SD-Net: Synergistic and Dynamic Learning for Small Object Detection in Aerial Images

Jiaxin Yang, Wenbo Liu, Tao Deng, Fei Yan
Southwest Jiaotong University
Chengdu 611756, P.R. China
1977104723@qq.com, liuwenbo@my.swjtu.edu.cn, tdeng@swjtu.edu.cn, fyan@home.swjtu.edu.cn

Abstract

The widespread deployment of Unmanned Aerial Vehicles (UAVs) in critical sectors such as public safety and environmental monitoring has made robust aerial object detection an essential technology. However, the unique top-down viewpoint of aerial imagery frequently results in objects that are small-scale and densely clustered. This presents a fundamental challenge for object detection, rooted in the feature sparsity of small-scale objects, where intrinsic details are insufficient for reliable recognition. Consequently, many prevailing detectors, which rely on static operators with fixed receptive fields, struggle to effectively address this task. In this paper, we propose the Synergistic and Dynamic Network (SD-Net), an object detector developed upon the YOLOv11 architecture, which systematically enhances feature representation by introducing adaptive computation mechanisms at the stages of feature extraction, multi-scale fusion, and feature refinement. Specifically, we design the Spatial-Context C2f (SC2f) block to introduce content-adaptive computation in the backbone, the Synergistic Scaling Block (S2B) to perform cross-scale feature modulation in the neck, and the Hybrid Convolution Block (HCB) for specialized refinement of high-resolution features. Extensive experiments on the challenging VisDrone and TinyPerson aerial imagery benchmarks demonstrate the effectiveness of our method. On these two datasets, our proposed SD-Net-s achieves substantial mAP_{50} improvements of 9.7% and 7.3% over its baseline, respectively, demonstrating its superior performance in complex aerial scenes.

***Keywords:** Object Detection, Aerial Imagery, Small Object Detection, Context-Aware Representation, Multi-Scale Feature Fusion*

1. Introduction

Unmanned Aerial Vehicles (UAVs) have become a critical platform for acquiring high-altitude visual data, serving pivotal roles in applications such as environmental monitoring, disaster management, and public safety [5, 43]. This has driven the development of object detection techniques tailored for aerial imagery. However, the unique perspective and imaging distance of aerial platforms present a distinct set of challenges for object detection compared to generic scenes, as illustrated in Figure 1.

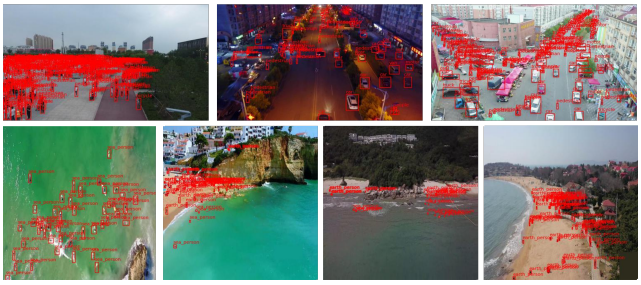

Figure 1. Examples of challenging scenes in aerial imagery datasets. The images are characterized by small, densely packed objects, complex backgrounds, and variable lighting conditions.

A prominent characteristic of objects in aerial images is their small pixel footprint on the imaging plane, which directly leads to a reduction in discernible information of the objects themselves. As the size of an object diminishes, its intrinsic shape and texture features can become ambiguous, increasing the likelihood of confusion with the background or other object classes [12]. Furthermore, scenes from a UAV perspective are often vast and complex, with objects frequently appearing in dense clusters and subject to occlusion, which adds to the difficulty of precise localization. These factors collectively constitute the core difficulties of the aerial small object detection task.

While existing detection methods have made progress in addressing these issues, for instance, by fusing multi-scale information through Feature Pyramid Networks (FPNs) [37] or by focusing on salient regions using attention mechanisms [51, 58], challenges persist. The effectiveness of these enhancement strategies can be limited when the objects themselves lack sufficient intrinsic features. In FPNs, for example, the direct fusion of high-level semantic features with low-level spatial details can lead to seman-

tic misalignment, potentially diluting the precise spatial information crucial for small object localization. Moreover, when regions containing small objects share similar textures with a complex background, the underlying feature representations themselves may not be sufficiently discriminative, making it difficult for any subsequent enhancement strategy to yield optimal results.

In response to these persistent challenges, this paper aims to improve upon the core computational paradigm itself by proposing a new network architecture, termed the Synergistic and Dynamic Network (SD-Net). Instead of designing a single add-on module, we systematically redesign the computational flow at three critical stages of the network. First, we design the Spatial-Context C2f (SC2f) block to serve as a core feature extraction block within the backbone. SC2f introduces a content-adaptive computation mechanism that allows it to adjust its convolutional operation based on input features and to expand its range of spatial information exchange. Second, at the cross-scale fusion stages of the feature pyramid, we propose the Synergistic Scaling Block (S2B). This module establishes an explicit feature modulation relationship, allowing features rich in semantic context to guide the recalibration of features carrying fine-grained spatial details. Finally, we introduce the Hybrid Convolution Block (HCB) for final feature refinement before the detection head, which dynamically fuses feature streams from different computational paradigms through a gating mechanism.

The main contributions of this work are as follows:

- We present a systematic detector architecture, SD-Net, that addresses the challenges in aerial image detection by introducing adaptive computation mechanisms at the stages of feature extraction, multi-scale fusion, and final refinement.
- We design the SC2f block, which introduces a weight generation mechanism based on spatial feature encoding to improve the information bottleneck associated with global pooling in some dynamic networks.
- We propose the S2B module, which establishes a cross-scale modulation relationship within the feature pyramid, utilizing context-rich features to guide the scaling of detail-carrying features for more effective feature enhancement.
- We introduce the HCB module as a feature refinement unit that adaptively fuses static structural and dynamic content-aware information via a gating mechanism.

2. Related Work

2.1. General Object Detectors

The development of modern object detectors has primarily evolved along three paradigms. Two-stage detectors, represented by Faster R-CNN [46], achieve high accuracy through a propose-then-classify pipeline. Single-stage detectors, such as the YOLO series [3, 20], attain higher computational efficiency by performing dense predictions directly. More recently, a third paradigm of query-based, end-to-end detectors has gained widespread attention. DETR [7] pioneered the use of the Transformer architecture, framing object detection as a set prediction problem. Its successors, such as DINO [64] and the real-time RT-DETR [66], have significantly improved convergence speed and performance by refining query initialization and attention mechanisms. Furthermore, an emerging research direction involves diffusion-based detectors, like DiffusionDet [11], which model the generation of bounding boxes as a denoising process from random noise to final predictions, showing potential in handling crowded scenes. Our work is built upon an efficient single-stage detector architecture, with a focus on enhancing its internal feature representation capabilities.

2.2. Object Detection in Aerial Images

To address the specific characteristics of aerial images, researchers have proposed improvements from multiple perspectives.

Multi-Scale Feature Representation: The vast scale variation of objects is a core challenge in aerial imagery. Feature Pyramid Networks (FPN) [37] serve as a foundational solution to this problem. To enhance FPN's fusion capabilities, PANet [38] augmented the top-down path of FPN with an additional bottom-up path. Recent works have explored more efficient cross-scale connections, such as the weighted feature fusion in BiFPN [48]. To make the fusion process more adaptive, some studies have begun to introduce attention gates to control the flow of features across scales [42], or to design dynamic FPN structures that can adjust their topology based on the input [56]. These methods aim for more powerful multi-scale features through more sophisticated and flexible structural designs.

Contextual Modeling and Geometric Learning: Enabling the model to understand the relationship between an object and its environment is an important research direction. Besides constructing specialized context modules [26], leveraging the self-attention mechanism in Transformers to capture long-range dependencies has become mainstream [65, 13]. Some works employ the Swin Transformer [40] as a backbone, utilizing its hierarchical windowed attention to efficiently model global information. To further address the drastic geometric variations common in aerial views, recent works have proposed sampling equivariant mechanisms [61] to improve feature alignment, or theoretically continuous representations [59] to resolve angle regression discontinuities. Moreover, to more explicitly learn inter-object relationships, some research utilizes Graph Neural

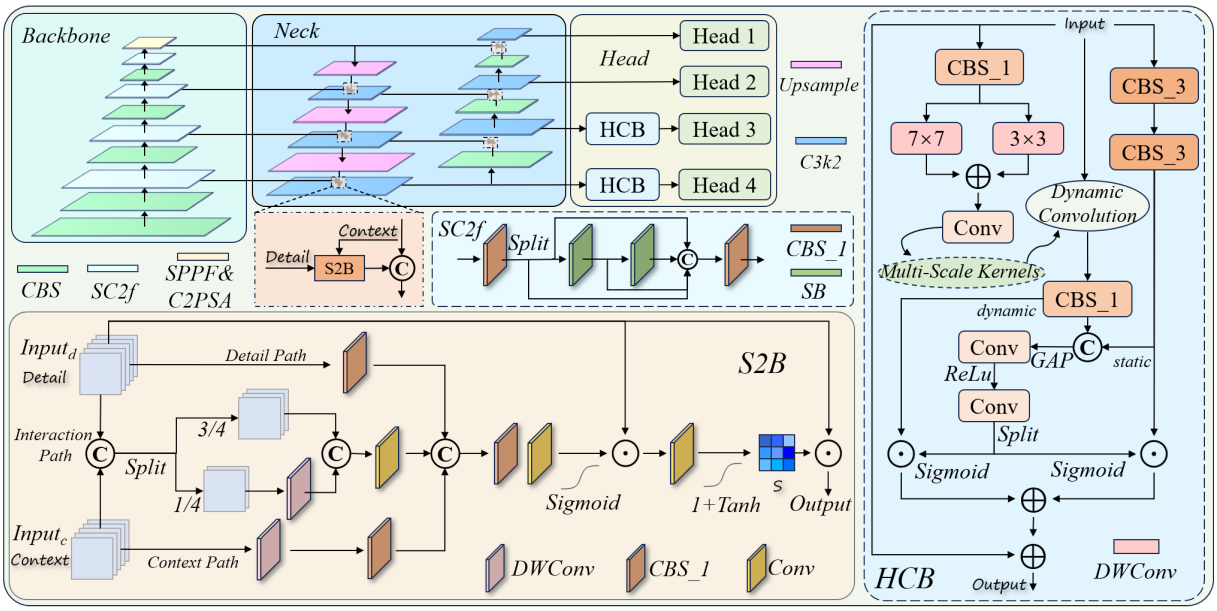

Figure 2. An overview of the proposed Synergistic and Dynamic Network (SD-Net) architecture. The network comprises a backbone, a neck, and four detection heads. The backbone is enhanced with our proposed SC2f blocks. In the neck, S2B modules are employed for multi-scale feature fusion. The detection heads for higher-resolution feature maps are augmented with our HCB modules for specialized feature refinement. The detailed structures of these three core components are illustrated in the corresponding insets. CBS denotes a standard convolutional block (Conv-BN-SiLU), while SB represents the Spatial-Context Bottleneck, the core of the SC2f block.

Networks (GNNs) [27] or pairwise relation modules [41] to model the spatial or semantic associations between objects, which is particularly effective for parsing dense and structured scenes.

Efficient Model Design and Data-Centric Approaches: Considering the resource constraints of UAV platforms, lightweight model design is a significant research avenue. Some efforts focus on designing efficient backbones [22] or leveraging knowledge distillation to transfer knowledge from large teacher models to smaller student models [32]. On the data front, given the scarcity of small object samples, data augmentation is key to improving generalization. In addition to methods like Copy-Paste [21], recent studies have started to explore the use of Generative Adversarial Networks (GANs) or diffusion models to synthesize high-quality, diverse small object samples to alleviate data scarcity and long-tail distribution issues [4].

2.3. Adaptive and Dynamic Computation

Adaptive computation mechanisms have been proposed to overcome the limitations of static computation in conventional convolutional networks. The core idea is to enable the network's computational process to adjust based on the input content. Dynamic convolution is one of the primary methods for this goal. CondConv [60] generates routing weights for each input to linearly combine multiple expert kernels. ODConv [31] extends this dynamic weighting concept to all dimensions of the kernel. These methods typically employ a module based on Global Average Pooling (GAP) to generate sample-specific weights, the efficacy of which has been discussed in tasks requiring fine-grained spatial information [23]. Beyond the kernel level, Mixture-of-Experts (MoE) models [18] implement adaptive computation at the module level, activating a subset of "expert" sub-networks for each input via a routing network. Recent research has further explored more efficient and stable routing strategies, such as introducing auxiliary losses to balance expert utilization [47, 29], which helps mitigate some of the training challenges associated with MoE models. These research efforts collectively explore how to make neural network computation more flexible and efficient.

3. Methods

This paper introduces the Synergistic and Dynamic Network (SD-Net), an object detector developed upon the YOLOv11 architecture with substantial modifications. SD-Net systematically enhances the network's capabilities in feature extraction and fusion by incorporating three purpose-built core components: SC2f, S2B, and HCB. This section first provides an overview of the SD-Net's overall architecture, followed by a detailed description of each component's design.

3.1. Overall Architecture

The overall architecture of the proposed SD-Net is illustrated in Figure 2, and it primarily consists of three parts: a backbone, a neck, and detection heads.The core design

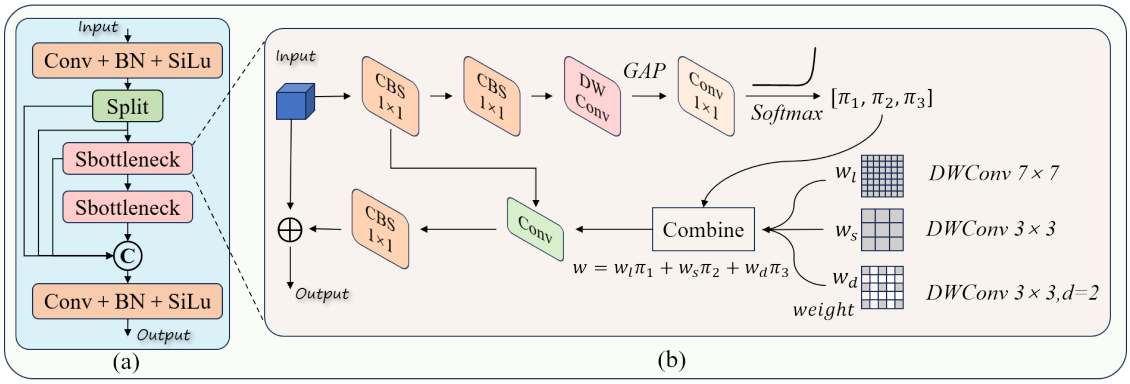

Figure 3. The detailed architecture of our proposed Spatial-Context C2f block. (a) The overall structure of the SC2f block. (b) The internal workflow of the SBottleneck.

philosophy is to introduce targeted, dynamic, and adaptive computation mechanisms at different stages of the network to enhance its feature representation capabilities.

In the backbone, we replace the original C3k2 modules with our proposed SC2f blocks. While retaining the CSP structure, the SC2f block substitutes the standard internal bottleneck with our SBottleneck, which is designed to introduce a content-adaptive computation mechanism and expand the receptive field. The backbone is responsible for extracting hierarchical features from the input image and passing feature maps from different levels to the neck.

The neck network inherits the bidirectional feature pyramid structure from YOLOv11 for effective multi-scale feature interaction. Its core modification is the introduction of the S2B to enhance cross-scale feature interaction prior to the standard concatenation and fusion. The function of the S2B is to mediate the interaction between two spatially-aligned input streams with different feature properties: one carrying finer-grained spatial details, and the other providing richer semantic context. The module enhances cross-scale feature interaction by performing a context-based modulation of the detail features.

The feature streams from the neck are finally passed to the detection heads for prediction. Our detection heads employ an asymmetric design. For the prediction branches that operate on lower-resolution feature maps (corresponding to medium-to-large objects), we retain the standard head structure. Conversely, for the branches that operate on higher-resolution feature maps (corresponding to small objects), the features are first refined by a HCB before entering the final prediction layer. The HCB dynamically fuses features from different computational paradigms via a gating mechanism, aiming to enhance the feature discrimination for small-sized objects.

3.2. Spatial-Context C2f Block

To introduce content-adaptiveness and expand the receptive field during the fundamental feature extraction stage in the backbone, we designed the Spatial-Context C2f (SC2f) block to replace the C3k2 modules in the YOLOv11 backbone. The overall structure of the SC2f block is depicted in Figure 3(a). It follows the Cross Stage Partial (CSP) design of the C2f module, which aims to achieve a rich gradient flow. Specifically, an input feature map is split into two parts: one part forms the main branch that sequentially passes through two of our designed SBottleneck (SB) modules, while the other part, along with the output of each stage in the main branch, serves as a skip connection that is fed into the final concatenation layer for fusion.

The functionality of the SC2f block is centered on its fundamental computational unit, the SBottleneck, whose detailed architecture is illustrated in Figure 3(b). The SBottleneck is designed to dynamically synthesize a customized convolutional kernel based on the input features. Upon receiving an input feature map $\mathbf{X} \in \mathbb{R}^{B\times C\times H\times W}$, the module first transforms it via a 1×1 CBS layer to produce an intermediate feature map $\mathbf{X}'$. This intermediate feature $\mathbf{X}'$ is then fed into a weight generation branch to compute a set of expert attention weights $\pi \in \mathbb{R}^{B\times K}$, where K is the number of experts. This process can be formulated as:

$$\pi = \text{Softmax}(\mathcal{F}_{wg}(\mathbf{X}')) \tag{1}$$

where $\mathcal{F}_{wg}(\cdot)$ represents the weight generation function. For an input $\mathbf{Z}$, its computation is defined as a sequence of operations:

$$\mathbf{Z}_1 = \text{CBS}_{1\times1}(\mathbf{Z}) \tag{2}$$
$$\mathbf{Z}_2 = \text{DWConv}_{7\times7}(\mathbf{Z}_1) \tag{3}$$
$$\mathbf{Z}_3 = \text{GAP}(\mathbf{Z}_2) \tag{4}$$
$$\mathcal{F}_{wg}(\mathbf{Z}) = \text{Conv}_{1\times1}(\mathbf{Z}_3) \tag{5}$$

Unlike methods that directly compress spatial dimensions with GAP, we introduce a large-kernel DWConv (Eq. 3) before the pooling operation (Eq. 4). This DWConv acts as a spatial encoder, enabling the weight generation process to be aware of the spatial layout of features, thereby avoiding the decision-making blindness caused by premature information compression.

Concurrently, we employ $K = 3$ parallel expert kernels with different receptive fields: a large 7×7 kernel $\mathbf{w}_l$, a standard 3×3 kernel $\mathbf{w}_s$, and a 3×3 dilated kernel $\mathbf{w}_d$ with a dilation rate of 2. All expert kernels are implemented as depth-wise convolutions. The final dynamic kernel $\mathbf{w}^{(b)}$ is synthesized for each sample b in the batch by a linear combination of these three expert kernels with the corresponding attention weights $\pi^{(b)}$:

$$\mathbf{w}^{(b)} = \sum_{k \in \{l,s,d\}} \pi_k^{(b)} \mathbf{w}_k, \quad \forall b \in \{1, ..., B\} \tag{6}$$

where $\pi_k^{(b)}$ is the attention weight for the k-th expert of the b-th sample. The synthesized dynamic kernel $\mathbf{w}^{(b)}$ is then applied to the corresponding feature map $\mathbf{X}'^{(b)}$. The output of this dynamic convolution operation passes through a final 1×1 CBS layer for channel fusion and is then added to the original input $\mathbf{X}$ via a residual connection to produce the final output of the SBottleneck.

3.3. Synergistic Scaling Block

In a feature pyramid network, the effective fusion of features from different levels is crucial for generating scale-robust representations. Conventional fusion methods, such as concatenation or addition, treat all features equally and lack a mechanism to discern feature importance. To achieve a more intelligent cross-scale feature interaction, we designed the Synergistic Scaling Block (S2B). The core idea of S2B is not to directly merge features, but to establish a cross-scale modulation relationship, where contextual features are used to guide the dynamic recalibration of detailed features.

The S2B is a dual-input module, as illustrated in Figure 2. It accepts two spatially-aligned feature maps as input: a feature map $\mathbf{X}_d \in \mathbb{R}^{B \times C_d \times H \times W}$ containing fine-grained spatial details, and a context feature map $\mathbf{X}_c \in \mathbb{R}^{B \times C_c \times H \times W}$ with richer semantic information. The objective of the module is to produce an enhanced detail feature map $\mathbf{X}'_d$, modulated by $\mathbf{X}_c$.

To achieve this, the internal architecture of S2B is designed with three parallel paths to fully exploit the synergistic information between the two inputs:

- Interaction Path: We concatenate $\mathbf{X}_d$ and $\mathbf{X}_c$ along the channel dimension and process them with a Partial Convolution (P-Conv) [8]. P-Conv applies a convolution to only a fraction of the channels (set to 1/4 in this work) while leaving the rest unchanged, aiming to efficiently learn the interaction patterns between the two feature streams at a low computational cost.
- Context Path: The context feature $\mathbf{X}_c$ independently passes through a large-kernel Depth-Wise Convolution to further enhance its long-range spatial context information.
- Detail Path: The detail feature $\mathbf{X}_d$ independently passes through a 1×1 convolution to preserve its original fine-grained structure.

The output features from the three paths are concatenated and fused through a series of convolutional layers. The fused feature map is then passed through a Sigmoid function to generate an attention gate $\mathbf{g}$ with values in the range $(0, 1)$. This gate is first multiplied with the original detail feature $\mathbf{X}_d$ for initial detail selection. Subsequently, we employ a $1 + \text{Tanh}$ activation function to transform the gated features into a scaling factor $\mathbf{s}$ with a range of $(0, 2)$. This scaling factor is finally applied to the original detail feature $\mathbf{X}_d$ via element-wise multiplication to produce the final output. The entire modulation process can be summarized as:

$$\mathbf{s} = 1 + \text{Tanh}(\mathcal{F}_{fuse}(\mathbf{g} \odot \mathbf{X}_d)) \tag{7}$$

$$\mathbf{X}'_d = \mathbf{s} \odot \mathbf{X}_d \tag{8}$$

where $\mathcal{F}_{fuse}$ represents the three-path fusion and gate generation network, and $\odot$ denotes element-wise multiplication. In this way, S2B leverages contextual information to generate a dynamic scaling coefficient for each spatial location of the detail feature map, thereby achieving enhancement of effective features and suppression of potential noise.

3.4. Asymmetric Detection Head

The design of the detection head is critical for the detection performance across different object scales. A standard detection head applies the same processing paradigm to all feature map scales, which may not be optimal for handling the extreme scale variations in aerial imagery. To more granularly process features of different scales, and particularly to enhance the discrimination of small objects, we modified the detection head of the baseline model, YOLOv11, to form an asymmetric structure. Our modifications comprise two aspects: extending the prediction scales and introducing a specialized feature refinement module.

3.4.1 P2-Level Prediction for Small Objects

In a Feature Pyramid Network, higher-resolution feature maps (i.e., feature levels with smaller strides) preserve finer spatial details, which are crucial for the precise localization of small objects. The highest-resolution prediction layer in the baseline model is P3 (feature map size of 80×80 for a 640×640 input). However, from a UAV perspective, a significant number of objects are extremely small. To better capture these objects, we incorporate an additional upsampling path in the neck to generate a higher-resolution P2-level feature map (feature map size of 160×160). Correspondingly, we add a new detection head (Head 4) dedicated to making predictions on this P2 level. The introduc-

tion of P2-level prediction enables our network to localize and classify minuscule objects at a finer granularity.

3.4.2 Hybrid Convolution Block (HCB)

To further enhance the discriminative power of the features, we designed the HCB as a powerful feature refinement unit. The core idea of the HCB is to dynamically fuse two complementary feature representations via a gating mechanism: static structural features extracted by standard convolutions, and content-adaptive features generated by a dynamic convolution. The detailed structure is illustrated in Figure 2.

The HCB takes an input feature map $X \in \mathbb{R}^{B\times C\times H\times W}$ and processes it through two parallel paths:

Static Path: This path is designed to extract robust, content-invariant structural features. The input feature X is first processed by two consecutive CBS_3 modules. Each CBS_3 module consists of a 3×3 convolution, a Batch Normalization layer, and a SiLU activation function. The output of this path is denoted as the static feature F_{static}.

Dynamic Path: In contrast to the static path, this path aims to generate instance-specific, adaptive features based on the content of the input. The implementation of this path involves two key parts: dynamic kernel generation and dynamic feature computation.

- Dynamic Kernel Generation: The input feature X is first passed through a CBS_1 module (1×1 Conv, BN, SiLU) for channel reduction. Subsequently, the reduced-channel feature is processed in parallel by two branches with different receptive fields: a 3×3 Depthwise Convolution (DWConv) and a 7×7 DWConv. The outputs of these two branches are summed and then passed through a point-wise convolution to finally generate the kernel weights, $K_{dynamic}$, for the dynamic convolution.

- Dynamic Feature Computation: We employ an efficient implementation based on the 'unfold' operation. The input feature X is first unfolded into a matrix of local blocks with a kernel size of 3×3, a stride of 1, and a padding of 1. This matrix is then element-wise multiplied with the generated dynamic kernels $K_{dynamic}$, and the result is summed over the local block dimension. Finally, the feature is reshaped back to its original spatial dimensions to obtain the dynamic feature $F_{dynamic}$.

To intelligently fuse these two feature streams, we introduce a gating mechanism. The static feature F_{static} and the dynamic feature F_{dynamic} are first concatenated along the channel dimension and then fed into a lightweight weight generation network. This network, composed of a Global Average Pooling (GAP) layer and two consecutive convolutional layers, generates two gate weights, g_{static} and g_{dynamic}, corresponding to the static and dynamic paths, respectively. The two weighted feature streams are first summed to produce a fused feature map, F_{fused}. The final output feature F_{out} is then obtained by adding this result to the original input X via a residual connection:

$$F_{\text{fused}} = \sigma(g_{\text{static}}) \odot F_{\text{static}} + \sigma(g_{\text{dynamic}}) \odot F_{\text{dynamic}} \quad (9)$$

$$F_{\text{out}} = X + F_{\text{fused}} \quad (10)$$

where σ denotes the Sigmoid function and $\odot$ represents element-wise multiplication. Through this mechanism, the HCB can adaptively determine the contribution of each path based on the fused features, achieving a refined feature representation.

3.4.3 Asymmetric Deployment Strategy

A key design choice when integrating the HCB is how to deploy the module across the prediction heads. A straightforward approach is to apply the HCB uniformly to all four prediction branches, a method we term a symmetric deployment. However, our empirical studies indicated that this symmetric deployment strategy was not optimal, and even yielded slightly lower detection accuracy compared to our proposed asymmetric strategy. We attribute this phenomenon to the intrinsic differences in the properties of feature maps at varying scales.

Specifically, higher-resolution feature maps (P2, P3) preserve rich spatial details but are relatively weak in semantic information. This makes them susceptible to interference from background noise and neighboring objects when processing feature-sparse small objects. Consequently, they benefit the most from the fine-grained, content-adaptive refinement provided by the HCB. Conversely, lower-resolution feature maps (P4, P5) possess stronger semantic information, and their corresponding medium-to-large objects inherently have more robust features. Applying an additional complex, dynamic transformation to these already abstract and semantically distinct features could introduce unnecessary computational redundancy and potentially disrupt the well-established stable representations.

Based on this insight, we propose an asymmetric deployment strategy. We insert the HCB for feature refinement only in the prediction branches that process the two highest-resolution feature maps, P2 and P3, just before the final prediction layer. For the P4 and P5 branches, we retain their original, more direct prediction paths. The superiority of this strategy lies in its provision of a differentiated and more targeted processing paradigm for features of different scales. It not only concentrates the powerful feature refinement capability of the HCB precisely on the most critical task of small object detection but also avoids the overprocessing of features for medium-to-large objects. Ulti-

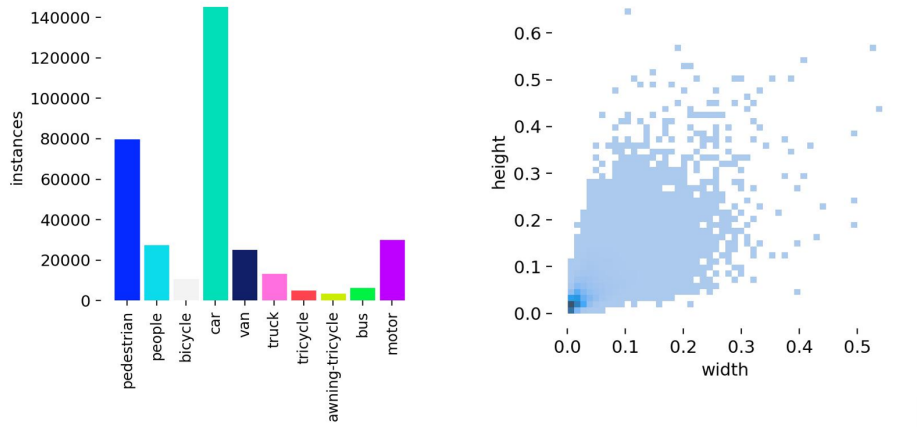

Figure 4. Object count and size distribution of the VisDrone-2019 dataset.

mately, this asymmetric design achieves a better overall accuracy than the symmetric deployment strategy while also keeping the additional computational cost within a reasonable range.

4. Experiments and Analysis

In this section, we conduct a series of experiments to evaluate the performance of our proposed Synergistic and Dynamic Network. We first introduce the experimental setup. Subsequently, we quantitatively compare SD-Net against several state-of-the-art object detection methods on two challenging aerial imagery datasets and provide a qualitative analysis. Finally, we validate the effectiveness of each of our designed core components through ablation studies.

4.1. Experimental Setup

4.1.1 Datasets

Our experiments are conducted on two public, representative, and large-scale aerial imagery datasets: VisDrone2019-DET and TinyPerson.

VisDrone2019-DET [15] is a benchmark dataset captured by UAVs under various weather and lighting conditions, focusing on object detection in aerial images. The dataset is annotated with ten object categories: pedestrian, people, bicycle, car, van, truck, tricycle, awning-tricycle, bus, and motor. It is divided into three subsets: 6,471 images for training, 548 for validation, and 1,610 for testing. As illustrated in Figure 4, this dataset presents two significant challenges. First, the class distribution is highly imbalanced. Second, the objects are predominantly small, with the normalized width and height of most instances being concentrated below 0.1, making it an ideal benchmark for evaluating multi-class small object detection performance in aerial scenes.

TinyPerson [62] is a dataset specifically constructed for the frontier challenge of extremely small person detection in the wild. It contains 1,610 images collected from the internet, with over 72,663 meticulously annotated person instances. The dataset is divided into 795 images for training and 815 for testing. The uniqueness of this dataset lies in its explicit focus on the "tiny" scale: the majority of person instances are less than 20 pixels in height, a definition far stricter than that for small objects in general-purpose benchmarks like COCO. These images encompass a wide variety of real-world scenarios, such as sea surfaces (people on boats), large crowds, streets, and sports events, where the tiny persons are often subject to severe occlusion, motion blur, and extreme lighting variations. Consequently, TinyPerson provides a highly challenging testbed for evaluating a model's robustness in handling severe feature sparsity caused by drastic scale reduction. We use VisDrone and TinyPerson in conjunction to assess our model's generalization capability in multi-class aerial scenarios and single-class extreme-scale scenarios, respectively.

4.1.2 Evaluation Metrics

To evaluate the model's performance, we employ a suite of metrics, including mean Average Precision (mAP), the number of parameters (Params), and GFLOPs.

We report three primary AP metrics: mAP_{50}, mAP_{75}, and $mAP_{.5:.95}$. The mAP_{50} is calculated at an Intersection over Union (IoU) threshold of 0.5 and reflects the model's overall detection capability across different objects. The mAP_{75} is calculated at a stricter IoU threshold of 0.75, placing a higher demand on the model's localization accuracy. The $mAP_{.5:.95}$, averaged over IoU thresholds from 0.5 to 0.95, provides a comprehensive measure of the model's bounding box regression capability. The number of parameters and GFLOPs are used to evaluate model complexity.

Table 1. Performance comparison with existing advanced methods on the **VisDrone2019-DET** dataset. The best and second-best results in each accuracy column are highlighted in **bold** and with an underline, respectively.

Method	Year	Params (M)	FLOPs (G)	mAP_{50}	mAP_{75}	$mAP_{.5:.95}$
YOLOv3 [45]	2018	103.7	282.3	45.0	28.4	27.7
Libra RCNN (ResNet101) [44]	2019	60.4	294.6	42.4	26.4	-
PISA (ResNet101) [6]	2020	60.2	293.5	44.2	27.9	-
YOLOv3-tiny [1]	2020	12.1	18.9	23.5	12.9	13.1
GFLv2 (ResNet101) [35]	2021	51.1	292.2	43.8	28.3	-
YOLOv6-s [30]	2022	16.2	43.7	36.0	21.8	21.5
YOLOv5-s [50]	2023	9.1	23.8	37.8	22.9	22.5
YOLOv8-s [24]	2023	11.1	28.5	38.3	23.3	22.8
CEASC (ResNet18) [14]	2023	19.3	-	44.5	26.4	-
SR-YOLOv8n [57]	2023	2.84	-	37.99	-	22.54
SR-YOLOv8 [57]	2023	7.61	-	41.6	-	23.9
YOLOv9-s [55]	2024	7.1	26.7	39.0	23.6	23.3
YOLOv10-s [53]	2024	8.0	24.5	38.8	23.5	23.2
YOLOv11-n [25]	2024	2.6	6.3	32.2	18.9	18.8
YOLOv11-s [25]	2024	9.4	21.3	37.9	23.4	22.9
YOLOv11-x [25]	2024	56.9	196.0	47.1	29.8	**29.3**
DTSSNet [10]	2024	10.1	-	39.9	25.2	-
DTSSNet* [10]	2024	10.1	-	41.1	26.9	-
SOD-UAV [36]	2024	32.2	126.2	45.7	26.8	-
TA-YOLO-n [34]	2024	3.8	14.1	40.1	-	24.1
TA-YOLO-s [34]	2024	13.9	43.3	45.4	-	27.7
TA-YOLO-o [34]	2024	21.4	64.6	<u>46.5</u>	-	<u>28.6</u>
RT-DETR-r18 [66]	2024	19.9	57.0	44.0	26.4	26.2
DMR-RTDETR [39]	2025	14.8	77.9	46.0	<u>28.7</u>	27.9
YOLOv12-n [49]	2025	2.5	6.2	30.6	17.3	-
YOLOv12-s [49]	2025	9.1	19.7	37.4	22.1	-
MSSEFPN + EFL (ResNet50) [28]	2025	-	-	40.2	27.9	-
MSSEFPN + EFL (ResNet101) [28]	2025	-	-	40.4	27.8	-
GFL+CFPT (ResNet18) [17]	2025	19.4	163.2	46.1	26.8	-
EDPDet-S [63]	2025	8.6	22.4	41.6	25.2	24.9
EViT-Net [16]	2025	4.1	18.9	39.1	-	22.3
LSOD-YOLO [9]	2025	3.8	33.9	37.0	-	-
DASSF [33]	2025	-	-	42.1	-	25.2
LUDY-N [52]	2025	2.81	-	35.2	-	-
LUDY-S [52]	2025	10.34	-	41.7	-	-
SD-Net-n	-	3.96	18.5	40.4	24.9	24.4
SD-Net-s	-	14.2	61.2	**47.6**	**30.0**	**29.3**

4.1.3 Implementation Details

Our proposed SD-Net is developed based on YOLOv11 [25]. We construct two versions, SD-Net-s and SD-Net-n, which correspond to the parameter scales of YOLOv11s and YOLOv11n, respectively.

Hyperparameters: All models are trained for 200 epochs with a batch size of 4. We use the SGD optimizer with an initial learning rate of 0.01, a weight decay of 0.0005, and a momentum of 0.937. Regarding the loss function, we directly adopt the default configuration of the YOLOv11 framework, which is applied uniformly across all detection heads, including the newly added P2 head. The input image size is uniformly resized to 640×640. Data augmentation strategies, including Mosaic, are employed during training and are turned off for the final 10 epochs.

Environment: All experiments are conducted on a single NVIDIA 4090 GPU. The specific software environment is as follows: Ubuntu 22.04, Python 3.10, and PyTorch 2.1.2.

4.2. Comparison with existing advanced methods

4.2.1 Results on VisDrone

To validate the effectiveness of our proposed SD-Net, we conduct extensive quantitative comparisons against a wide range of existing advanced object detection methods on the VisDrone2019-DET dataset. The results, encompassing multiple YOLO series baselines from YOLOv3 to YOLOv11, as well as various other two-stage and single-stage detectors, are presented in Table 1.

The experimental results in Table 1 indicate that our proposed SD-Net achieves a favorable balance between accuracy and model complexity.

First, compared to our baseline model, YOLOv11, SD-Net shows a substantial performance improvement across all primary metrics. Specifically, our SD-Net-s, with a moderate increase in parameters and computational cost, improves the mAP_{50} from 37.9% to 47.6% (+9.7%), the mAP_{75} from 23.4% to 30.0% (+6.6%), and the $mAP_{.5:.95}$ from 22.9% to 29.3% (+6.4%). Similarly, SD-Net-n also obtains comprehensive performance gains over YOLOv11-n. This clearly demonstrates the effectiveness of our proposed SC2f, S2B, and HCB modules in enhancing the baseline model.

Second, SD-Net demonstrates its superiority when compared against other recent methods that are also improved based on the YOLO architecture. Taking TA-YOLO-s as an example, at a similar parameter scale, our SD-Net-s achieves improvements of 2.2 and 1.4 percentage points in mAP_{50} and $mAP_{.5:.95}$, respectively. In the comparison with LSOD-YOLO, our lightweight version, SD-Net-n, with a similar parameter count but significantly lower FLOPs, obtains a 3.4 percentage point higher mAP_{50}.

Finally, SD-Net also shows strong competitiveness against other advanced lightweight detectors. For instance, compared to LUDY-S, our SD-Net-s, with a moderate increase in parameters, achieves a significant 5.9 percentage point gain in mAP_{50}. A similar conclusion can be drawn from the comparison with EDPDet-S, where SD-Net-s takes a comprehensive lead across all AP metrics, most notably a 4.2 percentage point advantage in the $mAP_{.5:.95}$ metric, which demands higher localization accuracy. These comparisons collectively demonstrate that our proposed approach achieves a competitive balance between accuracy and computational efficiency.

These data suggest that our proposed SD-Net, by introducing adaptive computation mechanisms at the feature extraction and fusion stages, can effectively boost detection performance in complex aerial scenes without incurring excessive computational overhead.

Table 2. Performance comparison with existing advanced methods on the **TinyPerson** dataset. The best and second-best results are highlighted in **bold** and with an underline, respectively.

Method	Year	P	R	mAP_{50}	$mAP_{.5:.95}$
YOLOv7 [54]	2023	47.7	25.5	24.9	7.1
YOLOv8-s [24]	2023	40.0	20.3	19.3	7.6
PS-YOLO-n [19]	2023	33.4	19.6	17.9	6.4
PS-YOLO-s [19]	2023	33.1	18.7	17.7	6.4
SR-YOLOv8n [57]	2024	-	33.4	29.2	9.7
SR-YOLOv8 [57]	2024	-	32.8	30.3	9.2
YOLOv11-n [25]	2024	33.3	24.8	18.4	5.7
YOLOv11-s [25]	2024	40.3	29.0	24.8	7.6
SFFEF-YOLO [2]	2025	45.1	29.3	27.5	**11.3**
FO-YOLO [67]	2025	35.9	27.3	22.3	6.5
SD-Net-n	-	39.2	31.0	25.8	8.1
SD-Net-s	-	**48.5**	**33.6**	**32.1**	10.1

4.2.2 Results on TinyPerson

To further evaluate the performance of our model in handling extremely small-sized objects, we also conducted comparative experiments on the TinyPerson dataset. TinyPerson is a benchmark specifically designed for small-scale pedestrian detection, with object sizes considerably smaller than those in most general-purpose datasets. It therefore places higher demands on the feature discrimination capabilities of a detector. The experimental results are presented in Table 2.

As shown by the results in Table 2, our SD-Net-s achieves the best performance across most key metrics. Compared to the baseline model, YOLOv11-s, SD-Net-s obtains significant improvements of 8.2%, 4.6%, 7.3%, and 2.5% in Precision (P), Recall (R), mAP_{50}, and $mAP_{.5:.95}$, respectively. This result strongly validates the effectiveness of our designed modules in enhancing the model's ability to discriminate and localize tiny objects.

In comparison with other state-of-the-art methods, SD-Net-s also demonstrates excellent performance. Although SFFEF-YOLO achieves the highest score on the $mAP_{.5:.95}$ metric, our SD-Net-s surpasses the second-best result among all compared methods (30.3% from SR-YOLOv8) by a margin of 1.8 percentage points on the mAP_{50} metric, which reflects overall detection capability. Furthermore, our model achieves the best balance between precision and recall.

The strong performance on TinyPerson, a benchmark specifically focused on extremely small human figures, particularly validates our architectural design choices. Specifically, the strategy of adding a detection head on a higher-resolution feature map (P2) and deploying the HCB module on the corresponding branches for feature refinement is crucial for improving the detection performance on tiny objects.

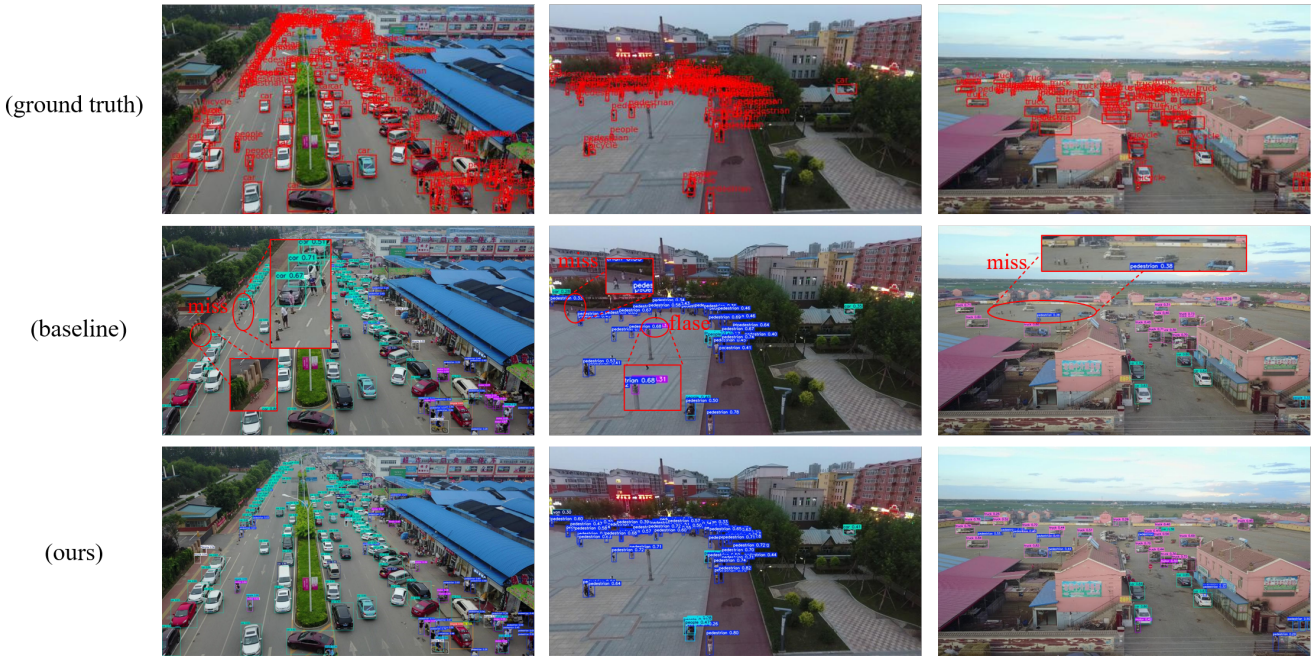

Figure 5. Qualitative detection results on challenging scenes from the VisDrone dataset. From top to bottom: Ground Truth, detection results of the baseline model YOLOv11s, and results of our proposed SD-Net-s.

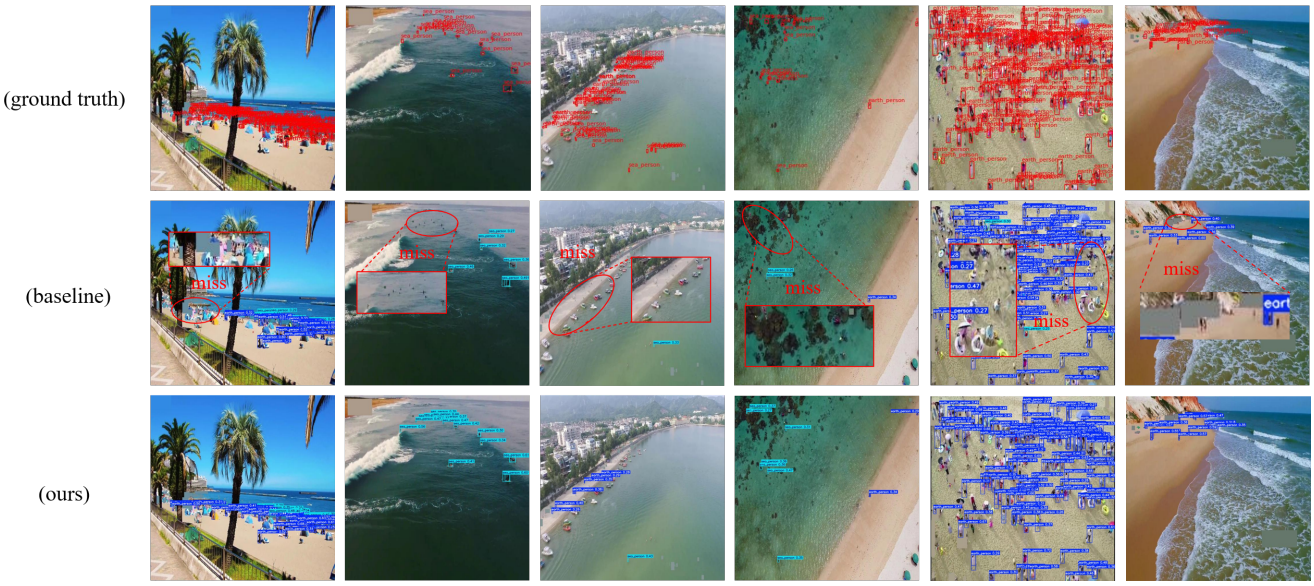

Figure 6. Qualitative detection results for tiny objects on the TinyPerson dataset. From top to bottom: Ground Truth, detection results of the baseline model YOLOv11s, and results of our proposed SD-Net-s.

4.3. Qualitative Analysis

To provide a more intuitive demonstration of the advantages of our proposed SD-Net in handling complex aerial scenes, we present a series of visual comparisons of detection results against the baseline model, YOLOv11.

Figure 5 illustrates the comparison results on three challenging scenes from the VisDrone dataset. The first row displays the ground-truth bounding boxes, highlighting the prevalent small-scale and dense distribution of objects in these scenes. The second row shows the detection results of the baseline model, YOLOv11s. It is evident that the baseline model suffers from a significant number of misses when faced with extremely dense crowds and traffic, and it struggles to distinguish heavily occluded objects. In contrast, as shown in the third row, our SD-Net-s is able to recall considerably more tiny objects in dense areas and successfully detects many individuals that were overlooked by the baseline. This suggests that the content-adaptive com-

Table 3. Ablation studies of the proposed components on the VisDrone dataset, based on YOLOv11-s.

Method	P2-Head	SC2f	HCB	S2B	Params (M)	FLOPs (G)	P	R	mAP_{50}	mAP_{75}	$mAP_{.5:.95}$
Baseline (YOLOv11-s)					9.4	21.3	49.1	37.1	37.9	23.4	22.9
	✓				9.6	28.6	52.7	40.8	42.9	26.5	25.9
	✓	✓			8.9	29.4	53.8	42.0	43.8	27.2	26.5
	✓		✓		9.8	34.0	53.5	42.5	43.6	27.5	26.7
	✓			✓	13.9	49.7	54.9	43.4	45.7	29.3	28.1
	✓	✓	✓		9.3	34.8	53.5	43.4	44.6	28.0	27.2
OURS (SD-Net-s)	✓	✓	✓	✓	14.2	61.2	56.5	45.9	47.6	30.0	29.3

Table 4. Ablation study on the deployment location of the HCB module. The experiments are based on a strong baseline model that includes P2 detection head, SC2f, and S2B. A ✓indicates that the HCB module is applied to the corresponding detection head (P2-P5).

P2	P3	P4	P5	Params (M)	FLOPs (G)	P	R	mAP_{50}	mAP_{75}	$mAP_{.5:.95}$
✓				13.9	58.5	**56.7**	44.2	46.6	29.5	28.7
✓	✓			14.2	61.2	56.5	**45.9**	**47.6**	**30.0**	**29.3**
✓	✓	✓		15.2	63.9	55.6	44.9	46.8	29.8	28.8
✓	✓	✓	✓	19.0	67.0	55.5	44.7	46.6	29.6	28.7

putation capability of our designed SC2f block, along with the cross-scale feature enhancement mechanism of the S2B block, effectively improves the model's ability to discriminate small objects in cluttered backgrounds.

Furthermore, we conducted visual comparisons on the TinyPerson dataset to further validate the model's performance on tiny object detection, with the results shown in Figure 6. A common characteristic of these scenes is the extremely small size of the objects, which often blend in with complex backgrounds such as ocean waves and sandy beaches. As can be seen from the results in the second row, the baseline model struggles to cope with these extreme cases, leading to severe misses. Our SD-Net (third row), however, successfully detects a large number of tiny pedestrians that were missed by the baseline. This performance improvement can be largely attributed to the additional detection head in our model that predicts on a higher-resolution feature map (P2), as well as the specialized refinement of high-resolution features by the HCB module. These qualitative results collectively demonstrate the effectiveness of our proposed method in enhancing the performance of aerial image object detection, particularly for tiny objects.

4.4. Ablation Studies

To meticulously validate the effectiveness of each proposed component in our SD-Net, we conduct a series of ablation studies on the VisDrone dataset. Our analysis begins with the YOLOv11-s baseline, and then we incrementally integrate our key modifications. The results are detailed in Table 3.

Our analysis begins with the YOLOv11-s baseline, which achieves an mAP_{50} of 37.9%. As a preliminary step, we introduce an additional P2 detection head to better accommodate the scale distribution of small objects in aerial images. As shown in the table, this structural modification provides a stronger starting point for the model, lifting the mAP_{50} to 42.9%.

Building upon this stronger baseline, we then evaluate the individual contributions of our three core modules. Among the three, the S2B block yields the largest performance increase, further improving the mAP_{50} by 2.8 percentage points. This demonstrates the superiority of our proposed cross-scale modulation mechanism in effectively fusing detail and context features. The SC2f block provides a 0.9 percentage point gain in mAP_{50}; notably, the introduction of SC2f even slightly reduces the model's total parameter count, highlighting its parameter-efficient design. The HCB module also contributes a solid 0.7 percentage point improvement, validating the effectiveness of feature refinement before the detection head.

Finally, when all components are integrated, our full model, SD-Net-s, achieves an mAP_{50} of 47.6%, the highest performance among all configurations. This result is not only higher than any single component's improvement but also surpasses the combinations of any two, demonstrating a positive synergistic effect among our proposed SC2f, S2B, and HCB modules. Overall, the results of the ablation studies strongly validate the soundness of our SD-Net's overall architectural design and the effectiveness of each individual component.

In addition to validating the individual contributions of each component, we further investigate the optimal deployment strategy for the HCB module. In our SD-Net-s architecture, HCBs are asymmetrically deployed on the P2 and P3 detection heads, which process high-resolution feature

maps. To verify the rationale behind this design choice, we conducted a series of experiments applying HCB to a varying number of detection heads. The results are presented in Table 4.

The results in Table 4 indicate that the model achieves the best performance across all primary AP metrics when the HCB modules are deployed on the P2 and P3 detection heads. Compared to using HCB on the P2 head alone, adding an HCB to the P3 head yields a 1.0 percentage point improvement in mAP_{50}, demonstrating the value of feature refinement on medium-scale features. However, as we continue to extend the HCB to the lower-resolution P4 and P5 heads, the model's performance does not continue to improve but instead shows a slight degradation, while the parameter count and computational cost steadily increase. This phenomenon suggests that for medium-to-large sized objects in aerial images, the features fused by the neck network are already sufficiently discriminative. The additional HCB refinement might introduce redundant computations and could even have a negative impact on the feature representation. Therefore, our final model adopts the asymmetric strategy of deploying HCBs on the P2 and P3 heads, as it strikes the optimal balance between accuracy and computational efficiency.

5. Conclusion

In this paper, we have addressed the challenges of object detection in aerial images, particularly the feature sparsity issue arising from small object sizes, by proposing the Synergistic and Dynamic Network (SD-Net). Our approach, built upon the YOLOv11 architecture, systematically enhances the model's feature representation capabilities by introducing purpose-built adaptive computation modules at three key stages of the network: feature extraction, multi-scale fusion, and the detection heads. Specifically, we designed the SC2f block to enrich the backbone's fundamental features through a context-aware dynamic convolution mechanism. We proposed the S2B block to facilitate more effective fusion of detail and semantic features in the neck via a cross-scale modulation relationship. We also introduced the HCB module to perform specialized refinement on high-resolution features intended for small object detection through a gating mechanism.

Extensive experiments on two challenging benchmarks, VisDrone and TinyPerson, have demonstrated the effectiveness of our proposed method. SD-Net achieves a competitive balance between accuracy and computational efficiency in comparison with a wide range of state-of-the-art detectors, including its baseline model, YOLOv11. Comprehensive ablation studies have also individually validated the rationale and effectiveness of each of our design choices.

Future work could explore extending the proposed dynamic computation mechanisms to a broader range of visual tasks. Furthermore, quantizing and optimizing these modules for specific hardware platforms to enable more efficient deployment on edge devices presents a promising direction for future research.

References

[1] P. Adarsh, P. Rathi, and M. Kumar. Yolo v3-tiny: Object detection and recognition using one stage improved model. In *2020 6th international conference on advanced computing and communication systems (ICACCS)*, pages 687–694. IEEE, 2020. 8

[2] C. Bai, K. Zhang, H. Jin, P. Qian, R. Zhai, and K. Lu. Sffef-yolo: Small object detection network based on fine-grained feature extraction and fusion for unmanned aerial images. *Image and Vision Computing*, 156:105469, 2025. 9

[3] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020. 2

[4] B. Bosquet, D. Cores, L. Seidenari, V. M. Brea, M. Mucientes, and A. Del Bimbo. A full data augmentation pipeline for small object detection based on generative adversarial networks. *Pattern Recognition*, 133:108998, 2023. 3

[5] E. V. Butilă and R. G. Boboc. Urban traffic monitoring and analysis using unmanned aerial vehicles (uavs): A systematic literature review. *Remote Sensing*, 14(3):620, 2022. 1

[6] Y. Cao, K. Chen, C. C. Loy, and D. Lin. Prime sample attention in object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11583–11591, 2020. 8

[7] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 2

[8] J. Chen, S.-h. Kao, H. He, W. Zhuo, S. Wen, C.-H. Lee, and S.-H. G. Chan. Run, don't walk: chasing higher flops for faster neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12021–12031, 2023. 5

[9] L. Chen, H. Deng, G. Liu, R. Law, D. Li, E. Q. Wu, and L. Zhu. Retinex-guided illumination recovery and progressive feature adaptation for real-world nighttime uav-based vehicle detection. *Expert Systems with Applications*, page 129476, 2025. 8

[10] L. Chen, C. Liu, W. Li, Q. Xu, and H. Deng. Dtssnet: Dynamic training sample selection network for uav object detection. *IEEE transactions on geoscience and remote sensing*, 62:1–16, 2024. 8

[11] S. Chen, P. Sun, Y. Song, and P. Luo. Diffusiondet: Diffusion model for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 19830–19843, 2023. 2

[12] G. Cheng, X. Yuan, X. Yao, K. Yan, Q. Zeng, X. Xie, and J. Han. Towards large-scale small object detection: Survey and benchmarks. *IEEE transactions on pattern analysis and machine intelligence*, 45(11):13467–13488, 2023. 1

[13] Z. Dai, B. Cai, Y. Lin, and J. Chen. Up-detr: Unsupervised pre-training for object detection with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1601–1610, 2021. 2

[14] B. Du, Y. Huang, J. Chen, and D. Huang. Adaptive sparse convolutional networks with global context enhancement for faster object detection on drone images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13435–13444, 2023. 8

[15] D. Du, P. Zhu, L. Wen, X. Bian, H. Lin, Q. Hu, T. Wang, J. He, L. Zhang, G. He, et al. Visdrone-det2019: The vision meets drone object detection in image challenge results. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 7

[16] Y. Du, L. Chen, and X. Hao. Evit-net: An efficient vision transformer-inspired network for enhanced multi-scale remote sensing image features. *Expert Systems with Applications*, page 129123, 2025. 8

[17] Z. Du, Z. Hu, G. Zhao, Y. Jin, and H. Ma. Cross-layer feature pyramid transformer for small object detection in aerial images. *IEEE Transactions on Geoscience and Remote Sensing*, 2025. 8

[18] W. Fedus, B. Zoph, and N. Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022. 3

[19] R. Fu, C. Chen, S. Yan, A. A. Heidari, X. Wang, J. Escorcia-Gutierrez, R. F. Mansour, and H. Chen. Gaussian similarity-based adaptive dynamic label assignment for tiny object detection. *Neurocomputing*, 543:126285, 2023. 9

[20] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021. 2

[21] G. Ghiasi, Y. Cui, A. Srinivas, R. Qian, T.-Y. Lin, E. D. Cubuk, Q. V. Le, and B. Zoph. Simple copy-paste is a strong data augmentation method for instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2918–2928, 2021. 3

[22] W. Guo, W. Li, Z. Li, W. Gong, J. Cui, and X. Wang. A slimmer network with polymorphic and group attention modules for more efficient object detection in aerial images. *Remote Sensing*, 12(22):3750, 2020. 3

[23] Q. Hou, D. Zhou, and J. Feng. Coordinate attention for efficient mobile network design. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13713–13722, 2021. 3

[24] G. Jocher, A. Chaurasia, and J. Qiu. Ultralytics YOLO, Jan. 2023. 8, 9

[25] G. Jocher and J. Qiu. Ultralytics yolo11, 2024. 8, 9

[26] S. D. Khan and S. Basalamah. Multi-scale and context-aware framework for flood segmentation in post-disaster high resolution aerial images. *Remote Sensing*, 15(8):2208, 2023. 2

[27] J. Kim, J. Baek, and S. J. Hwang. Object detection in aerial images with uncertainty-aware graph network. In *European Conference on Computer Vision*, pages 521–536. Springer, 2022. 3

[28] T. Kiobya, J. Zhou, and B. Maiseli. A multi-scale semantically enriched feature pyramid network with enhanced focal loss for small-object detection. *Knowledge-Based Systems*, 310:113003, 2025. 8

[29] D. Lepikhin, H. Lee, Y. Xu, D. Chen, O. Firat, Y. Huang, M. Krikun, N. Shazeer, and Z. Chen. Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668*, 2020. 3

[30] C. Li, L. Li, H. Jiang, K. Weng, Y. Geng, L. Li, Z. Ke, Q. Li, M. Cheng, W. Nie, et al. Yolov6: A single-stage object detection framework for industrial applications. *arXiv preprint arXiv:2209.02976*, 2022. 8

[31] C. Li, A. Zhou, and A. Yao. Omni-dimensional dynamic convolution. *arXiv preprint arXiv:2209.07947*, 2022. 3

[32] G. Li, W. Wang, X. Li, Z. Li, J. Yang, J. Dai, Y. Qiao, and S. Zhang. Distilling knowledge from large-scale image models for object detection. In *European Conference on Computer Vision*, pages 142–160. Springer, 2024. 3

[33] H. Li and H. Qu. Dassf: Dynamic-attention scale-sequence fusion for aerial object detection. In *International Conference on Computational Visual Media*, pages 212–227. Springer, 2025. 8

[34] M. Li, Y. Chen, T. Zhang, and W. Huang. Ta-yolo: a lightweight small object detection model based on multi-dimensional trans-attention module for remote sensing images. *Complex & Intelligent Systems*, 10(4):5459–5473, 2024. 8

[35] X. Li, W. Wang, X. Hu, J. Li, J. Tang, and J. Yang. Generalized focal loss v2: Learning reliable localization quality estimation for dense object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11632–11641, 2021. 8

[36] Y. Li, Y. Wang, Z. Ma, X. Wang, and Y. Tang. Sod-uav: Small object detection for unmanned aerial vehicle images via improved yolov7. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7610–7614. IEEE, 2024. 8

[37] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 1, 2

[38] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia. Path aggregation network for instance segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8759–8768, 2018. 2

[39] Y. Liu, Y. Wu, and Y. Yuan. Dmr-rtdetr: A multi-scale and context-aware measurement framework for quantitative characterization of photovoltaic module thermal anomalies with uncertainty analysis. *IEEE Transactions on Instrumentation and Measurement*, 2025. 8

[40] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 2

[41] Z. Liu, X. Zhang, C. Liu, H. Wang, C. Sun, B. Li, P. Huang, Q. Li, Y. Liu, H. Kuang, et al. Relationrs: Relationship rep-

resentation network for object detection in aerial images. *Remote Sensing*, 14(8):1862, 2022. 3

[42] H. Long, Y. Chung, Z. Liu, and S. Bu. Object detection in aerial images using feature fusion deep networks. *IEEE Access*, 7:30980–30990, 2019. 2

[43] S. A. H. Mohsan, N. Q. H. Othman, Y. Li, M. H. Alsharif, and M. A. Khan. Unmanned aerial vehicles (uavs): Practical aspects, applications, open challenges, security issues, and future trends. *Intelligent service robotics*, 16(1):109–137, 2023. 1

[44] J. Pang, K. Chen, J. Shi, H. Feng, W. Ouyang, and D. Lin. Libra r-cnn: Towards balanced learning for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 821–830, 2019. 8

[45] J. Redmon and A. Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 8

[46] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 2

[47] C. Riquelme, J. Puigcerver, B. Mustafa, M. Neumann, R. Jenatton, A. Susano Pinto, D. Keysers, and N. Houlsby. Scaling vision with sparse mixture of experts. *Advances in Neural Information Processing Systems*, 34:8583–8595, 2021. 3

[48] M. Tan, R. Pang, and Q. V. Le. Efficientdet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10781–10790, 2020. 2

[49] Y. Tian, Q. Ye, and D. Doermann. Yolov12: Attention-centric real-time object detectors. *arXiv preprint arXiv:2502.12524*, 2025. 8

[50] Ultralytics. Yolov5 by ultralytics. https://github.com/ultralytics/yolov5, 2023. Accessed: 2023-05-07. 8

[51] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1

[52] Z. Wan, S. Wang, W. Han, Y. Wang, X. Huang, X. Zhang, X. Chen, and Y. Chen. A systematic survey and meta-analysis of the segment anything model in remote sensing image processing: Challenges, advances, applications, and opportunities. *ISPRS Journal of Photogrammetry and Remote Sensing*, 229:436–466, 2025. 8

[53] A. Wang, H. Chen, L. Liu, K. Chen, Z. Lin, J. Han, et al. Yolov10: Real-time end-to-end object detection. *Advances in Neural Information Processing Systems*, 37:107984–108011, 2024. 8

[54] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7464–7475, 2023. 9

[55] C.-Y. Wang, I.-H. Yeh, and H.-Y. Mark Liao. Yolov9: Learning what you want to learn using programmable gradient information. In *European conference on computer vision*, pages 1–21. Springer, 2024. 8

[56] J. Wang, J. Yu, and Z. He. Arfp: A novel adaptive recursive feature pyramid for object detection in aerial images. *Applied Intelligence*, 52(11):12844–12859, 2022. 2

[57] L. Wang, Y. Shi, G. Mao, F. A. Dharejo, S. Javed, and M. Alathbah. Consumer-centric insights into resilient small object detection: Sciou loss and recursive transformer network. *IEEE Transactions on Consumer Electronics*, 70(1):2178–2187, 2023. 8, 9

[58] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018. 1

[59] Z. Xiao, G. Yang, X. Yang, T. Mu, J. Yan, and S. Hu. Theoretically achieving continuous representation of oriented bounding boxes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16912–16922, 2024. 2

[60] B. Yang, G. Bender, Q. V. Le, and J. Ngiam. Condconv: Conditionally parameterized convolutions for efficient inference. *Advances in neural information processing systems*, 32, 2019. 3

[61] G.-Y. Yang, X.-L. Li, Z.-K. Xiao, T.-J. Mu, R. R. Martin, and S.-M. Hu. Sampling equivariant self-attention networks for object detection in aerial images. *IEEE Transactions on Image Processing*, 32:6413–6425, 2023. 2

[62] X. Yu, Y. Gong, N. Jiang, Q. Ye, and Z. Han. Scale match for tiny person detection. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1257–1265, 2020. 7

[63] Y. Yu, W. Lyu, Q. Guo, Z. Deng, and W. Xu. Edpdet: Efficient dense pedestrian detectors with multi-scale feature extraction, enhancement and aggregation. *Expert Systems with Applications*, page 129933, 2025. 8

[64] H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, L. M. Ni, and H.-Y. Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022. 2

[65] Z. Zhang, X. Lu, G. Cao, Y. Yang, L. Jiao, and F. Liu. Vit-yolo: Transformer-based yolo for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2799–2808, 2021. 2

[66] Y. Zhao, W. Lv, S. Xu, J. Wei, G. Wang, Q. Dang, Y. Liu, and J. Chen. Detrs beat yolos on real-time object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16965–16974, 2024. 2, 8

[67] H. Zhou, W. Yin, K. Sun, T. Wu, and B. Deng. Fo-yolo for small object detection in drone aerial imagery: H. zhou et al. *The Journal of Supercomputing*, 81(12):1208, 2025. 9