Engineering Applications of Artificial Intelligence 163 (2026) 113016

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

Research paper

A real-time vehicle detection method in unmanned aerial vehicle images with selective contextual features

Wanxia Huang [a,1], Chaojun Dong [a,*,1], Xiankun Liu [a], Ye Li [a,1], Yikui Zhai [a,*,1], Kaitong Ou [a], Hao Quan [b]

[a] WuYi University, Jiang Men, 529020, Guang Dong, China
[b] Politecnico di Milano, Milan, 20133, Milano, Italy

ARTICLE INFO

Keywords:
Detection
Unmanned aerial vehicle images
Small object
Occluded object
Neck design

ABSTRACT

Detecting small and occluded objects in unmanned aerial vehicle (UAV) images remains a critical challenge. The inferior feature quality of these small and occluded objects leads to incomplete feature extraction, resulting in missed detections. To address this challenge, we propose an innovative detector based on ObjectBox to enhance detection performance and reduce missed detections of small and occluded objects by incorporating the neck called selective fused deformable context feature path aggregation network (SFDCFPAN) and the decoupled head. Firstly, we designed a neck called the selective feature path aggregation network (SFPAN) to fuse features and reduce the loss of spatial information. Subsequently, we provide a feature extraction module named fused deformable context feature extraction module (FDC) to model object shapes and then fuse context features to obtain the object's semantic and spatial information. We employ the FDC module as the feature extraction module at specific locations within four feature layers of SFPAN, denoted as SFDCFPAN, to enhance the detector's feature extraction and modeling capabilities. Lastly, we introduce a decoupled head structure to alleviate the mutual interference between classification and localization tasks. We conduct a comparative analysis of our detector with popular detectors on the VisDrone2019 and the UAVDT sub-dataset. Experimental results demonstrate the superior performance of our detector, achieving high accuracy on the two datasets while meeting real-time constraints. Furthermore, we integrate SFPAN and SFDCFPAN into various detectors. Experimental results exhibit the substantial enhancement in detector accuracy achieved by these feature fusion frameworks without compromising real-time performance, demonstrating the applicability to existing detectors.

1. Introduction

In recent years, deep learning techniques have been widely utilized in visual inspection tasks (Ren et al., 2015; Redmon et al., 2016; Zhang et al., 2025b, 2023c; Zhu et al., 2024a; Zhang et al., 2023a) to improve accuracy. Object detection algorithms accomplish detection tasks by extracting features of objects and utilizing them for localization and categorization. In the transportation field, object detection algorithms primarily perform detection tasks from the perspective of road monitoring, like UA-DETRAC (Wen et al., 2020) as shown in Fig. 1(a). Object detection algorithms from the perspective of road monitoring will encounter challenges such as vehicle mutual occlusion and limited detection range by the camera's mounting height. Unmanned aerial vehicles (UAV), known for their high level of automation, cost-effectiveness, and flexibility, are promised in transportation sectors. The flexibility of UAV enables them to navigate around high-rise buildings and highways for precise traffic monitoring, like VisDrone2019 (Du et al., 2019) as shown in Fig. 1(b). UAV extend the detection range of object detection algorithms and overcome the limitations of road monitoring perspectives with the expansive field of view at elevated altitudes, improving the efficiency of traffic management. However, small objects become in number with the extended shooting distance. Furthermore, the objects in UAV images are occluded with structures like buildings, trees, and vehicles when UAV are in intricate settings like urban streets and highways. Consequently, the captured images by the UAV at elevated altitudes contain numerous small and occluded objects that are susceptible to background interference and exhibit poor feature quality, posing challenges for

* Corresponding authors.
E-mail addresses: WanxiaHuang1998@outlook.com (W. Huang), cjun-dong@163.com (C. Dong), lxllml@163.com (X. Liu), leyelee@163.com (Y. Li), yikuizhai@163.com (Y. Zhai), 490729611@qq.com (K. Ou), hao.quan@polimi.it (H. Quan).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.engappai.2025.113016
Received 17 December 2024; Received in revised form 30 June 2025; Accepted 28 October 2025
Available online 8 November 2025

detectors to extract informative features. To address this challenge, current detection methods in UAV images fall into two main categories: CNN-based methods that rely on convolutional neural networks (LeCun et al., 2002), and Transformer-based methods that are built around the Transformer attention mechanism (Vaswani et al., 2017). In CNN-based methods, Faster R-CNN (Ren et al., 2015), a classic two-stage detection algorithm, reduces interference from negative samples by filtering candidate boxes. However, its two-stage process results in slow detection speed. You only look once (YOLO) (Redmon and Farhadi, 2018; Bochkovskiy et al., 2020; Jocher, 2020), a representative single-stage detector, offers high real-time performance and accuracy but suffers from a significant imbalance between positive and negative samples in scenarios involving small or dense objects due to its preset anchor boxes. Anchor-free algorithms such as FCOS (Tian et al., 2020), ObjectBox (Zand et al., 2022), and YOLOv11 (Khanam and Hussain, 2024) eliminate the need to match anchors with ground-truth bounding boxes by discarding anchors entirely. CNN-based methods excel at local feature extraction and offer high computational efficiency, but their limited receptive fields impede the capture of long-range contextual dependencies. Transformer-based methods can model long-range dependencies among objects. However, it has shortcomings in detecting small objects because it is difficult to capture detailed information of small objects due to its sparse representation on the feature map. To address this problem, RT-DETRv3 (Wang et al., 2025) introduces a dense positive sample auxiliary supervision strategy. Similarly, DEIM (Huang et al., 2025) further improves the detection performance by increasing the number of positive samples and optimizing the matching quality loss function. However, these Transformer-based methods lack integration with feature fusion frameworks, which hinders the fusion of semantic information and spatial details across scales, resulting in a high missed detection rate. Increasing input image resolution can partially mitigate the loss of small object information. However, this approach substantially raises computational complexity and inference time, hindering practical deployment. To balance real-time performance and small object detection accuracy, we leverage the real-time speed and computational efficiency of CNN-based methods and propose an innovative anchor-free detector to decrease missed detections of small and occluded objects in UAV images. The core of our detector lies in a unique feature fusion architecture called selective fused deformable context feature path aggregation network (SFDCFPAN). This architecture combines the fused deformable context feature extraction module (FDC) into the selective feature path aggregation network (SFPAN), which enhances the detector's ability to model shapes of small and occluded objects and extract features inclusive of contextual semantic and spatial information. SFDCFPAN and SFPAN differ from other feature fusion architectures in that they remove the largest-sized feature layer downsampling, preventing the loss of spatial information critical for object location. In designing the FDC module, we enhance feature information fusion and improve feature extraction capabilities by performing an element-wise addition of the extracted features with the original features in the feature extraction branch, followed by merging these with the features from the skip branch. Moreover, our detector adopts a decoupled head to reduce mutual interference between localization and classification tasks, enhancing its generalization capacity and detection performance. Notably, our method demonstrates excellent compatibility and scalability and can be seamlessly integrated into existing detectors like ObjectBox (Zand et al., 2022) and YOLOv5 (Jocher, 2020), which enhances their ability to detect small and occluded objects. Through adopting the decoupled head structure and the neck network SFDCFPAN, our detector achieves high detection accuracy on the VisDrone2019 test-set while meeting real-time requirements, with mAP_{50} reaching 38.4%, mAP reaching 21.8%, and the mAP for small objects reaching 13%, representing a 2.9% improvement over ObjectBox. SFDC-YOLOv8, which is YOLOv8 (Jocher, 2023) equipped with SFDCFPAN, maintains real-time performance while ensuring high detection accuracy, with the mAP_{50} reaching 41.0%, mAP reaching 23.6%, and the mAP for small objects reaching 14.0%, representing a 3.0% improvement over YOLOv8. Our approach has been proven effective in tackling detection challenges posed by small and occluded objects and achieving real-time conditions through experiments. Overall, our work's main contributions can be summarized as follows:

- We propose a detector to enhance the capability in detecting scenarios involving small and occluded objects, which employs SFDCFPAN as the neck network to improve shape modeling and feature extraction capacities for small and occluded objects by applying FDC modules in SFPAN's specific feature layers. Additionally, our detector introduces the decoupled head structure to minimize the conflict between classification and localization tasks by separating the two tasks, improving generalization capacity. By employing SFDCFPAN and a decoupled head, our detector displays remarkable detection performance in various scenarios, notably decreasing the occurrence of missed detections for small and occluded objects.
- We propose a novel feature extraction module called the fused deformable context feature extraction module (FDC) to boost the detector's feature extraction capability. The FDC module adopts two convolutional layers to split the features into two parts and feed them into the skip and the feature extraction branches, respectively. The skip branch transfers features, while the feature extraction branch extracts local features. Subsequently, the skip branch's features are concatenated with features from the feature extraction branch, followed by a deformable convolutional layer to model the object shape that acquires small and occluded objects' spatial information.
- We propose a feature fusion architecture named selected feature path aggregation feature pyramid network (SFPAN), which fuses multi-feature layers by upsampling and downsampling to obtain context information, but cancels the largest-sized feature layer's downsampling operation in the neck structure to reduce spatial information loss and enhance the detector's detection capability. Subsequently, we develop a selective fused deformable context feature path aggregation network (SFDCFPAN) to improve the detector's feature extraction capacities in scenarios of small and occluded objects by replacing the feature extraction module of the four feature output layers in SFPAN with the FDC module, further enhancing the detection capability of the detector. Significantly, SFPAN and SFDCFPAN can effortlessly integrate into diverse detection frameworks.

This paper's structure for the subsequent sections is as follows: Section 2 revisits previous related work, whereas Section 3 outlines the structures of our detector, FDC module, SFDCFPAN, and the decoupled head. Section 4 introduces ablation and comparison experiments conducted on the test-sets of VisDrone2019 and UAVDT (Du et al., 2018) sub-dataset. Concluding the paper, Section 5 provides the conclusions.

2. Related work

2.1. General object detection

Object detection algorithms accurately locate and classify objects in images to accomplish the detection task by extracting features containing object space and semantic information. CNN-based object detection algorithms are categorized into anchor-based or anchor-free detectors based on the presence of the anchor and classified as one-stage or two-stage detectors according to the detection steps. Faster R-CNN (Ren et al., 2015), a prominent two-stage anchor-based algorithm, employs a region proposal network (RPN) to handle the preset anchors for generating candidate bounding boxes. Subsequently, these boxes are forwarded to the detector for object localization and classification tasks. However, the candidate bounding boxes extracted by RPN in Faster R-CNN exhibit subpar quality. To address this issue, Cai et al. proposed Cascade R-CNN (Cai and Vasconcelos, 2018), which incorporates the cascade learning concept, progressively refining the quality

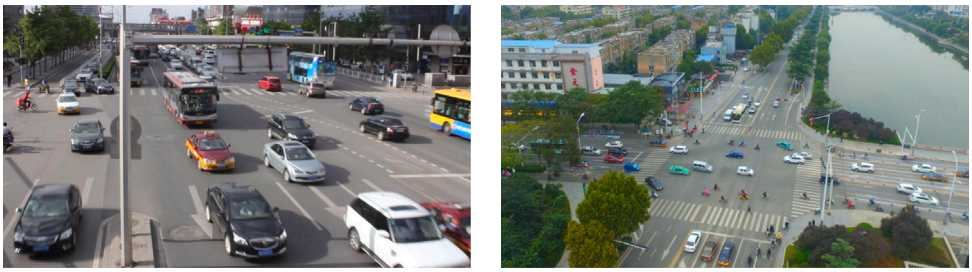

Fig. 1. A comparison of traffic monitoring perspectives. (a) Road-based. (b) UAV-based.

of candidate bounding boxes. Although the two-stage anchor-based algorithm has shown excellent performance in object detection tasks, its sequential division of the detection process into two stages leads to slow detection speed and cannot meet real-time requirements. You only look once(YOLO) series algorithms (Redmon et al., 2016; Redmon and Farhadi, 2017, 2018; Bochkovskiy et al., 2020), classic one-stage algorithms, have consistently achieved significant advancements and breakthroughs in recent years, offering efficient solutions for real-time object detection tasks by simplifying the detection task. The robustness of YOLOv5 has driven its broad adoption in practical engineering applications. However, YOLOv5 is affected by anchors, leading to poor detection performance. To mitigate the influence of anchors, Tian et al. proposed FCOS (Tian et al., 2020), which eliminates the reliance on the anchor by regressing the distances between each center position on the feature map and the bounding box. Inspired by FCOS, Ge et al. proposed YOLOX (Ge et al., 2021), which adopts a decoupled head that refines the classification and localization structure of YOLOv3 to circumvent the influence of anchors and employs the SimOTA label assignment method to optimize the distribution of positive samples. Lyu et al. proposed RTMDet (Lyu et al., 2022), which builds upon the foundation of YOLOX and incorporates a shared detection head structure and a backbone with large-deep convolutions, enhancing the model's capacity to capture global context and improving its representational prowess. Li et al. proposed YOLOv6 (Li et al., 2022), which employs EfficientRep and Rep-PAN architectures inspired by RepVGG (Ding et al., 2021) to augment the detector's feature expression capabilities. Xu et al. proposed PPYOLOE (Xu et al., 2022), drawing inspiration from the model concepts of RepVGG and CSPNet (Wang et al., 2020), enhanced the ResNet (He et al., 2016) architecture, proposing CSPRepResNet, and incorporating Task-aligned detection Head (T-Head) from TOOD (Feng et al., 2021) significantly improves model performance. Wang et al. proposed YOLOv7 (Wang et al., 2023), which adopts the auxiliary head during training to enhance model performance and integrates the feature extraction module ELAN into its architecture to augment feature extraction capability by controlling the gradient path. Reis et al. proposed YOLOv8, designing the feature extraction module C2F by referring to the C3 module of YOLOv5 and the ELAN module of YOLOv7, enhancing the model's feature extraction capability. Zand et al. proposed ObjectBox, which adopts a novel regression method that the bounding box regression on the object center cell's upper left and lower right corners to overcome the influence of anchor and treating objects from different feature layers equally in label assignment. Wang et al. proposed YOLOv9 (Wang et al., 2024c), which enhances detection efficiency through two key innovations: a Programmable Gradient Information (PGI) and a lightweight hierarchical feature extraction module. Wang et al. proposed YOLOv10 (Wang et al., 2024a), featuring a dynamic label assignment strategy, along with architectural refinements to improve detection accuracy. Khanam et al. proposed YOLOv11 (Khanam and Hussain, 2024), enhancing detection performance through a lightweight detection head and novel feature extraction modules C3k2 and C2PSA. In Transformer-based object detection algorithms, DETR (Carion et al., 2020) proposed by Carion et al. to cancel traditional post-processing steps in object detection algorithms by introducing Transformer (Vaswani et al., 2017). However, due to the limited processing area of the attention mechanism in the Transformer, DETR experiences prolonged training times for model convergence. Inspired by deformable convolutional networks (Dai et al., 2017), Zhu et al. proposed Deformable DETR (Zhu et al., 2020) to accelerate training by combining deformable convolution's sparse space sampling method and Transformer's relationship modeling method, enhancing detection performance while reducing training time. Zhang et al. proposed DINO (Zhang et al., 2022a), which enhances detector performance through a novel training method and hybrid query selection approach. Cai et al. proposed Align-DETR (Cai et al., 2024), which enhances detection performance by introducing a novel loss function termed Align Loss. Zhao et al. proposed RT-DETR (Zhao et al., 2024), improving the detector's training and inference by incorporating a hybrid encoder and IoU-aware Query Selection. Lv et al. proposed RT-DETRv2 (Lv et al., 2024) based on RT-DETR, which further enhances RT-DETR's performance without compromising speed by optimizing training procedures and sampling strategies. Wang et al. proposed RT-DETRv3, which further improved the detection performance through a series of dense positive supervision methods based on RT-DETRv2. Huang et al. proposed DEIM, which design a matching-aware Loss (MAL) to optimize the loss function of matching quality and increase the positive sample supervision density by increasing the number of objects in training images.

Although the aforementioned object detection algorithms demonstrate distinct advantages, their performance remains inadequate when applied to UAV images. This limitation primarily arises because UAV images frequently contain numerous small and occluded objects with incomplete feature information, which hinders general object detection algorithms from accurately identifying small and occluded objects in such images.

2.2. Object detection in UAV images

In recent years, researchers have investigated the object detection task of UAV images based on object detection algorithms, yielding significant outcomes. Zhu et al. proposed TPH-YOLOv5 (Zhu et al., 2021), a novel approach building upon YOLOv5, which integrates the CBAM (Woo et al., 2018) attention module into the neck network to enhance feature extraction capabilities and design a Transformer Prediction Head (TPH) structure to improve the detection accuracy in small and motion-blurred objects. Li et al. proposed an efficient method (Li et al., 2024) for detecting dense and small Objects based on YOLOv5 by incorporating kernel K-means (Dhillon et al., 2004) and two modules to enhance spatial information. Lu et al. proposed a convolutional neural network transformer hybrid model (Lu et al., 2023) that leverages the cross-shaped window (CSWin) transformer (Dong et al., 2022) as the backbone and implements a slicing-based inference (SI) method to enhance accuracy in detecting small and multi-scale objects. Meng et al. proposed SODCNN (Meng et al., 2023), which enhances the detector's performance for detecting multi-scale and small objects

by eliminating redundant detection heads and incorporating the EIoU loss (Zhang et al., 2022b) function and attention-based feature fusion module in YOLOv7. Jiang et al. proposed MFFSODNet (Jiang et al., 2024), which enhances the fusion capability of multi-scale features and improves the detection capability of small objects by designing a module for extracting multi-scale features and a neck network with bidirectional dense paths. Zhang et al. proposed a method (Zhang et al., 2024b) to improve multi-scale object detection by designing a full-scale feature aggregation (FFA) and a parallel super-resolution semantic enhancement (PSSE) module to enhance detection capabilities. Sun et al. proposed RSOD (Sun et al., 2022), enhancing the effect of detecting small objects through an algorithm capable of adaptive learning and distribution of output features. Song et al. proposed MHA-YOLOv5 (Song et al., 2024) to enhance the detection of small objects by designing a structure called multi-scale hybrid attention (MHA) into YOLOv5. Mo et al. proposed PVDet (Mo et al., 2023) to enhance pedestrian and vehicle detection in gigapixel images, which develops a backbone architecture known as DPRNet to improve the effective receptive field. Zhu et al. proposed STFNet (Zhu et al., 2024b), which enhances the detection accuracy of small moving objects by designing a feature extraction module that integrates spatiotemporal features. Zhang et al. proposed EANN (Zhang et al., 2024a), which enhances the detection of small objects by implementing a novel neck structure that improves feature richness and a dedicated loss function for regressing the position information of small objects. Wang et al. proposed SHRDet (Wang et al., 2024b) to enhance the detection accuracy of small objects on the sea surface by designing a spatial feature fusion module and a few-sample training strategy within HRNet (Sun et al., 2019). Zhang et al. proposed CFANet (Zhang et al., 2023b) to improve the detection of small objects by designing two novel feature extraction modules and proposing an adaptive overlapping slice (AOS) to obtain detailed object information.

The above-mentioned algorithms enhance the object detection accuracy in UAV or remote sensing images. However, they primarily apply in detecting small or dense objects, lacking suitability for occluded objects, necessitating further refinement and optimization. We propose a detector that employs the decoupled head structure and SFDCFPAN into the ObjectBox framework to enhance the detection capabilities for small and occluded objects, achieving detection accuracy and speed balance to meet the small and occluded object detection demands in UAV images.

3. Methodology and model

This section introduces our proposed detector structure, whose core components include CSPDarkNet (Bochkovskiy et al., 2020), SFDCFPAN, and the decoupled head, as shown in Fig. 2. After image preprocessing of the input image, the detector extracts basic features using CSPDarkNet and then combines features of different scales through SFDCFPAN. SFDCFPAN, a new fusion structure, enhances the modeling capacity of object shapes by adding FDC modules in SFPAN and fuses three feature layers to promote information exchange in different-scale objects. Subsequently, the decoupled head decouples the classification and localization tasks of object detection in UAV images. We employ scale-invariant distance-based IoU loss (SDIoU Loss) (Zand et al., 2022) for bounding box regression and binary cross-entropy loss (BCE Loss) for classification and objective prediction. The decoupled head's design synergizes with the optimization of SDIoU Loss and BCE Loss, enhancing the overall performance of the detector in UAV images. Finally, the detection results are output after image post-processing.

3.1. Overall framework

Due to the unique perspective of UAV, the UAV images often contain small or occluded objects, making feature extraction challenging and frequently causing missed detections. To enhance the detector's ability to identify such objects, we employ CSPDarknet as the backbone network for basic feature extraction, followed by SFDCFPAN for multi-scale feature fusion, and finally, a decoupled head to mitigate feature conflicts between classification and localization tasks. The overall framework is illustrated in Fig. 3. Specifically, CSPDarknet, based on the CSPNet and Darknet (Redmon and Farhadi, 2018) design, captures multi-scale features while preserving shallow detail and deep semantic information. This architecture significantly reduces computational complexity and improves feature extraction efficiency. SFDCFPAN subsequently refines and fuses the P2, P3, P4, and P5 feature maps extracted by CSPDarknet. Following the feature fusion structure of our SFPAN design, SFDCFPAN removes the downsampling operation on the largest-scale feature layer in the bidirectional feature pyramid, minimizing spatial information loss and facilitating effective multi-scale feature fusion. Additionally, the FDC module within SFDCFPAN precisely models object shapes, effectively capturing contours of occluded and small objects through its unique structure, enhancing localization accuracy. Finally, the fused feature map from SFDCFPAN is fed into the decoupled head, alleviating the feature conflict caused by sharing features between the classification and the localization tasks by separating the two tasks. This separation allows the classification branch to focus on discriminative semantic information, and the localization branch concentrates on extracting accurate spatial information. Consequently, this design improves classification and localization accuracy, enhancing overall detection performance.

3.2. FDC design and structure

To address the challenge of detecting small and occluded objects in UAV images, we draw inspiration from CSPNet and deformable convnets v2 (Zhu et al., 2019), and design the FDC module to enhance the detector's modeling and feature extraction capabilities for these objects, improving its ability to capture detailed information, as shown in Fig. 4. The FDC module consists of three 1×1 convolutional layers, three 3×3 convolutional layers, and a 3×3 deformable convolutional layer. The FDC module adopts a dual-branch design by segregating feature extraction into two distinct branches: a skip branch and a feature extraction branch. This architectural choice allows the FDC module to reduce computational burden and improve feature extraction efficiency. The FDC module begins with two independent 1×1 convolutional layers to reduce feature channels, then feeds them into the skip and feature extraction branches. The skip branch transfers feature, while the feature extraction branch focuses on feature extraction. The feature extraction branch initiates by extracting the local features of the object through three 3×3 convolutional layers. Subsequently, it element-wise adds the extracted local features with the preceding features to preserve the crucial feature information, followed by a 1×1 convolutional layer that introduces non-linear relationships. By iteratively applying the 3×3 convolutional layer three times and combining it with element-wise addition between feature maps, the feature extraction branch captures rich semantic and spatial information, enhancing the model's feature representation capabilities. The FDC module concatenates the features from the feature extraction and the skip branches. These features are directed to a 3×3 deformable convolutional layer for object shape modeling, further augmenting the model's capabilities in modeling and extracting features. Specifically, the deformable convolutional layer integrates deformable convolution v2, batch normalization (BN) (Ioffe and Szegedy, 2015), and the SILU activation function. Deformable convolution adapts to the spatial geometric variations of objects through offsets in the spatial sampling positions along the horizontal (x) and vertical (y) directions. This adaptation facilitates deformable convolution for accurately modeling object shapes and efficiently capturing spatial information, overcoming the limitations of standard convolution. Deformable convolution v2 extends the deformable convolution framework by introducing a modulation mechanism that broadens the deformation modeling scope,

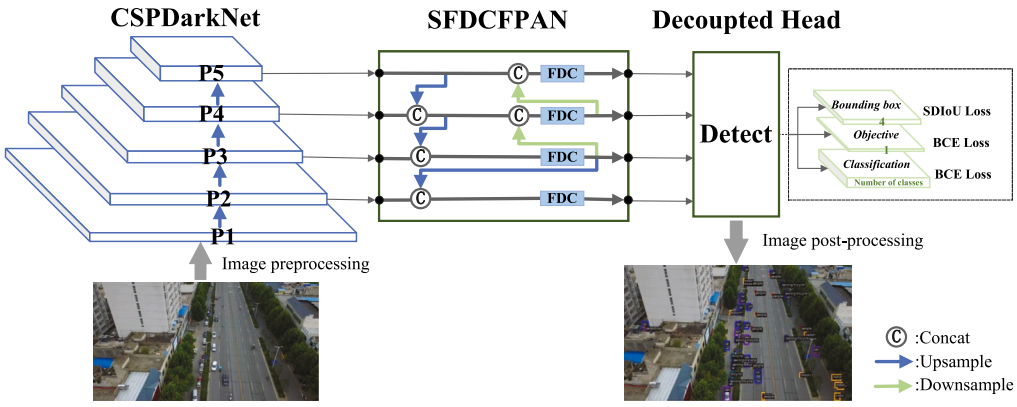

Fig. 2. Our detector structure.

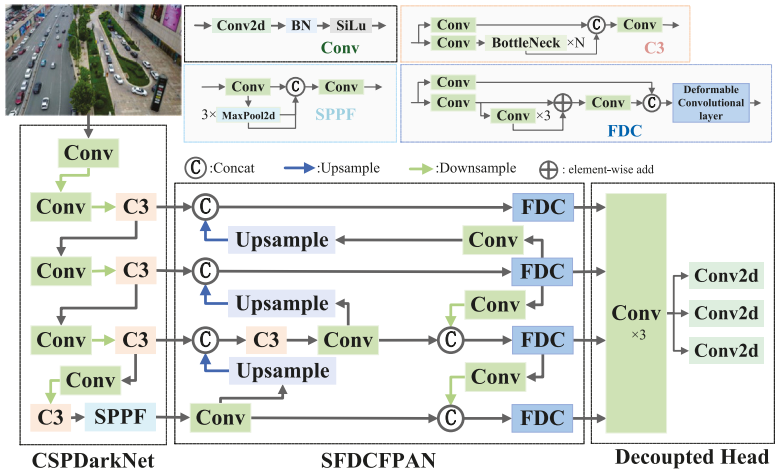

Fig. 3. Overall framework.

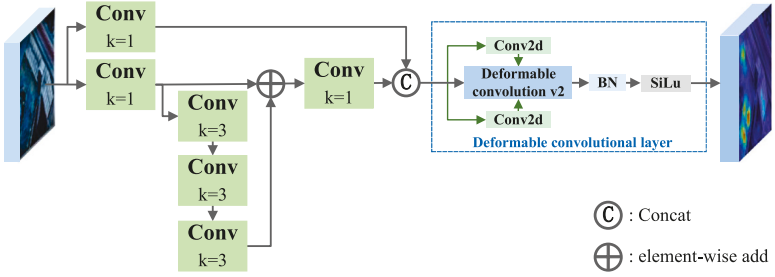

Fig. 4. FDC module structure.

enhancing model capabilities for modeling and feature extraction. In this process, the deformable convolutional layer generates the offset and modulation of the object using two 3 × 3 convolutions, respectively. These values are then passed to the deformable convolution v2, followed by BN and ReLU activations to enhance the model's expressiveness. Through this careful design, the FDC module combines the efficiency of CSPNet with the powerful geometric modeling capabilities of Deformable convents v2, which leverages deformable convolution layers to model object shapes and fully utilizes critical information extracted and preserved in the feature extraction branch to strengthen detail capture and key part representation of small and occluded objects, improving feature representation quality.

3.3. SFDCFPAN design and structure

Commonly, object detection algorithms employ the feature pyramid network (FPN) (Lin et al., 2017) or path aggregation network (PAN) (Liu et al., 2018) as the feature fusion architecture to integrate information across various scales extracted from the backbone, improving its feature expression capabilities. FPN consists of a top-down branch, while PAN adds a bottom-up branch based on FPN to facilitate the comprehensive integration of multi-scale features. Compared to FPN, PAN significantly enhances feature fusion capability, serving as the primary feature fusion framework. However, PAN exhibits constraints in detecting small objects. This limitation arises from PAN

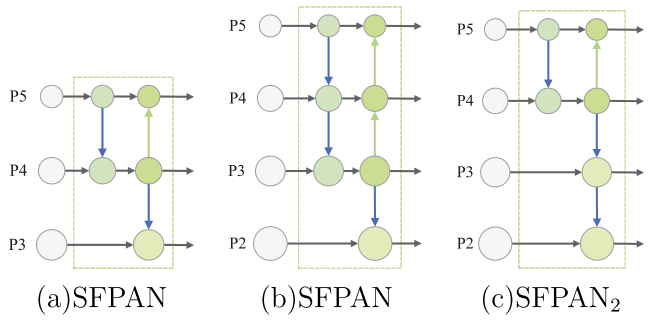

Fig. 5. SFPAN structures. (a) and (b) depict the SFPAN structures with three feature layers (P3 to P5) and four feature layers (P2 to P5), respectively, while (c) shows the structure that adds a P2 feature layer to the three-layer SFPAN, referred to as SFPAN$_2$.

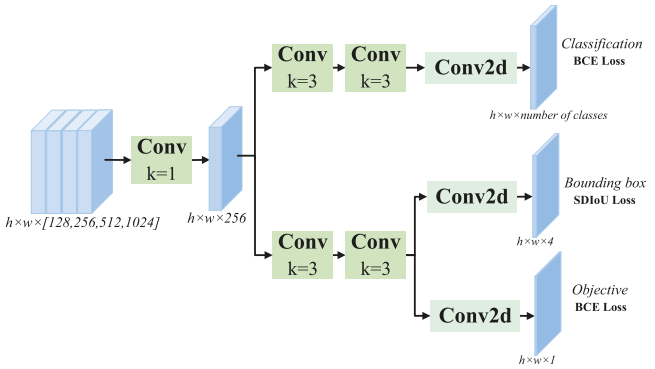

Fig. 6. Decoupled head structure.

directly downsampling the largest-sized feature layer and compressing the spatial position information of small-scale objects, leading to spatial information loss. To mitigate the model's spatial information loss, we design a neck network structure known as the selected feature path aggregation feature pyramid network (SFPAN), combining the advantages of both PAN and FPN. SFPAN structure is depicted in Fig. 5(a), which effectively fuses features of various scales and reduces the spatial information loss of the largest-sized feature maps, improving the model's localization capacity. Specifically, SFPAN cancels the downsampling operation on the largest-sized feature layer based on PAN to reduce spatial information loss due to downsampling the largest-sized feature layer and maximizes the retention of detailed information essential for detecting small and occluded objects, enhancing the detector's localization capability. Furthermore, SFPAN fuses the feature layers' features except for the largest-sized feature layer to enrich feature diversity and capture context information, improving the detector's representation capacity. Through this structural design, we preserve as much of the object's detailed spatial information and enhance feature richness by fusing contextual information of different scales, enhancing the detector's adaptability in various complex scenarios. To further evaluate the effectiveness of the SFPAN structure, Fig. 5(b) and (c) illustrate two versions of the SFPAN structure comprising four feature layers. Fig. 5(b) indicates that the SFPAN removes the downsampling of the largest-sized feature layer, specifically the P2 feature layer. In contrast, Fig. 5(c) depicts the addition of a P2 feature layer to the original SFPAN with three feature layers, referred to as SFPAN$_2$. Subsequent experimental results demonstrate that SFPAN exhibits higher detection efficiency than SFPAN$_2$. Specifically, eliminating the downsampling of the largest-sized feature layer enables SFPAN to attain optimal detection performance.

To develop a more robust feature fusion framework addressing the challenges of detecting small and occluded objects in UAV images, we introduced key improvements based on SFPAN, which employ the FDC module as the feature extraction module at the output position of the four feature layers in SFPAN, denoting this modified architecture as SFDCFPAN. The detailed structure of SFDCFPAN is illustrated in Fig. 3, which consists of SFPAN and FDC modules. SFDCFPAN enhances the preservation of object spatial information by employing the SFPAN structure and uses the features extracted by the FDC module as the output of each feature layer. With this design, SFDCFPAN preserves SFPAN's powerful multi-scale feature fusion capability and integrates the refined feature extraction and geometric adaptive modeling advantages introduced by the FDC module, improving detection accuracy for small and occluded objects. Consequently, as a unified feature fusion framework, SFDCFPAN systematically integrates the FDC module into the multi-scale output position of SFPAN, enhances the detector capabilities of feature extraction and modeling, and alleviates missed detections in small and occluded object scenarios, bolstering detection performance.

3.4. Decoupled head design and structure

The structure of the detector head typically falls into two categories: coupled head and decoupled head. If the detection head is designed as a coupled head structure, the detector must learn object classification and localization simultaneously during training, rendering it challenging for the model to train. Due to the conflicting nature of the object classification and localization tasks, where the localization task demands spatial information to locate, while the classification task relies more on semantic context to classify, the coupled head exacerbates task conflict, hindering the model from striking a balance between the two tasks and complicating model optimization. To mitigate the conflict between the two tasks and minimize the mutual interference between classification and localization, we introduce the decoupled head to decouple the detection head using multiple convolutional layers, as depicted in Fig. 6. Initially, a 1×1 convolutional layer is applied to adjust the number of feature channels. Subsequently, the modified features are sent to the classification and localization branches to accomplish the classification and localization tasks. The classification branch includes two 3×3 convolutional layers to capture semantic context information and a 1×1 convolutional layer to identify categories. The bounding box regression and the objective branches within the localization branch share two 3×3 convolutional layers that capture spatial information. The bounding box regression branch employs a 1×1 convolution to predict the bounding box, while the objective branch uses a 1×1 convolution to determine the objective confidence. In the loss function design, SDIoU Loss is employed for the bounding box regression loss in the localization task. This choice is caused by our detector's regression approach, which regresses based on the four distances from the upper-left and lower-right corners of the central cell containing the object to the bounding box. SDIoU Loss guides the detector in predicting accurate bounding boxes by minimizing the Euclidean distance between the predicted offsets and the ground truth. For classification and objectness confidence prediction, BCE Loss is used for optimization. BCE Loss facilitates more accurate classification and objective prediction. The structural design of the decoupled head cooperates with the effects of SDIoU Loss and BCE Loss, improving the training efficiency and overall performance of the detector.

4. Experiments and discussion

4.1. Experimental environment

The model optimizer utilized is Stochastic Gradient Descent (SGD) with a linear learning rate strategy. The initial learning rate is set to 0.01, with momentum of 0.937 and weight decay of 0.0005. Additionally, 200 epochs are configured to ensure that the model fully converges. The computational hardware employed is the NVIDIA 3060 graphics card. In the training strategy, we set the input image size

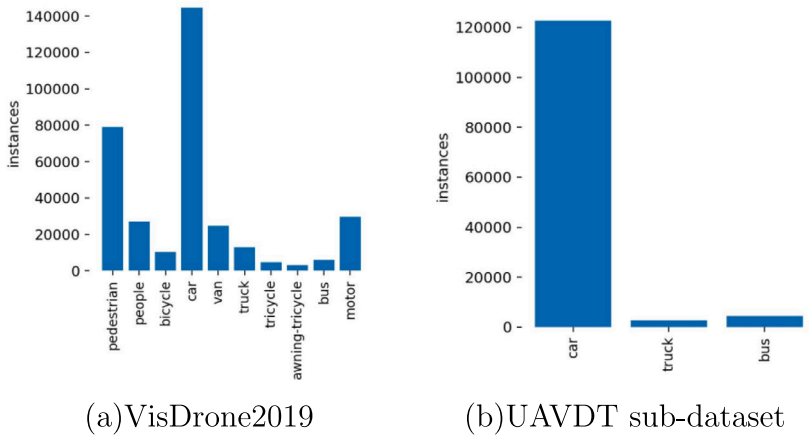

Fig. 7. Datasets label distribution chart. (a) and (b) show the category label distribution chart of VisDrone2019 and the UAVDT sub-dataset, respectively.

to 640 × 640 and the batch size to 8. Images are proportionally scaled, normalized, and converted from BGR to RGB format. We did not use pre-training and applied four data augmentation techniques: image flipping (horizontal flip with a probability of 0.5), random affine transformation (image translation with a probability of 0.1 and scaling with a ratio between 0.5 and 1.5), HSV image enhancement (gain factor of hue, saturation, and value adjustments of 0.015, 0.7, and 0.4, respectively), and mosaic. All experiments will be conducted in a standardized experimental environment, utilizing identical datasets and training strategies.

4.2. Datasets

In the experiment, we utilize VisDrone2019 and UAVDT sub-datasets as experimental datasets. Given the primary focus of the UAVDT on UAV image object-tracking tasks, we employ its sub-datasets as auxiliary validation datasets. The following is a brief description of the two datasets.

The VisDrone2019 (Du et al., 2019) dataset, released by Tianjin University, is a public dataset dedicated to UAV vision research, comprising 6471 training images, 548 verification images, and 1610 testing images. The dataset encompasses a wide range of scenes and varying object densities, focusing on typical pedestrians and vehicles in traffic environments. Affected by the flight altitude and special shooting angle of the UAV, VisDrone2019 various detection scenarios involving small objects, occluded objects, dense objects, and multi-scale objects, with a highly imbalanced category distribution, as shown in Fig. 7(a). Additionally, complex and variable weather and lighting conditions cause significant variations in image quality, posing considerable challenges to accurate target detection in UAV images.

The UAVDT (Du et al., 2018) dataset is centered on vehicle traffic scenes captured from UAV perspectives and comprises 50 video sequences. It targets high-altitude vehicle detection and tracking tasks, encompassing diverse weather conditions, flight altitudes, and viewing angles, supporting research on UAV-based object detection and tracking algorithms. Given its primary focus on object tracking, we extracted a sub-dataset of 8745 images featuring various scenarios from 24 UAVDT video sequences. Specifically, we randomly chose 7027 images from 16 video sequences as the training set and selected 1718 images from the remaining sequences as the validation and test sets, with a validation set of 418 images and a test set of 1300 images. The distribution chart depicted in Fig. 7(b) demonstrates the uneven distribution of category labels in the UAVDT sub-dataset with a concentration on the car category.

4.3. Evaluation indicators

Detection scenes from the drone perspective are complex and diverse, with varying levels of difficulty across categories. Average precision (AP) measures detection performance for a single category in UAV images by calculating the area under the precision–recall curve. Mean average precision (mAP) evaluates the overall detection performance across all categories, providing a comprehensive reflection of the detector's robustness and adaptability under varying conditions. The formulas for AP and mAP can be defined as follows: P, R, and N denote Precision, Recall, and category number, respectively. Precision measures the classification ability, while recall assesses its capability to locate.

$$AP = \int_0^1 P(R)\, dR \tag{1}$$

$$mAP = \frac{\sum_{i=1}^{N} AP(i)}{N} \tag{2}$$

To comprehensively evaluate the performance of the detector in UAV images, we employ the following metrics: AP_{50} (Average Precision of various categories at an IoU value of 0.5), mAP_{50} (mean Average Precision of all categories at an IoU value of 0.5), mAP (mean Average Precision of all categories at the range of IoU values 0.5 to 0.95), frames per second (FPS), and parameters. FPS indicates the detector's processing speed in UAV image detection, with real-time detection requiring FPS higher than 25. Parameters measure the complexity and computational resource requirements of the detector.

4.4. Ablation experiments and discussion

We conduct ablation experiments on our detector to validate the effectiveness of our proposed module. All experiments were performed using the VisDrone2019 test-set. We chose YOLOv5-L and ObjectBox as experimental benchmarks and configured the decoupled head and SFPAN to assess the models' detection speed. SFPAN performs feature fusion at the P2, P3, P4, and P5 layers. The experimental outcomes are presented in Table 1, which reveals that incorporating two modules into YOLOv5 significantly improves its accuracy, with the mAP_{50} by 4.2 percentage points from 34.5% to 38.7%. However, it unfortunately results in a notable decrease in detection speed, rendering it incapable of meeting real-time requirements. Conversely, ObjectBox enables maintaining a high detection speed while improving accuracy. Therefore, we select ObjectBox as the experimental benchmark and conduct experiments within the ObjectBox detection framework.

Table 1
Benchmark experiments.

Models	mAP_{50}	mAP	FPS
ObjectBox	33.7%	18.8%	**42.6**
ObjectBox+Decoupled Head+SFPAN	**37.3%**	**21.0%**	32.9
YOLOv5-L	34.5%	19.0%	**44.3**
YOLOv5-L+Decoupled Head+SFPAN	**38.7%**	**21.4%**	23.6

Table 2
Ablation experiments in the VisDrone2019.

Models	mAP_{50}	mAP	FPS	FLOPs
ObjectBox	33.7%	18.8%	**42.6**	**203.7**
+Decoupled Head	35.0%	19.5%	37.5	245.2
+Decoupled Head+SFPAN	37.3%	21.0%	32.9	280.6
+Decoupled Head+SFDCFPAN	**38.4%**	**21.8%**	29.5	281.1

Table 3
Ablation experiments in the UAVDT sub-dataset.

Models	mAP_{50}	mAP	FPS	FLOPs
ObjectBox	42.2%	26.6%	**44.4**	**203.7**
+Decoupled Head	42.5%	27.0%	37.2	245.1
+Decoupled Head+SFPAN	43.3%	27.1%	32.3	280.5
+Decoupled Head+SFDCFPAN	**44.2%**	**27.9%**	30.7	281.0

4.4.1. Overall ablation experiment

We conduct a comprehensive ablation experiment on our detector, systematically integrating various modules to analyze their contributions to the model's performance. The outcomes of ablation experiments in the VisDrone2019 test-set are presented in Table 2. Initially, we introduce a decoupled head structure on ObjectBox to diminish the interference between the classification and localization tasks. The experimental findings prove the decoupled head structure's usefulness, and the model's mAP_{50} increased from 33.7% to 35.0%, reflecting a 1.3 percentage point increase. Subsequently, we adopt SFPAN with four detection layers as the neck network. SFPAN stands out due to its elimination of the downsampling operation within the largest-sized layer, mitigating spatial information loss and augmenting the model's detection capability. The experimental results demonstrate SFPAN's effectiveness, and the model's mAP_{50} increases from 35.0% to 37.3%, marking a 2.3 percentage point increase. Next, we replace SFPAN with SFDCFPAN as the neck. SFDCFPAN incorporates the FDC module into the output positions of the four feature layers in SFPAN to improve detection performance. With its superior feature extraction and modeling capabilities, the FDC module enables the precise capture of objects' spatial information to enhance the model's feature extraction and modeling capabilities. By configuring the FDC module, SFDCFPAN can localize objects more accurately than SFPAN. Experimental findings indicate that the detector achieves optimal performance by integrating SFDCFPAN, increasing the mAP_{50} by 1.1 percentage points from 37.3% to 38.4%.

To further validate the efficacy of modules integrated into our detector, we perform a series of ablation experiments on the UAVDT sub-dataset, detailed in Table 3. The experiment outcomes show the detector performance improvement brought by our modules, with the mAP_{50} by 2.0 percentage points from 42.2% to 44.2%. Meanwhile, it meets the real-time requirements with an FPS of 31, reaffirming the outstanding performance of our method.

Through overall ablation experiments, we successfully validate the proposed modules' effectiveness, which significantly bolsters the detection capabilities while maintaining real-time performance. Besides, the visualization ablation experiment chart during training on the VisDrone2019 and UAVDT sub-dataset depicted in Fig. 8 proves the stability and efficiency of our detector in the training phase. Fig. 9 presents the precision–recall curve chart of our detector on the test-set of VisDrone2019 and UAVDT sub-dataset, providing an intuitive basis for performance evaluation and highlighting the exceptional performance of our detector.

Table 4
Comparative experiments of different feature extraction modules in SFPAN on VisDrone2019.

C3 module	C3 with DCN	FDC module	mAP_{50}	mAP	FPS	Params	FLOPs
✔	✗	✗	37.3%	21.0%	**32.9**	**53.6M**	280.6
✗	✔	✗	38.0%	21.6%	29.1	64.7M	**280.3**
✗	✗	✔	**38.4%**	**21.8%**	29.5	64.0M	281.1

4.4.2. FDC module ablation experiment

To evaluate the effectiveness of the FDC module, we sequentially configure the feature extraction modules of the four feature output layers within SFPAN to the C3 module, the C3 module with a deformable convolutional layer (C3 with DCN), and the FDC module. C3 with DCN, an extension of the C3 module, replaces the final 1×1 convolutional layer with a 3×3 deformable convolutional layer to enhance the model's object spatial information acquisition capability. The detailed experimental outcomes are presented in Table 4. Experimental findings indicate that the conventional convolution employed in the C3 module struggles to extract features with object spatial information in scenarios involving small and occluded objects, resulting in low detection accuracy. In contrast, the FDC module and C3 with DCN significantly enhance the model's competence in feature extraction containing object spatial information by introducing a deformable convolutional layer, with models' mAP_{50} reaching 38.4% and 38.0%, respectively. Experimental data show that the model configured with the FDC module achieves higher accuracy than the C3 with DCN configuration, which is 0.4% higher and a little faster, showcasing the FDC module's powerful feature extraction capabilities. This advantage arises from its extracting local image features with three 3×3 convolutional layers while preserving critical information through element-wise feature addition. This design enables the FDC module to accurately capture features that include semantic and spatial information, demonstrating its superior feature extraction performance. The experiments validate the effectiveness of the FDC module in acquiring semantic and spatial feature information, markedly enhancing the model's feature extraction capabilities. To scrutinize the distinctive benefits of the FDC module in feature extraction, particularly its exceptional efficacy in handling small and occluded objects, Fig. 10 showcases heatmaps generated by different feature extraction modules at SFPAN on the VisDrone2019 test-set. While the C3 with DCN excels in modeling small objects compared to the C3 module, its efficacy diminishes slightly in the presence of occluded objects. In contrast, the FDC module exhibits superior modeling capabilities. Whether discerning small or occluded objects, the FDC module accurately captures and models object shapes attributed to the unique feature extraction mechanism and the deformable convolutional layer. The heatmap results unequivocally validate the superior feature extraction capabilities of the FDC module for small and occluded objects, providing compelling visual substantiation.

To explore the performance benefits of the FDC module used in each feature layer of SFPAN, we conduct targeted experiments by sequentially replacing the feature extraction modules of the P2, P3, P4, and P5 output layers of SFPAN with FDC modules. Table 5 shows the detailed experimental outcomes. Interpretation of the findings reveals that replacing the feature extraction module of the P2 output layer in SFPAN with the FDC module leads to a decline in model performance, with the mAP decreasing from 37.3% to 36.7%. Although the FDC module enhances the spatial information richness of the P2 layer, the P2 output layer insufficiently fuses semantic and spatial information due to its exclusion from the feature fusion process in SFPAN. This deficiency may cause the model to confuse background spatial information with that of objects, leading to false detections

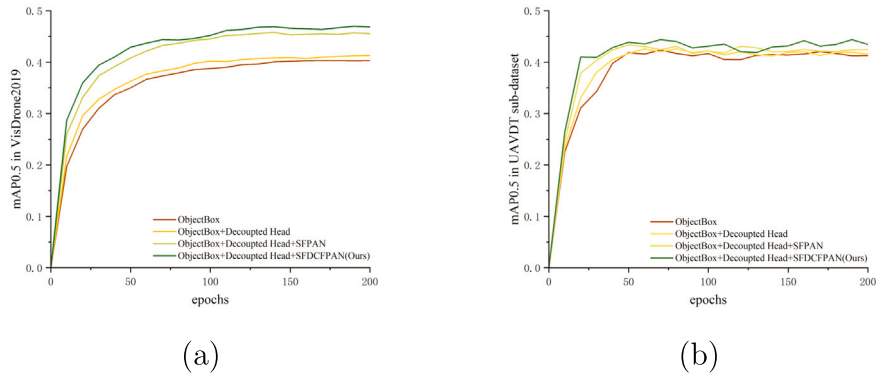

Fig. 8. Visualization training ablation experiment on VisDrone2019 and UAVDT sub-dataset. (a) VisDrone2019, (b) UAVDT sub-dataset.

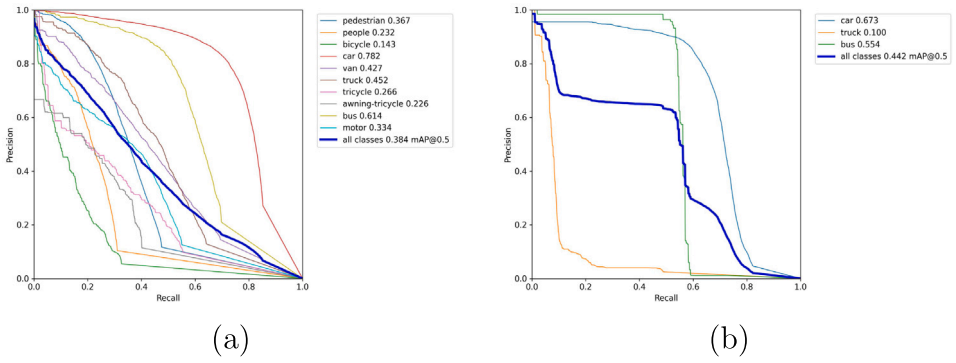

Fig. 9. PR curve chart of our detector in the test-sets of VisDrone2019 and UAVDT sub-dataset. (a) VisDrone2019, (b) UAVDT sub-dataset.

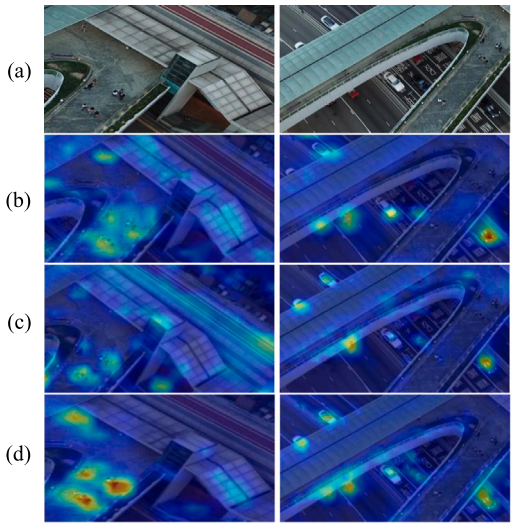

Fig. 10. Heatmaps using different feature extraction modules at SFPAN on VisDrone2019. (a) Original images, (b) C3 module, (c) C3 with DCN, (d) FDC module.

and ultimately reducing overall detection accuracy. Conversely, when feature extraction modules of the P2, P3, P4, and P5 output layers are replaced with FDC modules, forming SFDCFPAN, achieves optimal performance. This improvement occurs because the P3 output layer transmits features extracted from the FDC module that fuse abundant semantic and spatial information to the P2 output layer through the upsampling operation, significantly enhancing the feature expression capability of the P2 layer. Simultaneously, SFDCFPAN efficiently fuses features extracted by the FDC modules from P3, P4, and P5 output layers to augment feature representations, bolstering detection capability. Experimental data unequivocally demonstrate that employing FDC modules as feature extraction modules in four feature output layers of SFPAN achieves optimal detection performance, with mAP_{50} reaching 38.4%, mAP reaching 21.8%, and FPS reaching 30. Experiment results prove the effectiveness of the FDC module in four feature layers of SFPAN, which signifies a substantial improvement in accuracy while fulfilling real-time performance requirements.

4.4.3. Neck network ablation experiment

Most detection frameworks focus solely on detecting objects within the P3, P4, and P5 feature layers in PAN and FPN. It leaves the model vulnerable to missed detections and suboptimal performance in scenarios involving small objects. To ascertain the performance of SFDCFPAN, we undertake comprehensive comparative experiments using various feature fusion architectures. We sequentially employ FPN*, PAN*, SFPAN*, FPN, PAN, $SFPAN_2$, SFPAN, and SFDCFPAN as the model's neck network and meticulously document the experimental results, as presented in Tables 6 and 7. In this context, PAN*, FPN*, and SFPAN* represent that they perform detection tasks on the P3, P4, and P5 feature layers. PAN, FPN, and $SFPAN_2$ represent the addition of a P2 feature layer based on PAN*, FPN*, and SFPAN*, extending the detection feature layer's range and enhancing the model's detection capacity. While SFPAN stands for SFPAN* with four feature layers. Through a thorough analysis of experimental data, we observe substantial disparities in detection accuracy and speed among models employing diverse feature fusion architectures. When using FPN*, the model is prone to missed detections in scenarios with small-scale objects due to its constrained capability to detect them. Although FPN* demonstrates advantages in detection speed, it is challenging to meet the needs

Table 5
Comparative experiments of the FDC module at each detection layer of SFPAN in VisDrone2019.

P2 Layer	P3 Layer	P4 Layer	P5 Layer	mAP_{50}	mAP	FPS	Inference time	Params	FLOPs
✔	✗	✗	✗	36.7%	21.0%	**31.0**	**30.4 ms**	**53.8M**	281.8
✔	✔	✗	✗	38.0%	21.6%	29.5	32.0 ms	54.3M	282.2
✔	✔	✔	✗	37.5%	21.3%	29.7	31.8 ms	56.3M	281.8
✔	✔	✔	✔	**38.4%**	**21.8%**	29.5	32.0 ms	64.0M	**281.1**

Table 6
Comparative experiments using different feature fusion architectures in VisDrone2019.

Feature fusion architectures	mAP_{50}	mAP	FPS	FLOPs
FPN*	34.0%	18.9%	**38.5**	227.5
FPN	35.6%	19.8%	34.2	271.4
PAN*	35.0%	19.5%	37.5	245.2
PAN	37.1%	21.2%	29.9	290.5
Proposed SFPAN*	35.5%	19.6%	38.2	**226.5**
Proposed $SFPAN_2$	36.4%	20.3%	33.4	270.8
Proposed SFPAN	37.3%	21.0%	32.9	280.6
Proposed SFDCFPAN	**38.4%**	**21.8%**	29.5	281.1

Table 7
Comparative experiments using different feature fusion architectures in UAVDT sub-dataset.

Feature fusion architectures	mAP_{50}	mAP	FPS	FLOPs
FPN*	42.1%	26.3%	**38.8**	227.4
FPN	42.8%	26.6%	33.7	271.4
PAN*	42.5%	27.0%	37.2	245.1
PAN	43.0%	27.2%	30.1	290.5
Proposed SFPAN*	42.7%	26.8%	38.5	**226.4**
Proposed $SFPAN_2$	42.9%	27.0%	33.0	270.7
Proposed SFPAN	43.3%	27.1%	32.3	280.5
Proposed SFDCFPAN	**44.2%**	**27.9%**	30.7	281.0

of practical applications because of its limited detection capacity for small objects. When using FPN, FPN enhances the model's detection accuracy to a certain extent by adding a P2 feature layer based on FPN. However, FPN only performs feature fusion on a top-down branch, and it cannot comprehensively extract and fuse features of small-scale objects, resulting in lower detection accuracy compared with the model using PAN. When using PAN, PAN fuses features through top-down and bottom-up branches, significantly strengthening the feature expression capability. However, due to the high parameter count, the speed of model detection with PAN is relatively slow. To balance detection speed and accuracy, we adopt SFPAN* as the feature fusion architecture in our approach. SFPAN* cancels the P3 layer's downsampled operation to reduce spatial information loss, but it acquires less spatial information for detecting small and occluded objects due to its only three feature layers. $SFPAN_2$ adds a P2 layer based on SFPAN*, but only fuses the features of the P4 and P5 layers. SFPAN enhances detection accuracy by adding a P2 layer and eliminating the downsampling of the P2 layer to mitigate the loss of spatial information for small-scale and occluded objects and further fuses features from the P3, P4, and P5 layers through a bottom-up branch. Subsequently, we use SFDCFPAN as the feature fusion architecture. SFDCFPAN improves object location accuracy by incorporating FDC modules into four feature layers in SFPAN to extract features that contain object shape information, making the model obtain the highest detection accuracy. Experimental outcomes validate the advantage of SFDCFPAN, which improves the model's detection performance and meets in real-time.

To further validate the exceptional performance of SFDCFPAN in detecting small and occluded objects, Fig. 11 provides heatmaps generated using various feature fusion architectures on the VisDrone2019 test-set. The first row depicts the original images, while the subsequent rows display heatmaps generated using different feature fusion architectures: FPN*, FPN, PAN*, PAN, SFPAN*, SFPAN, and SFDCFPAN. The left side of Fig. 11 depicts scenarios with small objects under diverse lighting conditions. The heatmaps demonstrate that SFDCFPAN accurately identifies and locates small-scale pedestrians compared to other feature fusion architectures, enhancing the model's detection performance for small-scale objects, elevating detection accuracy, and minimizing missed detections. The right side of Fig. 11 showcases scenes with occlusions in varying lighting conditions. The heatmaps reveal SFDCFPAN's robust modeling capabilities in capturing and pinpointing occluded vehicle shapes against complex backgrounds. SFDCFPAN accurately locates vehicle positions where they are occluded by overpass or resemble the road color, enhancing the model's detection performance for occluded objects. A comparative analysis of heatmaps from different feature fusion architectures proves the significant advantages of SFDCFPAN in feature extraction and modeling. In situations involving small and occluded objects, SFDCFPAN effectively concentrates on all objects within the images, demonstrating outstanding feature extraction and modeling capabilities, mitigating missed detections, and improving the overall model performance.

4.4.4. Generalization ablation experiment

To comprehensively validate the effectiveness of SFPAN and SFDCFPAN within various detection frameworks, we chose the prevalent object detection frameworks YOLOv5, YOLOv8, and YOLOv11 as references. In this context, SFPAN denotes the feature fusion framework that executes detection tasks in the four feature layers: P2, P3, P4, and P5. Subsequently, we integrate SFPAN and SFDCFPAN into three frameworks individually and conduct a series of comparative experiments to assess performance at VisDrone2019 and UAVDT sub-dataset, shown in Tables 8 and 9. The modules in the two tables represent the feature extraction modules at each feature layer output position in the neck of the detection algorithm. The experiment outcomes in Table 8 illustrate significant improvements in the accuracy of detectors incorporating SFPAN or SFDCFPAN on the VisDrone2019 test-set. Particularly in detecting small objects, SFDCFPAN exhibits outstanding performance using FDC modules to model objects' shapes, enhancing the detectors' accuracy of small objects. To further verify the effectiveness of SFDCFPAN and SFPAN in detecting small objects within various detection frameworks, Fig. 12 provides heatmaps generated using different detection algorithms on the VisDrone2019 test-set. The first row depicts the original images, while subsequent rows show heatmaps generated using different detection algorithms: ObjectBox, our proposed detector, YOLOv5, SF-YOLOv5, SFD-YOLOv5, YOLOv8, SF-YOLOv8, SFD-YOLOv8, YOLOv11, SF-YOLOv11, and SFD-YOLOv11. The scenes depicted in Fig. 12 are all small object scenarios. The left side of the figure illustrates car detection in a dim environment, while the middle section presents pedestrian detection in a similarly dim setting. The right side shows pedestrian detection in an occluded scene. The heatmaps indicate that YOLOv5, YOLOv8, and YOLOv11, when configured with SFPAN and SFDCFPAN, significantly outperform the original detection algorithms in locating small objects, regardless of whether the scenes are dim or occluded. Notably, the algorithm configured with SFDCFPAN exhibits more substantial positioning capabilities, allowing for more accurate identification of small object locations compared to the algorithm configured with SFPAN. It demonstrates that in scenarios involving small and occluded objects, the SFDCFPAN assists the detector in accurately locating the object, reduces missed detections, and improves overall model performance. Besides, the experimental results presented in Table 9 confirm that the detection

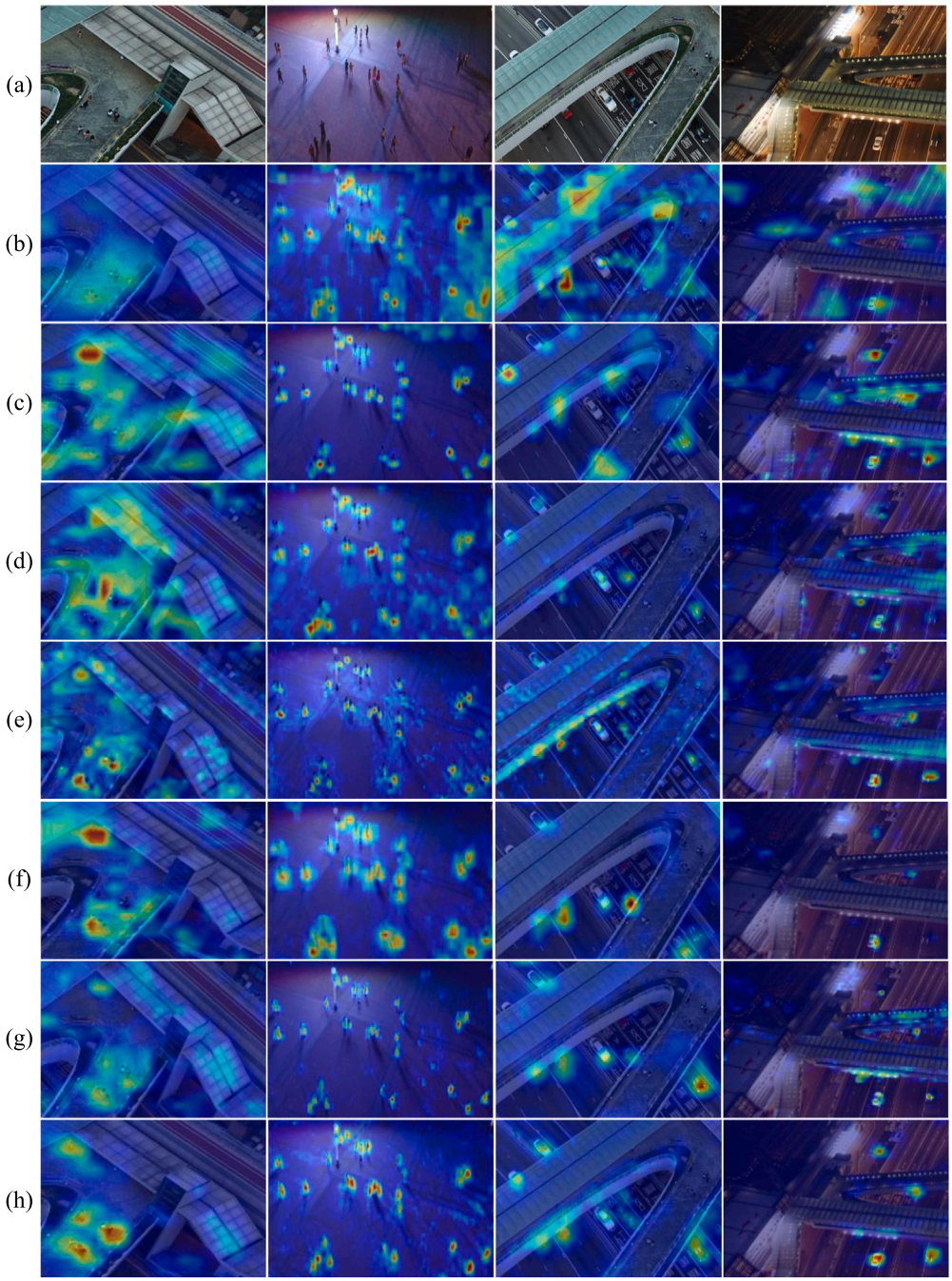

Fig. 11. Heatmaps using different feature fusion architectures on the VisDrone2019 test-set. (a) Original images, (b) FPN*, (c) FPN, (d) PAN*, (e) PAN, (f) SFPAN*, (g) SFPAN, (h) SFDCFPAN.

accuracy of the detectors utilizing SFPAN or SFDCFPAN on the UAVDT sub-dataset test-set is superior to the original detector. These findings unequivocally showcase the effectiveness of SFPAN and SFDCFPAN across diverse detection frameworks. By integrating these feature fusion architectures into existing frameworks, we can further elevate the performance of detectors.

4.4.5. Discussion

In this subsection, we will conduct an in-depth discussion of the above ablation experiments. In Section 4.4.1, we designed an ablation experiment to validate the effectiveness of our proposed method. This method introduces a decoupled head and proposes a distinctive feature fusion architecture SFDCFPAN consisting of a unique feature extraction module FDC and a feature fusion structure SFPAN. Since shared features can cause potential conflicts between classification and localization tasks, the decoupled head constructs independent branches for each task, effectively alleviating these conflicts and enabling each task to extract its required features independently. In feature fusion, our designed SFPAN combines the strengths of PAN and FPN. It enhances feature fusion and minimizes information loss caused by downsampling the highest-scale feature layer, significantly enriching the feature representation. Building on SFPAN, SFDCFPAN leverages the powerful feature modeling and extraction capabilities of the proposed FDC module, effectively improving the detection accuracy of small and occluded objects by providing high-quality features from SFPAN for deep extraction in FDC. The synergistic effect of these modules leads to a substantial improvement in detector performance.

Table 8
Performance of various detection frameworks integrated with SFPAN and SFDCFPAN on VisDrone2019.

Models	Modules	mAP_{50}	mAP	Small	FPS	Params
YOLOv5-L	C3	34.5%	19.0%	8.9%	**44.3**	**46.2M**
YOLOv5-L+SFPAN(SF-YOLOv5)	C3	37.1%	**20.1%**	10.9%	33.6	47.2M
YOLOv5-L+SFDCFPAN(SFD-YOLOv5)	FDC	**37.2%**	19.8%	**11.2%**	31.1	58.7M
YOLOv8-L	C2F	38.0%	22.2%	11.0%	**42.3**	43.6M
YOLOv8-L+SFPAN(SF-YOLOv8)	C2F	**41.0%**	**23.6%**	13.8%	36.5	42.9M
YOLOv8-L+SFDCFPAN(SFD-YOLOv8)	FDC	**41.0%**	**23.6%**	**14.0%**	35.3	**41.5M**
YOLOv11-L	C3k2	36.4%	21.8%	11.4%	**36.3**	**24.1M**
YOLOv11-L+SFPAN(SF-YOLOv11)	C3k2	38.1%	22.5%	13.0%	32.5	24.5M
YOLOv11-L+SFDCFPAN(SFD-YOLOv11)	FDC	**38.6%**	**22.7%**	**13.3%**	30.3	29.7M

Table 9
Performance of various detection frameworks integrated with SFPAN and SFDCFPAN on UAVDT sub-dataset.

Models	Modules	mAP_{50}	mAP	FPS	Params
YOLOv5-L	C3	41.8%	26.2%	**46.4**	**46.2M**
YOLOv5-L+SFPAN(SF-YOLOv5)	C3	42.5%	26.7%	38.8	47.2M
YOLOv5-L+SFDCFPAN(SFD-YOLOv5)	FDC	**43.0%**	**27.0%**	32.5	58.7M
YOLOv8-L	C2F	43.2%	27.4%	**45.6**	43.6M
YOLOv8-L+SFPAN(SF-YOLOv8)	C2F	44.1%	28.2%	40.3	42.9M
YOLOv8-L+SFDCFPAN(SFD-YOLOv8)	FDC	**44.6%**	**29.1%**	36.4	**41.5M**
YOLOv11-L	C3k2	43.4%	26.4%	**40.8**	**24.1M**
YOLOv11-L+SFPAN(SF-YOLOv11)	C3k2	44.0%	27.2%	35.7	24.5M
YOLOv11-L+SFDCFPAN(SFD-YOLOv11)	FDC	**44.3%**	**27.7%**	32.1	29.7M

In Section 4.4.2, we first verified the effectiveness of the FDC module we designed by comparing it with the C3 and C3 with DCN modules. Experimental results and heatmaps demonstrate that the FDC module, which incorporates a deformable convolution layer, captures local features through three successive 3 × 3 convolutional layers and employs element-wise addition to preserve key information in the feature extraction branch, superior to the C3 module and the C3 module with DCN in extracting features containing rich semantic and spatial information. We further analyzed the FDC module's performance gains across different feature layers of SFPAN. Results indicate that applying FDC to all four feature layers of SFPAN, which is the SFDCFPAN architecture, yields optimal detection accuracy. This advantage arises because FDC deeply extracts fused features from four feature layers to capture information beneficial for the accurate detection of objects, especially small objects and occluded objects, and transmits features containing this information to the detection head, significantly improving the detection accuracy.

In Section 4.4.3, we explored various feature fusion architectures and conducted a detailed analysis. Experimental results demonstrated the significant advantages of our proposed SFPAN and SFDCFPAN. SFPAN combines the strengths of FPN and PAN, enhancing feature fusion while reducing the risk of spatial information loss, outperforming both FPN and PAN. Building on SFPAN, SFDCFPAN integrates the FDC module for deep feature extraction. Through a deformable convolution layer and a specialized feature extraction structure, the FDC module captures spatial and semantic information, particularly for small and occluded objects, significantly enhancing the detection of these objects. Furthermore, the heatmap confirms that SFDCFPAN effectively improves our detector's capacity to localize small and occluded objects, reinforcing its advantages in practical applications.

In Section 4.4.4, to evaluate the generalization performance of the proposed feature fusion architectures SFPAN and SFDCFPAN, we conducted experiments using three mainstream frameworks: YOLOv5, YOLOv8, and YOLOv11, replacing their original feature fusion architectures. The results and heatmaps demonstrate that our two proposed feature fusion architectures significantly enhance detection performance on different detection frameworks, particularly in small object detection. SFDCFPAN retains SFPAN's advantages in enhancing feature fusion while reducing spatial information loss and integrating the modeling and extraction capabilities of the FDC module. It allows the detector to more accurately capture and identify features of small objects, significantly improving detection accuracy. These findings confirm that SFPAN and SFDCFPAN not only perform well within our detector but also exhibit strong generalization capabilities.

4.5. Comparative experiment

To verify the validity of our detector, we comprehensively compare it with the currently popular object detection algorithms. All experiments are conducted in a uniform environment, employing uniform data augmentation methods and training strategies, and tested on the VisDrone2019 test-set. The experimental results are depicted in Table 10. Upon meticulous comparison, we derive the following conclusions in CNN-based methods: In anchor-based detectors, two-stage detectors like Faster RCNN, Cascade RCNN, and Grid RCNN (Lu et al., 2019) exhibit slow detection speeds due to their complex detection processes and cannot meet real-time requirements. Single-stage detectors like ATSS (Zhang et al., 2020), YOLOv5, and YOLOv7 are influenced by the anchors' quantity, scale, and aspect ratio, leading to missed detections, particularly with small objects, diminishing detection accuracy. In anchor-free detectors, FCOS detects objects of specific scales in various feature layers, which predisposes the model towards detecting large-scale objects, resulting in suboptimal performance for smaller objects. In addition, FCOS, TOOD, FoevaBox (Kong et al., 2020), and Sparse RCNN (Sun et al., 2021) use ResNet as their backbone to extract features. The small receptive field of ResNet hinders the extraction of global features, consequently impacting model performance. YOLOX enhances model performance using a decoupled head structure and CSPDarkNet, which boasts a larger receptive field. However, in small object scenarios, the SimOTA label assignment method of YOLOX results in an uneven distribution of positive and negative samples, resulting in lower detection accuracy. RTMDet enhances the capability to extract global features via large-depth convolution and simultaneously incurs additional computing time, leading to a slower detection speed. YOLOv6 employs EfficientRep to improve detection accuracy. However, YOLOv6 uses plenty of heavy-parameterized structures, which results in slow detection speed and difficulty in achieving a balance between speed and accuracy. PPYOLOE relies on data augmentation techniques during the training phase and yields unsatisfactory performance using conventional four augmentation methods. YOLOv8 enhances the model's feature extraction capabilities by integrating C2F modules. However, the reliance on standard convolutional operations within the C2F architecture limits its adaptability to geometric variations in small and partially occluded objects, leading to suboptimal feature representation and reduced detection precision. This limitation persists across subsequent iterations: YOLOv9 employs GELAN, YOLOv10 adopts CIB, and YOLOv11 utilizes C3k2. Despite structural differences, all these modules inherit the fundamental constraint of static convolutional kernels that fail to capture spatial details critical for small object detection dynamically. The coupled head structure of ObjectBox complicates achieving a balance between classification and localization tasks, thus limiting detection accuracy. In Transformer-based methods, the DETR series of variants such as Deformable DETR,

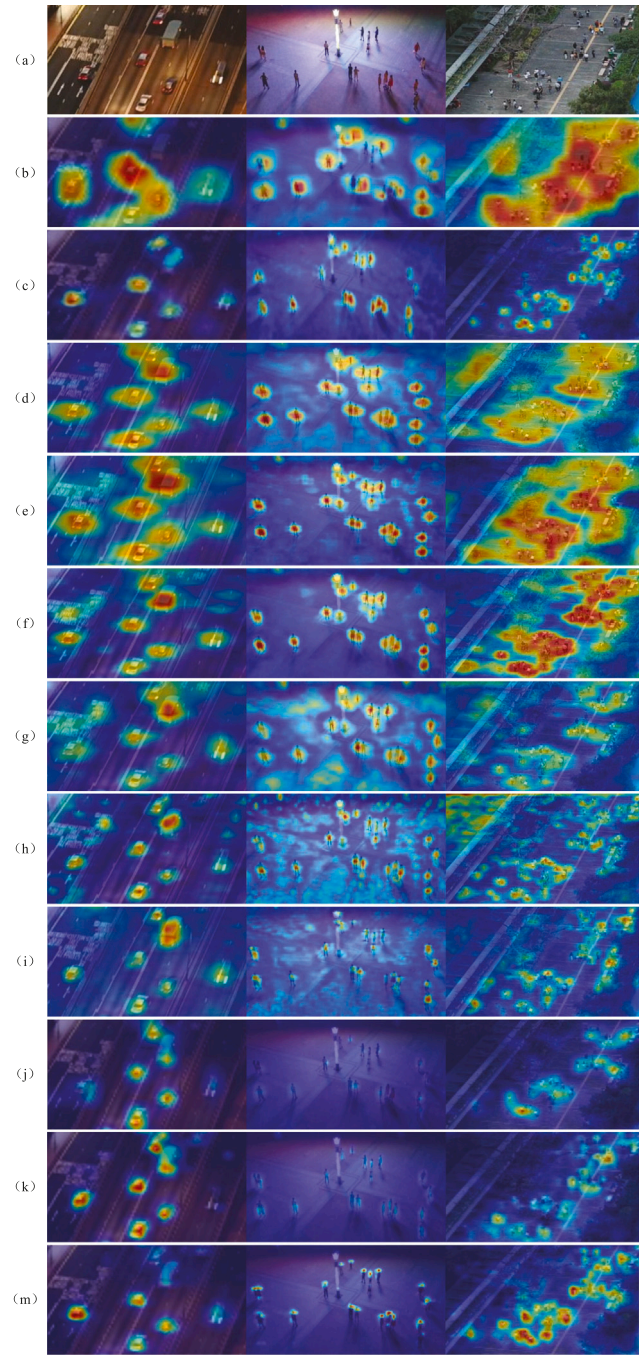

Fig. 12. Heatmaps using different detection algorithms on the VisDrone2019 test-set. (a) Original images, (b) ObjectBox, (c) ours, (d) YOLOv5, (e) SF-YOLOv5, (f) SFD-YOLOv5, (g) YOLOv8, (h) SF-YOLOv8, (i) SFD-YOLOv8, (j) YOLOv11, (k) SF-YOLOv11, (m) SFD-YOLOv11.

DINO, Align-DETR, RT-DETR, RT-DETRv2, RT-DETRv3, and DEIM effectively model long-range dependencies between objects by employing the attention mechanism. These variants have introduced various optimizations to address the defects of the DETR in both detection accuracy and convergence speed. However, they have not fully resolved a fundamental limitation of the DETR architecture: poor detection of small objects. This limitation arises primarily because small objects are sparsely represented on feature maps, and these methods are not integrated with feature fusion frameworks that include semantic information and spatial details across multiple scales. Consequently, these detectors struggle to capture critical details of small objects, leading to a high miss rate. To solve detection challenges of small and occluded objects, we propose a detector based on CNN that uses a decoupled head structure with a robust feature fusion architecture SFD-CFPAN. The experiment result shows that our detector achieves a high detection accuracy on the VisDrone2019 test-set while meeting real-time requirements, with mAP_{50} reaching 38.4%, representing a 2.9% improvement over ObjectBox. Besides, we integrated SFDCFPAN into YOLOv5, YOLOv8, and YOLOv11, named SFD-YOLOv5, SFD-YOLOv8, and SFD-YOLOv11 respectively. SFD-YOLOv5 achieved a 37.2% on the mAP_{50} indicator, reflecting a 2.7 percentage point enhancement compared to the YOLOv5-L. SFD-YOLOv8 achieves an excellent result of 41.0% on the mAP_{50} indicator, marking a 3.0 percentage point improvement over YOLOv8-L. Similarly, SFD-YOLOv11 achieves 38.6% on the mAP_{50}, reflecting a 2.2 percentage point enhancement compared to the YOLOv11-L. The experiments showcase the broad applicability and effectiveness of SFDCFPAN across various YOLO versions. Table 11 shows the detection accuracy comparison between different detectors within ten categories on the VisDrone2019 test-set. Comparative experiments prove the superior performance of our proposed detectors.

4.5.1. Comparison of detection results

To display the effectiveness of our proposed method in mitigating false and miss detections, we conduct testing experiments using ObjectBox and our detector on scenarios involving small objects, occluded objects, dense objects, and multi-scale objects from the VisDrone2019 test-set. The visualized experimental results are illustrated in Fig. 13. Through comparative analysis, we have drawn the following conclusions: In the scenarios involving small objects depicted in Fig. 13(a), ObjectBox exhibits noticeable missed detection issues, failing to detect numerous small-scale pedestrians. On the contrary, our detector only missed detecting a few individuals under trees, showcasing its superior detection capability for detecting small objects. In the scenarios involving occluded objects depicted in Fig. 13(b), ObjectBox exhibits significant shortcomings in detecting motors, tricycles, and a truck obscured by trees, resulting in a high missed detection rate. The missed detection rate of our detector is markedly lower than that of ObjectBox, and it only failed to detect two tricycles under trees, indicating its superior processing capabilities in detecting occluded objects. In the scenarios involving dense objects, illustrated in Fig. 13(c), ObjectBox exhibits a high rate of missed detections, particularly for pedestrians and cars. Our detector demonstrates a low missed detection rate, with only a few pedestrians missing detection, showcasing superior performance in dense environments. In the scenarios involving multi-scale objects depicted in Fig. 13(d), ObjectBox detects pillars as pedestrians erroneously and fails to detect all buses in poorly illuminated areas. However, our detector can accurately detect all buses in dim lighting conditions, demonstrating its superior capability in detecting multi-scale objects in poorly illuminated areas. The visualization of detection results conclusively validates the effectiveness and superiority of our proposed methods that substantially enhance the model's detection capabilities across diverse scenarios and minimize false and missed detections, enhancing overall detection performance.

5. Conclusion

We propose a detector that integrates the feature fusion framework SFDCFPAN and the decoupled head to enhance the detection of small and occluded objects in UAV images. Our designed feature fusion architecture, SFDCFPAN, consists of SFPAN and FDC modules. Specifically, SFPAN efficiently fuses multi-scale feature information while minimizing spatial information loss. The FDC module markedly enhances the detector's capability to extract spatial information for small and occluded objects. Consequently, SFDCFPAN enables the effective extraction of features on different scales to boost detection capability and performance, reducing missed detections. Besides, we introduce the

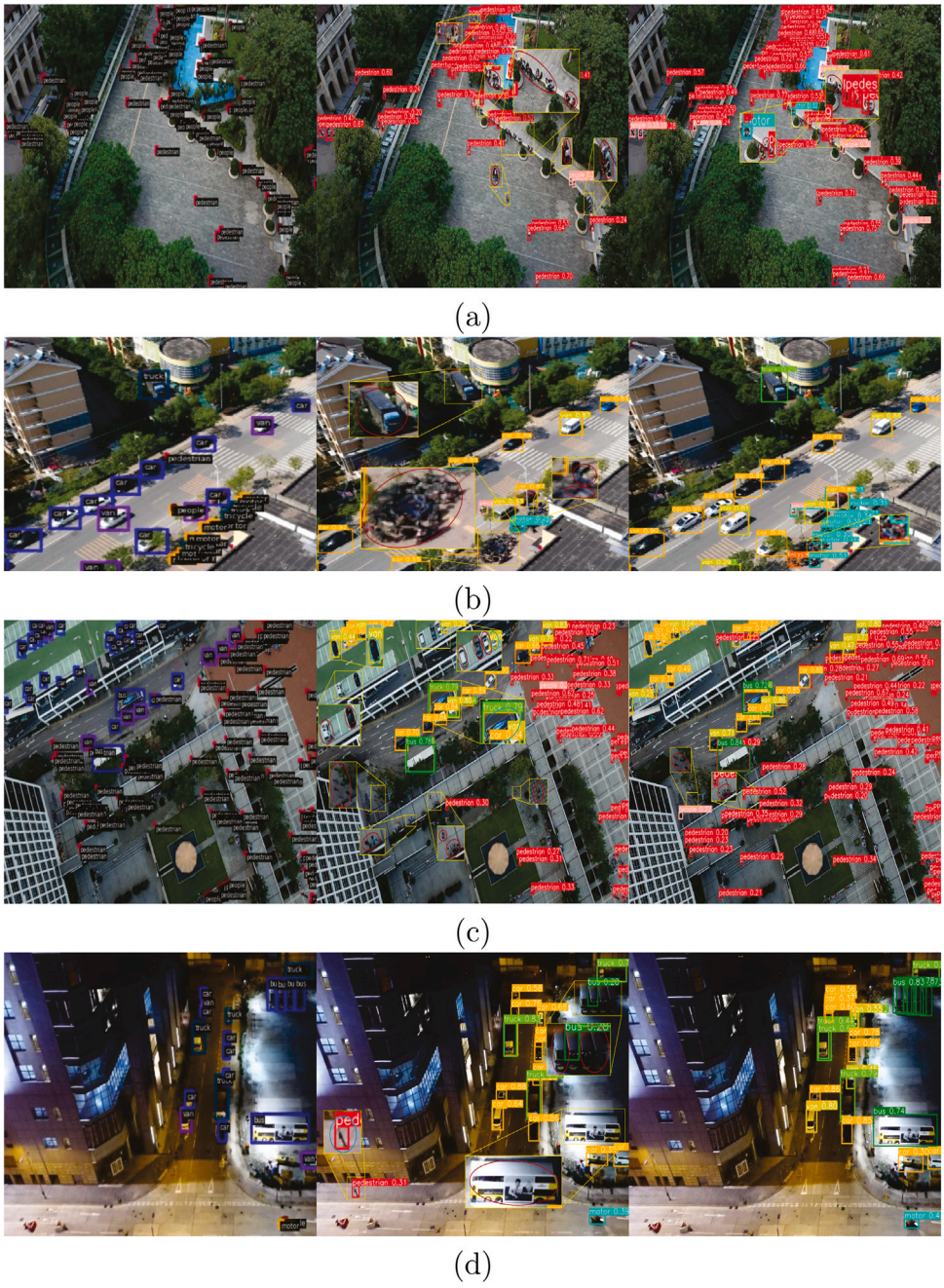

Fig. 13. Visualization of different detectors' detection results on the VisDrone2019 test-set. The left shows the original images, the center displays the ObjectBox detection results, and the right presents our method detection results. Red circles represent the missed detection, and blue circles represent the false detection. (a) In small objects scenarios, (b) In occluded objects scenarios, (c) In dense objects scenarios, and (d) In multi-scale objects scenarios.

decoupled head to alleviate mutual interference between classification and localization tasks by separating the two tasks, further improving detector performance. The experimental outcomes conclusively showcase that our proposed detector attains an outstanding performance on the VisDrone2019 and the UAVDT sub-dataset by employing the decoupled head structure and the robust neck network SFDCFPAN. It is noteworthy that SFPAN and SFDCFPAN can be readily integrated into existing detector frameworks. These detectors can meet real-time requirements while maintaining high detection accuracy, achieving a commendable balance between speed and precision.

Despite significant improvements in detection performance, our detector has a relatively high parameter count and increases model complexity and computational cost. Consequently, our future research will focus on decreasing the parameters of our detector to meet the lightweight requirements of UAV applications. To achieve this goal, we will explore the following technical routes: neural architecture search (NAS) (Zoph and Le, 2016) and pruning technology. NAS automatically searches for the optimal network structure within preset resource constraints to reduce parameters and computational complexity. Pruning technology for model optimization can eliminate redundant parameters and maintain high detection accuracy.

Table 10
Comparative experiments of different detectors on the VisDrone2019 test-set.

Models	Backbone	mAP_{50}	mAP	Small	Medium	Large	FPS	Params
Faster RCNN (Ren et al., 2015)	ResNet101	14.1%	7.8%	2.3%	12.7%	18.0%	15.8	60.4M
Cascade RCNN (Cai and Vasconcelos, 2018)	ResNet101	15.4%	8.9%	2.7%	14.6%	21.3%	13.1	88.2M
Grid RCNN (Lu et al., 2019)	ResNet101	12.8%	7.6%	2.3%	12.2%	19.2%	10.6	–
FCOS (Tian et al., 2020)	ResNet101	11.1%	5.5%	1.5%	8.6%	13.2%	**68.3**	51.1M
TOOD (Feng et al., 2021)	ResNet101	15.2%	8.2%	3.4%	13.1%	17.3%	68.0	51.0M
ATSS (Zhang et al., 2020)	ResNet101	11.8%	6.1%	2.1%	9.7%	12.8%	17.0	51.1M
FoevaBox (Kong et al., 2020)	ResNet101	10.2%	5.2%	1.2%	8.2%	12.9%	17.7	55.3M
Sparse RCNN (Sun et al., 2021)	ResNet101	14.3%	7.9%	3.6%	12.1%	15.8%	14.6	–
Deformable DETR (Zhu et al., 2020)	ResNet50	15.3%	7.4%	2.9%	11.8%	20.3%	15.7	40.1M
DINO (Zhang et al., 2022a)	ResNet50	16.7%	8.4%	3.7%	13.6%	20.3%	67.6	47.6M
Align-DETR (Cai et al., 2024)	ResNet50	21.5%	11.3%	5.1%	17.5%	27.9%	5.5	47.5M
RT-DETR (Zhao et al., 2024)	ResNet50	16.1%	8.3%	3.5%	12.7%	21.7%	16.4	40.8M
RT-DETRv2 (Lv et al., 2024)	ResNet50	16.7%	8.6%	3.7%	13.4%	23.4%	18.3	40.8M
RT-DETRv3 (Wang et al., 2025)	ResNet50	17.5%	9.3%	5.1%	13.6%	21.7%	30.7	43.5M
DEIM (Huang et al., 2025)	ResNet50	33.1%	18.7%	10.5%	27.1%	33.7%	23.9	40.1M
PPYOLOE-L (Xu et al., 2022)	CSPRepResNet	28.8%	16.3%	7.3%	25.1%	33.5%	27.6	51.4M
RTMDet-L (Lyu et al., 2022)	CSPDarknet	33.0%	19.6%	8.6%	30.5%	42.4%	24.3	52.3M
YOLOv5-S (Jocher, 2020)	CSPDarknet	27.5%	13.8%	5.7%	21.3%	34.0%	53.4	**7.1M**
YOLOv5-M (Jocher, 2020)	CSPDarknet	32.1%	17.2%	7.6%	26.7%	40.2%	48.7	20.9M
YOLOv5-L (Jocher, 2020)	CSPDarknet	34.5%	19.0%	8.9%	29.2%	41.2%	44.3	46.2M
YOLOv5-X (Jocher, 2020)	CSPDarknet	36.1%	20.2%	9.7%	31.0%	43.7%	32.2	86.3M
YOLOX-S (Ge et al., 2021)	CSPDarknet	34.4%	18.4%	9.3%	27.8%	35.1%	44.1	8.9M
YOLOX-M (Ge et al., 2021)	CSPDarknet	37.6%	20.7%	11.2%	30.8%	38.6%	39.6	25.3M
YOLOv6-S (Li et al., 2022)	EfficientRep	32.3%	18.4%	8.4%	28.3%	37.5%	27.5	17.2M
YOLOv6-M (Li et al., 2022)	EfficientRep	36.4%	21.0%	10.5%	31.9%	41.7%	23.6	34.2M
YOLOv6-L (Li et al., 2022)	EfficientRep	37.4%	21.8%	10.9%	33.3%	**48.7%**	22.1	58.5M
YOLOv7-L (Wang et al., 2023)	CSPDarknet	33.8%	18.0%	9.1%	26.9%	36.5%	46.5	36.6M
YOLOv8-S (Jocher, 2023)	CSPDarknet	33.1%	18.8%	8.8%	29.2%	39.7%	62.5	11.1M
YOLOv8-M (Jocher, 2023)	CSPDarknet	36.5%	21.1%	10.5%	32.5%	43.4%	51.6	25.9M
YOLOv8-L (Jocher, 2023)	CSPDarknet	38.0%	22.2%	11.0%	34.0%	43.1%	42.3	43.6M
YOLOv9-c (Wang et al., 2024c)	CSPDarknet	36.0%	21.4%	11.0%	30.7%	41.2%	36.8	24.2M
YOLOv10-L (Wang et al., 2024a)	CSPDarknet	36.3%	21.7%	10.2%	30.8%	43.2%	38.5	24.5M
YOLOv11-L (Khanam and Hussain, 2024)	CSPDarknet	36.4%	21.8%	11.4%	31.1%	40.1%	36.3	24.1M
ObjectBox (Zand et al., 2022)	CSPDarknet	33.7%	18.8%	10.1%	27.6%	36.6%	42.5	82.1M
Ours	CSPDarknet	**38.4%**	**21.8%**	13.0%	30.5%	41.2%	29.5	64.0M
SFD-YOLOv5(ours)	CSPDarknet	**37.2%**	**19.8%**	11.2%	28.8%	42.8%	31.1	58.7M
SFD-YOLOv8(ours)	CSPDarknet	**41.0%**	**23.6%**	**14.0%**	**34.8%**	40.2%	35.3	41.5M
SFD-YOLOv11(ours)	CSPDarknet	**38.6%**	**22.7%**	13.3%	31.4%	42.3%	30.3	29.7M

Table 11
Detection performance of various detectors across categories on the VisDrone2019 test-set.

Models	AP_{50}/%									
	Pedestrian	People	Bicycle	Car	Van	Truck	Tricycle	Awning-tricycle	Bus	Motor
Faster RCNN (Ren et al., 2015)	6.3	2.9	1.6	45.6	18.8	14.3	6.1	5.3	35.4	4.8
Cascade RCNN (Cai and Vasconcelos, 2018)	6.9	3.3	1.3	48.3	19.7	17.8	6.5	6.4	37.0	6.6
Grid RCNN (Lu et al., 2019)	5.9	2.4	0.5	43.8	16.6	12.5	4.8	4.2	32.1	4.5
FCOS (Tian et al., 2020)	5.4	2.0	1.1	42.1	13.5	10.1	3.2	2.1	27.8	3.4
TOOD (Feng et al., 2021)	10.2	4.4	2.6	51.3	18.3	11.4	4.9	6.4	33.2	9.4
ATSS (Zhang et al., 2020)	5.8	2.0	1.4	44.3	14.1	9.3	4.2	4.1	27.8	5.2
FoevaBox (Kong et al., 2020)	4.7	2.4	1.2	38.1	11.4	9.4	1.8	1.9	28.3	3.1
Sparse RCNN (Sun et al., 2021)	9.7	6.0	2.5	47.0	16.9	11.8	4.7	5.4	29.0	10.0
Deformable DETR (Zhu et al., 2020)	10.4	8.0	1.4	49.8	16.2	12.2	4.9	3.6	34.7	11.7
DINO (Zhang et al., 2022a)	11.6	9.2	2.9	52.6	17.5	12.1	8.0	6.4	32.9	13.3
Align-DETR (Cai et al., 2024)	15.8	12.6	5.7	59.1	21.8	18.2	11.5	10.6	41.5	18.7
RT-DETR (Zhao et al., 2024)	11.0	9.9	4.9	50.4	15.7	11.8	5.5	5.9	32.6	13.1
RT-DETRv2 (Lv et al., 2024)	12.4	10.4	4.7	52.5	13.5	13.6	7.2	5.8	32.8	14.1
RT-DETRv3 (Wang et al., 2025)	14.0	16.2	3.2	57.2	21.1	8.9	8.2	5.1	19.7	21.6
DEIM (Huang et al., 2025)	28.0	20.6	11.9	73.9	39.0	33.6	19.6	18.1	53.9	32.1
PPYOLOE-L (Xu et al., 2022)	22.3	10.1	8.2	69.0	36.2	35.5	13.1	15.2	53.9	24.6
RTMDet-L (Lyu et al., 2022)	18.8	7.7	7.4	69.0	40.0	38.8	14.8	17.0	55.7	20.8
YOLOv5-S (Jocher, 2020)	18.0	12.0	7.1	64.9	34.7	35.1	14.7	14.1	53.9	20.2
YOLOv5-M (Jocher, 2020)	22.0	13.9	10.0	69.4	39.8	42.1	20.9	18.7	59.0	25.4
YOLOv5-L (Jocher, 2020)	24.7	15.7	11.3	71.6	41.6	46.3	22.2	20.9	60.7	29.9
YOLOv5-X (Jocher, 2020)	25.7	16.3	12.9	72.8	43.0	49.3	24.8	22.5	61.8	32.1
YOLOX-S (Ge et al., 2021)	27.9	20.4	13.1	72.1	39.4	40.0	19.2	22.6	56.5	32.8

(*continued on next page*)

Table 11 (*continued*).

Models	AP_{50}/%									
	Pedestrian	People	Bicycle	Car	Van	Truck	Tricycle	Awning-tricycle	Bus	Motor
YOLOX-M (Ge et al., 2021)	32.2	23.3	15.0	74.8	43.7	44.1	21.6	23.7	60.7	36.5
YOLOv6-S (Li et al., 2022)	24.5	11.9	9.5	71.6	42.2	42.8	17.3	18.0	57.1	28.4
YOLOv6-M (Li et al., 2022)	28.2	13.9	13.4	74.4	45.7	47.5	23.0	23.4	61.3	32.9
YOLOv6-L (Li et al., 2022)	29.2	15.3	14.8	74.9	46.1	49.6	23.7	23.6	62.0	34.4
YOLOv7-L (Wang et al., 2023)	25.0	19.0	11.5	72.4	43.7	41.8	18.6	18.4	57.8	29.8
YOLOv8-S (Jocher, 2023)	25.4	12.7	11.7	71.4	40.3	41.6	19.4	20.5	58.0	30.0
YOLOv8-M (Jocher, 2023)	28.3	15.4	13.7	73.9	43.8	47.8	22.6	24.1	61.8	33.8
YOLOv8-L (Jocher, 2023)	29.4	15.8	15.4	75.5	46.5	**51.7**	24.1	22.7	**63.2**	35.7
YOLOv9-c (Wang et al., 2024c)	30.0	15.8	12.4	74.4	41.6	45.8	23.5	21.3	60.9	33.8
YOLOv10-L (Wang et al., 2024a)	30.0	17.0	12.1	75.0	43.2	46.3	21.3	23.7	61.4	32.5
YOLOv11-L (Khanam and Hussain, 2024)	30.7	16.0	13.1	75.2	43.8	46.1	22.9	22.6	60.8	33.1
ObjectBox (Zand et al., 2022)	29.5	16.0	10.7	73.9	39.8	41.7	18.9	19.4	58.9	28.2
Ours	36.7	23.2	14.3	78.2	42.7	45.2	26.6	22.6	61.4	33.4
SFD-YOLOv5(ours)	32.6	21.6	15.6	76.4	43.4	44.9	22.7	20.9	60.1	34.1
SFD-YOLOv8(ours)	**37.2**	**23.9**	**18.6**	**79.5**	**47.6**	46.9	**27.5**	24.0	61.5	**41.2**
SFD-YOLOv11(ours)	37.1	23.5	15.0	78.5	43.3	39.7	27.0	**24.3**	59.5	38.3

CRediT authorship contribution statement

Wanxia Huang: Writing – original draft, Software, Methodology, Investigation, Conceptualization. **Chaojun Dong:** Writing – review & editing, Supervision, Conceptualization. **Xiankun Liu:** Writing – review & editing, Software. **Ye Li:** Visualization, Software. **Yikui Zhai:** Writing – review & editing, Visualization, Supervision. **Kaitong Ou:** Visualization, Software. **Hao Quan:** Writing – review & editing, Visualization.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

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 study is supported by the 2017 Provincial Science and Technology Plan Fund project of the Guangdong Province of China (No. 2017A010101019).

Data availability

Data will be made available on request.

References

Bochkovskiy, A., Wang, C.-Y., Liao, H.-Y.M., 2020. Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934.

Cai, Z., Liu, S., Wang, G., Ge, Z., Zhang, X., Huang, D., 2024. Align-DETR: Enhancing end-to-end object detection with aligned loss. arXiv preprint arXiv:2304.07527.

Cai, Z., Vasconcelos, N., 2018. Cascade r-cnn: Delving into high quality object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6154–6162.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S., 2020. End-to-end object detection with transformers. In: European Conference on Computer Vision. Springer, pp. 213–229.

Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., Wei, Y., 2017. Deformable convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 764–773.

Dhillon, I.S., Guan, Y., Kulis, B., 2004. Kernel k-means: Spectral clustering and normalized cuts. In: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 551–556.

Ding, X., Zhang, X., Ma, N., Han, J., Ding, G., Sun, J., 2021. Repvgg: Making vgg-style convnets great again. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13733–13742.

Dong, X., Bao, J., Chen, D., Zhang, W., Yu, N., Yuan, L., Chen, D., Guo, B., 2022. Cswin transformer: A general vision transformer backbone with cross-shaped windows. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12124–12134.

Du, D., Qi, Y., Yu, H., Yang, Y., Duan, K., Li, G., Zhang, W., Huang, Q., Tian, Q., 2018. The unmanned aerial vehicle benchmark: Object detection and tracking. In: Proceedings of the European Conference on Computer Vision. ECCV, pp. 370–386.

Du, D., Zhu, P., Wen, L., Bian, X., Lin, H., Hu, Q., Peng, T., Zheng, J., Wang, X., Zhang, Y., et al., 2019. VisDrone-DET2019: The vision meets drone object detection in image challenge results. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops.

Feng, C., Zhong, Y., Gao, Y., Scott, M.R., Huang, W., 2021. Tood: Task-aligned one-stage object detection. In: 2021 IEEE/CVF International Conference on Computer Vision. ICCV, IEEE Computer Society, pp. 3490–3499.

Ge, Z., Liu, S., Wang, F., Li, Z., Sun, J., 2021. Yolox: Exceeding yolo series in 2021. arXiv preprint arXiv:2107.08430.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778.

Huang, S., Lu, Z., Cun, X., Yu, Y., Zhou, X., Shen, X., 2025. DEIM: DETR with improved matching for fast convergence. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 15162–15171.

Ioffe, S., Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International Conference on Machine Learning. pmlr, pp. 448–456.

Jiang, L., Yuan, B., Du, J., Chen, B., Xie, H., Tian, J., Yuan, Z., 2024. MFFSODNet: Multi-scale feature fusion small object detection network for UAV aerial images. IEEE Trans. Instrum. Meas..

Jocher, G., 2020. yolov5: https://github.com/ultralytics/yolov5.

Jocher, G., 2023. yolov8: https://github.com/ultralytics.

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

Kong, T., Sun, F., Liu, H., Jiang, Y., Li, L., Shi, J., 2020. Foveabox: Beyound anchor-based object detection. IEEE Trans. Image Process. 29, 7389–7398.

LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., 2002. Gradient-based learning applied to document recognition. Proc. IEEE 86 (11), 2278–2324.

Li, C., Li, L., Jiang, H., Weng, K., Geng, Y., Li, L., Ke, Z., Li, Q., Cheng, M., Nie, W., et al., 2022. YOLOv6: A single-stage object detection framework for industrial applications. arXiv preprint arXiv:2209.02976.

Li, C., Zhou, S., Yu, H., Guo, T., Guo, Y., Gao, J., 2024. An efficient method for detecting dense and small objects in uav images. IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens..

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

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

Lu, W., Lan, C., Niu, C., Liu, W., Lyu, L., Shi, Q., Wang, S., 2023. A CNN-transformer hybrid model based on CSWin transformer for UAV image object detection. IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens. 16, 1211–1231.

Lu, X., Li, B., Yue, Y., Li, Q., Yan, J., 2019. Grid r-cnn. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7363–7372.

Lv, W., Zhao, Y., Chang, Q., Huang, K., Wang, G., Liu, Y., 2024. Rt-detrv2: Improved baseline with bag-of-freebies for real-time detection transformer. arXiv preprint arXiv:2407.17140.

Lyu, C., Zhang, W., Huang, H., Zhou, Y., Wang, Y., Liu, Y., Zhang, S., Chen, K., 2022. Rtmdet: An empirical study of designing real-time object detectors. arXiv preprint arXiv:2212.07784.

Meng, L., Zhou, L., Liu, Y., 2023. SODCNN: A convolutional neural network model for small object detection in drone-captured images. Drones 7 (10), 615.

Mo, W., Zhang, W., Wei, H., Cao, R., Ke, Y., Luo, Y., 2023. PVDet: Towards pedestrian and vehicle detection on gigapixel-level images. Eng. Appl. Artif. Intell. 118, 105705.

Redmon, J., Divvala, S., Girshick, R., Farhadi, A., 2016. You only look once: Unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 779–788.

Redmon, J., Farhadi, A., 2017. YOLO9000: better, faster, stronger. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7263–7271.

Redmon, J., Farhadi, A., 2018. Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767.

Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. Adv. Neural Inf. Process. Syst. 28.

Song, G., Du, H., Zhang, X., Bao, F., Zhang, Y., 2024. Small object detection in unmanned aerial vehicle images using multi-scale hybrid attention. Eng. Appl. Artif. Intell. 128, 107455.

Sun, W., Dai, L., Zhang, X., Chang, P., He, X., 2022. RSOD: Real-time small object detection algorithm in UAV-based traffic monitoring. Appl. Intell. 1–16.

Sun, K., Xiao, B., Liu, D., Wang, J., 2019. Deep high-resolution representation learning for human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5693–5703.

Sun, P., Zhang, R., Jiang, Y., Kong, T., Xu, C., Zhan, W., Tomizuka, M., Li, L., Yuan, Z., Wang, C., et al., 2021. Sparse r-cnn: End-to-end object detection with learnable proposals. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14454–14463.

Tian, Z., Shen, C., Chen, H., He, T., 2020. FCOS: A simple and strong anchor-free object detector. IEEE Trans. Pattern Anal. Mach. Intell. 44 (4), 1922–1933.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Adv. Neural Inf. Process. Syst. 30.

Wang, C.-Y., Bochkovskiy, A., Liao, H.-Y.M., 2023. 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. pp. 7464–7475.

Wang, A., Chen, H., Liu, L., Chen, K., Lin, Z., Han, J., et al., 2024a. Yolov10: Real-time end-to-end object detection. Adv. Neural Inf. Process. Syst. 37, 107984–108011.

Wang, B., Jiang, P., Liu, Z., Li, Y., Cao, J., Li, Y., 2024b. An adaptive lightweight small object detection method for incremental few-shot scenarios of unmanned surface vehicles. Eng. Appl. Artif. Intell. 133, 107989.

Wang, C.-Y., Liao, H.-Y.M., Wu, Y.-H., Chen, P.-Y., Hsieh, J.-W., Yeh, I.-H., 2020. CSPNet: A new backbone that can enhance learning capability of CNN. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 390–391.

Wang, S., Xia, C., Lv, F., Shi, Y., 2025. RT-DETRv3: Real-time end-to-end object detection with hierarchical dense positive supervision. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision. WACV, IEEE, pp. 1628–1636.

Wang, C.-Y., Yeh, I.-H., Mark Liao, H.-Y., 2024c. Yolov9: Learning what you want to learn using programmable gradient information. In: European Conference on Computer Vision. Springer, pp. 1–21.

Wen, L., Du, D., Cai, Z., Lei, Z., Chang, M.-C., Qi, H., Lim, J., Yang, M.-H., Lyu, S., 2020. UA-DETRAC: A new benchmark and protocol for multi-object detection and tracking. Comput. Vis. Image Underst. 193, 102907.

Woo, S., Park, J., Lee, J.-Y., Kweon, I.S., 2018. Cbam: Convolutional block attention module. In: Proceedings of the European Conference on Computer Vision. ECCV, pp. 3–19.

Xu, S., Wang, X., Lv, W., Chang, Q., Cui, C., Deng, K., Wang, G., Dang, Q., Wei, S., Du, Y., et al., 2022. PP-YOLOE: An evolved version of YOLO. arXiv preprint arXiv:2203.16250.

Zand, M., Etemad, A., Greenspan, M., 2022. Objectbox: From centers to boxes for anchor-free object detection. In: European Conference on Computer Vision. Springer, pp. 390–406.

Zhang, S., Chi, C., Yao, Y., Lei, Z., Li, S.Z., 2020. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9759–9768.

Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L.M., Shum, H.-Y., 2022a. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. arXiv preprint arXiv:2203.03605.

Zhang, Y., Liu, Y., Kang, W., Tao, R., 2023a. Vss-net: Visual semantic self-mining network for video summarization. IEEE Trans. Circuits Syst. Video Technol. 34 (4), 2775–2788.

Zhang, X., Lu, T., Wang, J., Fu, S., Gao, F., 2024a. Small object detection by edge-aware neural network. Eng. Appl. Artif. Intell. 138, 109406.

Zhang, Y.-F., Ren, W., Zhang, Z., Jia, Z., Wang, L., Tan, T., 2022b. Focal and efficient IOU loss for accurate bounding box regression. Neurocomputing 506, 146–157.

Zhang, Y., Wu, C., Guo, W., Zhang, T., Li, W., 2023b. CFANet: Efficient detection of UAV image based on cross-layer feature aggregation. IEEE Trans. Geosci. Remote Sens. 61, 1–11.

Zhang, Y., Wu, C., Zhang, T., Zheng, Y., 2024b. Full-scale feature aggregation and grouping feature reconstruction based UAV image target detection. IEEE Trans. Geosci. Remote Sens..

Zhang, Y., Zhang, T., Wang, S., Yu, P., 2025b. An efficient perceptual video compression scheme based on deep learning-assisted video saliency and just noticeable distortion. Eng. Appl. Artif. Intell. 141, 109806.

Zhang, Y., Zhang, T., Wu, C., Tao, R., 2023c. Multi-scale spatiotemporal feature fusion network for video saliency prediction. IEEE Trans. Multimed. 26, 4183–4193.

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

Zhu, X., Hu, H., Lin, S., Dai, J., 2019. Deformable convnets v2: More deformable, better results. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9308–9316.

Zhu, S., Ji, L., Zhu, J., Chen, S., Ren, H., 2024a. Spatio-temporal fusion with motion masks for the moving small target detection from remote-sensing videos. Eng. Appl. Artif. Intell. 138, 109362.

Zhu, S., Ji, L., Zhu, J., Chen, S., Ren, H., 2024b. Spatio-temporal fusion with motion masks for the moving small target detection from remote-sensing videos. Eng. Appl. Artif. Intell. 138, 109362.

Zhu, X., Lyu, S., Wang, X., Zhao, Q., 2021. TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2778–2788.

Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J., 2020. Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159.

Zoph, B., Le, Q.V., 2016. Neural architecture search with reinforcement learning. arXiv preprint arXiv:1611.01578.