electronics

MDPI

Article

AeroPinWorld: Revisiting Stride-2 Downsampling for Zero-Shot Transferable Open-Vocabulary UAV Detection

Jie Li [1,†], Mingze Guan [1,†], Jincheng Xu [2], Xun Du [1], Haonan Chen [1] and Yang Liu [1,*]

[1] School of Telecommunications Engineering, Xidian University, Xi'an 710071, China; 24012100031@stu.xidian.edu.cn (J.L.); 24012100051@stu.xidian.edu.cn (M.G.); duxun@stu.xidian.edu.cn (X.D.); chenhaonan@stu.xidian.edu.cn (H.C.)
[2] Faculty of Integrated Circuit, Xidian University, Xi'an 710071, China; 24009190007@stu.xidian.edu.cn
* Correspondence: yangl@xidian.edu.cn
† These authors contributed equally to this work.

Abstract

Open-vocabulary object detection in unmanned aerial vehicle (UAV) imagery remains challenging under zero-shot cross-dataset transfer because tiny and cluttered targets are highly sensitive to early resolution reduction under domain shift. This study aims to improve transferable open-vocabulary UAV detection by revisiting stride-2 downsampling in YOLO-World v2 as a transfer-critical bottleneck. AeroPinWorld is introduced as a pinwheel-augmented YOLO-World v2 that selectively replaces four key stride-2 transitions with pinwheel-shaped convolution (PConv) so as to reduce aliasing, alleviate sampling-phase sensitivity, and preserve fine-grained local structures, while keeping the original detection head unchanged to ensure a fair and efficient comparison. The model is trained on COCO2017 for 24 epochs from official pretrained weights and directly evaluated, without target-domain fine-tuning, on VisDrone2019-DET and UAVDT using fixed offline prompt vocabularies. Compared with YOLO-World v2-S, AeroPinWorld improves zero-shot transfer performance on VisDrone from 0.112 to 0.135 mAP and from 0.054 to 0.063 AP_S, and it also yields consistent gains on UAVDT. Ablation studies show that both early backbone replacements and head bottom–up replacements contribute to the final gains, with their combination achieving the best accuracy–efficiency trade-off. These results indicate that selectively redesigning transfer-critical downsampling operators is an effective and lightweight way to improve zero-shot open-vocabulary UAV detection under aerial domain shift.

Keywords: open-vocabulary object detection; zero-shot transfer; cross-dataset generalization; UAV detection; domain shift; tiny object detection

check for updates

Academic Editors: Stefanos Kollias and Manohar Das

Received: 26 February 2026
Revised: 23 March 2026
Accepted: 24 March 2026
Published: 25 March 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.

1. Introduction

Open-vocabulary object detection (OVOD) [1–3] aims to detect objects beyond a fixed set of training categories by leveraging vision–language priors and text prompts [4]. This capability is especially attractive for unmanned aerial vehicle (UAV) perception, where data collection is relatively easy but exhaustive instance-level annotation is expensive and often lags behind rapidly changing deployment scenarios. In practical aerial applications, detectors are frequently required to recognize categories that are absent from the source training annotations while remaining efficient enough for real-time use.

UAV imagery also amplifies several long-standing detection challenges. Objects are often tiny, densely distributed, and embedded in cluttered backgrounds, while viewpoint

changes, altitude variation, motion blur, compression artifacts, and illumination changes introduce severe domain shift relative to natural-image corpora [5–8]. Consequently, transferring an open-vocabulary detector trained on generic images to UAV imagery is considerably more difficult than in-domain evaluation.

This study focuses on a strict and practically meaningful regime: zero-shot cross-dataset UAV detection. In this setting, the detector is trained only on a source dataset and is directly evaluated on a target UAV dataset without target-domain fine-tuning, prompt tuning, or target-domain training images. Although YOLO-World v2 [9] provides a strong real-time OVOD baseline by integrating language features into multi-scale visual representations and enabling prompt-driven inference with an offline vocabulary, its zero-shot transfer to UAV benchmarks remains fragile, especially for tiny objects.

The central observation of the present work is that the transfer problem is not determined only by cross-modal alignment but also by how visual evidence is formed and preserved before the text-conditioned head is applied. In particular, stride-2 downsampling is revisited as a transfer-critical operator. Standard stride-2 convolution performs aggressive resolution reduction and commits to a single sampling origin, which can induce aliasing-driven detail loss and sensitivity to sampling phase under small spatial shifts [10,11]. In UAV imagery, where many targets occupy only a few pixels, the loss of weak fine-grained cues at early reductions can irreversibly damage downstream multi-scale fusion and open-vocabulary matching. This work therefore treats stride-2 reduction as a key bottleneck for zero-shot transferable OV-UAV detection.

To address this issue, AeroPinWorld is proposed as a transfer-oriented variant of YOLO-World v2. Instead of globally redesigning the detector, AeroPinWorld selectively introduces pinwheel-shaped convolution (PConv) [12] as a phase-aware reduction operator at four transfer-critical stride-2 transitions: the first two backbone reductions (P1/2 and P2/4) and the two bottom–up stride-2 reductions in the head. This selective design preserves efficiency while strengthening local structure aggregation at the stages where tiny-object evidence is most vulnerable under aerial domain shift.

The main contributions of this study are summarized as follows:

- A strict zero-shot cross-dataset open-vocabulary UAV detection setting is studied, and stride-2 downsampling is identified as a transfer-critical bottleneck that disproportionately harms tiny-object representations under aerial domain shift.
- AeroPinWorld is proposed as a pinwheel-augmented open-vocabulary UAV detector for zero-shot cross-dataset transfer, which introduces PConv as a phase-aware stride-2 reduction operator via selective insertion at transfer-critical pyramid transitions.
- Extensive experiments on VisDrone2019-DET and UAVDT demonstrate consistent zero-shot transfer improvements over YOLO-World v2-S, together with a favorable accuracy–efficiency trade-off.

The remainder of this paper is organized as follows. Section 2 reviews related work on open-vocabulary detection, UAV detection under domain shift, and downsampling stability. A nomenclature table is then provided before Section 3 to clarify the main symbols used in the methodology. Section 3 presents the proposed AeroPinWorld framework, including the phase-aware stride-2 block and the selective integration strategy. Section 4 describes the datasets, protocol, implementation details, quantitative and qualitative results, ablation studies, sensitivity analysis, efficiency evaluation, and further discussion. Section 5 concludes the paper.

2. Related Works

2.1. Open-Vocabulary Detection and Zero-Shot Transfer

Traditional object detectors are typically trained under a *closed-set* assumption with a fixed label space, which limits their ability to recognize categories beyond the training vocabulary. Recent OVOD methods alleviate this restriction by aligning region-level visual representations with text embeddings from large-scale vision–language pre-training. Representative directions include region–text distillation and vocabulary expansion (e.g., ViLD and Detic [13,14]), phrase-grounding based pre-training (e.g., GLIP and GLIPv2 [15,16]), and transformer-based open-vocabulary detectors (e.g., OWL-ViT and Grounding DINO [17,18]). Among real-time OVOD methods, YOLO-World [9] emphasizes prompt-driven deployment with an offline vocabulary and provides a practical baseline for strict zero-shot evaluation.

The comparison in Table 1 indicates that existing OVOD studies mainly focus on semantic openness and vision–language alignment. By contrast, the present work investigates whether the *visual formation process itself*, especially stride-2 reduction, limits zero-shot transfer from natural-image corpora to UAV imagery.

Table 1. Representative open-vocabulary detection methods and their relevance to zero-shot UAV transfer.

Method Family	Representative Work	Main Strength	Remaining Limitation for Zero-Shot UAV Transfer
Region–text distillation	ViLD, Detic [13,14]	Expand category coverage beyond the training label space through language supervision or image-level labels	Strengthen semantic openness but do not explicitly address aerial domain shift or tiny-object feature fragility
Grounded pre-training	GLIP, GLIPv2 [15,16]	Improve phrase–region alignment with large-scale grounded language–image supervision	Strong grounding ability but not designed around real-time UAV deployment or transfer-critical reduction operators
Transformer-based OVOD	OWL-ViT, Grounding DINO [17,18]	Provide flexible open-vocabulary localization with strong zero-shot generality	Effective for open-set localization, but computationally heavy and not targeted at UAV-specific transfer stability
Real-time OVOD	YOLO-World/ YOLO-World v2 [9]	Enable prompt-driven real-time detection with online/offline vocabulary modes	Efficient and practical, but standard stride-2 reduction may still weaken tiny-object transfer under aerial domain shift

2.2. UAV Small-Object Detection Under Domain Shift

UAV object detection remains difficult because of dense layouts, severe scale variation, complex backgrounds, and the dominance of small targets [5,19]. Recent UAV-oriented detectors therefore strengthen multi-scale fusion, attention interaction, frequency-domain modeling, or lightweight neck redesign in order to improve small-object robustness and aerial generalization [8,20–23].

Table 2 suggests that UAV detection studies mainly improve feature fusion, lightweight design, or domain robustness. However, the specific role of stride-2 reduction in zero-shot open-vocabulary UAV transfer remains underexplored.

Table 2. Representative UAV-related detection studies and their remaining gaps relative to this work.

Representative Work	Main Focus	Main Design	Remaining Gap Relative to This Study
Frequency-domain disentanglement [20]	UAV domain generalization	Improve robustness by disentangling frequency-domain cues for UAV detection	Addresses domain shift but not open-vocabulary detection or selective stride-2 redesign
YOLO-Drone [21]	Tiny-object enhancement	Refine YOLO-style multi-scale representations for tiny UAV targets	Improves in-domain UAV detection but remains closed-set and does not study strict zero-shot transfer
Improved YOLOv7 UAV detector [8]	UAV-specific optimization	Strengthen multi-scale representation and improve aerial detection robustness	Focuses on in-domain UAV detection rather than source-only zero-shot transfer
RT-DETR drone detector [22]	Multi-scale interaction	Improve aerial detection with bidirectional feature attention fusion	Strengthens feature fusion but does not isolate downsampling as a transfer bottleneck
Lightweight UAV small-object detector [23]	Efficient UAV small-object detection	Improve upsampling and cross-modal feature fusion	Efficient and UAV-oriented but assumes target-domain training and does not analyze transfer-critical reduction operators

2.3. Downsampling, Aliasing, and Phase-Aware Reduction

Downsampling operators are essential in modern detectors, but they are not neutral with respect to transfer. Prior studies show that conventional convolutional networks can be sensitive to small translations and sampling shifts [11,24]. Anti-aliasing and shift-invariant designs have therefore been proposed to improve spatial stability [10]. More recently, phase-aware reduction modules such as PConv [12] have shown that directional kernels and asymmetric padding can preserve weak local structures more effectively during downsampling. Table 3 summarizes representative studies on downsampling stability and clarifies how they motivate the phase-aware reduction design adopted in AeroPinWorld.

Table 3. Representative studies on downsampling stability and their relevance to AeroPinWorld.

Representative Work	Main Concern	Key Idea	Relevance to This Study
Small image transformation sensitivity [11]	Sensitivity to small image transformations	Show that CNNs can generalize poorly under small spatial perturbations	Motivates the need to reduce phase sensitivity during cross-dataset transfer
Anti-aliased shift-invariant CNNs [10]	Shift invariance	Introduce anti-aliased reduction to recover shift robustness	Shows that the design of downsampling operators strongly influences representation stability
Translation invariance in CNNs [24]	Translation invariance of CNNs	Analyze why convolutional layers can still exploit absolute spatial location	Suggests that standard fixed-alignment convolution is not inherently phase-robust
PConv [12]	Weak local target preservation	Use asymmetric padding and directional kernels to probe multiple offset hypotheses before fusion	Provides a lightweight operator suitable for transfer-critical stride-2 redesign

In summary, existing OVOD methods mainly improve vocabulary generalization and vision–language alignment, UAV studies mainly strengthen multi-scale robustness under aerial scenes, and downsampling studies show that reduction operators affect spatial stability. However, the role of stride-2 downsampling as a *transfer-critical bottleneck* in zero-shot open-vocabulary UAV detection remains underexplored. This gap directly motivates the design of AeroPinWorld. For clarity, the main symbols used in Section 3 are summarized in Table 4.

Table 4. Main symbols and notations used in Section 3.

Symbol	Meaning
X	Input feature map before stride-2 reduction
Y	Output feature map after stride-2 reduction
B	Batch size
C	Number of input channels of X
C'	Number of output channels of Y
H, W	Spatial height and width of the input feature map
k	Kernel size used in the phase-aware reduction block
$\mathcal{D}$	Standard stride-2 reduction operator
$\mathcal{P}$	Proposed PConv-based phase-aware stride-2 reduction operator
$\mathcal{T}$	Set of transfer-critical stride-2 transitions to be replaced
λ_{box}	Weight of the IoU-based localization term
λ_{cls}	Weight of the classification term
λ_{dfl}	Weight of the Distribution Focal Loss term

3. Method

3.1. Overview: Revisiting Stride-2 Downsampling for Transferable OV-UAV Detection

AeroPinWorld is a transfer-oriented variant of YOLO-World v2 for open-vocabulary UAV detection. As illustrated in Figure 1, the overall prompt-driven detection pipeline and the vision–language fusion design are preserved, while stride-2 downsampling is revisited as a transfer-critical design axis for zero-shot cross-dataset transfer.

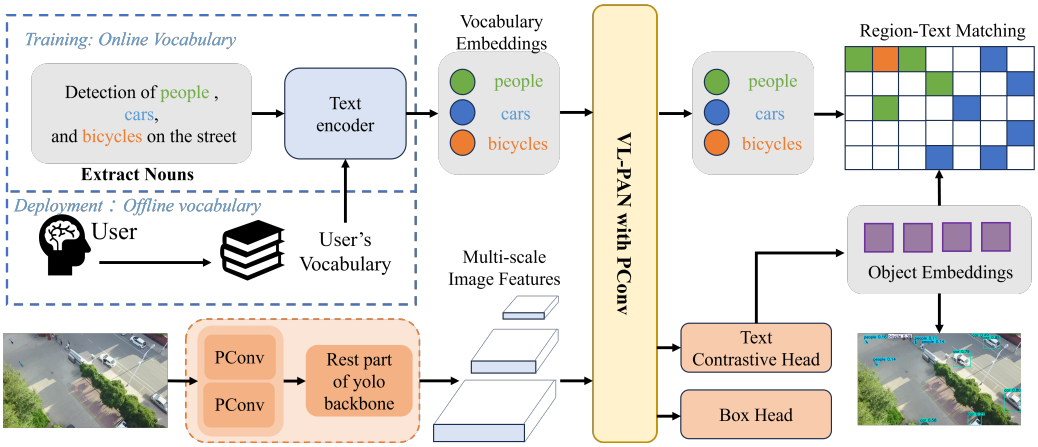


Figure 1. Overall architecture of *AeroPinWorld*. The detector preserves the prompt-driven YOLO-World v2 pipeline and selectively replaces four transfer-critical stride-2 reductions with PConv at P1/2, P2/4, P3–P4, and P4–P5. The resulting multi-scale features are then fed into the unchanged text contrastive head and box regression head for open-vocabulary detection.

Two architectural decisions define the method. First, selected stride-2 reduction operators are replaced with a phase-aware PConv-based block, described in Section 3.3. Second, these replacements are intentionally restricted to four transfer-critical transitions, described in Section 3.4 so that transfer robustness is improved without globally redesigning the detector.

3.2. Revisiting Stride-2 Reduction as a Transfer-Critical Operator

Let $X \in \mathbb{R}^{B\times C\times H\times W}$ denote an intermediate feature map, where B is the batch size, C is the number of input channels, and $H \times W$ denotes the spatial resolution. A standard stride-2 reduction maps X to an output feature map $Y \in \mathbb{R}^{B\times C'\times \frac{H}{2}\times \frac{W}{2}}$, where C' denotes the number of output channels after reduction.

In conventional detectors, this operation is efficient because sampling and aggregation are performed jointly. Under aerial domain shift, however, it is also risky: tiny targets

may occupy only a few pixels, and a fixed sampling origin can amplify aliasing and phase sensitivity. Once early weak cues are attenuated, later pyramid fusion and text-conditioned matching can no longer fully recover them. This is especially harmful in zero-shot transfer, where the target-domain appearance already differs from the source training distribution.

AeroPinWorld therefore treats stride-2 reduction as a transfer-critical operator. Rather than modifying the open-vocabulary head, the proposed method improves the stability of the multi-scale features *before* they are consumed by the prompt-conditioned prediction head.

3.3. PConv-Based Phase-Aware Stride-2 Block

The proposed reduction block adopts the plug-and-play PConv module [12] as a phase-aware stride-2 operator. Given $X \in \mathbb{R}^{B\times C\times H\times W}$, PConv first constructs multiple offset probes through asymmetric padding and directional kernels, and then fuses them into a downsampled feature.

3.3.1. Offset Probing with Asymmetric Padding

Let $\mathrm{Pad}(\cdot; l, r, t, b)$ denote zero-padding with (l, r, t, b) pixels on the left, right, top, and bottom. With kernel size k (set to 3 in this work), PConv forms four complementary probes:

$$\begin{aligned} Y_{w0} &= \mathrm{Conv}^{s=2}_{1\times k}(\mathrm{Pad}(X;\, k, 0, 1, 0)), \\ Y_{w1} &= \mathrm{Conv}^{s=2}_{1\times k}(\mathrm{Pad}(X;\, 0, k, 0, 1)), \\ Y_{h0} &= \mathrm{Conv}^{s=2}_{k\times 1}(\mathrm{Pad}(X;\, 0, 1, k, 0)), \\ Y_{h1} &= \mathrm{Conv}^{s=2}_{k\times 1}(\mathrm{Pad}(X;\, 1, 0, 0, k)). \end{aligned} \tag{1}$$

3.3.2. Channel Fusion into a Coarse-Grid Feature

The four branch outputs are concatenated and fused by a lightweight convolution:

$$Y \;=\; \phi(\mathrm{Concat}[Y_{w0}, Y_{w1}, Y_{h0}, Y_{h1}]), \tag{2}$$

where $\phi(\cdot)$ is implemented as a small convolution that mixes offset-specific evidence and outputs the stride-2 feature Y.

3.3.3. Why This Helps Transfer

By exposing multiple sub-sampling alignments before reduction, PConv yields a more stable coarse-grid representation that is less sensitive to the sampling phase and more robust to domain-specific texture or appearance shifts. This property is particularly important when transferring from generic natural-image corpora to UAV imagery dominated by tiny, cluttered targets.

3.4. Selective Integration into YOLO-World v2

AeroPinWorld integrates the proposed phase-aware stride-2 block *selectively* in order to maximize transfer gains under a constrained budget. Rather than replacing every downsampling layer, the method targets the transitions where fine-detail survival and bottom–up pyramid consistency matter most for tiny-object transfer.

Let $\mathcal{D}$ denote a standard stride-2 reduction operator and $\mathcal{P}$ denote the proposed phase-aware stride-2 block. AeroPinWorld replaces only a small set of transfer-critical transitions:

$$\mathcal{T} \;=\; \{\mathrm{P1/2},\ \mathrm{P2/4},\ \mathrm{P3–P4},\ \mathrm{P4–P5}\}, \tag{3}$$

where P1/2 and P2/4 refer to the first two backbone reductions, and P3–P4 and P4–P5 refer to the two bottom–up stride-2 reductions in the detection head.

Table 5 summarizes the rationale for each replacement. The first two backbone reductions operate on the highest-resolution internal maps and are therefore most sensitive to the disappearance of weak high-frequency cues. The two head reductions are responsible for bottom–up pyramid construction and directly influence the consistency of coarse-level features under domain shift. Later backbone reductions are kept unchanged to avoid unnecessary computational overhead.

Table 5. Transfer-critical stride-2 replacements used in AeroPinWorld.

Location	Baseline Operator	Functional Role	Design Rationale
P1/2	Standard stride-2 Conv	First backbone reduction	Highest-resolution internal map; weak tiny-object evidence is most vulnerable here
P2/4	Standard stride-2 Conv	Second backbone reduction	Consolidates fine-grained local structure before later pyramid formation
P3–P4	Bottom–up stride-2 Conv in head	Coarse feature construction	Stabilizes bottom–up multi-scale aggregation under domain shift
P4–P5	Bottom–up stride-2 Conv in head	High-level pyramid transition	Improves coarse-level consistency while keeping overhead limited

3.5. Open-Vocabulary Detection Head

AeroPinWorld inherits the open-vocabulary detection pipeline of YOLO-World [9]. Given an input image and a set of text prompts, a text encoder maps prompts into text embeddings, which are fused with multi-scale image features through a vision–language aggregation module. The detection head then outputs bounding box predictions and region-level logits for prompt-conditioned classification.

The head design is intentionally kept unchanged for two reasons. First, the goal of this work is to isolate the contribution of transfer-critical stride-2 redesign; modifying the head would introduce a confounding architectural factor. Second, retaining the original head preserves compatibility with the prompt-driven YOLO-World pipeline and keeps the efficiency profile close to the baseline.

3.6. Optimization Objective

The training objective follows the baseline YOLO-World implementation and combines a classification term with IoU-based regression and Distribution Focal Loss (DFL) [25,26]. The formulation is retained here for completeness and to support the later discussion of the loss-balance hyperparameters; the loss design itself is not changed by AeroPinWorld.

3.6.1. Assignment and Normalization

A task-aligned assigner produces a foreground mask $\mathcal{F}$, assigned target boxes $\{b_i^\star\}$, and soft target scores $\{y_{i,c}\}$ for each anchor i and class c. We denote the normalization factor by

$$S = \max\left(\sum_i \sum_c y_{i,c}, 1\right), \tag{4}$$

which normalizes all loss components by the total target-score mass.

3.6.2. Classification Loss $\mathcal{L}_{\text{con}}$

Let $z_{i,c}$ be the predicted logit for class c at anchor i, and $y_{i,c} \in [0,1]$ be the corresponding soft target score. We use the element-wise BCEWithLogits loss and normalize by S:

$$\mathcal{L}_{\text{con}} = \frac{1}{S} \sum_i \sum_c \Big(-y_{i,c} \log \sigma(z_{i,c}) - (1 - y_{i,c}) \log\big(1 - \sigma(z_{i,c})\big) \Big), \tag{5}$$

where $\sigma(\cdot)$ denotes the sigmoid function.

3.6.3. IoU Loss $\mathcal{L}_{\text{iou}}$ (CIoU)

For each positive anchor $i \in \mathcal{F}$, let b_i denote the decoded predicted box and $b_i^\star$ the assigned target box. We use Complete IoU (CIoU) [26] and define

$$\mathcal{L}_{\text{iou}} = \frac{1}{S} \sum_{i \in \mathcal{F}} w_i (1 - \text{CIoU}(b_i, b_i^\star)), \qquad w_i = \sum_c y_{i,c}, \tag{6}$$

where w_i weights each positive anchor by its aggregated target score mass.

3.6.4. Distribution Focal Loss $\mathcal{L}_{\text{dfl}}$

A box is parameterized by four distances from an anchor point, $\mathbf{t}_i = (l_i, t_i, r_i, b_i)$. For each side $s \in \{l, t, r, b\}$, the model predicts a distribution over M bins ($M =$ `reg_max`). Let $\mathbf{p}_{i,s} \in \mathbb{R}^M$ denote the predicted logits and $t_{i,s}$ the continuous target distance. We clamp the target to $[0, M - 1 - \epsilon)$ and define

$$\tilde{t}_{i,s} = \text{clip}(t_{i,s},\, 0,\, M - 1 - \epsilon), \quad t_{i,s}^l = \lfloor \tilde{t}_{i,s} \rfloor, \quad t_{i,s}^r = t_{i,s}^l + 1, \tag{7}$$

$$\omega_{i,s}^l = t_{i,s}^r - \tilde{t}_{i,s}, \qquad \omega_{i,s}^r = 1 - \omega_{i,s}^l. \tag{8}$$

The per-side DFL is then written as

$$\ell_{\text{dfl}}(i, s) = \omega_{i,s}^l \cdot \text{CE}(\mathbf{p}_{i,s}, t_{i,s}^l) + \omega_{i,s}^r \cdot \text{CE}(\mathbf{p}_{i,s}, t_{i,s}^r), \tag{9}$$

and the normalized DFL term becomes

$$\mathcal{L}_{\text{dfl}} = \frac{1}{S} \sum_{i \in \mathcal{F}} w_i \cdot \frac{1}{4} \sum_{s \in \{l,t,r,b\}} \ell_{\text{dfl}}(i, s). \tag{10}$$

Total Loss

Finally, the three components are combined with loss gains:

$$\mathcal{L} = \lambda_{\text{box}}\, \mathcal{L}_{\text{iou}} + \lambda_{\text{cls}}\, \mathcal{L}_{\text{con}} + \lambda_{\text{dfl}}\, \mathcal{L}_{\text{dfl}}, \tag{11}$$

where $\lambda_{\text{box}}, \lambda_{\text{cls}}, \lambda_{\text{dfl}}$ are scalar weights (hyperparameters). Since AeroPinWorld keeps the same detection head and objective as the baseline, performance differences can be more directly attributed to the enhanced transferability of multi-scale features induced by the selective stride-2 redesign.

4. Experiments

4.1. Datasets

4.1.1. Source Dataset (Training)

All models are trained on COCO2017 [27] as the source domain, following the standard YOLO-World v2 training pipeline.

4.1.2. Target Datasets (Zero-Shot Evaluation)

Evaluation is conducted under a strict cross-dataset zero-shot setting on two UAV benchmarks: VisDrone2019-DET [19], using the validation split as the primary target dataset, and UAVDT [5] as a complementary vehicle-centric benchmark. No target-domain images or annotations are used for training, adaptation, or fine-tuning.

4.2. Zero-Shot Transfer Protocol and Prompt Vocabulary

A train-on-source, test-on-target protocol is adopted. The detector is trained only on COCO2017 and then directly evaluated on the target UAV datasets without any target-domain fine-tuning, prompt tuning, or test-time adaptation. This protocol isolates transferability under aerial domain shift.

Prompt Vocabulary

Dataset-aligned prompt sets are used during evaluation.

VisDrone2019-DET (10 classes).

$$\mathcal{P}_{\text{VisDrone}} = \{\texttt{pedestrian}, \texttt{people}, \texttt{bicycle}, \texttt{car}, \texttt{van}, \texttt{truck}, \texttt{tricycle}, \texttt{awning tricycle}, \texttt{bus}, \texttt{motorcycle}\}. \tag{12}$$

UAVDT (3 classes).

$$\mathcal{P}_{\text{UAVDT}} = \{\texttt{car}, \texttt{truck}, \texttt{bus}\}. \tag{13}$$

4.3. Implementation Details

To ensure a fair comparison, the baseline and AeroPinWorld share the same source dataset, initialization, training schedule, optimizer, augmentation policy, prompt-vocabulary mode, and evaluation protocol; only the transfer-critical stride-2 operators are changed. Table 6 summarizes the exact configuration used in the experiments. This controlled setup is important because it ensures that any observed performance difference can be attributed to the architectural modification rather than to changes in optimization or evaluation conditions. In particular, the benchmark-specific evaluation resolutions are kept fixed at 1280 for VisDrone and 640 for UAVDT to match the target settings, while the original YOLO-World v2 loss gains are retained so that the architectural analysis is not confounded by task-specific loss re-tuning.

Table 6. Exact training and inference configuration.

Item	Setting
Source training dataset	COCO2017
Target evaluation datasets	VisDrone2019-DET (val), UAVDT
Initialization	Official YOLO-World v2-S pretrained weights
Training schedule	24 epochs
Training image size	640
VisDrone evaluation image size	1280
UAVDT evaluation image size	640
Batch size	16
Optimizer	SGD
Initial learning rate (lr_0)	0.001
Momentum	0.937
Weight decay	0.0005
Warm-up epochs	2.0
Loss weights	$\lambda_{\text{box}} = 7.5$, $\lambda_{\text{cls}} = 0.5$, $\lambda_{\text{dfl}} = 1.5$
Confidence threshold	0.001
NMS IoU threshold	0.7
Workers	0
Random seed	0
Augmentation policy	Standard YOLO-World v2 baseline augmentation recipe; same for baseline and AeroPinWorld
Vocabulary mode	Offline prompt vocabulary
Software environment	Python 3.12.12, PyTorch 2.6.0, CUDA 12.4, and the official YOLO-World codebase (https://github.com/AILab-CVC/YOLO-World, accessed on 22 March 2026)
Hardware	Intel Core Ultra 5 225H CPU (Intel, Santa Clara, CA, USA) and NVIDIA GeForce RTX 3090 Ti GPU (NVIDIA, Santa Clara, CA, USA)

4.4. Baselines and Compared Methods

The following methods are compared under the same protocol: (i) YOLO-World v2-S, which uses standard stride-2 operators; and (ii) AeroPinWorld-S, which selectively replaces four transfer-critical stride-2 transitions with PConv. Representative general and UAV-related detectors are also listed in the main result tables in order to contextualize the transfer performance.

4.5. Evaluation Metrics

The reported metrics include COCO-style box mAP, AP_{50}, and AP_{75}. Since UAV scenes are dominated by small objects, AP_S is also reported. Model efficiency is analyzed using parameter count, FLOPs, and runtime.

4.6. Main Results

4.6.1. Zero-Shot Transfer on VisDrone2019-DET (*val*, 1280)

Table 7 reports the strict train-on-COCO/test-on-VisDrone results. AeroPinWorld-S consistently improves YOLO-World v2-S across all reported metrics: mAP rises from **0.112** to **0.135** (+0.023), AP_{50} rises from **0.159** to **0.197** (+0.038), AP_{75} rises from **0.098** to **0.118** (+0.020), and AP_S rises from **0.054** to **0.063** (+0.009). The gains on both AP_{50} and AP_{75} indicate that revisiting a small set of stride-2 junctions improves both detection coverage and localization quality under aerial domain shift, while the AP_S gain confirms stronger robustness on tiny objects.

Table 7. Zero-shot transfer results on VisDrone2019-DET (*val*, input size 1280). The best results in each metric column are shown in bold.

Method	mAP	AP_{50}	AP_{75}	AP_S	Params (M)	FLOPs (G)
CAGE [28]	0.105	0.156	0.106	0.048	17.4	36.3
PMG-SAM [29]	0.109	0.157	0.110	0.051	136.0	296.2
SAM [30]	0.105	0.159	0.107	0.048	93.7	233.5
Faster R-CNN [31]	0.111	0.168	0.111	0.051	41.1	133.6
RetinaNet [32]	0.123	0.179	0.112	0.056	36.5	129.0
FCOS [33]	0.108	0.161	0.109	0.049	32.1	125.3
DINO [34]	0.129	0.183	0.111	0.051	47.6	146.0
ET-FSM [35]	0.131	0.192	0.114	0.052	41.1	133.6
YOLO-World [9]	0.126	0.179	0.109	0.056	46.8	180.4
YOLO-World+UMM [36]	0.128	0.182	0.113	0.057	79.3	276.0
YOLO-World v2-S [9]	0.112	0.159	0.098	0.054	16.4	34.2
AeroPinWorld-S (ours)	**0.135**	**0.197**	**0.118**	**0.063**	16.3	35.7

4.6.2. Comparison with Existing Baselines in Table 7

AeroPinWorld-S achieves the best mAP/AP_{50}/AP_{75} among the listed methods while remaining substantially more compact than many alternatives. For example, it outperforms ET-FSM (0.131 mAP) and DINO (0.129 mAP) while operating at only **16.3 M** parameters. This suggests that selectively strengthening transfer-critical stride-2 reductions can yield a better transfer gain–efficiency trade-off than merely increasing model capacity.

4.6.3. Zero-Shot Transfer on UAVDT (640)

Table 8 shows that AeroPinWorld also improves the baseline on UAVDT, increasing mAP from **0.144** to **0.146**, AP_{50} from **0.265** to **0.270**, AP_{75} from **0.140** to **0.148**, and AP_S from **0.060** to **0.068**. Although the absolute mAP gain is smaller than on VisDrone, the larger gains on AP_{75} and AP_S indicate improved localization and stronger robustness on small or distant vehicles.

Table 8. Zero-shot transfer results on UAVDT (input size 640, vehicle prompts: `car/truck/bus`, "s" model scale). The best results in each metric column are shown in bold.

Method	mAP	AP_{50}	AP_{75}	AP_S
CAGE [28]	0.135	0.259	0.134	0.056
PMG-SAM [29]	0.139	0.263	0.141	0.058
SAM [30]	0.137	0.258	0.138	0.056
Faster R-CNN [31]	0.142	0.265	0.136	0.059
RetinaNet [32]	0.139	0.255	0.142	0.060
FCOS [33]	0.141	0.263	0.141	0.058
DINO [34]	0.143	0.261	0.134	0.061
ET-FSM [35]	0.138	0.254	0.138	0.058
YOLO-World [9]	0.145	0.263	0.141	0.063
YOLO-World+UMM [36]	0.144	0.265	0.143	0.059
YOLO-World v2 [9]	0.144	0.265	0.140	0.060
AeroPinWorld (ours)	**0.146**	**0.270**	**0.148**	**0.068**

4.7. Ablation Studies

To understand where the transfer gains come from, ablations are conducted by selectively enabling the proposed phase-aware stride-2 block at different stages. "B" denotes the first two backbone reductions (P1/2 and P2/4), and "H" denotes the two head bottom–up reductions (P3→P4 and P4→P5). The results are reported on VisDrone in Table 9.

Table 9. Ablation on VisDrone2019-DET (*val*, input size 1280). "B" denotes early backbone reductions, and "H" denotes head bottom–up reductions.

Setting	mAP	AP_{50}	AP_{75}	AP_S	Params (M)	FLOPs (G)
Baseline (YOLO-World v2)	0.112	0.159	0.098	0.054	16.4027	34.2
Only B (P1/2 + P2/4)	0.124	0.171	0.109	0.057	16.4031	35.3
Only H (P3→P4 + P4→P5)	0.129	0.182	0.111	0.059	16.3745	35.6
B+H (Full AeroPinWorld)	0.135	0.197	0.118	0.063	16.3745	35.7

Both components provide measurable gains, and their effects are complementary. Replacing only the early backbone reductions already improves mAP and AP_S, suggesting that preserving weak details at the highest-resolution internal maps is beneficial. Replacing only the head bottom–up reductions yields a stronger gain, indicating that stabilizing pyramid construction under domain shift is also critical. Combining both provides the best overall performance.

4.8. Qualitative Results

To complement the quantitative results, qualitative comparisons on VisDrone2019-DET (*val*) are shown in Figure 2. The displayed scenes are representative examples containing dense tiny objects and cluttered backgrounds.

As shown in Figure 2, AeroPinWorld yields more complete coverage of small and distant targets while suppressing false positives in cluttered regions. The highlighted regions show two typical baseline failure modes in UAV imagery: missed detections caused by weak fine-grained cues and spurious responses triggered by repetitive background structures. By selectively strengthening transfer-critical stride-2 junctions, AeroPinWorld preserves more stable local evidence after resolution reduction, which improves the robustness of downstream multi-scale fusion.

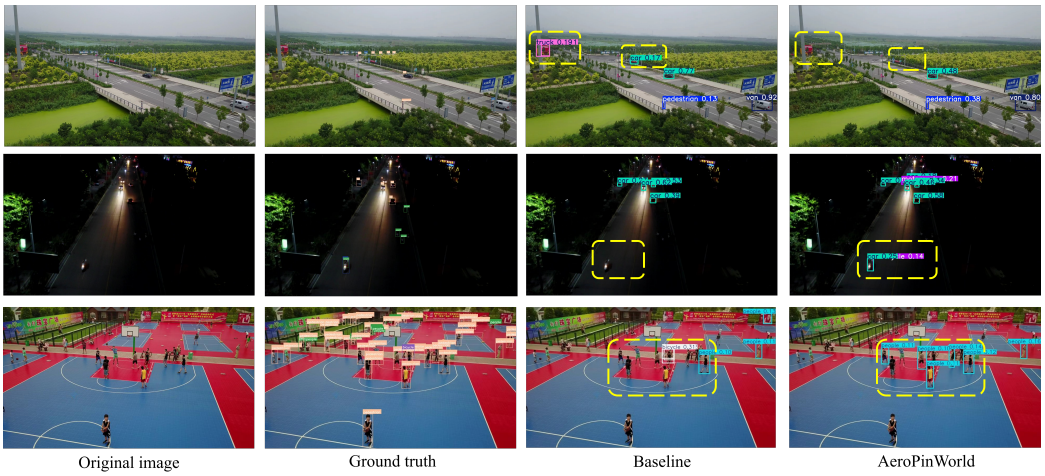


Figure 2. Qualitative comparison on VisDrone2019-DET (*val*). The panels present the input sample, the ground-truth annotations, predictions from the baseline, and predictions from AeroPinWorld. Yellow dashed regions highlight representative cases where AeroPinWorld reduces both missed detections and false positives.

4.9. Efficiency and Runtime Analysis

AeroPinWorld is designed to improve transfer robustness without substantially increasing computational burden. As shown in Tables 7 and 9, the parameter count remains nearly unchanged, while FLOPs increase from 34.2G to 35.7G for the full model on the VisDrone setting, corresponding to an absolute increase of 1.5G and a relative increase of only about 4.4%. This overhead remains limited because only four local stride-2 transitions are replaced: two early backbone reductions and two bottom–up reductions in the head. The later high-channel backbone stages, the text-conditioned fusion pipeline, and the original detection head are left unchanged. The ablation results further show that partial replacement incurs 35.3G FLOPs for "Only B" and 35.6G FLOPs for "Only H", while the full model reaches 35.7G, confirming that the complexity increase is bounded and arises from the selected transfer-critical locations rather than from a global redesign.

Table 10 further shows that the runtime overhead remains moderate. At 1280 resolution, AeroPinWorld introduces a small latency increase relative to the baseline, which is consistent with the modest FLOPs growth. At 640 resolution, the measured inference speed remains competitive. Overall, these results support the claim that transfer gains are obtained through a highly selective architectural revision rather than through a heavy global redesign.

Table 10. Runtime comparison measured with batch size 1 on an RTX 3090 Ti GPU after warm-up.

Method	Resolution	Latency (ms/Image)	FPS	Note
YOLO-World v2-S	1280	25.86	38.67	VisDrone setting
AeroPinWorld-S	1280	28.24	35.41	VisDrone setting
YOLO-World v2-S	640	20.50	48.78	UAVDT setting
AeroPinWorld-S	640	18.09	55.28	UAVDT setting

4.10. Further Discussion

4.10.1. Effect of the Loss-Balance Hyperparameters in Equation (11): A Baseline Control Study

Equation (11) contains three scalar weights, namely λ_{box}, λ_{cls}, and λ_{dfl}, which balance localization, classification, and distribution-based box regression. Increasing λ_{cls} generally strengthens category discrimination, but an excessively large value can shift optimization away from precise localization. Increasing λ_{dfl} typically sharpens the distance

estimation and can improve the localization quality, but over-emphasizing it may suppress classification learning or reduce optimization stability.

Since AeroPinWorld intentionally keeps the same detection head and loss design as YOLO-World v2-S, the effect of these hyperparameters is examined here through a control sensitivity analysis on the baseline configuration. The purpose of this control study is not to claim additional tuning gains but to verify that the default loss weighting inherited from YOLO-World v2-S already provides a reasonable classification–localization balance. Around the default setting $(\lambda_{\text{cls}}, \lambda_{\text{dfl}}) = (0.5, 1.5)$, all tested perturbations lead to slight mAP degradation.

As shown in Table 11, decreasing or increasing either λ_{cls} or λ_{dfl} in the baseline causes a consistent drop in performance. In practical terms, larger λ_{cls} emphasizes prompt-conditioned category discrimination, whereas larger λ_{dfl} emphasizes distribution-based box regression. The control result indicates that the default baseline weighting remains the most balanced choice among the tested settings. AeroPinWorld therefore keeps the same loss gains so that the improvements reported in Tables 7–9 can be attributed to the selective stride-2 redesign rather than to favorable loss re-weighting.

Table 11. Control sensitivity analysis of the loss-balance hyperparameters on the baseline YOLO-World v2-S model on VisDrone2019-DET (*val*, input size 1280). Since only the loss weights are changed while the architecture remains unchanged, the inference latency is identical across all settings.

Setting	Latency (ms/Image)	mAP	AP_{50}
Default $(\lambda_{\text{cls}}, \lambda_{\text{dfl}})$	25.86	0.1122	0.159
$0.5\lambda_{\text{cls}}, \lambda_{\text{dfl}}$	25.86	0.0997	0.149
$1.5\lambda_{\text{cls}}, \lambda_{\text{dfl}}$	25.86	0.0993	0.148
$\lambda_{\text{cls}}, 0.5\lambda_{\text{dfl}}$	25.86	0.0994	0.148
$\lambda_{\text{cls}}, 1.5\lambda_{\text{dfl}}$	25.86	0.0986	0.147

4.10.2. Why the Detection Head Is Kept Unchanged

The detection head is intentionally preserved for methodological clarity. The purpose of this work is to determine whether transfer-critical stride-2 reduction itself acts as a bottleneck for zero-shot aerial transfer. If the head were redesigned together with the downsampling operators, the source of the performance gain would become ambiguous. Keeping the head unchanged therefore isolates the contribution of the proposed stride-2 redesign and preserves compatibility with the prompt-driven YOLO-World pipeline.

4.10.3. Why the Extra FLOPs Remain Modest

The modest computational overhead is a direct consequence of the selective design. Only the first two backbone reductions and the two bottom–up head reductions are replaced, whereas later backbone stages with greater cumulative computational contribution remain unchanged. Moreover, the adopted PConv module is lightweight and introduces only local directional probing followed by a small fusion step. This explains why AeroPinWorld improves transfer performance while maintaining a competitive efficiency profile.

4.10.4. Limitations

The present study has several limitations. First, evaluation is currently limited to two UAV benchmarks and one small model scale. Second, the work focuses on architectural transferability and does not yet investigate prompt engineering, vocabulary optimization, or target-domain adaptation. Third, the runtime analysis is conducted on a desktop GPU rather than on embedded UAV hardware. Finally, only four manually selected transfer-critical locations are investigated; a broader search over candidate replacement locations may further improve the accuracy–efficiency trade-off.

5. Conclusions

AeroPinWorld is presented as a transfer-oriented redesign of YOLO-World v2 for zero-shot open-vocabulary UAV detection. The central idea is to revisit stride-2 downsampling as a transfer-critical bottleneck rather than treating it as a fixed implementation detail. By selectively replacing four stride-2 transitions with a PConv-based phase-aware reduction block, AeroPinWorld improves the stability of multi-scale feature formation under aerial domain shift while keeping the original detection head unchanged.

Experiments on VisDrone2019-DET and UAVDT demonstrate consistent gains over YOLO-World v2-S under a strict train-on-COCO/test-on-target protocol. In particular, the proposed design improves both general detection performance and tiny-object robustness, while introducing only a modest increase in computational cost. Ablation and discussion results further show that both early backbone reductions and head bottom–up reductions contribute to the final gain.

Future work will focus on extending the proposed idea to more model scales and UAV benchmarks, exploring prompt and vocabulary design for long-tail aerial categories, and validating the efficiency of AeroPinWorld on embedded UAV hardware.

Author Contributions: Conceptualization, J.L. and M.G.; Methodology, J.L. and M.G.; Software, J.L. and J.X.; Validation, X.D., H.C. and J.X.; Formal analysis, J.L. and J.X.; Investigation, X.D., H.C. and M.G.; Resources, Y.L.; Data curation, X.D. and H.C.; Writing—original draft preparation, J.L. and J.X.; Writing—review and editing, M.G., X.D., H.C. and Y.L.; Visualization, J.L.; Supervision, Y.L.; Project administration, M.G.; Funding acquisition, Y.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Natural Science Foundation of China under Grant 62376207, in part by the Open Research Fund of Anhui Province Key Laboratory of Machine Vision Detection and Perception (Grant No. KLMVI-2025-HIT-13), in part by the Open Project of Anhui Province Key Laboratory of Special and Heavy Load Robot (Grant No. TZJQR004-2026), and in part by the Fundamental Research Funds for the Central Universities, China.

Data Availability Statement: Publicly available datasets were analyzed in this study. The COCO2017 dataset is available at https://cocodataset.org/ (accessed on 22 March 2026). The VisDrone2019-DET dataset can be accessed at https://github.com/VisDrone/VisDrone-Dataset (accessed on 22 March 2026). The UAVDT dataset is available at https://sites.google.com/view/grli-uavdt/ (accessed on 22 March 2026).

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

References

1. Pham, C.; Vu, T.; Nguyen, K. LP-OVOD: Open-vocabulary object detection by linear probing. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2024; pp. 779–788.
2. Fang, R.; Pang, G.; Bai, X. Simple image-level classification improves open-vocabulary object detection. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 20–27 February 2024; Volume 38, pp. 1716–1725.
3. Zhu, C.; Chen, L. A survey on open-vocabulary detection and segmentation: Past, present, and future. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 8954–8975. [CrossRef] [PubMed]
4. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the International Conference on Machine Learning, Virtual, 18–24 July 2021; pp. 8748–8763.
5. 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.
6. Zhu, P.; Wen, L.; Du, D.; Bian, X.; Fan, H.; Hu, Q.; Ling, H. Detection and tracking meet drones challenge. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 7380–7399. [CrossRef] [PubMed]

7. Tang, G.; Ni, J.; Zhao, Y.; Gu, Y.; Cao, W. A survey of object detection for UAVs based on deep learning. *Remote Sens.* **2023**, *16*, 149. [CrossRef]
8. Qin, J.; Yu, W.; Feng, X.; Meng, Z.; Tan, C. A UAV Aerial Image Target Detection Algorithm Based on YOLOv7 Improved Model. *Electronics* **2024**, *13*, 3277. [CrossRef]
9. Cheng, T.; Song, L.; Ge, Y.; Liu, W.; Wang, X.; Shan, Y. Yolo-world: Real-time open-vocabulary object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 16901–16911.
10. Zhang, R. Making convolutional networks shift-invariant again. In Proceedings of the International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; pp. 7324–7334.
11. Azulay, A.; Weiss, Y. Why do deep convolutional networks generalize so poorly to small image transformations? *J. Mach. Learn. Res.* **2019**, *20*, 1–25.
12. Yang, J.; Liu, S.; Wu, J.; Su, X.; Hai, N.; Huang, X. Pinwheel-shaped convolution and scale-based dynamic loss for infrared small target detection. In Proceedings of the AAAI Conference on Artificial Intelligence, Singapore, 20–27 January 2025; Volume 39, pp. 9202–9210.
13. Gu, X.; Lin, T.Y.; Kuo, W.; Cui, Y. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv* **2021**, arXiv:2104.13921.
14. Zhou, X.; Girdhar, R.; Joulin, A.; Krähenbühl, P.; Misra, I. Detecting twenty-thousand classes using image-level supervision. In *Proceedings of the European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 350–368.
15. Li, L.H.; Zhang, P.; Zhang, H.; Yang, J.; Li, C.; Zhong, Y.; Wang, L.; Yuan, L.; Zhang, L.; Hwang, J.N.; et al. Grounded language-image pre-training. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 10965–10975.
16. Zhang, H.; Zhang, P.; Hu, X.; Chen, Y.C.; Li, L.; Dai, X.; Wang, L.; Yuan, L.; Hwang, J.N.; Gao, J. Glipv2: Unifying localization and vision-language understanding. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 36067–36080.
17. Minderer, M.; Gritsenko, A.; Stone, A.; Neumann, M.; Weissenborn, D.; Dosovitskiy, A.; Mahendran, A.; Arnab, A.; Dehghani, M.; Shen, Z.; et al. Simple open-vocabulary object detection. In *Proceedings of the European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 728–755.
18. Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Jiang, Q.; Li, C.; Yang, J.; Su, H.; et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *Proceedings of the European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2024; pp. 38–55.
19. 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 Workshops, Seoul, Republic of Korea, 27–28 October 2019.
20. Wang, K.; Fu, X.; Huang, Y.; Cao, C.; Shi, G.; Zha, Z.J. Generalized uav object detection via frequency domain disentanglement. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 1064–1073.
21. Zhai, X.; Huang, Z.; Li, T.; Liu, H.; Wang, S. YOLO-Drone: An optimized YOLOv8 network for tiny UAV object detection. *Electronics* **2023**, *12*, 3664. [CrossRef]
22. Li, J. Research on RT-DETR Drone Target Detection Method Based on Bidirectional Feature Attention Fusion. In Proceedings of the 2025 IEEE 7th International Conference on Civil Aviation Safety and Information Technology (ICCASIT), Yinchuan, China, 22–24 October 2025; pp. 222–227.
23. Gong, J.; Yuan, Z.; Li, W.; Li, W.; Guo, Y.; Guo, B. A Lightweight Upsampling and Cross-Modal Feature Fusion-Based Algorithm for Small-Object Detection in UAV Imagery. *Electronics* **2026**, *15*, 298. [CrossRef]
24. Kayhan, O.S.; Gemert, J.C.V. On translation invariance in cnns: Convolutional layers can exploit absolute spatial location. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 14274–14285.
25. Li, X.; Wang, W.; Wu, L.; Chen, S.; Hu, X.; Li, J.; Tang, J.; Yang, J. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 21002–21012.
26. 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, 9–11 February 2020; Volume 34, pp. 12993–13000.
27. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 740–755.
28. Weng, Z.; Li, X.; Wu, C.; He, W.; Lv, J.; Zhou, D.; Yu, Z. Light-Weight Cross-Modal Enhancement Method with Benchmark Construction for UAV-based Open-Vocabulary Object Detection. *arXiv* **2025**, arXiv:2509.06011.
29. Gao, J.; Jiang, X.; Wang, A.; Gao, Y.; Fang, Z.; Lew, M.S. PMG-SAM: Boosting Auto-Segmentation of SAM with Pre-Mask Guidance. *Sensors* **2026**, *26*, 365.

30. Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A.C.; Lo, W.Y.; et al. Segment anything. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023; pp. 4015–4026.
31. 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.* **2016**, *39*, 1137–1149. [CrossRef] [PubMed]
32. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
33. 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, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9627–9636.
34. 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. *arXiv* **2022**, arXiv:2203.03605.
35. Yu, Y.; Sun, X.; Cheng, Q. Expert teacher based on foundation image segmentation model for object detection in aerial images. *Sci. Rep.* **2023**, *13*, 21964. [CrossRef] [PubMed]
36. Sun, X.; Yu, Y.; Cheng, Q. Unified diffusion-based object detection in multi-modal and low-light remote sensing images. *Electron. Lett.* **2024**, *60*, e70093.

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.