IACD: Iterative Adversarial Collaborative Detection via Dual-Perspective Blind Spot Discovery

Xiaoyan Li[1]*, Cuicui Jiang[2], Jiaoping Chen[3], and Rumei Yang[4]

[1] Department of Automation, Tsinghua University, Beijing, China
xiaoyanli629@tsinghua.edu.cn
[2] College of Computer Science, Inner Mongolia University, Hohhot, China
jiangcuicui@mail.imu.edu.cn
[3] Merrick School of Business, University of Baltimore, Baltimore, MD, USA
jchen@ubalt.edu
[4] School of Nursing, Nanjing Medical University, Jiangsu, China
rumeiyang@njmu.edu.cn

Abstract. Modern object detectors systematically fail on concealed, camouflaged, or strategically hidden objects, not due to insufficient capacity, but because they reason solely from visual appearance with no mechanism to consider where objects might be hidden. We propose **IACD** (Iterative Adversarial Collaborative Detection), a dual-perspective framework that pairs a frozen pretrained YOLOv11m (Seeker) with a lightweight trainable module (Hider) that reasons inversely: given the environment, where would a target be hidden? The Hider analyzes environmental concealment suitability across texture, edge, and semantic dimensions, then identifies the Seeker's blind spots by contrasting suitability against detection coverage. The resulting blind spot maps residually amplify feature responses via attention gates before a second detection pass. The two branches interact across multiple rounds, with the Hider's blind spot estimates informing the attention gates that guide the second detection pass. Crucially, the Hider is trained under weak supervision derived solely from the detector's own false negatives, requiring no manual concealment annotations. We frame this as *detector blind-region recovery*—recovering a frozen detector's false negatives—rather than as standard camouflaged object detection. Experiments on the Poppy illegal crop detection dataset and the VisDrone aerial surveillance benchmark demonstrate consistent improvements over the strong YOLOv11m baseline, while introducing only ∼9M trainable parameters atop the frozen detector.

Keywords: Object detection · Blind-region recovery · Dual-perspective reasoning · Detector failure mining · Weakly supervised learning

* First and corresponding author. xiaoyanli629@tsinghua.edu.cn. Code: https://github.com/xiaoyanLi629/IACD

1 Introduction

Object detectors are increasingly deployed in applications where targets actively or passively conceal themselves to evade detection. In agricultural law enforcement, illegal opium poppy cultivations are deliberately interplanted with legitimate crops or sited beneath canopy cover [14, 27]. In aerial and urban surveillance, small targets blend into cluttered backgrounds and are further degraded by resolution limits [30]. More broadly, camouflaged organisms and artifacts exploit visual similarity with their surroundings to avoid detection [3]. Despite substantial progress in object detection [1, 18, 19], modern detectors continue to exhibit systematic failures on these concealed targets.

This failure is not a capacity limitation; it reflects a fundamental design assumption shared by all contemporary detectors: targets are localized solely from visual appearance, with no mechanism to reason about where objects might be hidden when those cues are suppressed. Existing remedies, including improved architectures [8], end-to-end transformer detectors [28], and multi-scale feature pyramids [12], all operate within the same single-perspective paradigm and thus cannot address this root cause. Camouflaged object detection (COD) methods [3, 5] approach the problem from a segmentation perspective but produce no image-specific blind spot maps to guide a second detection pass.

The missing ingredient is a complementary reasoning strategy: *thinking like the hider*. Rather than asking "does this region look like a target?", an observer can ask "if I wanted to conceal something here, where would I place it?", thereby directing attention toward likely hiding spots even when they offer minimal visual evidence. While this inverse perspective has been partially explored in camouflaged object detection [25], no prior work has applied it within a practical detection pipeline that requires no concealment annotations, integrates as a plug-in atop a pretrained YOLOv11m detector, and produces blind spot estimates across multiple rounds.

We propose **IACD** (Iterative Adversarial Collaborative Detection), a framework that formalizes this dual-perspective reasoning for object detection (Figure 1). IACD augments a frozen pretrained detector (the *Seeker*) with a lightweight *Hider* module that models environmental concealment suitability and predicts where the Seeker is likely to fail. The predicted blind spot maps are used to residually amplify YOLO neck features via learned attention gates before a second detection pass, enabling the detector to recover missed targets without modifying its original weights. The Hider is applied over multiple rounds, producing blind spot estimates that drive the attention gates for the second detection pass. Crucially, the Hider is trained through *weak supervision* derived entirely from the Seeker's own false negatives, requiring no manually annotated concealment labels.

This work makes the following contributions:

- We formulate **detector blind-region recovery**—recovering a frozen detector's false negatives—and introduce a **dual-perspective framework** that models the hider's inverse reasoning as a prior for it, complementing the standard forward appearance-based detection paradigm.

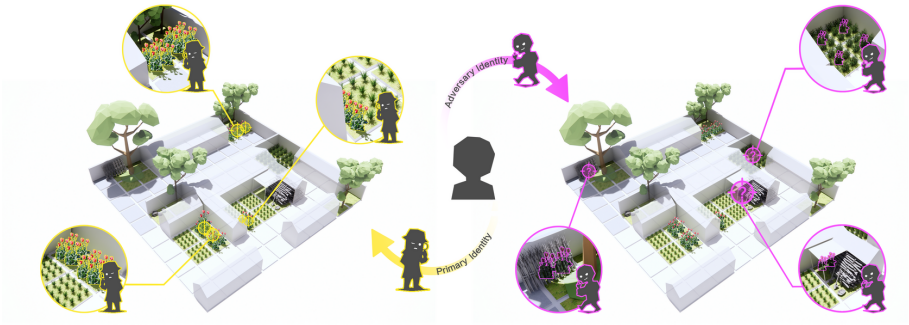

Fig. 1: Dual-perspective reasoning for detector blind-region recovery, illustrated on illegal poppy cultivation. **Left (Primary Identity)**: conventional detectors identify targets by visual appearance—objects that *look like* poppies are detected (yellow circles), but concealed plantings among legitimate crops are missed. **Right (Adversary Identity)**: by adopting the hider's perspective, reasoning about which locations *favor concealment* (dense vegetation, canopy cover, field edges), the system directs attention toward likely hiding spots (pink circles), recovering targets invisible to appearance-based detection alone. IACD formalizes this dual-perspective reasoning within a Seeker-Hider framework.

- We propose a **Hider branch** comprising an Environment Analyzer and a Blind Spot Predictor, paired with learned **Attention Gates** that use the predicted blind spot maps to amplify detector features in concealment-likely regions.
- We develop a **weakly supervised training scheme** that derives blind-spot supervision directly from the detector's false negatives, requiring no additional manual annotations.
- We demonstrate consistent improvements over a strong YOLOv11m baseline on two challenging benchmarks (Poppy illegal crop detection and VisDrone aerial surveillance) using only ∼9M trainable parameters atop a frozen 20.1M-parameter detector.

2 Related Work

2.1 Real-Time Object Detection

Real-time object detection has advanced along two complementary trajectories. The YOLO family [8, 9, 18, 21] established the dominant paradigm for single-stage detection, with successive versions progressively refining the accuracy/speed trade-off through decoupled prediction heads, anchor-free formulations, and C2f bottleneck modules. YOLOv9 [22] further addresses information loss in deep networks via Programmable Gradient Information (PGI), ensuring complete input context is preserved all the way into the prediction stage. In

parallel, transformer-based approaches introduced end-to-end detection without hand-crafted anchors: DETR [1] first demonstrated the viability of set-prediction for detection, and RT-DETR [28] extended this to real-time inference through an efficient hybrid encoder with IoU-aware query selection.

Despite diverging in architecture (convolutional vs. transformer, anchor-based vs. anchor-free), all of these methods share a fundamental assumption: objects are localized exclusively from their visual appearance in a single forward pass. When appearance cues are suppressed through camouflage or environmental blending, this shared assumption fails, and no architectural refinement within this paradigm can compensate for the lack of a complementary reasoning strategy.

2.2 Camouflaged Object Detection

Camouflaged object detection (COD) addresses the challenge of segmenting objects whose appearance closely mimics their surroundings [3]. The field has pursued progress along two parallel directions. On one front, methods have progressively enriched the visual reasoning process: SINet-V2 [3] introduced a search-and-identification strategy inspired by predatory behavior; FEDER [5] decomposes features and reconstructs edge cues to suppress background interference; and FSEL [20] jointly reasons in frequency and spatial domains to expose texture boundaries invisible to either domain alone. On another front, methods have sought to reduce reliance on expensive annotations: Noisy-COD [26] learns from noisy pseudo-labels, and VSCode [15] unifies salient and camouflaged detection under a single 2D prompt-learning framework, broadening applicability without task-specific retraining.

The work most directly related to IACD is UCOD-DPL [25], which introduces an adversarial dual-branch decoder that reasons about foreground and background through complementary perspectives via teacher-student learning, making it the first COD method to explicitly leverage a dual-perspective design. Yet UCOD-DPL shares the core constraints of all prior COD work: it is a standalone model trained from scratch on curated camouflage datasets with dense pixel-level annotations, and its dual-branch predictions are computed in isolation, without any feedback from intermediate detection outputs. Consequently, it cannot be grafted onto an existing pretrained detector, and its training signal depends entirely on manually annotated concealment masks.

In contrast to all prior COD work, IACD requires no curated concealment datasets or pixel-level annotations, operates as a plug-in module atop a frozen pretrained YOLOv11m detector rather than a standalone model, and produces image-specific blind spot maps that guide a second detection pass at inference time.

Positioning relative to COD, and why we do not compare directly. We stress that IACD does not target the standard COD benchmark task, and we therefore do not report head-to-head comparisons against specialized COD models such as SINet-V2 [3] or FEDER [5]. "Concealment" here denotes a reverse-thinking reasoning prior, not the canonical close-range camouflage task: our ob-

jective is *detector blind-region recovery*, i.e. recovering the instance-level false negatives of a frozen detector on aerial imagery. A direct mAP comparison would be neither well-defined nor fair, for three reasons. *(i) Incompatible outputs:* COD methods emit a single-channel pixel-level foreground mask, whereas detection mAP requires class-labelled instance boxes; bridging the two requires a mask-to-box heuristic that confounds the comparison with conversion artefacts. *(ii) Domain mismatch:* COD models are trained on close-range natural-camouflage datasets (COD10K, NC4K) and degrade sharply under the scale-, viewpoint-, and resolution-dominated concealment of UAV imagery. *(iii) Semantic scope:* COD is single-foreground binary segmentation, whereas our benchmarks require multi-class detection (ten categories on VisDrone), which lies outside the COD formulation. We thus treat COD as related but methodologically distinct prior art, and benchmark IACD against general-purpose detectors that share its instance-level, multi-class output space.

2.3 Weakly Supervised Object Detection

Obtaining supervision without exhaustive annotation is a persistent challenge. Weakly supervised detectors derive localization from image-level labels [11], while semi-supervised methods self-train on pseudo-labels, where ensuring label quality is key—e.g. Consistent-Teacher [23] stabilizes the teacher's targets across iterations. IACD instead mines supervision from the base detector's *failures* on already-labeled data: ground-truth boxes the frozen detector misses at training time form the false-negative signal driving the Hider, needing no annotation beyond the bounding boxes already present in any detection dataset.

2.4 Frozen Backbone Adaptation

A growing paradigm freezes large pretrained models and augments them with compact trainable modules instead of full fine-tuning. Adapters [2,6] and prompt-tuning [7] approach full fine-tuning with few extra parameters while preserving pretrained representations, and Frozen-DETR [4] extends this to detection by bridging frozen foundation-model features. IACD applies the same paradigm to a *complete* detection pipeline: the entire YOLOv11m is frozen, and lightweight modules driven by its own failure signals are inserted upstream of the head.

3 IACD: Iterative Adversarial Collaborative Detection

3.1 Problem Formulation

Let $\mathbf{I} \in \mathbb{R}^{B\times 3\times H\times W}$ denote a batch of input images, and let $\mathcal{G}$ denote the set of ground-truth bounding boxes. A standard detector f_θ produces detections $\mathcal{D} = f_\theta(\mathbf{I})$ that miss a subset $\mathcal{FN} = \mathcal{G}\backslash\mathcal{D}$ of false negative (FN) targets. Our goal is to train an auxiliary module that identifies the spatial regions responsible for these failures (the *blind spots*) and guides the detector to recover them, *without modifying* θ.

Formally, we seek a blind spot map $\mathbf{B} \in [0,1]^{B\times 1\times H'\times W'}$ such that high values of $\mathbf{B}$ coincide with the centers of FN targets. Given $\mathbf{B}$, we then modulate the detector's internal features to increase the response in those regions and run a second detection pass.

3.2 Framework Overview

The IACD framework (Figure 2) consists of four trainable components layered atop a **frozen** pretrained YOLOv11m [9]: (1) **Feature Projections** that map YOLO neck features to a uniform 256-channel space for the Hider; (2) the **Hider Branch** that predicts environmental suitability and blind spot maps; (3) **Attention Gates** that residually amplify neck features guided by the blind spot map; and (4) an **Auxiliary Detection Head** used only during training. The backbone and detection head of YOLOv11m are entirely frozen; only the ∼9M parameters of the IACD modules are trained.

Feature extraction relies on PyTorch's `register_forward_hook` mechanism. During a single YOLOv11m forward pass, neck outputs at three scales are intercepted: $\mathbf{P}_3 \in \mathbb{R}^{B\times c_3\times 80\times 80}$ (stride 8), $\mathbf{P}_4 \in \mathbb{R}^{B\times c_4\times 40\times 40}$ (stride 16), and $\mathbf{P}_5 \in \mathbb{R}^{B\times c_5\times 20\times 20}$ (stride 32), where $c_3 = 256$, $c_4 = 512$, $c_5 = 512$ are the native YOLOv11m neck channel dimensions. These features are then projected to 256 channels each via 1×1 Conv-BN-SiLU blocks, forming the shared feature set $\mathcal{F} = \{\mathbf{P}'_3, \mathbf{P}'_4, \mathbf{P}'_5\}$ used by the Hider.

3.3 Hider Branch

The Hider branch takes $\mathcal{F}$ and the Seeker's detection coverage map $\mathbf{D}$ as inputs and produces the blind spot map $\mathbf{B}$. It comprises two sequential modules.

Environment Analyzer. The Environment Analyzer answers the question: *given the scene content, which regions offer favorable concealment conditions?* It fuses the three scales by upsampling $\mathbf{P}'_4$ and $\mathbf{P}'_5$ to 80×80 and concatenating with $\mathbf{P}'_3$, producing a 768-channel fused representation compressed back to 256 channels via a 1×1 Conv-BN-ReLU block.

Three parallel branches analyze different concealment-relevant properties: (i) a **texture branch** using dilated convolutions (dilation $\in \{1,2,4\}$) to capture multi-scale texture complexity, which is strongly correlated with effective camouflage; (ii) an **edge density branch** using 3×3 convolutions to detect regions of high edge clutter that facilitate boundary confusion; (iii) a **context branch** combining local 3×3 convolution features with a global average-pooled context vector to capture semantic cues such as vegetation and shadow patterns. A lightweight weight network (global pooling $\rightarrow$ FC $\rightarrow$ ReLU $\rightarrow$ FC $\rightarrow$ Softmax) produces three learnable scalar weights $[w_t, w_e, w_c]$, and the environmental suitability map is:

$$\mathbf{S} = w_t \cdot \mathbf{S}_t + w_e \cdot \mathbf{S}_e + w_c \cdot \mathbf{S}_c \in [0,1]^{B\times 1\times 80\times 80} \tag{1}$$

Blind Spot Predictor. Given $\mathbf{S}$ and the detection coverage map $\mathbf{D}$ (a Gaussian-smoothed heatmap of YOLO's current detections), the Blind Spot Predictor computes:

$$\mathbf{B} = \lambda \cdot \hat{\mathbf{B}} + (1 - \lambda) \cdot \mathbf{S} \odot (1 - \mathbf{D}) \tag{2}$$

where $\hat{\mathbf{B}}$ is a learned fusion of $[\mathbf{P}'_3, \mathbf{S}, \mathbf{D}]$ through a 4-layer convolutional network, and $\lambda \in (0, 1)$ is a learnable scalar initialized to 0.5. The second term provides a principled heuristic: high environmental suitability combined with low detection confidence indicates a likely blind spot.

3.4 Attention-Guided Second Pass

Three Attention Gates ($\mathcal{A}_3$, $\mathcal{A}_4$, $\mathcal{A}_5$) modulate the original (non-projected) YOLO neck features $\mathbf{P}_k$ using the blind spot map $\mathbf{B}$:

$$\tilde{\mathbf{P}}_k = \mathbf{P}_k \odot \big(1 + \alpha_k \cdot \mathcal{G}_k(\mathbf{P}_k, \mathbf{B}) \odot \mathbf{B}\big) \tag{3}$$

where $\mathcal{G}_k$ is a small convolutional network ($\text{Conv}_{c_k+1 \to c_k}(3\times3) \to \text{BN} \to \text{SiLU} \to \text{Conv}_{c_k \to c_k}(1\times1) \to \text{Sigmoid}$) that learns which feature channels to amplify in blind-spot regions, and $\alpha_k = \sigma(s_k) \cdot \alpha_{\max}$ with learnable strength s_k (initialized to -4) and $\alpha_{\max} = 0.10$.

The residual form $(1 + \cdots)$ guarantees that easy images where $\mathbf{B} \approx 0$ experience zero perturbation. The strict bound $\alpha_k \leq 0.10$ caps the maximum feature amplification at $1.10\times$, preventing false positive inflation on dense scenes. The enhanced features $\{\tilde{\mathbf{P}}_3, \tilde{\mathbf{P}}_4, \tilde{\mathbf{P}}_5\}$ are passed to the frozen YOLO detection head for a second inference pass.

3.5 Iterative Collaboration

During training, IACD runs the Hider $T = 2$ times against the detection coverage map $\mathbf{D}$ constructed from the baseline YOLO detections (confidence ≥ 0.25, Gaussian diffusion $\sigma = 2.0$, map size 80×80). The final blind spot map $\mathbf{B}_T$ drives the attention gates for the second detection pass. The framework's gains are robust to this choice: $T = 1$ achieves performance comparable to $T = 2$, confirming that the core Seeker-Hider design, rather than the number of Hider applications, is the primary source of improvement (see Table 3).

3.6 Training Objectives

All YOLO parameters are frozen throughout training. Only the feature projections, Hider branch, attention gates, and auxiliary detection head are optimized. The total loss comprises three terms:

Hider Focal Loss $\mathcal{L}_{\text{hider}}$. The blind spot map should assign high probability to FN regions. For each image, we identify FN targets as ground-truth boxes with maximum IoU < 0.5 with any confident (≥ 0.25) YOLO detection. A soft target map $\mathbf{T}_{\text{FN}}$ is constructed by placing Gaussian blobs (radius $\sigma = 3.0$) at FN box centers. The loss applies focal weighting [13]:

$$\mathcal{L}_{\text{hider}} = \sum_{t} \mathbb{E}\big[\alpha_f (1 - p_t)^{\gamma} \mathcal{L}_{\text{BCE}}(\mathbf{B}_t, \mathbf{T}_{\text{FN}})\big] \tag{4}$$

with $\alpha_f = 0.75$, $\gamma = 2.0$.

Auxiliary Detection Loss $\mathcal{L}_{\text{aux}}$. A CenterNet-style head [29] on the projected feature $\mathbf{P}'_3$ provides additional supervision:

$$\mathcal{L}_{\text{aux}} = \mathcal{L}_{\text{hm}} + 0.1\,\mathcal{L}_{\text{wh}} + 0.1\,\mathcal{L}_{\text{off}} \tag{5}$$

where the heatmap loss $\mathcal{L}_{\text{hm}}$ uses focal weighting [10] with enhanced weight ($3\times$) on FN targets, and $\mathcal{L}_{\text{wh}}$, $\mathcal{L}_{\text{off}}$ are ℓ_1 regression losses on width/height and sub-pixel offset.

Attention Gate Loss $\mathcal{L}_{\text{gate}}$. To train the attention gates, enhanced features $\{\tilde{\mathbf{P}}_k\}$ are passed through the frozen YOLO detection head in training mode, and the standard Ultralytics detection loss [8] $\mathcal{L}_{\text{yolo}}$ is computed against ground-truth boxes. Since the YOLO head is frozen, gradients from this term flow only through the attention gates, denoted $\mathcal{L}_{\text{gate}} = \mathcal{L}_{\text{yolo}}$. The three terms are combined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{hider}} + \mathcal{L}_{\text{aux}} + 0.1\,\mathcal{L}_{\text{gate}} \tag{6}$$

3.7 Analysis

Dual-perspective complementarity. The Seeker's features $\mathbf{P}_k$ encode *appearance* evidence while the Hider's suitability map $\mathbf{S}$ encodes *environmental* evidence; a region with high $\mathbf{S}$ and low coverage $\mathbf{D}$ is the canonical signature of a concealed target, which Eq. (2) formalizes. Training is a coupled optimization—the Hider learns to predict FN regions, supervising the gates, which in turn learn to amplify the right features—run end-to-end within each mini-batch.

Stability under bounded amplification. From Eq. (3), since $\mathcal{G}_k, \mathbf{B} \in [0, 1]$, $\|\tilde{\mathbf{P}}_k - \mathbf{P}_k\|_F \leq \alpha_{\max} \|\mathbf{P}_k\|_F$; with $\alpha_{\max}$=0.10 features stay within 10% of the original, avoiding out-of-distribution inputs to the frozen head, and the residual form guarantees the second pass equals the first when $\mathbf{B} \approx \mathbf{0}$ (a strict lower bound of baseline).

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate on two benchmarks with different concealment regimes.

Poppy is a UAV aerial dataset for illegal opium poppy cultivation detection, assembled from three public Roboflow datasets [16,17,24] (all CC BY 4.0). All annotations are unified into a single "poppy" class, yielding 6,790 images partitioned into 4,723 / 1,034 / 1,033 train/val/test splits. Cultivators actively employ concealment strategies including intercropping with legitimate vegetation and canopy positioning, making this a natural benchmark for concealed object detection.

VisDrone [30] is a large-scale aerial detection benchmark captured from UAVs across diverse scenarios, with 6,471 / 548 / 1,610 train/val/test splits over 10 object categories (pedestrian, car, bus, *etc.*). Objects are small and frequently obscured by backgrounds, representing natural environmental concealment.

Baselines. We compare against nine representative detection architectures spanning three model families: YOLOv8 (n/s/m) [8], YOLOv10 (s/m) [21], YOLOv11 (s/m) [9], and RT-DETR (L/X) [28]. All baselines are trained from pretrained weights for 300 epochs with early stopping (patience 50). Following standard practice [3], all experiments are repeated with 5 independent random seeds, and we report mean ± standard deviation.

Implementation Details. IACD wraps a frozen YOLOv11m checkpoint (the best baseline on each dataset). Its ∼9M trainable parameters are optimized with AdamW (lr 10^{-3}, weight decay 10^{-4}, cosine annealing to 10^{-5}), batch size 16, up to 300 epochs (patience 50), $\alpha_{\max}$=0.10, T=2; other settings follow Sec. 3. We report 3 seeds (the Hider is costly to train atop a fixed baseline). Inference injects a uniform blind map ($\mathbf{B}$=0.3, chosen by grid search over $\{0.1, 0.2, 0.3, 0.5\}$ on validation) through the training-time hooks and uses the same Ultralytics `model.val()` pipeline as the baselines. All runs use PyTorch on one NVIDIA A100 with mixed precision.

4.2 Comparison with Baselines

We report five metrics on Poppy and VisDrone (Tables 1 and 2); YOLOv11m is the strongest single baseline on both. IACD, built atop the frozen YOLOv11m, attains the best mAP@0.5 and mAP@0.5:0.95 across all runs: on Poppy 88.87±0.02 vs. 88.40±0.18 (Recall 79.79→80.07), on VisDrone 38.65±0.00 vs. 38.59±0.24 (Precision 50.81→51.81). IACD's std ($\leq$ 0.02%) is far below the baseline range (0.08–1.06%), indicating blind-spot attention stabilizes detection across seeds.

Recall gains matter most here, since false negatives are the primary failure mode: the +0.28% Poppy Recall gain reflects recovered targets, while on VisDrone the gain is mainly in Precision (+1.00%) with Recall comparable (39.78 vs. 39.95), i.e. the gates here mostly suppress false positives in clutter. Qualitative examples (Figure 3) confirm the blind-spot maps highlight missed targets, and IACD recovers many baseline false negatives.

Table 1: Detection performance on the **Poppy** test set (mean±std over 5 seeds; IACD over 3 seeds). **Bold**: best result.

Model	mAP@0.5 (%)	mAP@0.5:0.95 (%)	P (%)	R (%)	F1 (%)
YOLOv8n [8]	85.95±0.19	76.38±0.22	92.14±0.58	77.46±0.76	84.16±0.30
YOLOv8s [8]	86.98±0.15	80.06±0.15	93.11±0.72	79.66±0.49	85.86±0.19
YOLOv8m [8]	87.86±0.21	82.19±0.17	93.93±0.91	79.50±0.70	86.11±0.27
YOLOv10s [21]	86.82±0.39	80.01±0.35	93.39±0.83	79.37±0.52	85.81±0.28
YOLOv10m [21]	87.72±0.13	82.05±0.30	**94.31**±0.64	79.25±0.63	86.13±0.17
YOLOv11s [9]	87.21±0.21	79.54±0.19	92.96±0.37	79.66±0.52	85.79±0.25
YOLOv11m [9]	88.40±0.20	82.41±0.21	93.80±0.66	79.79±0.32	**86.23**±0.21
RT-DETR-L [28]	86.12±0.24	79.37±0.23	91.12±0.19	79.73±0.40	85.04±0.19
RT-DETR-X [28]	86.21±0.17	79.30±0.27	91.17±0.97	79.93±0.79	85.18±0.16
IACD (Ours)	**88.87**±0.02	**82.83**±0.02	93.41±0.01	**80.07**±0.00	**86.23**±0.00

Table 2: Detection performance on the **VisDrone** test set (mean±std over 5 seeds†; IACD over 3 seeds). **Bold**: best result.

Model	mAP@0.5 (%)	mAP@0.5:0.95 (%)	P (%)	R (%)	F1 (%)
YOLOv8n [8]	29.87±0.15	16.98±0.07	41.07±0.91	31.81±0.45	35.84±0.31
YOLOv8s [8]	34.16±0.43	19.92±0.28	46.32±0.63	35.59±0.46	40.25±0.39
YOLOv8m [8]	36.99±0.33	21.85±0.15	50.34±0.62	38.47±0.19	43.61±0.19
YOLOv10s [21]	34.21±0.08	19.81±0.03	46.05±0.38	35.59±0.18	40.15±0.13
YOLOv10m [21]	37.32±0.36	22.08±0.23	50.44±0.88	38.10±0.42	43.40±0.45
YOLOv11s [9]	34.06±0.16	19.89±0.12	45.83±0.41	36.09±0.26	40.37±0.09
YOLOv11m [9]	38.59±0.27	**22.93**±0.14	50.81±0.55	**39.95**±0.65	44.73±0.30
RT-DETR-L [28]	30.77±1.06†	16.64±0.87	46.01±1.10	34.56±1.35	39.46±1.15
RT-DETR-X [28]	31.65±0.78	17.58±0.44	46.29±0.49	34.84±0.74	39.75±0.62
IACD (Ours)	**38.65**±0.00	**22.93**±0.01	**51.81**±0.01	39.78±0.01	**45.00**±0.01

†RT-DETR-L: 4 valid runs; one seed diverged and was excluded.

4.3 Ablation Study

To isolate the contribution of each IACD component, we conduct ablation experiments across four metrics on both datasets (Table 3; mean±std over 3 seeds each). All IACD variants consistently surpass the frozen YOLOv11m baseline on mAP@0.5 and mAP@0.5:0.95, confirming that the Seeker-Hider framework is the source of improvements regardless of specific component choices.

Removing the Environment Analyzer or Blind Spot Predictor yields slightly lower mAP@0.5, with the Blind Spot Predictor showing the largest degradation (-0.06%). The mAP@0.5:0.95 metric confirms this trend, indicating that these components contribute to localization quality beyond threshold-level detection. Single-round ($T = 1$) performance is comparable to $T = 2$, confirming that the

Table 3: Ablation study (mean±std over 3 seeds; Baseline over 5 seeds). **Bold**: best result per dataset.

Configuration	mAP@0.5 (%)	mAP@.5:.95 (%)	R (%)	F1 (%)
Poppy				
Baseline (YOLOv11m)	88.40±0.20	82.41±0.21	79.79±0.32	86.23±0.21
w/o Env. Analyzer	88.83±0.01	82.73±0.02	80.11±0.02	86.12±0.02
w/o Blind Spot Pred.	88.81±0.02	82.68±0.03	80.12±0.02	86.13±0.01
w/o Iterative ($T=1$)	88.84±0.01	82.74±0.01	80.11±0.02	86.12±0.02
Conservative ($\alpha_{\max}=0.05$)	88.83±0.00	82.72±0.01	80.10±0.00	86.11±0.02
IACD full ($T=2$, $\alpha_{\max}=0.10$)	**88.87**±0.02	**82.83**±0.02	80.07±0.00	**86.23**±0.00
VisDrone				
Baseline (YOLOv11m)	38.59±0.27	22.93±0.14	39.95±0.65	44.73±0.30
w/o Env. Analyzer	38.64±0.01	22.94±0.01	39.82±0.03	44.93±0.02
w/o Blind Spot Pred.	38.64±0.00	22.93±0.00	39.79±0.03	44.93±0.04
w/o Iterative ($T=1$)	38.64±0.00	22.93±0.00	39.84±0.02	44.89±0.01
Conservative ($\alpha_{\max}=0.05$)	38.64±0.00	22.93±0.01	39.80±0.02	44.90±0.04
IACD full ($T=2$, $\alpha_{\max}=0.10$)	**38.65**±0.00	**22.93**±0.01	39.78±0.01	**45.00**±0.01

improvements stem from the Seeker-Hider dual-perspective design itself rather than from the number of Hider applications.

Reducing $\alpha_{\max}$ to 0.05 slightly lowers accuracy but achieves the lowest variance ($\sigma = 0.003\%$ vs. 0.016%); we adopt $\alpha_{\max} = 0.10$ as default since both are far more stable than baselines ($\sigma \geq 0.08\%$).

4.4 Model Scaling and Parameter-Neutral Comparison

A natural concern is whether IACD's gains stem from added capacity rather than from blind-region reasoning: its ∼9M trainable parameters are a non-trivial increase over the frozen 20.1M-parameter host. To control for this, we compare IACD against simply scaling the vanilla YOLO backbone (s → m → l) under matched parameter budgets (Table 4; 3 seeds each).

Two observations rule out the "larger-vanilla-wins" explanation. First, on Poppy, IACD on a *frozen* YOLOv11m with only 9.05M trainable parameters (88.87%) outperforms a *fully trainable* 25.3M-parameter YOLOv11l (88.21%): it beats a larger vanilla model while optimizing 2.8× fewer parameters, and collapses seed variance by an order of magnitude (±0.02 vs. ±0.20–0.28). Second, the IACD module is *identical* in size for the m and l hosts (9,046,430 trainable parameters in both, since YOLOv11m and YOLOv11l share the same neck channel widths) and smaller for the narrower s host (6,485,918): the added module scales with neck width rather than being tuned per backbone, and IACD improves Poppy mAP@0.5 at all three scales (+0.17/ + 0.47/ + 0.34 for s/m/l). On

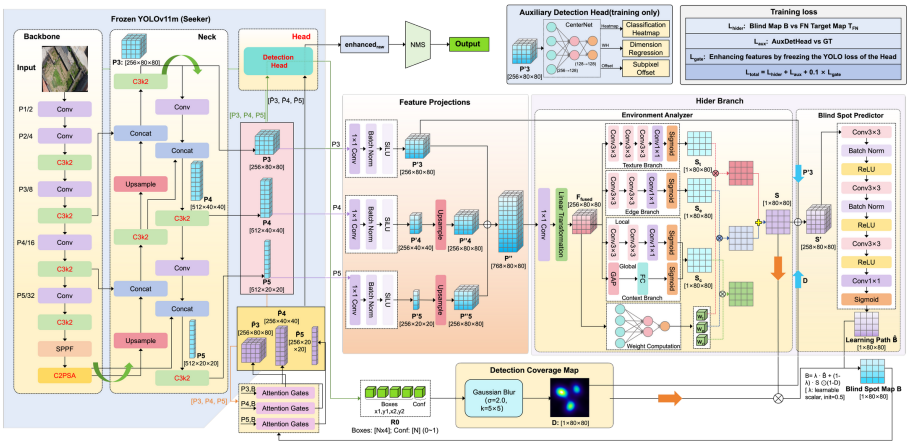

Fig. 2: IACD architecture overview. *Seeker (left):* a frozen YOLOv11m extracts multi-scale Neck features {P3, P4, P5} (strides 8/16/32); the Detection Head produces first-pass detections R0. *Feature Projections:* 1×1 Conv-BN-SiLU modules project P3/P4/P5 to a uniform 256-channel space. *Hider Branch:* the *Environment Analyzer* fuses the projected features through Texture, Edge, and Context branches into a Suitability Map **S**; the *Detection Coverage Map* turns R0 into a Gaussian heatmap **D**; the *Blind Spot Predictor* combines a learned estimate with the heuristic $\mathbf{S}\odot(1-\mathbf{D})$ into the Blind Spot Map **B** (Eqs. (1) and (2)). *Attention Gates:* three residual gates $\tilde{P}_k = P_k\odot(1+\alpha_k G_k \mathbf{B})$ amplify the Neck features for a second pass through the frozen head. *Auxiliary Head* (training only): a CenterNet-style head adding false-negative supervision. Total loss $\mathcal{L} = \mathcal{L}_{\text{hider}} + \mathcal{L}_{\text{aux}} + 0.1\,\mathcal{L}_{\text{gate}}$.

Table 4: Model scaling vs. IACD under matched parameter budgets. Params in millions; IACD freezes its host (only the IACD modules are trainable). mAP@0.5 (%), mean±std over 3 seeds. **Bold**: best per dataset. The IACD module is identical for the m/l hosts (same neck width, 9,046,430 params) and smaller for s (6,485,918), as discussed in the text.

Model	Train. (M)	Total (M)	Poppy (%)	VisDrone (%)
YOLOv11s	9.43	9.43	87.21±0.21	34.06±0.16
YOLOv11m	20.05	20.05	88.40±0.20	38.59±0.27
YOLOv11l	25.31	25.31	88.21±0.28	**39.90**±0.13
v11s+IACD	6.49	15.91	87.38±0.04	34.13±0.02
v11m+IACD	9.05	29.10	**88.87**±0.02	38.65±0.00
v11l+IACD	9.05	34.36	88.55±0.07	39.78±0.01

VisDrone, scaling the vanilla backbone to YOLOv11l reaches 39.90% by relearning all 25.3M weights—a capacity-driven axis orthogonal to IACD's frozen-host design.

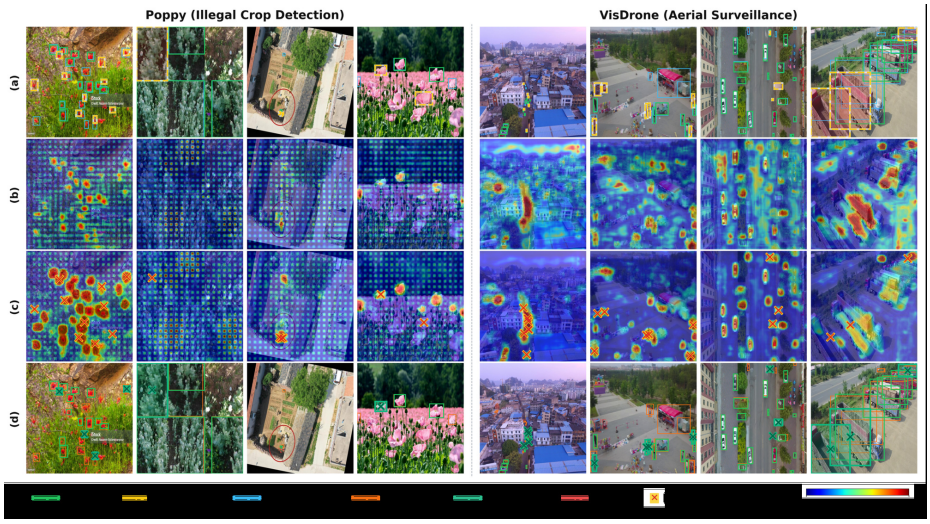

Fig. 3: Qualitative results on Poppy (left) and VisDrone (right). **(a)** baseline YOLOv11m detections (cyan), detected GT (green), false negatives (yellow dashed); **(b)** suitability map S; **(c)** blind-spot map B ($\times$ at FN centers); **(d)** IACD-enhanced detections (orange) with recovered FN (green) and remaining misses (red dashed). IACD recovers 20/37 baseline FN (54%) across the eight examples.

Table 5: Dynamic Hider vs. static blind map on VisDrone test set (1,610 images, 75,102 GT boxes). Images stratified by baseline false-negative count.

	FN count (↓)		Image-level wins		
Stratum	Static	Dynamic	Dyn. better	Static better	Tie
Easy (0 BL-FN, 67 imgs)	1	0	1	0	66
Medium (1–2 BL-FN, 109 imgs)	175	171	7	3	99
Hard (≥3 BL-FN, 1434 imgs)	36,813	**36,062**	**505**	70	859
All (1,610 imgs)	36,989	**36,233**	**513**	73	1,024

4.5 Dynamic vs. Static Inference

A natural question is whether the Hider branch is needed at inference time, or whether a fixed blind spot map suffices. We compare two deployment modes: **(i)** *Hook mode*: the Hider is not executed; a uniform blind map $\mathbf{B} = 0.3$ is injected through forward hooks, adding only the three lightweight Attention Gates to the YOLO forward pass. **(ii)** *Full pipeline*: the Hider dynamically generates a per-image blind spot map through two rounds before gating.

On Poppy, both modes perform nearly identically (5 vs. 3 image-level wins across 1,033 test images), consistent with the relatively low baseline false-negative rate on this dataset. On VisDrone (Table 5), the full pipeline recovers **756 additional false negatives** over the static mode (36,233 vs. 36,989 FN against

Table 6: Inference overhead (batch size 1, 640×640 input).

Mode	Latency (ms)	FPS	GPU Mem. (MB)
Baseline YOLOv11m	13.4	74.5	227
IACD hook mode	14.3 (+6%)	70.0	263
IACD full pipeline	23.4 (+74%)	42.7	263

75,102 GT boxes), corresponding to a per-image FN recovery rate of 51.75% vs. 50.75% at conf ≥ 0.25. Note that this FN-count metric uses a fixed low confidence threshold and differs from the standard mAP Recall reported in Table 2, which is computed at the optimal threshold by the Ultralytics `model.val()` pipeline. The advantage concentrates on hard images ($\geq$3 baseline FN): the dynamic Hider wins on 505 images versus 70 for the static mode (7:1 ratio), recovering 855 baseline FN compared to only 104 for the static mode, a **7.2**× improvement.

These results validate the Hider's dual role: during *training* it supplies the weakly-supervised signal that teaches the gates which features to amplify, and during *inference* it generates per-image blind spot maps that recover targets the static gates miss. The resulting accuracy/speed trade-off (hook mode +6% latency, full pipeline ∼74%) is navigable via the confidence-gated activation of Sec. 4.6.

4.6 Computational Overhead

Deployment efficiency is critical for UAV platforms. We benchmark inference latency on a single NVIDIA A100 GPU at 640×640 resolution (Table 6). Hook mode adds only 0.9 ms (the cost of three Attention Gate convolutions), preserving near-baseline throughput at 70 FPS. The full pipeline, which additionally runs the Hider branch and a second detection-head pass, operates at 42.7 FPS. Peak GPU memory increases by only 36 MB (+16%) in both modes, as the IACD modules are lightweight relative to the frozen YOLO backbone.

Confidence-gated activation. The full pipeline need not run on every image. Since the Hider only helps where the first pass is uncertain, we gate its activation on the first-pass maximum confidence, invoking the Hider and the second detection pass only when this confidence falls below a threshold τ (otherwise the baseline detections are returned unchanged). On VisDrone this exposes a tunable accuracy/latency frontier between the hook and full-pipeline extremes above: at $\tau = 0.9$ the Hider fires on 85.5% of images and preserves 81% of the full pipeline's recall gain, whereas $\tau = 0.8$ fires on only 10.6% of images while still preserving 13.5% of the gain. The reachable trade-off is bounded by YOLOv11m's already-high baseline confidence, which leaves few images below τ at stricter thresholds.

4.7 Limitations

First, the improvement margins remain modest—e.g. +0.47% mAP@0.5 on Poppy and +0.06% on VisDrone for the YOLOv11m host—and they are not uniform across hosts: on one of the six host×dataset settings (YOLOv11l on VisDrone) IACD is marginally below the vanilla baseline (−0.12%, Table 4). The full pipeline also incurs increased latency (Table 6), which the confidence-gated activation of Section 4.6 only partially mitigates. Second, although we validate the plug-in across the YOLOv11 family (s/m/l; Table 4), all three hosts share the same single-stage YOLO architecture; the "plug-and-play" claim is therefore established *within the YOLO family only*, and generalization to other paradigms—RT-DETR, the DETR family, or two-stage detectors—remains future work requiring head-specific adaptation. Third, both benchmarks are aerial; other domains (underwater, satellite) would further test generality.

5 Conclusion

We presented IACD, a dual-perspective framework for *detector blind-region recovery* that addresses a fundamental limitation of appearance-based detectors: their inability to reason about *where* their own blind spots—the targets they miss—are likely to lie. By pairing a frozen pretrained detector (the Seeker) with a lightweight Hider module that analyzes environmental concealment suitability across texture, edge, and semantic dimensions, IACD recovers targets the Seeker misses without modifying its original weights. Crucially, the Hider is trained entirely from the detector's own false negatives, requiring no concealment annotations beyond standard bounding-box labels. On two challenging aerial benchmarks, IACD improves over the strong YOLOv11m baseline by +0.47% mAP@0.5 on Poppy and +0.06% on VisDrone, using only ∼9M trainable parameters atop a frozen 20.1M-parameter detector. Future work will explore stronger gating mechanisms, broader backbone architectures, and extending IACD to domains such as medical imaging and security screening where environmental concealment reasoning may help.

Broader Impact and Ethics Statement. IACD targets narrow law-enforcement tasks such as illegal crop detection and is **not** intended for mass surveillance or individual tracking; any deployment must comply with regulations, have institutional oversight, and keep a human in the loop.

References

1. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European conference on computer vision. pp. 213–229. Springer (2020)
2. Chen, Z., Duan, Y., Wang, W., He, J., Lu, T., Dai, J., Qiao, Y.: Vision transformer adapter for dense predictions. arXiv preprint arXiv:2205.08534 (2022)

3. Fan, D.P., Ji, G.P., Cheng, M.M., Shao, L.: Concealed object detection. IEEE transactions on pattern analysis and machine intelligence **44**(10), 6024–6042 (2021)
4. Fu, S., Yan, J., Yang, Q., Wei, X., Xie, X., Zheng, W.S.: Frozen-detr: Enhancing detr with image understanding from frozen foundation models. Advances in Neural Information Processing Systems **37**, 105949–105971 (2024)
5. He, C., Li, K., Zhang, Y., Tang, L., Zhang, Y., Guo, Z., Li, X.: Camouflaged object detection with feature decomposition and edge reconstruction. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 22046–22055 (2023)
6. Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., Gelly, S.: Parameter-efficient transfer learning for nlp. In: International conference on machine learning. pp. 2790–2799. PMLR (2019)
7. Jia, M., Tang, L., Chen, B.C., Cardie, C., Belongie, S., Hariharan, B., Lim, S.N.: Visual prompt tuning. In: European conference on computer vision. pp. 709–727. Springer (2022)
8. Jocher, G., Chaurasia, A., Qiu, J.: Ultralytics YOLO. GitHub (2023), `https://github.com/ultralytics/ultralytics`
9. Khanam, R., Hussain, M.: YOLOv11: An overview of the key architectural enhancements. arXiv preprint arXiv:2410.17725 (2024)
10. Law, H., Deng, J.: Cornernet: Detecting objects as paired keypoints. In: Proceedings of the European conference on computer vision (ECCV). pp. 734–750 (2018)
11. Lin, J., Shen, Y., Wang, B., Lin, S., Li, K., Cao, L.: Weakly supervised open-vocabulary object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 3404–3412 (2024)
12. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2117–2125 (2017)
13. 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. pp. 2980–2988 (2017)
14. Liu, X., Tian, Y., Yuan, C., Zhang, F., Yang, G.: Opium poppy detection using deep learning. Remote Sensing **10**(12), 1886 (2018)
15. Luo, Z., Liu, N., Zhao, W., Yang, X., Zhang, D., Fan, D.P., Khan, F., Han, J.: Vscode: General visual salient and camouflaged object detection with 2d prompt learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 17169–17180 (2024)
16. opium poppy: poppy dataset. `https://universe.roboflow.com/opium-poppy/poppy-qgghf` (Jun 2024), `https://universe.roboflow.com/opium-poppy/poppy-qgghf`, visited on 2026-02-28
17. project pvbts: poppy dataset. `https://universe.roboflow.com/project-pvbts/poppy-x8njl` (Apr 2025), `https://universe.roboflow.com/project-pvbts/poppy-x8njl`, visited on 2026-02-28
18. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 779–788 (2016)
19. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems **28** (2015)
20. Sun, Y., Xu, C., Yang, J., Xuan, H., Luo, L.: Frequency-spatial entanglement learning for camouflaged object detection. In: European Conference on Computer Vision. pp. 343–360. Springer (2024)

21. Wang, A., Chen, H., Liu, L., Chen, K., Lin, Z., Han, J., et al.: Yolov10: Real-time end-to-end object detection. Advances in neural information processing systems **37**, 107984–108011 (2024)
22. Wang, C.Y., Yeh, I.H., Mark Liao, H.Y.: Yolov9: Learning what you want to learn using programmable gradient information. In: European conference on computer vision. pp. 1–21. Springer (2024)
23. Wang, X., Yang, X., Zhang, S., Li, Y., Feng, L., Fang, S., Lyu, C., Chen, K., Zhang, W.: Consistent-teacher: Towards reducing inconsistent pseudo-targets in semi-supervised object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3240–3249 (2023)
24. xiaolong: poppy dataset. `https://universe.roboflow.com/xiaolong-dpol1/poppy-dcv8q` (Dec 2025), `https://universe.roboflow.com/xiaolong-dpol1/poppy-dcv8q`, visited on 2026-02-28
25. Yan, W., Chen, L., Kou, H., Zhang, S., Zhang, Y., Cao, L.: Ucod-dpl: Unsupervised camouflaged object detection via dynamic pseudo-label learning. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 30365–30375 (2025)
26. Zhang, J., Zhang, R., Shi, Y., Cao, Z., Liu, N., Khan, F.S.: Learning camouflaged object detection from noisy pseudo label. In: European Conference on Computer Vision. pp. 158–174. Springer (2024)
27. Zhang, Z., Xia, W., Xie, G., Xiang, S.: Fast opium poppy detection in unmanned aerial vehicle (uav) imagery based on deep neural network. Drones **7**(9), 559 (2023)
28. Zhao, Y., Lv, W., Xu, S., Wei, J., Wang, G., Dang, Q., Liu, Y., Chen, J.: Detrs beat yolos on real-time object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16965–16974 (2024)
29. Zhou, X., Wang, D., Krähenbühl, P.: Objects as points. arXiv preprint arXiv:1904.07850 (2019)
30. Zhu, P., Wen, L., Du, D., Bian, X., Ling, H., Hu, Q., Nie, Q., Cheng, H., Liu, C., Liu, X., et al.: Visdrone-det2018: The vision meets drone object detection in image challenge results. In: Proceedings of the European conference on computer vision (ECCV) workshops. pp. 0–0 (2018)