the input image. Subsequently, deep features are extracted from these regions to perform precise classification and bounding box regression. Representative algorithms in this paradigm include Faster RCNN [20], FPN [21], and Mask RCNN [22]. Conversely, one-stage detectors forgo the region proposal step to directly regress class probabilities and bounding box coordinates from the image, thereby significantly reducing inference latency. Prominent examples include SSD [23], RetinaNet [24], CenterNet [25], EfficientDet [26], and the YOLO series [27]. Notably, recent iterations such as YOLO11 [28] and YOLOv12 [29] have further optimized the architecture by introducing advanced anchor mechanisms and multi-scale prediction strategies. These improvements represent a state-of-the-art balance between detection accuracy and computational efficiency. Despite the success of generic detectors in visible light, their direct application to infrared pedestrian detection — particularly in UAV scenarios — remains challenging. Our proposed IPeDet specifically targets the dual challenges of IR image degradation and small-scale targets, achieving a superior balance between lightweight deployment and high-precision detection in complex aerial environments.

2.2. Research of infrared pedestrian recognition

A significant amount of research has been dedicated to pedestrian recognition and detection in infrared (IR) scenes. Fu et al. [30] present a feature-augmented long-range attention fusion network, with the objective of improving detection accuracy through the integration of long-range dependencies inherent in visible and infrared images. Wei et al. [31] introduce an improved UNet and YOLO method which shares the visible light information from multiple related datasets. To fully exploit infrared–visible information in both modalities. Wu et al. [32] develop implicit modality knowledge alignment and uncertainty estimation network, strengthening robustness of learned common embedding space and leading better detection results. To enhance the pedestrian detection capabilities of fused images while preserving pixel-level information, Zheng et al. [33] introduces an innovative two-stage cascaded network architecture specifically designed to thoroughly explore the relationship between image fusion outcomes and high-level visual details. To address the limitations of the Kaniadakis entropy thresholding method, which struggles with noisy images and complex backgrounds despite its effectiveness in segmenting infrared images with long-tail histogram distributions, Lei et al. [34] propose a novel infrared pedestrian segmentation algorithm based on two-dimensional Kaniadakis entropy thresholding. Li et al. [35] develop a novel cross-modality disentanglement and shared feedback learning framework to boost infrared–visible person re-identification performance, through

efficient cross-modality images disentanglement network and a dual-path shared feedback learning network. Farhat et al. [36] present a novel multi-object tracking model combining the YOLOv9 detection algorithm with DeepSORT tracking to addresses the need for enhanced detection of pedestrians under occlusion, scale variations and complex backgrounds, improving detection accuracy for real-time autonomous systems. Our proposed IPeDet significantly improves pedestrian detection performance, particularly for small-scale targets in infrared urban environments, while preserving high computational efficiency and a compact model architecture.

2.3. Feature enhancement and fusion

There are two common types of multi-scale feature fusion networks: the first is a parallel multi-branch network. SPPNet developed a strategy that could generates the uniform output vector according different inputs with spatial pyramid pooling. ASPP [37] can accurately and efficiently classify regions of varying proportions by resampling convolutional features extracted at a single scale. PSPNet [38] integrates features from four different scales to create a comprehensive representation of various regions, thereby enhancing feature extraction in segmentation tasks. The second is a serial layer-hopping connectivity structure. Both types perform feature extraction across different receptive fields. FPN [39,40] enhances the network's recognition and representation capabilities across different scales by fusing high-level semantic information with low-level detailed features through a top-down pathway. BiFPN [26] conceptualizes each bi-directional path as an individual feature network layer and through iterative replication of this layer structure, it facilitates a progressive enhancement of feature fusion efficacy. In contrast, our proposed method incorporates a fine-grained feature aggregation strategy reinforced by Large Separable Kernel Attention (LSKA) and Multi-Path Coordinate Attention (MPCA). Unlike standard fusion, our approach explicitly expands the receptive field in shallow layers to capture spatial context and adaptively refines features during fusion to mitigate aliasing effects, thereby significantly enhancing the detectability of small infrared targets.

2.4. Lightweight networks

Designing a lightweight neural architecture that achieves both rapid inference speed and high performance presents a significant challenge. MobileNet [41] proposed Depth-wise separable convolution (DSConv) to replace traditional 3×3 and 1×1 filters, which largely reduce the overall model parameters and complexity. ShuffleNet [42] introduced channel shuffle operation to enhance the information transmission across different groups. GhostNet [43] introduced ghost module to generate more related features by the integration of 1×1 convolution and DWConv, which largely reduce the model's redundancy. FasterNet [44] developed partial convolution architecture to obtain effective feature extraction with less memory access and lower model complexity. MobileViT [45] proposes an integrated framework that merges lightweight MobileNet components — characterized by point-wise and depth-wise convolution operations — with Multi-Head Self-Attention (MHSA) modules. MobileMamba [46] leverages the efficient long-range modeling capabilities of State Space Models (SSMs) while optimizing for mobile constraints through a lightweight architecture. Lu et al. [47] propose a lightweight CNN–Transformer network incorporating a novel Laplacian loss and an efficient QVSATA decoder to solve the challenge of real-time, high-accuracy semantic segmentation for low-altitude UAV imagery on resource-constrained embedded systems. Wu et al. [48] propose AirboardNet, a lightweight multi-task network leveraging a teacher-student framework and hybrid distillation loss, to solve the challenge of limited onboard computational resources for real-time, high-accuracy inspection of high-speed railway bridge girders. Yuan et al. [49] propose a distributed Edge-Cloud collaborative framework utilizing an Edge-Embedded Lightweight (E2L) algorithm and a fuzzy neural network-based decision mechanism to solve the challenges of real-time, accurate ground moving target detection on UAVs with limited resources. Sang et al. [50] propose a lightweight thermal image super-resolution (LTSR) model that combines a multiscale knowledge distillation network with latent neural representations to solve the challenges of insufficient thermal image resolution and the deployment constraints of resource-limited UAVs. Zhang et al. [51] propose SGMFNet, a network utilizing global–local feature guidance and parallel sampling feature fusion, to solve the challenges of complex backgrounds, significant scale differences, and densely arranged small objects in UAV image object detection. While substantial progress characterizes the field of lightweight UAV vision, IPeDet uniquely addresses the specific challenges of infrared pedestrian detection, delivering superior accuracy within a compact architectural framework.

3. Methodology

3.1. Framework of IPeDet

The primary advantage of the YOLO algorithm lies in its high speed and efficiency, which facilitate real-time object detection and significantly enhance detection speed by predicting bounding boxes and categories simultaneously through a single forward pass. In this work, we choose the lightweight YOLOv10s, for its satisfactory parameters and competitive detection performance, as the baseline with several improvement strategies to build IPeDet. The overall architecture of IPeDet is presented in Fig. 1 and consists several key components: lightweight HGNetv2-CS functions as the main backbone. LSKA is embedded in the shallow stage of model with large receptive fields for spatial feature exploration. OCRP equips with re-parameterization topology for fast training and advanced multi-scale feature fusion in the deep layers. IPeDet adopts 4 different scales MPCA enhanced heads to perform final classification and localization.

3.2. Efficient feature extraction backbone

Currently, lightweight convolutional networks are of great necessity in the fields of computer vision tasks, especially in scenarios with limited processing resources. Despite the performance gains achieved by deep learning-enhanced CNNs, their exorbitant computational overhead and memory consumption severely constrain real-world adoption, particularly in resource-constrained environments. Lightweight convolutional networks can provide good performance in environments with limited computational resources by designing more efficient models to reduce computation, memory requirements and inference latency.

The architecture of HGNetv2 [52] is shown in Fig. 2(a), we will elaborate each module in the follows. The *Stem* block is the preprocessing layers for extracting feature from raw data, including convolutional layers, batch normalization and non-linear activation function. As the core components, HG block aims to process data in a hierarchical manner shown in Fig. 2(b), each could deal with a different level of abstraction of the data, allowing the network to learn from both low-level and high-level features. Between HG blocks, layer downsampling (LDS) operations can be applied to compress the feature map dimensions, easing computational burdens and enhancing the receptive field capacity of deeper network stages. The global average pooling (GAP) condenses each feature map into a single vector by aggregating spatial information, thereby enhancing the network's invariance to input data spatial variations. Finally, the network ends with a series of fully connected and convolutional layers that perform the final classification task.

As shown in Fig. 2(b), The HG block operates on the principle of hierarchical feature extraction, where multiple convolutional pathways process input data using different filter sizes. This parallel processing captures features at varying scales and levels of abstraction, which

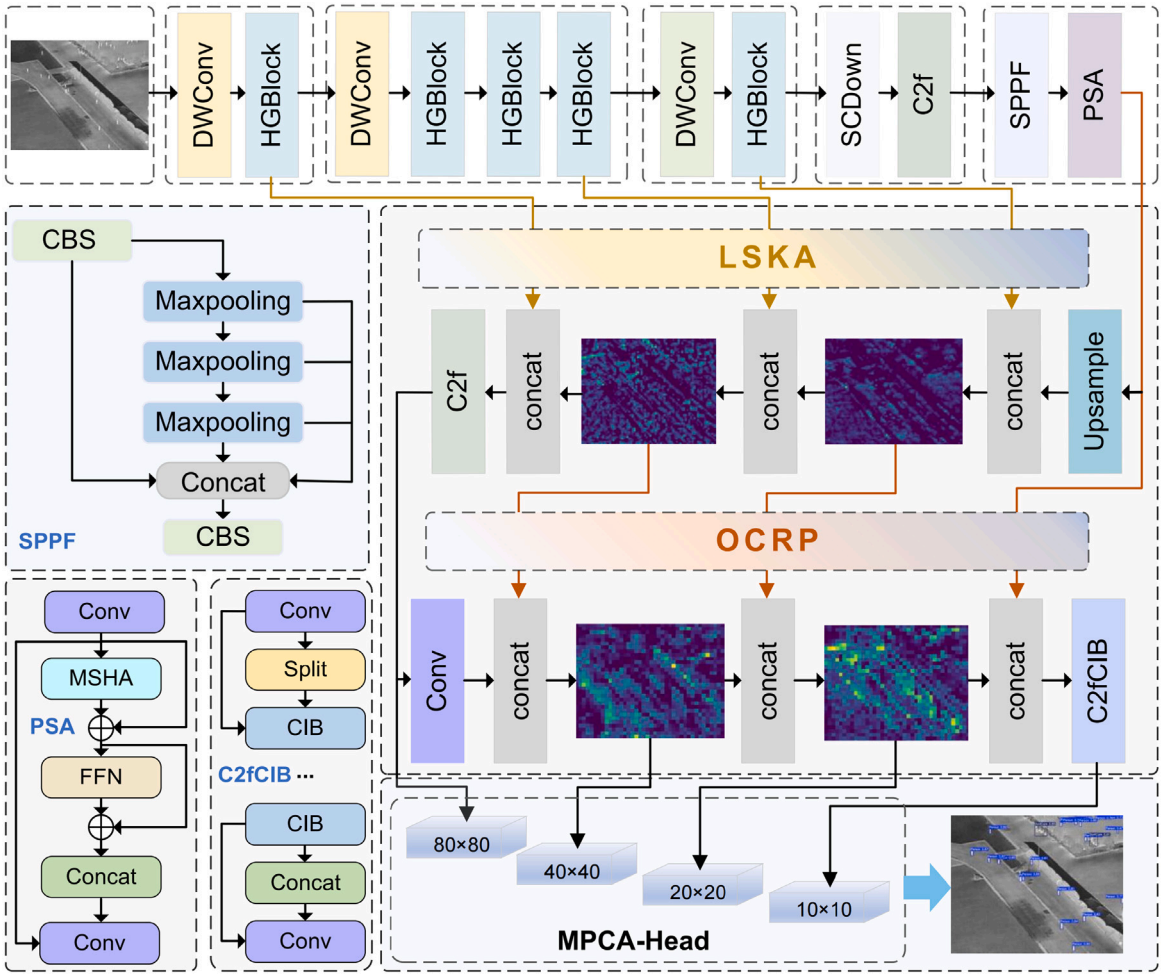


Fig. 1. Overview architecture of IPeDet. Channel slimmed HGNetv2 with combination of DWConv and HGBlock as the main backbone. Shallow stage layers multi-scale feature fusion with large separable kernel attention (LSKA). Deep stage layers multi-scale feature fusion with online convolutional re-parameterization (OCRP) learning. MPCA-Head: multi-path coordinate attention (MPCA) refinement with 4 scales detection heads 80 × 80, 40 × 40, 20 × 20 and 10 × 10.

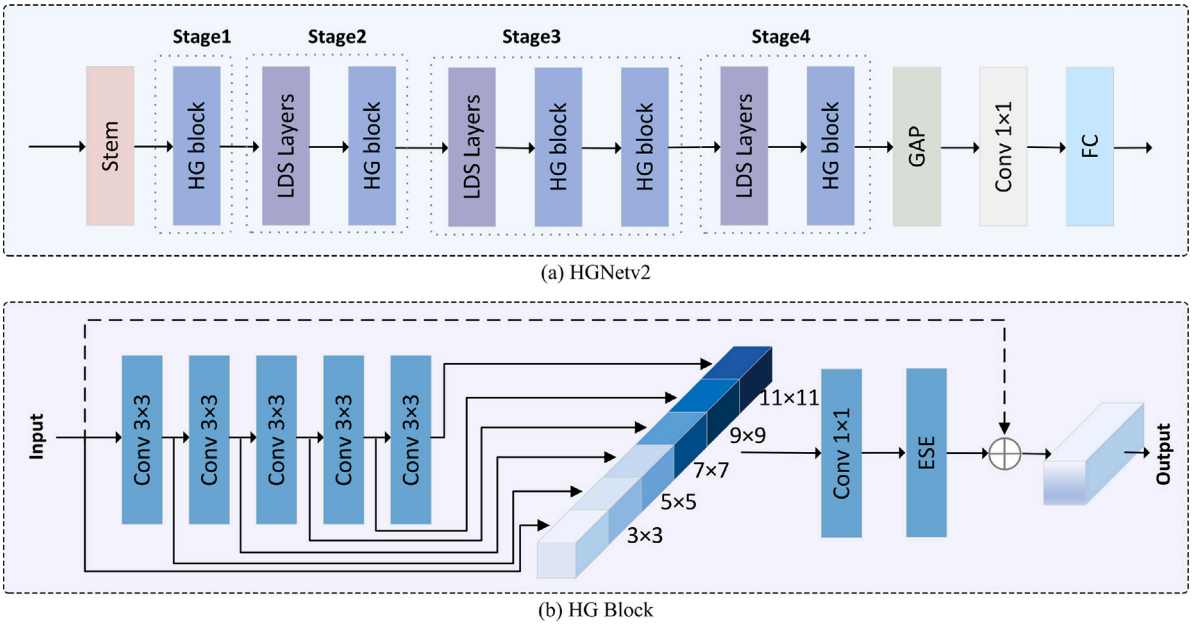


Fig. 2. Architecture of HGNetv2. (a) HGNetv2 includes 4 stages with coalition of LDS layers and HGBlock. (b) HGBlock is composed of multi-branch by stack of 3 × 3 convolution and then concatenated by 5 different kernel size convolution for further fusion.

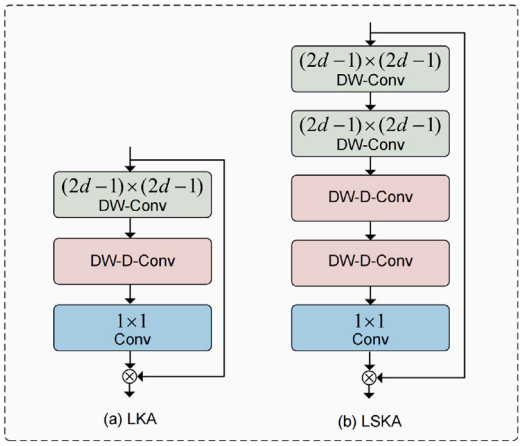


Fig. 3. Architecture of Large Separable Kernel Attention.

are then integrated through either summation or concatenation operations to form a unified feature map for the next network stage. Consequently, This hierarchical architecture captures intricate patterns across multiple scales and abstraction levels, significantly improving the model's ability to process complex visual data. The multi-scale feature extraction is especially beneficial for demanding tasks like image recognition, where detecting diverse patterns and multi-resolution features is critical. To optimize computational efficiency, we implemented channel slimming in HGNetv2's deep layers, reducing redundant parameters while preserving infrared feature extraction capabilities—a critical requirement for edge deployment.

3.3. Receptive field with Large Separable Kernel Attention

Modern deep learning-based object detection methods have demonstrated remarkable performance by leveraging multi-scale feature representations, which effectively capture visual information at varying resolutions. This approach proves particularly advantageous for detecting small objects in complex scenes [53]. Research indicates that early-stage multi-scale feature fusion primarily utilizes low-level visual cues, including edge features, texture patterns, color distributions, basic geometric shapes, and local structural information. While these low-level features effectively encode local image characteristics, they typically fail to capture global structural relationships and high-level semantic understanding. In contrast, deep-stage features derived from subsequent processing layers inherently incorporate richer semantic representations of objects through hierarchical feature abstraction.

Our approach incorporates Large Separable Kernel Attention (LSKA) modules in the early fusion stage to simultaneously achieve two critical objectives: (1) learning globally-aware semantic representations, and (2) facilitating dynamic cross-layer feature interactions. As shown in Fig. 3(b), LSKA is designed for model long-range dependencies and exhibits strong spatial and channel adaptability, allowing its scale to extreme large kernel, consequent offering a larger receptive field [54]. Large convolutional kernels can establish receptive fields that are functionally equivalent to those generated by self-attention mechanisms in terms of spatial coverage. Technically, a large convolution kernel can be built with relative less parameters and computation by using depth-wise (DW) convolution and dilated DW and consecutive 1×1 convolution. Like LKA attention, shown in Fig. 3(a), employs depthwise convolution for local feature extraction, complemented by large-kernel operations to establish long-range dependencies and mitigate gridding artifacts.

Given an input feature map $\mathbf{F} \in \mathbb{R}^{C\times H\times W}$, where C denotes the input channels, H and W represents the height ans width of feature map respectively. To mitigate the high computation cost by larger kernel size, an efficiency method is introduced to decompose the DW convolution with a small kernel followed by dilated DW convolution with a fairly large kernel, the output of LKA can be expressed by,

$$\bar{\mathbf{Z}}^C = \sum_{H,W} \mathbf{W}^C_{(2d-1)\times(2d-1)} * \mathbf{F}^C \tag{1}$$

$$\mathbf{Z}^C = \sum_{H,W} \mathbf{W}_{\left\lfloor \frac{k}{d} \right\rfloor \times \left\lfloor \frac{k}{d} \right\rfloor} * \bar{\mathbf{Z}}^C \tag{2}$$

$$\mathbf{A}^C = \mathbf{W}_{1\times1} * \mathbf{Z}^C \tag{3}$$

$$\bar{\mathbf{F}}^C = \mathbf{A}^C \circ \mathbf{F}^C \tag{4}$$

where d the dilation rate. * and ∘ denote convolution and Hadamard product respectively. $\bar{\mathbf{Z}}^C$ is the output of depth-wise convolution generated by kernel size $(2d-1)\times(2d-1)$. $\lfloor\cdot\rfloor$ is the floor operation.

By splitting the 2D weight kernels of original DW convolution and dilated DW convolution into two cascade separable weight kernels, we can obtain an parameter friendly LSKA block as shown in Fig. 3(b). Compared with LKA, LSKA could realize similar performance while being computationally efficient and enhances the long range dependency of the input image without incurring high computational and memory footprints. Meanwhile, LSKA achieves large-kernel modeling capabilities while retaining the runtime efficiency of LKA implementations, the output of LSKA can be expressed by follows,

$$\bar{\mathbf{Z}}^C = \sum_{H,W} \mathbf{W}^C_{(2d-1)\times1} * \left(\sum_{H,W} W^C_{1\times(2d-1)} * \mathbf{F}^C \right) \tag{5}$$

$$\mathbf{Z}^C = \sum_{H,W} \mathbf{W}^C_{\left\lfloor \frac{k}{d} \right\rfloor \times 1} * \left(\sum_{H,W} \mathbf{W}^C_{1\times\left\lfloor \frac{k}{d} \right\rfloor} * \bar{\mathbf{Z}}^C \right) \tag{6}$$

$$\mathbf{A}^C = \mathbf{W}_{1\times1} * \mathbf{Z}^C \tag{7}$$

$$\bar{\mathbf{F}}^C = \mathbf{A}^C \circ \mathbf{F}^C \tag{8}$$

3.4. Multi-scale feature aggregation via online convolutional re-parameterization

While deep layer features inherently encode critical abstract representations — including semantic information and contextual object relationships that are essential for inter-category discrimination — extracting these high-level features remains challenging. Conventional deep convolutional architectures face optimization difficulties in processing such abstract features, often resulting in inefficient training convergence.

To reduce the computational complexity of deep layer multi-scale feature fusion and simplify training workflow, we adopt online convolutional re-parameterization (OCRP) [55] to minimize the additional training overhead. OCRP, shown in Fig. 4, optimizes the convolution operation by dynamically adjusting the parameters of kernels, which minimizes redundant computations by employing direct optimization to adapt the kernels to current distribution of input data. Specifically, OCRP integrates dynamic convolutional kernel tuning with optimization techniques to efficiently reduce unnecessary computations and memory usage. Through efficient parametric transformations of high-level semantic feature, OCRP effectively models inter-object conceptual relationships while eliminating feature redundancy. This dual optimization simultaneously improves training efficiency and enhances object detection performance. Intermediate normalization layers are the essential elements for multi-layers and multi-branch structures in the re-parameterization, while all the intermediate operations in the re-parameterization block at the inference stage are linear, thus can be merged into one layer, producing a simpler network structure [56]. There needs two steps to realize OCRP: *Block Linearization* and *Block Squeezing*.

Block Linearization. During training stage, the intermediate layers may hinder the merging the separate layers, a scaling layers with a learnable vector was introduced to tackle this problem and scales

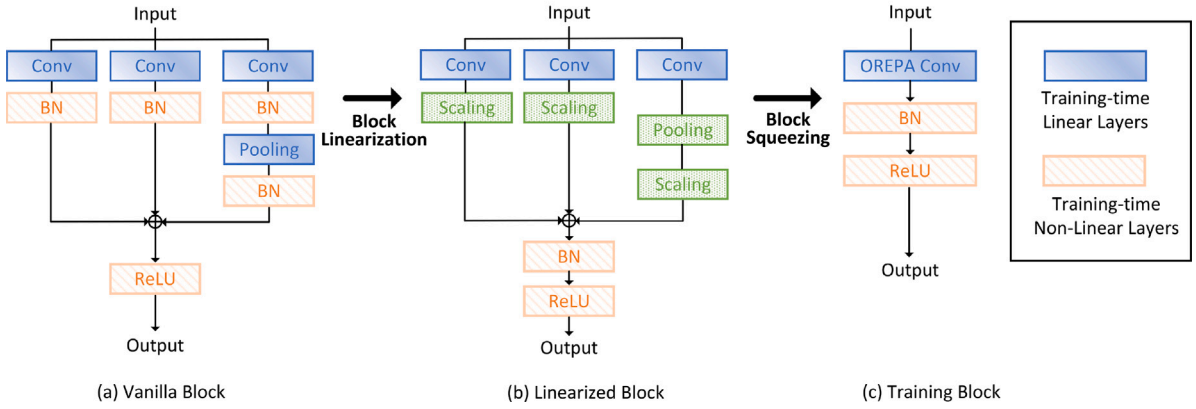


Fig. 4. Two stage pipeline of Online Re-Parameterization (OREPA). In the first stage (Block Linearization), non-linear components are removed from the basic block. In the second stage (Block Squeezing), the block are then merged into a single convolution layer.

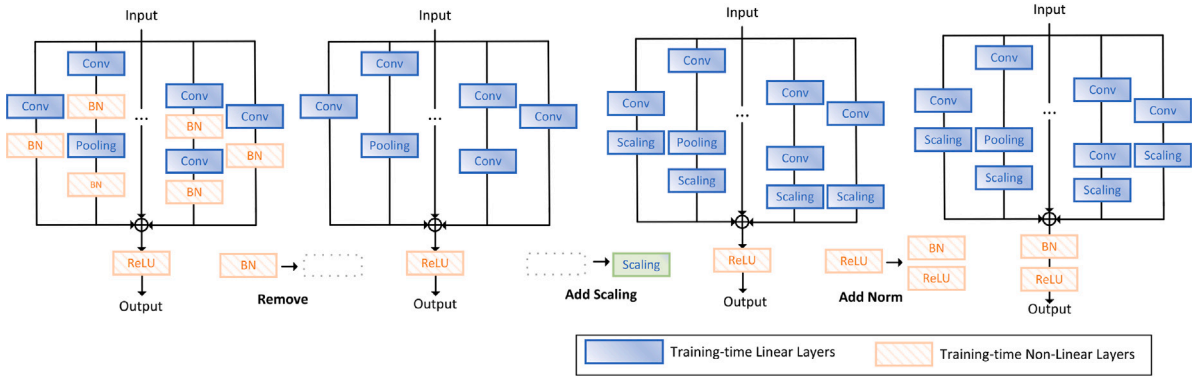


Fig. 5. Block linearization needs three steps: (1) Removing all the training-time non-linear norm layers. (2) Adding a linear scaling layer at the end of each branch. (3) Last, a post-normalization layer is added after each block for stabilizing training.

the features in the channel dimension. The Block Linearization process comprises three key steps: (1) Removal of all non-linear layers (e.g., normalization layers in re-parameterization blocks); (2) Introduction of a scaling layer to preserve optimization diversity; (3) Incorporation of a post-addition normalization layer following branch aggregation to ensure training stability, shown in Fig. 5.

Block Squeezing. After finishing the block linearization, the non-linear components were removed thoroughly, there only exists linear layers in the block, the block squeezing are used to squeeze the train-time linear block into a single convolution layer, which including two common structure types: sequential structure Fig. 6(a) and parallel structure Fig. 6(b). The process can be expressed by the following equations,

Let the C_i, C_o denote the input and output channel number with $K_H \times K_W$ kernel size, $\mathbf{X} \in \mathbb{R}^{C_i \times H \times W}$, $\mathbf{Y} \in \mathbb{R}^{C_i \times H \times W}$ mean the input and output result, for the basic convolutional feed forward process is denoted by,

$$\mathbf{Y} = \mathbf{W} * \mathbf{X} \tag{9}$$

Simplify a sequential structure. For a stack of convolutional layers,

$$\mathbf{Y} = \mathbf{W}_N(\mathbf{W}_{N-1} * \cdots (\mathbf{W}_2 * (\mathbf{W}_1 * \mathbf{X}))) \tag{10}$$

where $\mathbf{W}_j \in \mathbb{R}^{C_j \times C_{j-1} \times K_{H_j} \times K_{W_j}}$ is the weight of jth layer, satisfies $C_0 = C_i$, $C_N = C_o$.

$$\begin{aligned} \mathbf{Y} &= \mathbf{W}_N(\mathbf{W}_{N-1} * \cdots (\mathbf{W}_2 * (\mathbf{W}_1 * \mathbf{X}))) \\ &= \mathbf{W}_e * \mathbf{X} \end{aligned} \tag{11}$$

$\mathbf{W}_e$ denotes the end-to-end mapping matrix.

Simplify a parallel structure can be expressed by,

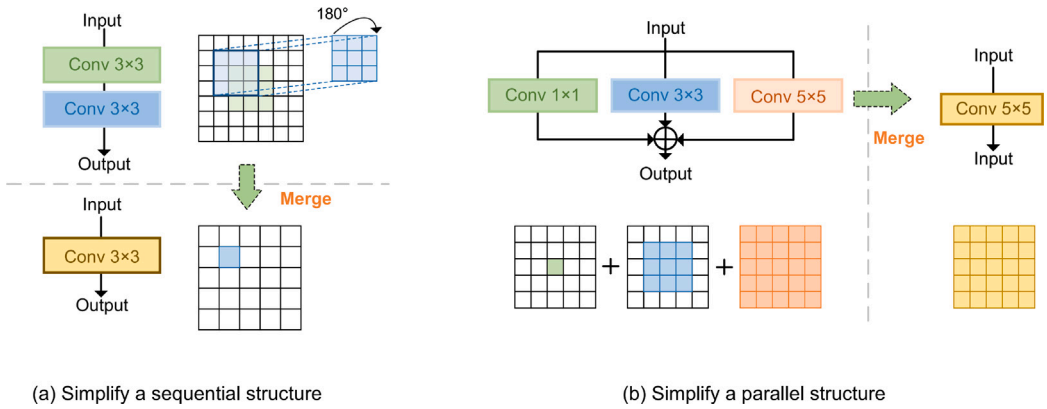
$$\mathbf{Y} = \sum_{m=0}^{M-1} (\mathbf{W}_m * \mathbf{X}) = (\sum_{m=0}^{M-1} \mathbf{W}_m) * \mathbf{X} \tag{12}$$

where $\mathbf{W}_m$ is the weight of the mth branch and the $\sum_{m=0}^{M-1} \mathbf{W}_m$ is the unified weight, and the spatial centers of different kernel sizes need be aligned.

3.5. Accurate detection head with multi-path coordinate attention

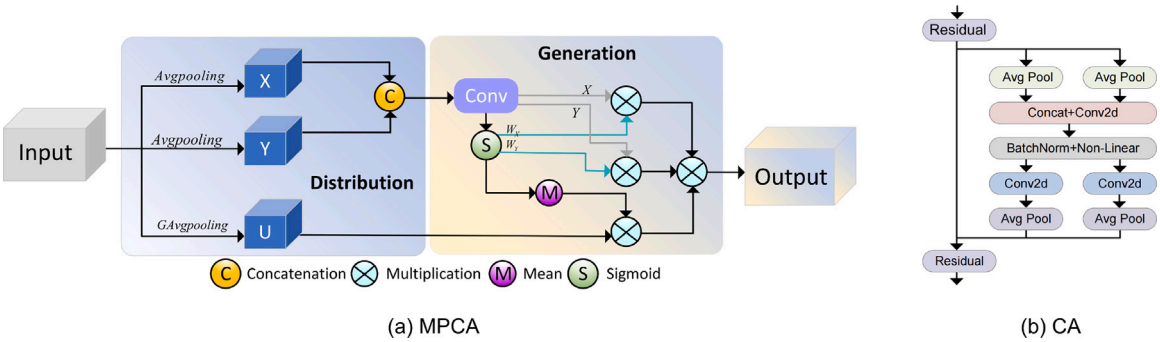
Prior to being processed by the detection head, the fused feature map often exhibits information redundancy due to the sequential stacking of 1 × 1 and 3 × 3 convolutional operations. This architectural characteristic not only introduces inevitable aliasing artifacts but also adversely impacts detection performance. To address this challenge, we propose a novel multi-path coordinate attention (MPCA) mechanism, specifically designed to mitigate aliasing artifacts in fused multi-scale feature representations, shown in Fig. 7(a). Different from original CA [57] attention shown in Fig. 7(b). MPCA dynamically assigns weights to channels, thereby highlighting salient features, suppressing redundancy and enhancing semantic representation. MPCA including two main multipath steps to generate the final outputs with enhanced denoising features: **Distribution** and **Generation**, which are employed for global information encoding and adaptive recalibration of channel relations.

Multi-Path Information Distribution. For any intermediate feature tensor $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_C] \in \mathbb{R}^{H \times W \times C}$ as input and output a transformed tensor with augmented representations $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_C]$ of the same



(a) Simplify a sequential structure (b) Simplify a parallel structure

Fig. 6. Two type structures of block squeezing: (a) simplification of sequential structure (b) simplification of parallel structures.



(a) MPCA (b) CA

Fig. 7. Architecture of attention mechanism. (a) Our proposed MPCA contains two steps for feature processing. Distribution: multiple pooling split the input into three branches with multi-dimension information. Generation: dimension X and Y are then multiplied with their refined weights respectively to produce distinctive patterns. (b) Coordinate attention owes two separate branches to refine features.

size to $\mathbf{X}$. Previous studies have shown vanilla convolution is difficult to build relations across channels, simply stacking the channel encoding information will result in model computation to the informative channels. By contrast, applying average pooling can facilitate the model in catching global knowledge and further reducing channel redundancy. Give the input $\mathbf{X}$, the information distribution step of MPCA can be formulated as follows:

$$\mathbf{z}_x = \frac{1}{k^2}\sum_{i=1}^{k}\sum_{j=1}^{k}\mathbf{x}_c(i,j) \tag{13}$$

$$\mathbf{z}_y = \frac{1}{k^2}\sum_{i=1}^{k}\sum_{j=1}^{k}\mathbf{x}_c(i,j) \tag{14}$$

where Eq. (13) and (14) denote the average pooling into x and y separate branch. c means the channel of the input $\mathbf{X}$.

$$\mathbf{z}_u = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}\mathbf{x}_c(i,j) \tag{15}$$

The above denotes global average pooling operation of the generated output $\mathbf{z}_u$, which is the output associated with cth channel.

Multi-Path Information Generation. This step aims to fully capture channel-wise dependencies and aggregates global spatial information with efficient excitation by multiplication among generated weights for modeling-channel interaction.

$$\begin{aligned}\mathbf{z}_\alpha &= Concat(\mathbf{z}_x,\mathbf{z}_y)\\ \mathbf{z}_{x'},\mathbf{z}_{y'} &= Conv(\mathbf{z}_\alpha)\end{aligned} \tag{16}$$

$$\mathbf{W}_x,\mathbf{W}_s = Sigmoid(Conv(\mathbf{z}_\alpha)) \tag{17}$$

$$\begin{aligned}\mathbf{z}''_x &= \mathbf{z}_{x'}\otimes\mathbf{W}_x\\ \mathbf{z}''_y &= \mathbf{z}_{y'}\otimes\mathbf{W}_y\end{aligned} \tag{18}$$

$$\mathbf{z}_\beta = Mean(Sigmoid(Conv(\mathbf{z}_\alpha)))\otimes\mathbf{z}_u \tag{19}$$

$$\mathbf{z}_\gamma = \mathbf{z}_\beta\otimes\mathbf{z}''_x\otimes\mathbf{z}''_y \tag{20}$$

where $Sigmoid$ and $Mean$ denote the non-linear activation function and the mean value of all the channel elements respectively, $\otimes$ means the element multiplication.

MPCA enhances the expression of key features while suppressing redundant information by dynamically adjusting the feature weights of each channel. This process improves the model's performance and robustness. It allows the network to flexibly modify the importance of each channel based on the specific task and input image, thereby enhancing the efficiency and accuracy of feature extraction.

4. Experiments

4.1. Datasets and experimental settings

To evaluate the effectiveness of our proposed method on infrared pedestrian detection task, we have adopted two public datasets to test the performance of IPeDet.

HIT-UAV [58]: This dataset is specifically designed for UAV high-altitude object detection with infrared scenario, which contains 2898 infrared images extracted from 43,470 frames and is capable of capturing thermal radiation information from targets under different lighting conditions (day and night) to effectively identify potential objects. There are 2008 images in the training set, 287 images in the validation set and 571 images in the test set, with main 5 categories including People, Bicycles, Cars, OtherVehicles and DonCare.

M^3FD [59]: This dataset is developed for a multimodal image fusion and target detection with rich annotation information and diverse

Table 1
Overall comparison with the End-to-End YOLO algorithms on the HIT-UAV *test* set.

Methods	*Precision*	*Recall*	*mAP*@50	*mAP*@50 : 95	*Params.*(*M*)	*FLOPs*(*G*)
YOLOv5-N [60]	0.871	0.766	0.838	0.526	1.76	4.1
YOLOv5-S [60]	0.889	0.784	0.845	0.546	7.02	15.8
YOLOv5-M [60]	0.862	0.885	0.841	0.553	20.86	47.9
YOLOv5-L [60]	0.889	0.813	0.843	0.553	46.12	107.7
YOLOv6-N [61]	0.839	0.726	0.793	0.503	4.63	11.34
YOLOv6-S [61]	0.845	0.768	0.815	0.512	18.5	45.17
YOLOv7-T [62]	0.825	0.729	0.793	0.491	6	13.1
YOLOv7 [62]	0.777	0.725	0.75	0.466	36.5	103.2
YOLOv8-N [63]	0.796	0.805	0.817	0.543	3	8.1
YOLOv8-S [63]	0.891	0.8	0.846	0.565	11.12	28.4
YOLOv9-T [64]	0.827	0.732	0.777	0.505	2.61	10.7
YOLOv9-S [64]	0.831	0.768	0.821	0.558	9.6	38.7
YOLOv10-N [65]	0.829	0.756	0.801	0.53	2.26	6.5
YOLOv10-S [65]	0.903	0.792	0.848	0.544	8.03	24.5
YOLOv10-M [65]	0.849	0.802	0.838	0.557	24.2	16.4
YOLO11-N [28]	0.863	0.788	0.82	0.535	2.58	6.3
YOLO11-S [28]	0.859	0.785	0.842	0.563	9.4	21.3
YOLOv12-N [29]	0.862	0.715	0.797	0.517	2.55	6.3
YOLOv12-S [29]	0.86	0.744	0.819	0.542	9.23	21.2
IPeDet	0.872	0.796	**0.852**	**0.576**	12.1	15.7

scenes, which is applicable to monitoring and automatic driving and other fields, especially for object detection tasks in low-light conditions or complex environments, such as night-time driving safety, surveillance systems, UAV navigation, etc. The dataset contains 8400 images paired for fusion and detection, including 34,407 tokens that have been manually tagged with 6 categories: People, Car, Bus, Motorcycle, Lamp and Truck. In this work, we adopt 4200 infrared images for pedestrian detection with 3360 images for training and 840 images for testing.

Our experiments were conducted under CPU Intel Core i7-13700KF and one single GPU NVIDIA RTX 4090, with python version 3.10, deep learning framework pytorch version 2.3 and cuda toolkit 12.1. We have set 200 epochs during training for IPeDet, with image input size 640 × 640, batch size 32, learning rate 0.01, momentum 0.937 and weight decay 0.0005 during training stage, the batch size was set to 1 for fast inference processing, other specific hyperparameters remain consistent with the default settings of the YOLOv10 framework.

4.2. Evaluation metrics

To verify the effectiveness of our proposed IPeDet in the pedestrian detection task, we utilize six primary evaluation metrics to comprehensively assess the model: Precision Rate (P), Recall Rate (R), mean Average Precision (mAP) at IoU threshold from 0.50 to 0.95, number of Parameters (params.), Floating Point Operations (FLOPs),

$$P = \frac{TP}{TP + FP} \tag{21}$$

$$R = \frac{TP}{TP + FN} \tag{22}$$

$$AP_i = \int_0^1 P_i(R_i) dR_i \tag{23}$$

$$mAP = \frac{1}{n} \sum_{i=1}^{n} AP_i \tag{24}$$

where true positive (TP), false positive (FP) and false negative (BN) bounding box samples are vital metrics. Average Precision (AP) measures the area under the precision–recall (P–R) curve, precisely reflects model detection performance. mean Average Precision (mAP) gathers the average precision across all the categories, offering a comprehensive evaluation of detection model.

Table 2
Comparison with the mainstream detectors on HIT-UAV *test* set.

Methods	*mAP*@50	*mAP*@50 : 95	*Params.*(*M*)	*FLOPs*(*G*)
Faster RCNN [20]	0.902	0.55	41.36	168
Cascade RCNN [66]	0.82	0.453	32.12	161
Deformable DETR [67]	0.788	0.398	40.1	154
Sparse RCNN [68]	0.609	0.321	106	122
DINO [69]	0.865	0.526	47.55	221
RT-DETR [52]	0.73	0.442	41.9	125.6
HyperYOLOs [70]	0.851	0.575	14.81	38.9
PHSI-RTDETR [71]	0.826	0.516	13.87	47.5
CSFPR-RTDETR [72]	0.831	–	14.09	63.6
Freq-DETR [73]	0.816	0.527	–	80
DFAS-YOLO [74]	0.785	0.541	7.52	–
DEIM-HGNetv2s [75]	0.826	0.557	10.1	24.8
IPeDet	0.852	**0.576**	**12.1**	**15.7**

4.3. Experimental analysis

Table 1 presents a comparative analysis of our proposed IPeDet against mainstream end-to-end YOLO series models on the HIT-UAV *test* set. It is evident that IPeDet achieves top-tier detection accuracy, recording 0.852 and 0.576 for *mAP*@50 and *mAP*@50:95, respectively, while maintaining a compact footprint of only 12.1M parameters and 15.7G FLOPs. While the super-lightweight YOLOv5-N 1.76M parameters achieves an *mAP*@50:95 of 0.526, a significant performance gap remains compared to IPeDet. Similarly, although YOLOv5-M achieves a comparable 0.533 its parameter count exceeds 20M—substantially larger than that of IPeDet. Compared to YOLOv8-N and YOLOv8-S, IPeDet improves detection accuracy by 6.1% and 1.95%, respectively. Furthermore, YOLOv9-S and YOLOv9-T yield *mAP*@50:95 scores of 0.558 and 0.505, falling behind IPeDet by 3.22% and 14.1%, respectively. Against the baseline YOLOv10-S, IPeDet demonstrates a 5.88% improvement on the strict *mAP*@50:95 metric while reducing model complexity by 35.9%. Finally, IPeDet outperforms the recent YOLO11N/S and YOLO12N/S models across the first four key evaluation metrics. In conclusion, IPeDet delivers superior accuracy for infrared pedestrian detection while maintaining an efficient architecture with competitive model complexity.

Table 2 lists the detection results between IPeDet and prevailing general object detectors. Comparative studies reveal that our method consistently outperforms existing approaches on all evaluation criteria. Specifically, IPeDet greatly improve the detection accuracy on *mAP*@50:95 by 4.72% and 27.2% with 12.1M parameters compared with two-stage detector Faster RCNN and Cascade RCNN. Compared

Table 3
Experimental results towards lightweight CNNs backbones on HIT-UAV *test* set.

Methods	*Precision*	*Recall*	*mAP*@50	*mAP*@50 : 95	*Params.*(*M*)	*FLOPs*(*G*)
MobileNetv4 [76]	0.838	0.67	0.77	0.502	47.2	27.7
EfficientViT [77]	0.868	0.74	0.791	0.517	7.9	18.3
ShuffleNetv2 [78]	0.842	0.757	0.801	0.518	6.42	16.3
MobileViTxxs [45]	0.893	0.711	0.819	0.52	7.16	18.9
GhostNet [43]	0.835	0.724	0.796	0.52	9.22	19.5
HGNetv2-CS	0.853	0.745	0.81	**0.532**	**4.17**	19.5

Table 4
Experimental results the choices of detection heads with lightweight design for HGNetv2 on HIT-UAV *test* set.

Methods	*Precision*	*Recall*	*mAP*@50	*mAP*@50 : 95	*Params.*(*M*)	*FLOPs*(*G*)
HGNetv2 [52]	0.903	0.751	0.808	0.527	5.93	21
HGNetv2-CS	0.853	0.745	0.81	0.532	4.17	19.5
HGNetv2-CS+2HS(P3+P4)	0.903	0.723	0.812	0.529	3.92	18.7
HGNetv2-CS+2HS(P3+P5)	0.87	0.747	0.815	0.53	3.75	19.2
HGNetv2-CS+4HS(P2)	0.854	0.767	0.8	0.52	4.73	34.4
HGNetv2-CS+4HS(P6)	0.862	0.766	**0.827**	**0.535**	4.6	19.3

Table 5
Inter-relation of shallow stage layers for processing multi-scale feature fusion on HIT-UAV *test* set.

Methods	*Precision*	*Recall*	*mAP*@50	*mAP*@50 : 95	*Params.*(*M*)	*FLOPs*(*G*)
EMA [79]	0.802	0.794	0.824	0.544	6.8	14.3
SE [80]	0.862	0.793	0.839	0.546	6.79	13
CBAM [81]	0.861	0.743	0.816	0.55	7	13
GSConv [82]	0.828	0.777	0.826	0.55	6.84	13.2
DyConv [83]	0.897	0.749	0.826	0.551	7.4	12.8
LSKA	0.898	0.768	0.833	**0.558**	6.97	14.2

Table 6
Experimental results of the deep stage multi-scale feature fusion methods on HIT-UAV *test* set.

Methods	*Precision*	*Recall*	*mAP*@50	*mAP*@50 : 95	*Params.*(*M*)	*FLOPs*(*G*)
RFAConv [84]	0.811	0.698	0.783	0.512	8.86	15.4
AKConv [85]	0.814	0.695	0.778	0.509	8.78	15
CAFM [86]	0.807	0.714	0.782	0.508	8.64	15.1
FasterC2f [44]	0.816	0.702	0.777	0.502	8.23	14.8
OCRP	**0.893**	**0.76**	**0.839**	**0.566**	8.4	15.1

with latest UAV-oriented detectors, IPeDet still exhibits competitive performance regarding across main evaluation metrics, exhibiting obvious benefits against DETR detectors DEIM. The results indicates that IPeDet operates as lightweight style with high detection performance and is competent for efficient and accurate infrared detection.

Lightweight model architecture serves as a critical enabler for efficient deployment on computation-constrained platforms. We report basic feature extraction detection results with the current prevailing lightweight CNNs architecture and our channel slimmed HGNetv2 termed as HGNetv2-CS in Table 3. Compared with latest MobileNet variant MobileNetv4, HGNetv2-CS achieves 5.98% improvements at mAP@50:95 with only around 4M parameters. We also conducted experiments to testify the effectiveness of attention-based ViT methods. For example, EfficientViT and MobileViTxxs produced 0.517 and 0.52 at mAP@50:95 respectively, which exists a considerable disparity between CNNs based HGNetv2-CS. Our evaluation reveals that the channel-slimmed HGNetv2-CS preserves robust feature extraction performance, effectively handling the complexities of infrared pedestrian images.

Table 4 lists the experimental results towards choices of detection head of IPeDet. Channel slimming was strategically applied to the deeper layers of HGNetv2 to achieve model compression. However, the detection accuracy was boosted to 0.527 from 0.532 at mAP@50:95 and parameters were reduced from 5.93M to 4.17M. We have tested choices by utilizing only 2 heads for detection, such as combination noted of (P3 80 × 80 + P4 40 × 40) and (P3 80 × 80 + P5 20 × 20) based on the HGNetv2-CS, accomplishing unpleasant improvement at mAP@50:95. Consequently, we have attempted by adding one extra head P2 (160 × 160) or P6 (10 × 10) in total of 4 heads to perform detection. Experiments results indicates that HGNetv2-CS with 4 HS (P6) achieves the best detection results at mAP@50:95 0.535 over 0.527 baseline, which also behaves at relative lower model parameters 4.6M and complexity 19.3G. Our experiments demonstrate that auxiliary smaller heads effectively improve small infrared target detection by fusing local details with global context, countering the challenges of weak signatures and background clutter in thermal images.

Table 5 shows the multi-scale feature fusion processing during shallow stage layers. Several mainstream attention methods were adopted to testify the efficiency for feature exploration. Channel refinement based SE attention realize detection accuracy 0.546 at mAP@50:95. By Contrast, CBAM processes the features from channel and spatial dimension, generating slight higher detection accuracy with 0.55 than SE attention. Large kernel designed LSKA obtained 0.558 at mAP@50:95, explaining that the detailed spatial information, such as contour, texture and edges could be fully exploited by large receptive field. Moreover, LSKA outperforms other attention mechanisms in the main evaluation metrics, all while operating with a significantly lower computational overhead.

Table 6 list the experimental results towards deep layers multi-scale feature optimization methods. OCRP strategically simplifies complex feature learning by online reparameterization specifically for deep feature processing, thereby simultaneously reducing the computational

Table 7
Experimental results optimization for detection head on HIT-UAV *test* set.

Methods	*Precision*	*Recall*	*mAP*@50	*mAP*@50 : 95	*Params.*(*M*)	*FLOPs*(*G*)
SPD [87]	0.854	0.587	0.677	0.458	24.82	18.1
CAAHead [88]	0.828	0.706	0.787	0.511	10.9	16.7
MixedConv [89]	0.842	0.686	0.781	0.516	12.6	18.9
SEHead [80]	0.809	0.72	0.792	0.519	10	15.6
MPCAHead	**0.872**	**0.796**	**0.852**	**0.576**	12.1	15.7

Table 8
Ablation study of the improvement modules on HIT-UAV *test* set.

Baseline	HGNetv2-CS	LSKA	OCRP	MPCA	*mAP*@50	*mAP*@50 : 95	*Params.*(*M*)	*FLOPs*(*G*)
✓					0.848	0.544	8.03	24.5
✓	✓				0.827	0.535	4.6	19.3
✓	✓	✓			0.833	0.558	6.97	14.2
✓	✓		✓		0.829	0.551	5.3	13.1
✓	✓	✓	✓		0.839	0.566	8.4	15.1
✓	✓	✓	✓	✓	0.852	0.576	12.1	15.7

cost of multi-scale feature extraction and alleviating the optimization challenges inherent in deep semantic feature learning. Compared with RFAConv, AKConv, CAFM and FasterC2f, OCRP improves the detection by 10.54%, 11.2%, 11.4% and 12.75% at $mAP@50{:}95$ respectively, and also show leading detection performance at *Precision*, *Recall* and $mAP@50$ evaluation metrics with lightweight model scale.

Table 7 lists the experimental results of optimization strategy for the detection head. As for the proposed IPeDet, we devise 4 different scales head 80×80, 40×40, 20×20 and 10×10 to fully cover various object size. We adopt MPCA embedded into 4 heads to reduce the information aliasing effects caused by the previous feature fusion step. Specifically, MPCA realizes 0.576 at $mAP@50{:}95$, 12.72%, 11.63% and 10.98% higher compared with attention method CAAHead, MixedConv and SEHead respectively. MPCA aggregates fused features through multi-branch recalibration and residual distributed projection, significantly mitigating aliasing effects between layers and ensuring stable training and inference.

We have conducted ablation study to further evaluate the effectiveness of our proposed module with IPeDet, the results were shown in Table 8. For the baseline, the detection accuracy of $mAP@50{:}95$ is 0.544, after employing HGNetv2-CS as the main backbone and with extra detection head P6, the model scale of parameters and complexity were reduced by 42.7% and 21.2% respectively. Notable improvements can be obtained after LSKA block were embedded with the shallow inter-layers, implying the spatial context features benefit from large kernel receptive field with extra model complexity decrease. OCRP functions as structural re-parameterization has boost the efficiency of fusion process, reaching the detection accuracy to 0.566 at $mAP@50{:}95$. Finally, MPCA was introduced to optimize detector with more accurately object localization and classification, boosting the detection to 0.576 at $mAP@50{:}95$, 5.88% improvement compared to baseline. The ablation study confirms that the proposed modules are essential to IPeDet, enabling the model to achieve satisfactory performance in pedestrian detection tasks.

To verify the robustness of our proposed method, we conducted additional experiments on the M^{3}FD *val* dataset. Table 9 presents a comparison of the detection results between IPeDet and the YOLO series on the validation set. Notably, IPeDet achieved the highest accuracy, with an $mAP@50{:}95$ of 0.545, outperforming most YOLO variants. Furthermore, IPeDet demonstrates an excellent trade-off between detection accuracy and model efficiency (in terms of parameters and complexity). In contrast, the recently released YOLO11 and YOLOv12 variants failed to yield satisfactory results across these evaluation metrics. Table 10 presents a comparative analysis between IPeDet and mainstream object detectors. The Transformer-based RT-DETR achieves an accuracy of 0.483 at $mAP@50{:}95$, while its substantial parameter count and computational complexity make it unsuitable for lightweight applications. Similarly, the two-stage Faster RCNN yields lower detection accuracy across both primary evaluation metrics. These results further demonstrate that IPeDet achieves high performance in infrared detection tasks while maintaining strong robustness across related infrared vision scenarios.

Ablation study was also reported on the M^{3}FD *val* set to test the effectiveness of each module in the IPeDet, shown in Table 11. When adopted the HGNetv2-CS as the main backbone, the model parameter and complexity dropped significantly compared with the baseline. Meanwhile, the detection accuracy both from $mAP@50$ and $mAP@50{:}95$ are all presented growth to some extent, implying that lightweight model is suitable for dealing with infrared image features. Same growth also can be observed when adding the LSKA, OCRP and MPCA gradually to construct IPeDet, the final detection accuracy were reached to 0.815 and 0.545 at $mAP@50$ and $mAP@50{:}95$, realizing 3.69% and 6.23% compared with baseline. This ablation study also testify the feasibility and robustness of each essential component of our proposed method, showcasing the IPeDet could generalized well on the pedestrian detection task.

4.4. Quantitative visualization

Fig. 8 shows the visualization comparison of IPeDet with real-time SOTA methods. It can be observed that IPeDet exhibits competitive performance regarding the main evaluation metrics while keeping lower model complexity. With inference speed 218 FPS, IPedet meets the standard of real-time detection, indicating a great potential when deployed on real world applications. We present visualization comparison between IPeDet and baseline YOLOv10s in Fig. 9 on the HIT-UAV *test* set. While YOLOv10s demonstrates balanced performance in general pedestrian detection, its efficacy significantly degrades for small targets in infrared imagery, particularly under low-contrast conditions characteristic of thermal environment. As shown in the dotted circles, IPeDet enhances the perception of small targets by introducing a specially well designed feature enhancement module. More visualization examples with real-time methods YOLOv8s, YOLOv10s and YOLO11s are shown in Fig. 12 on the HIT-UAV and Fig. 13 on the M^{3}FD dataset respectively. We can be informed that IPeDet could classify and locate objects accurately with much higher confidence threshold compare with baseline. The model pays special attention to the low-contrast information in infrared images and adopts advanced context-aware techniques, which can effectively extract and recognize small targets in complex backgrounds.

Table 9
Overall comparison with the End-to-End YOLO algorithms on the M^{3}FD *val* set.

Methods	*Precision*	*Recall*	*mAP*@50	*mAP*@50 : 95	*Params.*(*M*)	*FLOPs*(*G*)
YOLOv5-N [60]	0.82	0.666	0.737	0.444	1.76	4.2
YOLOv5-S [60]	0.887	0.715	0.808	0.508	7	15.8
YOLOv6-N [61]	0.801	0.61	0.669	0.431	4.63	11.34
YOLOv6-S [61]	0.823	0.657	0.716	0.463	18.5	45.18
YOLOv7-T [62]	0.932	0.632	0.72	0.439	6	13.1
YOLOv7 [62]	0.875	0.756	0.822	0.519	36.5	103.2
YOLOv8-N [63]	0.831	0.668	0.738	0.478	3	8.1
YOLOv8-S [63]	0.864	0.725	0.899	0.825	11.1	28.4
YOLOv9-T [64]	0.832	0.604	0.681	0.459	2.6	10.7
YOLOv9-S [64]	0.832	0.704	0.768	0.518	9.6	38.7
YOLOv10-N [65]	0.826	0.676	0.748	0.493	2.69	8.2
YOLOv10-S [65]	0.837	0.722	0.786	0.513	8	24.5
YOLO11-N [28]	0.803	0.658	0.735	0.479	2.58	6.3
YOLO11-S [28]	0.837	0.727	0.79	0.523	9	21.3
IPeDet	0.81	0.755	0.815	**0.545**	12.1	15.7

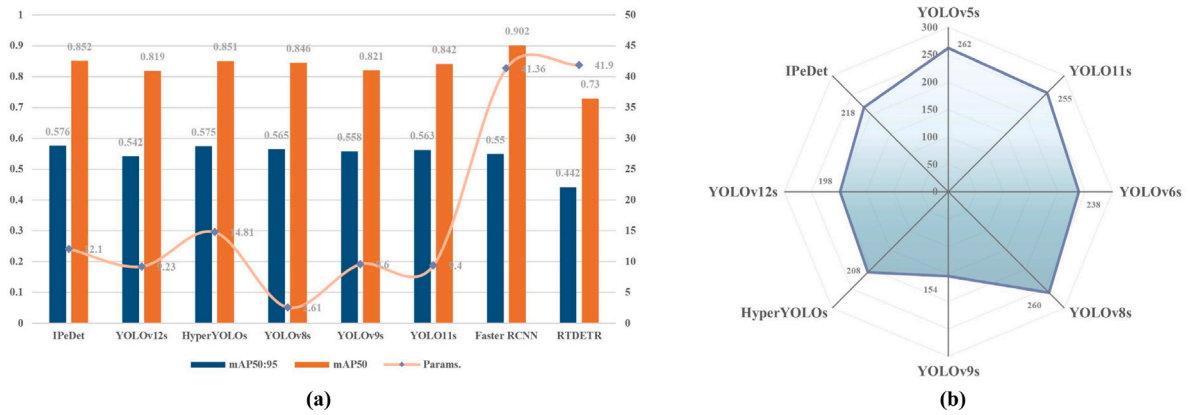


Fig. 8. Detection performance comparison with real-time SOTA methods. (a) The evaluation metrics of *mAP*@50, *mAP*@50:95 and *Params*. (b) Radar graph of FPS.

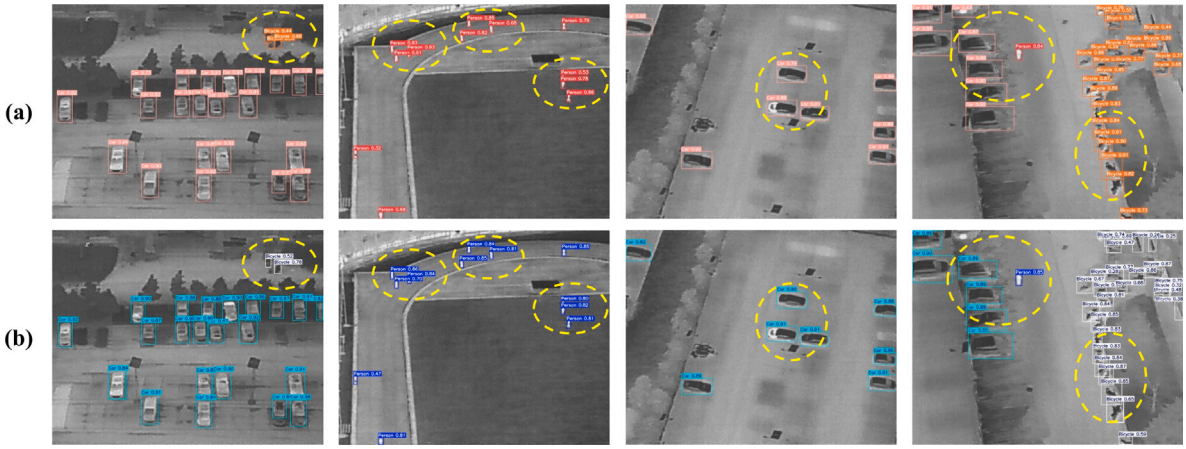


Fig. 9. Visualization of detection results on HIT-UAV *test* set. (a) The first row presents baseline (b) The second row of IPeDet.

To better perceive the understanding of infrared object of our proposed method, we visualize the comparison results between baseline and IPeDet through class activation map generated by Grad-CAM [92]. Fig. 10 row (a) shows the results of baseline on the HIT-UAV *test* set. It is clearly that row (b) from IPeDet could distinguish well on the dense scattered infrared pedestrian objects and the heat response regions focus accurately to the tiny objects. Fig. 11 shows the experimental on the M^{3}FD set, similarity, the heat map of baseline row (a) presented unpleasant results, most of regions are labeled imprecisely and some infrared pedestrian are omitted to a large extent. On the contrary, row (b) results from IPeDet is capable to perceive accurate response to pedestrian targets, indicating that IPeDet could operate with high performance on the infrared recognition task.

5. Limitations and future work

While IPeDet strikes a favorable balance between accuracy and efficiency, there are limitations that motivate our future work. Firstly,

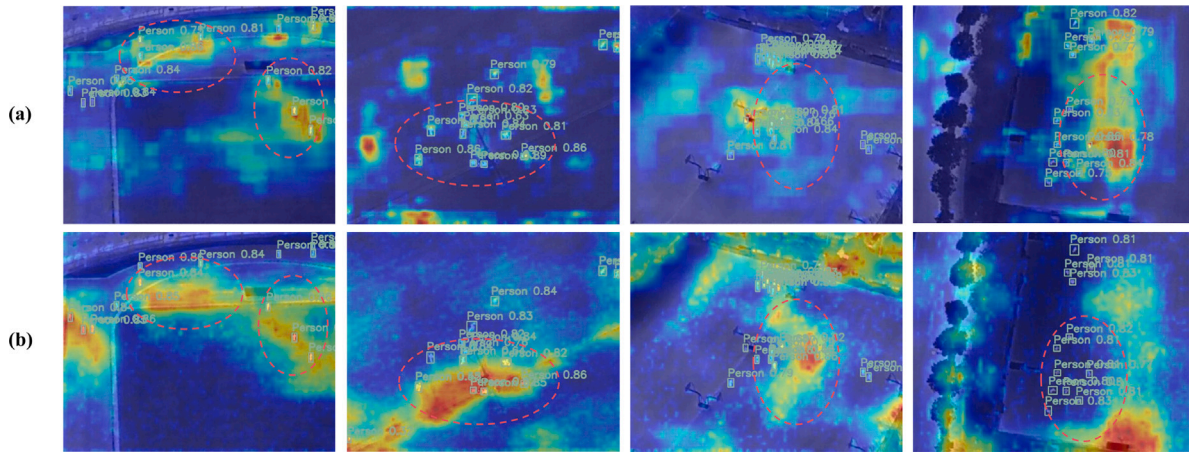


Fig. 10. Visualization of class activation maps by Grad-CAM on HIT-UAV *test* set. (a) The first row presents baseline. (b) The second row of IPeDet.

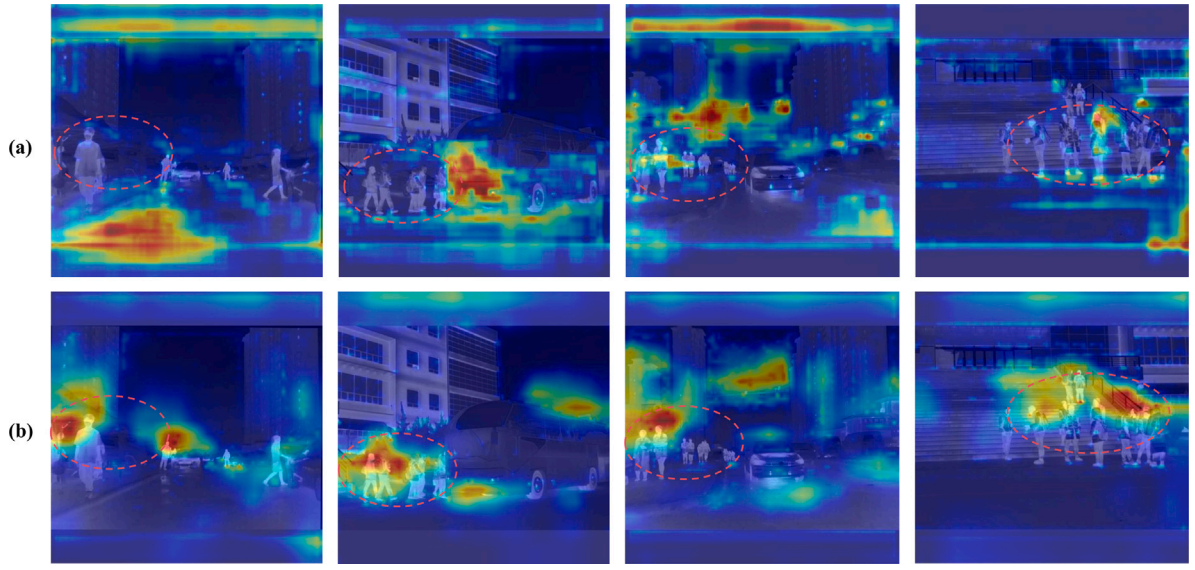


Fig. 11. Visualization of class activation maps by Grad-CAM on M^{3}FD val set. (a) The first row presents baseline. (b) The second row of IPeDet.

Table 10
Comparison with the mainstream algorithms on M^{3}FD *val* set.

Methods	$mAP@50$	$mAP@50:95$	$Params.(M)$	$FLOPs(G)$
Faster RCNN [20]	0.689	0.413	41.37	178
DCNv2 [90]	0.696	0.42	149	204
DINO [69]	0.762	0.454	47.55	235
Sparse RCNN [68]	0.532	0.309	106	130
FCOS [91]	0.337	0.179	32.12	167
RT-DETR [52]	0.759	0.483	41.94	125.6
IPeDet	**0.815**	**0.545**	**12.1**	**15.7**

our current evaluation is limited to specific pedestrian datasets. Future research will aim to enhance the model's generalization by incorporating more diverse scenes and expanding detection categories beyond pedestrians. Secondly, to address the challenge of data annotation in thermal imaging, we plan to explore few-shot learning strategies to enable robust training with limited samples. Lastly, although our experiments on the NVIDIA RTX 4090 validate the model's theoretical efficiency, practical deployment on edge devices remains a critical next step. We intend to further optimize the model using quantization and knowledge distillation to ensure real-time performance on resource-limited embedded hardware in real-world UAV applications.

6. Conclusion

In this work, we propose a high-accuracy pedestrian detection framework tailored for UAV infrared imagery. Specifically, we design the HGNetv2-CS as the fundamental backbone using a channel-slimming approach to ensure efficiency. To effectively capture substantial spatial features during the shallow fusion stage, we employ large separable kernel attention, which introduces minimal parameter overhead. Furthermore, an online convolution re-parameterization strategy is adopted to optimize multi-scale feature fusion in the deep stage and stabilize the training process. We also introduce a novel channel attention module MPCA, to mitigate the aliasing effects caused by preceding fusion operations. Extensive experiments validate the effectiveness of our proposed method; IPeDet achieves detection accuracy improvements of 5.88% and 6.23% on infrared detection dataset HIT-UAV and M^{3}FD, respectively. Overall, IPeDet demonstrates superior performance while maintaining a compact model scale, effectively striking

Table 11
Ablation study of the improvement modules on M^{3}FD *val* set.

Baseline	HGNetv2-CS	LSKA	OCRP	MPCA	$mAP@50$	$mAP@50:95$	$Params.(M)$	$FLOPs(G)$
✓					0.786	0.513	8	24.5
✓	✓				0.808	0.532	6.8	13.1
✓	✓	✓			0.809	0.538	6.97	14.2
✓	✓	✓	✓		0.811	0.541	8.43	15.1
✓	✓	✓	✓	✓	0.815	0.545	12.1	15.7

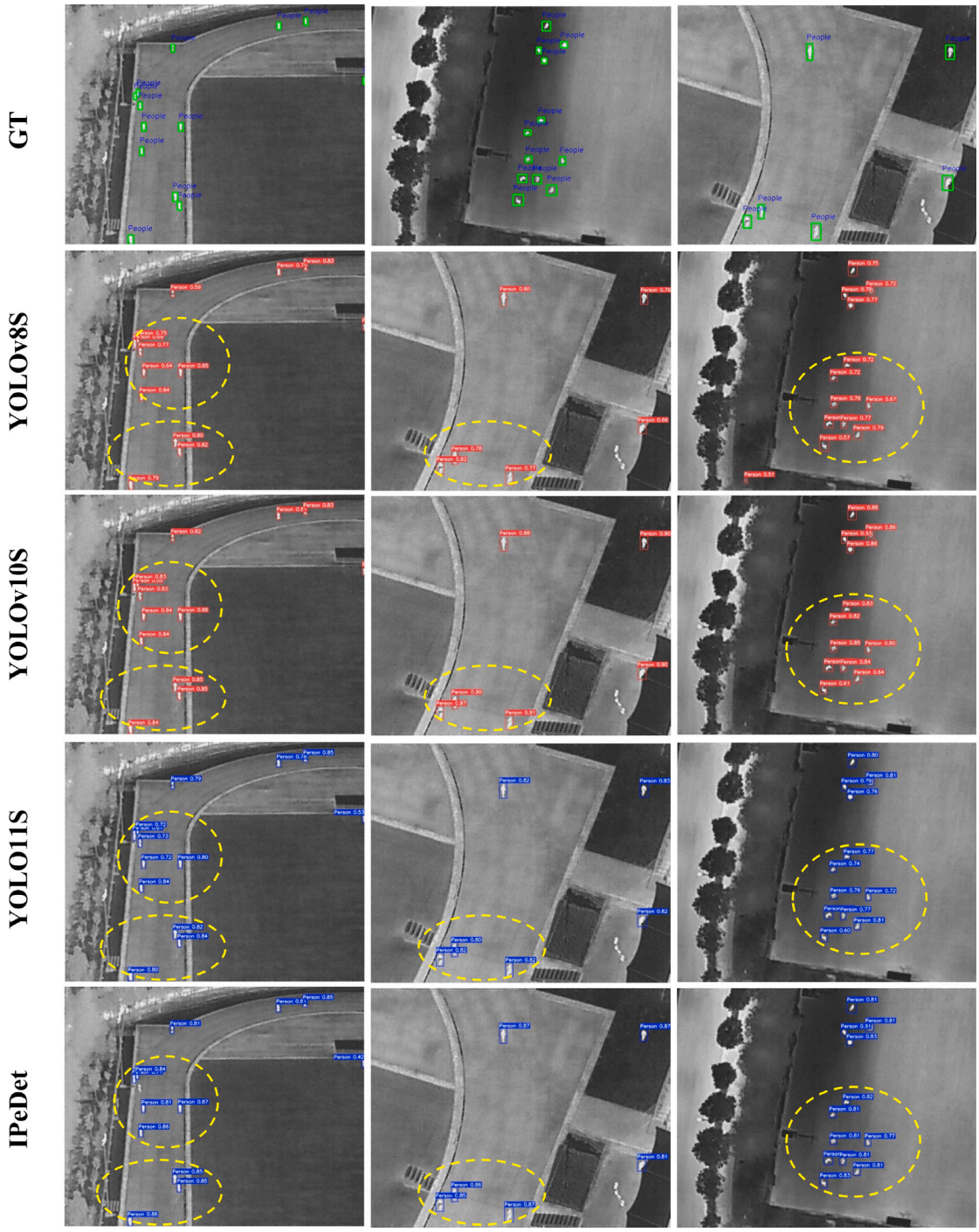


Fig. 12. Visualization comparison of End-to-End mainstream detectors on HIT-UAV *test* set.

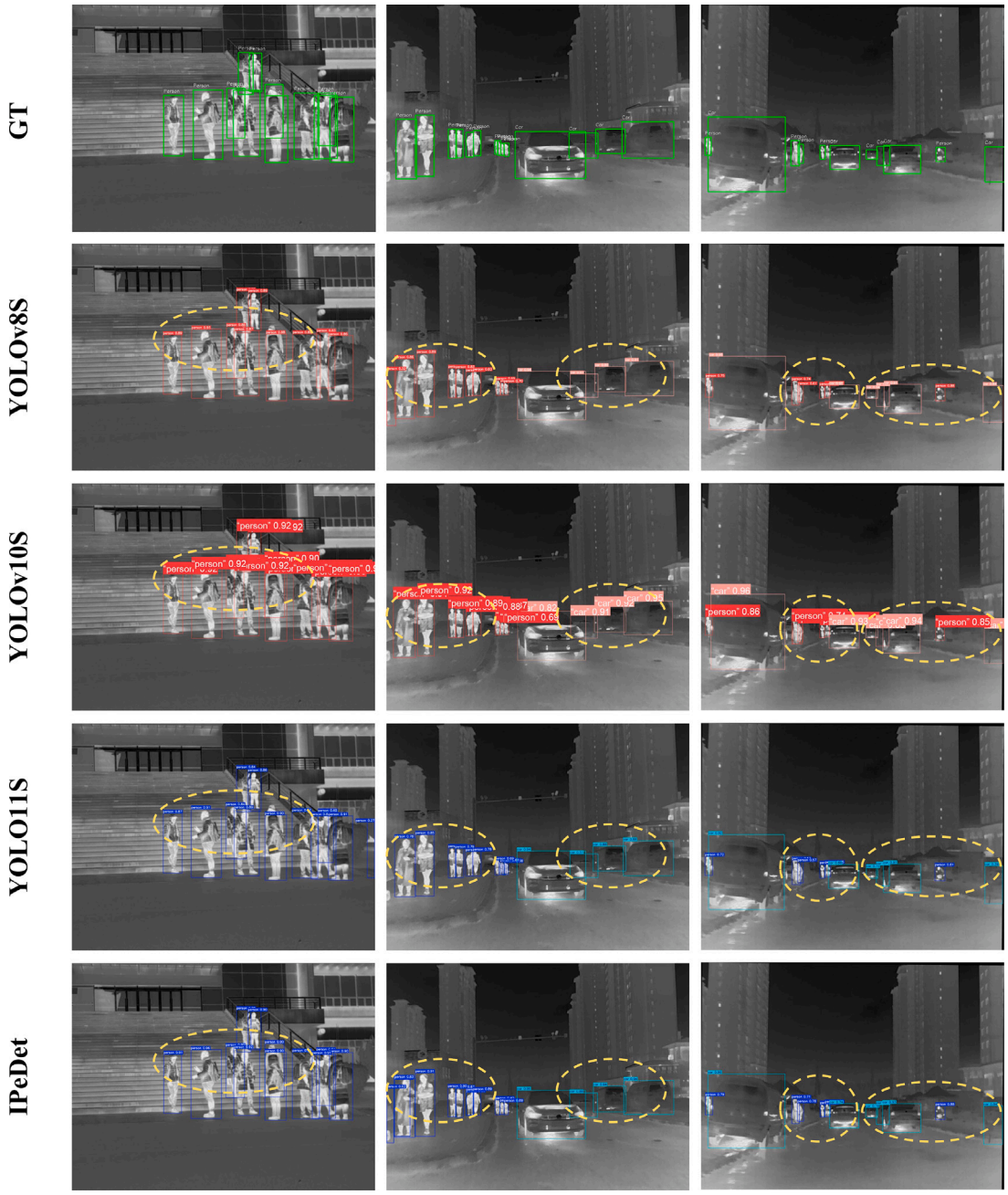


Fig. 13. Visualization comparison of End-to-End mainstream detectors on M^3FD *val* set.

a balanced trade-off between detection accuracy and computational complexity.

CRediT authorship contribution statement

Yi Li: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Huiying Xu:** Writing – review & editing, Visualization, Validation, Methodology, Investigation. **Xinzhong Zhu:** Validation, Supervision, Software, Investigation, Funding acquisition, Formal analysis. **Hongbo Li:** Validation, Supervision, Software, Resources, Investigation. **Yiming Sun:** Validation, Supervision, Project administration, Investigation, Formal analysis, Data curation. **Ruidong Wang:** Visualization, Validation, Software, Resources, Investigation. **Lingling Xu:** Writing – original draft, Project administration, Methodology, Investigation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (62376252); Key Project of Natural Science Foundation of Zhejiang Province (LZ22F030003); Zhejiang Province Leading Geese Plan (2025C02025, 2025C01056).

Data availability

Data will be made available on request.

References

[1] S. Rajendar, D. Rathinasamy, R. Pavithra, V.K. Kaliappan, S. Gnanamurthy, Prediction of stopping distance for autonomous emergency braking using stereo camera pedestrian detection, Mater. Today: Proc. 51 (2022) 1224–1228.

[2] Z. Shi, Z. Xu, T. Wang, A method for detecting pedestrian height and distance based on monocular vision technology, Measurement 199 (2022) 111418.

[3] P.K.-Y. Wong, H. Luo, M. Wang, P.H. Leung, J.C. Cheng, Recognition of pedestrian trajectories and attributes with computer vision and deep learning techniques, Adv. Eng. Informatics 49 (2021) 101356.

[4] Y. Ji, K. Song, H. Wen, X. Xue, Y. Yan, Q. Meng, UAV applications in intelligent traffic: RGBT image feature registration and complementary perception, Adv. Eng. Informatics 63 (2025) 102953.

[5] Y. Zhang, Y. Yang, W. Kang, J. Zhen, Cross-erasure enhanced network for occluded person re-identification, Pattern Recognit. Lett. 193 (2025) 108–114.

[6] G. Xu, W. Liao, X. Zhang, C. Li, X. He, X. Wu, Haar wavelet downsampling: A simple but effective downsampling module for semantic segmentation, Pattern Recognit. 143 (2023) 109819.

[7] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR'05, Vol. 1, IEEE, 2005, pp. 886–893.

[8] T. Ojala, M. Pietikainen, D. Harwood, Performance evaluation of texture measures with classification based on Kullback discrimination of distributions, in: Proceedings of 12th International Conference on Pattern Recognition, Vol. 1, IEEE, 1994, pp. 582–585.

[9] D.G. Lowe, Object recognition from local scale-invariant features, in: Proceedings of the Seventh IEEE International Conference on Computer Vision, Vol. 2, IEEE, 1999, pp. 1150–1157.

[10] H. Li, X.-J. Wu, J. Kittler, Infrared and visible image fusion using a deep learning framework, in: 2018 24th International Conference on Pattern Recognition, ICPR, IEEE, 2018, pp. 2705–2710.

[11] S. Park, D.H. Choi, J.U. Kim, Y.M. Ro, Robust thermal infrared pedestrian detection by associating visible pedestrian knowledge, in: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2022, pp. 4468–4472.

[12] X. Deng, P.L. Dragotti, Deep convolutional neural network for multi-modal image restoration and fusion, IEEE Trans. Pattern Anal. Mach. Intell. 43 (10) (2020) 3333–3348.

[13] M. Zhao, W. Li, L. Li, J. Hu, P. Ma, R. Tao, Single-frame infrared small-target detection: A survey, IEEE Geosci. Remote. Sens. Mag. 10 (2) (2022) 87–119.

[14] Y. Zhang, J. Zhen, S. Sun, T. Liu, L. Huo, T. Wang, SCAFNet: A semantic compensated adaptive fusion network for remote sensing images change detection, IEEE Geosci. Remote. Sens. Lett. 23 (2026) 1–5.

[15] Y. Zhang, T. Liu, J. Zhen, Y. Kang, Y. Cheng, Adaptive downsampling and scale enhanced detection head for tiny object detection in remote sensing image, IEEE Geosci. Remote. Sens. Lett. 22 (2025) 1–5.

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

[17] J. Xu, X. Fan, H. Jian, C. Xu, W. Bei, Q. Ge, T. Zhao, YoloOW: A spatial scale adaptive real-time object detection neural network for open water search and rescue from UAV aerial imagery, IEEE Trans. Geosci. Remote Sens. 62 (2024) 1–15.

[18] B. Yang, X. Zhang, J. Zhang, J. Luo, M. Zhou, Y. Pi, EFLNet: Enhancing feature learning network for infrared small target detection, IEEE Trans. Geosci. Remote Sens. 62 (2024) 1–11.

[19] T. Yue, X. Lu, J. Cai, Y. Chen, S. Chu, SDS-Net: Shallow–deep synergism-detection network for infrared small target detection, IEEE Trans. Geosci. Remote Sens. 63 (2025) 1–13.

[20] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: Towards real-time object detection with region proposal networks, IEEE Trans. Pattern Anal. Mach. Intell. 39 (6) (2016) 1137–1149.

[21] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2117–2125.

[22] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask r-cnn, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2961–2969.

[23] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A.C. Berg, Ssd: Single shot multibox detector, in: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, the Netherlands, October 11–14, 2016, ProceediFngs, Part I 14, Springer International Publishing, 2016, pp. 21–37.

[24] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2980–2988.

[25] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, Q. Tian, Centernet: Keypoint triplets for object detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 6569–6578.

[26] M. Tan, R. Pang, Q.V. Le, Efficientdet: Scalable and efficient object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 10781–10790.

[27] J. Redmon, You only look once: Unified, real-time object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016.

[28] R. Khanam, M. Hussain, Yolov11: An overview of the key architectural enhancements, 2024, arXiv preprint arXiv:2410.17725.

[29] Y. Tian, Q. Ye, D. Doermann, Yolov12: Attention-centric real-time object detectors, 2025, arXiv preprint arXiv:2502.12524.

[30] H. Fu, S. Wang, P. Duan, C. Xiao, R. Dian, S. Li, Z. Li, Lraf-net: Long-range attention fusion network for visible–infrared object detection, IEEE Trans. Neural Networks Learn. Syst. (2023).

[31] J. Wei, S. Su, Z. Zhao, X. Tong, L. Hu, W. Gao, Infrared pedestrian detection using improved UNet and YOLO through sharing visible light domain information, Measurement 221 (2023) 113442.

[32] S. Wu, S. Shan, G. Xiao, M.S. Lew, X. Gao, Implicit modality knowledge alignment and uncertainty estimation for visible-infrared person re-identification, Expert Syst. Appl. 259 (2025) 125291.

[33] B. Zheng, H. Huo, X. Liu, S. Pang, J. Li, Pedestrian detection-driven cascade network for infrared and visible image fusion, Signal Process. 225 (2024) 109620.

[34] B. Lei, J. Fan, Infrared pedestrian segmentation algorithm based on the two-dimensional Kaniadakis entropy thresholding, Knowl.-Based Syst. 225 (2021) 107089.

[35] K. Li, X. Wang, Y. Liu, B. Zhang, M. Zhang, Cross-modality disentanglement and shared feedback learning for infrared-visible person re-identification, Knowl.-Based Syst. 252 (2022) 109337.

[36] W. Farhat, O.B. Rhaiem, H. Faiedh, C. Souani, Pedestrian detection and tracking using an enhanced YOLOv9 model for automotive vehicles, Measurement (2025) 118009.

[37] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A.L. Yuille, Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, IEEE Trans. Pattern Anal. Mach. Intell. 40 (4) (2017) 834–848.

[38] H. Zhao, J. Shi, X. Qi, X. Wang, J. Jia, Pyramid scene parsing network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2881–2890.

[39] S. Liu, L. Qi, H. Qin, J. Shi, J. Jia, Path aggregation network for instance segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8759–8768.

[40] Y. Zhang, Y. Zhang, Y. Xiao, T. Wang, Spatiotemporal dual-branch feature-guided fusion network for driver attention prediction, Expert Syst. Appl. 292 (2025) 128564.

[41] A.G. Howard, Mobilenets: Efficient convolutional neural networks for mobile vision applications, 2017, arXiv preprint arXiv:1704.04861.

[42] X. Zhang, X. Zhou, M. Lin, J. Sun, Shufflenet: An extremely efficient convolutional neural network for mobile devices, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6848–6856.

[43] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, C. Xu, Ghostnet: More features from cheap operations, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 1580–1589.

[44] J. Chen, S.-h. Kao, H. He, W. Zhuo, S. Wen, C.-H. Lee, S.-H.G. Chan, Run, don't walk: Chasing higher FLOPS for faster neural networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 12021–12031.

[45] S. Mehta, M. Rastegari, Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer, 2021, arXiv preprint arXiv:2110.02178.

[46] H. He, J. Zhang, Y. Cai, H. Chen, X. Hu, Z. Gan, Y. Wang, C. Wang, Y. Wu, L. Xie, MobileMamba: Lightweight multi-receptive visual mamba network, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2025, pp. 4497–4507.

[47] W. Lu, Z. Zhang, M. Nguyen, A lightweight CNN–transformer network with Laplacian loss for low-altitude UAV imagery semantic segmentation, IEEE Trans. Geosci. Remote Sens. 62 (2024) 1–20.

[48] Y. Wu, T. Xu, Y. Qin, F. Guo, Z. Zhao, T. Yang, C. Wang, AirboardNet: A UAV onboard girder inspection approach for high-speed railroad bridge using multi-task knowledge distillation, Adv. Eng. Informatics 67 (2025) 103544.

[49] Y. Yuan, S. Gao, Z. Zhang, W. Wang, Z. Xu, Z. Liu, Edge-cloud collaborative UAV object detection: Edge-embedded lightweight algorithm design and task offloading using fuzzy neural network, IEEE Trans. Cloud Comput. 12 (1) (2024) 306–318.

[50] Y. Sang, T. Liu, Y. Liu, T. Ma, S. Wang, X. Zhang, J. Sun, A lightweight network with latent representations for UAV thermal image super-resolution, IEEE Trans. Geosci. Remote Sens. 62 (2024) 1–11.

[51] Y. Zhang, C. Wu, T. Zhang, Y. Liu, Y. Zheng, Self-attention guidance and multiscale feature fusion-based UAV image object detection, IEEE Geosci. Remote. Sens. Lett. 20 (2023) 1–5.

[52] Y. Zhao, W. Lv, S. Xu, J. Wei, G. Wang, Q. Dang, Y. Liu, J. Chen, Detrs beat yolos on real-time object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 16965–16974.

[53] Y. Zhang, T. Wang, L. Xue, W. Lian, R. Tao, ORSI salient object detection via progressive interaction and saliency-guided enhancement, IEEE Geosci. Remote. Sens. Lett. 23 (2026) 1–5.

[54] K.W. Lau, L.-M. Po, Y.A.U. Rehman, Large separable kernel attention: Rethinking the large kernel attention design in cnn, Expert Syst. Appl. 236 (2024) 121352.
[55] M. Hu, J. Feng, J. Hua, B. Lai, J. Huang, X. Gong, X.-S. Hua, Online convolutional re-parameterization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 568–577.
[56] S. Santurkar, D. Tsipras, A. Ilyas, A. Madry, How does batch normalization help optimization? Adv. Neural Inf. Process. Syst. 31 (2018).
[57] Q. Hou, D. Zhou, J. Feng, Coordinate attention for efficient mobile network design, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 13713–13722.
[58] J. Suo, T. Wang, X. Zhang, H. Chen, W. Zhou, W. Shi, HIT-UAV: A high-altitude infrared thermal dataset for unmanned aerial vehicle-based object detection, Sci. Data 10 (1) (2023) 227.
[59] J. Liu, X. Fan, Z. Huang, G. Wu, R. Liu, W. Zhong, Z. Luo, Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 5802–5811.
[60] G. Jocher, YOLOv5 by ultralytics, 2020, http://dx.doi.org/10.5281/zenodo.3908559, URL https://github.com/ultralytics/yolov5.
[61] C. Li, L. Li, H. Jiang, K. Weng, Y. Geng, L. Li, Z. Ke, Q. Li, M. Cheng, W. Nie, et al., YOLOv6: A single-stage object detection framework for industrial applications, 2022, arXiv preprint arXiv:2209.02976.
[62] C.-Y. Wang, A. Bochkovskiy, H.-Y.M. Liao, YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 7464–7475.
[63] G. Jocher, A. Chaurasia, J. Qiu, Ultralytics YOLO, 2023, URL https://github.com/ultralytics/ultralytics.
[64] C.-Y. Wang, I.-H. Yeh, H.-Y.M. Liao, YOLOv9: Learning what you want to learn using programmable gradient information, 2024, arXiv preprint arXiv:2402.13616.
[65] A. Wang, H. Chen, L. Liu, K. Chen, Z. Lin, J. Han, G. Ding, Yolov10: Real-time end-to-end object detection, 2024, arXiv preprint arXiv:2405.14458.
[66] Z. Cai, N. Vasconcelos, Cascade r-cnn: Delving into high quality object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6154–6162.
[67] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, J. Dai, Deformable detr: Deformable transformers for end-to-end object detection, 2020, arXiv preprint arXiv:2010.04159.
[68] P. Sun, R. Zhang, Y. Jiang, T. Kong, C. Xu, W. Zhan, M. Tomizuka, L. Li, Z. Yuan, C. Wang, et al., Sparse r-cnn: End-to-end object detection with learnable proposals, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 14454–14463.
[69] H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, L.M. Ni, H.-Y. Shum, Dino: Detr with improved denoising anchor boxes for end-to-end object detection, 2022, arXiv preprint arXiv:2203.03605.
[70] Y. Feng, J. Huang, S. Du, S. Ying, J.-H. Yong, Y. Li, G. Ding, R. Ji, Y. Gao, Hyper-yolo: When visual object detection meets hypergraph computation, IEEE Trans. Pattern Anal. Mach. Intell. (2024).
[71] S. Wang, H. Jiang, Z. Li, J. Yang, X. Ma, J. Chen, X. Tang, PHSI-RTDETR: A lightweight infrared small target detection algorithm based on UAV aerial photography, Drones 8 (6) (2024).
[72] L. Hu, J. Yuan, B. Cheng, Q. Xu, CSFPR-RTDETR: Real-time small object detection network for UAV images based on cross-spatial-frequency domain and position relation, IEEE Trans. Geosci. Remote Sens. 63 (2025) 1–19.
[73] J. Chen, N. Liu, H. Sun, Y. Wang, Freq-DETR: Frequency-aware transformer for real-time small object detection in unmanned aerial vehicle imagery, Expert Syst. Appl. 298 (2026) 129710.
[74] X. Liu, S. Zhou, J. Ma, Y. Sun, J. Zhang, H. Zuo, DFAS-YOLO: Dual feature-aware sampling for small-object detection in remote sensing images, Remote. Sens. 17 (20) (2025).
[75] H. Shihua, L. Zhichao, C. Xiaodong, Y. Yongjun, Z. Xiao, S. Xi, DEIM: DETR with improved matching for fast convergence, 2025.
[76] D. Qin, C. Leichner, M. Delakis, M. Fornoni, S. Luo, F. Yang, W. Wang, C. Banbury, C. Ye, B. Akin, et al., MobileNetV4: Universal models for the mobile ecosystem, in: European Conference on Computer Vision, Springer, 2025, pp. 78–96.
[77] X. Liu, H. Peng, N. Zheng, Y. Yang, H. Hu, Y. Yuan, Efficientvit: Memory efficient vision transformer with cascaded group attention, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 14420–14430.
[78] N. Ma, X. Zhang, H.-T. Zheng, J. Sun, Shufflenet v2: Practical guidelines for efficient cnn architecture design, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 116–131.
[79] D. Ouyang, S. He, G. Zhang, M. Luo, H. Guo, J. Zhan, Z. Huang, Efficient multi-scale attention module with cross-spatial learning, in: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2023, pp. 1–5.
[80] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132–7141.
[81] S. Woo, J. Park, J.-Y. Lee, I.S. Kweon, Cbam: Convolutional block attention module, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 3–19.
[82] H. Li, J. Li, H. Wei, Z. Liu, Z. Zhan, Q. Ren, Slim-neck by GSConv: A lightweight-design for real-time detector architectures, J. Real-Time Image Process. 21 (3) (2024) 62.
[83] Y. Chen, X. Dai, M. Liu, D. Chen, L. Yuan, Z. Liu, Dynamic convolution: Attention over convolution kernels, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 11030–11039.
[84] X. Zhang, C. Liu, D. Yang, T. Song, Y. Ye, K. Li, Y. Song, RFAConv: Innovating spatial attention and standard convolutional operation, 2023, arXiv preprint arXiv:2304.03198.
[85] X. Zhang, Y. Song, T. Song, D. Yang, Y. Ye, J. Zhou, L. Zhang, AKConv: Convolutional kernel with arbitrary sampled shapes and arbitrary number of parameters, 2023, arXiv preprint arXiv:2311.11587.
[86] S. Hu, F. Gao, X. Zhou, J. Dong, Q. Du, Hybrid convolutional and attention network for hyperspectral image denoising, IEEE Geosci. Remote. Sens. Lett. (2024).
[87] R. Sunkara, T. Luo, No more strided convolutions or pooling: A new CNN building block for low-resolution images and small objects, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2022, pp. 443–459.
[88] X. Cai, Q. Lai, Y. Wang, W. Wang, Z. Sun, Y. Yao, Poly kernel inception network for remote sensing detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 27706–27716.
[89] W. Lin, Z. Wu, J. Chen, J. Huang, L. Jin, Scale-aware modulation meet transformer, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, ICCV, 2023, pp. 6015–6026.
[90] X. Zhu, H. Hu, S. Lin, J. Dai, Deformable convnets v2: More deformable, better results, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 9308–9316.
[91] A.-F.O. Detector, Fcos: A simple and strong anchor-free object detector, IEEE Trans. Pattern Anal. Mach. Intell. 44 (4) (2022).
[92] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 618–626.