remote sensing

MDPI

Article

SSABNet: Spatial-Semantic Aggregation and Balancing Network for Small-Target Detection in UAV Remote Sensing Images

Hongxing Zhang [1], Zhonghong Ou [1,*], Siyuan Yao [1], Shigeng Wang [1], Yang Guo [1] and Meina Song [1,2]

[1] School of Computer Science, Beijing University of Posts and Telecommunications, Beijing 100876, China; hongxingzhang@bupt.edu.cn (H.Z.); yaosiyuan@bupt.edu.cn (S.Y.); wangshigeng@bupt.edu.cn (S.W.); love_a_wild_horse@bupt.edu.cn (Y.G.); mnsong@bupt.edu.cn (M.S.)

[2] Department of Computer Science, School of Petroleum, China University of Petroleum-Beijing at Karamay, Karamay 834000, China

* Correspondence: zhonghong.ou@bupt.edu.cn

Highlights

What are the main findings?

- We propose SSABNet, a holistic framework tailored for UAV-based small-target detection that effectively tackles the dual challenges of irreversible spatial detail loss and scale-insensitive optimization bias.
- A Spatial-Semantic Aggregation (SSA) module is introduced to construct a high-fidelity restoration pathway, while a Scale-Aware Balancing Loss (SABL) employs a scale-dependent modulation mechanism to ensure stable training.

What are the implications of the main findings?

- Extensive experiments demonstrate state-of-the-art performance, achieving gains of 1.3% in overall Average Precision and 2.5% in Average Precision for small objects on the VisDrone benchmark, alongside an improvement of 0.5% in overall Average Precision and 16.9% in Average Precision for small objects on the UAVDT dataset.
- This study validates the effectiveness of jointly addressing representational fidelity and gradient balancing, offering a robust solution for precision aerial surveillance tasks.

Abstract

The precise localization of small objects in UAV-captured remote sensing imagery remains a formidable challenge due to their limited spatial support, coarse resolution, and severe background clutter. These factors often cause weak target cues to be progressively overwhelmed during deep feature extraction. Existing deep learning-based detectors typically suffer from two fundamental limitations: the irreversible loss of fine-grained spatial details during hierarchical feature fusion and the scale-insensitive optimization of conventional loss functions, which inadequately emphasize hard-to-detect small targets. To address these issues, we propose a novel Spatial-Semantic Aggregation and Balancing Network (SSABNet) tailored for UAV-based small-target detection. First, a Spatial-Semantic Aggregation (SSA) module is introduced to establish a high-fidelity restoration pathway that recovers fine-grained texture and boundary information from shallow layers. By employing content-aware operators, SSA effectively reconciles the structural discrepancy between spatial details and semantic abstractions, enabling precise cross-scale feature fusion while suppressing aliasing artifacts. Second, we design a Scale-Aware Balancing Loss (SABL) to mitigate the gradient instability and vanishing-gradient issues commonly encountered when optimizing non-overlapping small targets. SABL adopts a scale-dependent modulation mechanism that smoothly transitions from Wasserstein distance for distributional alignment of small objects to Euclidean distance for geometric refinement of larger targets,

Check for updates

Academic Editor: Francesco Nex

Received: 8 January 2026
Revised: 31 January 2026
Accepted: 6 February 2026
Published: 9 February 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

thereby ensuring stable and balanced optimization across object scales. Extensive experiments on the VisDrone benchmark demonstrate that SSABNet outperforms state-of-the-art detectors, achieving gains of 1.3% in overall AP and 2.5% in AP_s. Further evaluation on the UAVDT dataset confirms its strong generalization capability, yielding improvements of 0.5% in AP and 16.9% in AP_s. These results validate the effectiveness of jointly addressing feature representation and scale-aware optimization for UAV small-target detection.

Keywords: UAV remote sensing imagery; small-target detection; Spatial-Semantic Aggregation; Scale-Aware Balancing Loss

1. Introduction

With the rapid expansion of Unmanned Aerial Vehicle (UAV) applications in low-altitude airspace [1], UAVs have transcended their traditional roles to become critical infrastructure supporting a wide range of intelligent sensing and monitoring tasks [2]. Modern UAV platforms are increasingly deployed in complex autonomous missions, such as continuous visual tracking [3,4] and large-scale aerial surveillance, where reliable visual perception is indispensable. In these scenarios, accurate object detection [5] serves as a fundamental prerequisite for subsequent high-level decision-making and scene understanding.

However, compared with conventional ground-based vision benchmarks, UAV-based visual perception faces a unique set of challenges [6]. As illustrated in Figure 1, aerial imagery is characterized by nadir viewing perspectives, large-scale scene variations, and highly cluttered backgrounds, which significantly increase the difficulty of interpretation tasks such as building extraction and semantic segmentation. These challenges are further exacerbated in small-target detection, where objects of interest, such as vehicles and pedestrians, typically appear as diminutive instances occupying fewer than 32×32 pixels and lack discriminative geometric and texture cues.



Figure 1. Representative UAV remote sensing images from the VisDrone datasets. Objects of interest, such as vehicles and pedestrians, often occupy fewer than 32×32 pixels and are embedded in highly cluttered backgrounds. The combination of nadir viewing perspectives, dense object distributions, and limited spatial resolution highlights the inherent difficulty of small-target detection in UAV scenarios.

A primary impediment in parsing UAV-captured imagery stems from three fundamental challenges [6] that standard detection frameworks fail to reconcile. These challenges arise from the intrinsic conflict between the design principles of generic object detectors and the unique characteristics of aerial data.

The first challenge is the irreversible erosion of features during the downsampling process. To guarantee extensive ground coverage, UAVs often operate at high altitudes [7], inevitably reducing targets to pixel-limited instances [8]. Standard CNN architectures, such as the YOLO series [9], DINO [10], FCOS [11], and Faster R-CNN [12], rely on aggressive strided convolutions to expand the receptive field. While effective for general object detection, this mechanism is detrimental for UAV scenarios [13,14]. Specifically, for an object occupying fewer than 32×32 pixels, a typical $32\times$ downsampling factor effectively collapses its representation into a single sub-pixel feature point or eliminates it entirely. Once these high-frequency spatial details are discarded in the early stages, they become theoretically unrecoverable in deep semantic layers, rendering the target mathematically indistinguishable from the background.

The second challenge is the Spatial-Semantic misalignment. Although Feature Pyramid Networks and recent Transformer-based architectures [15] introduce mechanisms to fuse multi-scale features, a severe semantic gap persists. Shallow layers in UAV imagery are often replete with high-frequency background noise, such as complex vegetation textures or building edges, which lack semantic discriminability. Naively fusing these noisy shallow features with abstract deep features leads to semantic aliasing, where the weak signal of the small target is overwhelmed by background clutter during aggregation. Without explicit alignment and filtering, the network struggles to determine which features to trust, leading to false alarms in texture-rich regions.

The third challenge is the gradient dominance and instability inherent in the optimization process. There exists a severe scale imbalance in aerial datasets. Standard geometric metrics, including GIoU [16], DIoU, and CIoU [17], are hypersensitive for small objects, where a deviation of only one or two pixels can cause drastic loss fluctuations. Moreover, when predicted boxes fail to overlap with minute ground truth targets, the IoU-based gradients often vanish completely. Consequently, the backpropagation process becomes dominated by the stable and large-magnitude gradients of large objects [5]. This optimization bias causes the network to prioritize fitting easier, large instances while treating errors on hard-to-detect small targets as negligible noise, thereby stalling the learning progress for the minority class.

To address these multifaceted challenges, we propose the Spatial-Semantic Aggregation and Balancing Network (SSABNet), a holistic framework designed specifically for robust small-target detection in aerial imagery. By harmonizing feature representation and optimization objectives, our method systematically resolves the dilemmas of feature erosion and optimization bias.

The main contributions of this paper are summarized as follows:

- We propose a Spatial-Semantic Aggregation (SSA) module that constructs a high-fidelity restoration pathway utilizing high-resolution shallow features. By integrating content-aware reassembly with topology-aware feature injection, the module explicitly recovers fine-grained spatial textures lost during downsampling while maintaining semantic consistency to suppress aliasing artifacts.
- We develop a Scale-Aware Balancing Loss (SABL) that implements a scale-adaptive internal rectification strategy. It utilizes the Wasserstein distance for the distributional alignment of small objects and Euclidean metrics for the geometric refinement of larger targets, thereby resolving the gradient vanishing issue and ensuring balanced optimization across varying object scales.

- We achieve state-of-the-art (SOTA) performance on aerial benchmarks. Comprehensive evaluations on the VisDrone2019 and UAVDT datasets demonstrate that SSABNet significantly outperforms existing detectors, particularly in improving detection recall for pixel-limited targets within dense and complex scenarios.

The paper is structured as follows: Section 2 reviews related work in small-target detection and feature fusion techniques. Section 3 provides a detailed formulation of the proposed SSABNet, including the architecture of the SSA and the derivation of the SABL. Section 4 presents the experimental setup, comprehensive ablation studies, and comparative results. Section 5 analyzes the underlying mechanisms, computational trade-offs, generalization capability, and limitations of the framework. Finally, Section 6 summarizes the findings and outlines future research directions.

2. Related Work

2.1. Small-Target Detection in Aerial Imagery

Small-target detection has been extensively studied in the broader field of object detection [18]; however, methods developed for natural images or ground-based benchmarks often exhibit limited effectiveness when applied to UAV remote sensing scenarios. Unlike conventional datasets, UAV imagery is characterized by extreme scale imbalance, dense object distributions, and complex background clutter, which impose distinct requirements on both feature representation and optimization strategies. Consequently, recent research has increasingly focused on designing UAV-oriented detection frameworks that explicitly account for the unique challenges of aerial small-target perception.

Recent advances can be broadly categorized into two paradigms: CNN-based enhancements and Transformer-based architectures. Transformer-based models, such as the Hierarchical Graph Interaction Transformer [15], leverage global self-attention to capture long-range contextual dependencies, which can benefit small-target discrimination in cluttered scenes. However, their quadratic computational complexity and substantial memory overhead impose a prohibitive burden on UAV platforms with limited onboard resources. In contrast, CNN-based methods, such as ESOD [19], offer superior inference efficiency and deployment feasibility but often struggle to preserve fine-grained spatial details during successive downsampling operations. This trade-off highlights a fundamental tension between representational fidelity and computational efficiency in UAV small-target detection.

From the perspective of feature processing, existing approaches have explored different strategies to enhance small-target representations. For instance, methods such as [8] focus on removing redundant features to emphasize salient responses. However, for pixel-limited small targets in UAV imagery, the dominant challenge often lies in feature scarcity rather than redundancy. Excessive suppression may further weaken already fragile signals. This observation motivates approaches that emphasize accurate feature compensation and restoration, aiming to explicitly recover lost spatial details from shallow layers rather than filtering them out.

2.2. Feature Pyramid Networks and Sampling Strategies

Multi-scale feature fusion is pivotal for detecting objects with drastic scale variations. The classic Feature Pyramid Network (FPN) establishes a top-down pathway to propagate semantic information, but its rigid architectural design often suffers from semantic inconsistency across layers [20]. While recent variants such as AFPN [21] attempt to alleviate this issue through adaptive integration, they generally treat upsampling and downsampling operations as independent processes, overlooking their coupled influence on feature consistency.

Traditional interpolation methods, such as nearest neighbor upsampling, lack content awareness and often introduce feature misalignment and semantic ambiguity. To address this limitation, learnable sampling mechanisms have been proposed. CARAFE [22] introduces content-aware reassembly for upsampling, while DySample [23] adopts point-based sampling to improve efficiency and spatial alignment. Conversely, for downsampling, ScDown [24] proposes a decoupled strategy to preserve structural information during resolution reduction. Despite these advances, existing methods typically apply upsampling and downsampling operators in isolation, without explicitly modeling their internal consistency.

This lack of coordinated design often leads to information leakage or distortion during cross-scale feature transformation. Motivated by this observation, recent efforts have begun to explore more unified feature processing pipelines that jointly consider restoration and compression behaviors within multi-scale architectures.

2.3. Label Assignment and Loss Functions

The training stability of object detectors relies heavily on label assignment (LA) strategies and loss function design. Modern detectors have progressively transitioned from static IoU thresholding to dynamic assignment mechanisms. Methods such as OTA [25] and SimOTA [26] formulate label assignment as an optimal transport problem, allocating positive samples based on global cost matrices. However, these cost formulations are often dominated by large objects, causing small targets—which naturally exhibit lower IoU values—to be under-represented in the positive sample pool.

Regarding loss functions, widely adopted IoU-based metrics, including GIoU [16], DIoU, and CIoU [17], are highly sensitive to positional deviations for small objects. Even minor localization errors can lead to drastic IoU fluctuations. While these variants improve geometric robustness, they remain scale-agnostic in terms of optimization magnitude, failing to address the gradient imbalance problem where large objects dominate the back-propagation process [5]. As a result, the optimization objective becomes misaligned with the requirements of small-target detection.

To alleviate the limitations of geometric overlap metrics, distribution-based approaches have emerged as a promising alternative. Normalized Wasserstein Distance (NWD) [27] models bounding boxes as 2D Gaussian distributions and measures similarity via optimal transport theory, providing smoother gradients for non-overlapping tiny objects. However, existing methods typically combine NWD and IoU through linear weighting, implicitly treating distributional and geometric objectives as independent. Such additive formulations may introduce gradient conflicts, where competing optimization signals hinder stable convergence. This motivates the development of loss functions that intrinsically reconcile geometric precision and distributional alignment in a scale-aware manner.

3. Methodology

In this section, we present the proposed SSABNet, designed to address the intrinsic challenges of small-target detection in UAV-captured imagery. The overall framework enhances the conventional detection pipeline by integrating structural improvements for high-fidelity feature restoration and a tailored optimization strategy. Specifically, SSABNet introduces two key components: the SSA for recovering fine-grained details and ensuring feature consistency, and the SABL for rebalancing the gradient contribution of hard-to-detect small targets. These architectural components are not introduced as isolated enhancements; instead, FARUp and CPDown are specifically designed to address feature misalignment and structural degradation during scale transformation, while GAR and CGAR serve as lightweight activation refinements to stabilize weak signal propagation.

Together, they form a coherent representation-oriented solution, which is complemented by the optimization-oriented SABL described later.

3.1. Overall Architecture

The proposed SSABNet framework is designed to tackle the challenges of UAV small-target detection through a holistic approach. The overall network architecture is illustrated in Figure 2, while the associated joint optimization strategy is depicted in Figure 3.

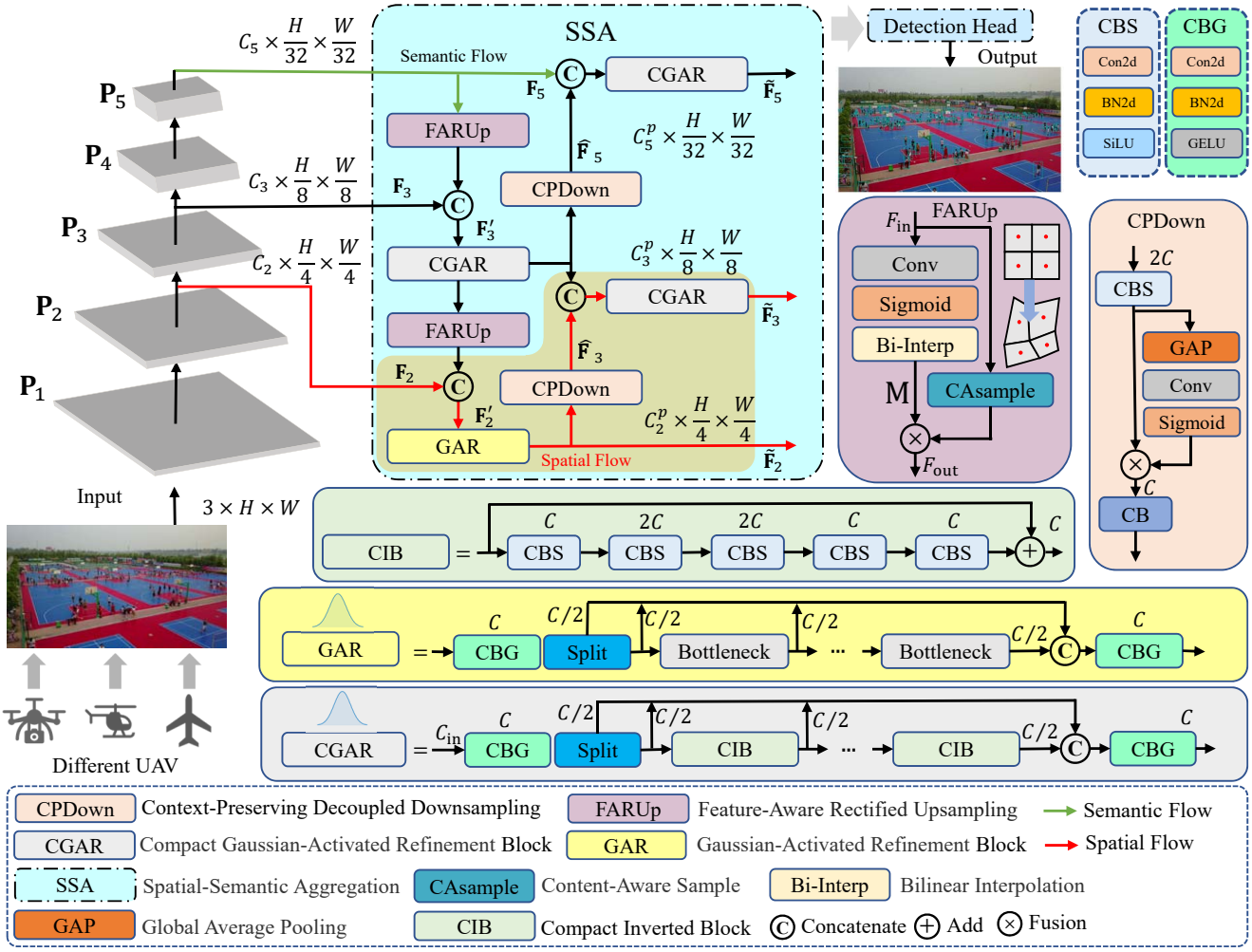


Figure 2. Schematic illustration of the proposed SSABNet architecture. The network integrates the SSA module to reconcile the representational divergence between shallow geometric details and deep abstract features. Note the topology-aware selective injection design: the Spatial Flow (Red) transmits $\mathbf{P}_2$ features only to $\mathbf{P}_3$ to avoid noise contamination, while the Semantic Flow (Green) ensures semantic consistency. FARUp and CPDown are employed for content-aware upsampling and downsampling, respectively, while GAR and CGAR blocks are utilized for refined feature activation.

To mitigate the irreversible loss of small object information during feature aggregation, we redesign the neck using the proposed SSA. Within SSA, we introduce Feature-Aware Rectified Upsampling (FARUp) and Context-Preserving Decoupled Downsampling (CPDown). Specifically, FARUp incorporates the proposed Content-Aware Sampler (CAsample), which adapts the dynamic sampling principles of [23] to strictly align low-level spatial details with high-level semantics. Meanwhile, CPDown utilizes the decoupled strategy to minimize structural information loss. These content-aware operators effectively suppress aliasing effects and preserve boundary information during resolution changes.

Additionally, we employ a Gaussian-Activated Refinement (GAR) block to enhance feature extraction stability. Unlike the conventional SiLU activation which may suppress weak feature signals in the negative regime, the integrated Gaussian Error Linear Units (GELU) provides a smoother probabilistic mapping. This design mitigates the gradient vanishing problem for pixel-limited targets, thereby enhancing the network's capability to capture fine-grained patterns without increasing architectural depth.

Furthermore, to balance efficiency and representation capability, we design the Compact Gaussian-Activated Refinement (CGAR) block. By integrating the Compact Inverted Block (CIB) structure into the GAR framework, CGAR significantly reduces channel redundancy and parameter overhead while maintaining the smoothness benefits of Gaussian activation.

Parallel to the architectural enhancements, we address the optimization bias through the proposed SABL. As illustrated in Figure 3, standard geometric metrics often fail on pixel-limited targets due to gradient vanishing. To overcome this, SABL introduces a scale-adaptive mechanism that dynamically modulates the regression objective: it prioritizes Wasserstein-based distributional alignment for small objects to ensure stable supervision, while seamlessly transitioning to Euclidean-based geometric refinement for larger targets. This dual-strategy ensures robust convergence across the full spectrum of object scales.

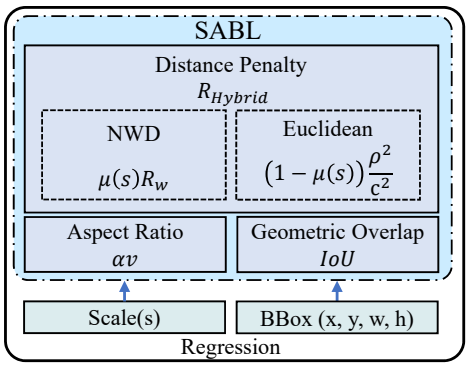


Figure 3. Schematic illustration of the proposed SABL. As the core regression objective of SSAB-Net, SABL implements a scale-adaptive internal rectification strategy. The mechanism seamlessly transitions from Wasserstein-based distributional alignment for pixel-limited targets (NWD) to Euclidean-based geometric refinement for larger objects via the modulation factor $\mu(s)$, ensuring balanced gradient optimization across varying object scales.

The core innovation lies in the synergistic effect of the SSA and the SABL, which we detail in the following subsections.

3.2. Spatial-Semantic Aggregation Module

Standard FPN-based architectures typically adopt a dense connection strategy, propagating low-level features to all upper layers. However, simply incorporating the high-resolution shallow layer ($\mathbf{P}_2$) into the feature pyramid is not straightforward. In the context of UAV imagery, although $\mathbf{P}_2$ contains rich spatial details, it often carries significant high-frequency noise derived from complex backgrounds, such as intricate vegetation textures and sharp building edges. Indiscriminately fusing these noisy features with deep semantic layers can introduce interference, potentially overwhelming the target signals and degrading detection performance.

To address this, we propose a topology-aware selective injection strategy within the SSA module. As illustrated in Figure 2, we decouple the feature flow into two streams:

- Spatial Flow (Red Path): The high-resolution features from $\mathbf{P}_2$ are exclusively injected into $\mathbf{P}_3$ after passing through the CPDown and GAR modules. This strict truncation prevents noise propagation to higher levels.

- Semantic Flow (Green Path): The semantic flow (Green Path) propagates high-level contextual cues from $\mathbf{P}_5$ to guide feature calibration at $\mathbf{P}_3$, while the influence on $\mathbf{P}_2$ is implicitly constrained through the intermediate aggregation at $\mathbf{P}_3$, thereby avoiding direct semantic contamination of the shallowest layer.

This design ensures that high-level layers remain semantically pure, while $\mathbf{P}_3$ serves as the critical aggregation node balancing spatial precision and semantic richness.

The SSA fuses features from adjacent scales to generate an enhanced, aligned representation. Let $\mathbf{F}_i$ denote the feature map at $\mathbf{P}_i$. The fusion process integrates high-level semantics with low-level details through dynamic aggregation.

Specifically, for the feature fusion at the $\mathbf{P}_3$ and $\mathbf{P}_2$ scales, the process employs our proposed FARUp operator:

$$\begin{aligned} \mathbf{F}'_3 &= \text{Concat}(\mathbf{F}_3, \mathcal{U}_{FA}(\mathbf{F}_5)) \\ \widetilde{\mathbf{F}}_2 &= \text{Concat}\big(\mathbf{F}_2, \mathcal{U}_{FA}(\mathbf{F}'_3)\big) \end{aligned} \tag{1}$$

where $\text{Concat}(\cdot)$ denotes channel concatenation, and $\mathcal{U}_{FA}(\cdot)$ represents the proposed FARUp operation.

Specifically, the FARUp module employs a dual-stage context-aware strategy, formally defined as follows:

Content-Aware Dynamic Sampling: Let $\mathbf{X} \in \mathbb{R}^{C\times H\times W}$ denote the input feature map. First, we utilize a dynamic sampling operator to perform content-aware upsampling. Unlike fixed interpolation, it generates point-wise offsets $\Delta\mathcal{P}$ derived from the input features to retrieve information from optimal locations:

$$\mathbf{X}_{dyn} = \text{Sampling}(\mathbf{X}, \Delta\mathcal{P}) \tag{2}$$

where $\mathbf{X}_{dyn} \in \mathbb{R}^{C\times sH\times sW}$ represents the initial upsampled features aligned with target structures.

Residual Rectification Mechanism: To suppress the background noise often exacerbated by upsampling, we introduce a rectification branch. We first generate a content-adaptive saliency mask $\mathbf{M}$ from the input features using a lightweight gating network, followed by bilinear alignment:

$$\mathbf{M} = \mathcal{U}_{bilinear}(\sigma(\mathcal{W}_{1\times 1} * \mathbf{X} + \mathbf{b})) \tag{3}$$

where $\sigma(\cdot)$ is the Sigmoid activation, $\mathcal{W}_{1\times 1}$ denotes the 1×1 convolution, and $\mathcal{U}_{bilinear}$ represents bilinear interpolation to match the scale of $\mathbf{X}_{dyn}$.

Finally, the rectified output $\mathbf{X}_{out}$ is obtained via a residual attention formulation:

$$\mathbf{X}_{out} = \mathbf{X}_{dyn} \odot (1 + \mathbf{M}) \tag{4}$$

where $\odot$ denotes element-wise multiplication. This formulation allows the network to adaptively emphasize target-relevant regions while preserving the base structural information, effectively filtering out aliasing artifacts.

This dynamic alignment is crucial for maintaining the structural integrity of small objects during the restoration of $\mathbf{P}_2$ features.

Subsequently, the restored features are propagated to other scales. For downsampling operations within the fusion block, we employ CPDown to minimize information loss:

$$\begin{aligned} \hat{\mathbf{F}}_3 &= \mathcal{D}_{CP}(\mathbf{P}'_2) \\ \widetilde{\mathbf{F}}_3 &= \text{Concat}\big(\mathbf{F}'_3, \hat{\mathbf{F}}_3\big) \end{aligned} \tag{5}$$

where $\mathcal{D}_{CP}(\cdot)$ denotes the proposed CPDown operation.

Specifically, unlike standard strided convolutions that simultaneously reduce resolution and mix channels (often causing structural destruction of small targets), CPDown adopts a decoupled strategy. It prioritizes the retention of high-frequency cues before semantic mapping, formulated as:

$$\mathcal{D}_{CP}(\mathbf{X}) = \mathcal{W}_{mix} * (\mathcal{P}_{context}(\mathbf{X})) \tag{6}$$

where $\mathcal{P}_{context}(\cdot)$ represents a structure-locking spatial reduction operator (e.g., maximizing local texture retention) that preserves the geometric boundaries of small objects, and $\mathcal{W}_{mix}$ denotes a point-wise convolution for inter-channel semantic integration.

This design ensures that both semantic context and spatial details are effectively compensated and preserved across all detection scales.

Finally, the feature fusion at $\widetilde{\mathbf{F}}_5$ is formulated as follows:

$$\begin{aligned} \hat{\mathbf{F}}_5 &= \mathcal{D}_{CP}(\mathbf{F}'_3) \\ \widetilde{\mathbf{F}}_5 &= \text{Concat}(\mathbf{F}_5, \hat{\mathbf{F}}_5) \end{aligned} \tag{7}$$

3.3. Scale-Aware Balancing Loss (SABL)

Unlike conventional loss fusion strategies that combine multiple objectives via external linear weighting, SABL adopts an internal rectification paradigm, where unstable components of the original loss are selectively replaced based on object scale, ensuring a single, non-conflicting optimization objective.

Complete Intersection over Union (CIoU) loss is the standard for bounding box regression due to its comprehensive consideration of overlap area, central point distance, and aspect ratio. Formally, for a predicted box b and a ground truth box b^{gt}, the CIoU loss [17] is defined as:

$$\mathcal{L}_{\text{CIoU}} = 1 - \text{IoU} + \frac{\rho^2(b, b^{gt})}{c^2} + \alpha v \tag{8}$$

where $\rho(\cdot)$ denotes the Euclidean distance between center points, c is the diagonal length of the smallest enclosing box, and the aspect ratio consistency terms α and v are preserved from CIoU to maintain morphological correctness:

$$v = \frac{4}{\pi^2}\left(\arctan\frac{w^{gt}}{h^{gt}} - \arctan\frac{w}{h}\right)^2, \quad \alpha = \frac{v}{(1-\text{IoU}) + v} \tag{9}$$

The term αv imposes a penalty on aspect ratio inconsistencies. While effective for general object detection, $\mathcal{L}_{\text{CIoU}}$ exhibits a inherent drawback when applied to UAV aerial imagery dominated by tiny objects. Specifically, the Euclidean penalty term $\frac{\rho^2}{c^2}$ suffers from pixel-level sensitivity. For a micro-scale target, the normalizing factor c^2 is inherently negligible. Consequently, a positional deviation of merely 1–2 pixels can cause the penalty term to fluctuate violently, resulting in gradient explosion and optimization instability. Furthermore, when the predicted box and ground truth are spatially disjoint (a common occurrence for small targets during early training), the IoU gradient vanishes, leaving the optimization solely dependent on this unstable Euclidean term.

To address these challenges, previous approaches often resorted to a linear combination of different loss functions. However, we argue that such additive strategies essentially treat geometric regression and distribution alignment as independent supervision signals. This leads to a Gradient Conflict: the unstable geometric gradients from the CIoU term are still propagated through the network, interfering with the distributional gradients from metrics like NWD.

In this work, we propose a fundamental internal rectification strategy. We posit that the optimization objective should dynamically transition based on the object scale:

1. Micro-Scale Regime (Distributional Alignment): For pixel-limited targets, geometric exactness is ill-posed due to quantization error. The optimization should prioritize distributional overlap using Wasserstein distance, which possesses infinite support and provides non-vanishing, smooth gradients.
2. Macro-Scale Regime (Geometric Refinement): As the object scale increases, distinctive geometric features become available. The optimization should seamlessly shift towards Euclidean regression to ensure precise boundary localization.

Based on this motivation, our SABL constructs a unified gradient field that rectifies the intrinsic defects of CIoU by replacing its unstable penalty term with a scale-modulated hybrid metric.

The core of SABL is the proposed hybrid penalty term, which integrates Wasserstein distance into the geometric framework via a non-linear modulation.

Wasserstein Penalty Modeling. We first model the bounding boxes as 2D Gaussian distributions. For a bounding box $B = (cx, cy, w, h)$, its corresponding Gaussian distribution $\mathcal{N}(\mathbf{m}, \mathbf{\Sigma})$ [28] can be modeled as:

$$\mathbf{m} = \begin{bmatrix} cx \\ cy \end{bmatrix}, \quad \mathbf{\Sigma} = \begin{bmatrix} (w/2)^2 & 0 \\ 0 & (h/2)^2 \end{bmatrix} \tag{10}$$

where (cx, cy) represents the center coordinates, and w, h denote the width and height. The second-order Wasserstein distance W_2^2 [29] between the predicted distribution $\mathcal{N}_p$ and ground truth $\mathcal{N}_{gt}$ is calculated as:

$$W_2^2(\mathcal{N}_p, \mathcal{N}_{gt}) = ||\mathbf{m}_p - \mathbf{m}_{gt}||_2^2 + ||\mathbf{\Sigma}_p^{1/2} - \mathbf{\Sigma}_{gt}^{1/2}||_F^2 \tag{11}$$

Using Equation (11), we formulate the Normalized Wasserstein Penalty $\mathcal{R}_W$ [27]:

$$\mathcal{R}_W = 1 - \exp\left(-\frac{\sqrt{W_2^2}}{C}\right) \tag{12}$$

where C is a normalization constant set to 12.0, which tunes the sensitivity of the metric to distribution deviations. $\mathcal{R}_W$ provides robust, long-range gradient attraction for tiny objects without the sensitivity issues of Euclidean distance.

Exponential Scale Modulation. To achieve a smooth manifold transition between distributional and geometric metrics, we introduce a scale modulation factor $\mu(s)$. Distinct from step-wise weighting, we employ a generalized exponential decay function to model the "intensity" of the small-object characteristic:

$$\mu(s) = \exp\left(-\left(\frac{s}{\kappa}\right)^{\beta}\right) \tag{13}$$

In this equation:

- $s = \sqrt{w \cdot h}$ denotes the geometric mean scale of the target in pixels.
- κ is the scale constant, acting as the soft boundary of the small object definition.
- β controls the steepness of the transition. A higher β creates a "plateau" effect, ensuring consistent NWD dominance for targets smaller than κ before a rapid decay.

For extremely small targets ($s < \kappa$), the high-order exponent ensures $\mu(s) \approx 1$, indicating a dominance of distributional alignment. As the target size grows beyond κ, $\mu(s)$ decays rapidly towards 0, naturally handing over control to geometric regression.

We formulate the final SABL for the regression branch by internally rectifying the CIoU equation. Specifically, we replace the standard Euclidean penalty $\frac{\rho^2}{c^2}$ with a scale-adaptive hybrid penalty $\mathcal{R}_{Hybrid}$:

$$\mathcal{R}_{Hybrid} = \mu(s) \cdot \mathcal{R}_W + (1 - \mu(s)) \cdot \frac{\rho^2(b, b^{gt})}{c^2} \tag{14}$$

Regression Loss. Substituting this into the IoU-based framework, the SABL is defined as:

$$\mathcal{L}_{reg} = 1 - \text{IoU} + \mathcal{R}_{Hybrid} + \alpha v \tag{15}$$

This formulation ensures that for small objects, the gradient contribution from the unstable Euclidean term is physically suppressed by the coefficient $(1 - \mu(s)) \approx 0$, while the robust Wasserstein term $\mathcal{R}_W$ guides the optimization. Simultaneously, the aspect ratio penalty αv remains active with full weight, ensuring shape consistency even when spatial localization relies on distributional metrics.

Classification Loss. We employ the Binary Cross-Entropy (BCE) loss for the classification task. For each category, the loss is calculated independently to handle the multi-label classification problem. The formulation is defined as:

$$L_{cls} = -\frac{1}{N_{pos}} \sum_i (y_i \ln(\hat{p}_i) + (1 - y_i) \ln(1 - \hat{p}_i)) \tag{16}$$

where N_{pos} denotes the number of positive samples, $y_i \in \{0, 1\}$ represents the ground truth label for the i-th category (1 for the target class and 0 otherwise), and $\hat{p}_i \in [0, 1]$ is the predicted probability after the sigmoid activation.

Total Loss Function. To jointly optimize the classification and localization tasks, the final objective function of SSABNet is formulated as the weighted sum of the classification loss and the proposed regression loss:

$$L_{total} = \lambda_{cls} L_{cls} + \lambda_{reg} L_{reg} \tag{17}$$

where L_{reg} is the SABL proposed in Equation (15) for bounding box regression, and L_{cls} is the classification loss defined in Equation (16). λ_{cls} and λ_{reg} are the hyperparameters to balance the two tasks. In our experiments, we set $\lambda_{cls} = 0.5$ and $\lambda_{reg} = 7.5$ following the standard configuration of the baseline to ensure stable convergence.

4. Experiments

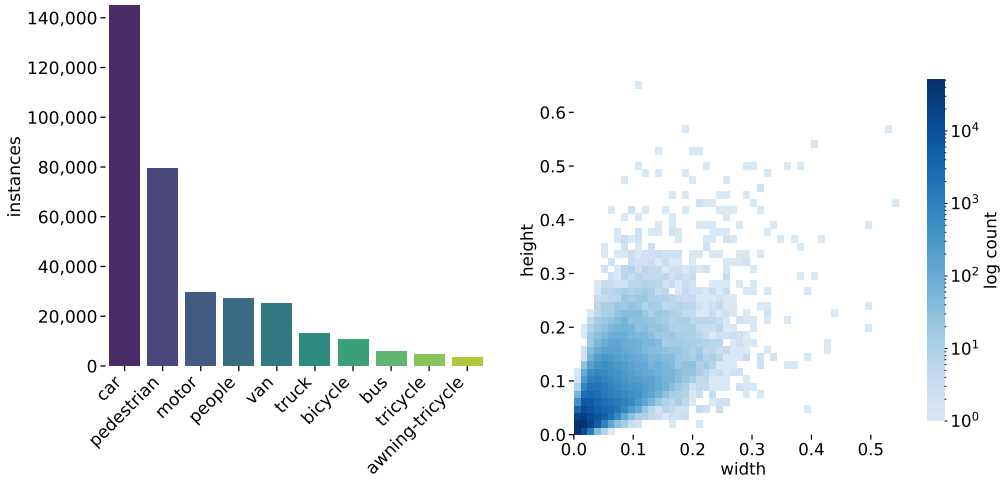
4.1. Datasets Description and Analysis

To verify the robustness of our method in diverse aerial scenarios, we utilize the VisDrone and UAVDT datasets, both of which are characterized by high-density small objects and complex background clutter.

4.1.1. VisDrone Dataset

The VisDrone [30] benchmark serves as a comprehensive platform for drone-based vision tasks. It consists of a large-scale collection of images and video sequences acquired by various UAV platforms across fourteen different cities in China. This archive spans a broad spectrum of real-world settings encompassing crowded city avenues and open plazas alongside peaceful residential neighborhoods. Regarding the detection benchmark, it comprises 10,209 high-resolution static samples. The official partition allocates 6471 images to the training subset and 548 to the validation subset while reserving the remaining 3190 images for testing purposes. The annotations encompass ten distinct object classes

critical for aerial monitoring namely pedestrians and various types of vehicles and cycles. A significant challenge of VisDrone lies in its significant scale variation; image resolutions fluctuate between 960 × 540 and 2000 × 1500 pixels. As illustrated in Figure 4, the distribution of object instances exhibits a long-tail characteristic, and the majority of targets occupy a very small pixel area relative to the image size. This extreme scale disparity poses stringent requirements on the detector's multi-scale representation capability.



(**a**) Category Distribution (**b**) Scale Distribution

Figure 4. Statistical analysis of the VisDrone dataset. Subfigure (**a**) illustrates the instance frequency across ten categories, highlighting class imbalance. Subfigure (**b**) depicts the distribution of object scales (normalized width and height), demonstrating the dominance of small-scale targets.

4.1.2. UAVDT Dataset

Designed specifically for traffic surveillance tasks, the UAVDT benchmark [31] serves as a specialized platform for vehicle analysis. This dataset comprises a total of 38,327 frames with a standardized resolution of 1080 × 540 pixels. Unlike datasets with broad categories, UAVDT focuses exclusively on three vehicle types namely car, bus, and truck. For the experimental setup, we utilize the standard partition which assigns 23,258 images to the training set and reserves the remaining 15,069 images for testing. Although the category list is concise, the dataset presents substantial challenges including high traffic density and severe occlusion alongside diverse weather conditions. Figure 5 outlines the detailed statistical distribution of these instances.

4.2. Evaluation Protocols

To facilitate a comprehensive quantitative comparison, we adhere to the standard evaluation protocols defined by the MS COCO benchmark. The detection performance is primarily gauged using Average Precision (AP), which represents the mean precision averaged over ten Intersection-over-Union (IoU) thresholds ranging from 0.50 to 0.95 (step size 0.05). To characterize localization accuracy under varying strictness, we also report AP_{50} and AP_{75}, corresponding to IoU thresholds of 0.5 and 0.75, respectively.

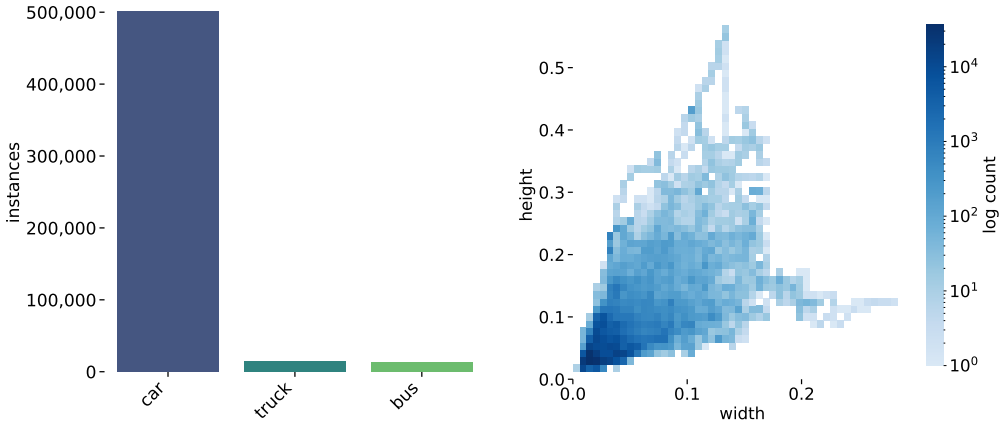
Considering the multi-scale nature of aerial imagery, specific attention is paid to scale-stratified metrics: AP_s, AP_m, and AP_l. These denote the AP for objects with pixel areas less than 32^2, between 32^2 and 96^2, and greater than 96^2, respectively. Given the specific challenges of this study, AP_s serves as a critical indicator for evaluating small object recovery.

Furthermore, to scrutinize the classifier's ability to distinguish targets from the background, we utilize Precision (P) and Recall (R). As formalized in Equation (18), Precision

quantifies the proportion of true positives among all predicted instances, while Recall measures the fraction of ground truth targets successfully identified:

$$\text{Precision} = \frac{N_{TP}}{N_{TP} + N_{FP}}, \quad \text{Recall} = \frac{N_{TP}}{N_{TP} + N_{FN}} \tag{18}$$

where N_{TP}, N_{FP}, and N_{FN} represent the counts of True Positives, False Positives, and False Negatives, respectively.



(**a**) Category Distribution (**b**) Scale Distribution

Figure 5. Statistical overview of the UAVDT dataset. Subfigure (**a**) shows the instance count for the three vehicle categories. Subfigure (**b**) displays the distribution of object dimensions relative to image size, confirming the prevalence of small objects.

4.3. Implementation Details

To ensure the reproducibility of our results, we provide a detailed description of the hardware and software environment. Experimental validation was conducted utilizing a computational workstation powered by eight NVIDIA GeForce RTX 3090 GPUs, each featuring 24 GB of VRAM, manufactured by NVIDIA Corporation (Santa Clara, CA, USA). The central processing unit consists of dual AMD EPYC 7773X 64-Core Processors. On the software side, the framework is implemented in Python 3.9.19 using PyTorch 2.0.1. GPU acceleration is provided by CUDA 11.7.0 and CUDNN libraries. A detailed summary of the experimental environment is provided in Table 1.

For all experiments, input images were resized to a uniform resolution of 1024×1024 pixels. We employed the Adam optimizer to update network parameters, with a total batch size of 16 distributed across GPUs. The training process spanned 200 epochs. A cosine annealing strategy was adopted for learning rate decay, initializing at 0.001 and gradually reducing to 1×10^{-6}, with a linear warm-up phase during the first 3 epochs to stabilize early convergence.

To bolster the robustness of the model against varying aerial conditions, we employed a comprehensive data augmentation pipeline. This regimen integrates geometric transformations including stochastic flipping and scaling and cropping alongside photometric distortions such as color jittering. Furthermore, Mosaic augmentation complements these techniques to enrich sample diversity for minute objects.

Table 1. Detailed configuration of the hardware and software environment.

Component	Specification
Hardware	
GPU	8 × NVIDIA GeForce RTX 3090 (24 GB)
CPU	2 × AMD EPYC 7773X (64-Core)
Software	
OS	Ubuntu 22.04.2 LTS
Language	Python 3.9.19
Framework	PyTorch 2.0.1
CUDA Version	11.7.0

The hyperparameters for the proposed SABL are configured to ensure a smooth transition between distributional alignment and geometric refinement. Specifically, the scale decay constant κ in Equation (13) is set to 32, which aligns with the standard definition of small objects (area $< 32 \times 32$ pixels) in the COCO protocol. To regulate the transition dynamics, we introduce a steepness factor β set to 6.0. This parameter creates a "plateau" effect for micro-scale targets ($s < \kappa$), ensuring that the Wasserstein penalty remains dominant and stable before decaying rapidly for larger objects. Additionally, the normalization constant C in the Wasserstein penalty term (Equation (12)) is set to 12.0. This value was empirically determined to normalize the magnitude of the Wasserstein gradient, preventing it from overwhelming the optimization process. A uniform hyperparameter configuration was maintained for both the VisDrone and UAVDT benchmarks. By abstaining from dataset-specific customization, we empirically validate the inherent generalization capability of the proposed framework.

4.4. Main Results

We utilize metrics directly from the original literature, given that official implementations and comprehensive training logs are not universally available for all comparison methods. Although negligible variations in computational environments may exist, the uniform evaluation on standard VisDrone and UAVDT benchmarks preserves the integrity and equity of this comparative analysis.

Quantitative Analysis on VisDrone. To empirically verify the architectural superiority of SSABNet, we executed an extensive benchmarking campaign against contemporary state-of-the-art methods on the challenging VisDrone dataset. The resulting quantitative metrics are tabulated in Table 2. Our evaluation protocol leverages standard indicators—namely AP, AP_{50}, and AP_{75}—to provide a holistic assessment of detection precision. Furthermore, to explicitly gauge the model's robustness across varying scales, we incorporated scale-specific metrics (AP_s, AP_m, and AP_l), which are critical for analyzing aerial imagery containing objects of drastically different sizes.

As evidenced by the data in Table 2, SSABNet demonstrates superior detection performance, outperforming competing methods across multiple evaluation dimensions. In terms of overall accuracy, our method achieves a consistent improvement of 1.3% in AP compared with the strongest competing approach, highlighting its consistent superiority. More notably, a substantial improvement is observed in the high-precision regime: the AP_{75} score exceeds the runner-up by 3.6%. This indicates that our model is far more capable of generating precise bounding boxes with high IoU, rather than merely making loose predictions. Crucially, the improvements are most prominent in scale-specific metrics, which directly address the challenges of drone vision. Specifically, AP_s and AP_m are increased by 2.5% and 3.6%, respectively. These quantitative improvements corroborate SSABNet's efficacy in

retrieving feature representations for small and medium-scale instances, which represent the predominant target categories within aerial imagery.

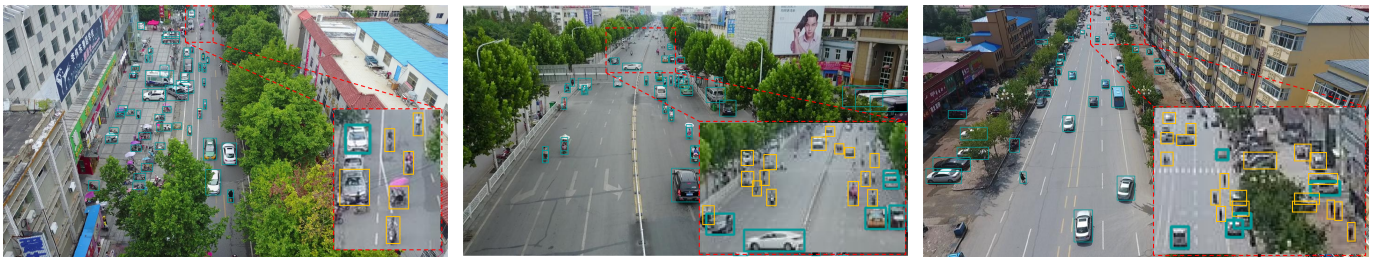
Furthermore, to strictly validate the effectiveness of our design, we conducted a controlled comparison with DINO+UGS using the identical DarkNet backbone (denoted by *). As shown in Table 2, SSABNet outperforms this baseline by 3.1% in AP, confirming that the gains stem from our unique architecture rather than the backbone. Regarding efficiency, SSABNet achieves a superior balance compared to representative methods. It operates at 21.1 FPS, which is significantly faster than YOLC (1.8 FPS) and DINO+UGS (11.5 FPS), while simultaneously achieving higher detection accuracy.

Table 2. Statistical summary of detection performance on the VisDrone. Entries in bold indicate the superior results. P. denotes the number of parameters, and GFL. represents Giga Floating-point Operations Per Second (GFLOPS).

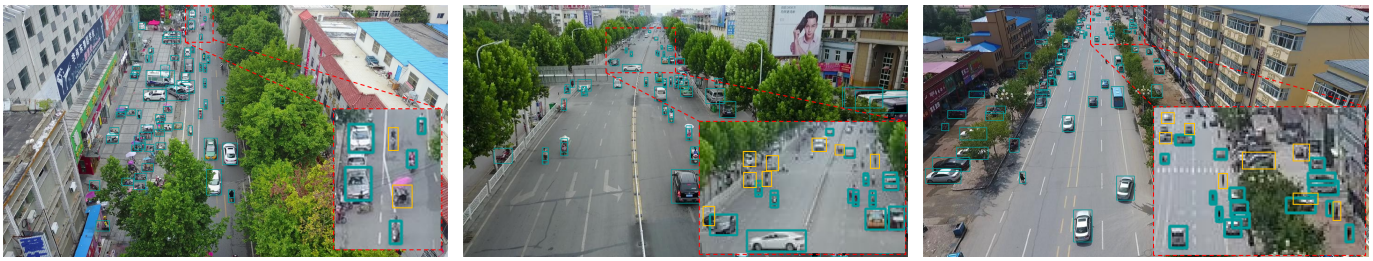
Method	Year & Venue	Backbone	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l	P.(M)	GFL.	FPS
FRCNN [12]	2017 TPAMI	ResNeXt101	24.4	47.8	21.8	17.8	34.8	34.3	-	-	1.1
ClusDet [32]	2019 ICCV	ResNeXt101	32.4	56.2	31.6	-	-	-	-	-	1.3
DREN [33]	2019 ICCVW	ResNeXt152	30.3	-	-	-	-	-	-	-	-
CenterNet [34]	2019 arXiv	Hourglass104	27.8	47.9	27.6	21.3	42.1	49.8	-	-	4.7
DMNet [35]	2020 CVPRW	ResNeXt101	29.4	49.3	30.6	21.6	41.0	56.7	-	-	1.6
AMRNet [36]	2020 arXiv	ResNeXt101	32.1	-	-	23.2	43.9	**60.5**	-	-	-
AFSM [37]	2020 arXiv	CBResNet50+DCN	39.5	**67.0**	39.5	-	-	-	-	-	-
YOLOX [26]	2021 arXiv	CSPv5-M	27.6	47.7	27.5	17.6	41.0	46.1	25.3	**73.8**	67.2
GLSAN [38]	2021 TIP	ResNet101	32.5	55.8	30.0	-	-	-	-	-	-
HRDNet [39]	2021 ICME	ResNeXt50+101	35.5	62.0	35.1	-	-	-	-	-	-
CDMNet [40]	2021 ICCVW	ResNeXt101	30.7	51.3	32.0	22.2	42.4	44.7	-	-	-
FCOS+SAHI [41]	2022 ICIP	-	-	38.5	-	-	-	-	-	-	-
VFNet+SAHI [41]	2022 ICIP	-	-	42.2	-	-	-	-	-	-	-
TOOD+SAHI [41]	2022 ICIP	-	-	43.5	-	-	-	-	-	-	-
QueryDet [42]	2022 CVPR	ResNet50	28.3	48.1	28.8	-	-	-	-	-	2.7
UFPMP-Det [43]	2022 AAAI	ResNeXt101	40.1	66.8	41.3	-	-	-	-	-	6.6
CEASC [44]	2023 CVPR	ResNet18	28.7	50.7	28.4	-	-	-	-	-	21.7
YOLOv9 [45]	2024 ECCV	GELAN	29.5	49.9	29.4	20.2	41.7	47.8	**20.1**	76.8	**67.4**
YOLC [46]	2024 TITS	ResNeXt101	36.3	60.1	37.4	28.9	47.5	51.8	130.2	536.9	1.8
GLSDet [2]	2025 TGRS	CSPv5-M	32.3	54.2	32.8	23.3	42.9	50.2	60.9	100.4	48.6
DINO+UGS [5]	2025 ICCV	ResNet-50	38.1	61.9	-	24.2	47.3	56.6	48.1	128.1	11.5
DINO+UGS * [5]	2025 ICCV	DarkNet	38.3	62.3	-	24.3	47.7	56.4	50.3	136.8	10.7
SSABNet	Ours	DarkNet	**41.4**	63.2	**44.9**	**31.4**	**51.1**	57.4	64.1	365.4	21.1

Note. For the sake of accuracy, the performance metrics for the comparison methods are adopted directly from their respective publications. Due to the absence of publicly accessible source code for certain baselines, we retain their originally reported experimental configurations, such as input dimensions. Notwithstanding these minor variations, the utilization of the standardized VisDrone benchmark guarantees a fair and valid assessment. Specifically, the method marked with * denotes our re-implementation using the same DarkNet backbone as SSABNet to ensure a strictly controlled comparison.

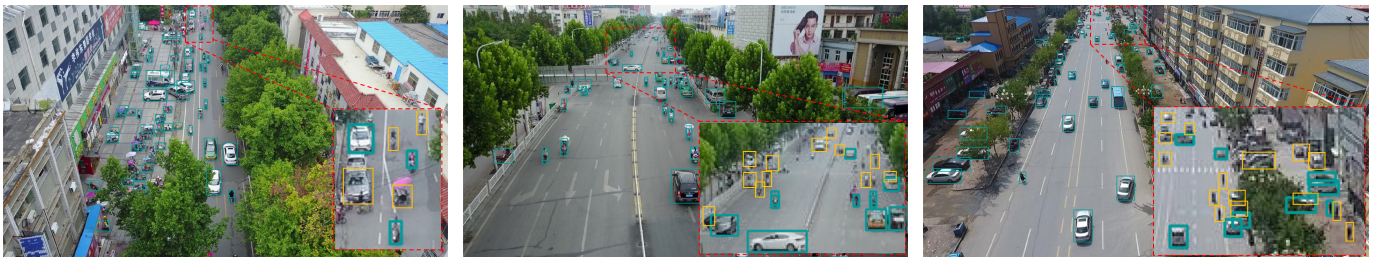
Qualitative Visualization on VisDrone. To empirically validate the superiority of SSABNet, Figure 6 presents a qualitative comparison against representative and recently proposed algorithms, including YOLOv9, YOLC, and GLSDet. The visualization covers three representative scenarios selected from the test set. For clarity, the zoomed-in regions are marked with red dashed boxes. Within these regions, the orange bounding boxes in the baseline results explicitly indicate missed detections, highlighting the limitations of current methods in capturing weak feature signals. In contrast, green bounding boxes represent the targets correctly localized by the models. As shown in the bottom row, SSABNet successfully localizes these difficult targets without omission. These comparative results strongly corroborate that the synergistic integration of the SSA architecture and the SABL optimization strategy enables the network to maintain robust detection capabilities even in demanding aerial scenes.



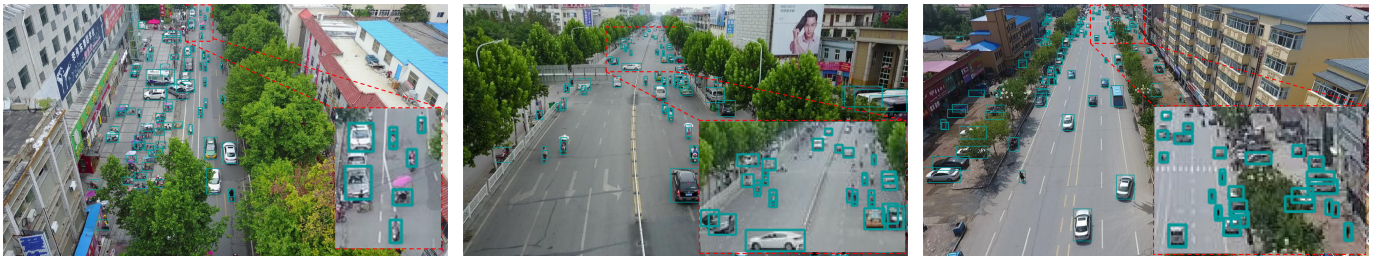
(**a**) YOLOv9



(**b**) YOLC



(**c**) GLSDet



(**d**) SSABNet (Ours)

Figure 6. Visual comparison of detection results on the VisDrone dataset under three representative scenarios. Rows (**a**–**d**) correspond to the detection outputs of YOLOv9, YOLC, GLSDet, and our proposed SSABNet, respectively. The red dashed boxes indicate zoomed-in regions. Inside these regions, orange bounding boxes denote missed detections, while green bounding boxes represent correctly detected targets. As observed, SSABNet demonstrates superior capability in detecting targets compared to other representative methods.

Category-Specific Performance Analysis on VisDrone. To offer a more granular perspective, Table 3 delineates the performance metrics of SSABNet across individual object categories within the VisDrone benchmark. An examination of these results reveals a distinct performance stratification that correlates strongly with both the physical scale and the structural rigidity of the targets.

Table 3. Class-wise detection metrics achieved by SSABNet across the VisDrone dataset.

Class	Instances	P	R	AP	AP_{50}	AP_{75}
pedestrian	8844	75.9	66.7	40.8	74.0	39.7
people	5125	72.7	56.5	30.5	62.5	25.6
bicycle	1287	51.4	47.6	25.6	47.1	23.8
car	14,064	84.8	87.7	70.4	91.5	80.5
van	1975	66.4	62.3	49.3	64.9	57.3
truck	750	66.3	55.2	43.8	60.4	48.8
tricycle	1045	61.7	52.5	33.8	53.2	38.1
awning-tricycle	532	44.4	29.9	20.3	30.0	23.8
bus	251	77.5	70.9	59.6	76.1	70.1
motor	4886	70.2	70.5	40.1	72.8	40.7

On one end of the spectrum, the model demonstrates exceptional robustness in detecting macro-scale, rigid objects. For instance, the categories car and bus achieve high AP scores of 70.4% and 59.6%, respectively. This superior performance can be ascribed to the prominent geometric signatures and relatively substantial pixel coverage of these objects, which facilitate the extraction of discriminative features even from high altitudes.

Conversely, categories characterized by extreme spatial sparsity, such as bicycle and awning-tricycle, exhibit lower AP scores. However, a closer examination of the Recall rates reveals the effectiveness of our proposed optimization strategy.

For small and vulnerable classes like pedestrian and motor, SSABNet achieves robust Recall of 66.7% and 70.5%, respectively. This directly substantiates the theoretical contribution of the SABL. Unlike standard IoU-based losses where gradients from these small objects (occupying $<0.1\%$ of the image) would vanish compared to large buses or trucks, SABL dynamically amplifies their gradient contribution.

Regarding the long-tail categories like awning-tricycle (only 532 instances), while the absolute AP (20.3%) is constrained by data scarcity, the model still manages to localize a significant portion of targets (29.9% Recall). This suggests that SABL helps prevent the model from completely ignoring these rare, small classes during the dominance of majority classes, forcing the network to learn discriminative features even from limited samples.

Synthesizing the quantitative metrics with the qualitative visual evidence, it is evident that SSABNet is highly tailored to the specific constraints and demands of aerial surveillance. By effectively addressing the dual challenges of feature loss and scale imbalance, the framework not only advances the SOTA in detection accuracy but also maintains computational feasibility. These attributes are pivotal for deploying robust vision systems on UAV platforms, thereby significantly enhancing their capability for autonomous scene understanding and ground target monitoring in complex, real-world environments.

Quantitative Analysis on UAVDT. To further verify the generalization capability of the proposed method in traffic surveillance scenarios, we conducted a rigorous comparative evaluation on the UAVDT dataset. The quantitative results, benchmarked against recent SOTA detectors, are summarized in Table 4. To ensure a strict and consistent evaluation protocol, we adhere to the standard COCO definitions for object scales.

As evidenced by the statistical data, SSABNet demonstrates exceptional performance advantages over existing solutions. In terms of overall detection accuracy, our method achieves a new SOTA AP of 25.1%, outperforming the second-best approach by a margin of 0.5%. Similar trends are observed in localization precision metrics, where SSABNet leads by 0.8% in AP_{50} and 1.6% in the stricter AP_{75} metric.

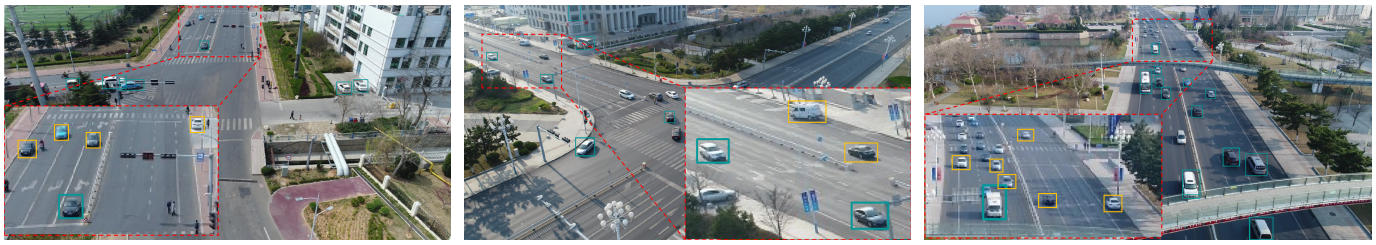
Table 4. Statistical summary of detection performance on the UAVDT. Entries in bold indicate the superior results.

Method	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l
R-FCN [47]	7.0	17.5	3.9	4.4	14.7	12.1
SSD [48]	9.3	21.4	6.7	7.1	17.1	12.0
FRCNN [12]	5.8	17.4	2.5	3.8	12.3	9.4
FRCNN [12] + FPN	11.0	23.4	8.4	8.1	20.2	26.5
ClusDet [32]	13.7	26.5	12.5	9.1	25.1	31.2
DREN [33]	17.1	-	-	-	-	-
CenterNet [34]	13.2	26.7	11.8	7.8	26.6	13.9
DMNet [35]	14.7	24.6	16.3	9.3	26.2	35.2
AMRNet [36]	18.2	30.4	19.8	10.3	31.3	33.5
GLSAN [38]	17.1	28.3	18.8	-	-	-
CDMNet [40]	16.8	29.1	18.5	11.9	29.0	15.7
YOLOX [26]	15.8	27.7	16.3	10	27.3	24.8
UFPMP-Det [43]	24.6	38.7	28.0	-	-	-
CEASC [44]	17.1	30.9	17.8	-	-	-
YOLOv9 [45]	16.1	28.1	16.5	10.6	27.9	25.3
YOLC [46]	19.3	30.9	20.1	10.9	**32.2**	**35.5**
GLSDet [2]	18.3	29.8	17.6	11.8	29.3	26.8
SSABNet (Ours)	**25.1**	**39.5**	**29.6**	**28.8**	26.8	20.4

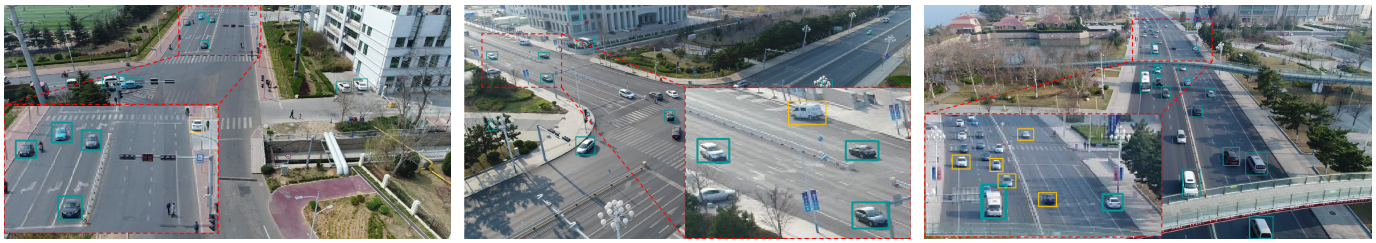
Note. Performance metrics for the comparison methods are extracted directly from their primary publications. Due to the unavailability of official implementations for certain baselines, we adhere to the experimental configurations as originally reported. Despite potential minor variances in setup, the consistent evaluation on the UAVDT benchmark guarantees the validity of this comparative assessment.

The most remarkable breakthrough, however, is observed in the small object regime. SSABNet achieves an unprecedented AP_s of 28.8%, representing a remarkable 16.9% absolute improvement in AP_s compared with the strongest competing methods that report AP_s on UAVDT. This effectively doubles the detection capability for pixel-limited targets compared to competing SOTA methods. This dominance is not attributed to a single component but validates the holistic synergy of the SSABNet framework: the SSA reconstructs critical spatial details, while the SABL ensures these restored features receive sufficient gradient focus during optimization. Together, they significantly push the performance boundary for small-target detection in aerial imagery.

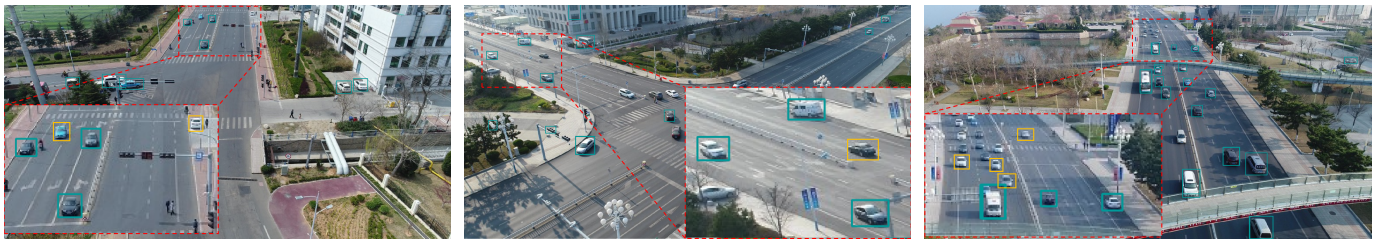
Qualitative Visualization on UAVDT. The empirical evidence from the UAVDT benchmark further attests to the robust generalization capability of SSABNet in aerial traffic surveillance. To visually demonstrate this, Figure 7 presents a qualitative comparison against representative and recently proposed algorithms, specifically YOLOv9, YOLC, and GLSDet. The visualization encompasses three representative scenarios characterized by varying traffic densities and viewing angles. For clarity, the zoomed-in regions are marked with red dashed boxes. Within these regions, the orange bounding boxes in the baseline results explicitly indicate missed detections, revealing the difficulty existing methods face in crowded environments. In contrast, green bounding boxes represent the targets correctly localized by the models. As observed in Row (d), SSABNet effectively addresses these failures, demonstrating exceptional sensitivity to pixel-limited vehicle targets. This visual evidence confirms that our framework significantly reduces false negatives in high-density traffic environments compared to other competitive detectors.



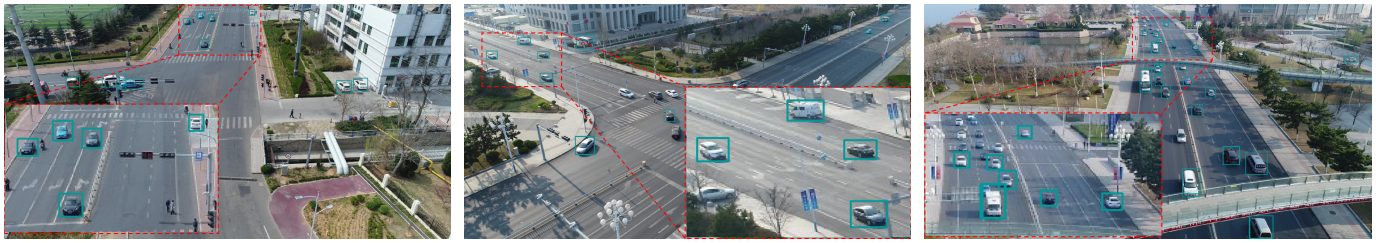
(**a**) YOLOv9



(**b**) YOLC



(**c**) GLSDet



(**d**) SSABNet (Ours)

Figure 7. Visual comparison of detection results on the UAVDT dataset under three challenging scenarios (Columns Left to Right). Rows (**a**–**d**) correspond to the detection outputs of YOLOv9, YOLC, GLSDet, and our proposed SSABNet, respectively. The red dashed boxes indicate zoomed-in regions. Inside these regions, orange bounding boxes denote missed detections, while green bounding boxes represent correctly detected targets. As observed, SSABNet demonstrates superior capability in detecting densely packed and small-scale vehicles compared to other representative methods.

Category-Specific Performance Analysis on UAVDT. Table 5 presents class-specific performance indicators, offering a detailed breakdown of detection results for each object type within the UAVDT benchmark. As observed, SSABNet achieves remarkable performance on the car class, with an AP_{50} of 77.0% and a Precision of 80.0%. This can be attributed to the abundant training samples (over 290 k instances), which allow the network to learn robust feature representations. In contrast, the performance on truck and bus categories is relatively lower. This disparity is primarily caused by the extreme class imbalance inherent in the dataset, where trucks and buses account for a very small fraction of the total instances compared to cars. Nevertheless, our model still demonstrates the capability to distinguish these distinct vehicle types within complex traffic flows.

Table 5. Class-wise detection metrics achieved by SSABNet across the UAVDT dataset.

Class	Instances	P	R	AP	AP_{50}	AP_{75}
car	293,483	80.0	71.4	47.6	77.0	54.4
truck	10,488	39.4	18.8	12.9	20.4	15.8
bus	3952	33.8	30.8	14.6	21.1	18.5

4.5. Ablation Study

To rigorously assess the contribution of each individual component within the SSABNet framework, we performed a systematic ablation study on the VisDrone benchmark. The quantitative outcomes, summarized in Table 6, were derived by progressively incorporating the SSA and the SABL into the baseline architecture.

The results in Table 6 reveal a clear hierarchy of performance gains. The initial integration of the SSA yielded a substantial boost in small-target detection, elevating the AP_s metric by a significant margin of 5.3%.

In parallel, the introduction of the SABL yields a distinct performance uplift, increasing AP_s by 2.2% compared to the baseline. Cumulatively, the joint deployment of these two innovations resulted in a remarkable 6.3% improvement in AP_s and a solid 5.0% increase in overall AP, confirming their combined efficacy.

Table 6. Component-wise impact analysis performed on the VisDrone benchmark. "Baseline" denotes the baseline model, "SSA" indicates the inclusion of the Spatial-Semantic Aggregation module, and "SABL" represents the Scale-Aware Balancing Loss function. Entries in bold indicate the superior results.

Baseline	SSA	SABL	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l
✓	-	-	36.4	56.8	34.8	25.1	47.7	**58.8**
✓	✓	-	40.6	62.1	43.7	30.4	50.7	55.4
✓	-	✓	38.6	59.5	41.1	27.3	50.2	58.2
✓	✓	✓	**41.4**	**63.2**	**44.9**	**31.4**	**51.1**	57.4

Figure 8 provides a visual benchmarking of detection performance. As revealed in the magnified insets, the baseline detector exhibits a susceptibility to omission errors, evidenced by the prevalence of unrecovered targets which are demarcated by orange bounding boxes. In contrast, SSABNet successfully localizes the vast majority of challenging instances, with the exception of a few extremely difficult targets. The heatmaps corresponding to the baseline model and SSABNet are displayed above their respective detection images. The comparison of these heatmaps reveals that SSABNet demonstrates superior capability in focusing on the targets. This visual evidence corroborates the quantitative metrics, demonstrating that our proposed method effectively mitigates feature erosion in complex environments.

The proposed SSABNet strikes a practical balance between detection accuracy and computational efficiency. The model comprises 64.1 M parameters and entails a computational load of 365.4 GFLOPs at an input resolution of 1024×1024. In terms of runtime performance, it achieves an inference speed of 21.1 FPS on a single NVIDIA RTX 3090 GPU (NVIDIA Corporation, Santa Clara, CA, USA), with a memory footprint of approximately 3.2 GB.



(**a**) Baseline (**b**) Ground Truth (**c**) SSABNet (Ours)

Figure 8. Qualitative comparison on the VisDrone dataset. Top Row: Heatmap visualizations of the corresponding columns. Red colors indicate regions with higher activation values, while blue colors represent lower activation. The green bounding boxes denote correctly detected targets, while the orange bounding boxes indicate the targets missed by the algorithms. Bottom Row: Detection results. Column (**a**) shows that the blurred feature activations in the baseline lead to missed detections (orange boxes). Column (**c**) demonstrates that SSABNet generates sharp, focused feature activations (top), resulting in accurate detection of small targets (bottom).

While onboard testing on physical UAV hardware was not conducted in this study due to equipment constraints, the current metrics indicate that SSABNet is a viable candidate for high-resolution aerial analysis. Although the inference speed is 21.1 FPS on a high-end GPU, the reduced parameter count compared to Transformer-based models suggests potential for optimization on edge devices. Future work will prioritize model compression techniques such as quantization to satisfy the stringent efficiency constraints of UAV hardware.

In summary, the ablation study confirms that the integration of SSA and SABL is not merely additive but synergistic. This combination effectively bridges the semantic-spatial gap in feature representation and corrects the optimization imbalance inherent in multi-scale detection tasks.

To further analyze classification performance, Figure 9 compares the normalized confusion matrices. A closer inspection of the diagonal elements reveals that SSABNet consistently achieves higher True Positive rates across all categories compared to the baseline. This improvement suggests that the feature aggregation provided by SSA, combined with the gradient rebalancing of SABL, significantly reduces semantic ambiguity. Consequently, the model not only localizes objects better but also classifies them with higher confidence.

The superiority of our approach is further substantiated by the Precision-Recall (PR) curves presented in Figure 10. SSABNet exhibits PR curves that enclose a larger area under the curve (AUC) compared to the baseline, particularly for small-scale categories such as pedestrians, bicycles, and motorcycles. The smoother trajectory of the SSABNet curves indicates a more stable precision-recall trade-off across varying confidence thresholds. This stability is attributed to the SSA's ability to preserve structural integrity and the SABL function's role in alleviating the bias towards easy samples. Overall, these results confirm SSABNet's robustness in complex aerial sensing scenarios.

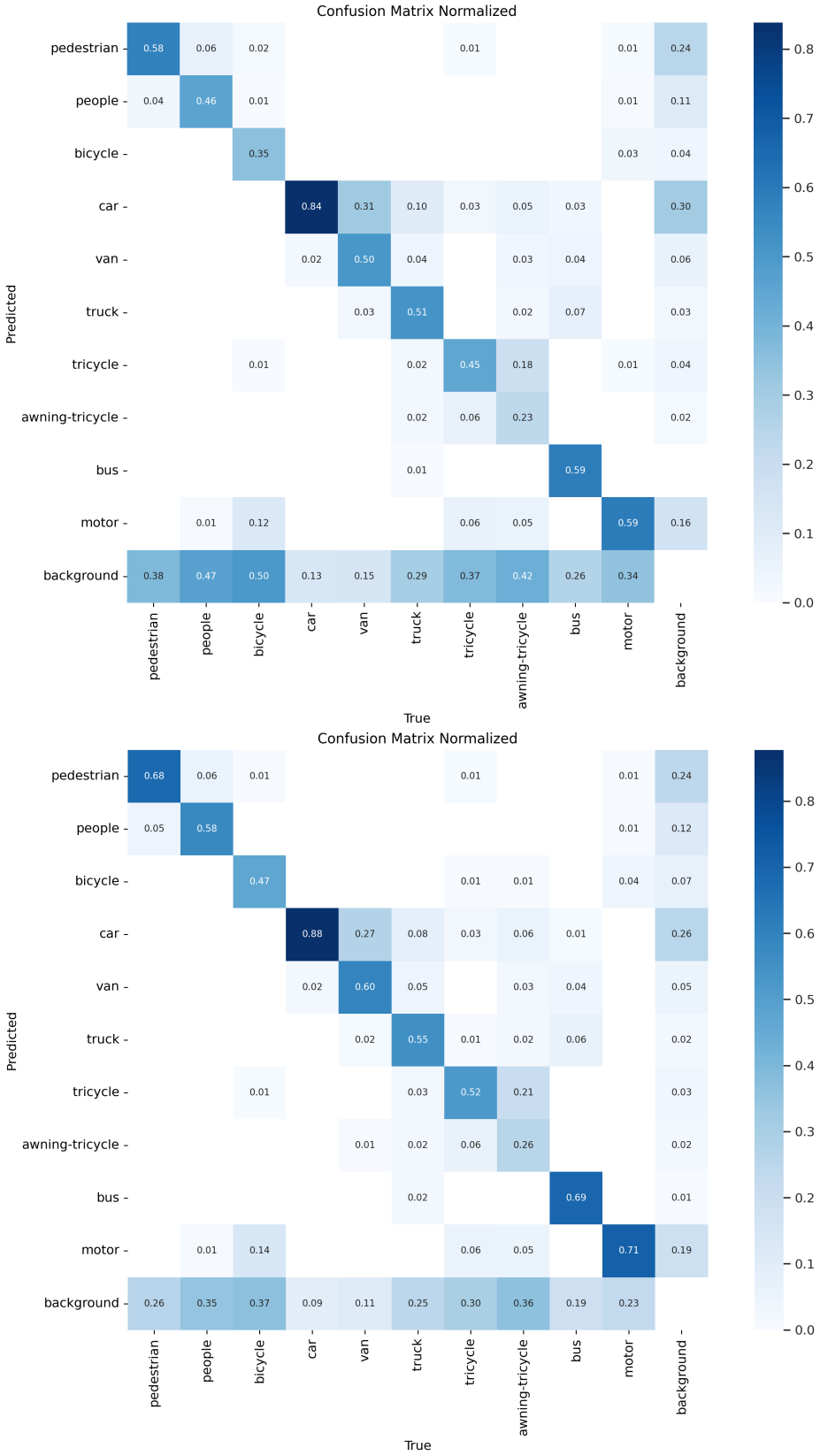


Figure 9. Normalized confusion matrices for the baseline model (**top**) and SSABNet (**bottom**) on the VisDrone dataset. The diagonal values represent classification accuracy. SSABNet demonstrates consistently improved discrimination capability across diverse categories.

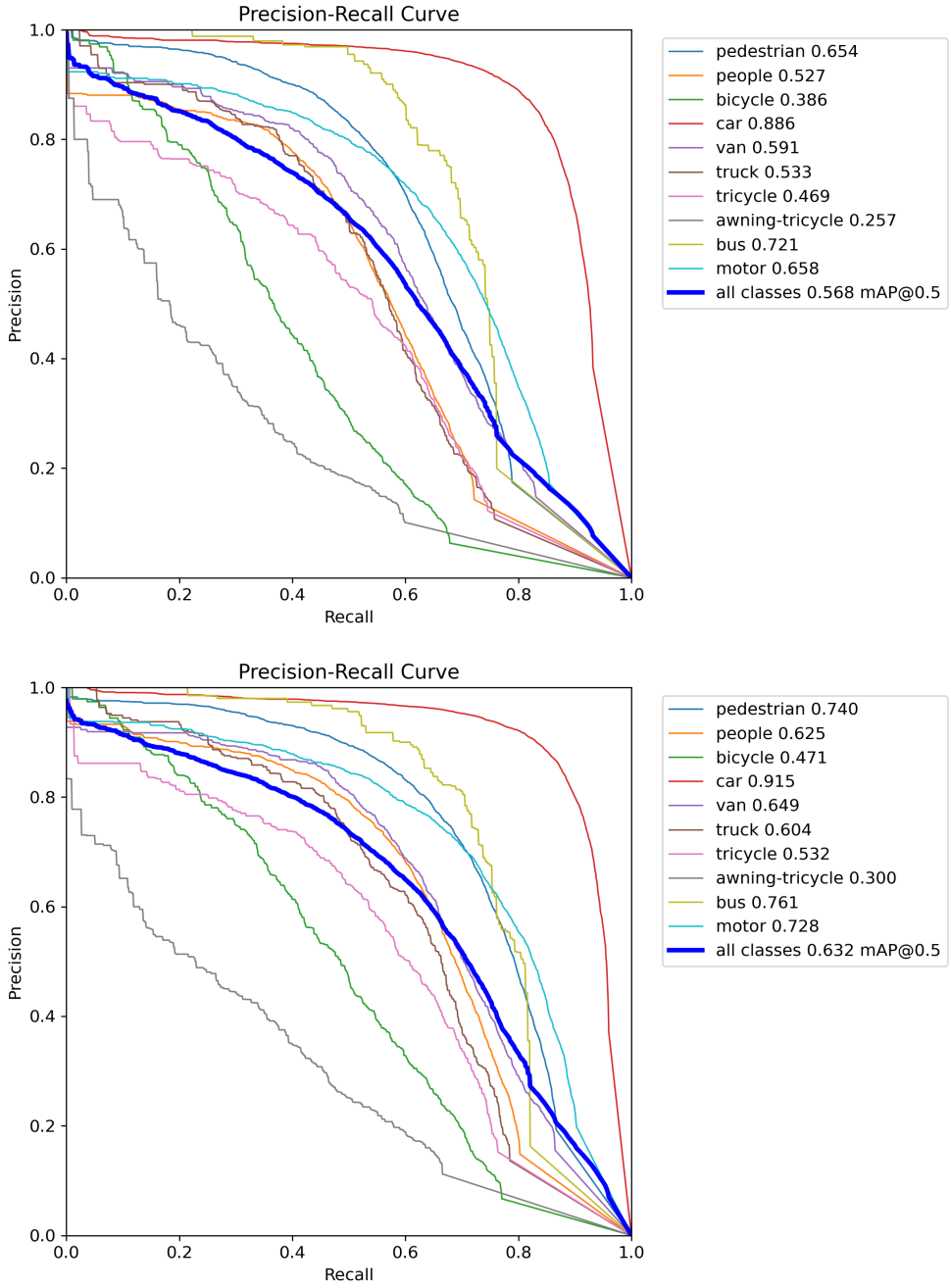


Figure 10. Comparison of Precision-Recall (PR) curves between the baseline model (**top**) and SSABNet (**bottom**). The curves demonstrate that SSABNet achieves superior envelope coverage, indicating better overall detection performance, especially for small object classes.

5. Discussion

The experimental results presented above demonstrate the superiority of SSABNet in detecting small targets. In this section, we delve deeper into the underlying mechanisms of these improvements, analyze the trade-off between computational cost and accuracy, and discuss the limitations of the current framework.

5.1. Mechanism Analysis

The substantial performance gains, particularly in the small-object regime (AP_s), can be attributed to the specific architectural and optimization designs of SSABNet.

First, regarding feature representation, the empirical evidence from the ablation study underscores the critical function of the SSA module. By establishing high-fidelity restoration pathways for feature compensation, SSA effectively recovers high-frequency spatial details that are typically attenuated in deep convolutional layers. This restoration is pivotal for distinguishing small targets from background noise, as evidenced by the sharp activation maps shown in Figure 8. Second, regarding optimization, the improvements verify the complementary relationship between architectural and objective-level enhancements. While SSA enhances the quality of the feature representation via aggregation, the proposed SABL optimizes the learning focus. By dynamically re-weighting the gradient contribution of pixel-limited samples using the Wasserstein penalty, SABL prevents tiny objects from being overwhelmed by the dominant loss of larger backgrounds, ensuring balanced convergence.

5.2. Trade-Off Between Accuracy and Efficiency

A critical concern in UAV-based detection is the balance between detection performance and computational overhead. As shown in Table 2, we included the parameter count, GFLOPS, and inference time for a comprehensive evaluation.

It is observed that SSABNet incurs a higher computational cost with 64.1 million parameters and 365.4 GFLOPS compared to general-purpose real-time detectors such as YOLOv9 which utilizes 57.3 million parameters. This increase is primarily driven by the sophisticated feature aggregation pathways and the intensive spatial restoration mechanisms specifically designed to resolve tiny details, which naturally require more computational resources than standard coarse-grained detectors.

However, this additional cost is well-justified by the performance breakthrough. Compared to YOLC which relies on an extremely heavy backbone utilizing 130.2 million parameters and 536.9 GFLOPS, SSABNet achieves significantly higher accuracy of 41.4% AP against 36.3% AP while requiring much lower computational demand. Furthermore, with an inference speed of 21.1 frames per second on a single RTX 3090, our model maintains a quasi-real-time processing capability, making it a viable solution for high-end UAV platforms where precision is the priority.

5.3. Generalization Capability

Beyond the numerical superiority on a single dataset, the empirical success on the UAVDT benchmark serves as a critical testament to the model's robust generalization capabilities. The consistent and substantial breakthroughs in small-target detection—persisting across both VisDrone and UAVDT benchmarks—imply that our framework has successfully learned scale-invariant representations. This enables the network to adeptly navigate the heterogeneous and complex conditions typical of UAV-captured footage, such as varying altitudes and dynamic traffic flows.

5.4. Limitations

Despite the promising performance, SSABNet has certain limitations that merit further investigation. First, regarding the trade-off between efficiency and accuracy, our method prioritizes the recovery of extremely small targets through sophisticated feature aggregation. While this design significantly boosts detection recall, it entails a slightly higher computational cost compared to ultra-lightweight detectors designed purely for speed. Therefore, further optimization via techniques like model pruning or quantization would be beneficial for deployment on hardware with extremely limited power budgets, such as embedded chips on small drones. Second, the current framework utilizes horizontal bounding boxes for object localization. In aerial scenarios where targets often appear with arbitrary orientations, such as vehicles turning at intersections, horizontal boxes may

introduce background noise within the detection region or cause overlap in dense clusters. Extending the framework to support oriented bounding boxes could further refine localization accuracy in such cases. Lastly, the current model is primarily verified on daylight scenarios. Its robustness in extreme weather conditions, including heavy fog or night-time low light where visual features are further degraded, remains to be fully explored.

6. Conclusions

In this work, we presented SSABNet to address the long-standing challenges of small-target detection in UAV-captured aerial imagery, where extreme scale imbalance, feature sparsity, and optimization bias jointly degrade detection performance. The proposed framework systematically tackles these issues from both the representation and optimization perspectives through two complementary components: the Spatial-Semantic Aggregation (SSA) module and the Scale-Aware Balancing Loss (SABL).

The SSA module constructs a high-fidelity feature restoration pathway that effectively reconciles fine-grained spatial cues with high-level semantic representations. By enforcing topology-aware feature aggregation, SSA preserves critical spatial details of pixel-limited targets while maintaining semantic consistency across scales. In parallel, the SABL introduces a scale-adaptive internal rectification mechanism to stabilize the regression process for small objects. By dynamically transitioning the supervision signal from distributional alignment for micro-scale targets to geometric refinement for larger objects, SABL alleviates gradient instability and mitigates scale-induced optimization bias.

Extensive experiments conducted on the VisDrone and UAVDT benchmarks demonstrate that SSABNet consistently outperforms existing state-of-the-art detectors, particularly in complex aerial scenarios dominated by small and densely distributed targets. These results validate the effectiveness of jointly enhancing feature fidelity and rebalancing gradient contributions for robust UAV-based small-object detection.

Future work will explore lightweight variants of SSABNet to further reduce computational latency for practical edge deployment on UAV platforms. In addition, we plan to extend the proposed framework to oriented object detection by incorporating rotation-aware modeling, enabling more accurate localization in densely packed aerial scenes.

Author Contributions: Conceptualization, H.Z.; methodology, H.Z.; software, Y.G.; validation, S.W.; formal analysis, H.Z.; investigation, H.Z.; resources, Z.O.; data curation, S.W.; writing—original draft preparation, H.Z.; writing—review and editing, H.Z., S.Y., and M.S.; visualization, Y.G. and S.Y.; supervision, Z.O.; project administration, Z.O. and M.S.; funding acquisition, Z.O. and S.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the Algorithm Cost R&D Project of the AI Capability Center, Digitalization Department, China United Network Communications Co., Ltd. Beijing Branch under Grant A2024054, the National Key Research and Development Program of China under Grant 2024YFC3308500, the National Natural Science Foundation of China under Grant 62406036, and the National Natural Science Foundation of China under Grant 62402055.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found in the VisDrone repository at https://github.com/VisDrone/VisDrone-Dataset (accessed on 26 January 2025) and the UAVDT repository at https://sites.google.com/view/grli-uavdt/ (accessed on 28 January 2025).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Sunkara, R.; Luo, T. No More Strided Convolutions or Pooling: A New CNN Building Block for Low-Resolution Images and Small Objects. In *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Grenoble, France, 19–23 September 2022*; Springer: Cham, Switzerland, 2022; pp. 443–459.
2. Chen, Y.; Ye, Z.; Sun, H.; Gong, T.; Xiong, S.; Lu, X. Global–Local Fusion with Semantic Information Guidance for Accurate Small Object Detection in UAV Aerial Images. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 4701115. [CrossRef]
3. Wu, H.; Yao, S.; Huang, F.; Wang, S.; Zhang, L.; Zheng, Z.; Ren, W. LVPTrack: High Performance Domain Adaptive UAV Tracking with Label Aligned Visual Prompt Tuning. In Proceedings of the AAAI Conference on Artificial Intelligence, Philadelphia, PA, USA, 25 February–4 March 2025; Volume 39, pp. 8395–8403.
4. Yao, S.; Han, X.; Zhang, H.; Wang, X.; Cao, X. Learning Deep Lucas-Kanade Siamese Network for Visual Tracking. *IEEE Trans. Image Process.* **2021**, *30*, 4814–4827. [CrossRef]
5. Sun, H.; Li, Y.; Yang, L.; Cao, X.; Zhang, B. Uncertainty-Aware Gradient Stabilization for Small Object Detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Honolulu, HI, USA, 19–23 October 2025; pp. 8407–8417.
6. Du, Z.; Hu, Z.; Zhao, G.; Jin, Y.; Ma, H. Cross-Layer Feature Pyramid Transformer for Small Object Detection in Aerial Images. *arXiv* **2024**, arXiv:2407.19696. [CrossRef]
7. Yao, S.; Zhu, R.; Wang, Z.; Ren, W.; Yan, Y.; Cao, X. UMDATrack: Unified Multi-Domain Adaptive Tracking Under Adverse Weather Conditions. *arXiv* **2025**, arXiv:2507.00648. [CrossRef]
8. Yang, Y.; Zang, B.; Song, C.; Li, B.; Lang, Y.; Zhang, W.; Huo, P. Small Object Detection in Remote Sensing Images Based on Redundant Feature Removal and Progressive Regression. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5629314. [CrossRef]
9. Sapkota, R.; Flores-Calero, M.; Qureshi, R.; Badgujar, C.; Nepal, U.; Poulose, A.; Zeno, P.; Vaddevolu, U.B.P.; Khan, S.; Shoman, M.; et al. YOLO Advances to Its Genesis: A Decadal and Comprehensive Review of the You Only Look Once (YOLO) Series. *Artif. Intell. Rev.* **2025**, *58*, 274. [CrossRef]
10. Zhang, H.; Li, F.; Liu, S.; Zhang, L.; Su, H.; Zhu, J.; Ni, L.M.; Shum, H.Y. DINO: DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection. In Proceedings of the Eleventh International Conference on Learning Representations, Kigali, Rwanda, 1–5 May 2023.
11. Tian, Z.; Shen, C.; Chen, H.; He, T. FCOS: Fully Convolutional One-Stage Object Detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9627–9636.
12. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef] [PubMed]
13. Liu, J.; Zhang, J.; Ni, Y.; Chi, W.; Qi, Z. Small-Object Detection in Remote Sensing Images with Super-Resolution Perception. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 15721–15734. [CrossRef]
14. Yao, S.; Liu, D.; Li, T.; Li, S.; Ren, W.; Cao, X. UAGLNet: Uncertainty-Aggregated Global-Local Fusion Network with Cooperative CNN-Transformer for Building Extraction. *arXiv* **2025**, arXiv:2512.12941. [CrossRef]
15. Yao, S.; Sun, H.; Xiang, T.Z.; Wang, X.; Cao, X. Hierarchical Graph Interaction Transformer with Dynamic Token Clustering for Camouflaged Object Detection. *IEEE Trans. Image Process.* **2024**, *33*, 5936–5948. [CrossRef]
16. Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; Savarese, S. Generalized Intersection over Union: A Metric and a Loss for Bounding Box Regression. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 658–666.
17. Zheng, Z.; Wang, P.; Liu, W.; Li, J.; Ye, R.; Ren, D. Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 12993–13000.
18. Pang, D.; Shan, T.; Ma, Y.; Ma, P.; Hu, T.; Tao, R. LRTA-SP: Low-Rank Tensor Approximation with Saliency Prior for Small Target Detection in Infrared Videos. *IEEE Trans. Aerosp. Electron. Syst.* **2025**, *61*, 2644–2658. [CrossRef]
19. Liu, K.; Fu, Z.; Jin, S.; Chen, Z.; Zhou, F.; Jiang, R.; Chen, Y.; Ye, J. ESOD: Efficient Small Object Detection on High-Resolution Images. *IEEE Trans. Image Process.* **2025**, *34*, 183–195. [CrossRef] [PubMed]
20. Liang, T.; Wang, Y.; Tang, Z.; Hu, G.; Ling, H. OPANAS: One-Shot Path Aggregation Network Architecture Search for Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 10195–10203.
21. Yang, G.; Lei, J.; Zhu, Z.; Cheng, S.; Feng, Z.; Liang, R. AFPN: Asymptotic Feature Pyramid Network for Object Detection. In *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics (SMC), Honolulu, HI, USA, 1–4 October 2023*; IEEE: Piscataway, NJ, USA, 2023; pp. 2184–2189.
22. Wang, J.; Chen, K.; Xu, R.; Liu, Z.; Loy, C.C.; Lin, D. CARAFE: Content-Aware ReAssembly of FEatures. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 3007–3016.

23. Liu, W.; Lu, H.; Fu, H.; Cao, Z. Learning to Upsample by Learning to Sample. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023; pp. 6027–6037.
24. Wang, A.; Chen, H.; Liu, L.; Chen, K.; Lin, Z.; Han, J. YOLOv10: Real-Time End-to-End Object Detection. *Adv. Neural Inf. Process. Syst. (NeurIPS)* **2024**, *37*, 107984–108011.
25. Ge, Z.; Liu, S.; Li, Z.; Yoshie, O.; Sun, J. OTA: Optimal Transport Assignment for Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 303–312.
26. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. YOLOX: Exceeding Yolo Series in 2021. *arXiv* **2021**, arXiv:2107.08430. [CrossRef]
27. Xu, C.; Wang, J.; Yang, W.; Yu, H.; Yu, L.; Xia, G.S. Detecting Tiny Objects in Aerial Images: A Normalized Wasserstein Distance and a New Benchmark. *ISPRS J. Photogramm. Remote Sens.* **2022**, *190*, 79–93. [CrossRef]
28. Wang, J.; Yang, W.; Li, H.C.; Zhang, H.; Xia, G.S. Learning Center Probability Map for Detecting Objects in Aerial Images. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 4307–4323. [CrossRef]
29. Peyré, G.; Cuturi, M. Computational Optimal Transport: With Applications to Data Science. *Found. Trends® Mach. Learn.* **2019**, *11*, 355–607. [CrossRef]
30. Du, D.; Zhu, P.; Wen, L.; Bian, X.; Lin, H.; Hu, Q.; Peng, T.; Zheng, J.; Wang, X.; Zhang, Y.; 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 (ICCV) Workshops, Seoul, Republic of Korea, 27–28 October 2019.
31. Du, D.; Qi, Y.; Yu, H.; Yang, Y.; Duan, K.; Li, G.; Zhang, W.; Huang, Q.; Tian, Q. The Unmanned Aerial Vehicle Benchmark: Object Detection and Tracking. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 370–386.
32. Yang, F.; Fan, H.; Chu, P.; Blasch, E.; Ling, H. Clustered Object Detection in Aerial Images. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 8311–8320.
33. Zhang, J.; Huang, J.; Chen, X.; Zhang, D. How to Fully Exploit the Abilities of Aerial Image Detectors. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops, Seoul, Republic of Korea, 27–28 October 2019.
34. Zhou, X.; Wang, D.; Krähenbühl, P. Objects as Points. *arXiv* **2019**, arXiv:1904.07850.
35. Li, C.; Yang, T.; Zhu, S.; Chen, C.; Guan, S. Density Map Guided Object Detection in Aerial Images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 190–191.
36. Wei, Z.; Duan, C.; Song, X.; Tian, Y.; Wang, H. Amrnet: Chips Augmentation in Aerial Images Object Detection. *arXiv* **2020**, arXiv:2009.07168. [CrossRef]
37. Gong, Z.; Li, D. Towards Better Object Detection in Scale Variation with Adaptive Feature Selection. *arXiv* **2020**, arXiv:2012.03265. [CrossRef]
38. Deng, S.; Li, S.; Xie, K.; Song, W.; Liao, X.; Hao, A.; Qin, H. A Global-Local Self-Adaptive Network for Drone-View Object Detection. *IEEE Trans. Image Process.* **2020**, *30*, 1556–1569. [CrossRef] [PubMed]
39. Liu, Z.; Gao, G.; Sun, L.; Fang, Z. HRDNet: High-Resolution Detection Network for Small Objects. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME), Shenzhen, China, 5–9 July 2021*; IEEE: Piscataway, NJ, USA, 2021; pp. 1–6.
40. Duan, C.; Wei, Z.; Zhang, C.; Qu, S.; Wang, H. Coarse-Grained Density Map Guided Object Detection in Aerial Images. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops, Montreal, BC, Canada, 11–17 October 2021; pp. 2789–2798.
41. Akyon, F.C.; Altinuc, S.O.; Temizel, A. Slicing Aided Hyper Inference and Fine-Tuning for Small Object Detection. In *Proceedings of the IEEE International Conference on Image Processing (ICIP), Bordeaux, France, 16–19 October 2022*; IEEE: Piscataway, NJ, USA, 2022; pp. 966–970.
42. Yang, C.; Huang, Z.; Wang, N. QueryDet: Cascaded Sparse Query for Accelerating High-Resolution Small Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 13658–13667.
43. Huang, Y.; Chen, J.; Huang, D. UFPMP-Det: Toward Accurate and Efficient Object Detection on Drone Imagery. In Proceedings of the AAAI Conference on Artificial Intelligence, Online, 22 February–1 March 2022; Volume 36, pp. 1026–1033.
44. Du, B.; Huang, Y.; Chen, J.; Huang, D. 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 (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 13435–13444.
45. Wang, C.Y.; Yeh, I.H.; Mark Liao, H.Y. YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information. In *Proceedings of the European Conference on Computer Vision (ECCV), Milan, Italy, 29 September–4 October 2024*; Springer: Cham, Switzerland, 2024; pp. 1–21.
46. Liu, C.; Gao, G.; Huang, Z.; Hu, Z.; Liu, Q.; Wang, Y. YOLC: You Only Look Clusters for Tiny Object Detection in Aerial Images. *IEEE Trans. Intell. Transp. Syst.* **2024**, *25*, 13863–13875. [CrossRef]

47. Dai, J.; Li, Y.; He, K.; Sun, J. R-FCN: Object Detection via Region-Based Fully Convolutional Networks. In Proceedings of the 30th International Conference on Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; Volume 29, pp. 379–387.
48. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single Shot Multibox Detector. In *Proceedings of the 14th European Computer Vision Conference (ECCV), Amsterdam, The Netherlands, 11–14 October 2016*; Springer: Cham, Switzerland, 2016; pp. 21–37.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.