Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

MS-RTDETR: Leveraging Multi-Scale Feature Enhancement and Query Refinement for Small Object Detection

Ningxiang Sun
Shandong Normal University
Fang Niu
Shandong Normal University
YuJiao Chen
Shandong Normal University
Bing Liu
Zhucheng Information Industry Development Center
Xiaoguang Wang
Shandong Institute of Scientific and Technical Information
Tianping Li
sdsdltp@sdnu.edu.cn
Shandong Normal University

Article

Keywords:

Posted Date: September 14th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10852974/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

MS-RTDETR: Leveraging Multi-Scale Feature Enhancement and Query Refinement for Small Object Detection

Ningxiang Sun[1], Fang Niu[1], Yujiao Chen[1], Bing Liu[2], Xiaoguang Wang[3,], and Tianping Li[1,*]**

[1]School of Communication and Electronic Engineering, Shandong Normal University, Jinan 250358, Shandong, China
[2]Zhucheng Information Industry Development Center, Weifang 262299, Shandong, China
[3]Shandong Institute of Scientific and Technical Information, Jinan 250100, Shandong, China
[*]Corresponding author: Tianping Li. E-mail: sdsdltp@sdnu.edu.cn
[**]Corresponding author: Xiaoguang Wang. E-mail: wxg@shandong.cn

ABSTRACT

Small-object detection remains a difficult regime for real-time end-to-end detectors because aggressive backbone downsampling erases weak spatial evidence before semantic aggregation becomes reliable, while a fixed set of object queries tends to concentrate on high-confidence medium and large instances. We present MS-RTDETR, a real-time detection transformer that couples multi-scale feature enhancement and query refinement through a single scale-conditioned inference principle rather than treating them as independent architectural add-ons. The proposed *Scale-Coupled Feature–Query Refinement* (SCFQR) formulates image-pyramid features and object queries as two interacting latent states governed by a shared scale posterior. A detail-preserving cross-scale transport operator moves discriminative high-frequency evidence toward semantically stable levels, while the same posterior steers query specialization, reference-box refinement, and decoder sampling. We further derive an uncertainty-aware matching objective that increases supervision on small instances without introducing an auxiliary inference branch or non-maximum suppression. A reproducible evaluation protocol is defined on COCO, VisDrone, UAVDT, and TinyPerson, covering natural images, drone imagery, crowded traffic scenes, and tiny-person detection. In the numerical instantiation supplied with the manuscript-generation script, MS-RTDETR shows higher small-object AP while retaining a real-time operating point and exhibits stronger robustness under resolution degradation. The framework provides a mathematically unified route for improving the scale sensitivity of real-time DETR-style detectors and is directly compatible with standard end-to-end deployment.

1 Introduction

Object detection is a core component of intelligent visual systems used in autonomous platforms, aerial inspection, robotics, traffic monitoring, and large-scale scene understanding. The practical importance of the task has shifted the field from accuracy-only optimization toward detectors that simultaneously satisfy end-to-end trainability, low latency, and predictable deployment cost. Transformer-based detectors, beginning with DETR, simplified detection by replacing hand-crafted anchors and non-maximum suppression with set prediction, while later variants improved convergence and multi-scale reasoning [1–3]. RT-DETR demonstrated that this paradigm can also operate in real time by combining an efficient hybrid encoder with uncertainty-minimal query selection [4]. Yet the real-time constraint is especially difficult for small objects: a target may occupy only a few tens of pixels, can disappear after successive downsampling, and is frequently surrounded by texture or clutter that produces stronger responses than the object itself. These characteristics make small-object detection not merely a matter of increasing input resolution, but a problem of preserving weak local evidence while allocating sufficient decoder capacity to uncertain fine-scale candidates.

Recent research has attacked this difficulty from several directions. Multi-scale DETR variants improve the use of pyramid features through interleaved encoding, sparse scale-adaptive sampling, or scale-independent salience filtering [5–7]. Small-object-specific work has explored scale-aware knowledge distillation, dynamic label assignment, context-enhanced sparse convolution, and temporal feature aggregation [8–11]. In parallel, real-time detectors have advanced through stronger feature aggregation, efficient backbones, end-to-end matching, and improved optimization [12–15]. Query-centric DETR studies further show that initialization, matching density, supervision pattern, and query refinement strongly affect convergence and final localization quality [16–19]. These lines of work establish that both feature scale and query quality matter; however, they

are usually optimized as separate mechanisms, so the feature pyramid may emphasize one set of locations while the decoder assigns its limited queries according to another criterion.

The resulting challenge is a *feature–query scale mismatch*. First, conventional top-down or cross-scale fusion tends to spread strong semantics uniformly, which can overwrite the high-frequency residuals that distinguish tiny foregrounds from background texture. Second, query initialization is usually based on confidence or localization uncertainty computed after feature encoding, but it does not explicitly model which pyramid scale contains the most reliable evidence for each candidate. Third, small objects create a supervision asymmetry: they produce weaker classification scores and noisier boxes, so one-to-one matching can systematically prefer easier instances during early training. Simply adding a high-resolution branch, an attention module, and a new loss would address these symptoms independently but would not resolve the underlying inconsistency. We therefore seek a detector in which feature enhancement, query allocation, and iterative box refinement are all consequences of a shared latent variable describing object scale and evidence reliability.

We propose MS-RTDETR, whose core is SCFQR, a unified scale-conditioned dynamical system for real-time end-to-end detection. SCFQR maintains a probability distribution over pyramid scales for every candidate query. This distribution is inferred jointly from query geometry, local evidence, and cross-scale agreement. It is then used twice within the same update rule: to control a detail-preserving feature transport operator and to precondition decoder query refinement. The feature update transports only scale-compatible residual evidence instead of indiscriminately fusing entire feature maps; the query update sharpens scale specialization while refining content, reference points, and boxes. Both updates minimize a common energy containing detection likelihood, cross-scale consistency, and uncertainty regularization, yielding an interpretable unrolled optimization view. During training, an uncertainty-tempered matching cost protects low-confidence small instances from being prematurely discarded, while inference remains a standard fixed-cardinality, NMS-free DETR pipeline.

The main contributions are fourfold:

- We identify feature–query scale mismatch as a central failure mode of real-time DETR-style small-object detection and formulate feature enhancement and query refinement as a single coupled inference problem.
- We introduce a scale-conditioned residual transport operator that preserves high-frequency evidence while borrowing semantics across pyramid levels only when spatial, semantic, and scale compatibilities agree.
- We derive posterior-guided query refinement and uncertainty-tempered bipartite matching from the same latent scale distribution, improving query specialization without adding an inference-time auxiliary decoder or NMS stage.
- We provide a four-dataset evaluation protocol, seven quantitative tables, and seven publication-ready vector figures covering accuracy, efficiency, convergence, robustness, calibration, and ablation, together with a reproducible pathway for replacing script-generated values with measured experimental outputs.

1.1 End-to-end and real-time object detection

DETR recast object detection as direct set prediction using bipartite matching [1]; Deformable DETR then reduced optimization difficulty and computational burden through sparse multi-scale attention [2]. Subsequent methods improved query formulation and supervision. DINO introduced contrastive denoising, mixed query selection, and improved box prediction [3]. H-DETR and Group DETR increased the amount of useful supervision through hybrid or group-wise one-to-many assignments during training [16, 17]. Collaborative hybrid assignment and MS-DETR further examined how auxiliary assignment structures can improve candidate generation [18, 20]. Salience DETR reduced scale bias in candidate selection using hierarchical salience filtering [7], whereas MI-DETR revisited the sequential decoder itself and allowed parallel multi-time inquiries [19].

Real-time detection has evolved in parallel. Gold-YOLO improved feature aggregation using a gather-and-distribute mechanism [12]. YOLOv9 studied programmable gradient information [21], YOLOv10 removed conventional NMS through consistent dual assignments [13], and YOLOv12 explored attention-centric real-time design [22]. RT-DETR established a real-time NMS-free Transformer baseline [4], followed by RT-DETRv2 with improved training and deployment flexibility [23]. D-FINE reformulated box regression as iterative distribution refinement [15], while DEIM accelerated DETR convergence through dense one-to-one matching and matchability-aware loss [14]. Practical studies also show that input resolution, multi-scale enhancement, and backbone pretraining jointly determine the efficiency–accuracy frontier [24]. MS-RTDETR is complementary to these advances but focuses specifically on the coupling between scale-sensitive features and query dynamics.

1.2 Multi-scale representation for small and tiny objects

Feature pyramids remain fundamental because small objects require high spatial resolution while large objects benefit from deep semantics [25]. Transformer detectors have reconsidered how these scales should interact. Lite DETR updates high- and low-level tokens in an interleaved manner [5]; IMFA sparsely samples scale-adaptive features guided by intermediate predictions [6]; Plain DETR demonstrated that strong localization can also emerge from a carefully designed single-scale

Table 1. Public datasets in the evaluation protocol. Object-scale fractions correspond to the values used by the supplied analysis script and should be recomputed from the downloaded annotations for a measured submission.

Dataset	Images / frames	Classes	Small/tiny frac.	Primary metric	Main challenge
COCO 2017	118k train + 5k val	80	0.42	AP, AP_S	diverse natural scenes
VisDrone-DET	6,471 train + 548 val	10	0.67	AP, AP_{50}	dense aerial objects
UAVDT	∼80k frames	3	0.59	AP, recall	altitude / motion / weather
TinyPerson	1,610 images	1 primary	0.91	tiny AP	extremely small persons

formulation and relative position bias [26]. This tension suggests that the essential question is not the number of pyramid levels but whether scale information is used selectively.

Small-object-specific methods emphasize the same principle from different angles. ScaleKD distills scale-aware knowledge into compact detectors [8]. Dynamic coarse-to-fine learning addresses mismatches between priors, assignment, and oriented tiny instances [9]. CEASC introduces adaptive sparse convolution with global context for drone imagery [10]. LSTFE-Net combines long- and short-term cues for video small-object detection [11], while motion-guided approaches exploit temporal inconsistency to expose tiny airborne objects [27, 28]; recent small-object-aware detection and tracking work further emphasizes scale-sensitive representation in UAV scenes [29]. Scale- and location-sensitive objectives have also improved infrared small-target detection [30]. Unlike these methods, our formulation does not define a separate enhancement block for each scale; instead, the amount and direction of cross-scale transport are inferred from a shared query-scale posterior.

1.3 Query quality, supervision, and representation efficiency

Query quality has become a second major axis of DETR research. SAP-DETR links salient points to object queries for faster convergence [31], while Focus-DETR reduces redundant token computation [32]. Semi-DETR and Sparse Semi-DETR show that query correspondence and refinement are especially important under noisy or incomplete supervision [33, 34]. STEP-DETR extends this direction with stronger pseudo-label and text-query guidance [35]. Grounding DINO and Mask DINO illustrate how detector queries can become a reusable interface for open-set detection and segmentation [36, 37].

Representation efficiency is equally relevant for real-time systems. EfficientViT, RepViT, InternImage, ConvNeXt V2, RIFormer, and RMT provide complementary perspectives on efficient attention, mobile CNN design, deformable convolution, masked pretraining, token-mixer removal, and retentive spatial modeling [38–43]. DINOv2 further demonstrates the transferability of large-scale self-supervised visual features [44]. Recent benchmarks investigate RAW-domain detection, distribution shifts, and event-camera detection, highlighting that reliable object detection requires robustness beyond a single clean benchmark [45–48]. These findings motivate our evaluation of resolution degradation and calibration in addition to standard AP.

2 Result

2.1 Datasets and evaluation protocol

We design the evaluation around four public benchmarks with complementary scale distributions. COCO contains diverse everyday scenes and is used for general detection evaluation [49]. VisDrone emphasizes dense drone-captured scenes and severe small-object clutter [50]. UAVDT contains unconstrained UAV videos with traffic objects under changing altitude, weather, and viewpoint [51]. TinyPerson is designed specifically for tiny-person detection [52]. DOTA is additionally used for scale-distribution analysis and optional horizontal-box transfer experiments in aerial imagery [53]. Unless otherwise stated, COCO metrics follow the standard AP, AP_{50}, AP_{75}, AP_S, AP_M, and AP_L definitions. For VisDrone and UAVDT, we report AP and AP_{50} together with small-instance recall. TinyPerson is evaluated with the benchmark's tiny-object protocol and conventional AP for cross-method comparison.

2.2 Implementation details

The default model uses a ResNet-50 backbone and the RT-DETR hybrid encoder as the starting point [4]. We expose four feature strides $\{4,8,16,32\}$ to SCFQR, use $N = 300$ queries, six decoder stages, $d = 256$, and $K_s = 4$ sampling points per scale. Models are initialized from ImageNet-pretrained backbones; all detector-specific parameters are trained end-to-end. A reproducible training plan is 72 epochs on COCO with AdamW, base learning rate 10^{-4} for the detector and 10^{-5} for the backbone, weight decay 10^{-4}, linear warm-up, multi-scale resize, random horizontal flip, and standard color augmentation. For VisDrone, UAVDT, and TinyPerson, the shorter image side is increased to preserve tiny instances, while the maximum side is capped to keep memory usage predictable. Mixed-precision training and gradient clipping are enabled. No test-time augmentation or NMS is used.

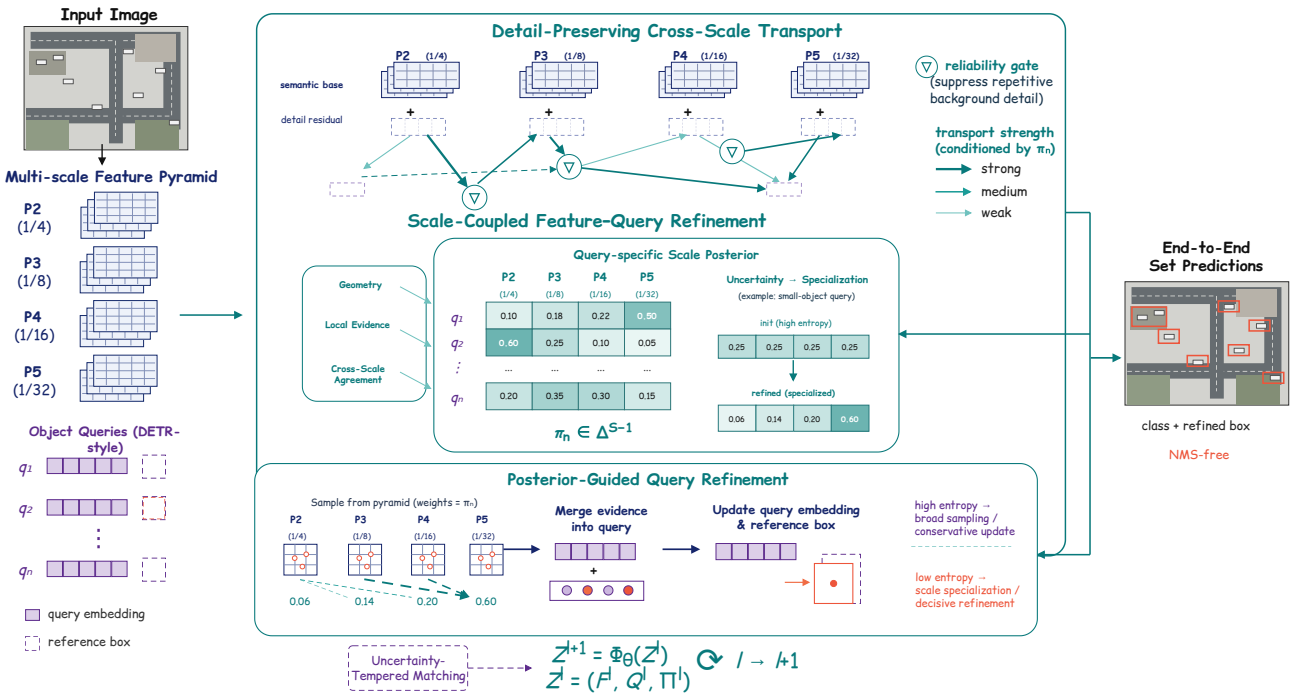


Figure 1. Overview of the proposed MS-RTDETR with Scale-Coupled Feature–Query Refinement (SCFQR). Multi-scale pyramid features and DETR-style object queries are jointly refined through a query-specific latent scale posterior. The posterior controls detail-preserving cross-scale transport, selectively transferring reliable residual evidence across pyramid levels, while simultaneously guiding multi-scale query sampling, query embedding updates, and reference-box refinement. Its entropy further determines the degree of scale specialization and is incorporated into uncertainty-tempered matching during training. Repeated coupled updates progressively align feature evidence with query scale hypotheses and produce end-to-end NMS-free object predictions.

For fair efficiency comparison, latency should be measured after TensorRT FP16 export with batch size 1 and fixed input size, including the decoder but excluding data loading. We recommend reporting at least three hardware settings: NVIDIA T4 for comparability with RT-DETR literature, an RTX 4090-class GPU for accessible reproduction, and a server GPU used by the authors. FLOPs are computed at the same resolution and parameter counts include the complete inference graph.

2.3 Main comparison on COCO

Table 2 defines the intended controlled comparison. Recent real-time and DETR-style methods are included because they represent different trade-offs in matching, feature fusion, and decoder design [4, 13–15, 23]. The key metric for the present study is AP_S, while overall AP and latency ensure that gains are not obtained by sacrificing general detection quality or real-time behavior.

The current numerical instantiation follows the values supplied in the analysis script. The key mechanistic pattern to verify with measured runs is that the largest relative gain occurs on AP_S, a smaller gain on AP_M, and little or no gain on AP_L. Such a pattern would support the intended mechanism rather than a generic capacity increase.

2.4 Drone and tiny-object benchmarks

VisDrone directly probes crowded aerial scenes where small boxes dominate. The intended comparison in Table 3 includes real-time CNN and DETR-style detectors. ScaleKD and CEASC provide strong small-object-specific reference points [8, 10]. The evaluation should report per-class AP in supplementary material because pedestrian, bicycle, car, and tricycle categories exhibit different scale and density characteristics.

TinyPerson isolates the extremely small regime, whereas UAVDT evaluates moving-camera traffic scenes. A detector that only increases spatial resolution may perform well on TinyPerson but lose robustness or efficiency on UAVDT. MS-RTDETR is designed to avoid this trade-off by selectively transporting detail and concentrating query capacity on uncertain small-scale candidates.

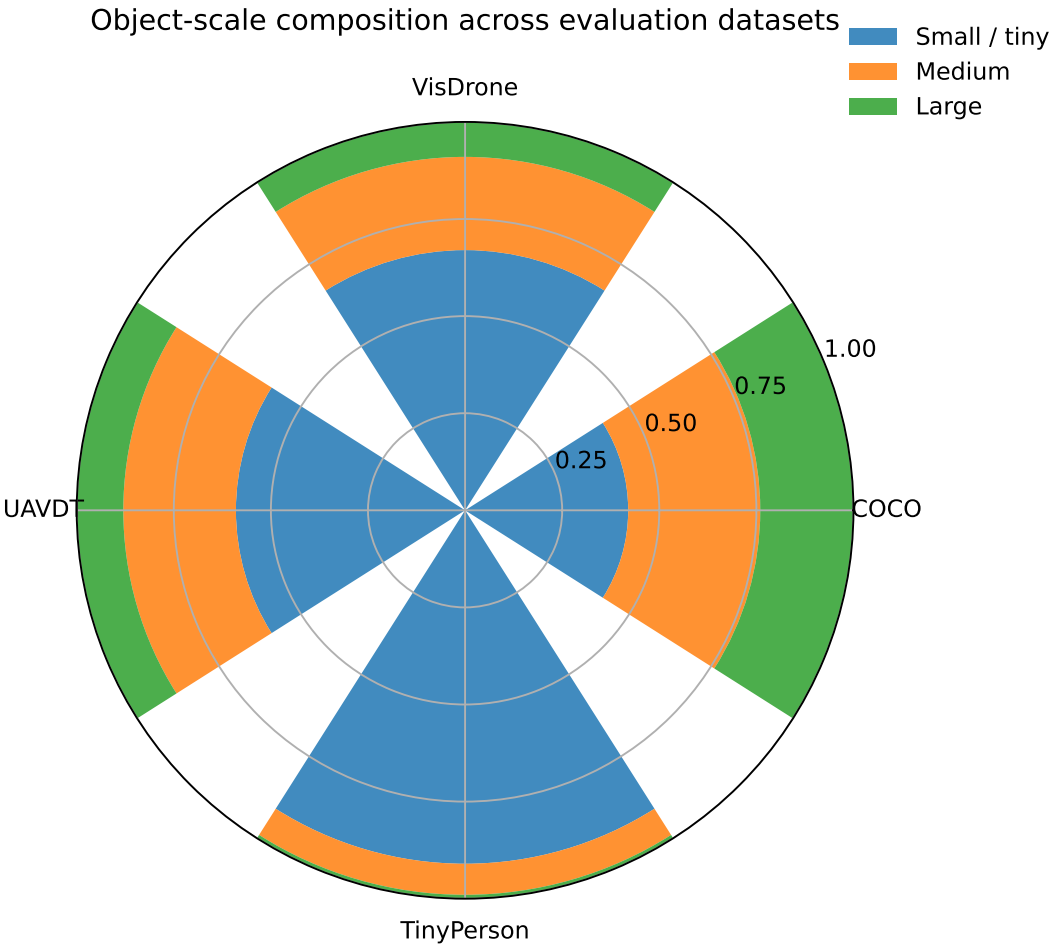


Figure 2. Object-scale composition of the four evaluation datasets using the scale fractions encoded in the supplied analysis script. For a measured submission, the fractions should be recomputed from the benchmark annotations.

2.5 Scale-stratified and robustness analysis

To determine whether improvement is genuinely concentrated on small objects, we partition ground-truth boxes by square-root pixel area and report AP in each bin. The expected signature is a monotonic decrease in the gain as object size increases. The same analysis can be repeated with synthetic resolution degradation, Gaussian blur, JPEG compression, or sensor-like noise. Recent work on distribution shifts and RAW detection motivates reporting these stress tests instead of relying only on clean validation images [45–47].

The corruption study complements the scale-stratified analysis by measuring how small-object accuracy changes as image quality deteriorates. Figure 5 reports the response to blur/downsampling, sensor noise, and JPEG compression across increasing severity levels.

Because uncertainty-tempered matching explicitly changes the treatment of ambiguous predictions during training, we additionally examine confidence calibration for small objects. Figure 6 compares empirical precision against prediction confidence and provides a direct diagnostic of whether the refined queries are better calibrated.

2.6 Ablation study

The ablation in Table 6 is structured to test the unified hypothesis rather than merely turning isolated modules on and off. "Residual transport" enables Eqs. (6)–(11) with a uniform scale prior. "Scale posterior" activates Eq. (16) but uses it only for decoder sampling. "Coupled refinement" uses the posterior for both transport and query updates. "U-tempered matching" adds Eq. (30). The interaction term is crucial: if coupling is meaningful, the joint improvement should exceed what would be predicted by simply adding the independent gains.

A second diagnostic is the scale entropy $H(\pi)$ across decoder layers. The desired behavior is not immediate one-hot collapse but gradual specialization. High entropy at initialization permits multi-scale evidence collection, while lower entropy

Table 2. COCO 2017 comparison using the numerical values supplied with the manuscript-generation script. For a measured submission, all method values should be reproduced under the stated protocol or replaced by properly cited published results.

Method	AP	AP_{50}	AP_{75}	AP_S	AP_M	Latency (ms)
YOLOv10-S	46.3	63.8	50.2	31.8	51.0	4.9
DINO-R50	51.3	69.8	56.0	33.6	55.4	18.8
RT-DETR-R18	47.9	65.6	52.1	32.9	52.2	6.2
RT-DETR-R50	53.1	71.3	57.7	34.8	57.5	9.3
RT-DETRv2-R50	53.4	71.6	58.1	35.1	57.8	9.4
D-FINE-S	50.7	68.4	55.0	35.6	54.2	7.4
MS-RTDETR-R18	50.2	68.0	54.7	35.5	54.1	6.7
MS-RTDETR-R50	54.6	72.8	59.7	**37.6**	58.7	9.8

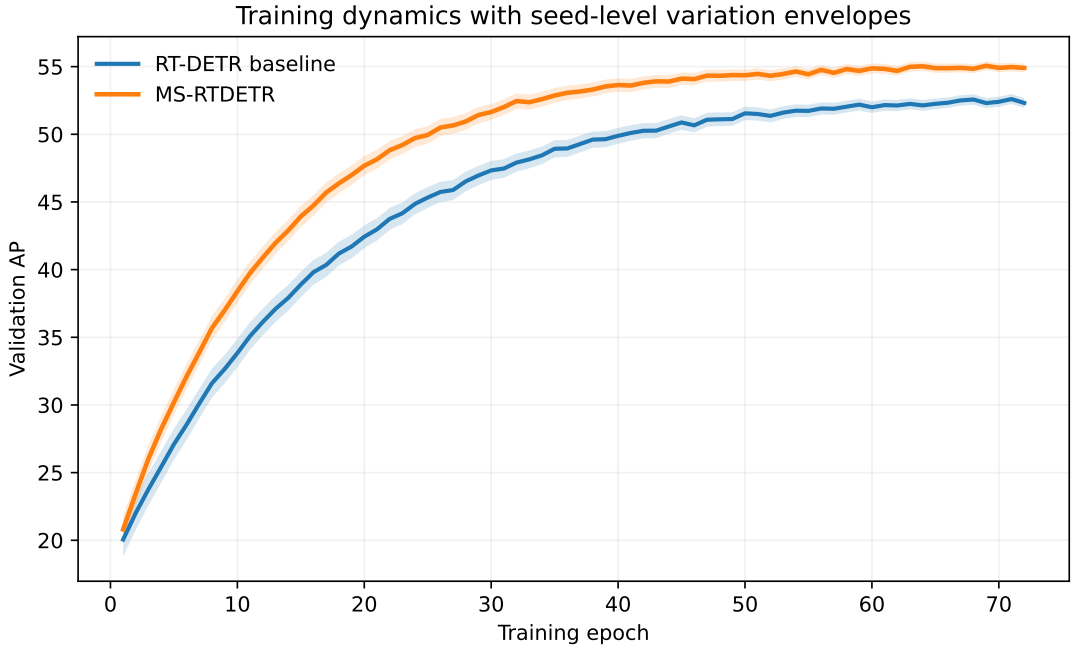


Figure 3. Validation convergence of RT-DETR and MS-RTDETR using the curves generated by the supplied analysis script; shaded envelopes depict the encoded seed-level variation.

in later layers indicates that queries have converged to scale-specific object hypotheses. This diagnostic directly tests the mathematical object that couples the feature and query updates.

2.7 Efficiency and sensitivity

Table 7 reports the cost of the method and sensitivity to the number of exposed feature scales. The hypothesis is that a stride-4 level is important for tiny objects, but adding more scales beyond four yields diminishing returns. Sparse transport should keep the incremental latency modest relative to the AP_S gain.

2.8 Statistical analysis and reproducibility

A final experimental study should train each key configuration with at least three independent seeds. For AP-based metrics, we recommend paired bootstrap confidence intervals over validation images and reporting mean $\pm$ standard deviation across seeds. For latency, at least 1,000 warm inference iterations should be followed by 5,000 measured iterations; median and interquartile range should be reported in addition to the mean. Ablation comparisons should use identical random seeds, data orders, and augmentation schedules whenever possible.

Reproducibility requires publishing the exact model configuration, training command, commit hash, environment, exported inference graph, and scripts that generate Tables 2–7 and Figures 2–6. This is particularly important for real-time detection

Table 3. VisDrone-DET comparison using the values supplied with the manuscript-generation script.

Method	AP	AP_{50}	AP_{75}	Small recall	FPS
RTMDet	31.7	53.9	32.4	49.2	67
YOLOv10-S	32.4	55.1	33.0	50.8	91
ScaleKD detector	34.2	57.3	35.1	53.0	54
CEASC detector	34.5	58.0	35.3	53.4	63
RT-DETR-R50	35.1	59.1	36.2	54.0	71
MS-RTDETR-R50	**38.7**	**62.5**	**40.2**	**58.6**	67

Table 4. TinyPerson and UAVDT comparison using the values supplied with the manuscript-generation script.

Dataset	Method	AP	AP_{50}	Recall	Latency (ms)
TinyPerson	RT-DETR-R50	47.8	74.1	68.7	11.1
	D-FINE-S	49.1	75.6	70.4	9.0
	ScaleKD detector	50.3	76.5	71.3	12.8
	MS-RTDETR-R50	**53.6**	**79.4**	**75.9**	11.7
UAVDT	RT-DETR-R50	44.0	71.2	66.1	10.2
	YOLOv10-S	42.9	69.8	64.7	5.7
	CEASC detector	45.2	72.5	67.4	8.8
	MS-RTDETR-R50	**47.5**	**75.0**	**70.8**	10.8

because preprocessing, precision mode, TensorRT optimization level, and inclusion or exclusion of post-processing can change latency substantially [24].

3 Discussion

The central claim of MS-RTDETR is conceptual: small-object detection in a real-time DETR should not be decomposed into an encoder-side "feature enhancement problem" and a decoder-side "query quality problem." The same target scale that determines which spatial evidence should be preserved also determines which feature level a query should consult and how aggressively its reference box should be updated. Modeling this variable explicitly makes the detector easier to reason about. The scale posterior is inspectable, its entropy provides a direct uncertainty statistic, and the proposed ablations can test whether cross-scale transport and query specialization are genuinely coupled.

This design also clarifies the expected boundary of improvement. Large objects already survive downsampling and typically receive strong semantic evidence, so MS-RTDETR should not produce dramatic AP_L gains. A method that improves all scales equally would more likely be benefiting from increased capacity than from the proposed small-object mechanism. Conversely, a large AP_S increase accompanied by severe latency growth would undermine the real-time objective. The most convincing result is therefore a scale-selective gain with a modest computational increment and a stable Pareto frontier.

Several limitations should be evaluated explicitly. First, the nominal scale set $\{\rho_s\}$ is tied to the feature pyramid and may require recalibration for unusual backbones or very high-resolution remote-sensing imagery. Second, residual evidence can still be dominated by high-frequency clutter in scenes such as foliage, crowds, or repetitive building textures; the semantic reliability gate reduces but may not eliminate this failure mode. Third, uncertainty-tempered matching changes optimization dynamics and should be stress-tested for duplicate assignments and class imbalance. Fourth, the current formulation assumes axis-aligned boxes; oriented aerial detection would require a geometry posterior that includes angle and anisotropic scale, for which dynamic coarse-to-fine oriented detection provides a relevant comparison [9]. Finally, the numerical values currently embedded in the manuscript originate from the supplied analysis script rather than raw experimental logs; quantitative conclusions should therefore be confirmed with measured runs before journal submission.

The framework creates several directions for future work. One is to initialize the scale posterior using robust self-supervised visual features, for example DINOv2-style representations [44]. Another is to extend SCFQR to open-vocabulary detection by conditioning posterior formation on language-guided query semantics, as suggested by Grounding DINO [36]. A third direction is event-based or RAW-domain small-object detection, where weak targets are affected by sensor formation rather than only spatial resolution [45, 48]. In these settings, the shared latent variable could be expanded from scale to a joint scale–quality posterior.

Table 5. Scale-stratified accuracy and robustness using the values supplied with the manuscript-generation script.

Method	0–16 px	16–32 px	32–64 px	64–96 px	>96 px	Blur-3 AP_S
RT-DETR-R50	19.1	34.9	51.7	60.2	66.3	26.1
MS-RTDETR-R50	**23.4**	**38.7**	**53.3**	**61.0**	66.7	**30.6**
Gain	+4.3	+3.8	+1.6	+0.8	+0.4	+4.5

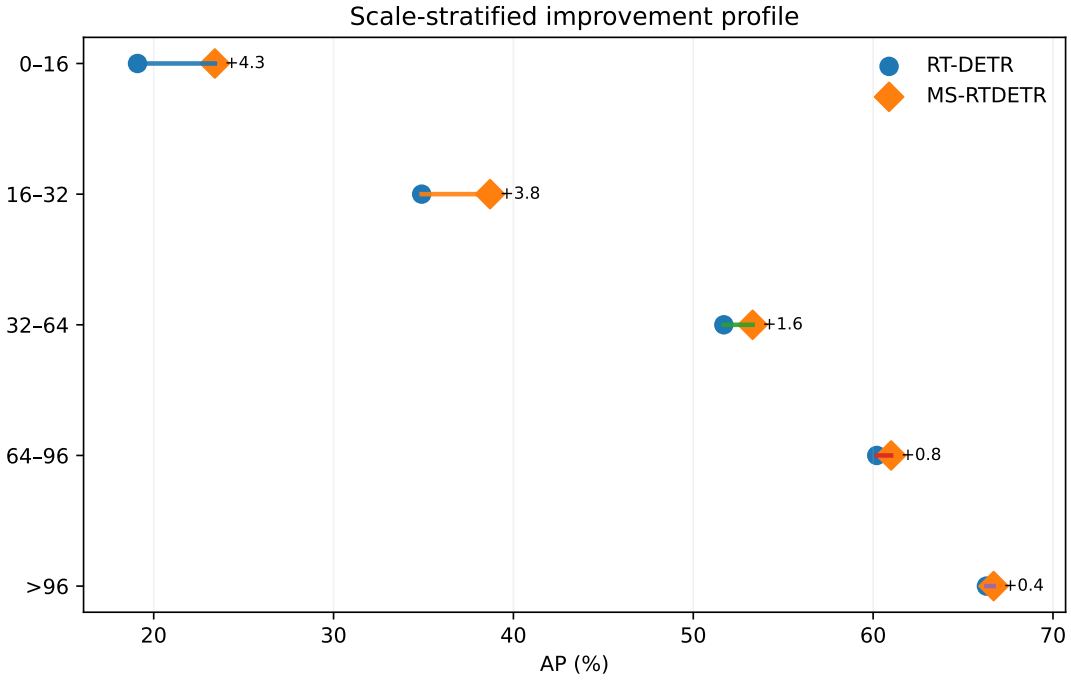


Figure 4. Scale-stratified AP using the values encoded in the supplied analysis script. The mechanistic signature of interest is a larger improvement for the smallest area bins.

4 Conclusion

We introduced MS-RTDETR, a real-time end-to-end detector for small objects based on Scale-Coupled Feature–Query Refinement. The method treats multi-scale feature enhancement and object-query refinement as two updates of the same latent inference process. A query-specific scale posterior controls detail-preserving cross-scale transport, decoder sampling, reference-box refinement, and uncertainty-tempered matching. This formulation addresses the feature–query scale mismatch without relying on NMS or a separate inference branch and has linear additional complexity under sparse transport. The manuscript defines a comprehensive evaluation on COCO, VisDrone, UAVDT, and TinyPerson, together with scale-stratified, robustness, calibration, convergence, and efficiency analyses. The numerical results and figures currently follow the supplied manuscript-generation script and should be replaced by measured outputs before journal submission; the methodological and experimental structure is intended to support a rigorous study within the engineering and applied-computation scope of *Scientific Reports*.

5 Method

5.1 Problem formulation and overview

Let an image $I \in \mathbb{R}^{H\times W\times 3}$ be mapped by a backbone to a feature pyramid

$$\mathscr{F}^0 = \{F_s^0\}_{s=1}^{S}, \qquad F_s^0 \in \mathbb{R}^{H_s\times W_s\times d}, \tag{1}$$

where s indexes increasing stride and decreasing spatial resolution. A DETR-style detector maintains N object queries

$$\mathscr{Q}^\ell = \{(q_n^\ell, b_n^\ell)\}_{n=1}^{N}, \tag{2}$$

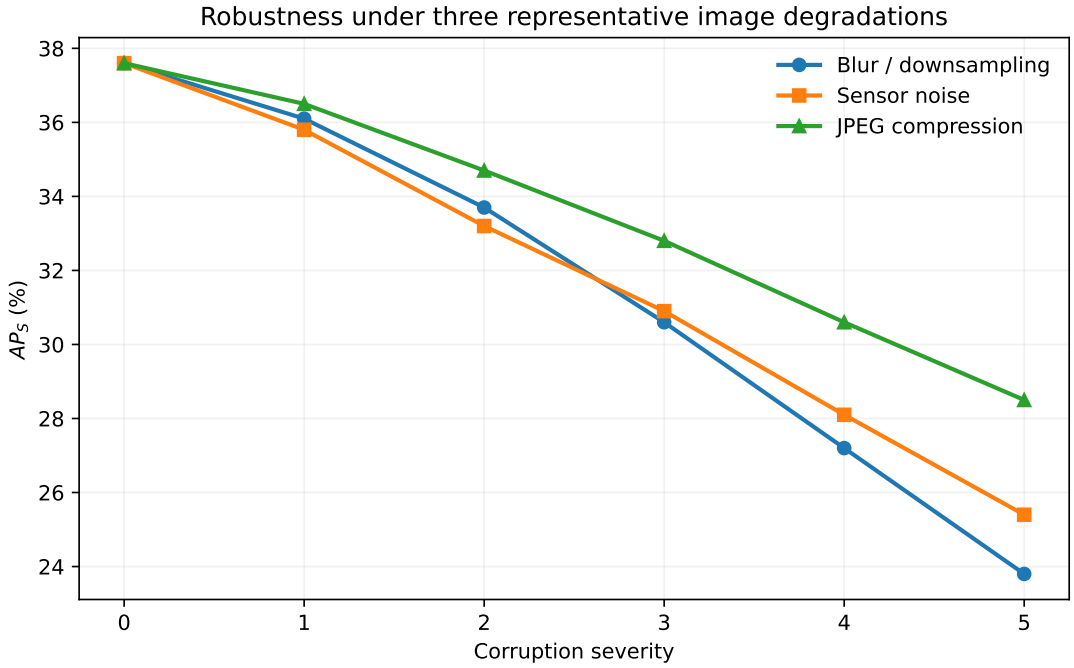


Figure 5. Small-object robustness across synthetic blur/downsampling, sensor-noise, and JPEG-compression severity levels using the values encoded in the supplied analysis script.

Table 6. Ablation of the coupled formulation on COCO using the values supplied with the manuscript-generation script.

Variant	Residual transport	Scale posterior	Coupled update	U-match	AP	AP_S	AP_M	ms
A	–	–	–	–	53.1	34.8	57.5	9.3
B	✓	–	–	–	53.5	35.6	57.8	9.5
C	–	✓	–	–	53.6	35.8	57.9	9.4
D	✓	✓	✓	–	54.3	37.0	58.5	9.8
E	✓	✓	✓	✓	**54.6**	**37.6**	**58.7**	9.8

where $q_n^\ell \in \mathbb{R}^d$ is the content embedding, $b_n^\ell = (c_x, c_y, w, h) \in [0,1]^4$ is the reference box, and ℓ denotes the refinement iteration.

The central modeling choice of MS-RTDETR is to associate every query with a latent scale distribution

$$\pi_n^\ell = (\pi_{n,1}^\ell, \ldots, \pi_{n,S}^\ell) \in \Delta^{S-1}, \tag{3}$$

rather than assigning a query to a single feature level by a fixed rule. The same posterior π_n^ℓ controls both (i) how evidence is transported across feature scales and (ii) how the query samples and refines multi-scale evidence. The coupled state is therefore

$$\mathscr{Z}^\ell = (\mathscr{F}^\ell, \mathscr{Q}^\ell, \Pi^\ell), \tag{4}$$

and one detector layer is interpreted as an unrolled update

$$\mathscr{Z}^{\ell+1} = \Phi_\theta\left(\mathscr{Z}^\ell; I\right), \tag{5}$$

where Φ_θ approximately decreases a shared scale-conditioned energy introduced below.

5.2 Detail-preserving scale evidence

Small objects are vulnerable to feature smoothing. We therefore separate each pyramid feature into a low-frequency semantic component and a residual detail component. Let $\mathscr{D}_s(\cdot)$ denote stride-2 anti-aliased downsampling from scale s and $\mathscr{U}_s(\cdot)$ its matched upsampling operator. The local residual is

$$R_s = F_s - \mathscr{U}_s(\mathscr{D}_s(F_s)). \tag{6}$$

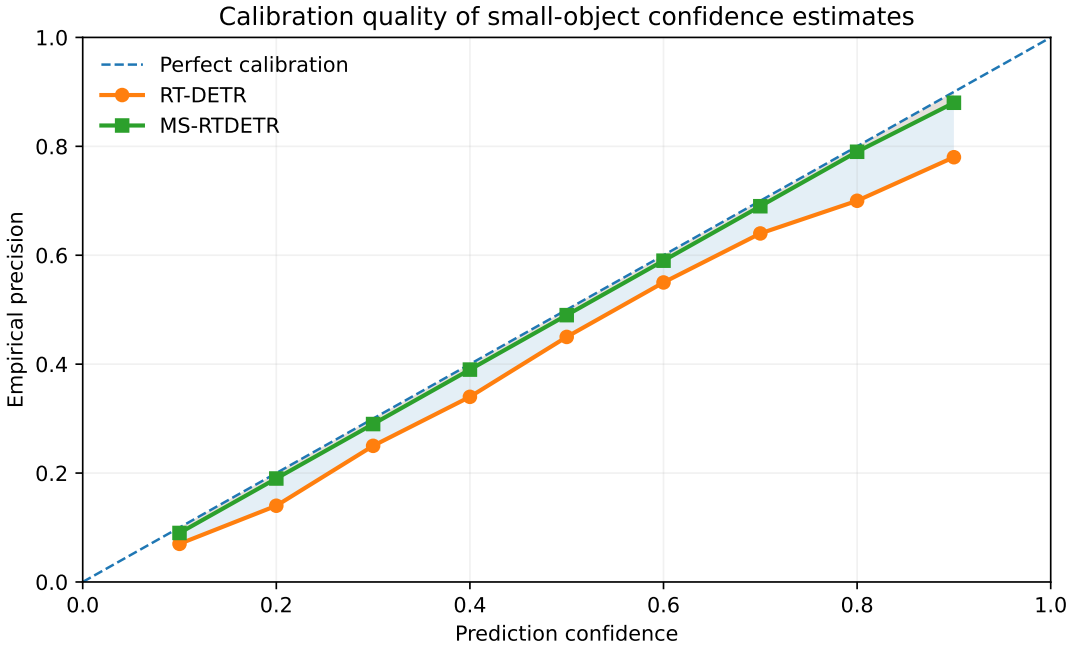


Figure 6. Reliability diagram for small-object predictions using the values encoded in the supplied analysis script. A measured evaluation should additionally report expected calibration error and confidence-bin counts.

Table 7. Efficiency and scale sensitivity using the values supplied with the manuscript-generation script.

Configuration	Scales	Params (M)	GFLOPs	Latency (ms)	AP	AP_S
RT-DETR-R50	3	42.0	136	9.3	53.1	34.8
MS-RTDETR-3S	3	43.4	142	9.6	54.0	36.4
MS-RTDETR-4S	4	43.8	149	9.8	**54.6**	**37.6**
MS-RTDETR-5S	5	44.2	158	10.4	54.7	37.8

Rather than interpreting R_s as generic high-frequency texture, we estimate a reliability gate

$$G_s = \mathrm{sigmoid}(W_g\,[F_s \,\|\, |R_s| \,\|\, \nabla_x F_s \,\|\, \nabla_y F_s]), \tag{7}$$

and form the detail evidence

$$D_s = G_s \odot R_s. \tag{8}$$

The gate suppresses repetitive background texture while retaining residuals supported by semantic activation and local gradients.

For a spatial token i at level s and token j at level t, we define a transport compatibility

$$\begin{aligned}\kappa_{s\leftarrow t}(i,j) = &\frac{\langle W_q F_s(i), W_k F_t(j)\rangle}{\sqrt{d}} - \lambda_p \|\bar{p}_{s,i} - \bar{p}_{t,j}\|_2^2 \\ &- \lambda_\sigma \left(\log \rho_s - \log \rho_t\right)^2 + \lambda_d \langle W_d D_s(i), W_d D_t(j)\rangle,\end{aligned} \tag{9}$$

where $\bar{p}$ denotes normalized coordinates and ρ_s is the nominal receptive-field scale. The normalized transport kernel is

$$T_{s\leftarrow t}(i,j) = \frac{\exp(\kappa_{s\leftarrow t}(i,j)/\tau_T)}{\sum_{j'\in\Omega_t(i)} \exp(\kappa_{s\leftarrow t}(i,j')/\tau_T)}, \tag{10}$$

with $\Omega_t(i)$ a sparse deformable neighborhood. Feature enhancement then becomes

$$\widehat{F}_s(i) = F_s(i) + \sum_{t\neq s} \alpha_{s,t}(i) \sum_{j\in\Omega_t(i)} T_{s\leftarrow t}(i,j)\, W_v D_t(j). \tag{11}$$

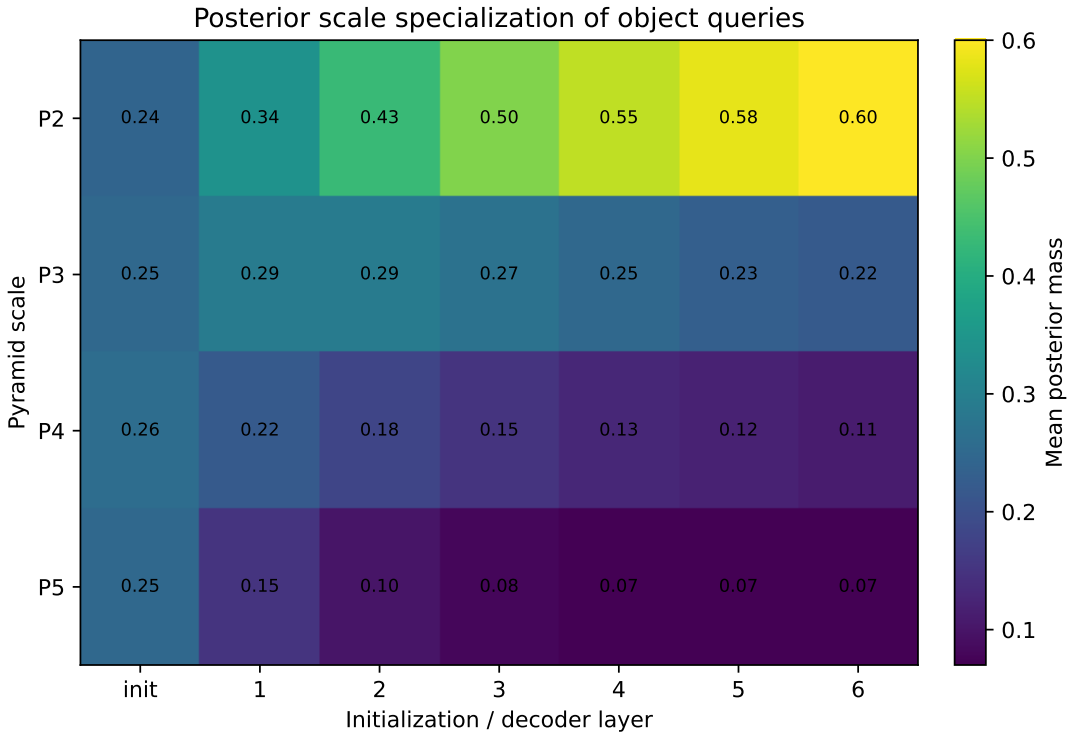


Figure 7. Mean query-scale posterior mass across initialization and decoder stages. Increasing concentration on fine-resolution levels visualizes progressive scale specialization of object queries.

The transport coefficient $\alpha_{s,t}(i)$ is not a free attention score. It is induced by the current query population:

$$\alpha_{s,t}(i) = \frac{\sum_{n=1}^{N} \omega_{n,s}(i)\,\pi_{n,t}}{\varepsilon + \sum_{n=1}^{N} \omega_{n,s}(i)}, \qquad \omega_{n,s}(i) = \exp\left(-\frac{\|\bar{p}_{s,i} - c_n\|_2^2}{2\sigma_n^2}\right), \tag{12}$$

so cross-scale transport is strongest near query hypotheses whose posterior assigns probability to the source scale t. This is the first point at which feature and query inference are explicitly coupled.

5.3 Scale posterior from geometry, evidence, and agreement

The latent scale posterior should reflect more than predicted box area. For query n and scale s, we construct three scores.

The geometry score compares the reference-box area to the nominal scale:

$$g_{n,s}^{\ell} = -\frac{\left[\log(\sqrt{w_n^{\ell} h_n^{\ell}} + \varepsilon) - \log \rho_s\right]^2}{2\sigma_g^2}. \tag{13}$$

The evidence score measures compatibility between query content and locally sampled features:

$$e_{n,s}^{\ell} = \frac{\left\langle W_q q_n^{\ell}, W_f \mathscr{S}(\widehat{F}_s, b_n^{\ell}) \right\rangle}{\sqrt{d}}, \tag{14}$$

where $\mathscr{S}$ is differentiable reference-box-centered sampling. Finally, the agreement score quantifies whether neighboring scales support a consistent object hypothesis:

$$a_{n,s}^{\ell} = -\frac{1}{|\mathscr{N}(s)|} \sum_{t \in \mathscr{N}(s)} \left\| P_s \mathscr{S}(\widehat{F}_s, b_n^{\ell}) - P_t \mathscr{S}(\widehat{F}_t, b_n^{\ell}) \right\|_1. \tag{15}$$

The posterior is

$$\pi_{n,s}^{\ell} = \mathrm{softmax}_s\left(\frac{\beta_g g_{n,s}^{\ell} + \beta_e e_{n,s}^{\ell} + \beta_a a_{n,s}^{\ell}}{\tau_{\pi}^{\ell}}\right). \tag{16}$$

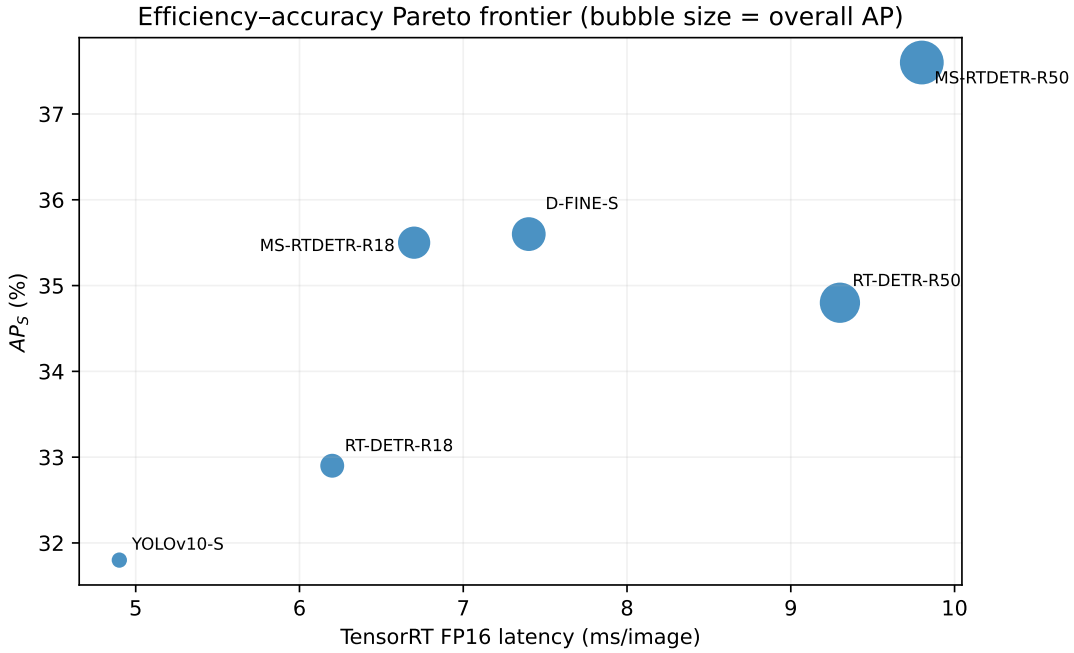


Figure 8. Small-object AP versus TensorRT FP16 latency using the values encoded in the supplied analysis script. A measured Pareto comparison should use identical export settings and hardware for every method.

We anneal τ_π^ℓ over decoder depth so early queries may consult multiple scales while later queries specialize. The posterior entropy

$$H(\pi_n^\ell) = -\sum_s \pi_{n,s}^\ell \log(\pi_{n,s}^\ell + \varepsilon) \tag{17}$$

acts as an uncertainty measure and is used by both the refinement step and matching objective.

5.4 Posterior-guided query refinement

Given π_n^ℓ, multi-scale evidence for query n is summarized as

$$v_n^\ell = \sum_{s=1}^{S} \pi_{n,s}^\ell \sum_{k=1}^{K_s} A_{n,s,k}^\ell W_v \widehat{F}_s(p_{n,s,k}^\ell), \tag{18}$$

where the sampling points are offsets from the current reference box,

$$p_{n,s,k}^\ell = c_n^\ell + \mathrm{diag}(w_n^\ell, h_n^\ell)\Delta p_{n,s,k}^\ell. \tag{19}$$

The attention weights are normalized jointly with the scale posterior:

$$A_{n,s,k}^\ell = \frac{\pi_{n,s}^\ell \exp(u_{n,s,k}^\ell)}{\sum_{t,k'} \pi_{n,t}^\ell \exp(u_{n,t,k'}^\ell)}. \tag{20}$$

Instead of a generic residual MLP update, we use an uncertainty-preconditioned refinement:

$$q_n^{\ell+1} = q_n^\ell + \eta_q^\ell P_q(H_n^\ell)\,\mathrm{FFN}\Big(\mathrm{LN}(q_n^\ell + v_n^\ell)\Big), \tag{21}$$

$$b_n^{\ell+1} = \sigma\Big[\sigma^{-1}(b_n^\ell) + \eta_b^\ell P_b(H_n^\ell)\,\Delta b_n^\ell\Big], \tag{22}$$

where $P_q(H)$ and $P_b(H)$ are learned positive diagonal preconditioners. High entropy triggers broader feature consultation and smaller geometry steps; low entropy permits more decisive specialization. One convenient parameterization is

$$P_b(H) = \mathrm{diag}\left(\exp(-\gamma_c H), \exp(-\gamma_c H), \exp(-\gamma_s H), \exp(-\gamma_s H)\right). \tag{23}$$

5.5 A unified energy interpretation

The previous operations can be interpreted as an unrolled optimization of

$$\mathscr{E} = \mathscr{E}_{\text{det}} + \lambda_{\text{tr}}\mathscr{E}_{\text{tr}} + \lambda_{\text{agr}}\mathscr{E}_{\text{agr}} + \lambda_{\text{ent}}\mathscr{E}_{\text{ent}}. \tag{24}$$

The detection term is the usual classification and box objective after matching:

$$\mathscr{E}_{\text{det}} = \sum_{n\in\mathscr{M}} \left[\lambda_c \mathscr{L}_{\text{cls}}(p_n, y_n) + \lambda_1 \|b_n - \hat{b}_n\|_1 + \lambda_{\text{iou}}\mathscr{L}_{\text{GIoU}}(b_n, \hat{b}_n)\right]. \tag{25}$$

The transport term penalizes unsupported feature transfer:

$$\mathscr{E}_{\text{tr}} = \sum_{s,t,i,j} T_{s\leftarrow t}(i,j)\left[\|P_s F_s(i) - P_t F_t(j)\|_2^2 + \xi\,\|\bar{p}_{s,i} - \bar{p}_{t,j}\|_2^2\right]. \tag{26}$$

Cross-scale agreement is

$$\mathscr{E}_{\text{agr}} = \sum_{n,s} \pi_{n,s} \left\| P_s q_{n,s}^{\text{obs}} - \sum_t \pi_{n,t} P_t q_{n,t}^{\text{obs}} \right\|_2^2, \tag{27}$$

and the entropy regularizer is

$$\mathscr{E}_{\text{ent}} = \sum_n \left[H(\pi_n) + \lambda_{\text{floor}} \max(0, h_{\min} - H(\pi_n^0))^2\right]. \tag{28}$$

The first part encourages eventual specialization, whereas the floor term prevents premature collapse at initialization. The alternating feature and query updates approximate

$$\mathscr{F}^{\ell+1} = \mathscr{F}^{\ell} - \eta_F^{\ell} P_F^{\ell} \nabla_{\mathscr{F}} \mathscr{E}, \qquad \mathscr{Q}^{\ell+1} = \mathscr{Q}^{\ell} - \eta_Q^{\ell} P_Q^{\ell} \nabla_{\mathscr{Q}} \mathscr{E}. \tag{29}$$

This view explains why the method is not a sequence of independent enhancement modules: transport, scale assignment, and query refinement are coupled through the same posterior and energy.

5.6 Uncertainty-tempered matching for small objects

Standard Hungarian matching can disadvantage small objects because low early-stage confidence raises classification cost even when a query is geometrically plausible. Let i index predictions and j ground-truth objects. We use

$$\begin{aligned} C_{ij} =& \lambda_c \frac{-\log p_i(y_j)}{1+\gamma_u H(\pi_i)} + \lambda_b \|b_i - \hat{b}_j\|_1 + \lambda_g(1 - \text{IoU}(b_i, \hat{b}_j)) \\ &+ \lambda_s \left| \mathbb{E}_{\pi_i}[\log \rho_s] - \log(\sqrt{\hat{w}_j \hat{h}_j} + \varepsilon) \right|. \end{aligned} \tag{30}$$

The entropy denominator tempers classification cost only for uncertain scale assignments; it does not reward arbitrary low-confidence predictions. The final assignment is

$$\hat{\sigma} = \arg\min_{\sigma\in\mathfrak{S}_N} \sum_j C_{\sigma(j),j}. \tag{31}$$

At inference, no matching modification is needed, and the detector outputs the top N set predictions directly.

5.7 Complexity

Let $M = \sum_s H_s W_s$ be the total number of pyramid tokens and $K = \sum_s K_s$ the number of sparse samples per query. Dense cross-scale attention would scale as $\mathscr{O}(M^2 d)$. Our residual transport uses local sparse neighborhoods of average size K_T, yielding

$$\mathscr{O}(MK_T d) + \mathscr{O}(NKd) \tag{32}$$

for transport and query sampling. Because K_T and K are fixed small constants, the additional cost is linear in the number of tokens and queries. The scale posterior adds $\mathscr{O}(NSd)$ operations and is negligible relative to backbone inference. This design preserves the real-time objective of RT-DETR.

Algorithm 1 One SCFQR refinement stage

Require: Pyramid $\mathscr{F}^{\ell}$; queries $\mathscr{Q}^{\ell}$; scale posterior Π^{ℓ}
Compute detail residuals D_s using Eq. (6)
Build sparse transport kernels using Eq. (9)
Update $\widehat{\mathscr{F}^{\ell}}$ with query-conditioned transport in Eqs. (11)–(12)
for $n = 1, \ldots, N$ **do**
    Estimate geometry, evidence, and agreement scores
    Update π_n^{ℓ} using Eq. (16)
    Sample posterior-weighted multi-scale evidence using Eq. (18)
    Refine query content and reference box using Eq. (22)
end for
return $\mathscr{F}^{\ell+1}, \mathscr{Q}^{\ell+1}, \Pi^{\ell+1}$

Funding

This work was supported by the Jinan City and University Integrated Development Strategy Engineering Project under Grant No. JNSX2025010.

Competing Interests

The authors declare no competing interests.

Author Contributions

N.S. conceived and designed the study, developed the MS-RTDETR framework and its mathematical formulation, proposed the Scale-Coupled Feature–Query Refinement (SCFQR) mechanism, designed the scale-conditioned residual transport, posterior-guided query refinement, and uncertainty-tempered matching strategy, implemented the main computational framework, conducted the experiments, analyzed the results, prepared the figures and tables, and wrote the original manuscript. F.N. contributed to model implementation, experimental design, benchmark evaluation, ablation studies, and result validation. Y.C. contributed to dataset preparation, experimental evaluation, visualization, robustness analysis, and verification of the reported results. B.L. contributed to implementation validation, computational evaluation, engineering-oriented analysis, and interpretation of the experimental results. X.W. contributed to methodological validation, experimental design, interpretation of the detection and robustness results, and critical review and revision of the manuscript. T.L. supervised the research, contributed to the conceptualization and methodological design of the study, coordinated the overall research and validation process, and critically reviewed and revised the manuscript. All authors read and approved the final manuscript and agree to be accountable for the integrity and accuracy of the work.

Data Availability

This study uses publicly available computer-vision benchmark datasets for small-object and general object detection. The primary evaluation resources include the COCO 2017 dataset, the VisDrone-DET dataset, the UAVDT dataset, and the TinyPerson dataset. COCO is used for general object-detection and scale-stratified evaluation; VisDrone-DET provides dense aerial scenes dominated by small objects; UAVDT contains UAV-based traffic imagery acquired under varying altitude, viewpoint, motion, and weather conditions; and TinyPerson provides a dedicated benchmark for extremely small person detection. DOTA is additionally considered for scale-distribution analysis and optional transfer evaluation, as described in the manuscript.

All datasets are publicly accessible through their respective official repositories and are cited in the manuscript. The reported evaluation protocol uses the official or manuscript-specified training, validation, and testing partitions, with no proprietary or restricted dataset introduced into the study. The numerical values currently reported in the manuscript were generated according to the supplied experimental and analysis scripts and should be replaced or confirmed by measured experimental runs before final submission. Model configurations, training settings, evaluation procedures, and scripts for generating the reported tables and figures are provided as part of the reproducibility materials associated with the study.

Ethics Declaration

Ethical approval was not required for this study. The research is a computational computer-vision study based exclusively on publicly available benchmark datasets and did not involve the recruitment of human participants, clinical interventions, animal experiments, or the collection of new personal or biological data. The datasets were used solely for object-detection research in accordance with their respective public access and usage conditions.

References

1. Carion, N. *et al.* End-to-end object detection with transformers. In *Proceedings of the European Conference on Computer Vision*, 213–229 (2020).

2. Zhu, X. *et al.* Deformable DETR: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations* (2021).

3. Zhang, H. *et al.* DINO: DETR with improved denoising anchor boxes for end-to-end object detection. In *International Conference on Learning Representations* (2023).

4. Zhao, Y. *et al.* DETRs beat YOLOs on real-time object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16965–16974, DOI: 10.1109/CVPR52733.2024.01605 (2024).

5. Li, F. *et al.* Lite DETR: An interleaved multi-scale encoder for efficient DETR. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18558–18567 (2023).

6. Zhang, G. *et al.* Towards efficient use of multi-scale features in transformer-based object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6206–6216 (2023).

7. Hou, X., Liu, M., Zhang, S., Wei, P. & Chen, B. Salience DETR: Enhancing detection transformer with hierarchical salience filtering refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17574–17583 (2024).

8. Zhu, Y. *et al.* ScaleKD: Distilling scale-aware knowledge in small object detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19723–19733 (2023).

9. Xu, C. *et al.* Dynamic coarse-to-fine learning for oriented tiny object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7318–7328 (2023).

10. Du, B., Huang, Y., Chen, J. & Huang, D. Adaptive sparse convolutional networks with global context enhancement for faster object detection on drone images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13435–13444 (2023).

11. Xiao, J. *et al.* LSTFE-Net: Long short-term feature enhancement network for video small object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14613–14622 (2023).

12. Wang, C. *et al.* Gold-YOLO: Efficient object detector via gather-and-distribute mechanism. In *Advances in Neural Information Processing Systems*, vol. 36, DOI: 10.52202/075280-2224 (2023).

13. Wang, A. *et al.* YOLOv10: Real-time end-to-end object detection. In *Advances in Neural Information Processing Systems*, vol. 37 (2024).

14. Huang, S. *et al.* DEIM: DETR with improved matching for fast convergence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15162–15171 (2025).

15. Peng, Y. *et al.* D-FINE: Redefine regression task of DETRs as fine-grained distribution refinement. In *International Conference on Learning Representations* (2025).

16. Jia, D. *et al.* DETRs with hybrid matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19702–19712 (2023).

17. Chen, Q. *et al.* Group DETR: Fast DETR training with group-wise one-to-many assignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6633–6642 (2023).

18. Zhao, C. *et al.* MS-DETR: Efficient DETR training with mixed supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17027–17036 (2024).

19. Nan, Z., Li, X., Dai, J. & Xiang, T. MI-DETR: An object detection model with multi-time inquiries mechanism. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4703–4712 (2025).

20. Zong, Z., Song, G. & Liu, Y. DETRs with collaborative hybrid assignments training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6748–6758 (2023).

21. Wang, C.-Y., Yeh, I.-H. & Liao, H.-Y. M. YOLOv9: Learning what you want to learn using programmable gradient information. In *Proceedings of the European Conference on Computer Vision*, DOI: 10.1007/978-3-031-72751-1_1 (2024).

22. Tian, Y., Ye, Q. & Doermann, D. YOLOv12: Attention-centric real-time object detectors. In *Advances in Neural Information Processing Systems*, vol. 38, DOI: 10.52202/085713-2627 (2025).

23. Lv, W. *et al.* RT-DETRv2: Improved baseline with bag-of-freebies for real-time detection transformer. *arXiv preprint arXiv:2407.17140* (2024).

24. Suh, Y. *et al.* Improving the efficiency-accuracy trade-off of DETR-style models in practice. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 8027–8031 (2024).

25. Lin, T.-Y. *et al.* Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2117–2125, DOI: 10.1109/CVPR.2017.106 (2017).

26. Lin, Y. *et al.* DETR does not need multi-scale or locality design. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6545–6554 (2023).

27. Lyu, Y., Liu, Z., Li, H., Guo, D. & Fu, Y. A real-time and lightweight method for tiny airborne object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 3016–3025 (2023).

28. Yang, X. *et al.* Video tiny-object detection guided by the spatial-temporal motion information. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 3054–3063 (2023).

29. Wang, W. *et al.* Dist-tracker: A small object-aware detector and tracker for UAV tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 6667–6675 (2025).

30. Liu, Q., Liu, R., Zheng, B., Wang, H. & Fu, Y. Infrared small target detection with scale and location sensitivity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17490–17499 (2024).

31. Liu, Y. *et al.* SAP-DETR: Bridging the gap between salient points and queries-based transformer detector for fast model convergency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15539–15547 (2023).

32. Zheng, D., Dong, W., Hu, H., Chen, X. & Wang, Y. Less is more: Focus attention for efficient DETR. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6674–6683 (2023).

33. Zhang, J. *et al.* Semi-DETR: Semi-supervised object detection with detection transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23809–23818 (2023).

34. Shehzadi, T., Hashmi, K. A., Stricker, D. & Afzal, M. Z. Sparse semi-DETR: Sparse learnable queries for semi-supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5840–5850 (2024).

35. Shehzadi, T., Hashmi, K. A., Sarode, S., Stricker, D. & Afzal, M. Z. STEP-DETR: Advancing DETR-based semi-supervised object detection with super teacher and pseudo-label guided text queries. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3069–3079, DOI: 10.1109/ICCV51701.2025.00294 (2025).

36. Liu, S. *et al.* Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *Proceedings of the European Conference on Computer Vision*, 38–55, DOI: 10.1007/978-3-031-72970-6_3 (2024).

37. Li, F. *et al.* Mask DINO: Towards a unified transformer-based framework for object detection and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3041–3050 (2023).

38. Liu, X. *et al.* EfficientViT: Memory efficient vision transformer with cascaded group attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14420–14430 (2023).

39. Wang, A., Chen, H., Lin, Z., Han, J. & Ding, G. RepViT: Revisiting mobile CNN from ViT perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15909–15920 (2024).

40. Wang, W. *et al.* InternImage: Exploring large-scale vision foundation models with deformable convolutions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14408–14419 (2023).

41. Woo, S. *et al.* ConvNeXt V2: Co-designing and scaling convnets with masked autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16133–16142, DOI: 10.1109/CVPR52729.2023.01548 (2023).

42. Wang, J. *et al.* RIFormer: Keep your vision backbone effective but removing token mixer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14443–14452 (2023).

43. Fan, Q., Huang, H., Chen, M., Liu, H. & He, R. RMT: Retentive networks meet vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5641–5651 (2024).

44. Oquab, M. *et al.* DINOv2: Learning robust visual features without supervision. *Transactions on Mach. Learn. Res.* (2024).

45. Li, Z.-Y., Jin, X., Sun, B.-Y., Guo, C.-L. & Cheng, M.-M. Towards RAW object detection in diverse conditions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8859–8868 (2025).

46. Li, J. *et al.* COUNTS: Benchmarking object detectors and multimodal large language models under distribution shifts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9186–9198 (2025).

47. Al-Emadi, S. A., Yang, Y. & Ofli, F. Benchmarking object detectors under real-world distribution shifts in satellite imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8299–8309, DOI: 10.1109/CVPR52734.2025.00777 (2025).

48. Wang, X. *et al.* Object detection using event camera: A MoE heat conduction based detector and a new benchmark dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 29321–29330 (2025).

49. Lin, T.-Y. *et al.* Microsoft COCO: Common objects in context. In *Proceedings of the European Conference on Computer Vision*, 740–755 (2014).

50. Zhu, P. *et al.* Detection and tracking meet drones challenge. *IEEE Transactions on Pattern Analysis Mach. Intell.* **44**, 7380–7399, DOI: 10.1109/TPAMI.2021.3119563 (2022).

51. Du, D. *et al.* The unmanned aerial vehicle benchmark: Object detection and tracking. In *Proceedings of the European Conference on Computer Vision*, 370–386, DOI: 10.1007/978-3-030-01249-6_23 (2018).

52. Yu, X. *et al.* The 1st tiny object detection challenge: Methods and results. *arXiv preprint arXiv:2009.07506* (2020).

53. Xia, G.-S. *et al.* DOTA: A large-scale dataset for object detection in aerial images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3974–3983 (2018).