Expert Systems With Applications 299 (2026) 130240

Contents lists available at ScienceDirect

Expert Systems With Applications

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

# KSCNet: Exploring KAN and state space model collaboration network for small object detection from UAV imagery

Yi Li [a,b], Huiying Xu [a,b,*], Yiming Sun [c], Pengfei Zhu [d], Lingling Xu [e], Xinzhong Zhu [a,b]

[a] Key Laboratory of Intelligent Education Technology and Application, Zhejiang Normal University, Jinhua, Zhejiang, 321004, China
[b] School of Computer Science and Technology of Zhejiang Normal University, Jinhua, Zhejiang, 321004, China
[c] School of Automation, Southeast University, Nanjing, Jiangsu, 210096, China
[d] College of Intelligence and Computing, Tianjin University, Tianjin, 300350, China
[e] School of Computer Science of Hangzhou Dianzi University, Hangzhou, Zhejiang, 310018, China

ARTICLE INFO

*Keywords:*
Unmanned aerial vehicle
Small object detection
Kolmogorov-Arnold networks
State space model

ABSTRACT

Detecting small objects in aerial images taken by unmanned aerial vehicles (UAVs) has become a crucial research challenge in the field of computer vision. This challenge is attributable to the following primary factors: the small size of the targets, the complexity of the background and the inadequate feature fusion, which makes small targets more susceptible to limited effective information and inferior detection performance. To address this issues, we propose a collaboration network that integrates Kolmogorov-Arnold Networks (KAN) and State Space Model (SSM) to improve the small target detection performance from UAV imagery. Specifically, we employ the KAN inserted into original YOLO11 architecture as primary backbone for feature extraction, which is sufficient to decompose complex high-dimensional data into simple one-dimensional function combinations so as to efficiently explore features with strong expressive power. We design Semantic Aggregation Network (SAN) to perform highly-effective multiscale feature fusion of global patterns. The SSM module plays a crucial role in SAN, which has been demonstrated to exhibit a superior capacity to adapt to a variety of input data types through its distinctive scanning strategy and dynamic weighting mechanism, especially for the complicated UAV images. An efficient Depthwise Channel Attention (DCA) is developed to reduce the aliasing effect generated from fused feature via lightweight channel dimension refinement. Extensive experiments on the public UAV datasets have been conducted to validate the effectiveness of KSCNet. Concretely, KSCNet performs 0.844 $mAP@50$ and 0.691 $mAP@50:95$ on the SIMD dataset, achieving 1.69% and 2.22% accuracy improvement compared with baseline. Moreover, KSCNet also accomplishes 4.82% and 5.11% accuracy increase on the VisDrone validation set and 4.43% and 6.16% boost on the VisDrone test set at $mAP@50$ and $mAP@50:95$ respectively, indicating that the KSCNet demonstrates excellent performance in the UAV small object detection task, providing substantial technical support for applications in related domains.

## 1. Introduction

Recent years have seen a rapid development of UAV technology, the UAV remote sensing platform is assuming a progressively crucial role across diverse domains, including agricultural monitoring (Zhang et al., 2024b), search and rescue (Martinez-Alpiste et al., 2021), traffic managing (Wu et al., 2021), environmental monitoring (Motlagh et al., 2023), satellite monitoring (Gagliardi et al., 2023) etc. The advantages of this technology include its flexibility and mobility, low cost and high resolution of acquired data. The UAV platform can quickly acquire large-scale surface information, providing timely and accurate data support for decision-making in related fields (Zhang et al., 2021). However, the intrinsic attributes of UAV images, characterized by their extensive coverage, diminutive target scale and dense spatial distribution, present formidable challenges to conventional object detection methodologies, especially when it comes to the detection of small targets (Zhang et al., 2024a). This can be attributed to the limitations in the available information regarding target features, the presence of substantial background interference and the suboptimal detection accuracy.

* Corresponding author.
*E-mail addresses:* leeyee@zjnu.edu.cn (Y. Li), xhy@zjnu.edu.cn (H. Xu), sunyiming@seu.edu.cn (Y. Sun), zhupengfei@tju.edu.cn (P. Zhu), linglingxu@hdu.edu.cn (L. Xu), zxz@zjnu.edu.cn (X. Zhu).

https://doi.org/10.1016/j.eswa.2025.130240
Received 1 June 2025; Received in revised form 10 September 2025; Accepted 28 October 2025
Available online 2 November 2025

In the field of computer vision, small object detection in UAV remote sensing imagery has attracted considerable research attention due to its critical role in various applications (Li et al., 2017). This task remains highly challenging, primarily owing to three major obstacles: the small size of objects, complex environmental conditions such as varying illumination, and the stringent demand for computational efficiency in real-world deployment. Small objects in UAV-captured images often occupy only minimal pixel regions, making them easily indistinguishable from background clutter and texture noise. Moreover, inconsistent lighting conditions-including shadows, overexposure, and weather-related variations-further degrade the discernibility of such objects. These issues collectively hinder the performance of conventional detection models, necessitating more robust and adaptive feature extraction techniques (Wang et al., 2020). In addition to these perceptual challenges, the inherent limitations of UAV platforms-such as constrained computational resources and the need for real-time processing-impose strong requirements for lightweight and efficient model architectures. It is essential to develop detection systems that not only achieve high accuracy under challenging visual conditions, but also maintain low computational overhead. Therefore, the integration of deep learning approaches that balance performance and efficiency becomes imperative for enabling reliable on-board detection in UAV applications.

In recent years, object detection task has witnessed remarkable progress through advances in deep learning (LeCun et al., 2015), particularly with the development of convolutional neural network (CNN)-based detection architectures. Prominent examples include Faster R-CNN (Ren et al., 2016), YOLO (Redmon, 2016), and SSD (Liu et al., 2016), which have demonstrated remarkable performance on general-purpose object detection datasets. These algorithms are capable of automatically extracting target features and achieving outstanding localization and classification through an end-to-end learning approach. However, they usually encounter the some critical issues when directly applied to small target detection in UAV remote sensing images: the insufficient process of feature extraction and the presence of extreme small targets within image results in a reduced number of pixels, thereby limiting the amount of feature information used for the subsequent deep neural networks.

It is evident that conventional neural networks typically employ downsampling operations to expand the sensory field. However, this process concomitantly results in the loss of small target presentation (Chen et al., 2025b). Additionally, the capacity to detect targets across multiple scales is inadequate. The scale of objects in UAV remote sensing images varies significantly and existing algorithms are challenging to utilize for the accurate detection of both large and small targets. Albeit the substantial advancements in deep learning methodologies for small object detection, several critical challenges remain unresolved. Foremost among these is the development of a network architecture that enables effective multi-scale perception, which is essential to accommodate the wide range of sizes exhibited by small targets. Secondly, the intricate nature of image backgrounds, the high degree of similarity between targets and their surroundings and the frequent occlusion of targets collectively exacerbate the complexity of the detection task. Furthermore, the lack of effective feature fusion is a key factor that degrades small objects detection performance. Due to their limited pixel coverage, small targets have sparse and fragile feature representations that are easily overshadowed by background noise. This makes it difficult for models to accurately extract features and distinguish small targets from the background, thereby reducing detection accuracy (Yue et al., 2024).

To tackle the significant challenges in UAV small object detection and boost the detection performance in complex scenarios, we have developed an innovative detection network called KSCNet. This network leverages the hybrid collaboration of KAN and SSM to achieve remarkable detection performance especially for small objects. Notably, KAN doesn't hold fixed activation functions but with learnable univariate functions, allowing it to excel in data fitting and complex feature learning tasks. SSM exhibits greate potential for managing long-range dependencies with linear computational complexity, and also enables model to filter out irrelevant information while retaining necessary and relevant feature indefinitely through effective selection mechanism. Besides, we also design a new FPN variants named Semantic Aggregation Networks (SAN), which consists aggregation and expansion step to fully explore and utilize multiscale features. Finally, a depthwise channel attention (DCA) mechanism was introduced to mitigate the aliasing effect for fused features. In summary, we conclude our contribution of this work as follows,

- We propose a innovative high performance UAV small object detection architecture named KSCNet, which collaborates prevalent methodologies KAN and SSM, enabling effective modeling of the intricate relationship between small targets and their background in detection tasks, while simultaneously facilitating flexible feature extraction and largely improve the UAV small object detection performance.
- Several strategies were adopted to build KSCNet, including KAN-based backbone for efficient feature extraction, Semantic Aggregation Network (SAN) for processing multiscale features, SSM acts as core function within SAN as its superior cost-effective ability for long range dependancy relation, a channel pattern Depthwise Channel Attention (DCA) to reduce the aliasing effect for fused features.
- Extensive experiments conducted on the public UAV datasets SIMD and VisDrone demonstrate that KSCNet achieves superior detection performance in comparison with mainstream object detectors, exhibiting a consistent improvement in accuracy across various object detection metrics, particularly for small targets.

## 2. Related works

### 2.1. Deep learning object detectors

In the field of deep learning, detection algorithms can be classified into two categories: two-stage and one stage types based on the workflow and complexity of the involved processing tasks. The typical representative algorithms of the two stage methods are RCNN (Girshick et al., 2014), Faster RCNN (Ren et al., 2016), FPN (Lin et al., 2017a), Mask RCNN (He et al., 2017) and so on. Meanwhile, for the one stage counterpart like SSD (Liu et al., 2016), RetinaNet (Lin et al., 2017b), CenterNet (Duan et al., 2019), EfficientDet (Tan & Le, 2019) and YOLO series algorithms. Particularly, the YOLO family techniques have been developed to the YOLOv12 (Tian et al., 2025) version after continuous updating and iteration since its introduction. The network structure has been extensively optimized to significantly enhance the performance and efficiency of object detection, garnering considerable attention from numerous researchers and scholars. DETR (Zhao et al., 2024) is notable for introducing the transformer encoder-decoder architecture into the detection task for the first time, and replacing the traditional anchor and NMS with the global attention mechanism and end-to-end design. These represents a significant innovation in the detection paradigm. Subsequent models, including Deformable DETR (Zhu et al., 2020), DINO (Zhang et al., 2022) and others, leading to substantial improvements in training efficiency and detection accuracy. These models have been shown to combine both global context modeling capability and end-to-end simplicity, with outstanding performance in dense occlusion and small target scenes. Conventional object detection methods often fail to deliver satisfactory performance when applied to UAV small object detection due to the unique challenges of this domain. We propose KSCNet, a detection framework specifically designed for UAV scenarios, achieves superior detection performance by enhancing feature extraction and localization accuracy for small objects in aerial imagery.

### 2.2. Optimized strategies for UAV detectors

Researchers have proposed numerous enhanced approaches to address the critical challenges for small objects and boost the detection performance in UAV imagery, primarily encompassing the following domains: The enhancement of feature extraction backbone, the design of deeper network structures, the introduction of attention mechanisms and other optimization techniques. To address extreme variation of small target scale faced in the UAV aerial images, Chen (Chen et al., 2025a) *et.al* proposed semantic information guided fusion module through high-level semantic information to guide and align texture patterns for enhancing the representation of small targets, significantly improve the detection performance. Xiao (Xiao et al., 2024) *et.al* introduced lightweight fusion strategy by rethinking interlayer feature correlation within FPN architecture and proposed grouped feature focus unit and multi-level feature reconstruct module to improve small detection performance in the complex backgrounds and densely populated areas. Jing (Jing et al., 2024) *et.al* proposed Feature Aggregation Network to fully explore different scale features by introducing top-down pathway and feature-aware modules, which contribute to narrow semantic information gap within architecture and boost the small object detection. Xu (Xue et al., 2024) *et.al* introduced EL-YOLO aimed for low-altitude aerial small object detectors, by developing sparsely connected asymptotic FPN and cross stage multi-head self attention mechanism, EL-YOLO realized excellent performance on NVIDIA Jetson harware platform with lightweight model pattern. Fan (Fan et al., 2025) *et.al* introduced LUD-YOLO to improve the unmanned aerial vehicle detection by designing new feature fusion mode and dynamic sparse attention into C2f to achieve flexible computation location and content awareness of features with excellent detection accuracy. Nevertheless, the existing methods still exhibit certain limitations in feature fusion, leading to features of small objects to be submerged in the complex background, thereby diminishing detection accuracy. We propose the SAN feature fusion network. The SAN network employs the SSM unit as its core component to model the features of small objects over long distances with linear modol complexity, achieving substantial improvements in UAV small detection performance.

### 2.3. Feature enhancement of small object detection

In the field of small object detection, the inherent challenges posed by diminutive target sizes and limited pixel representation often hinder conventional detection approaches from effectively capturing discriminative features. To address these limitations, contemporary methodologies employ sophisticated multi-scale feature fusion architectures that integrate high-resolution spatial details from early network layers with semantically rich information from deeper levels, thereby constructing more robust feature representations of multi-scale features (Lin et al., 2017a)(Ghiasi et al., 2019)(Tan et al., 2020)(Hu et al., 2021).

The incorporation of attention mechanisms has proven particularly valuable, as these modules automatically identify and emphasize the most salient channels and spatial regions while suppressing irrelevant background interference (Hassanin et al., 2024). Furthermore, context augmentation techniques (Liu et al., 2018a), including expanded receptive fields through dilated convolutions or global relationship modeling via Transformer architectures, establish crucial connections between small targets and their surrounding environment to compensate for their inherently weak semantic signatures (Wu et al., 2022). To counteract the inevitable information degradation caused by progressive downsampling, modern approaches implement various high-resolution preservation strategies (Noh et al., 2019). These include maintaining shallow feature maps throughout the network hierarchy and integrating super-resolution reconstruction modules to enhance feature clarity. During model optimization, carefully designed data augmentation protocols work in concert with specialized loss functions to address class imbalance issues and refine localization precision.

## 3. The methodology of KSCNet

Our proposed framework follows the overall architecture of efficient typical YOLO methods, shown in Fig. 1, including backbone, neck and detection head. For an input image, the KSCNet architecture employs specifically designed CKN blocks as its primary computational components. This strategic integration significantly enhances the feature discernment capabilities of the conventional C3k2 module that is utilized in YOLO11. During the neck stage, the SSM act as core along with channel attention DCA for driving the basic features into high advanced semantic information, producing feature maps at three distinct scales: $80\times80$, $40\times40$ and $20\times20$. These hierarchically structured semantic features are then employed for the final object detection task.

### 3.1. KAN-based feature extractor

Recent advancements in Kolmogorov-Arnold Networks (KAN) have catalyzed a paradigm shift in machine learning, offering novel solutions for integrating domain-specific knowledge into deep learning models. KAN (Kolmogorov, 1961) offer a compelling alternative to Multi-Layer Perceptrons (MLP) (Hornik et al., 1989) by leveraging the Kolmogorov-Arnold theorem. These theorem proves that any continuous multivariate function can be decomposed into a finite composition of single-variable functions. This principle underpins KANs, which replace traditional neural networks' fixed linear weights with learnable univariate activation functions. With this paradigm, KANs gain superior flexibility in modeling complex patterns and enhanced interpretability through their mathematically structured architecture.

Unlike conventional MLPs that employ fixed linear transformations, KANs implement parametric spline functions as their fundamental computational units, shown in Fig. 2. This architectural innovation achieves dual advantages: significant parameter efficiency through optimized function representation and enhanced network generalization by encoding smoothness priors in the spline formulations. The Kolmogorov-Arnold representation theorem considers that a continuous multivariate function $f(\mathbf{x}_1,\dots,\mathbf{x}_n)$ can be presented as,

$$f\left(\mathbf{x}_1,\dots,\mathbf{x}_n\right)=\sum_{q=1}^{2n+1}\Phi_q\left(\sum_{p=1}^{n}\phi_{q,p}\left(\mathbf{x}_p\right)\right) \tag{1}$$

Here, the $\Phi_q$ and $\phi_{q,p}$ are a set of continues univariate functions.

Each layer in the KANs can be regarded as a matrix of these learnable 1D functions:

$$\Phi=\phi_{q,p},\ p=1,2,\dots,n_{in},\ q=1,2,\dots,n_{out} \tag{2}$$

The $\phi_{q,p}$ can be defined as a B-spline, which is a type of function defined by a linear combination of basis splines. $n_{in}$ denotes the number of input features to a particular layer, while $n_{out}$ is the number of output features generated by that layer. The activation functions $\phi_{l,j,i}$ in this metric are such learnable spline functions,

$$\text{spline}(x)=\sum_i c_i B_i(\mathbf{x}) \tag{3}$$

$c_i$ are trainable coefficients.

With stacking of the complex functional mapping, we can get the overall structural of KAN,

$$KAN(\mathbf{x})=(\Phi_{L-1}\circ\Phi_{L-2}\circ\dots\circ\Phi_0)(\mathbf{x}) \tag{4}$$

where $\Phi_l$ acts on the input $x_l$ to produce the next layer's input $x_{l+1}$, presented as,

$$x_{l+1}=\Phi_l\left(x_l\right)=\begin{pmatrix}\phi_{l,1,1}(\cdot) & \cdots & \phi_{l,1,n_l}(\cdot)\\ \vdots & \ddots & \vdots\\ \phi_{l,n_{l+1},1}(\cdot) & \cdots & \phi_{l,n_{l+1},n_l}(\cdot)\end{pmatrix}x_l \tag{5}$$

In this work, we employ KAN convolution (Bodner et al., 2024) into C3K2 (Khanam & Hussain, 2024) architecture shown in Fig. 3(b) to build CKN block as the fundamental backbone in Fig. 1. The KAN convolution

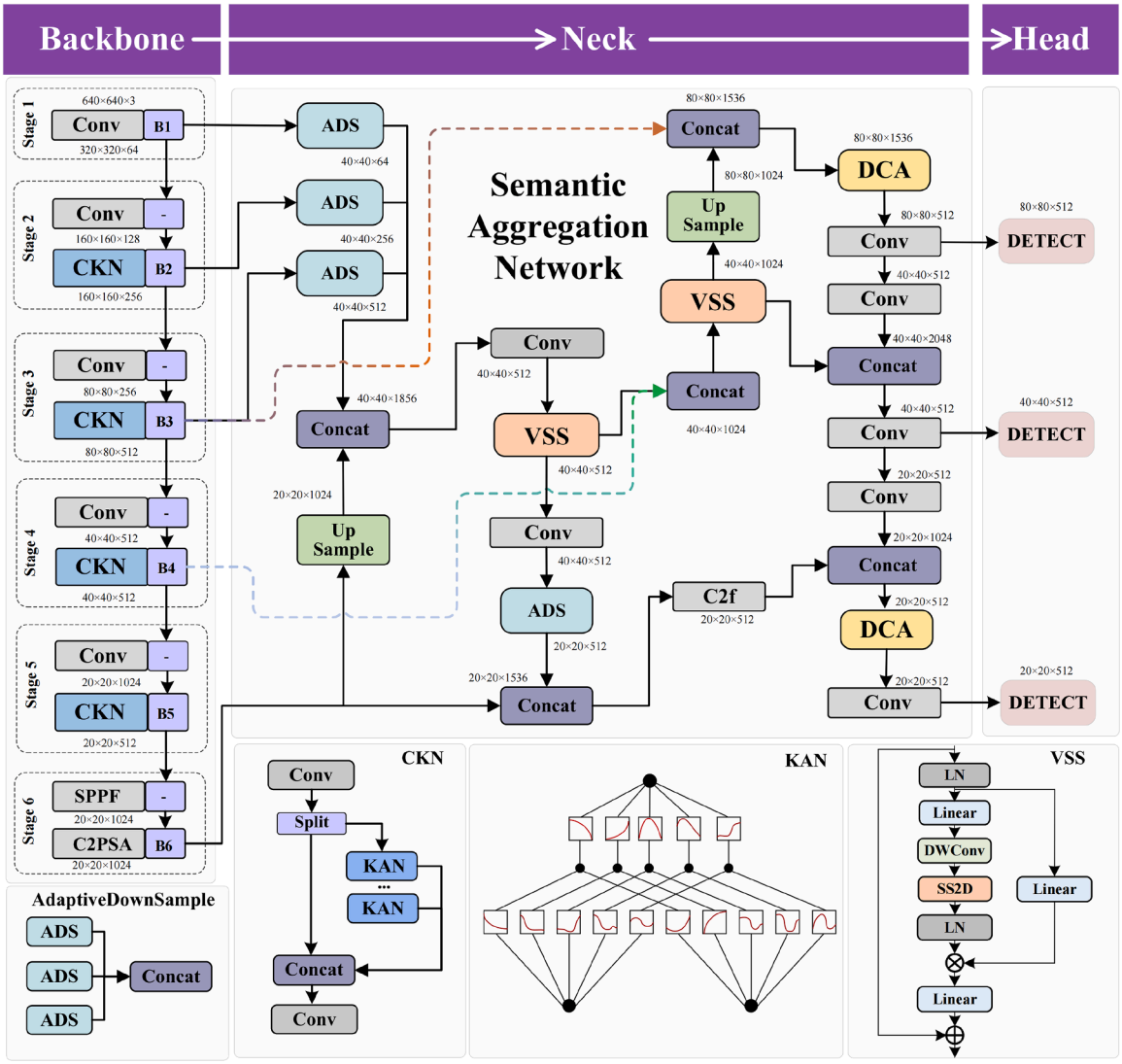


**Fig. 1.** Main architecture of proposed KSCNet for UAV small object detection. KAN-style integrated convolution backbone for elementary feature extraction. Semantic Aggregation Network with core VSS within neck for efficient multiscale feature fusion. A channel attention DCA for reducing aliasing effect for fused features. Finally, 3 scales detection head $80 \times 80$, $40 \times 40$, $20 \times 20$ to execute final classification and localization.

integrates a basis functions $b(x)$ and the output $O(x)$ can been concluded the sum of $b(x)$ and spline function $spline(x)$, defined as,

$$\begin{aligned} O(\mathbf{x}) &= w(b(\mathbf{x}) + spline(\mathbf{x})) \\ b(\mathbf{x}) &= SiLU(\mathbf{x}) \\ spline(\mathbf{x}) &= \sum_i c_i B_i(\mathbf{x}) \end{aligned} \tag{6}$$

where $\mathbf{w}$ is the training weight of the network, $c_i$ is the coefficient to optimize the training loss function, $B_i(\mathbf{x})$ is B-spline function.

CKN apply kernels composed of learnable non-linear functions, allowing each kernel element to adapt dynamically during training, enabling greater flexibility and expressiveness. Furthermore, CKN share fundamental architectural similarities with conventional CNNs, while they incorporate two key modifications to the standard CNN framework: (1) The replacement of traditional convolutional layers with parametric KAN-based convolutional layers.(2) The option to employ either a KAN layer or a standard MLP following the flattening operation.

The principal advantage of CKN lies in their superior parameter efficiency compared to conventional architectures. This efficiency stems from the network's unique construction, particularly its utilization of B-spline basis functions to model activation patterns. Unlike fixed activation functions such as ReLU that employ piecewise linear approximations, B-splines enable the smooth representation of complex, non-linear activation functions through learnable parameters. This adaptive approach allows for more efficient function approximation while maintaining strong representational capacity, enabling the convolution kernel to learn optimized, smooth functions that are highly effective at encoding the complex textures and structures found in UAV imagery.

### 3.2. Semantic aggregation network

The Feature Pyramid Networks (FPN) (Lin et al., 2017a) has been identified as a seminal innovation in the domain of object detection, playing a pivotal role in enhancing the robustness of the model to scaling variations via the multi-scale feature fusion mechanism, particularly in the case of small targets. Basic FPN architecture, shown in Fig. 4(a), constructs a feature pyramid with both high-resolution details and strong semantic expression through top-down semantic propagation and cross-layer fusion with lateral connection. PAFPN (Liu et al., 2018b) notably improved information flow while keeping the framework simple and generalizable, shown in Fig. 4(b), the core innovation lies in its double bidirectional (top-down + bottom-up) multi-scale feature fusion

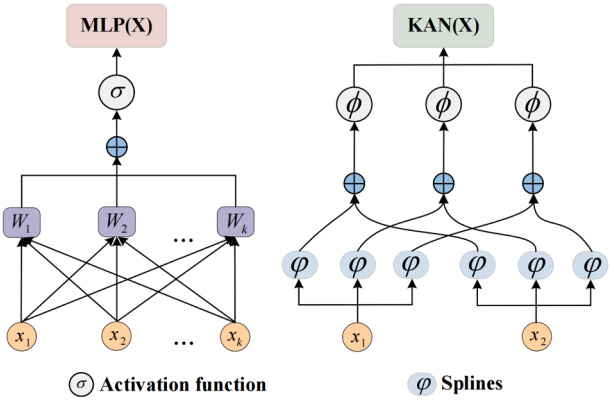


**Fig. 2.** Comparison of multilayer perception (MLP) and Kolmogorov-Arnold Network (KAN).

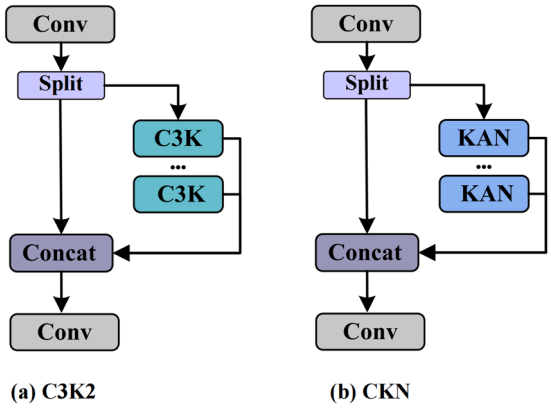


**Fig. 3.** Architecture of C3K2 and CKN.

approach, which became influential in subsequent detection architectures. While FPN and PAFPN effectively propagate semantic information, their fixed fusion pathways (e.g., simple addition or complex concatenation) struggle to adaptively model the non-linear relationships and semantic gaps between features from vastly different scale.

We introduces a redesigned FPN architecture, the Semantic Aggregation Network, which implements an enhanced feature fusion mechanism for combining multi-scale representations from various network stages, shown in Fig. 4(c). Through hierarchical feature aggregation and semantically guided feature expansion, the SAN network significantly improves the characterization of the feature pyramid while maintaining computational efficiency. SAN consists of two steps: **Aggregation** and **Expansion**. The Aggregation step is designed to construct a centralized, multi-scale feature context by hierarchically integrating information from different backbone levels. This is achieved through a set of parallel adaptive average pooling operations, which systematically gather and unify feature representations across varying spatial resolution. The Expansion step aims to semantically and structurally enrich the aggregated multi-scale features through a guided upsampling and refinement process. This step utilizes the globally-aware context obtained from the Aggregation phase to intelligently guide the feature reconstruction and enhancement across resolutions. By integrating holistic semantic guidance with local feature refinement, the Expansion step effectively generates a new pyramid of enhanced features that possess both high spatial fidelity and strong semantic consistency. The resulting feature context serves as a rich, globally aware foundation for subsequent processing, effectively mitigating the semantic fragmentation and scale misalignment issues commonly encountered in complex scenes such as UAV imagery. The whole process is easy to follow and can be implemented with 2 steps,

*step1:*

$$\alpha = AVG(\mathbf{C}_i), i \in 2,3,4$$
$$\mathbf{X}_{agg} = Concat[\alpha, \mathbf{C}_i]$$

*step2:*

$$\mathbf{X}_{exp} = Expand(\mathbf{X}_{agg})$$
$$\mathbf{M}_i = \mathbf{P}_i \times \mathbf{X}_{exp}, i \in 1,2,3,4 \quad (7)$$

where $AVG$ denotes the 2-D averaging pooling, $P_i$ mean the features before fusion expansion, $M_i$ is the final detection head.

The hierarchical aggregation is explicitly linked to creating a more coherent multi-scale feature representation, reducing the risk of semantic ambiguity for small objects. The expansion process is analyzed not just as an upsampling operation, but as a feature refinement step that uses the aggregated semantic context to guide the reconstruction of high-resolution features, thereby preserving crucial spatial details for localization.

### 3.3. Vision state space model for handling multiscale features

Vision Transformers (ViTs) Vaswani et al. (2017), Liu et al. (2021), Yu et al. (2022) have revolutionized visual representation learning, demonstrating the critical role of large-scale pre-training in advancing image classification performance. However, their practical deployment faces a fundamental challenge: the quadratic computational complexity of self-attention mechanisms, which becomes prohibitive for long-sequence inputs. To address this limitation, Mamba (Gu & Dao, 2023) emerges as an innovative evolution of State Space Models (Gu et al., 2020) (Gu et al., 2021), introducing Selective Structured State Space Models that achieve two key breakthroughs: (1) linear computational complexity scaling, (2) enhanced capacity for modeling long-range dependencies. This paradigm shift enables efficient processing of high-resolution visual data while maintaining the representational power of traditional ViTs.

SSMs has sparkled renewed enthusiasm with remarkable progress from both academic and industrial communities in the recent years (Wang et al., 2024c). Evolving from their classical predecessors like the Kalman filter, contemporary SSM have demonstrated exceptional capabilities in modeling long-range dependencies while maintaining efficient parallelizability during training. These advantages have positioned modern SSMs as a compelling alternative to traditional sequence modeling approaches, particularly in handling large-scale sequential data.

**Preliminaries**. The SSM-based models, can be considered as linear time-invariant system that maps the 1-D function or sequence $x(t) \in \mathbb{R} \rightarrow y(t) \in \mathbb{R}$ through a hidden state $h(t) \in \mathbb{R}^N$. The process uses $\mathbf{A} \in \mathbb{R}^{N\times N}$ as the evolution parameter and $\mathbf{B} \in \mathbb{R}^{N\times 1}$, $\mathbf{C} \in \mathbb{1} \times \mathbb{N}$ as the projection parameters.

$$h'(t) = \mathbf{A}h(y) + \mathbf{B}x(t)$$
$$y(t) = \mathbf{C}h(t) \quad (8)$$

To adapt the model into deep networks, continuous-time SSM need to undergo discretization in advance, including a times scale parameter $\Delta$ to transform $\mathbf{A}$ and $\mathbf{B}$ into discrete pattern $\overline{\mathbf{A}}$ and $\overline{\mathbf{B}}$, among this process the common employed technique is zero-order hold (ZOH),

$$\overline{\mathbf{A}} = exp(\Delta)$$
$$\overline{\mathbf{B}} = (\Delta\mathbf{A})^{-1}(exp(\Delta\mathbf{A} - \mathbf{I})) \cdot \Delta\mathbf{B} \quad (9)$$
$$h'(t) = \overline{\mathbf{A}}h(y) + \overline{\mathbf{B}}x(t)$$
$$y(t) = \mathbf{C}h(t) \quad (10)$$

Finally, the output can be reached by global convolution,

$$\overline{\mathbf{K}} = (\mathbf{C}\overline{\mathbf{B}}, \mathbf{C}\overline{\mathbf{A}\mathbf{B}}, \ldots, \mathbf{C}\overline{\mathbf{A}}^{L-1}\overline{\mathbf{B}})$$
$$y = x * \overline{\mathbf{K}} \quad (11)$$

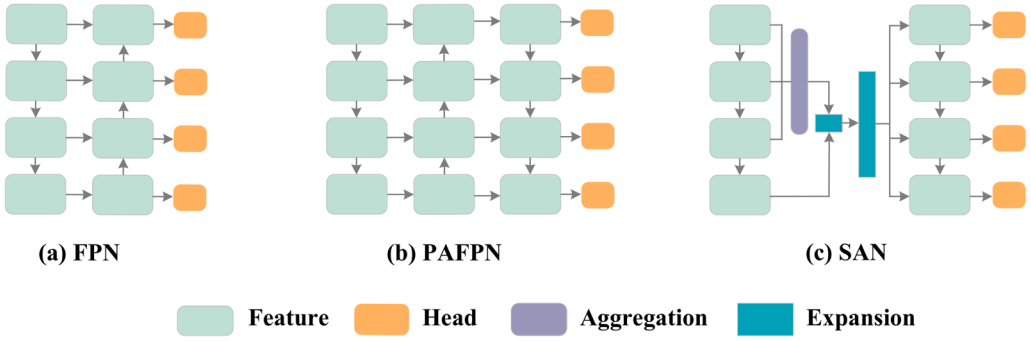


**Fig. 4.** Different Feature Pyramid Networks. (a) the vanilla FPN. (b) PANet with extra top-down information flows. (c) Our proposed Semantic Aggregation Networks with feature aggregation and expansion steps for enriched knowledge from a wide range scale features.

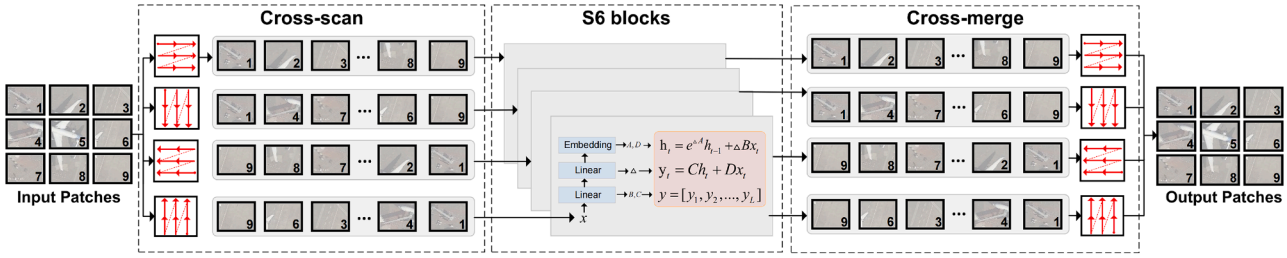


**Fig. 5.** Architecture of 2D-Selective-Scan (SS2D). The input patches are transmitted from four different scanning paths and then processed by distinct S6 blocks, the results are subsequently merged together to build 2D feature map for eventual output.

where $L$ denotes the length of the input sequence $x$ and $\overline{\mathbf{K}} \in \mathbb{R}^L$.

We employ cost-effective selective mechanism named Vision State Space (VSS) block from Vmamba (Liu et al., 2024) as the core expansion method in the SAN architecture, shown in Fig. 6, VSS serves as the fundamental exchange center for processing the multi-scale features to construct hierarchical vision representation with linear computation and complexity. As shown in Fig. 5, the workflow of SS2D contains four steps: Split, Cross-scan, S6 blocks (Gu & Dao, 2023), Cross-merge. The input features are first split into several patches, these data are generated into sequences pattern from four distinct traversal paths, and then further process by consecutive parallel S6 blocks for managing long-range dependency information, the treated features are then sent into Cross-merge produced to output.

SS2D employs a sophisticated cross-scanning strategy that processes the 2D image through complementary 1D traversal paths. This innovative approach enables each spatial location to dynamically aggregate and integrate features from all regions of the input image across multiple orientations. By systematically combining information from these diverse scanning directions, the model effectively constructs comprehensive global receptive fields while maintaining computational efficiency. The omni-directional information flow along different axes ensures robust feature representation that captures both local details and long-range dependencies in the visual space.

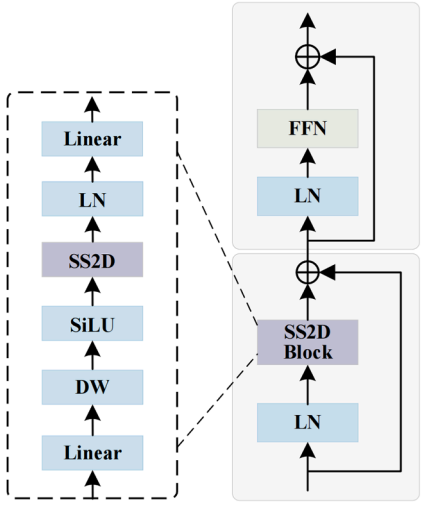


**Fig. 6.** Visual state space block.

### 3.4. Depthwise channel attention

A channel attention mechanism named Depthwise Channel Attention (DCA) is designed to eliminate the aliasing effect, which denotes a distortion phenomenon where erroneous and misleading information is introduced into feature maps due to continual signal sampling and directly combined multi-scale feature fusion process (Li et al., 2022b). The architecture of DCA, shown in Fig. 7, consists two main branches to recalibrate input features for more discriminative information objects. We employed depthwise convolution (Chollet, 2017) along with relatively shorted information flow path, mitigating channel information loss due to the dimension reduction compared with SE (Hu et al., 2018). Besides, Harsh-Sigmoid activation function was also adopted to improve the feature non-linear representation, which is a variant of the sigmoid function and due to the strong gradient around its saturation interval (i.e. the part of the input away from 0), the optimization process can be faster in terms of more gradient new, which can improve model training efficiency, especially in deep neural networks. The calculation of DCA can summarized as follows,

$$\mu, \nu = Split(\mathbf{X}) \tag{12}$$

$$\mathbf{O1} = \sigma(Linear(AVGPool(\nu))) \tag{13}$$

$$\tau 1, \tau 2, \tau 3 = DW(split(\mu)) \tag{14}$$

$$\mathbf{O2} = Concat[\tau 1; \tau 2; \tau 2] \tag{15}$$

$$\mathbf{X}_{refine} = O1 \times O2 \times \mathbf{X} \tag{16}$$

For input $\mathbf{X} \in \mathbb{R}^{C \times H \times W}$, where $C$,$H$ and $W$ denote the channel, height and width respectively. $\sigma$ means the Harsh-Sigmoid activation function.

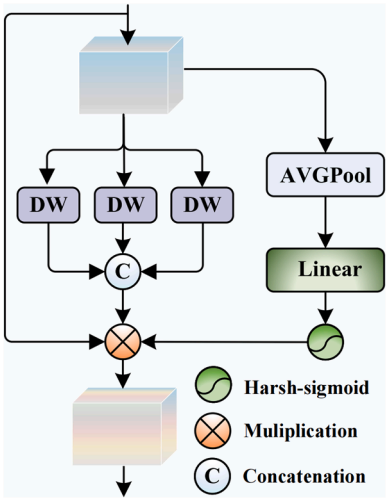


**Fig. 7.** Depthwise Channel Attention (DCA).

DW is depthwise convolution with kernel size $3 \times 3$, AVGPool is global average pooling. *Concat* represents that all the outputs are stacked along the channel dimension. The DCA module addresses aliasing not by directly filtering spatial frequencies, but by dynamically recalibrating channel-wise feature responses to suppress channels that are most corrupted by these spatial dimension. It uses DWConv to sense spatial inconsistencies and then employs channel recalibration to mute the channels that are most polluted by these effects, making it more robust for accurate localization and classification.

## 4. Experiments

### 4.1. Dataset and settings

In this work, we adopt two public UAV dataset to testify the effectiveness for KSCNet on the UAV small detection:

**SIMD**: This dataset is an open source dataset for small object detection in remote sensing imagery, released by a team of researchers at the National University of Science and Technology (NUST) (Haroon et al., 2020). The images in SIMD are mainly acquired from multiple locations in the EU and the US from the public Google Earth satellite imagery, which is mainly used for multi-size and multi-category vehicle detection tasks in high-resolution remote sensing imagery. It comprises 5000 images of resolution $1024 \times 768$ and collectively contains 45 096 objects in 15 different classes of vehicles including cars, trucks, buses, long vehicles, various types of aircrafts and boats.

**VisDrone**: The dataset is a large benchmark dataset created by the AISKYEYE team at the Machine Learning and Data Mining Laboratory of Tianjin University, China, designed for the analysis of images and videos captured by drones (Du et al., 2019). This dataset contains 288 video clips, 261,908 frames and 10,209 still images captured by various drone cameras. For the task for UAV detection, there are 6471 images for training, 548 for validation and 3190 for testing. The dataset covers a variety of aspects including different locations, environments, objects and densities and also are under different weather and lighting conditions.

For more detail, all experiments were conducted under CPU Intel Core i7-13700KF and one single GPU NVIDIA RTX 4090, with deep learning framework pytorch version 2.2 and cuda toolkit 11.8. We set the total 300 training epochs with training input image size $640 \times 640$, batch size 16, learning rate 0.01, momentum 0.937 and weight decay 5e-5 during training time, the batch size was set to 1 for fast processing during inference stage.

### 4.2. Evaluation metrics

In order to thoroughly validate the performance of the model, we follow the standard evaluation metrics commonly used in object detection. These metrics assess the model's performance in various aspects, including Precision (P), Recall (R), and the balance between them, such as Mean Average Precision (mAP). By analyzing these comprehensive performance metrics, we gain a deeper understanding of the model's detection capabilities under different scenarios and conditions, allowing for a more accurate evaluation.

$$P = \frac{TP}{TP + FP} \tag{17}$$

$$R = \frac{TP}{TP + FN} \tag{18}$$

$$AP_i = \int_0^1 P_i(R_i) dR_i \tag{19}$$

$$mAP = \frac{1}{n} \sum_{i=1}^{n} AP_i \tag{20}$$

Ture Positive (TP), False Negative (FN) and False Positive (FP) were used to measure the accuracy and effectiveness of the model in detecting the target, which represented successful detection for real targets, failed for targets and misrecognition for true targets respectively.

### 4.3. Experimental analysis

We report the detection results of KSCNet with the real-time YOLO series methods on the SIMD dataset in Table 1. It's notable that KSCNet achieve the best detection performance across $Precision$, $mAP@50$ evaluation metric, realizing 2.68% and 1.69% increase respectively compared with baseline YOLO11s (Khanam & Hussain, 2024). For the strict and comprehensive metric $mAP@50:95$, KSCNet also realizes 2.22% detection improvements and also exhibits the top grades among all other algorithms, indicating that KSCNet could fully utilize the advantage of the model design and address the small detection issues. For the latest released YOLO variant YOLOv12 (Tian et al., 2025), KSCNet also provides the noticeable leading detection results at most of evaluation metrics. Obviously, model complexity matters overall detection results to some extent, this can be validated from YOLO-$s$ model, such as YOLOv5s (Jocher, 2020), YOLOv6s (Li et al., 2022a), YOLOv8s (Jocher et al., 2023) and YOLOv10s (Wang et al., 2024a), they generally produce higher detection results compared with their $n$ versions counterparts, which highlights more complex models have greater feature extraction capabilities, capturing subtle patterns and high-level features in the data to improve recognition of complex targets. In conclusion, KSCNet has been demonstrated to achieve optimal object detection performance with minimal increase in model complexity and parameter. The network performs particularly well on the comprehensive evaluation metric $mAP@50:95$, fully demonstrating its unique advantage in balancing model lightness and detection accuracy. This efficient performance enhancement not only underscores the innovation of its architectural design, additionally offers an optimal solution for object detection applications in contexts where resources are constrained.

To fully understand the detection performance of KSCNet, Table 2 presents the results of KSCNet with mainstream detectors including two-stage, one-stage and transformer based methods. Faster RCNN, as the classic milestone two-stage detectors, realizes 0.868 detection accuracy at $mAP@50$, a slightly higher than that of KSCNet. As for the evaluation metric $mAP@50:95$, KSCNet excels by 1.17% of 0.691 with less parameters and model complexity, explaining that heavy parameter model is not suitable for realtime device applications and will become computation burden for fast inference situation, which is also adaptable for RT-DETR detectors holds inferior detection results on $mAP@50$ and $mAP@50:95$ compared with KSCNet. Recent HyperYOLO exhibits outstanding detection performance on the COCO dataset, whereas HyperYOLOs produces relatively

**Table 1**
Overall comparison with the YOLO algorithms on the SIMD dataset. * means the baseline method.

| Methods | *Precision* | *Recall* | *mAP*@50 | *mAP*@50: 95 | *Params.*(*M*) | *FLOPs*(*G*) |
|---|---|---|---|---|---|---|
| YOLOv5n (Jocher, 2020) | 0.774 | 0.777 | 0.801 | 0.616 | 1.77 | 4.2 |
| YOLOv5s (Jocher, 2020) | 0.803 | 0.777 | 0.811 | 0.639 | 7 | 15.9 |
| YOLOv6n (Li et al., 2022a) | 0.807 | 0.759 | 0.804 | 0.637 | 4.63 | 11.35 |
| YOLOv6s (Li et al., 2022a) | 0.771 | 0.773 | 0.792 | 0.633 | 18.51 | 45.19 |
| YOLOv7t (Wang et al., 2023) | 0.768 | 0.784 | 0.816 | 0.643 | 6.04 | 13.1 |
| YOLOv8n (Jocher et al., 2023) | 0.814 | 0.764 | 0.813 | 0.651 | 3 | 8.1 |
| YOLOv8s (Jocher et al., 2023) | 0.804 | 0.803 | 0.836 | 0.677 | 11.1 | 28.5 |
| YOLOv9c (Wang et al., 2024b) | 0.834 | 0.811 | 0.855 | 0.702 | 50.73 | 236.8 |
| YOLOv10n (Wang et al., 2024a) | 0.764 | 0.757 | 0.796 | 0.633 | 2.7 | 8.3 |
| YOLOv10s (Wang et al., 2024a) | 0.814 | 0.764 | 0.822 | 0.664 | 8.04 | 24.5 |
| YOLO11n (Khanam & Hussain, 2024) | 0.738 | 0.798 | 0.813 | 0.65 | 2.58 | 6.3 |
| YOLO11s* (Khanam & Hussain, 2024) | 0.783 | 0.812 | 0.83 | 0.676 | 9.41 | 21.3 |
| YOLOv12n (Tian et al., 2025) | 0.74 | 0.792 | 0.812 | 0.66 | 2.55 | 6.3 |
| YOLOv12s (Tian et al., 2025) | 0.773 | 0.823 | 0.828 | 0.672 | 9.23 | 21.2 |
| **KSCNet** | 0.804 | 0.809 | **0.844** | **0.691** | 11.71 | 32.8 |

**Table 2**
Experimental results with the SOTA methods on the SIMD dataset.

| Methods | *mAP*@50 | *mAP*@50: 95 | *Params.*(*M*) | *FLOPs*(*G*) |
|---|---|---|---|---|
| Faster RCNN (Ren et al., 2016) | 0.868 | 0.683 | 41.42 | 178 |
| RT-DETR-R50 (Zhao et al., 2024) | 0.75 | 0.609 | 411.96 | 125.7 |
| RetinaNet (Lin et al., 2017b) | 0.603 | 0.439 | 36.62 | 179 |
| SSD300 (Liu et al., 2016) | 0.714 | 0.448 | 25.61 | 32.12 |
| SSD512 (Liu et al., 2016) | 0.752 | 0.496 | 26.45 | 89.5 |
| EfficientNet (Tan & Le, 2019) | 0.874 | 0.680 | 18.62 | 83.42 |
| HyperYOLOs (Feng et al., 2024) | 0.833 | 0.669 | 14.82 | 38.9 |
| DINO (Zhang et al., 2022) | 0.818 | 0.382 | 47.56 | 235 |
| **KSCNet** | **0.844** | **0.691** | 11.71 | 32.8 |

**Table 3**
Different methods for building effective feature extraction block with vanilla C3K2.

| Methods | *Precision* | *Recall* | *mAP*@50 | *mAP*@50: 95 | *Params.*(*M*) | *FLOPs*(*G*) |
|---|---|---|---|---|---|---|
| C2fCIB (Wang et al., 2024a) | 0.767 | 0.808 | 0.822 | 0.657 | 9.15 | 25.4 |
| FMB (Zheng et al., 2024) | 0.772 | 0.806 | 0.729 | 0.665 | 9.36 | 21.2 |
| DeConv (Chen et al., 2024b) | 0.818 | 0.778 | 0.824 | 0.668 | 9.41 | 21.3 |
| FasterConv (Chen et al., 2023) | 0.785 | 0.812 | 0.828 | 0.677 | 9 | 23.3 |
| **CKN** | 0.815 | **0.825** | **0.836** | **0.681** | 11.44 | 26.7 |

**Table 4**
Experimental results of CKN employment in different position within backbone.

| Baseline | Stage2 | Stage3 | Stage4 | Stage5 | *mAP*@50 | *mAP*@50: 95 | *Params.* | *FLOPs*(*G*) |
|---|---|---|---|---|---|---|---|---|
| ✓ | | | | | 0.834 | 0.672 | 9.41 | 21.3 |
| ✓ | ✓ | ✓ | | | 0.833 | 0.679 | 10.4 | 23.7 |
| ✓ | | | ✓ | ✓ | 0.831 | 0.678 | 10.1 | 23.3 |
| ✓ | ✓ | ✓ | ✓ | ✓ | 0.836 | 0.681 | 11.44 | 26.7 |

lower detection accuracy compared with KSCNet. This experiments validate the effectiveness KSCNet on the UAV small object detection with leading detection performance among the current detectors.

To validate the effectiveness of CKN, we compare it with other similar efficient model blocks in Table 3. The results shows the CKN provides the excellent detection accuracy across the four main evaluation metrics *Precision Recall*, *mAP*@50, *mAP*@50: 95 with 0.815, 0.825, 0.836 and 0.681. Compared with the lightweight method FasterConv and DeConv, CKN realizes higher detection accuracy results at *mAP*@50: 95 while with slight model parameter and complexity surge. CKN exhibits a strong performance boost compared to C2fCIB by 1.7% and 3.65% increase at *mAP*@50 and *mAP*@50: 95. CKN demonstrates superior efficacy in elucidating the complex relational structures between nodes. By effectively learning the intricate non-linear patterns embedded within the data, it achieves enhanced predictive accuracy. Notably, in the context of high-dimensional data, KAN convolution exhibits pronounced advantages in terms of flexibility, parameter and optimization efficacy. These attributes collectively underscore its potential as a powerful tool for advanced data analysis and modeling.

We perform an ablation study to examine the impact of CKN on model performance at different stages within the backbone, as presented in Table 4. As shown in the Fig. 1, the core components of backbone can be separated into 6 stages. The initial and final stages constitute the basic module responsible for downsampling input images and collecting the comprehensive semantic information. CKN blocks deployed range 2 to 5 stages offers the best detection accuracy with 0.836 and 0.681 at *mAP*@50 and *mAP*@50: 95 respectively, confirming the need of consecutive of KAN-base approaches to capture the complex patterns in the data. Employing CKN only in the first two stages or in the last two stages achieves insufficient performance, illustrating that although KAN convolution has a powerful representation and it requires a more complex structure to realize its theoretical potential in the face of scarce feature representations when dealing with weak targets in UAVs.

**Table 5**
Results of current prevailing feature pyramid networks with our proposed SAN on the SIMD dataset.

| Methods | *Precision* | *Recall* | *mAP*@50 | *mAP*@50: 95 | *Params.*(*M*) | *FLOPs*(*G*) |
|---|---|---|---|---|---|---|
| PANet* (Liu et al., 2018b) | 0.78 | 0.823 | 0.834 | 0.672 | 9.41 | 21.3 |
| HSFPN (Shi et al., 2025) | 0.821 | 0.782 | 0.826 | 0.659 | 6.6 | 18.7 |
| GDFPN (Xu et al., 2022) | 0.788 | 0.806 | 0.825 | 0.669 | 12.2 | 25.6 |
| CGRFPN (Ni et al., 2024) | 0.816 | 0.759 | 0.826 | 0.655 | 3.29 | 8.6 |
| ERepFPN (Li et al., 2022a) | 0.764 | 0.801 | 0.813 | 0.663 | 12.07 | 28.2 |
| BiFPN (Tan & Le, 2019) | 0.805 | 0.773 | 0.828 | 0.674 | 7.77 | 24.1 |
| **SAN** | **0.829** | 0.787 | **0.838** | **0.681** | 9.77 | 26.3 |

**Table 6**
Comparative results of long-range modeling methods employed within SAN on the SIMD dataset.

| Methods | *Precision* | *Recall* | *mAP*@50 | *mAP*@50: 95 | *Params.*(*M*) | *FLOPs*(*G*) |
|---|---|---|---|---|---|---|
| CAA (Cai et al., 2024) | 0.805 | 0.77 | 0.817 | 0.651 | 9.76 | 26.3 |
| C2fCIB (Wang et al., 2024a) | 0.767 | 0.808 | 0.822 | 0.657 | 9.15 | 25.4 |
| GLSA (Tang et al., 2023) | 0.803 | 0.793 | 0.83 | 0.667 | 10.6 | 27.5 |
| CA (Hou et al., 2021) | 0.763 | 0.826 | 0.836 | 0.668 | 9.11 | 24.1 |
| **SAN(VSS)** | **0.847** | 0.79 | **0.842** | **0.683** | 11.42 | 30.8 |

**Table 7**
Ablation study of core module for VSS and DCA at different locations.

| Neck | VSS | Head | DCA | *mAP*@50 | *mAP*@50: 95 | *Params.*(*M*) | *FLOPs*(*G*) |
|---|---|---|---|---|---|---|---|
| - | - | - | - | 0.834 | 0.672 | 9.41 | 21.3 |
| ✓ | 1 | ✓ | 3 | 0.845 | 0.683 | 12.71 | 34.2 |
| ✓ | 3 | - | - | 0.845 | 0.685 | 11.57 | 36.2 |
| ✓ | 3 | ✓ | 1(Up) | 0.834 | 0.674 | 11.68 | 31.8 |
| ✓ | 3 | ✓ | 1(Down) | 0.837 | 0.683 | 11.68 | 31.8 |
| ✓ | 2 | ✓ | 2 | 0.844 | 0.691 | 11.71 | 32.8 |

The construction of feature pyramids constitutes an essential procedure in a wide array of computer vision tasks and represents an indispensable component of modern detection frameworks. It lays the groundwork for addressing challenges associated with multiscale object recognition and analysis. Table 5 presents results of our proposed SAN with current prevalent FPN methods. It's noticeable that SAN achieves leading detection results across listed methods at main evaluation metric $mAP@50$ and $mAP@50:95$. Compared the baseline PANet architecture, SAN improves the accuracy at $mAP@50:95$ by 1.34% with 0.681. BiFPN achieves 0.674 at $mAP@50:95$, placing it at the pinnacle in comparison to alternative methods. This superior outcome of BiFPN is attributed to its innovative and efficient cross-connection design, facilitating effective feature information transfer from disparate layers. Nevertheless, it exhibits suboptimal performance in comparison to SANs. Thanks to its aggregation and expansion modules design, SAN could fully explore the multi-scale distinctive information and then generate differentiable details of small objects in deep networks.

Table 6 offers comparison results between VSS block with mainstream long-range modeling methods employed in the our proposed SAN architecture. SS2D plays a vital role in VSS block, by traversing the input image patches along four scanning paths, SS2D bridges the gap between the sequential nature of one-dimensional selective scanning and the non-sequential structure of two-dimensional visual data. This design enables the model to aggregate contextual information from diverse sources and perspectives, thereby enhancing its capability to process two-dimensional visual data more effectively. Compared with channel attention method CA, VSS block shows visible detection accuracy increase at $mAP@50:95$ with 0.683. CAA module acts as core feature procedure unit in the PKINet (Cai et al., 2024) which achieves similar effects to larger convolutional kernels and obtains 0.651 at $mAP@50:95$ by 4.92% lower than VSS, indicating that CAA suffers information loss when applied in the SAN architecture. The VSS module plays a pivotal role in the SAN, it enhances the model's capacity to process 2D visual data and optimizes the architecture to boost computational efficiency.

A series of ablation experiments were conducted to ascertain the impact of employment of VSS module and DCA attention mechanism at different position within the SAN network towards the utilization of multiscale features, shown in Table 7. The experimental positions were differentiated by the application of the color blue and green, respectively, in order to denote the variation numbers of VSS and DCA in each respective position.As the results shown in the bottom line, the optimal solution for VSS and DCA follows two VSS modules and two DCA modules in Neck for $80\times80$ and $20\times20$ head optimization, by contrast to $3+1$ combination strategy. This underscores the pivotal function of VSS in capturing the dependencies of detail information across multiple scales, while concurrently demonstrating proficiency in effectively managing consistent spatial and semantic feature representations of diminutive targets. Conversely, the DCA channel mechanism has been shown to be more effective in correcting and condensing the fused information, thereby reducing the feature blending effect.

Table 8 lists the comparison of the detection performance of KSCNet with the same level of parametric quantity detectors on $mAP@50:95$. It can be seen that YOLOv5s has the smallest number of model parameters, but achieves the lowest accuracy. Compared to YOLOv8s, YOLOv10s and YOLO11s, although KSCNet has a slight increase in the number of parameters, the improvement in accuracy achieved is satisfactory. RT-DETR exhibits a substantial disparity in detection accuracy when compared to KSCNet, attribute to its intricate structural design and the substantial parameters resulting from self-attention. It is noteworthy that KSCNet achieves an inference speed of 164 FPS, fulfilling the criteria for real-time object detection and demonstrating strong potential for practical deployment in real-world applications.

To further validate the effectiveness of the different components in KSCNet on model performance, we conducted detailed ablation experiments as shown in Table 9. When CKN blocks were employed as the basic backbone, the detection accuracy at $mAP@50:95$ surged from 0.676 to 0.681. Based on the CKN, SAN fusion strategy further enhance the whole detection performance to 0.685, emphasizing that the SAN is capable to refine multi-stages feature than baseline method PANet. Solely

**Table 8**
Comparison with the SOAT algorithms on the SIMD dataset.

| Methods | $mAP@50$ | $mAP@50:95$ | $Params.(M)$ | $FLOPs(G)$ | $FPS$ |
|---|---|---|---|---|---|
| SSD512 (Liu et al., 2016) | 0.752 | 0.496 | 26.45 | 89.5 | 154 |
| RT-DETR-R50 (Zhao et al., 2024) | 0.75 | 0.609 | 411.96 | 125.7 | 72 |
| YOLOv5s (Jocher, 2020) | 0.811 | 0.639 | 7 | 15.9 | 295 |
| YOLOv8s (Jocher et al., 2023) | 0.836 | 0.677 | 11.1 | 28.5 | 280 |
| YOLOv10s (Wang et al., 2024a) | 0.822 | 0.664 | 8.04 | 24.5 | 253 |
| YOLO11s (Khanam & Hussain, 2024) | 0.83 | 0.676 | 9.41 | 21.3 | 232 |
| YOLO12s (Tian et al., 2025) | 0.828 | 0.672 | 9.23 | 21.2 | 243 |
| **KSCNet** | **0.844** | **0.691** | 11.71 | 32.8 | 164 |

**Table 9**
Ablation study of the improvement modules on the SIMD dataset.

| Baseline | CKN | SAN | VSS | DCA | $mAP@50$ | $mAP@50:95$ | $Params.$ | $FLOPs(G)$ |
|---|---|---|---|---|---|---|---|---|
| ✓ | | | | | 0.83 | 0.676 | 9.41 | 21.3 |
| ✓ | ✓ | | | | 0.836 | 0.681 | 11.44 | 26.7 |
| ✓ | ✓ | ✓ | | | 0.841 | 0.685 | 11.5 | 28.3 |
| ✓ | ✓ | | ✓ | | 0.831 | 0.679 | 11.56 | 31.5 |
| ✓ | ✓ | ✓ | ✓ | ✓ | 0.844 | 0.691 | 11.71 | 32.8 |

**Table 10**
Experimental results on VisDrone *val* set with YOLO series methods.

| Methods | $Precision$ | $Recall$ | $mAP@50$ | $mAP@50:95$ | $Params.(M)$ | $FLOPs(G)$ |
|---|---|---|---|---|---|---|
| YOLOv5n (Jocher, 2020) | 0.341 | 0.272 | 0.248 | 0.125 | 1.77 | 4.2 |
| YOLOv5s (Jocher, 2020) | 0.444 | 0.326 | 0.323 | 0.171 | 7 | 15.8 |
| YOLOv8n (Jocher et al., 2023) | 0.452 | 0.342 | 0.339 | 0.195 | 3 | 8.1 |
| YOLOv8s (Jocher et al., 2023) | 0.504 | 0.393 | 0.399 | 0.236 | 11.12 | 28.5 |
| YOLOv10n (Wang et al., 2024a) | 0.452 | 0.362 | 0.332 | 0.193 | 2.69 | 8.2 |
| YOLOv10s (Wang et al., 2024a) | 0.499 | 0.38 | 0.386 | 0.23 | 8.04 | 24.5 |
| YOLO11n (Khanam & Hussain, 2024) | 0.45 | 0.34 | 0.335 | 0.193 | 2.58 | 6.3 |
| YOLO11s* (Khanam & Hussain, 2024) | 0.507 | 0.385 | 0.394 | 0.235 | 9.4 | 21.3 |
| YOLOv12n (Tian et al., 2025) | 0.455 | 0.334 | 0.336 | 0.195 | 2.55 | 6.3 |
| YOLOv12s (Tian et al., 2025) | 0.504 | 0.387 | 0.395 | 0.236 | 9.23 | 21.2 |
| **KSCNet** | **0.534** | **0.401** | **0.413** | **0.247** | 12.1 | 33 |

using VSS without SAN will not bring improvement, explaining that the efficacy of VSS is contingent upon its integration within a meticulously designed FPN network, wherein it can facilitate the efficient processing of features. Ultimately, the amalgamation of all the aforementioned design elements culminates in the formation of KSCNet. This network demonstrates a substantial enhancement in detection accuracy, registering a relative improvement of 1.69% and 2.22% in terms of $mAP@50$ and $mAP@50:95$, respectively compared with the baseline. These results serve as a robust validation of the efficacy of the proposed modifications in facilitating efficient and stable detection of small targets by UAVs.

In order to further validate the robustness of KSCNet on the UAV small target detection task, a number of comparative experiments were conducted on the VisDrone dataset. In Table 10, the results of experimental investigation are presented in which the KSCNet model was compared with the family of algorithms on the VisDrone *val* set. The experimental results reveal that KSCNet attains the highest detection scores across the four primary evaluation metrics, significantly outperforming other YOLO-based algorithms. Specifically, compared with the baseline model, KSCNet achieves respective increments of 5.32%, 4.16%, 4.82%, and 5.11% in the evaluation metrics $Precision$, $Recall$, $mAP@50$, and $mAP@50:95$. Similarly, Table 11 presents the comparative experiments of KSCNet on the VisDrone *test* dataset, where it demonstrates remarkable performance enhancements of 5.67%, 2.1%, 4.43% and 6.15% relative to the baseline YOLO11s, outperforms the rest of YOLO algorithms across the evaluation metrics. Table 12 presents the experimental results comparing KSCNet with current state-of-the-art (SOTA) methods. Compared to the latest object detectors, KSCNet achieves the highest detection performance with fewer parameters and lower model complexity when given smaller input image size. The aforementioned experiments have demonstrated that KSCNet, owing to its highly efficient model design, is capable of conducting in-depth exploration and focusing on small target features within UAV imagery and leads to its robust performance across various datasets. The strong robustness exhibited by KSCNet can be attributed to the stable synergistic processing of the KAN, VSS, SAN as well as DCA mechanism. The collaboration of these components endows KSCNet with superior detection accuracy and stability for small UAV object detection.

### 4.4. Quantitative results and analysis

In order to demonstrate the excellent performance of KSCNet in the UAV small target detection, a series of representative algorithms will be selected for comparative analysis in this section. To offer a more intuitive illustration of model performance, we visualize the precision, recall, mAP@50, and mAP@50-95 curves of KSCNet alongside real-time YOLO methods in Figs. 8 and 9, which correspond to the SIMD dataset and the VisDrone dataset, respectively. Among the evaluation indexes listed above, KSCNet is denoted by the brown curve, which demonstrates a gradual improvement during the training process. The final experimental results are notably higher than those of other algorithms following the completion of total 300 training epochs, indicating that KSCNet exhibits consistent and stable performance without significant fluctuations during the training. This stability is attributed to the collaborative processing among its internal modules. As illustrated in Fig. 9, KSCNet demonstrates superior performance results at the end of the training, a notable distinction from its comparable algorithms of equivalent size and magnitude. This observation underscores KSCNet's remarkable scalability within the VisDrone dataset.

Fig. 10 illustrates the detection results of KSCNet in comparison with the baseline model on the SIMD dataset. The empirical results reveal that the proposed KSCNet exhibits enhanced performance relative to the

**Table 11**
Experimental results on VisDrone *test* set with YOLO series methods.

| Methods | *Precision* | *Recall* | *mAP*@50 | *mAP*@50: 95 | *Params.*(*M*) | *FLOPs*(*G*) |
|---|---|---|---|---|---|---|
| YOLOv5n (Jocher, 2020) | 0.306 | 0.252 | 0.208 | 0.102 | 1.77 | 4.2 |
| YOLOv5s (Jocher, 2020) | 0.444 | 0.326 | 0.323 | 0.171 | 7 | 15.8 |
| YOLOv8n (Jocher et al., 2023) | 0.39 | 0.299 | 0.267 | 0.147 | 3 | 8.1 |
| YOLOv8s (Jocher et al., 2023) | 0.443 | 0.332 | 0.314 | 0.178 | 11.12 | 28.5 |
| YOLOv10n (Wang et al., 2024a) | 0.396 | 0.289 | 0.268 | 0.149 | 2.69 | 8.2 |
| YOLOv10s (Wang et al., 2024a) | 0.444 | 0.328 | 0.311 | 0.177 | 8.04 | 24.5 |
| YOLO11n (Khanam & Hussain, 2024) | 0.386 | 0.297 | 0.266 | 0.147 | 2.58 | 6.3 |
| YOLO11s* (Khanam & Hussain, 2024) | 0.441 | 0.339 | 0.316 | 0.179 | 9.4 | 21.3 |
| YOLOv12n (Tian et al., 2025) | 0.39 | 0.298 | 0.27 | 0.151 | 2.55 | 6.3 |
| YOLOv12s (Tian et al., 2025) | 0.45 | 0.341 | 0.321 | 0.186 | 9.23 | 21.2 |
| **KSCNet** | **0.466** | **0.346** | **0.33** | **0.19** | 12.1 | 33 |

**Table 12**
Experimental results on VisDrone *val* with state of the art methods.

| Methods | input size | *mAP*@50 | *mAP*@50: 95 | *Params.*(*M*) | *Params.*(*M*) |
|---|---|---|---|---|---|
| Faster RCNN (Ren et al., 2016) | 1333 × 800 | 0.384 | 0.232 | 41.39 | 208 |
| Cascade RCNN (Cai & Vasconcelos, 2018) | 1333 × 800 | 0.392 | 0.245 | 29.18 | 236 |
| ATSS (Zhang et al., 2020) | 1333 × 800 | 0.362 | 0.221 | 32 | 203 |
| FCOS (Detector, 2022) | 1344 × 768 | 0.316 | 0.188 | 32.13 | 198 |
| DTSSNet (Chen et al., 2024a) | 1333 × 800 | 0.399 | 0.242 | 10.1 | 50.3 |
| HyperYOLOs (Feng et al., 2024) | 640 × 640 | 0.409 | 0.244 | 14.8 | 38.9 |
| RT-DETR-R18 (Zhao et al., 2024) | 640 × 640 | 0.409 | 0.241 | 19.88 | 57 |
| **KSCNet** | 640 × 640 | **0.413** | **0.247** | 12.1 | 33 |

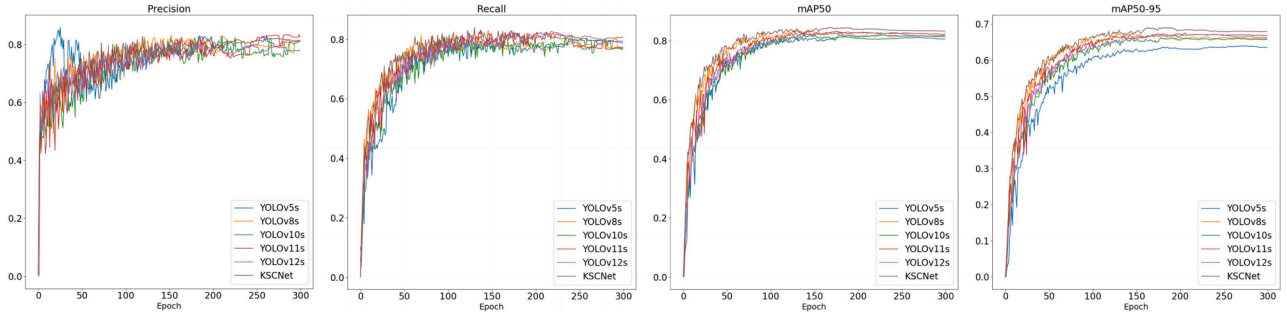


**Fig. 8.** Training curves of Precision, Recalll, *mAP*@50 and *mAP*@50: 95 for YOLO representative algorithms and KSCNet on SIMD dataset.

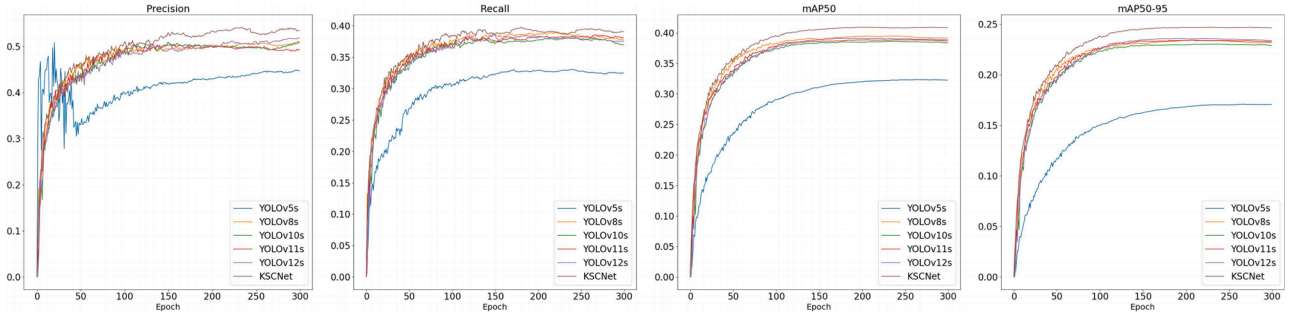


**Fig. 9.** Training curves of *Precision*, *Recall*, *mAP*@50 and *mAP*@50: 95 for YOLO representative algorithms and KSCNet on VisDrone dataset.

baseline counterpart, achieving higher classification accuracy and superior object localization capabilities across diverse scenarios. Notably, KSCNet consistently demonstrates the ability to accurately pinpoint and correctly classify aircraft models. This robust performance is attributed to its advanced feature extraction mechanisms, which effectively capture salient characteristics of aircraft even under complex background conditions. Moreover, the effectiveness of SAN and DCA attention strategies within KSCNet further enhances its generalization capability, ensuring high accuracy across diverse datasets and scenarios. Conversely, the baseline model occasionally encounters difficulties in accurately identifying the models under certain conditions.

Heatmaps can serve as a powerful tool to elucidate the inner workings of a model and demonstrate that the network has effectively captured meaningful information. Grad-CAM (Selvaraju et al., 2017), a widely recognized visualization technique, pinpoints the regions within the feature map of a deep neural network that are most influential in shaping the prediction outcomes. In this study, we employ this technique to visualize the feature maps of the middle layers in both the baseline and KSCNet. As shown in Fig. 11, KSCNet pays more attention to small target area than baseline. Therefore, the proposed KSCNet exhibits enhanced capability in extracting key features from UAV images. It demonstrates superior robustness in the presence of substantial

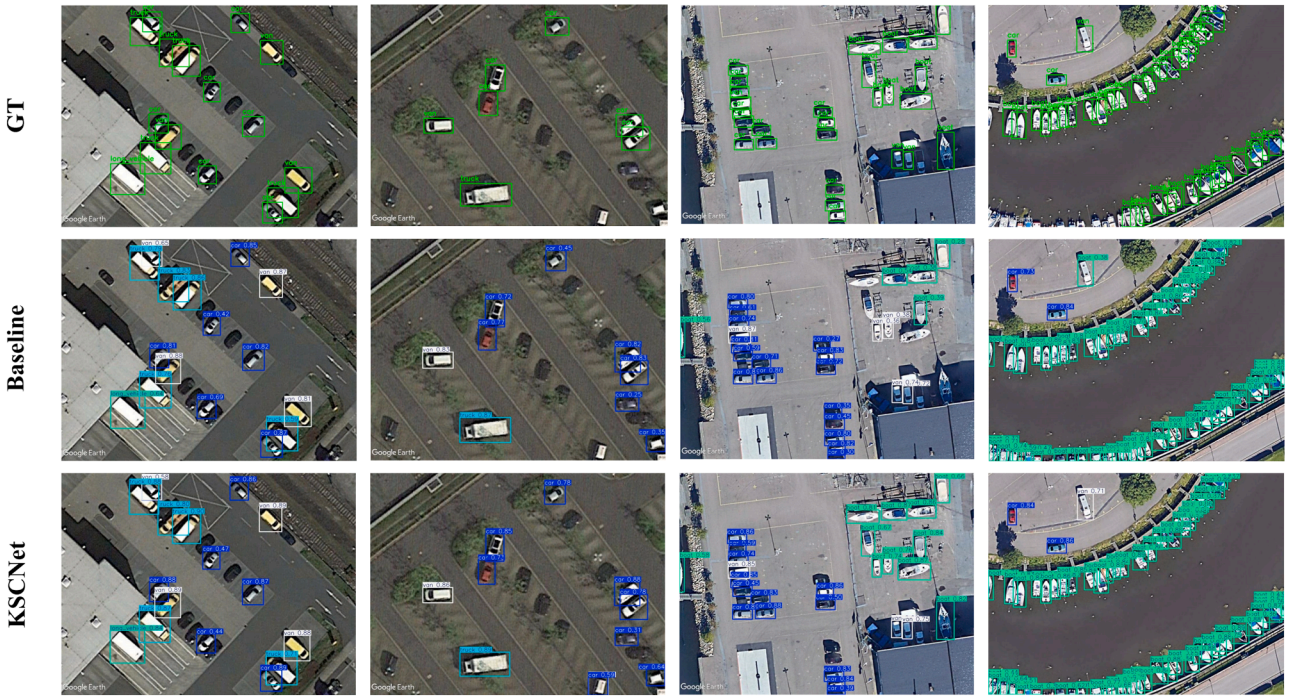


**Fig. 10.** Qualitative visualization comparisons on the SIMD dataset of KSCNet and baseline model.

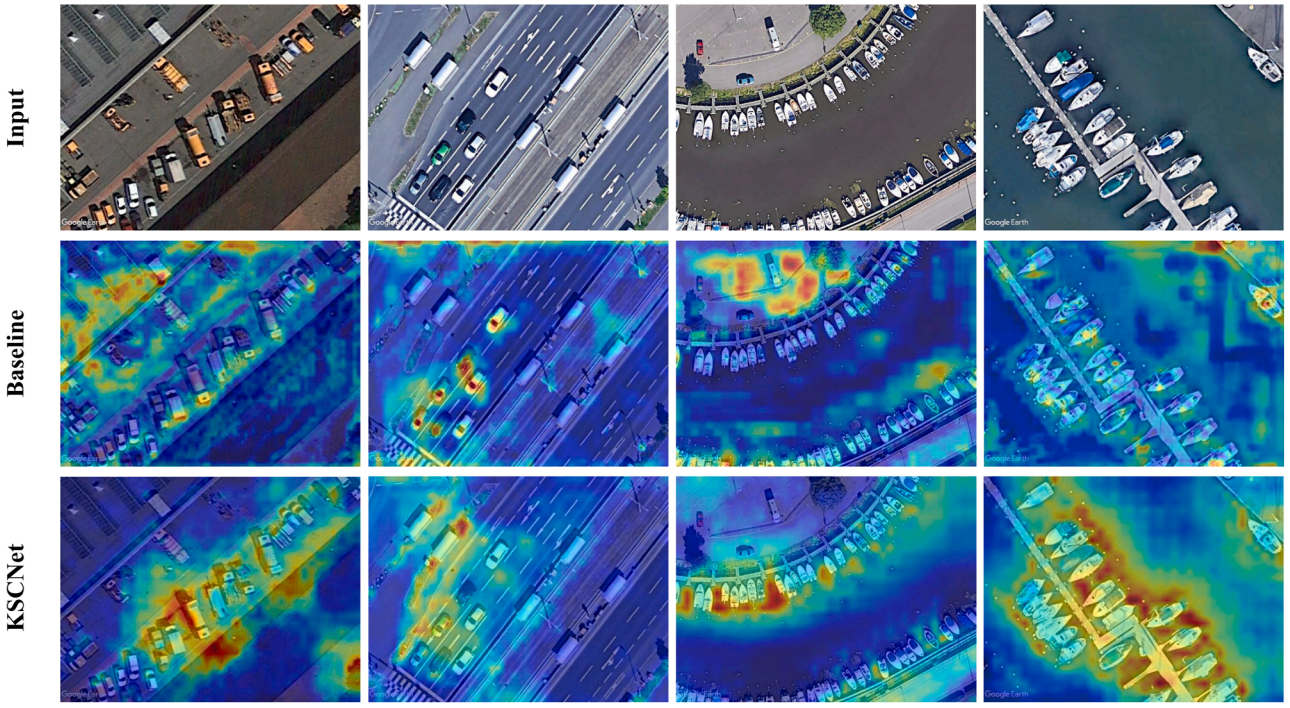


**Fig. 11.** Visualization comparison of Heatmap by Grad-CAM (Selvaraju et al., 2017) between baseline and KSCNet.

background interference, effectively mitigating the impact of such distractions on detection performance.

To further clarify the effectiveness of the proposed KSCNet in enhancing small object detection from UAV images, we meticulously selects four representative and challenging samples from the VisDrone dataset for detailed analysis. In Fig. 12, we present aerial images captured by UAV depicting a major urban thoroughfare. This image showcases bustling vehicular traffic and an extensive array of intricate objects, thereby rendering the scene highly complex and multifaceted. Experiments were conducted by KSCNet with recent released YOLO realtime detectors, YOLOv10s, YOLO11s and YOLOv12s, which execute constant iterative optimization of version updates based on YOLO. It is evident that the remaining three models exhibit relatively weaker detection capabilities for distant small objects. These models frequently demonstrate missed detections or lower class confidence scores for the targets. They particularly struggle with detecting pedestrians that have

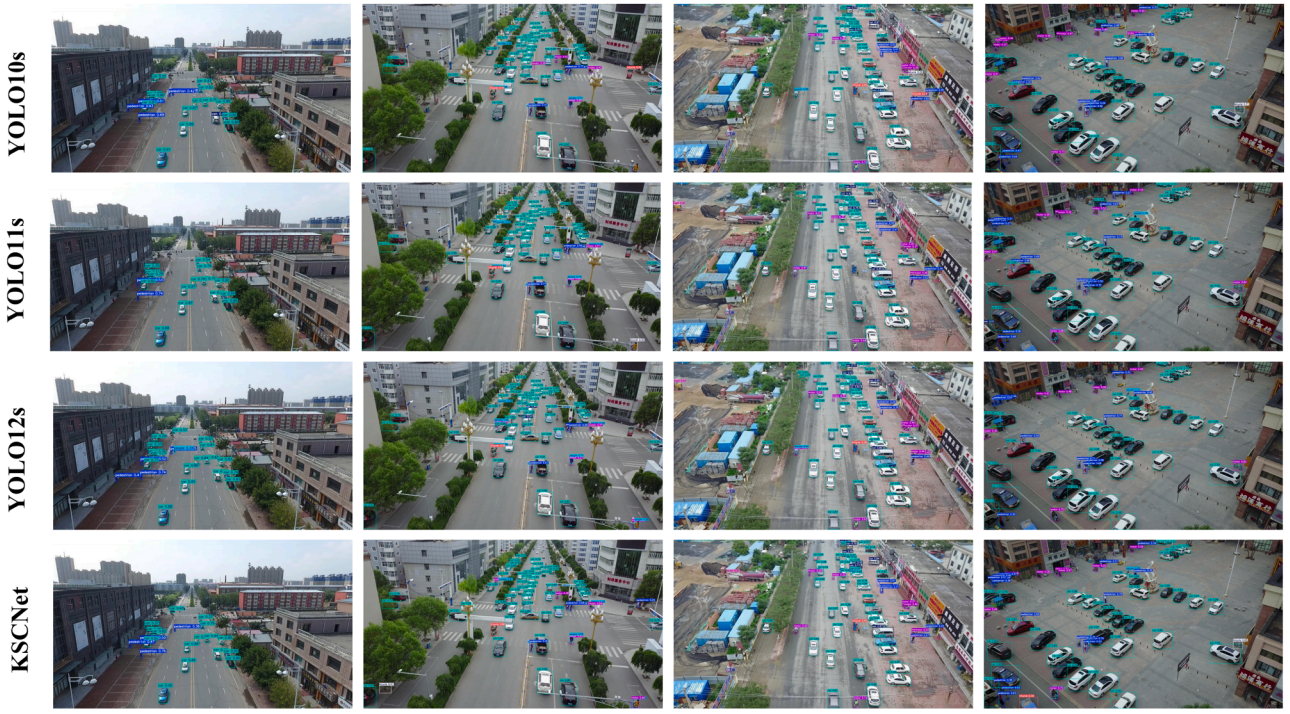


**Fig. 12.** Visualization of detection performance of KSCNet with advanced realtime YOLO detectors YOLO10s, YOLO11s and YOLOv12s on the VisDrone dataset.

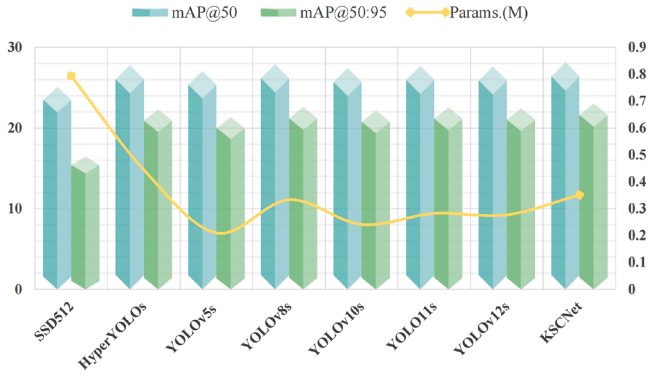


**Fig. 13.** Comparisons of mainstream detectors in $mAP@50$, $mAP@50{:}\,95$ and model parameters.

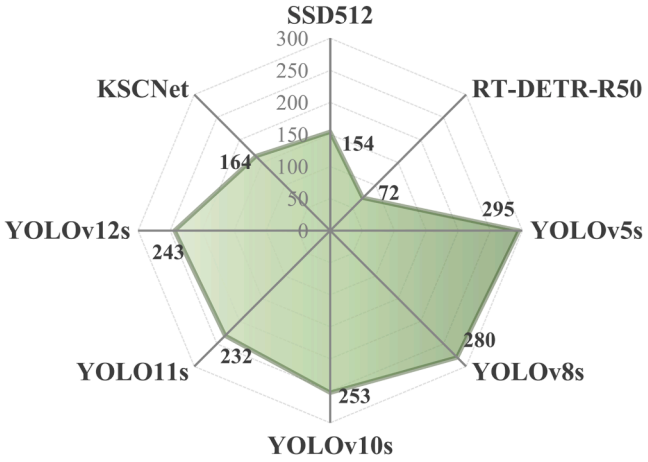


**Fig. 14.** Radar graph of FPS inference speed of KSCNet with mainstream real-time detectors.

minimal pixel coverage and occasionally misclassify vehicle categories. Conversely, the proposed KSCNet exhibits enhanced overall detection performance, successfully identifying a substantial proportion of the dense crowd traversing the pedestrian crossing.

To facilitate an intuitive understanding of the comparative performance between KSCNet and other models, Fig. 13 presents a comparison across key evaluation metrics, including $mAP@50$, $mAP@50{:}\,95$, and the number of model parameters. The results clearly demonstrate that KSCNet achieves a superior balance between detection precision and model complexity, outperforming most counterparts in accuracy while maintaining a compact structure. Furthermore, Fig. 14 provides a radar graph comparison of the FPS inference speed between KSCNet and current real-time object detectors, highlighting its competitive computational efficiency across different operational profiles. From the results presented, it can be observed that the proposed KSCNet not only achieves high detection accuracy but also attains an inference speed of 164 FPS and meeting the requirements for real-time object detection.

## 5. Limitations

Despite the promising performance achieved by the proposed KSCNet on standard UAV benchmarks, we acknowledge several limitations that present opportunities for future work. First, the incorporation of the KAN and SSM modules increases the computational complexity and inference latency of the model. This elevated cost may hinder its deployment on resource-constrained UAV platforms for real-time applications. Future work will focus on model compression, knowledge distillation, or the design of more efficient operators to achieve a better trade-off between accuracy and speed. Second, the model's validation is primarily confined to standard public datasets. Its robustness and generalization performance under extreme real-world conditions, such as adverse weather or severe motion blur, require further investigation. The model's robustness and generalization capability in these challenging and niche environments require further investigation. Third, the

design and integration of the proposed SAN and DCA modules are tailored for a specific YOLO-based architecture. The transferability and effectiveness of these components to other detector families, such as two-stage or query-based models, remain an open question. Adapting and evaluating the proposed modules across a wider range of architectural backbones would be a valuable direction to explore their universal efficacy. Addressing these limitations will be the focus of our subsequent research efforts to make KSCNet more robust, efficient, and applicable to a broader spectrum of real-world aerial vision tasks.

## 6. Conclusion

In this work, we proposes KSCNet, an innovative detection framework designed to enhance small target detection in UAV aerial imagery. By integrating Kolmogorov-Arnold Networks (KAN) and State Space Model (SSM), KSCNet effectively addresses challenges posed by small target size, complex backgrounds and limited imaging resolution in complex environment. Concretely, CKN efficiently decomposes high-dimensional data into simpler functions, thereby extracting robust features for subsequent processing. The Semantic Aggregation Network (SAN), augmented with SSM, optimizes multiscale feature fusion and generalization for complex UAV images. A new channel attention mechanisms further refine feature integration within SAN. Extensive experiments on public UAV datasets validate KSCNet's efficacy. On the SIMD dataset, KSCNet achieves 0.844 $mAP@50$ and 0.691 $mAP@50:95$, with respective improvements of 1.69% and 2.22% over the baseline. On the VisDrone dataset, KSCNet demonstrates significant accuracy increases of 4.82% and 5.11% on the validation set and 4.43% and 6.16% on the test set at $mAP@50$ and $mAP@50:95$. These results highlight KSCNet's superior performance in UAV small target detection, providing valuable technical support for related applications.

## CRediT authorship contribution statement

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

## Data availability

Data will be made available on request.

## 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.

## Acknowledgment

This work was supported by the National Natural Science Foundation of China (62376252); Key Project of Natural Science Foundation of Zhejiang Province (LZ22F030003); Zhejiang Province Leading Geese Plan (2025C02025, 2025C01056); Zhejiang Province Province-Land Synergy Program(2025SDXT004-3).

## References

Bodner, A. D., Tepsich, A. S., Spolski, J. N., & Pourteau, S. (2024). Convolutional Kolmogorov-Arnold networks. arXiv preprint arXiv:2406.13155.

Cai, X., Lai, Q., Wang, Y., Wang, W., Sun, Z., & Yao, Y. (2024). Poly Kernel inception network for remote sensing detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 27706–27716).

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).

Chen, J., Kao, S.-h., He, H., Zhuo, W., Wen, S., Lee, C.-H., & Chan, S.-H. G. (2023). Run, don't walk: Chasing higher FLOPS for faster neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 12021–12031).

Chen, L., Liu, C., Li, W., Xu, Q., & Deng, H. (2024a). Dtssnet: Dynamic training sample selection network for uav object detection. *IEEE Transactions on Geoscience and Remote Sensing*, *62*, 1–16.

Chen, Y., Ye, Z., Sun, H., Gong, T., Xiong, S., & Lu, X. (2025a). Global-local fusion with semantic information-guidance for accurate small object detection in UAV aerial images. *IEEE Transactions on Geoscience and Remote Sensing*, *63*, 1–15. https://doi.org/10.1109/TGRS.2025.3532612

Chen, Y., Yuan, X., Wang, J., Wu, R., Li, X., Hou, Q., & Cheng, M.-M. (2025b). Yolo-ms: Rethinking multi-scale representation learning for real-time object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *47*(6) 4240–4252. https://doi.org/10.1109/TPAMI.2025.3538473

Chen, Z., He, Z., & Lu, Z.-M. (2024b). DEA-Net: Single image dehazing based on detail-enhanced convolution and content-guided attention. *IEEE Transactions on Image Processing*, *33*, 1002–1015.

Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1251–1258).

Du, D., Zhu, P., Wen, L., Bian, X., Lin, H., Hu, Q., Peng, T., Zheng, J., Wang, X., Zhang, Y., Bo, L., Shi, H., Zhu, R., Kumar, A., Li, A., Zinollayev, A., Askergaliyev, A., Schumann, A., Mao, B., …Liu, Z. (2019). Visdrone-DET2019: The vision meets drone object detection in image challenge results. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)* (pp. 213–226).

Duan, K., Bai, S., Xie, L., Qi, H., Huang, Q., & Tian, Q. (2019). CenterNet: Keypoint triplets for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 6569-6578).

Fan, Q., Li, Y., Deveci, M., Zhong, K., & Kadry, S. (2025). Lud-YOLO: A novel lightweight object detection network for unmanned aerial vehicle. *Information Sciences*, *686*, 121366.

Feng, Y., Huang, J., Du, S., Ying, S., Yong, J.-H., Li, Y., Ding, G., Ji, R., & Gao, Y. (2024). Hyper-yolo: When visual object detection meets hypergraph computation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *47*, 2388–2401. https://doi.org/10.1109/TPAMI.2024.3524377

Gagliardi, V., Tosti, F., Bianchini Ciampoli, L., Battagliere, M. L., D'Amato, L., Alani, A. M., & Benedetto, A. (2023). Satellite remote sensing and non-destructive testing methods for transport infrastructure monitoring: Advances, challenges and perspectives. *Remote Sensing*, *15*(2), 418.

Ghiasi, G., Lin, T.-Y., & Le, Q. V. (2019). NAS-FPN: Learning scalable feature pyramid architecture for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 7036–7045).

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 580–587).

Gu, A., & Dao, T. (2023). Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752.

Gu, A., Dao, T., Ermon, S., Rudra, A., & Ré, C. (2020). Hippo: Recurrent memory with optimal polynomial projections. *Advances in Neural Information Processing Systems*, *33*, 1474–1487.

Gu, A., Johnson, I., Goel, K., Saab, K., Dao, T., Rudra, A., & Ré, C. (2021). Combining recurrent, convolutional, and continuous-time models with linear state space layers. *Advances in Neural Information Processing Systems*, *34*, 572–585.

Haroon, M., Shahzad, M., & Fraz, M. M. (2020). Multisized object detection using space-borne optical imagery. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *13*, 3032–3046.

Hassanin, M., Anwar, S., Radwan, I., Khan, F. S., & Mian, A. (2024). Visual attention methods in deep learning: An in-depth survey. *Information Fusion*, *108*, 102417.

He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision* (pp. 2961–2969).

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, *2*(5), 359–366.

Hou, Q., Zhou, D., & Feng, J. (2021). Coordinate attention for efficient mobile network design. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 13713–13722).

Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7132–7141).

Hu, M., Li, Y., Fang, L., & Wang, S. (2021). A2-FPN: Attention aggregation based feature pyramid network for instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 15343–15352).

Jing, R., Zhang, W., Li, Y., Li, W., & Liu, Y. (2024). Feature aggregation network for small object detection. *Expert Systems with Applications*, *255*, 124686.

Jocher, G. (2020). Yolov5 by ultralytics. https://github.com/ultralytics/yolov5. https://doi.org/10.5281/zenodo.3908559

Jocher, G., Chaurasia, A., & Qiu, J. (2023). Ultralytics YOLO. https://github.com/ultralytics/ultralytics.

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

Kolmogorov, A. N. (1961). On the representation of continuous functions of several variables by superpositions of continuous functions of a smaller number of variables. American Mathematical Society.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

Li, C., Li, L., Jiang, H., Weng, K., Geng, Y., Li, L., Ke, Z., Li, Q., Cheng, M., Nie, W. et al. (2022a). Yolov6: A single-stage object detection framework for industrial applications. arXiv preprint arXiv:2209.02976.

Li, J., Liang, X., Wei, Y., Xu, T., Feng, J., & Yan, S. (2017). Perceptual generative adversarial networks for small object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1222–1230).

Li, Y., Wang, L., & Wang, Z. (2022b). Single-shot object detection via feature enhancement and channel attention. *Sensors*, *22*(18), 6857.

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017a). Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2117–2125).

Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017b). Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision* (pp. 2980–2988).

Liu, S., Huang, D. et al. (2018a). Receptive field block net for accurate and fast object detection. In *Proceedings of the european conference on computer vision (ECCV)* (pp. 385–400).

Liu, S., Qi, L., Qin, H., Shi, J., & Jia, J. (2018b). Path aggregation network for instance segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 8759–8768).

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., & Berg, A. C. (2016). SSD: Single shot multibox detector. In *Computer vision–ECCV 2016: 14th European conference, Amsterdam, the Netherlands, October 11–14, 2016, proceedings, Part I 14* (pp. 21–37). Springer International Publishing.

Liu, Y., Tian, Y., Zhao, Y., Yu, H., Xie, L., Wang, Y., Ye, Q., Jiao, J., & Liu, Y. (2024). Vmamba: Visual state space model. *Advances in Neural Information Processing Systems*, *37*, 103031–103063.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 10012–10022).

Martinez-Alpiste, I., Golcarenarenji, G., Wang, Q., & Alcaraz-Calero, J. M. (2021). Search and rescue operation using UAVs: A case study. *Expert Systems with Applications*, *178*, 114937.

Motlagh, N. H., Kortoçi, P., Su, X., Lovén, L., Hoel, H. K., Haugsvær, S. B., Srivastava, V., Gulbrandsen, C. F., Nurmi, P., & Tarkoma, S. (2023). Unmanned aerial vehicles for air pollution monitoring: A survey. *IEEE Internet of Things Journal*, *10*(24), 21687–21704.

Ni, Z., Chen, X., Zhai, Y., Tang, Y., & Wang, Y. (2024). Context-guided spatial feature reconstruction for efficient semantic segmentation. arXiv preprint arXiv:2405.06228.

Noh, J., Bae, W., Lee, W., Seo, J., & Kim, G. (2019). Better to follow, follow to be better: Towards precise supervision of feature super-resolution for small object detection. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 9725–9734).

Redmon, J. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Ren, S., He, K., Girshick, R., & Sun, J. (2016). Faster r-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(6), 1137–1149.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision* (pp. 618–626).

Shi, Z., Hu, J., Ren, J., Ye, H., Yuan, X., Ouyang, Y., He, J., Ji, B., & Guo, J. (2025). HS-FPN: High frequency and spatial perception fpn for tiny object detection. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 6896–6904). (*vol. 39*).

Tian, Z., Shen, C., Chen, H., & He, T. (2022). Fcos: A simple and strong anchor-free object detector. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(4), 1922–1933. https://doi.org/10.1109/TPAMI.2020.3032166

Tan, M., & Le, Q. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning* (pp. 6105–6114). PMLR.

Tan, M., Pang, R., & Le, Q. V. (2020). Efficientdet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10781–10790).

Tang, F., Xu, Z., Huang, Q., Wang, J., Hou, X., Su, J., & Liu, J. (2023). DuAT: Dual-aggregation transformer network for medical image segmentation. In *Chinese conference on pattern recognition and computer vision (PRCV)* (pp. 343–356). Springer.

Tian, Y., Ye, Q., & Doermann, D. (2025). Yolov12: Attention-centric real-time object detectors. arXiv preprint arXiv:2502.12524.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In Guyonxsnm, I., Von Luxburg, U., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., & Garnett, R., (eds.), *Curran Associates, Inc.*, (vol. *30*). https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Wang, A., Chen, H., Liu, L., Chen, K., Lin, Z., Han, J., & Ding, G. (2024a). Yolov10: Real-time end-to-end object detection. arXiv preprint arXiv:2405.14458.

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, C.-Y., Yeh, I.-H., & Liao, H.-Y. M. (2024b). Yolov9: Learning what you want to learn using programmable gradient information. arXiv preprint arXiv:2402.13616.

Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., Liu, W., & Xiao, B. (2020). Deep high-resolution representation learning for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *43*(10), 3349–3364 https://doi.org/10.1109/TPAMI.2020.2983686.

Wang, X., Wang, S., Ding, Y., Li, Y., Wu, W., Rong, Y., Kong, W., Huang, J., Li, S., Yang, H. et al. (2024c). State space model for new-generation network alternative to transformers: A survey. arXiv preprint arXiv:2404.09516.

Wu, X., Li, W., Hong, D., Tao, R., & Du, Q. (2021). Deep learning for unmanned aerial vehicle-based object detection and tracking: A survey. *IEEE Geoscience and Remote Sensing Magazine*, *10*(1), 91–124.

Wu, Y.-H., Liu, Y., Zhan, X., & Cheng, M.-M. (2022). P2t: Pyramid pooling transformer for scene understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(11), 12760–12771.

Xiao, Y., Xu, T., Yu, X., Fang, Y., & Li, J. (2024). A lightweight fusion strategy with enhanced inter-layer feature correlation for small object detection. *IEEE Transactions on Geoscience and Remote Sensing*, *62*, 1–11, https://doi.org/10.1109/TGRS.2024.3457155.

Xu, X., Jiang, Y., Chen, W., Huang, Y., Zhang, Y., & Sun, X. (2022). Damo-yolo: A report on real-time object detection design. arXiv preprint arXiv:2211.15444.

Xue, C., Xia, Y., Wu, M., Chen, Z., Cheng, F., & Yun, L. (2024). El-yolo: An efficient and lightweight low-altitude aerial objects detector for onboard applications. *Expert Systems with Applications*, *256*, 124848.

Yu, W., Luo, M., Zhou, P., Si, C., Zhou, Y., Wang, X., Feng, J., & Yan, S. (2022). Metaformer is actually what you need for vision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10819–10829).

Yue, M., Zhang, L., Huang, J., & Zhang, H. (2024). Lightweight and efficient tiny-object detection based on improved YOLOv8n for UAV aerial images. *Drones*, *8*(7), 276.

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

Zhang, H., Wang, L., Tian, T., & Yin, J. (2021). A review of unmanned aerial vehicle low-altitude remote sensing (UAV-LARS) use in agricultural monitoring in china. *Remote Sensing*, *13*(6), 1221.

Zhang, H., Wen, S., Wei, Z., & Chen, Z. (2024a). High-resolution feature generator for small ship detection in optical remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, *62*, 1–11, https://doi.org/10.1109/TGRS.2024.3377999.

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, Z., Huang, L., Wang, Q., Jiang, L., Qi, Y., Wang, S., Shen, T., Tang, B.-H., & Gu, Y. (2024b). UAV hyperspectral remote sensing image classification: A systematic review. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *18*, 3099–3124, https://doi.org/10.1109/JSTARS.2024.3522318.

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).

Zheng, M., Sun, L., Dong, J., & Pan, J. (2024). SMFANet: A lightweight self-modulation feature aggregation network for efficient image super-resolution. In *European conference on computer vision* (pp. 359–375). Springer.

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.