GATESLICE: SCALE-SELECTIVE SLICED FUSION FOR TRAINING-FREE SMALL-OBJECT DETECTION WITH FROZEN DETECTORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Frozen detectors can miss small objects when high-resolution images are globally resized. Tiling restores local detail, but incomplete context, truncated objects, and shifted confidence scores make tiled predictions unreliable when merged indiscriminately with global predictions. We propose GateSlice, a training-free dual-view framework that selects complementary evidence according to object scale. Global predictions serve as an anchor; after mapping tiled detections to image coordinates, an area gate admits only small-scale candidates. A fixed detector-specific score factor aligns confidence across views, and class-aware nonmaximum suppression resolves duplicates. GateSlice requires no learnable parameters or weight updates. On COCO, TinyPerson, and VisDrone2019, it consistently improves small-object detection across five frozen detectors spanning Transformer, one-stage, and two-stage architectures. Ablations and instance-level analyses show that the gains primarily come from recovering objects missed after global resizing, while gating and score calibration limit interference with global predictions. These results cast inference-time enhancement as selective fusion of views with scale-dependent reliability.

1 INTRODUCTION

Small-object detection aims to recognize and localize objects that occupy only a minute fraction of an image. It is indispensable to long-range perception applications such as autonomous driving, video surveillance, and unmanned aerial inspection. Despite sustained progress in object detection, small objects remain difficult to detect because they provide inherently limited texture, boundary, and local-contrast cues. This difficulty is further amplified in high-resolution imagery: modern detectors commonly resize the full image to a prescribed input resolution, compressing small objects to only a few pixels before feature extraction. Their weak responses may subsequently be overwhelmed by background features or assigned lower priorities during detector-specific candidate selection and ranking. Consequently, fixed-resolution global resizing creates an input-side evidence bottleneck shared by different detection architectures.

Existing methods primarily address this bottleneck through training-time modifications, including higher input resolutions, specialized small-object heads, and multiscale representations (Lin et al., 2017; Kim et al., 2018; Yuan et al., 2023), as well as redesigned matching, assignment, or localization objectives (Xu et al., 2022a;b). Although effective, they require parameter optimization, retraining, or architectural adaptation, limiting their applicability to deployed or frozen detectors. More fundamentally, they leave unanswered whether an already trained detector can be improved solely by reorganizing the evidence presented at inference time.

Sliced inference offers a direct way to recover visual evidence suppressed by global resizing: it partitions a high-resolution image into local regions and enlarges each region before detection (Yang et al., 2023; Koyun et al., 2022; Shen et al., 2022). Although this transformation increases the effective resolution of small objects, it also induces a view-dependent prediction shift. Cropping removes global context, truncates objects at tile boundaries, and produces duplicate or false-positive detections; moreover, confidence scores from global and tiled views are not necessarily comparable. Consequently, directly pooling their predictions allows unreliable tiled candidates to interfere with

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.

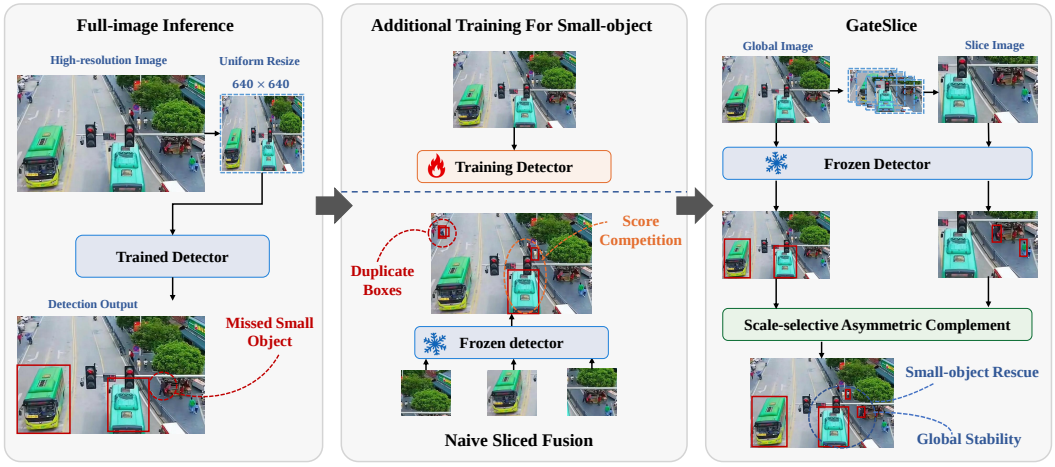

Figure 1: Conceptual comparison. Global inference preserves context but suppresses small-object evidence under resizing, whereas naive sliced fusion recovers local detail at the cost of unreliable predictions. GateSlice uses global predictions as an anchor and selectively incorporates scale-appropriate tiled detections.

otherwise stable global detections. As illustrated in Fig. 1, this conflict reveals a scale-dependent asymmetry between the two views. The global view preserves complete scene context and the detector's original prediction behavior, making it a reliable prediction anchor, whereas tiled views recover fine-grained evidence but are trustworthy primarily for small objects. Thus, the two views should not be treated as interchangeable ensemble members. Instead, training-free enhancement of frozen detectors should be formulated as reliability-aware evidence selection: transformed predictions are admitted only when their scale-dependent benefits outweigh the prediction shift they introduce.

Based on this formulation, we propose Training-Free Scale-Gated Sliced Fusion (GateSlice), a detector-agnostic framework that improves small-object detection without updating model parameters. GateSlice processes a global view and overlapping tiled views using the same frozen detector, assigning them asymmetric roles: global predictions form a stable anchor, while tiled predictions provide complementary evidence for small objects. After tiled detections are mapped to the original image coordinates, an area gate admits only candidates within the scale regime where local enlargement is beneficial. A detector-specific fixed score factor then calibrates cross-view confidence, and class-aware Non-Maximum Suppression resolves duplicates. By jointly controlling which tiled predictions enter the candidate set and how strongly they compete with global predictions, GateSlice converts sliced inference from unconstrained aggregation into reliability-aware, scale-selective fusion.Because GateSlice operates only on input views and output predictions, it is applicable across heterogeneous detector architectures.

Our main contributions are summarized as follows:

- We formulate training-free small-object enhancement as reliability-aware evidence selection and identify the scale-dependent complementarity between global and tiled views.
- We propose GateSlice, which retains global predictions as an anchor and selectively fuses area-gated, score-calibrated tiled candidates without training or model modification.
- We demonstrate consistent improvements across three datasets and five frozen detectors, with analyses showing that GateSlice recovers missed small objects while limiting interference from tiled predictions.

2 Related Work

2.1 Small-Object Detection

Small-object detection is challenging because limited texture and boundary cues are easily overwhelmed by background responses during downsampling. Existing methods mainly improve small-object representations through input-scale adaptation based on image pyramids or multiscale training (Lin et al., 2017; Singh & Davis, 2018; Singh et al., 2018), cross-scale feature aggregation (Liu

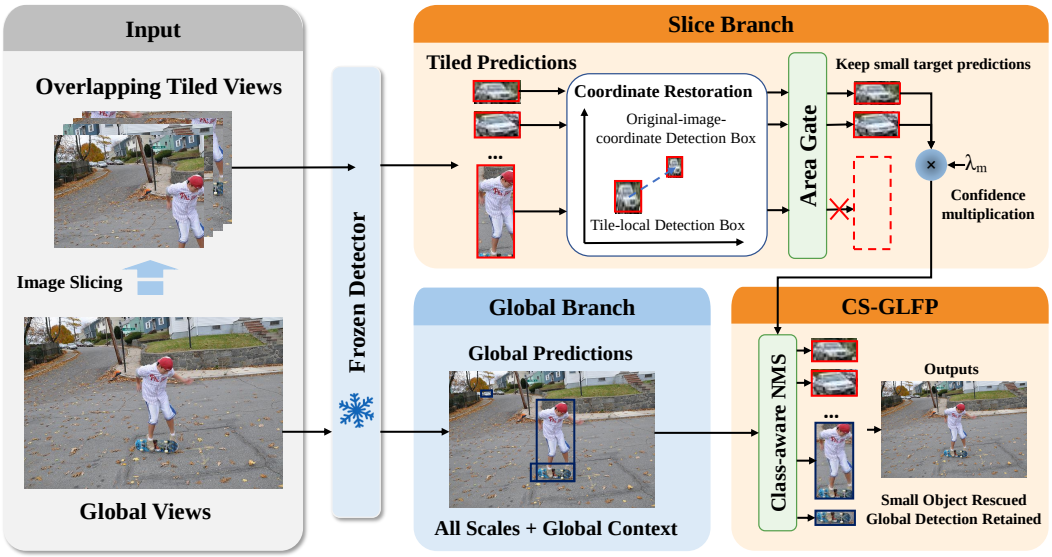

Figure 2: Overview of GateSlice. A frozen detector processes global and tiled views; area gating and score calibration retain reliable small-object candidates, which are merged with the global predictions using class-aware NMS.

et al., 2018; Ghiasi et al., 2019; Tan et al., 2020; Guo et al., 2020), and scale-aware supervision, localization, or sample assignment (Yu et al., 2020; Xu et al., 2022a;b). Selective computation has also been explored to reduce the cost of processing high-resolution features (Yang et al., 2022; Sharma, 2025). Although effective, these approaches generally require retraining or architectural modification and thus cannot be directly applied to deployed frozen detectors. In contrast, we investigate whether fixed-resolution inference suppresses small-object capabilities already encoded in a detector. GateSlice restores local evidence through overlapping tiled views while using area gating, score calibration, and class-aware fusion to restrict local predictions to supplementing stable global detections.

2.2 Sliced Inference for Object Detection

Sliced inference enlarges local regions to improve small-object visibility. Existing methods use scale-constrained pyramids or contextual crops (Singh & Davis, 2018; Singh et al., 2018; Najibi et al., 2019), content-dependent regions selected through clustering, density, attention, or foreground statistics (Yang et al., 2019; Li et al., 2020; Duan et al., 2021; Plastiras et al., 2018; Varga & Zell, 2021), and detector-agnostic regular tiling with coordinate restoration and post-processing (Akyon et al., 2022; Keles et al., 2022). Recent approaches further reduce computational cost through sparse computation, cascaded enlargement, and adaptive patch selection (Yang et al., 2022; Meethal et al., 2023; Zhang et al., 2024). However, learned region selection often requires additional training, while regular tiling introduces truncated, duplicate, and miscalibrated predictions. Moreover, existing methods commonly merge global and local predictions symmetrically or replace the global view. GateSlice instead preserves global predictions as an unrestricted anchor and selectively fuses area-gated, score-calibrated local candidates.

3 Proposed Method

3.1 Overview

This section presents the complete GateSlice inference pipeline. As illustrated in Fig. 2, given a frozen detector and an input image, GateSlice constructs a global view and a set of overlapping tiled views, performs forward inference on each, and conducts scale gating, branch-wise score calibration, and class-aware fusion in the original image coordinate system. The two branches share the same model parameters; no gradients are computed, no weights are updated, and no additional learnable modules are introduced. The global branch preserves the baseline detector's scene-level context and original output distribution, whereas the sliced branch serves only as a supplementary source of

small-object candidates. All additional computation is confined to inference, enabling GateSlice to be applied directly to pretrained detectors.

3.2 Overlapping Tiling and Scale Remapping

This module mitigates the reduction in effective pixels and local detail for small objects caused by global resizing, while overlapping regions reduce the risk of truncating objects at tile boundaries. Let the width and height of the original image $\boldsymbol{I}$ be W and H, respectively. We use a $g \times g$ grid to generate local views and set the overlap ratio between adjacent tiles to o. To ensure complete coverage of boundary regions, the horizontal and vertical tile sizes are defined by Eqs. (1) and (2), respectively. Each tile is then independently resized to $H_0 \times W_0$, increasing the effective pixel footprint of small objects relative to the detector input.

$$w_{\mathrm{t}} = \left\lceil \frac{W}{g-(g-1)o} \right\rceil, h_{\mathrm{t}} = \left\lceil \frac{H}{g-(g-1)o} \right\rceil, \tag{1}$$

where w_{t} and h_{t} denote the tile width and height, respectively, and $\lceil \cdot \rceil$ denotes the ceiling operation, which prevents the rightmost or bottommost regions from being left uncovered. Overlap reduces the probability that an object is truncated at a tile boundary, but it also introduces duplicate candidates across tiles; these duplicates are handled by the subsequent class-aware fusion stage.

3.3 Dual-Branch Frozen Inference

This module preserves both the scene-level context of the global branch and the local scale benefit of the sliced branch. Given an original high-resolution image $\boldsymbol{I}$, GateSlice constructs the global and sliced branches in parallel. The global branch resizes the complete image to $H_0 \times W_0$ and produces the prediction set D_{full}, which serves as the baseline. The sliced branch resizes each of the $g \times g$ overlapping tiles to $H_0 \times W_0$ and processes them with the same frozen detector, producing the tiled prediction set D_{s}.

The coordinates in D_{s} differ from those in the original image and must first be restored. Let the upper-left corner of a local tile be (x_0, y_0) and its width and height be $(w_{\mathrm{t}}, h_{\mathrm{t}})$. The bounding box mapped to the original image coordinates is

$$\boldsymbol{B}_g = \left(x_0 + \frac{x_1^r}{s_x}, y_0 + \frac{y_1^r}{s_y}, x_0 + \frac{x_2^r}{s_x}, y_0 + \frac{y_2^r}{s_y} \right), \tag{2}$$

where $(x_1^r, y_1^r, x_2^r, y_2^r)$ are pixel coordinates in the resized tile, and s_x, s_y are the horizontal and vertical scale factors, respectively: $s_x = \frac{W_0}{w_{\mathrm{t}}}, s_y = \frac{H_0}{h_{\mathrm{t}}}$. This transformation maintains geometric consistency among the original-image crop, the model input resize, and the final detection boxes.

We designate the sliced branch exclusively as a supplement for small objects, while the global branch remains responsible for medium and large objects. Without gating, the sliced branch generates candidates at all scales, and high local confidence can displace valid detections under a limited candidate budget. We therefore introduce an area gate to restrict the role of the sliced branch. Global candidates are not subject to a scale constraint, whereas a tiled candidate is retained only if its area after mapping to the original image satisfies:

$$A_{\mathrm{tile}}(x_1, y_1, x_2, y_2) = (x_2 - x_1)(y_2 - y_1) \leq 32^2 = 1024, \tag{3}$$

where (x_1, y_1, x_2, y_2) are the coordinates of the detection box mapped to the original image. The threshold 1024 follows the COCO definition of a small object.

The confidence distributions of the same object can differ between the global and locally enlarged views. Directly comparing raw scores may give one branch an unwarranted advantage during NMS and Top-K ranking. We therefore multiply only the sliced-candidate scores by a detector-specific fixed factor λ_m, leaving box coordinates and class labels unchanged. λ_m = 1 denotes no calibration, whereas a smaller value lowers the ranking priority of sliced candidates relative to global candidates.

3.4 Cross-Scale Global-Local Prediction Fusion and Post-Processing(CS-GLFP)

After coordinate restoration and area gating, tiled predictions and global candidates are represented in the original image pixel coordinate system. We then merge the two candidate sets and apply non-

Table 1: Detection results on COCO val2017 (COCO-style bounding-box metrics, %).

Method	AP	AP_{50}	AP_s	AR_s
Faster R-CNN (Ren et al., 2015)	46.06	66.50	29.23	42.49
Faster R-CNN+GateSlice	46.42 (+0.36)	67.11 (+0.61)	30.35 (+1.12)	46.12 (+3.63)
FCOS (Tian et al., 2019)	37.35	56.01	18.61	29.93
FCOS+GateSlice	38.51 (+1.17)	57.65 (+1.64)	21.29 (+2.68)	37.35 (+7.42)
Deformable-DETR (Zhu et al., 2021)	43.72	62.78	24.36	39.77
Deformable-DETR+GateSlice	44.11 (+0.39)	64.07 (+1.29)	26.40 (+2.04)	46.67 (+6.89)
RT-DETR (Zhao et al., 2024)	46.41	63.72	28.39	49.71
RT-DETR+GateSlice	46.87 (+0.45)	65.12 (+1.40)	30.11 (+1.72)	53.28 (+3.57)
YOLO11n (Jocher & Qiu, 2024)	38.67	54.09	19.46	33.17
YOLO11n+GateSlice	39.17 (+0.50)	55.88 (+1.78)	20.93 (+1.47)	37.40 (+4.23)

Table 2: Cross-domain results on shared categories of the TinyPerson test set (COCO-style bounding-box metrics, %).

Method	AP	AP_{50}	AP_s	AR_s
Faster R-CNN (Ren et al., 2015)	10.18	26.71	6.91	11.27
Faster R-CNN+GateSlice	12.98 (+2.80)	34.12 (+7.41)	10.04 (+3.13)	18.22 (+6.96)
FCOS (Tian et al., 2019)	4.63	12.61	2.52	4.60
FCOS+GateSlice	7.76 (+3.13)	21.11 (+8.50)	6.31 (+3.79)	11.76 (+7.16)
Deformable-DETR (Zhu et al., 2021)	5.62	16.51	2.79	7.11
Deformable-DETR+GateSlice	8.55 (+2.93)	25.15 (+8.64)	6.06 (+3.27)	16.29 (+9.17)
RT-DETR (Zhao et al., 2024)	6.89	19.38	4.10	10.04
RT-DETR+GateSlice	11.30 (+4.41)	30.81 (+11.43)	9.25 (+5.15)	19.17 (+9.13)
YOLO11n (Jocher & Qiu, 2024)	3.12	8.11	1.19	2.68
YOLO11n+GateSlice	5.44 (+2.32)	15.28 (+7.17)	3.85 (+2.66)	8.22 (+5.54)

maximum suppression separately by predicted class to remove duplicate boxes for the same object and limit the final number of candidates. Specifically, the area-gated tiled set is

$$D_s^g = \{(\boldsymbol{b}_i, c_i, s_i) \in D_s \mid A_{\text{tile}}(\boldsymbol{b}_i) \leq 1024\}, \tag{4}$$

where b_i, c_i, s_i denote the bounding-box coordinates, predicted class label, and confidence score for that class, respectively.

The candidates from the two branches are merged as $D_m = D_{\text{full}} \cup D_s^g$. For class k, we extract $D_k = \{(b_i, c_i, s_i) \in D_m \mid c_i = k\}$ and sort it in descending order of confidence. The intersection over union of two predicted boxes b_i, b_j is defined as $\text{IoU}(\boldsymbol{b}_i, \boldsymbol{b}_j) = \frac{|\boldsymbol{b}_i \cap \boldsymbol{b}_j|}{|\boldsymbol{b}_i \cup \boldsymbol{b}_j|}$. NMS iteratively retains the highest-scoring box and removes all remaining candidates of the same class that satisfy $\text{IoU}(\boldsymbol{b}_i, \boldsymbol{b}_j) > 0.6$. Candidates from different classes do not suppress one another. This operation removes duplicate detections of the same object both across overlapping tiles and between the sliced and global branches.

After NMS has been applied to every class, the retained predictions are merged and reranked by confidence: $D_{\text{out}} = \text{TopK}\left(\bigcup_{k=1}^{C} \text{NMS}_{0.6}(D_k), 300\right)$, where C is the number of classes. TopK retains at most the 300 highest-confidence candidates after fusion, limiting the output size while avoiding premature removal of valid predictions.

4 Experiment and Analysis

Datasets. We evaluate GateSlice on three complementary detection benchmarks. COCO 2017 (Lin et al., 2014) measures in-domain general-purpose detection, while TinyPerson (Yu et al., 2020) and VisDrone2019 (Zhu et al., 2022) evaluate the cross-domain transfer of frozen COCO-pretrained detectors to small-object scenarios. The primary experiments use the complete COCO val2017 split. Cross-domain experiments use the TinyPerson test split and the VisDrone2019 validation split, with no fine-tuning on any dataset.

Metrics. We use pycocotools to report AP@[0.50:0.95], AP_{50}, AP_s, and AR_s. For TinyPerson, sea_person and earth_person are merged as person, while uncertain, ignore, and logo annotations are excluded, leaving 13,687 instances. For VisDrone, we retain the six categories shared with COCO:

Table 3: Cross-domain results on shared categories of the VisDrone2019 validation set (COCO-style bounding-box metrics, %).

Method	AP	AP_{50}	AP_s	AR_s
Faster R-CNN (Ren et al., 2015)	15.39	28.08	6.97	14.13
Faster R-CNN+GateSlice	16.19 (+0.80)	30.20 (+2.12)	8.28 (+1.31)	17.12 (+2.99)
FCOS (Tian et al., 2019)	8.88	16.03	3.09	6.91
FCOS+GateSlice	10.07 (+1.19)	18.86 (+2.83)	5.05 (+1.95)	11.45 (+4.54)
Deformable-DETR (Zhu et al., 2021)	9.33	17.96	2.90	7.69
Deformable-DETR+GateSlice	11.09 (+1.76)	23.07 (+5.12)	5.18 (+2.27)	16.86 (+9.16)
RT-DETR (Zhao et al., 2024)	12.22	22.07	5.71	15.59
RT-DETR+GateSlice	14.22 (+2.00)	27.57 (+5.50)	8.75 (+3.04)	23.66 (+8.07)
YOLO11n (Jocher & Qiu, 2024)	7.41	13.12	1.98	4.68
YOLO11n+GateSlice	8.62 (+1.21)	16.54 (+3.42)	3.88 (+1.90)	9.05 (+4.36)

person (merging pedestrian and people), bicycle, car, motorcycle, bus, and truck, yielding 35,207 instances after excluding van, tricycle, and awning-tricycle.

Implementation details. We use official COCO-pretrained RT-DETR-R18, Deformable-DETR-R50, YOLO11n, FCOS R50-FPN, and Faster R-CNN R50-FPN-v2 weights, frozen throughout evaluation. Images use a short side of 640 pixels and a maximum long side of 1,067 pixels, except for models with native 640×640 inputs. GateSlice uses a 2×2 grid, 0.2 overlap, a $1{,}024\ \mathrm{px}^2$ area threshold, and class-aware NMS at IoU 0.6. The tiled-score factors are 1.0, 0.7, 0.7, 0.85, and 0.7 for the five detectors listed above, respectively. All experiments run on one NVIDIA GeForce RTX 3060 Laptop GPU.

4.1 Results Across Detectors and Datasets

Tables 1–3 compare full-image inference with GateSlice across five frozen detectors. GateSlice consistently improves detection performance on all three datasets: on COCO, it yields gains of 0.36–1.17 AP, 1.12–2.68 AP_s, and 3.57–7.42 AR_s; on TinyPerson, where extremely small pedestrians dominate, the improvements increase to 2.32–4.41 AP and 5.54–9.17 AR_s; and on VisDrone2019, it achieves gains of 0.80–2.00 AP and 2.99–9.16 AR_s despite the greater inter-class competition in dense multi-category scenes. These consistent improvements across transformer-based, one-stage, and two-stage detectors demonstrate that GateSlice generalizes across architectures and domains by recovering small-object evidence through tiled enlargement while limiting interference from unreliable local predictions through constrained fusion.

4.2 Ablation Studies and Analysis

All ablations use frozen Deformable-DETR-R50 on COCO val2017 with shared cached predictions. Variants differ only in area gating, score calibration, and class-aware NMS. We set the area threshold to $1{,}024\ \mathrm{px}^2$, $\lambda_m = 0.7$, the NMS threshold to 0.6, and $\mathrm{maxDets} = 100$.As shown in Table 4, unconstrained fusion improves recall but substantially reduces AP. Area gating provides the largest recovery, while score calibration and NMS further regulate cross-view competition and remove duplicates. The full model improves the baseline by 0.39 AP, 2.04 AP_s, and 6.89 AR_s.

Table 4: Strict component ablation using frozen Deformable-DETR-R50 on the complete COCO val2017 split (%).

Setting	Global branch	Sliced branch	Area gate	Score calib.	Class NMS	AP	AP_s	AR_s
Global baseline	✓	×	×	×	×	43.72	24.36	39.77
Direct concatenation	✓	✓	×	×	×	29.41	24.28	53.06
Symmetric NMS fusion	✓	✓	×	×	✓	35.46	27.76	49.49
Gate only	✓	✓	✓	×	✓	43.54	24.73	47.45
Score only	✓	✓	×	✓	✓	42.06	27.57	48.39
Gate + score, no NMS	✓	✓	✓	✓	×	43.82	24.91	50.85
GateSlice	✓	✓	✓	✓	✓	**44.11**	26.40	46.67

4.3 Visualization Experiment

Figure 3 compares five frozen detectors with and without GateSlice on COCO, TinyPerson, and VisDrone2019. Across different architectures and datasets, GateSlice consistently recovers small objects missed or poorly localized after full-image resizing by restoring their effective scale through



Figure 3: Qualitative comparisons on COCO, TinyPerson, and VisDrone2019. Green dashed boxes denote ground truth, while red and cyan boxes show baseline and GateSlice predictions, respectively. GateSlice consistently recovers small objects missed by the frozen detectors. Best viewed in color and zoomed in.

tiled views. Meanwhile, area gating and score calibration restrict unreliable local predictions, preserving the stable outputs of the global branch. These qualitative results support the quantitative improvements in $\mathrm{AP_s}$ and $\mathrm{AR_s}$ and indicate that fixed-resolution resizing is a general inference-time bottleneck rather than a detector-specific limitation.

4.4 Distributional and Feature-Level Analysis of Scale-Conditioned View Reliability

Motivation and protocol. We hypothesize that view reliability is scale dependent: tiled views recover small-object evidence suppressed by global resizing but may be less reliable for larger objects because of incomplete context and boundary truncation. We test this hypothesis on a fixed 500-image COCO val2017 subset using frozen RT-DETR-R18. Before fusion, we compare the two views using the highest-confidence same-class match at IoU ≥ 0.5 and measure each object's footprint after global resizing. No gradients, parameter updates, or learnable modules are introduced. Full metric definitions are provided in Appendix A.2.

Instance-level response distribution. Figure 4(a) compares the empirical cumulative distributions of matched confidence for COCO small objects. The tiled-view distribution shifts toward higher confidence relative to the global view. Across 1,667 small instances, the mean matched response increases by 0.0669, with a 95% image-clustered bootstrap confidence interval of $[0.0567, 0.0765]$. The matched detection rate increases from 84.28% under global inference to 91.90% under tiled inference, corresponding to an improvement of 7.62 percentage points (95% CI: $[5.70, 9.44]$). These results show that local scale remapping strengthens the output response to small objects without modifying the detector parameters.

Evidence bottleneck under global resizing. Figure 4(c) stratifies ground-truth instances according to their footprints after global resizing. For objects occupying less than $64\ \mathrm{px}^2$ at the detector input, the matched detection rate increases from 50.4% under the global view to 71.9% under tiled views. For objects occupying 64–256 px^2, the corresponding rate increases from 77.9% to 89.5%. The advantage of tiled inference becomes smaller when the input footprint exceeds $256\ \mathrm{px}^2$. This pattern directly connects the input-side evidence bottleneck proposed in the introduction to the observed detector outputs: local scale remapping is most beneficial when global resizing leaves insufficient spatial support for an object.

Scale-dependent view reliability. Figure 4(d) shows that tiled views increase the mean response for small objects by 0.0669, have little effect on medium objects (0.0043), and reduce it for large objects by 0.2099. This scale-dependent reliability supports using the global branch as the prediction anchor and restricting tiled candidates through area gating.

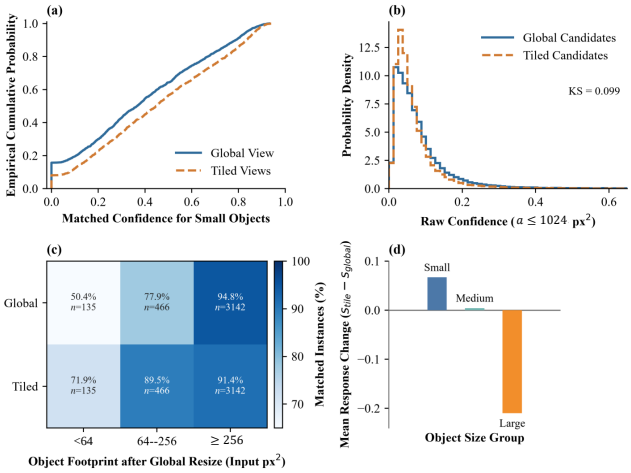

Figure 4: Scale-dependent view reliability: (a) matched-confidence distributions for small objects; (b) cross-view confidence distributions; (c) detection rates by resized object footprint; and (d) response changes across object scales.

Cross-view confidence shift. Figure 4(b) shows different confidence distributions for global and tiled candidates under the same 1024 px^2 area constraint, with a descriptive Kolmogorov–Smirnov distance of 0.099. Because the two views are correlated, this statistic is used only as a distributional measure. The observed shift indicates that their scores are not directly comparable, motivating the detector-specific score calibration in GateSlice.

Feature-level activation visualization. To examine how scale remapping affects internal representations, we visualize gradient-free channel-energy maps from the highest-resolution P3 feature of frozen RT-DETR, with implementation details provided in Appendix A.1. As shown in Fig. 5, tiled views make the target responses more distinguishable from their surrounding context. Across the three examples, the target-to-context contrast ratio increases from 1.277 to 1.387, 1.571 to 1.888, and 1.470 to 1.662, respectively, accompanied by improved matching IoUs in the fused output. These results suggest that local enlargement reorganizes spatial feature responses and improves target–context separation, rather than uniformly increasing feature magnitude.

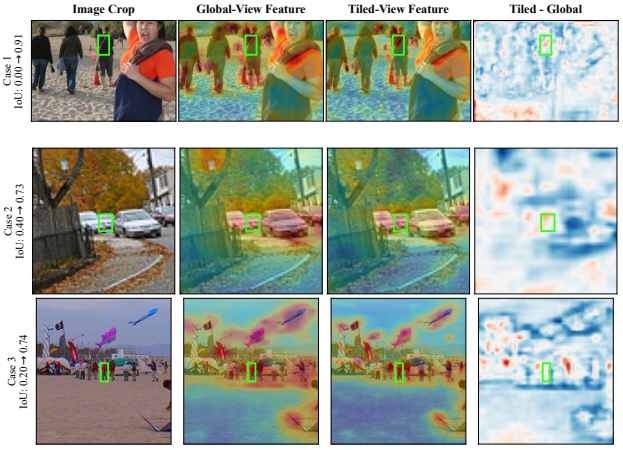

Figure 5: Feature activations for small objects rescued by GateSlice. Columns show the image crop, global-view activation, tiled-view activation, and their difference. Tiled inference improves target–context feature contrast.

4.5 Parameter Settings and Robustness.

Area-gating threshold. As shown in Figure 6 shows the trade-off between excluding beneficial small-object candidates and admitting noisy larger predictions. We use 1024 px^2, the COCO small-object upper bound, without dataset-specific tuning. The no-limit variant removes only area filtering, providing a single-variable ablation of the gate.

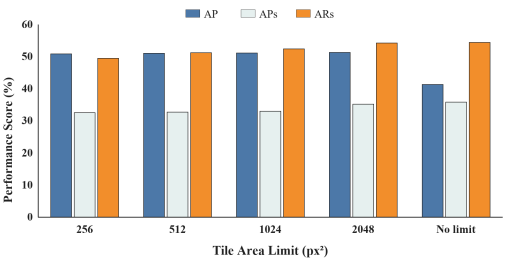

Figure 6: Area-threshold sensitivity. Removing the area gate improves small-object recall but substantially reduces overall AP.

Parameter sensitivity. As shown in Fig. 7, performance is generally stable with respect to score calibration, confidence threshold, and overlap ratio, whereas the NMS threshold and

candidate budget have a greater effect on AR_s. Increasing either retains more small-object candidates but may weaken duplicate suppression. We therefore use a sliced-score factor of 1.0, a confidence threshold of 0, an NMS threshold of 0.6, a candidate limit of 300, and an overlap ratio of 0.2 consistently across datasets.

4.6 Post-fusion instance attribution.

Figure 8 compares the global baseline with the final GateSlice output at IoU ≥ 0.5. GateSlice rescues 107 instances while losing 42 baseline detections, yielding a net gain of 65; 98 of the rescued instances are small objects, compared with 32 lost small-object detections, and 97.96% of these rescues originate from the tiled branch. For the 1,263 small objects detected by both methods, the mean matched IoU increases from 0.763 to 0.783 and the mean confidence from 0.492 to 0.557. These results show that constrained fusion primarily converts tiled-view evidence into newly recovered and better-localized small-object detections while largely preserving the global predictions.

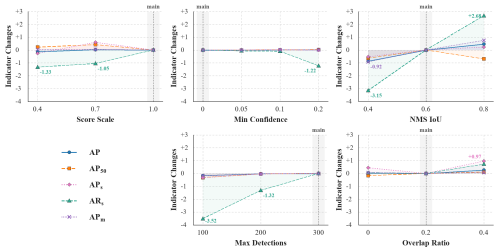

Figure 7: Parameter sensitivity relative to the default configuration. Small-object recall is most affected by the NMS threshold and candidate budget.

4.7 Effective-resolution control.

Figure 9 compares crop-before-resize with downsample-before-crop while keeping all other settings fixed. The two variants achieve similar AP_s (33.028 vs. 32.977) and AR_s (52.442 vs. 51.980), indicating that the gains primarily arise from mapping small objects to a more favorable scale rather than recovering additional image detail.

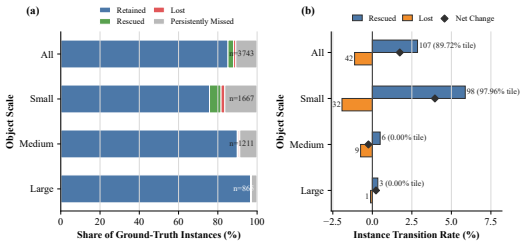

Figure 8: Instance-level transitions at IoU ≥ 0.5: (a) detection states by object scale and (b) rescued, lost, and net rates. Percentages denote tiled-branch contributions.



Figure 9: Effective-resolution control. Similar gains from the two sliced variants indicate that scale remapping is the primary source of improvement.

5 Conclusion

We presented GateSlice, a training-free framework that improves frozen detectors by combining context-preserving global inference with scale-gated tiled predictions. Area gating, score calibration, and class-aware fusion recover small-object evidence while limiting interference from unreliable local predictions. Consistent improvements across detectors and datasets demonstrate the generality of this approach. Future work will investigate adaptive tiling and more efficient confidence alignment and fusion.

AI Use Statement

Large language models (LLMs) were used solely to improve the writing of this paper, including grammar, clarity, and readability. They were not used to generate research ideas, design the methodology or experiments, conduct analyses, interpret results, or produce scientific claims. All AI-assisted revisions were reviewed by the authors, who take full responsibility for the final content of this work.

REFERENCES

Fatih Cagatay Akyon, Sinan Onur Altinuc, and Alptekin Temizel. Slicing aided hyper inference and fine-tuning for small object detection. In *Proceedings of the IEEE International Conference on Image Processing*, pp. 966–970, 2022. doi: 10.1109/ICIP46576.2022.9897990.

Chengzhen Duan, Zhiwei Wei, Chi Zhang, Siying Qu, and Hongpeng Wang. Coarse-grained density map guided object detection in aerial images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 2789–2798, 2021.

Golnaz Ghiasi, Tsung-Yi Lin, and Quoc V. Le. NAS-FPN: Learning scalable feature pyramid architecture for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7036–7045, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Ghiasi_NAS-FPN_Learning_Scalable_Feature_Pyramid_Architecture_for_Object_Detection_CVPR_2019_paper.html.

Chaoxu Guo, Bin Fan, Qian Zhang, Shiming Xiang, and Chunhong Pan. AugFPN: Improving multi-scale feature learning for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12595–12604, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Guo_AugFPN_Improving_Multi-Scale_Feature_Learning_for_Object_Detection_CVPR_2020_paper.html.

Glenn Jocher and Jing Qiu. Ultralytics yolo11. https://github.com/ultralytics/ultralytics, 2024. Version 11.0.0, accessed August 13, 2026.

Muhammed Can Keles, Batuhan Salmanoglu, Mehmet Serdar Guzel, Baran Gursoy, and Gazi Erkan Bostanci. Evaluation of YOLO models with sliced inference for small object detection. *arXiv preprint arXiv:2203.04799*, 2022.

Seung-Wook Kim, Hyong-Keun Kook, Jee-Young Sun, Mun-Cheon Kang, and Sung-Jea Ko. Parallel feature pyramid network for object detection. In *Proceedings of the European Conference on Computer Vision*, pp. 234–250, 2018.

Onur Can Koyun, Reyhan Kevser Keser, İbrahim Batuhan Akkaya, and Behçet Uğur Töreyin. Focus-and-detect: A small object detection framework for aerial images. *Signal Processing: Image Communication*, 104:116675, 2022.

Changlin Li, Taojiannan Yang, Sijie Zhu, Chen Chen, and Shanyue Guan. Density map guided object detection in aerial images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 190–191, 2020.

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 *Computer Vision – ECCV 2014*, volume 8693 of *Lecture Notes in Computer Science*, pp. 740–755. Springer, 2014. doi: 10.1007/978-3-319-10602-1_48.

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2117–2125, 2017.

Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8759–8768, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Liu_Path_Aggregation_Network_CVPR_2018_paper.html.

Akhil Meethal, Eric Granger, and Marco Pedersoli. Cascaded zoom-in detector for high resolution aerial images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 2046–2055, 2023.

Mahyar Najibi, Bharat Singh, and Larry S. Davis. AutoFocus: Efficient multi-scale inference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9745–9755, 2019.

George Plastiras, Christos Kyrkou, and Theocharis Theocharides. Efficient ConvNet-based object detection for unmanned aerial vehicles by selective tile processing. In *Proceedings of the 12th International Conference on Distributed Smart Cameras*, pp. 1–6, 2018. doi: 10.1145/3243394.3243692.

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*, volume 28, pp. 91–99, 2015.

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. URL https://openaccess.thecvf.com/content/ICCV2025/html/Sharma_DM-EFS_Dynamically_Multiplexed_Expanded_Features_Set_Form_for_Robust_and_ICCV_2025_paper.html.

Yanyun Shen, Di Liu, Feizhao Zhang, and Qingling Zhang. Fast and accurate multi-class geospatial object detection with large-size remote sensing imagery using CNN and truncated NMS. *ISPRS Journal of Photogrammetry and Remote Sensing*, 191:235–249, 2022. doi: 10.1016/j.isprsjprs.2022.07.019.

Bharat Singh and Larry S. Davis. An analysis of scale invariance in object detection – SNIP. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3578–3587, 2018.

Bharat Singh, Mahyar Najibi, and Larry S. Davis. Sniper: efficient multi-scale training. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, pp. 9333–9343, Red Hook, NY, USA, 2018. Curran Associates Inc.

Mingxing Tan, Ruoming Pang, and Quoc V. Le. EfficientDet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10778–10787, Seattle, WA, USA, 2020. IEEE. doi: 10.1109/CVPR42600.2020.01079.

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*, pp. 9627–9636, 2019. doi: 10.1109/ICCV.2019.00972.

Leon Amadeus Varga and Andreas Zell. Tackling the background bias in sparse object detection via cropped windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 2768–2777, 2021.

Chang Xu, Jinwang Wang, Wen Yang, Huai Yu, Lei Yu, and Gui-Song Xia. Detecting tiny objects in aerial images: A normalized wasserstein distance and a new benchmark. *ISPRS Journal of Photogrammetry and Remote Sensing*, 190:79–93, August 2022a. doi: 10.1016/j.isprsjprs.2022.06.002.

Chang Xu, Jinwang Wang, Wen Yang, Huai Yu, Lei Yu, and Gui-Song Xia. RFLA: Gaussian receptive field based label assignment for tiny object detection. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision – ECCV 2022*, volume 13669 of *Lecture Notes in Computer Science*, pp. 526–543, Cham, 2022b. Springer Nature Switzerland. doi: 10.1007/978-3-031-20077-9_31.

Chenhongyi Yang, Zehao Huang, and Naiyan Wang. QueryDet: Cascaded sparse query for accelerating high-resolution small object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13668–13677, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Yang_QueryDet_Cascaded_Sparse_Query_for_Accelerating_High-Resolution_Small_Object_Detection_CVPR_2022_paper.html.

Fan Yang, Heng Fan, Peng Chu, Erik Blasch, and Haibin Ling. Clustered object detection in aerial images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8311–8320, 2019.

Xin Yang, Yong Song, Ya Zhou, Yizhao Liao, Jinqi Yang, Jinxiang Huang, Yiqian Huang, and Yashuo Bai. An efficient detection framework for aerial imagery based on uniform slicing window. *Remote Sensing*, 15(17):4122, 2023. doi: 10.3390/rs15174122.

Xuehui Yu, Yuqi Gong, Nan Jiang, Qixiang Ye, and Zhenjun Han. Scale match for tiny person detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1257–1265, 2020. URL https://openaccess.thecvf.com/content_WACV_2020/html/Yu_Scale_Match_for_Tiny_Person_Detection_WACV_2020_paper.html.

Xiang Yuan, Gong Cheng, Kebing Yan, Qinghua Zeng, and Junwei Han. Small object detection via coarse-to-fine proposal generation and imitation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6317–6327, 2023.

Tianyi Zhang, Kishore Kasichainula, Yaoxin Zhuo, Baoxin Li, Jae-Sun Seo, and Yu Cao. Patch-based selection and refinement for early object detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 729–738, 2024.

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*, pp. 16965–16974, 2024.

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. doi: 10.1109/TPAMI.2021.3119563.

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.

A Appendix

A.1 Feature-Activation Visualization Details

We extract the highest-resolution P3 feature tensor

$$F_v \in \mathbb{R}^{C \times H_f \times W_f} \tag{5}$$

from the frozen RT-DETR encoder for each view v. The activation map is computed without gradients as the channel-wise root-mean-square energy:

$$A_v(x, y) = \sqrt{\frac{1}{C} \sum_{c=1}^{C} F_{v,c}(x, y)^2}. \tag{6}$$

For tiled inference, we select the tile containing the center of the target instance and map its activation to the original image coordinates. The global and tiled maps for each example share the same robust normalization bound to ensure a comparable color scale. We measure the target-to-context contrast as the ratio between the mean activation inside the target box and that in its surrounding region. This procedure requires neither backpropagation nor model modification and therefore preserves the training-free setting.

A.2 Instance-Level Response Definition

For each non-crowd ground-truth instance g and inference view $v \in \{\text{global}, \text{tile}\}$, the matched detection response is defined as

$$r_v(g) = \max_{d_i \in \mathcal{D}_v} \left\{ s_i \mid c_i = c_g,\ \text{IoU}(\boldsymbol{b}_i, \boldsymbol{b}_g) \geq 0.5 \right\}, \tag{7}$$

where s_i, c_i, and $\boldsymbol{b}_i$ denote the confidence score, category, and bounding box of prediction d_i, respectively. We set $r_v(g) = 0$ when no valid match exists. Predictions are analyzed before cross-view fusion to isolate the effect of the input view.

The footprint of instance g after global resizing is defined as

$$a_{\mathrm{in}}(g) = a_{\mathrm{orig}}(g)\frac{H_0 W_0}{HW}, \tag{8}$$

where $H \times W$ and $H_0 \times W_0$ are the original and detector-input resolutions, respectively. This quantity measures the spatial support retained after global resizing.