QAD-D-FINE: QUERY-ADAPTIVE SAMPLING AND REDUNDANCY SUPPRESSION FOR DENSE SMALL OBJECT DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Dense small object detection is challenging because weak fine-grained cues and crowded layouts jointly cause missed objects and redundant responses. We propose QAD-D-FINE, an extension of D-FINE that combines query-adaptive sampling and redundancy suppression for dense small-object scenes. QAD-D-FINE contains two complementary components. Query-Adaptive Sampling (QAS) augments the decoder with a high-resolution P2 feature and continuously adjusts deformable sampling offsets according to the incoming query scale and confidence, making sampling more local for confident small-object queries while keeping the scale factor closer to one for low-confidence queries. Matching-based Redundant Query Suppression (MRS) uses the standard one-to-one matching result as an anchor, assigns unmatched queries to same-class ground-truth instances using a geometry-normalized ownership criterion, and jointly suppresses redundant classification and localization responses during training. On VisDrone, QAD-D-FINE reaches 33.1 AP and 23.6 AP_{small}, improving D-FINE by 1.7 and 3.3 points, respectively, while retaining 166.7 FPS with 10.1M parameters at 640×640 input. Experiments on SODA-D and UAVDT further demonstrate consistent gains in dense small-object scenarios.

1 INTRODUCTION

Dense small objects are common in aerial imagery, traffic monitoring, and automated visual inspection. Small objects often occupy only a few pixels in an image, so texture and boundary cues are easily weakened by downsampling. At the same time, closely spaced instances make local aggregation fragile: a small spatial deviation can mix neighboring objects or background responses. Consequently, the difficulty is not only object scale, but also the coupling among detail preservation, context selection, and instance separation.

Modern detectors have improved both accuracy and efficiency. Convolutional detectors such as Faster R-CNN and FCOS (Ren et al., 2015; Tian et al., 2019), the YOLO family (Terven et al., 2023; Wang et al., 2024), and real-time transformer detectors such as RT-DETR (Zhao et al., 2024) serve as widely used general-purpose baselines. DETR-style models formulate detection as set prediction with bipartite matching (Carion et al., 2020); subsequent methods improve sparse sampling, query priors, denoising, and initialization (Zhu et al., 2021; Liu et al., 2022; Li et al., 2022; Zhang et al., 2023). DQ-DETR explicitly targets tiny objects with dynamic queries (Huang et al., 2024), while D-FINE improves localization through fine-grained distribution refinement and localization self-distillation (Peng et al., 2025).

Despite its strong real-time performance, D-FINE has two limitations relevant to dense small-object scenes. First, its reference-box offset rule is conditioned on the reference box but lacks a continuous query-state-dependent mechanism that adapts sampling range to scale and confidence. Second, after one-to-one matching, some unmatched high-response queries can remain clustered around an already matched object and produce redundant responses. These two issues arise at different stages: feature acquisition and post-matching supervision.

To address these issues, we introduce **QAD-D-FINE**, which improves D-FINE along these two axes. Our contributions are threefold: (1) we propose a dense small-object detection framework that

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

combines query-adaptive feature sampling with matching-aware redundancy suppression; (2) we design **QAS**, which adds P2 features and continuously rescales decoder offsets using the incoming query state, and **MRS**, which regularizes redundant unmatched queries without changing the inference graph; and (3) we conduct controlled ablations, mechanism analyses, efficiency measurements, multi-seed experiments, and cross-dataset evaluations on VisDrone, SODA-D, and UAVDT.

2 Related Work

Small-object detection. Small-object detection is affected by scale, annotation density, class imbalance, and complex background structure (Cheng et al., 2023). VisDrone and UAVDT provide challenging aerial and low-altitude traffic benchmarks with dense layouts, occlusion, and motion (Zhu et al., 2022; Du et al., 2018). High-resolution processing has been explored through slicing-based inference (Akyon et al., 2022) and region-aware feature slicing (Liu et al., 2025). Recent work also improves sampling, shallow-feature reuse, and optimization stability for small targets (Huang et al., 2025; Sharma, 2025; Sun et al., 2025). Our focus is complementary: rather than globally increasing image resolution or redesigning the detector, we adapt the decoder sampling range per query and regularize redundant queries after matching.

Query-based detectors. DETR introduced end-to-end set prediction with Hungarian matching (Carion et al., 2020). Deformable DETR uses sparse multi-scale sampling (Zhu et al., 2021); DAB-DETR introduces dynamic anchor boxes (Liu et al., 2022); DN-DETR and DINO improve training with denoising and query selection (Li et al., 2022; Zhang et al., 2023). DQ-DETR further adapts queries for tiny objects (Huang et al., 2024). D-FINE provides a particularly strong real-time basis through distributional box refinement (Peng et al., 2025). QAD-D-FINE retains this end-to-end structure and addresses two remaining failure modes that become pronounced in dense small-object scenes.

3 Method

3.1 Overview

Figure 1 shows the overall architecture. QAS operates inside the decoder forward pass. It adds the P2 feature as the highest-resolution sampling level and uses the state entering each decoder layer to determine a continuous scale factor for that layer's deformable offsets. MRS is training-only: after standard D-FINE Hungarian matching, matched small-object queries serve as anchors, while unmatched queries are assigned to same-class ground-truth instances and weighted according to class confidence, normalized center distance, and IoU redundancy. Only a small number of high-redundancy candidates per object are penalized. MRS contributes no additional operations to inference; QAS is the only proposed component in the inference graph.

3.2 Query-Adaptive Sampling

Let $\hat{w}_j^{(l-1)}$ and $\hat{h}_j^{(l-1)}$ be the normalized width and height of query j entering decoder layer l, and let $z_{j,c}^{(l-1)}$ be its class logit. We first define the query scale, the image-specific normalized small-object reference scale, and the query confidence:

$$\begin{aligned} a_j^{(l-1)} &= \sqrt{\hat{w}_j^{(l-1)} \hat{h}_j^{(l-1)}}, \\ a_{s,b} &= \sqrt{\frac{\kappa_b A_s}{H_{\text{in}} W_{\text{in}}}}, \\ \rho_j^{(l-1)} &= \max_c \sigma\left(z_{j,c}^{(l-1)}\right), \end{aligned} \tag{1}$$

where $A_s = 32^2$ pixels^2 is the COCO small-object area boundary in the original annotation coordinates, $\kappa_b = s_{x,b} s_{y,b}$ is the total area scaling from the original image to the network input, and $H_{\text{in}}, W_{\text{in}}$ denote the input size.

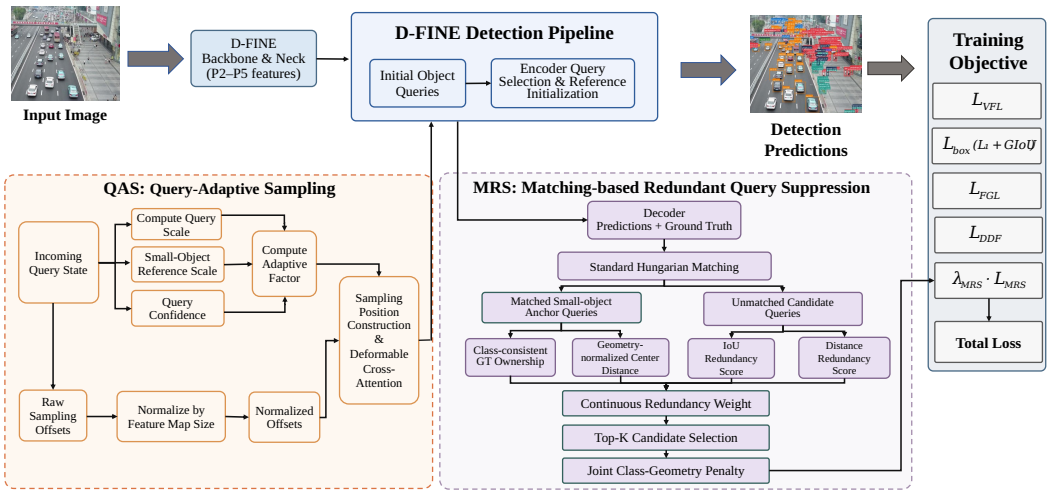


Figure 1: Overview of QAD-D-FINE. QAS performs query-adaptive multi-scale sampling in the decoder, while MRS uses the post-matching relation to suppress redundant unmatched queries during training.

To prevent the detection head from indirectly controlling the sampling range by changing its own predicted scale or confidence, we apply stop-gradient $\mathrm{sg}(\cdot)$:

$$\gamma_j^{(l)} = 1 + \mathrm{sg}(\rho_j^{(l-1)}) \tanh\Big(\beta[\mathrm{sg}(a_j^{(l-1)}) - a_{s,b}]\Big). \tag{2}$$

Because $0 \le \rho_j^{(l-1)} \le 1$ and $|\tanh(\cdot)| < 1$, the adaptive factor satisfies $0 < \gamma_j^{(l)} < 2$. A confident query smaller than the reference scale receives $\gamma < 1$ and samples more locally; a larger query receives $\gamma > 1$ and aggregates broader context. As query confidence decreases, γ approaches 1, which reduces sensitivity to unstable early predictions.

For feature level m of width W_m and height H_m, the raw predicted offset $\Delta p_{j,m,k}^{(l)}$ is normalized by the feature-map size and then scaled by $\gamma_j^{(l)}$:

$$\widetilde{\Delta p}_{j,m,k}^{(l)} = \left(\frac{\Delta p_{j,m,k}^{x,(l)}}{W_m}, \frac{\Delta p_{j,m,k}^{y,(l)}}{H_m}\right), \quad p_{j,m,k}^{(l)} = \mathrm{ref}_j^{(l-1)} + \gamma_j^{(l)} \widetilde{\Delta p}_{j,m,k}^{(l)}. \tag{3}$$

The first layer uses the encoder-selected query state and initial reference box from D-FINE; later layers use the previous decoder output. Setting $\gamma = 1$ disables query-dependent range adjustment but still retains feature-map-size normalization, and therefore is not identical to the original D-FINE reference-box offset rule. Operationally, every decoder layer follows the same feed-forward sequence: it reads the incoming query state, computes $\gamma_j^{(l)}$, rescales that layer's normalized offsets, and then updates the query representation and reference box for the next layer. Thus, the adaptive factor never depends on the output of the layer whose sampling positions it controls, avoiding an intra-layer circular dependence.

3.3 Matching-based Redundant Query Suppression

After standard Hungarian matching, let $m(i)$ denote the query matched to a small ground-truth object g_i. MRS acts only on unmatched queries and does not modify the primary one-to-one assignment. For an unmatched box b_j, we normalize its center distance by half the diagonal length of g_i (using a consistent coordinate system for boxes and distances):

$$r_i = \frac{1}{2}\sqrt{(w_i^g)^2 + (h_i^g)^2}, \qquad q_{ij} = \frac{d(\mathrm{ctr}(b_j), \mathrm{ctr}(g_i))}{r_i + \epsilon}. \tag{4}$$

Define $d(\cdot,\cdot)$ as Euclidean distance, $\mathrm{ctr}(\cdot)$ as the box center, w_i^g, h_i^g as the ground-truth box dimensions, and $\epsilon > 0$ as a numerical stabilizer. The notation $[x]_+ = \max(x, 0)$ is used below. The distance and IoU redundancy strengths are

$$m_{ij}^d = [1 - q_{ij}]_+, \qquad m_{ij}^o = \left[\frac{\mathrm{IoU}(b_j, g_i) - \tau_{\mathrm{IoU}}}{1 - \tau_{\mathrm{IoU}}}\right]_+. \tag{5}$$

To reduce cross-suppression between adjacent instances of the same class, let $\hat{y}_j = \arg\max_c p_j(c)$ be the predicted class of unmatched query j, where $p_j(c)$ is its predicted probability for class c. We assign each candidate to the ground-truth instance with the highest compatibility among those of class $\hat{y}_j$ (if no such ground truth exists, the candidate is excluded from MRS):

$$o(j) = \arg\max_{r:y_r=\hat{y}_j} \left[\mathrm{IoU}(b_j, g_r) + \exp(-q_{rj})\right]. \tag{6}$$

Let y_i and A_i denote the class label and area of ground-truth object i, and let $p_{m(i)}(y_i)$ denote the matched anchor query's confidence for that class. For a small object ($A_i \le A_s$) and a candidate j with $o(j) = i$, we define a continuous redundancy weight

$$u_{ij} = \mathrm{sg}[p_{m(i)}(y_i)]\, p_j(y_i)\left(m_{ij}^d + m_{ij}^o\right). \tag{7}$$

Let $\mathcal{S} = \{i : A_i \le A_s\}$ be the set of small ground-truth instances in the current image; $\mathcal{M}$ is the matched-query set. For $i \in \mathcal{S}$, define $\mathcal{C}_i = \{j : j \notin \mathcal{M},\ o(j) = i,\ m_{ij}^d + m_{ij}^o > 0\}$ and let $\mathcal{T}_i \subseteq \mathcal{C}_i$ contain up to K candidates with the largest detached weights $\mathrm{sg}(u_{ij})$. Empty candidate sets contribute zero to the loss. The loss is normalized first per object and then over small objects:

$$\mathcal{L}_{\mathrm{MRS}} = \frac{1}{\max(1, |\mathcal{S}|)} \sum_{i\in\mathcal{S}} \frac{1}{\max(1, |\mathcal{T}_i|)} \sum_{j\in\mathcal{T}_i} u_{ij}, \qquad |\mathcal{T}_i| \le K. \tag{8}$$

The candidate assignment and top-K selection are discrete and are not differentiated. Because u_{ij} contains both class probability and geometric redundancy, MRS back-propagates to the selected unmatched queries' class responses and box coordinates. Minimizing $\mathcal{L}_{\mathrm{MRS}}$ therefore discourages a selected unmatched query from simultaneously keeping a high target-class response and remaining geometrically concentrated around the same ground-truth instance. The total objective is

$$\mathcal{L} = \mathcal{L}_{\text{D-FINE}} + \lambda_{\mathrm{MRS}} \mathcal{L}_{\mathrm{MRS}}, \tag{9}$$

where $\mathcal{L}_{\text{D-FINE}}$ includes the original classification, ℓ_1/GIoU (Rezatofighi et al., 2019), FGL, DDF, and auxiliary losses.

4 Experiments

4.1 Datasets, Metrics, and Implementation

We evaluate on VisDrone (Zhu et al., 2022), SODA-D (Cheng et al., 2023), and UAVDT (Du et al., 2018). VisDrone contains 10 categories and is our primary dense aerial benchmark; SODA-D contains 9 driving-scene categories; UAVDT contains 3 vehicle categories. We use the D-FINE COCO evaluator with bbox IoU and `maxDets`=$[1, 10, 500]$. VisDrone reports COCO AP/AR metrics (Lin et al., 2014); SODA-D follows its ES/RS/GS/N scale groups; UAVDT reports AP metrics and AR@500.

We additionally measure a dataset-level *duplicate prediction rate*. Each retained prediction is uniquely assigned to the same-class non-ignore ground truth with the largest IoU if that IoU is at least 0.5. If n_i predictions are assigned to object i, one is treated as the primary prediction and the remaining $d_i = \max(n_i - 1, 0)$ as duplicates:

$$R_{\mathrm{dup}} = \frac{\sum_i \max(n_i - 1, 0)}{\sum_i n_i + 10^{-12}} \times 100\%. \tag{10}$$

Thus each prediction can contribute to at most one object even in crowded same-class regions.

The backbone is HGNetv2-B0. P3–P5 enter the Hybrid Encoder with channels $[256, 512, 1024]$; P2 is projected by a 1×1 convolution to dimension 256 and directly serves as the fourth decoder

sampling level. We use 3 decoder layers and 640×640 inputs. Unless otherwise stated, $\beta = 8$, $\tau_{\text{IoU}} = 0.30$, $K = 3$, and $\lambda_{\text{MRS}} = 0.03$. MRS is applied only at the final decoder layer. Training uses AdamW for 132 epochs, learning rates 10^{-4} for the backbone and 2×10^{-4} elsewhere, weight decay 10^{-4}, D-FINE-S COCO pretraining, AMP, and the official D-FINE augmentations. Only the new P2 projection introduces learnable parameters beyond D-FINE; QAS scaling and MRS themselves are parameter-free.

All VisDrone comparison numbers are reproduced from public implementations under the same split, 640×640 input, and evaluation code. D-FINE and QAD-D-FINE additionally share initialization, optimizer, schedule, augmentation, and hardware.

4.2 Main Results on VisDrone

Table 1 shows that QAD-D-FINE obtains 33.1 AP, improving D-FINE by 1.7 points. The gain is concentrated on small targets: AP_s rises from 20.3 to 23.6 (+3.3), and AR_s from 38.8 to 46.5 (+7.7), while AP_l remains essentially unchanged (52.8 vs. 52.4). Compared with DINO, QAD-D-FINE improves AP by 0.7 and AP_s by 1.2 points, and also achieves higher small-object recall.

Table 1: VisDrone comparison under the unified evaluation protocol (%).

Method	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l	AR	AR_s
Faster R-CNN (Ren et al., 2015)	23.9	41.1	20.1	15.9	34.7	38.7	35.5	26.4
FCOS (Tian et al., 2019)	19.8	34.0	16.1	11.2	30.5	41.7	31.1	19.0
DAB-DETR (Liu et al., 2022)	26.7	47.9	22.4	19.1	35.6	41.7	45.8	33.5
DQ-DETR (Huang et al., 2024)	27.4	48.8	23.6	20.7	37.7	43.4	43.2	37.5
DINO (Zhang et al., 2023)	32.4	51.3	28.8	22.4	41.1	53.9	50.1	43.1
RT-DETR (Zhao et al., 2024)	30.5	50.7	25.9	19.5	38.1	52.1	45.3	35.4
YOLOv8s (Terven et al., 2023)	23.7	39.8	19.1	14.0	35.8	43.6	39.9	29.2
YOLOv10s (Wang et al., 2024)	24.5	42.3	20.5	15.3	36.6	45.5	41.4	31.0
D-FINE (Peng et al., 2025)	31.4	49.9	27.1	20.3	38.8	52.8	45.6	38.8
QAD-D-FINE	**33.1**	**52.3**	**29.2**	**23.6**	40.0	52.4	49.8	**46.5**

These results support the intended effect of improving small-object coverage rather than uniformly increasing performance across all scales. Figure 2 further compares representative dense regions. Across the three zoomed areas, QAD-D-FINE generally recovers more true positives than D-FINE and DQ-DETR although false-positive counts increase in one region; the local trade-off varies by region, so we use the dataset-level metrics rather than a single image as the primary evidence.

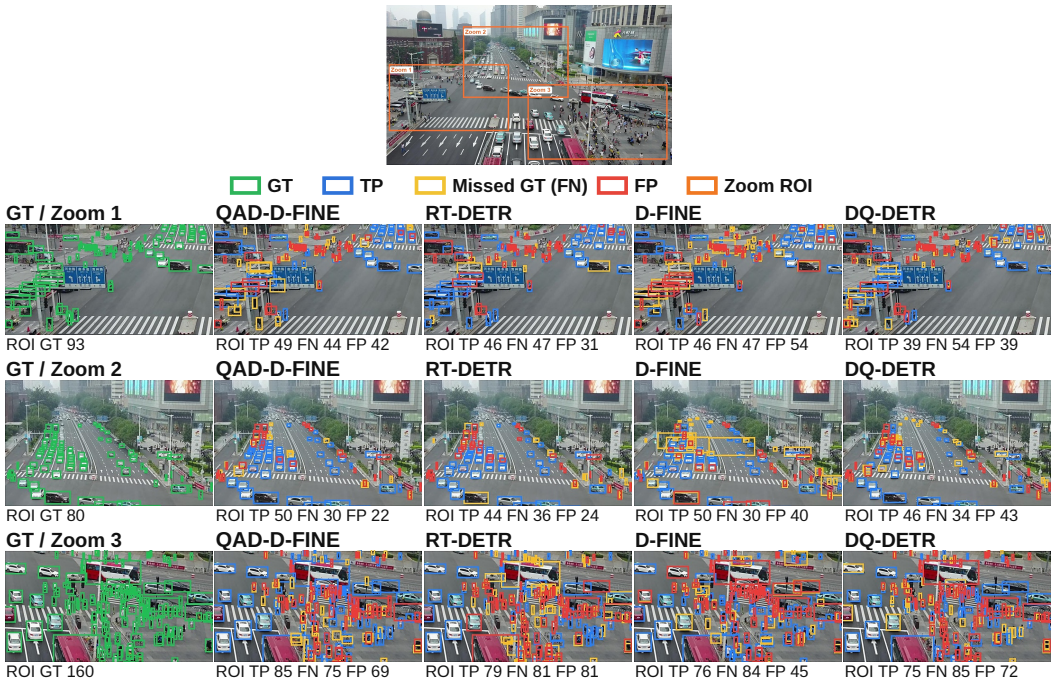


Figure 2: Qualitative comparison on dense VisDrone regions. Green boxes denote ground truth (GT), blue boxes true positives (TP), yellow boxes missed ground truth / false negatives (FN), and red boxes false positives (FP). Orange boxes in the overview mark the three zoom regions. The numbers below each panel report the corresponding ROI counts.

4.3 Efficiency and Module Ablation

Table 2 traces both cost and accuracy as the components are introduced. P2 accounts for most of the extra compute. The full QAS model increases parameters only from 10.0M to 10.1M and FLOPs from 25.0G to 28.0G; latency increases from 5.1 to 6.0 ms, corresponding to 166.7 FPS on an RTX 4090. MRS has no inference-time effect on parameters, FLOPs, latency, or FPS because it is training-only. It increases training time from 8.7 to 9.1 min/epoch and peak memory from 10.0 to 10.4 GB relative to full QAS.

The module ablations further show complementary accuracy gains. With QAS disabled, full MRS in Table 4 improves AP/AP_s/AR_s by 1.0/1.5/4.9 points over D-FINE; full QAS without MRS in Table 2 improves them by 1.1/1.6/5.2 points. Combining both reaches 33.1 AP, 23.6 AP_s, and 46.5 AR_s.

Table 2: Efficiency and module ablation on VisDrone. Speed uses batch size 1, AMP, 100 warm-up iterations and 1000 timed forwards on one RTX 4090.

Configuration	Params (M)	FLOPs (G)	Lat. (ms)	FPS	Train min/ep.	Mem. GB	AP	AP_s	AR_s
D-FINE	10.0	25.0	5.1	196.1	7.8	8.7	31.4	20.3	38.8
D-FINE + P2	10.1	27.9	5.8	172.4	8.5	9.8	31.6	20.7	41.3
+ continuous adaptive sampling (QAS)	10.1	28.0	6.0	166.7	8.7	10.0	32.5	21.9	44.0
QAD-D-FINE (+MRS)	10.1	28.0	6.0	166.7	9.1	10.4	**33.1**	**23.6**	**46.5**

Figure 3 provides a module-level qualitative comparison on the same local regions. The D-FINE baseline misses vehicles in both ROIs and produces overlapping false positives near adjacent objects in ROI B. QAS and MRS each recover part of the missed detections, but through different mechanisms: QAS improves feature acquisition for small targets, whereas MRS reduces redundant unmatched responses after assignment. The full model preserves the true positives recovered by the

individual modules and further removes local false responses, which is consistent with the complementary gains in Table 2 and Table 4.



Figure 3: Detection comparison for the module ablation. Columns show ground truth, the D-FINE baseline, +QAS, +MRS, and full QAD-D-FINE. Green boxes mark true positives recovered relative to the baseline. The two ROIs illustrate both recovered small objects and changes in local false responses.

4.4 What Makes QAS Effective?

We first separate the two switchable ingredients of QAS. P2 alone raises AP_s from 20.3 to 20.7 and AR_s from 38.8 to 41.3. Continuous adaptive sampling without P2 reaches 21.1 AP_s and 42.8 AR_s. Their combination reaches 21.9 AP_s and 44.0 AR_s, showing that high-resolution evidence and query-dependent range adjustment are complementary.

We then retain P2–P5 and perform a more controlled comparison (Table 3). Q0 uses the original D-FINE reference-box offset rule. Q1 switches to Eq. 3 with $\gamma = 1$ to isolate coordinate normalization. Q2 uses a query-independent constant scale, Q3 uses scale adaptation with confidence fixed to one, Q4 is full QAS, and Q5 randomly permutes the full-QAS scale factors among queries within the same image and decoder layer. Q4 outperforms Q5 by 0.5 AP, 0.7 AP_s, and 1.8 AR_s, even though the set and marginal distribution of scale factors are preserved within each permutation. Thus, the gain cannot be explained only by offset normalization or globally changing the sampling radius; matching the scale factor to the query state matters.

Table 3: Controlled analysis of query-adaptive sampling on VisDrone (MRS disabled).

Cfg.	Sampling rule	AP	AP_{50}	AP_s	AR_s
Q0	original D-FINE offset + P2	31.6	50.3	20.7	41.3
Q1	Eq. 3, $\gamma = 1$	31.9	50.7	21.0	41.9
Q2	Eq. 3, fixed $\gamma = c$	32.1	51.0	21.3	42.6
Q3	Eq. 3, $\rho = 1$	32.3	51.4	21.6	43.4
Q4	full QAS	**32.5**	**51.8**	**21.9**	**44.0**
Q5	Eq. 3, permuted γ	32.0	50.9	21.2	42.2

4.5 How Does MRS Suppress Redundancy?

Table 4 decomposes MRS on the D-FINE baseline with QAS disabled. The distance term alone reduces the duplicate rate from 12.6% to 11.0%, whereas the IoU term lowers it to 8.2% and provides

a larger detection gain. Combining both reaches 32.4 AP, 21.8 AP_s, 43.7 AR_s, and a 6.9% duplicate rate, a 45.2% relative reduction from baseline.

Table 4: Fine-grained MRS ablation on VisDrone (QAS disabled).

Distance	IoU	AP	AP_s	AR_s	R_{dup}
		31.4	20.3	38.8	12.6
✓		31.9	21.0	40.8	11.0
	✓	32.1	21.3	42.0	8.2
✓	✓	**32.4**	**21.8**	**43.7**	**6.9**

To examine how MRS affects optimization, we compare three variants with identical candidate construction and top-K selection: *Class-only* stops gradients through the geometric terms, *Geometry-only* stops gradients through the unmatched candidate's target-class probability, and *Full MRS* keeps both paths. For each ground-truth small object, we define the average candidate class response C_i and average geometric redundancy $R_i = \frac{1}{|\mathcal{T}_i|}\sum_{j\in\mathcal{T}_i}(m^d_{ij}+m^o_{ij})/2$, and plot changes relative to the baseline. Objects without selected candidates are excluded from these per-candidate averages. Figure 4 shows that Class-only shifts more strongly toward lower class scores, Geometry-only toward lower geometric redundancy, and Full MRS more often occupies the lower-left region. The distributions overlap, so this analysis is not intended to assign independent AP gains to the two gradient paths; rather, it shows that they act differently and complementarily on the final query state.

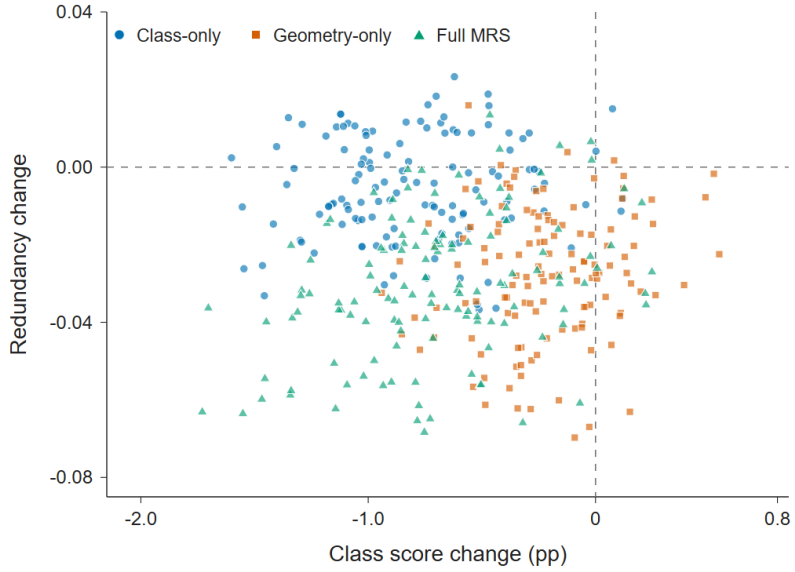


Figure 4: Per-object change in redundant-query class response and geometric redundancy relative to D-FINE. Negative values indicate suppression. Full MRS more often combines reductions along both axes.

Multi-seed stability and the complete one-factor hyperparameter sweeps are reported in Appendix B. The observed gains are substantially larger than the seed-to-seed variation, and a broad neighborhood around the default hyperparameters remains stable.

4.6 Cross-Dataset Generalization

Table 5 summarizes SODA-D and UAVDT. On SODA-D, QAD-D-FINE improves D-FINE from 30.7 to 32.3 AP. The gains are larger on the extremely-small group: AP_{ES} improves by 2.0 points and AR_{ES} by 3.8. On UAVDT, QAD-D-FINE reaches 41.1 AP and 29.6 AP_s, improving D-FINE

by 1.2 and 3.7 points, respectively, while AR@500 rises from 58.4 to 60.9. These results indicate that the gains transfer beyond the VisDrone aerial setting to driving and traffic-monitoring scenes with dense small objects.

Table 5: Cross-dataset results (%). SODA-D uses its extremely-small (ES) group; UAVDT reports AR@500.

	SODA-D					UAVDT			
Method	AP	AP_{ES}	AR	AR_{ES}	Method	AP	AP_s	AP_{50}	AR@500
D-FINE	30.7	15.8	43.8	27.6	D-FINE	39.9	25.9	68.7	58.4
DINO	28.8	13.2	41.5	24.4	DINO	38.4	23.5	63.1	55.1
RT-DETR	29.5	15.0	42.6	26.5	RT-DETR	38.2	24.5	67.0	56.6
YOLOv10s	25.9	11.4	38.9	21.3	YOLOv10s	38.4	27.3	66.2	55.3
YOLOv8s	28.2	13.6	41.2	25.1	YOLOv8s	36.6	24.5	64.0	51.5
DQ-DETR	30.1	15.5	43.2	27.1	DQ-DETR	40.6	27.7	69.5	59.2
QAD-D-FINE	**32.3**	**17.8**	**46.2**	**31.4**	**QAD-D-FINE**	**41.1**	**29.6**	**70.4**	**60.9**

5 Conclusion

We presented QAD-D-FINE for dense small object detection. QAS adapts decoder sampling to the incoming query scale and confidence while incorporating high-resolution P2 features, and MRS uses the standard matching relation to suppress redundant unmatched queries through complementary classification and geometry paths. On VisDrone, the method improves D-FINE by 1.7 AP, 3.3 AP_s, and 7.7 AR_s while maintaining 166.7 FPS and 10.1M parameters. Controlled ablations show that query-state correspondence is important for QAS and that distance and IoU redundancy terms are complementary in MRS. Multi-seed results and cross-dataset experiments on SODA-D and UAVDT further support the stability and transferability of the gains. Future work will study more flexible query interactions and lighter high-resolution feature integration.

AI use statement

Generative AI tools assisted with translation, language editing, consistency checks, manuscript organization, and LaTeX formatting. The authors reviewed AI-assisted material against the original research materials, implementation, and experimental records. The authors take responsibility for the final content of this work, including text or artifacts produced with the aid of generative AI.

Ethics statement

This work evaluates object-detection methods on existing aerial and traffic-scene benchmarks and does not introduce a new dataset or newly collected human-subject data. Potential real-world uses of dense detection in monitoring systems may raise privacy, surveillance, bias, and misuse concerns; deployments should therefore follow applicable laws, dataset licenses, privacy protections, and domain-specific oversight.

Reproducibility statement

Sections 3.2–3.3 specify the QAS and MRS computations, while Section 4.1 reports the datasets, evaluator, duplicate-prediction metric, model configuration, default hyperparameters, optimizer, learning rates, training duration, initialization, AMP setting, and augmentations. Table 2 states the inference timing protocol and hardware. Appendices A–C provide the layer-wise QAS procedure, the complete two-module ablation, multi-seed results, complete hyperparameter sweeps, and full cross-dataset metric breakdowns. All VisDrone comparisons use the same split, input resolution, and evaluation code, with D-FINE and QAD-D-FINE additionally sharing initialization, optimizer, schedule, augmentation, and hardware.

REFERENCES

Fatih Cagatay Akyon, Sinan Onur Altinuc, and Alptekin Temizel. Slicing aided hyper inference and fine-tuning for small object detection. In *IEEE International Conference on Image Processing*, pp. 966–970, 2022.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, 2020.

Gong Cheng, Xiang Yuan, Xiwen Yao, Kebing Yan, Qinghua Zeng, Xingxing Xie, and Junwei Han. Towards large-scale small object detection: Survey and benchmarks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11):13467–13488, 2023.

Dawei Du, Yuankai Qi, Hongyang Yu, Yifan Yang, Kaiwen Duan, Guorong Li, Weigang Zhang, Qingming Huang, and Qi Tian. The unmanned aerial vehicle benchmark: Object detection and tracking. In *European Conference on Computer Vision*, 2018.

Yi-Xin Huang, Hou-I Liu, Hong-Han Shuai, and Wen-Huang Cheng. DQ-DETR: DETR with dynamic query for tiny object detection. In *European Conference on Computer Vision*, 2024.

Zile Huang, Chong Zhang, Mingyu Jin, Fangyu Wu, Chengzhi Liu, and Xiaobo Jin. Better sampling, towards better end-to-end small object detection. In *Computer Animation and Social Agents*, pp. 319–335, 2025.

Feng Li, Hao Zhang, Shilong Liu, Jian Guo, Lionel M. Ni, and Lei Zhang. DN-DETR: Accelerate DETR training by introducing query denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

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 *European Conference on Computer Vision*, 2014.

Shilong Liu, Feng Li, Hao Zhang, Xiao Yang, Xianbiao Qi, Hang Su, Jun Zhu, and Lei Zhang. DAB-DETR: Dynamic anchor boxes are better queries for DETR. In *International Conference on Learning Representations*, 2022.

Kai Liu, Zhihang Fu, Sheng Jin, Ze Chen, Fan Zhou, Rongxin Jiang, Yaowu Chen, and Jieping Ye. ESOD: Efficient small object detection on high-resolution images. *IEEE Transactions on Image Processing*, 34:183–195, 2025.

Yansong Peng, Hebei Li, Peixi Wu, Yueyi Zhang, Xiaoyan Sun, and Feng Wu. D-FINE: Redefine regression task of DETRs as fine-grained distribution refinement. In *International Conference on Learning Representations*, 2025.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, 2015.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

Aashish Sharma. DM-EFS: Dynamically multiplexed expanded features set form for robust and efficient small object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 24569–24579, 2025.

Huixin Sun, Yanjing Li, Linlin Yang, Xianbin Cao, and Baochang Zhang. Uncertainty-aware gradient stabilization for small object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8407–8417, 2025.

Juan Terven, Diana-Margarita Cordova-Esparza, and Julio-Alejandro Romero-Gonzalez. A comprehensive review of YOLO architectures in computer vision: From YOLOv1 to YOLOv8 and YOLO-NAS. *Machine Learning and Knowledge Extraction*, 5(4):1680–1716, 2023.

Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. FCOS: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.

Ao Wang, Hui Chen, Lihao Liu, Kai Chen, Zijia Lin, Jungong Han, and Guiguang Ding. YOLOv10: Real-time end-to-end object detection. In *Advances in Neural Information Processing Systems*, 2024.

Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M. Ni, and Heung-Yeung Shum. DINO: DETR with improved denoising anchor boxes for end-to-end object detection. In *International Conference on Learning Representations*, 2023.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*, 2021.

Pengfei Zhu, Longyin Wen, Dawei Du, Xiao Bian, Heng Fan, Qinghua Hu, and Haibin Ling. Detection and tracking meet drones challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):7380–7399, 2022.

A Additional Implementation and Ablation Results

A.1 Layer-wise QAS Procedure

For completeness, we summarize the layer-wise implementation of QAS as follows. The input consists of multi-scale features F together with the D-FINE initialized query state $Q^{(0)}$, class responses $z^{(0)}$, and reference boxes $b^{(0)}$. For each decoder layer $l = 1, \ldots, L$: (1) read the incoming class logits $z^{(l-1)}$ and normalized box scale $b^{(l-1)}$; (2) compute the query scale $a^{(l-1)}$, image-specific small-object reference scale $a_{s,b}$, and confidence $\rho^{(l-1)}$ using Eq. 1; (3) compute the bounded adaptive factor $\gamma^{(l)}$ using Eq. 2; (4) normalize the raw offsets by the feature-map dimensions and rescale them according to Eq. 3; and (5) sample over P2–P5 and update the query representation, class responses $z^{(l)}$, and reference boxes $b^{(l)}$. The decoder returns the classification and box-regression outputs from its layers.

A.2 Complete Two-Module Ablation

The complete switch-level comparison below complements the efficiency table in the main paper by explicitly separating the standalone and joint contributions of QAS and MRS.

Table A.1: Two-module ablation on VisDrone (%). Standalone and joint module configurations are reported.

Configuration	MRS	QAS	AP	AP_{50}	AP_s	AP_m	AP_l	AR_s
Baseline			31.4	49.9	20.3	38.8	52.8	38.8
+ MRS	✓		32.4	51.6	21.8	39.6	52.7	43.7
+ QAS		✓	32.5	51.8	21.9	39.6	52.5	44.0
Full	✓	✓	**33.1**	**52.3**	**23.6**	**40.0**	52.4	**46.5**

B Additional Stability and Sensitivity Results

We repeat Baseline, +MRS, and Full training with random seeds 0, 1, and 2 while keeping all other training settings fixed. This directly tests whether the gains attributed to MRS and the full model are larger than ordinary training variation.

Table B.1: Three-seed stability on VisDrone (%). Values are mean ± standard deviation.

Configuration	AP	AP_s
Baseline	31.4 ± 0.09	20.3 ± 0.12
+MRS	32.4 ± 0.08	21.8 ± 0.10
Full	33.1 ± 0.08	23.6 ± 0.10

The +MRS gain of 1.0 AP and the Full gain of 1.7 AP over Baseline are much larger than the corresponding standard deviations. The same pattern holds for AP_s, supporting that the reported improvements are not driven by a single favorable random seed.

For sensitivity analysis, we vary one hyperparameter at a time around the default configuration while keeping the other three fixed. The full sweep is reported below.

Table B.2: Complete one-factor hyperparameter sensitivity analysis on VisDrone (%). When one parameter is varied, the others remain at $\beta = 8$, $\tau_{\mathrm{IoU}} = 0.30$, $K = 3$, and $\lambda_{\mathrm{MRS}} = 0.03$.

Parameter	Value	AP	AP_{50}	AP_s	AR_s	R_{dup}
β	2	32.5	51.5	22.6	44.8	8.0
	4	32.8	51.9	23.1	45.7	7.5
	8	**33.1**	**52.3**	**23.6**	**46.5**	6.8
	12	33.0	52.1	23.4	46.1	6.9
	16	32.7	51.8	23.0	45.3	7.2
τ_{IoU}	0.20	32.7	51.8	23.1	45.6	6.3
	0.25	32.9	52.0	23.4	46.1	6.5
	0.30	**33.1**	**52.3**	**23.6**	**46.5**	6.8
	0.35	33.0	52.1	23.5	46.2	7.2
	0.40	32.8	51.9	23.2	45.8	7.8
K	1	32.6	51.7	22.9	45.1	8.0
	2	32.9	52.0	23.3	46.0	7.3
	3	**33.1**	**52.3**	**23.6**	**46.5**	6.8
	4	32.9	52.0	23.3	46.0	6.5
	5	32.6	51.6	22.8	45.4	6.3
λ_{MRS}	0.01	32.7	51.8	23.0	45.5	7.8
	0.02	32.9	52.0	23.3	46.1	7.2
	0.03	**33.1**	**52.3**	**23.6**	**46.5**	6.8
	0.04	32.9	52.0	23.3	46.0	6.4
	0.05	32.6	51.7	22.8	45.2	**6.1**

The sweeps show a broad stable neighborhood around the default setting. For β, increasing the scale response from 2 to 8 improves AP/AP_s/AR_s, while a stronger response at 16 slightly degrades performance, indicating that overly aggressive scale adaptation amplifies box-scale fluctuations. For τ_{IoU}, the default 0.30 gives the best accuracy trade-off: a lower threshold suppresses more duplicates but also lowers AP, whereas a higher threshold leaves more redundant responses. Similarly, $K = 3$ balances coverage and over-suppression; $K = 5$ reduces R_{dup} to 6.3% but lowers AP_s to 22.8. Increasing λ_{MRS} beyond 0.03 continues to reduce duplicates but begins to hurt small-object accuracy. Overall, the neighborhood $\beta \in [4, 12]$, $\tau_{\mathrm{IoU}} \in [0.25, 0.35]$, $K \in [2, 4]$, and $\lambda_{\mathrm{MRS}} \in [0.02, 0.04]$ maintains AP of at least 32.8, supporting the robustness of the default setting.

C Full Cross-Dataset Results

Tables C.1–C.3 provide the full scale-specific precision and recall breakdowns that are summarized more compactly in Table 5 of the main paper.

Table C.1: Full SODA-D precision breakdown (%).

Method	AP	AP_{50}	AP_{75}	AP_{ES}	AP_{RS}	AP_{GS}	AP_{N}
D-FINE	30.7	61.6	26.9	15.8	27.9	36.4	45.6
DINO	28.8	59.4	24.1	13.2	24.0	34.2	43.1
RT-DETR	29.5	60.5	25.6	15.0	26.9	35.5	44.7
YOLOv10s	25.9	57.0	21.7	11.4	23.1	33.6	43.2
YOLOv8s	28.2	58.6	23.9	13.6	29.0	37.6	43.5
DQ-DETR	30.1	61.0	26.3	15.5	27.6	36.0	45.1
QAD-D-FINE	**32.3**	**63.2**	**28.5**	**17.8**	**30.4**	**39.1**	**46.9**

Table C.2: Full SODA-D recall breakdown (AR@500, %).

Method	AR	AR_{ES}	AR_{RS}	AR_{GS}	AR_{N}
D-FINE	43.8	27.6	39.5	49.2	59.8
DINO	41.5	24.4	35.8	46.7	57.1
RT-DETR	42.6	26.5	38.4	47.8	58.7
YOLOv10s	38.9	21.3	33.2	44.1	55.4
YOLOv8s	41.2	25.1	40.6	49.5	56.3
DQ-DETR	43.2	27.1	39.1	48.6	59.2
QAD-D-FINE	**46.2**	**31.4**	**43.6**	**52.5**	**61.7**

Table C.3: Full UAVDT detection performance (%).

Method	AP	AP_{50}	AP_{75}	AP_{s}	AP_{m}	AP_{l}	AR@500
YOLOv8s	36.6	64.0	37.1	24.5	44.9	54.3	51.5
DINO	38.4	63.1	39.2	23.5	47.8	58.7	55.1
RT-DETR	38.2	67.0	39.8	24.5	48.8	61.9	56.6
YOLOv10s	38.4	66.2	38.2	27.3	47.4	57.9	55.3
D-FINE	39.9	68.7	41.2	25.9	51.1	61.7	58.4
DQ-DETR	40.6	69.5	40.8	27.7	49.5	57.7	59.2
QAD-D-FINE	**41.1**	**70.4**	**42.2**	**29.6**	**51.8**	60.3	**60.9**