What Limits Small Object AP in Drone/Bird Detection?

SeungJun Gwon Hyun Lee Seowoo Park
Seoul National University
{kylesjgwon6, hyun86, seowoo0805}@snu.ac.kr

Abstract

Small aerial objects are often visible enough to be detected coarsely but remain difficult to localize precisely. We study this gap in two-class drone/bird detection with a COCO-pretrained RT-DETRv2-S detector. After fine-tuning, RT-DETRv2-S reaches $\mathrm{AP}^{\mathrm{S}}_{50} = 0.8274$ *but only* $\mathrm{AP}^{\mathrm{S}} = 0.5531$*, yielding a much larger AP50-to-AP drop for small objects than for medium or large objects. Matched-box statistics show that small-object predictions have lower IoU and less stable box scale relative to ground truth. Motivated by this diagnosis, we compare a family of loss-level interventions for small-object localization, including scale-aware reweighting, scale-normalized localization, quality-aware localization reweighting, and box-scale bias regularization. The combined scale-aware and scale-normalized loss obtains the highest* AP^{S} *of* 0.5697*, while box-scale bias regularization gives the best overall trade-off, improving* AP^{S} *to* 0.5692 *and AP to* 0.8156 *without additional inference cost. These results indicate that strict-IoU localization and box-scale calibration are important residual errors in this drone/bird setting.*

1. Introduction

Detecting small aerial objects is important for surveillance, airport safety, disaster response, and autonomous monitoring. Although radar, RF, and infrared sensors can complement visual systems, RGB cameras remain attractive because they are inexpensive, dense, and easy to deploy. Their main weakness is scale: distant drones and birds often occupy merely a few pixels, making them hard to distinguish and even harder to localize tightly.

The challenge is not only recognizing the category. In many images, the target is visible as a compact blob with limited texture, so a detector may assign the correct class and still produce a loose box. This matters because detection is evaluated by both semantic correctness and bounding-box overlap, and the tolerance for boundary error is not the same across object sizes. Detectors trained on broad datasets such as MS COCO encounter objects spanning many scales, and the learned box regression generally performs well for medium and large instances. For small objects, however, the same AP evaluation standard imposes a much stricter absolute constraint: a one- or two-pixel boundary error that is negligible for a large object can substantially reduce IoU for a small one, pushing a qualitatively correct prediction below strict thresholds. This asymmetry is compounded by any global box-scale tendency inherited from pretraining. Even if such a tendency is sub-pixel in magnitude, its effect on IoU is disproportionately large for small objects, where the box area is already small and the denominator of the IoU calculation offers little tolerance for additional error.

Modern detectors, including DETR-style transformer detectors [2, 6, 14], perform strongly on generic benchmarks, yet small objects remain a persistent failure case. Existing work often improves small-object detection by increasing feature resolution, using multi-scale features, slicing images, or changing data augmentation [1, 4]. Loss-side alternatives such as IoU-balanced loss [12], ARUBA [9], Inner-IoU [13], and Wise-IoU [11] directly modify localization training. These strategies are effective when the detector misses small objects or needs stronger regression pressure. In this work, we focus on a different regime: the detector often finds the object at IoU 0.50, but its box is not accurate enough for higher thresholds.

A scale-wise discrepancy emerges after fine-tuning COCO-pretrained detectors on a two-class drone/bird dataset. A fine-tuned RT-DETRv2-S baseline achieves 0.9475 AP50 and 0.8130 AP overall, but its small-object metrics are $\mathrm{AP}^{\mathrm{S}}_{50} = 0.8274$ and $\mathrm{AP}^{\mathrm{S}} = 0.5531$. The corresponding gap, 0.2743, is much larger than the medium-object gap (0.0975) and the large-object gap (0.0603). This motivates the central question of this paper: when small aerial objects are already detected coarsely, what limits their strict-IoU AP?

Figure 1 visualizes this gap and motivates our diagnosis-driven study of loss-level small-object localization. We first quantify the scale-wise AP50-to-AP gap and analyze matched prediction boxes. The analysis shows that small-object matches have lower mean IoU and less stable box

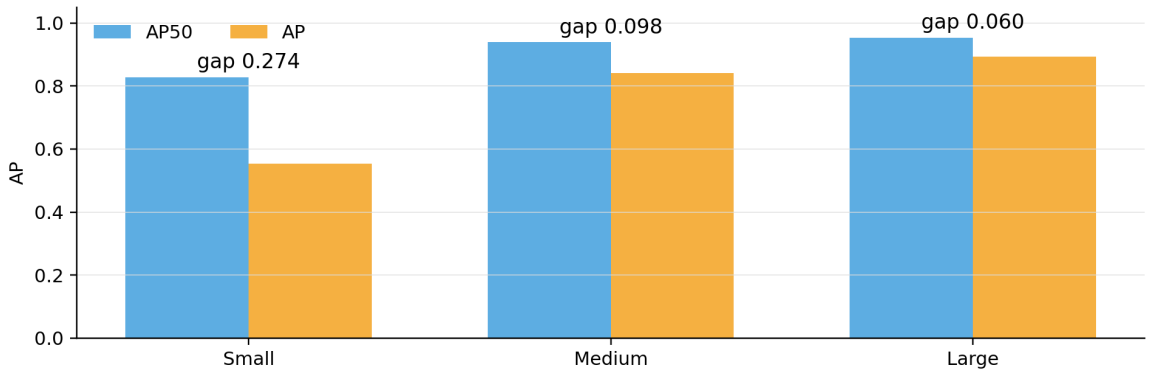

Figure 1. Motivating metric gap on the fine-tuned RT-DETRv2-S baseline. Small objects already achieve high AP^{S}_{50}, but their AP50-to-AP drop is much larger than that of medium and large objects. This figure establishes the central failure mode studied in the paper: small aerial objects are often detected coarsely but not localized tightly enough for strict IoU thresholds.

scale than medium and large objects. Based on this observation, we construct and compare several loss interventions that target different parts of the failure: ARUBA-style scale reweighting [9], Inner-IoU-style scale-normalized localization [13], quality-aware localization reweighting [11, 12], and box-scale bias regularization in RT-DETRv2.

The box-scale bias variant follows directly from RT-DETRv2 box refinement. The final width and height are predicted as logit-space offsets added to reference boxes. The bias of the final box-head layer therefore creates a feature-independent tendency to expand or shrink predicted boxes. Constraining only the width/height bias targets this global box-scale drift while leaving feature-dependent box refinement intact.

Our contributions are as follows:

- We identify a large AP50-to-AP gap for small drone/bird objects after fine-tuning, showing that strict localization remains a bottleneck even when coarse detection is strong.
- We quantify the box geometry behind this gap with scale-wise no-detection rate, matched IoU, and predicted-to-ground-truth area ratio.
- We study multiple loss-level interventions for small-object localization under a shared evaluation protocol. The comparison shows modest but consistent gains: scale-aware normalized localization gives the largest AP^{S} improvement, while box-scale bias regularization gives the strongest overall trade-off.

2. Related Work

Object detection models. Modern object detectors are commonly built around either dense one-stage prediction or query-based set prediction. YOLO-style detectors provide a representative one-stage line of work for real-time detection [7], while DETR introduced object detection as set prediction with transformer decoders and Hungarian matching [2], replacing duplicate dense predictions with a one-to-one assignment objective. RT-DETR adapts this formulation for real-time detection through efficient hybrid encoding and query selection [14], while RT-DETRv2 further adds training and deployment refinements [6]. We focus on RT-DETRv2-S because its iterative box refinement and decoder box-head structure make box-scale errors directly analyzable, allowing loss-side calibration to be studied without changing the inference graph.

Small object detection. Small object detection is commonly addressed by improving multi-scale representations, preserving high-resolution details, or changing the effective object scale. Feature Pyramid Networks [4] construct semantically strong features across resolutions and remain a standard design choice for scale-robust detection. In aerial imagery, datasets such as VisDrone [16] highlight the difficulty of dense and small targets, while slicing-based methods such as SAHI [1] improve small-object visibility by evaluating cropped high-resolution regions. These approaches are effective but often modify the architecture, input resolution, or inference procedure. Our study instead fixes the inference graph and asks whether the training objective alone can improve strict localization AP for small drone/bird objects.

Localization losses and box calibration. Detection losses often reweight examples or align training objectives with evaluation metrics. Focal Loss [5] addresses dense classification imbalance, while bounding-box losses focus on localization quality. GIoU [8] extends IoU loss to non-overlapping boxes, and DIoU and CIoU [15] add center-distance and aspect-ratio terms. Balanced-loss variants such as Scale-Balanced Loss [10], IoU-balanced loss [12], and ARUBA [9] reweight training contributions according to object scale or localization difficulty. Auxiliary-box and

scale-normalized variants such as Inner-IoU [13], together with dynamic focusing losses such as Wise-IoU [11], further adapt bounding-box regression according to localization quality. These losses improve generic localization, but they do not explicitly separate coarse detection success from strict-IoU failure on small objects. For small boxes, the same absolute boundary error produces a larger relative IoU penalty, motivating the scale-aware, scale-normalized, quality-aware, and box-scale regularized objectives compared in this work.

Bias and parameter regularization. Weight decay is usually applied broadly to neural network parameters to control model capacity, but task-specific parameter regularization can encode more targeted assumptions. In detection heads, the final prediction layer can introduce global tendencies that are independent of image evidence. For RT-DETRv2 box heads, width and height bias terms directly shift box-size logits after reference boxes are incorporated. We therefore include width/height bias regularization as one member of the loss-intervention family: it targets a measurable box-scale failure mode while preserving the rest of the model's representational capacity.

Evaluation of small-object localization. Small-object performance is often summarized by AP^{S}, but this single value hides whether the main failure comes from missed detections, classification confusion, or localization error. COCO-style evaluation [3] reports AP averaged over IoU thresholds together with AP50 and scale-wise AP, which makes it possible to separate coarse detection from strict localization. We use this convention not only as an evaluation metric but also as a diagnostic tool: comparing $\mathrm{AP}^{\mathrm{S}}_{50}$ with AP^{S} identifies a localization gap, and box-relation statistics then measure whether the gap corresponds to missed objects, loose boxes, or scale bias.

3. Preliminaries

3.1. COCO-Style AP and Scale-Wise AP

A detector outputs a set of predicted class labels, confidence scores, and boxes,

$$D = \{(\hat{c}_i, \hat{s}_i, \hat{b}_i)\}_{i=1}^{N}. \tag{1}$$

For a fixed IoU threshold t, predictions are sorted by confidence and matched to ground-truth boxes of the same class. A prediction is a true positive when its IoU with an unmatched ground-truth box is at least t. The AP at threshold t is the area under the corresponding precision–recall curve,

$$\mathrm{AP}_t = \int_0^1 p_t(r)\, dr. \tag{2}$$

COCO-style AP [3] averages over ten IoU thresholds,

$$\mathrm{AP} = \frac{1}{10} \sum_{t \in \{0.50, 0.55, \ldots, 0.95\}} \mathrm{AP}_t. \tag{3}$$

We also report AP_{50} and AP_{75} for thresholds 0.50 and 0.75.

Scale-wise AP is computed by filtering ground-truth boxes according to their area $A = wh$ in pixels:

$$\begin{aligned} \text{small}: &\quad A < 32^2, \\ \text{medium}: &\quad 32^2 \le A < 96^2, \\ \text{large}: &\quad A \ge 96^2. \end{aligned} \tag{4}$$

The resulting metrics are denoted by AP^{S}, AP^{M}, and AP^{L}. The same filtering can be applied at a single IoU threshold, yielding $\mathrm{AP}^{\mathrm{S}}_{50}$, $\mathrm{AP}^{\mathrm{M}}_{50}$, and $\mathrm{AP}^{\mathrm{L}}_{50}$ for the AP at IoU threshold 0.50 within each scale group.

3.2. IoU Sensitivity of Small Boxes

The sensitivity of AP to localization error increases as object size decreases. Consider two square boxes of side length s where the prediction is shifted by δ pixels along one axis and has the same size as the ground truth. If $0 \le \delta < s$, their IoU is

$$\mathrm{IoU}(s, \delta) = \frac{s(s-\delta)}{s^2 + s\delta} = \frac{s-\delta}{s+\delta}. \tag{5}$$

For a fixed absolute error δ, this quantity decreases faster when s is small. A similar effect occurs for width and height errors. If a predicted box has the same center as the ground truth but its side length is scaled by $\rho > 1$, then

$$\mathrm{IoU}(s, \rho) = \frac{1}{\rho^2}. \tag{6}$$

Thus a modest over-expansion can already prevent a small object from satisfying high IoU thresholds. These simple cases explain why AP averaged over thresholds up to 0.95 is much stricter than AP50 for small objects.

3.3. RT-DETRv2 Box Refinement

RT-DETRv2 predicts boxes by iteratively refining decoder reference boxes. At decoder layer ℓ, let h_ℓ be the query feature and $R_\ell \in [0,1]^4$ be the reference box in normalized (c_x, c_y, w, h) format. The box MLP predicts a four-dimensional logit offset,

$$\Delta_\ell = f_\ell(h_\ell) \in \mathbb{R}^4. \tag{7}$$

The refined prediction is

$$\hat{B}_\ell = \sigma\left(\Delta_\ell + \sigma^{-1}(R_\ell)\right), \tag{8}$$

where σ is the sigmoid function and σ^{-1} is the inverse sigmoid.

Writing the last linear layer of the box head as

$$\Delta_\ell = W_\ell \phi_\ell(h_\ell) + b_\ell, \tag{9}$$

the width and height channels are

$$\hat{w}_\ell = \sigma(\Delta_\ell^w + \sigma^{-1}(R_\ell^w)),$$
$$\hat{h}_\ell = \sigma(\Delta_\ell^h + \sigma^{-1}(R_\ell^h)). \tag{10}$$

Thus the bias terms b_ℓ^w and b_ℓ^h provide feature-independent offsets to the predicted width and height logits. When $\Delta_\ell^w = \Delta_\ell^h = 0$, the predicted width and height equal the reference width and height.

3.4. Detection Loss

RT-DETRv2 uses Hungarian matching to assign predictions to ground-truth boxes:

$$\pi = \arg\min_\sigma \sum_j \mathcal{C}(y_j, \hat{y}_{\sigma(j)}). \tag{11}$$

Throughout the loss definitions, ground-truth boxes and labels are written without decoration, while predicted quantities are denoted with hats. For a matched pair $(\hat{B}, B)$, the base detection loss is

$$\mathcal{L}_{\text{det}} = \lambda_{\text{cls}}\mathcal{L}_{\text{cls}} + \lambda_1\|\hat{B} - B\|_1 + \lambda_{\text{giou}}\mathcal{L}_{\text{giou}}(\hat{B}, B). \tag{12}$$

In our RT-DETRv2 configuration, $\lambda_{\text{cls}} = 1$, $\lambda_1 = 5$, and $\lambda_{\text{giou}} = 2$. Auxiliary losses are applied to intermediate decoder outputs, encoder outputs, and denoising-query outputs with the same form.

4. Method

4.1. Problem Formulation and Diagnosis

Our goal is to improve strict localization for small aerial objects while keeping the detector architecture and inference procedure unchanged. Let $a \in \{S, M, L\}$ denote the COCO scale group. We use the difference between AP at IoU 0.50 and AP averaged over stricter thresholds as the primary diagnostic signal.

Scale-wise localization gap. For each scale group $a \in \{S, M, L\}$, we define the AP50-to-AP gap as

$$G_a = \text{AP}_{a,50} - \text{AP}_a. \tag{13}$$

Since $\text{AP}_{a,50}$ accepts coarse localization while AP_a averages stricter IoU thresholds up to 0.95, a large G_a indicates that predictions often reach the object region but fail precise box localization.

Box-relation statistics. For each ground-truth box g_j, we select the same-class prediction with the largest IoU,

$$p_j = \arg\max_{i:\hat{c}_i = c_j} \text{IoU}(\hat{b}_i, g_j). \tag{14}$$

If no prediction reaches IoU 0.50, the object is counted as not detected. For matched pairs, we compute mean IoU and area ratio

$$r_j = \frac{\text{area}(\hat{b}_{p_j})}{\text{area}(g_j)}. \tag{15}$$

We summarize whether $r_j > 1$, $r_j < 1$, and whether the match reaches IoU 0.75. These statistics separate missed detections from inaccurate localization and box-scale errors.

The diagnosis produces two actionable signals. First, a large G_S indicates that small objects need stronger optimization at strict IoU thresholds, not merely higher coarse recall. Second, scale-dependent changes in r_j indicate a possible box-scale calibration problem. We therefore study loss-level interventions along two axes: increasing strict-localization pressure for small objects and calibrating predicted box scale.

4.2. Loss-Level Interventions

We consider loss modifications that keep the inference graph unchanged, act primarily on box localization, and avoid adding a small-object-specific detection branch. The point is not to redesign the detector but to test whether the training objective can better handle the observed small-object localization failure. Many alternative methods improve recall through high-resolution branches, slicing, or multi-stage refinement, but those changes also alter runtime and memory. Here we keep the architecture fixed and compare only training objectives. The names used below denote our implementations of broader loss families, not exact reimplementations of every cited method.

Scale-aware regression reweighting. The first intervention increases the contribution of small objects in the regression term. For matched object j with scale-dependent weight α_j, the weighted box loss is

$$\mathcal{L}_{\text{box}} = \sum_j \alpha_j \left(\lambda_1\|\hat{B}_j - B_j\|_1 + \lambda_{\text{giou}}\mathcal{L}_{\text{giou}}(\hat{B}_j, B_j)\right). \tag{16}$$

Here α_j is larger for small objects and closer to one for medium and large objects. This tests whether the low AP^{S} is mainly caused by small objects being under-weighted during training.

In implementation, boxes are represented in normalized (c_x, c_y, w, h) coordinates after resizing to 640×640. Motivated by scale-balanced reweighting losses for aerial objects [9], we use a scale-sensitive gain coefficient

$$\alpha_j = 1 + \alpha \cdot \text{clip}\left(\frac{\tau - w_j h_j}{\tau}, 0, 1\right), \tag{17}$$

where $\tau = 0.0025$ corresponds to the COCO small-object boundary $32^2/640^2$. The scale-aware setting uses $\alpha = 0.25$ and adds a normalized Wasserstein distance term with weight 0.25 plus a relative width/height scale term with weight 0.02.

Scale-normalized localization. The second intervention changes the unit of coordinate error. Let s_j denote a size

statistic of ground-truth box B_j, such as $\sqrt{w_j h_j}$. A normalized localization term can be written as

$$\mathcal{L}_{\text{snl}} = \sum_j \frac{\|\hat{B}_j - B_j\|_1}{s_j + \epsilon}. \tag{18}$$

This gives a larger penalty to the same absolute coordinate error when the ground-truth object is small. It directly matches the observation that a few pixels of error are more harmful for small boxes under strict IoU thresholds.

Motivated by scale-sensitive localization losses such as Inner-IoU [13], the implemented SNL term uses normalized boxes and decomposes the error into four normalized components:

$$\mathcal{L}_{\text{snl}} = \sum_j \alpha_j \sum_{u \in \{x,y,w,h\}} \text{SmoothL1}(\ell_j^u), \tag{19}$$

where the normalized residuals are

$$\begin{aligned} \ell_j^x &= \frac{\hat{c}_j^x - c_j^x}{w_j}, \quad \ell_j^y = \frac{\hat{c}_j^y - c_j^y}{h_j}, \\ \ell_j^w &= \log \frac{\hat{w}_j}{w_j}, \quad \ell_j^h = \log \frac{\hat{h}_j}{h_j}. \end{aligned} \tag{20}$$

The standalone SNL run uses $\alpha = 0.5$ and SNL weight 0.25. The combined scale-aware+SNL run uses $\alpha = 0.25$, NWD weight 0.25, relative-scale weight 0.02, and SNL weight 0.10.

Quality-aware localization reweighting. The third intervention follows the broader idea of reweighting localization examples according to localization quality. In experiments, we instantiate this family with Wise-IoU [11]. This baseline tests whether a generic quality-aware localization loss is sufficient, or whether the drone/bird failure requires scale- or box-bias-specific calibration.

Our Wise-IoU implementation reweights the GIoU loss by a detached quality-dependent factor [11]. Let ℓ_{iou} be the per-match IoU loss and $\beta = \ell_{\text{iou}}/\bar{\ell}_{\text{iou}}$. The focusing factor is clipped to $[0.25, 3.0]$ after applying the Wise-IoU dynamic form with $\delta = 3.0$ and $\gamma = 1.9$. This variant does not explicitly use object size.

Box-scale bias regularization. The final intervention targets the box-scale bias suggested by the matched-box analysis. Let $\mathcal{B}_{wh}$ be the set of width and height bias parameters from RT-DETRv2 box heads, and let $b^{(0)}$ denote the corresponding pretrained value. We introduce a pretrained-bias-anchored regularizer:

$$\mathcal{L}_{\text{bias}} = \lambda_b \sum_{b \in \mathcal{B}_{wh}} \left\| \text{ReLU}\left(b - b^{(0)}\right) \right\|_2^2 \tag{21}$$

to the detection objective. This penalizes feature-independent positive drift in width/height logits while preserving the standard box regression loss. For the bias-regularized variant, the final training objective is

$$\mathcal{L} = \mathcal{L}_{\text{det}} + \mathcal{L}_{\text{bias}}. \tag{22}$$

For the reported bias-regularized runs, $\lambda_b \in \{0.25, 0.5\}$. The one-sided penalty avoids forcing the pretrained bias to zero and instead discourages additional positive drift in width and height during target-domain fine-tuning. All RT-DETRv2 loss modifications are applied to the main decoder output and to the auxiliary decoder, encoder, and denoising outputs when those losses are present in the baseline criterion.

Design rationale. The four interventions cover complementary hypotheses. Scale-aware reweighting tests whether small objects are under-optimized. Scale-normalized localization tests whether absolute coordinate loss underpenalizes small-box boundary errors. Quality-aware localization tests whether a generic localization-quality weighting is sufficient. Bias regularization tests whether a detector-level width/height prior contributes to loose small-object boxes. Together, these variants form a controlled loss-level comparison rather than a change to the detector architecture.

The variants also differ in how strongly they encode the small-object prior. Wise-IoU is the least dataset-specific because it depends on localization quality rather than object scale. Scale-aware reweighting and SNL explicitly encode object size and therefore directly target AP^{S}. Bias regularization does not identify small objects individually; instead, it constrains a global box-size degree of freedom that can affect all predictions. Comparing these variants helps distinguish whether the observed failure is best addressed by per-object scale weighting or by detector-level box calibration.

Why regularize bias instead of all box parameters. Regularizing all box-head weights would constrain feature-dependent localization capacity. In contrast, the final width/height bias affects every prediction through the same logit offset and can encode a global box-size prior. When matched-box analysis indicates scale-dependent box-size error, constraining this prior is more targeted than penalizing the entire box head. The regularizer is applied only to width and height channels, not to center coordinates, because the measured failure is related to box scale rather than a consistent shift direction.

Applicability. The proposed loss modifications are used only during fine-tuning and leave the detector architecture, decoding procedure, and inference cost unchanged. After training, the model is evaluated with the standard RT-DETRv2 prediction pipeline.

5. Experiments

5.1. Dataset and Metrics

We evaluate on a two-class drone/bird object detection dataset. The dataset contains 7,165 training images, 796 validation images, and 2,691 test images. All results are reported on the same drone/bird test set. The two classes are visually related in the sense that both can appear as compact aerial objects, but the detector must still distinguish drone and bird categories while localizing each instance.

The training split contains 8,773 annotated instances, with 6,828 birds and 1,945 drones. The test split contains 3,061 instances, with 1,711 birds and 1,350 drones. By the COCO area convention [3], the test set contains 828 small, 1,001 medium, and 1,232 large objects. This distribution is useful for interpreting the scale-wise metrics: small objects are not a negligible tail of the benchmark, but a substantial portion of the evaluation set.

The main metrics are COCO-style AP, AP50, AP75, and scale-wise AP for small, medium, and large objects. We report both AP^{S}, AP^{M}, AP^{L} and their AP50 counterparts. Since the central question is strict localization, we emphasize the difference between AP50 and AP for each scale group. All AP values are computed on the same test annotations, so comparisons across loss variants are not affected by a change in evaluation split.

5.2. Implementation Details

We use a COCO-pretrained RT-DETRv2-S [6] checkpoint as initialization and fine-tune it on the drone/bird training split. Unless stated otherwise, all comparisons use the same test split and class set.

The model is trained and evaluated at 640×640 input resolution. RT-DETRv2-S is fine-tuned for 50 epochs using AdamW with learning rate 10^{-4}, batch size 16, mixed precision, EMA, and the standard RT-DETRv2 augmentation pipeline of photometric distortion, zoom-out, IoU crop, horizontal flip, and resize [6]. Validation is performed on the held-out validation split during training, and all reported results are computed on the test split.

Loss modifications are applied during fine-tuning from the same pretrained initialization. The bias-regularized variant is applied to the width and height bias channels of the decoder box heads. It does not introduce new inference-time operations, and the trained model is evaluated with the standard RT-DETRv2 decoding pipeline.

5.3. Transfer and Fine-Tuning Baselines

Table 1 compares pretrained-only evaluation and the fine-tuned baseline. Fine-tuning substantially improves the detector, but small-object AP remains the weakest scale-wise metric.

The COCO-pretrained model achieves non-trivial detection on this dataset for two reasons: MS COCO includes a bird class, so the pretrained model transfers directly to bird instances, and drones share visual appearance with small aerial objects in COCO at comparable scales. Despite this, the pretrained AP remains far below the fine-tuned baseline, confirming that domain-specific fine-tuning is necessary. All subsequent loss analyses in this paper use the same fine-tuned RT-DETRv2-S as the common starting point, so that differences between loss variants reflect only the training objective and not the initialization.

5.4. Diagnostic Analysis

The AP50-to-AP gap for the fine-tuned RT-DETRv2-S baseline is 0.2743 for small objects, compared with 0.0975 for medium objects and 0.0603 for large objects. If small objects were mostly missed, both $\mathrm{AP}^{\mathrm{S}}_{50}$ and AP^{S} would be low. Instead, $\mathrm{AP}^{\mathrm{S}}_{50}$ is already above 0.82, while AP^{S} is only 0.5531. This pattern indicates that the failure is not one of classification or coarse localization — the detector regularly assigns a high-confidence prediction to the correct object region — but rather one of bounding box regression precision. The predicted box finds the object but does not tighten around it accurately enough to survive strict IoU thresholds such as 0.75, 0.85, or 0.95 [3]. This is consistent with the IoU sensitivity analysis in Section 3.2: a model trained across a broad range of object sizes optimizes regression over the full scale distribution, but the same boundary deviation that is tolerable for a medium or large prediction incurs a disproportionately large IoU penalty when the object is small. The much smaller medium and large gaps confirm that this is a scale-dependent localization failure rather than a general weakness of the detector.

We further analyze matched prediction boxes for the fine-tuned RT-DETRv2-S baseline. Figure 2 summarizes scale-wise box relations, and Figure 3 visualizes small-object cases with loose predicted boxes. Small objects have a higher no-detection rate, lower matched IoU, and less stable predicted box scale than medium and large objects.

The no-detection rate decreases monotonically with object size, from 3.9% for small objects to 1.1% for large objects. More importantly, even detected small objects have lower mean IoU. Their mean predicted-to-ground-truth area ratio is close to one, but the rates of both larger and smaller predictions are much higher than those of medium and large objects. This pattern motivates box-scale calibration in the loss comparison.

5.5. Comparison of Loss-Level Interventions

Table 3 compares the loss-level interventions considered in this work: generic quality-aware reweighting (Wise-IoU [11]), explicit small-object reweighting motivated by scale-balanced losses [9], scale-normalized localization

Setting	AP	AP50	AP75	AP^S	AP^S_{50}	AP^M	AP^M_{50}	AP^L	AP^L_{50}
COCO pretrained	0.3869	0.6910	0.2961	0.1839	0.3232	0.4039	0.5592	0.5279	0.8257
Fine-tuned	0.8130	0.9475	0.8616	0.5531	0.8274	0.8418	0.9393	0.8940	0.9543

Table 1. Pretrained-only and fine-tuned RT-DETRv2-S performance on the drone/bird test set.

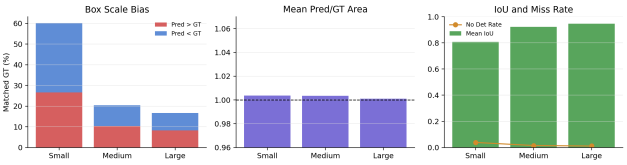

Figure 2. Scale-wise box-relation diagnosis for the fine-tuned RT-DETRv2-S baseline. The figure connects the AP50-to-AP gap to box geometry: small objects have lower matched IoU, a higher no-detection rate, and less stable predicted box scale than medium and large objects.

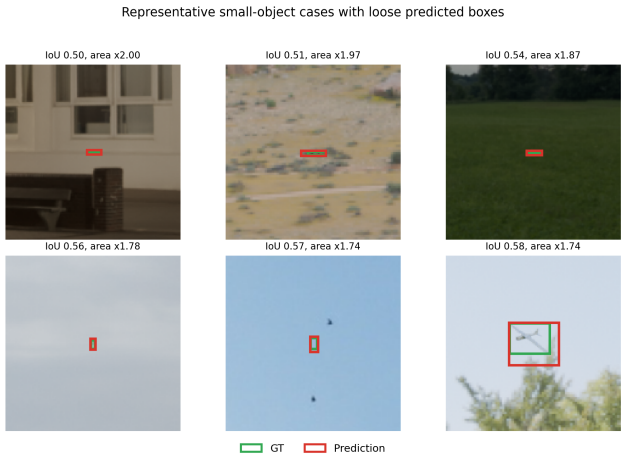

Figure 3. Representative small-object localization failures. The predictions cover the target and would often be counted at coarse IoU thresholds, but the boxes are loose enough to fall below stricter localization thresholds.

Scale	GT	No Det.	Mean IoU	Area	Larger	Smaller
Small	828	3.9%	0.809	1.004	26.6%	33.5%
Medium	1001	1.5%	0.923	1.003	10.3%	10.0%
Large	1232	1.1%	0.947	1.001	8.3%	8.4%

Table 2. Box-relation statistics for the fine-tuned RT-DETRv2-S baseline. "Area" denotes mean predicted-to-ground-truth area ratio for matched detections.

motivated by Inner-IoU [13], their combination, and box-scale bias regularization. This comparison is intended to identify which kind of loss modification is most effective for small drone/bird localization under a fixed inference graph.

Scale-aware and scale-normalized losses individually provide only small AP^S gains. Their combination is more effective, improving AP^S from 0.5531 to 0.5697. Wise-IoU does not improve this dataset, while bias regularization gives a more favorable trade-off: $\lambda_b = 0.25$ obtains the best overall AP and AP50, and $\lambda_b = 0.5$ obtains nearly the best AP^S and the best large-object AP.

The strongest absolute AP^S comes from the scale-aware plus SNL variant, while the strongest overall AP comes from mild bias regularization. Relative to the fine-tuned RT-DETRv2-S baseline, Wise-IoU [11] slightly lowers AP^S, scale-aware reweighting [9] and SNL [13] provide small gains, and the combined scale-aware+SNL setting gives the largest AP^S improvement. Bias regularization with $\lambda_b = 0.25$ gives the best overall AP and AP50, while $\lambda_b = 0.5$ gives nearly the best AP^S and the best large-object AP.

Table 4 ablates the bias regularizer form. Pulling the width/height bias directly toward zero is weaker: L1 improves AP^S but lowers AP^S_{50}, and L2 slightly reduces AP^S. Our anchored regularizer is more stable because it keeps the pretrained width/height prior as the anchor and only penalizes additional positive drift.

A class-wise check gives the same caution. For RT-DETRv2-S, the baseline already has much higher drone AP^S than bird AP^S (0.6533 vs. 0.4529), reflecting that small birds are particularly difficult despite the larger number of bird instances. The scale-aware+SNL variant improves both class-specific small-object AP values to 0.6602 for drone and 0.4792 for bird. Thus the aggregate AP^S improvement is not only a class-prior artifact; it also improves the harder bird-small subset.

5.6. Qualitative Diagnostic

Figure 3 is used as a diagnostic visualization rather than a detector showcase. It focuses on matched small-object predictions where the object is found but the box is spatially loose. This is the visual counterpart of the quantitative box-relation statistics in Table 2: the failure is not simply a missing detection, but a localization and box-scale error that becomes costly under high IoU thresholds.

This diagnostic also explains why AP50 alone is insufficient for the task. A prediction can clearly cover the object and still fail at stricter thresholds if the box includes too much background. For safety-sensitive drone detection, coarse localization may be enough for alerting, but precise localization is important for tracking, handoff to downstream modules, and estimating object motion. Therefore, improving AP^S rather than only AP^S_{50} is a meaningful ob-

Method	AP	AP50	AP75	AP^S	AP^S_{50}	AP^M	AP^M_{50}	AP^L	AP^L_{50}
Baseline	0.8130	0.9475	0.8616	0.5531	0.8274	0.8418	0.9393	0.8940	0.9543
Wise-IoU [11]	0.8091	0.9430	0.8561	0.5496	0.8020	0.8373	0.9345	0.8884	0.9506
Scale-aware	0.8125	0.9453	0.8580	0.5559	0.8203	0.8403	0.9363	0.8945	0.9549
SNL	0.8130	0.9465	**0.8642**	0.5573	0.8221	0.8351	0.9333	0.8917	0.9510
Scale-aware + SNL	0.8139	0.9456	0.8587	**0.5697**	0.8351	0.8418	0.9308	0.8906	0.9487
Bias reg. $\lambda_b = 0.25$	**0.8169**	**0.9493**	0.8627	0.5660	**0.8469**	**0.8470**	0.9412	0.8918	0.9538
Bias reg. $\lambda_b = 0.5$	0.8156	0.9476	0.8634	0.5692	0.8320	0.8454	0.9362	**0.8963**	**0.9583**

Table 3. Comparison of RT-DETRv2-S loss-level interventions on the drone/bird test set. Scale-aware and SNL denote our implementations motivated by scale-balanced reweighting [9] and Inner-IoU [13], respectively. Bias reg. denotes our proposed pretrained-bias-anchored width/height regularizer in Section 4.

Bias regularizer	AP	AP^S	AP^S_{50}	ΔAP^S
None	0.8130	0.5531	0.8274	–
L1	0.8145	0.5621	0.8256	+0.0090
L2	0.8134	0.5501	0.8208	-0.0030
Ours, $\lambda_b = 0.25$	**0.8169**	0.5660	**0.8469**	+0.0129
Ours, $\lambda_b = 0.5$	0.8156	**0.5692**	0.8320	**+0.0161**

Table 4. Ablation of width/height bias regularization forms on RT-DETRv2-S. L1 and L2 regularize the width/height bias toward zero. Our regularizer anchors the bias at the pretrained value and penalizes only positive width/height drift.

jective.

6. Discussion

The experiments indicate that small-object AP is limited by strict localization more than by AP50-level detection alone. The tested losses improve this bottleneck in different ways: scale-aware and scale-normalized losses increase the relative penalty for small-object coordinate error, while box-scale bias regularization constrains a global width/height prior in the detector head. The combined scale-aware and scale-normalized variant gives the largest AP^S gain, and bias regularization gives the strongest overall AP trade-off without increasing runtime.

The comparison also suggests a useful design principle. If the main bottleneck is missing small objects, architecture or data changes may be necessary. If the main bottleneck is the AP50-to-AP gap, loss-level localization calibration can be effective. In the drone/bird dataset, the high AP^S_{50} value indicates that the detector often has enough semantic evidence to find small objects. The remaining performance gap is therefore more naturally addressed by regression losses and box calibration.

7. Limitations

The experiments focus on one detector family and one drone/bird dataset. The bias-regularized variant is formulated specifically for RT-DETRv2-style box heads. Applying the same idea to other architectures requires identifying the parameter that plays an analogous role in box-scale prediction.

The current analysis also focuses on bounding-box geometry rather than temporal consistency or multi-sensor information. In real drone monitoring systems, video tracking, camera motion, and additional sensor cues can provide complementary information. These directions are beyond the loss-level calibration considered here.

Finally, all loss-variant results are single-run measurements. We use the same initialization, split, and training schedule to make comparisons controlled, but we do not report multi-seed variance. The numerical gains should therefore be interpreted as diagnostic trends rather than statistically definitive improvements.

8. Conclusion

Scale-wise AP gaps and box-relation statistics reveal that fine-tuned drone/bird detectors achieve high AP50 but still suffer large $AP^S_{50} - AP^S$ gaps. This is not because the model fails to classify or coarsely locate small objects, but because the bounding box precision that is acceptable for medium and large objects is insufficient at small scales. The AP^S_{50} above 0.82 confirms that the detector regularly finds them. The same degree of box imprecision that passes strict IoU thresholds for larger predictions incurs a disproportionately large penalty for small ones, resulting in less stable box scale and less precise coordinate predictions for the small-object subset. Small-object predictions are less tightly localized than medium and large objects and show less stable box scale. We therefore evaluate several loss-level interventions for small-object localization, including scale-aware reweighting, scale-normalized localization, quality-aware localization, and box-scale bias regularization. The best AP^S comes from combining scale-aware and scale-normalized losses, while width/height bias regularization gives the best overall trade-off in this single-run study. These results suggest that strict-IoU localization and box-scale calibration are useful directions for improving small aerial object detectors beyond coarse recall.

References

[1] Fatih Cagatay Akyon, Sinan Onur Altinuc, and Alptekin Temizel. Slicing aided hyper inference and fine-tuning for small object detection. In *IEEE Int. Conf. Image Process.*, 2022. 1, 2

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Eur. Conf. Comput. Vis.*, 2020. 1, 2

[3] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *Eur. Conf. Comput. Vis.*, 2014. 3, 6

[4] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017. 1, 2

[5] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Int. Conf. Comput. Vis.*, 2017. 2

[6] Wenyu Lv, Yian Zhao, Qinyao Chang, Kui Huang, Guanzhong Wang, and Yi Liu. RT-DETRv2: Improved baseline with bag-of-freebies for real-time detection transformer. *arXiv preprint arXiv:2407.17140*, 2024. 1, 2, 6

[7] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016. 2

[8] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. 2

[9] Rebbapragada V C Sairam, Monish Keswani, Uttaran Sinha, Nishit Shah, and Vineeth N Balasubramanian. ARUBA: An architecture-agnostic balanced loss for aerial object detection. *arXiv preprint arXiv:2210.04574*, 2022. 1, 2, 4, 6, 7, 8

[10] Kai Shuang, Zhiheng Lyu, Jonathan Loo, and Wentao Zhang. Scale-balanced loss for object detection. *Pattern Recognition*, 117:107997, 2021. 2

[11] Zanjia Tong, Yuhang Chen, Zewei Xu, and Rong Yu. Wise-iou: Bounding box regression loss with dynamic focusing mechanism. *arXiv preprint arXiv:2301.10051*, 2023. 1, 2, 3, 5, 6, 7, 8

[12] Shengkai Wu, Jinrong Yang, Xinggang Wang, and Xiaoping Li. Iou-balanced loss functions for single-stage object detection. *arXiv preprint arXiv:1908.05641*, 2019. 1, 2

[13] Hao Zhang, Cong Xu, and Shuaijie Zhang. Inner-iou: More effective intersection over union loss with auxiliary bounding box. *arXiv preprint arXiv:2311.02877*, 2023. 1, 2, 3, 5, 7, 8

[14] Yian Zhao, Wenyu Lv, Shangliang Xu, Jinman Wei, Guanzhong Wang, Qingqing Dang, Yi Liu, and Jie Chen. DETRs beat YOLOs on real-time object detection. *arXiv preprint arXiv:2304.08069*, 2023. 1, 2

[15] Zhaohui Zheng, Ping Wang, Wei Liu, Jinze Li, Rongguang Ye, and Dongwei Ren. Distance-iou loss: Faster and better learning for bounding box regression. In *AAAI*, 2020. 2

[16] Pengfei Zhu, Longyin Wen, Xiao Bian, Haibin Ling, and Qinghua Hu. Vision meets drones: A challenge. *arXiv preprint arXiv:1804.07437*, 2018. 2