

Zero-Harm Feature Calibration and Adaptive Boundary Query for Tiny Object Detection

Yingying Zhai

Northeastern University

Chang He

Northeastern University

Zhi Wang

Northeastern University

Zezheng Feng

fengzezheng@swc.neu.edu.cn

Northeastern University

Bin Wang

Northeastern University

Xiaochun Yang

Northeastern University

Research Article

Keywords: Tiny object detection, End-to-end object detection, Feature calibration, Dynamic query allocation, Machine vision

Posted Date: September 4th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10475931/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Zero-Harm Feature Calibration and Adaptive Boundary Query for Tiny Object Detection

Yingying Zhai¹, Chang He^{1,2}, Zhi Wang¹, Zezheng Feng^{1*}, Bin Wang¹,
Xiaochun Yang¹

^{1*}Northeastern University, Shenyang, Liaoning, China.

²Ningxia Institute of Science and Technology, Shizuishan, Ningxia, China.

*Corresponding author(s). E-mail(s): fengzezheng@swc.neu.edu.cn;

Abstract

Transformer-based detectors for tiny object detection suffer from a *coupled failure mode* rather than two independent shortcomings: *representation collapse*, in which Sigmoid-based attention can suppress neutral foreground activations and erode faint tiny-object signals; and *assignment misallocation*, in which a static or hard-thresholded query budget fails to track continuous, long-tailed density variation across images. These failures can reinforce each other: degraded features may bias the density estimate that drives query allocation, while a mismatched query budget can starve regions where feature evidence is already weak. We therefore reformulate tiny object detection as *joint optimization* over a feature-calibration branch and a density-allocation branch sharing a common representation. The feature branch, Zero-Harm Feature Calibration (ZFC), replaces Sigmoid gating with a Tanh-centered mechanism designed to preserve neutral activations and mitigate representation collapse. The assignment branch, Statistical Boundary Allocator (SBA), regresses continuous per-image log-space density boundaries under a cumulative distribution function (CDF)-based True Adaptive Boundary Loss, yielding differentiable soft query allocation for the main NMS-free inference setting. The two branches are explicitly coupled: SBA’s density prior guides ZFC’s spatial recalibration, and ZFC’s calibrated features in turn refine SBA’s density estimate. On AI-TOD-V2, the joint framework achieves 32.0% average precision (AP; +1.8 percentage points over DQ-DETR) with a 49% reduction in giga floating point operations (GFLOPs; 1805.4 $\rightarrow$ 921.0) at identical parameter count; gains are consistent on VisDrone-DET2019 (38.2% AP, +1.2 percentage points).

Keywords: Tiny object detection, End-to-end object detection, Feature calibration, Dynamic query allocation, Machine vision

1 Introduction

Accurate object detection in aerial and surveillance imagery is increasingly critical for real-world systems such as UAV-based monitoring, remote-sensing scene understanding, and automated industrial inspection. These applications share a common constraint: they must reliably

detect targets under resource limitations, across varying scales, and in cluttered environments. To meet this demand, the computer vision community has made substantial strides through deep learning-based detection frameworks. Among these, Transformer-based end-to-end detectors—most notably DETR [1]—have emerged as a

principled alternative to conventional anchor-based pipelines, eliminating hand-crafted post-processing components such as non-maximum suppression (NMS) and recasting detection as a direct set prediction problem via bipartite matching. Follow-up works including Deformable DETR [2], DAB-DETR [3], and DINO [4] have further improved convergence speed, multi-scale representation, and detection accuracy, establishing DETR-family models as strong baselines across a broad spectrum of detection tasks.

Despite this progress, a practically important and technically demanding sub-problem remains underexplored: the detection of *tiny objects*—instances occupying only a few pixels in high-resolution imagery. As documented in the AI-TOD-V2 benchmark [5], the average object footprint is just 12.7 pixels, with 86% of instances falling below 16 pixels and per-image counts ranging from 1 to over 2,600. Such characteristics are far from the distribution that standard DETR variants are implicitly optimized for. Tiny objects appear pervasively in safety-critical downstream tasks—detecting drones in restricted airspace, localizing vehicles from satellite imagery, or identifying micro-defects in manufactured parts—where missing even a small fraction of instances can have serious operational consequences.

We argue that the performance gap on tiny objects does not stem from two independent shortcomings of standard DETR variants, but from a *single coupled failure mode* that manifests along two axes of the same underlying problem: how faithfully an image is represented, and how the resulting representation is used to decide where to allocate detection capacity. We term these two manifestations *representation collapse* and *assignment misallocation*, and contend that treating them in isolation—calibrating features without regard to allocation, or allocating queries without regard to feature quality—necessarily leaves performance on the table, because each failure mode actively reinforces the other (see Fig. 1).

Representation collapse arises because localizing tiny objects requires high-resolution shallow features, which are simultaneously the features most susceptible to background interference. The signal-to-noise ratio of a 6×6 -pixel object is orders of magnitude lower than that of a 60×60 -pixel object, leaving its feature activations fragile and easily eroded. Conventional

attention mechanisms—typified by SENet [6] and CBAM [7]—rely on Sigmoid gating, which maps zero-valued activations to exactly 0.5, attenuating neutral feature energy by half. For large objects with robust activations this is inconsequential, but for tiny objects whose signals are inherently faint and frequently fall in this neutral regime, Sigmoid gating degrades residual boundary and texture cues, elevating the false-positive rate. Conversely, replacing Sigmoid with a naive residual activation mitigates suppression but amplifies background noise, causing excessive false alarms. Designing *neutral-preserving calibration* that preserves fragile foreground signals while suppressing background noise therefore remains an important open challenge.

Assignment misallocation arises because the number of tiny instances per image varies wildly and follows a severe long-tailed distribution, yet standard DETR architectures rely on a static query budget (e.g., 300 or 900 queries). This rigid design introduces a severe scalability bottleneck: it incurs redundant computational overhead in sparse scenes—a critical issue for resource-constrained deployment—while bounding recall in densely packed scenarios. DQ-DETR [8] partially addresses this by inferring spatial density from visual features to modulate query count dynamically. However, it discretizes density into categorical levels using hard thresholds derived from global dataset statistics (i.e., mean and variance), implicitly assuming a unimodal Gaussian count distribution. This assumption is frequently violated in real-world aerial datasets like AI-TOD-v2 [5], where instance counts exhibit severe long-tailed distributions. Crucially, DQ-DETR’s density estimate is itself read out from the same Sigmoid-gated features subject to representation collapse, so the two failure modes are not merely co-occurring but mechanistically coupled in practice: eroded features can yield a biased density estimate, and a misallocated query budget can in turn starve the sparse, low-confidence regions where feature evidence is already weakest, compounding the original signal loss. A principled treatment therefore requires addressing feature degradation and query allocation jointly rather than as separable engineering problems.

The connection between the two failure modes therefore runs through the shared feature stream: the features affected by Sigmoid attention are also

the features from which any density estimator—including a continuous one—must read out its allocation signal. Fixing only the representation leaves allocation to operate on a cleaner but still separately optimized signal; fixing only the allocation still reads from Sigmoid-suppressed features. Joint treatment is therefore a more coherent way to reduce both sides of the feedback loop simultaneously.

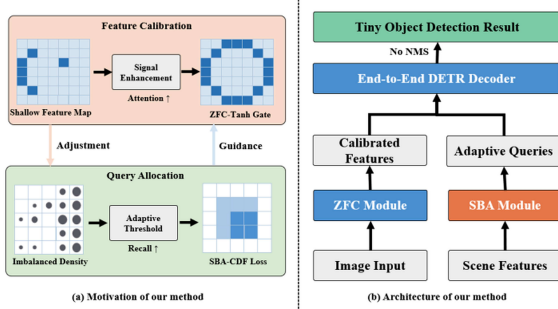


Fig. 1 Motivation and architecture of the proposed method. (a) At the feature level, conventional Sigmoid-based attention suppresses neutral channel responses toward 0.5, eroding faint target signals; the proposed ZFC module employs a Tanh-centered gate for neutral-preserving signal calibration. At the query-allocation level, rigid hard-threshold segmentation misfires on long-tail density distributions; the SBA module replaces hard thresholds with CDF-guided soft boundaries to improve recall in dense scenes. (b) The full framework feeds image inputs through the ZFC module to obtain calibrated multi-scale features and through the SBA module to generate density-adaptive object queries, which are jointly decoded by an end-to-end DETR decoder in the main NMS-free inference setting.

Motivated by this coupling, we reformulate tiny object detection as joint optimization over two branches that operate on a shared representation rather than as independent modules bolted onto a baseline detector (Fig. 1). The feature branch, *Zero-Harm Feature Calibration* (ZFC), replaces the Sigmoid gate with a Tanh-centered formulation: since $\tanh(0) = 0$, neutral activations receive a multiplicative weight of exactly $1 + \tanh(0) = 1$, so neutral inputs are not attenuated by the channel gate under this formulation while background responses are softly suppressed. The assignment branch, *Statistical Boundary Allocator* (SBA), regresses continuous logarithmic-space density boundaries via a True Adaptive Boundary Loss based on cumulative distribution functions,

enabling differentiable soft query allocation that dynamically adapts to per-image density in the main non-TTA inference path without NMS post-processing. The two branches are coupled rather than independent: SBA’s density representation supplies the spatial prior that guides where ZFC applies its level-adaptive residual correction, while ZFC’s calibrated features in turn sharpen the density signal that SBA allocates queries against. Together, ZFC and SBA form a single adaptive system that is both more accurate and substantially more computationally efficient than prior methods, yielding a 49% reduction in GFLOPs with no increase in parameter count, and the joint configuration empirically exceeds the sum of each branch’s individual gain (Section 4.6), supporting that the coupling is functional rather than merely architectural.

The main contributions of this paper are as follows:

- We identify representation collapse and density-variant assignment misallocation as a coupled failure mode in tiny object detection—rather than two independent deficiencies to be patched separately—and reformulate detection as joint optimization over a feature-calibration branch and a density-allocation branch sharing a common representation.
- We instantiate the feature branch as Zero-Harm Feature Calibration (ZFC), a Tanh-centered gating mechanism that addresses the dual failure modes of conventional Sigmoid attention—neutral feature suppression and background amplification—while preserving neutral weak foreground responses for tiny object detection.
- We instantiate the assignment branch as the Statistical Boundary Allocator (SBA), governed by a CDF-based True Adaptive Boundary Loss, which replaces rigid statistical thresholds with fully differentiable soft query allocation and supports density-adaptive query assignment in the main NMS-free detector.

2 Related Work

Rather than surveying tiny object detection by method family, we organize this section around the two axes of the coupled failure mode introduced in Section 1—feature-side degradation and assignment-side misallocation—and conclude by

showing that the DETR architectures connecting them have so far addressed each axis independently, leaving the gap that this paper targets.

2.1 Feature-Side Gap: Representation Solved Partially, Not Jointly with Allocation

Tiny object detection is a critical challenge in aerial surveillance, UAV monitoring, and industrial inspection [5, 9], driven by the fragile feature representations of objects spanning only a few pixels. Wang et al. established the AI-TOD benchmark [5], which highlights an average object footprint of 12.7 pixels and severe long-tail instance-count distributions. To address metric instability under extreme scale, Xu et al. proposed the Normalized Wasserstein Distance [10] and Dot Distance [11] as scale-aware localization measures, and later introduced RFLA [12] to resolve feature-pyramid mismatch via Gaussian receptive-field label assignment. At the representation level, SR-TOD [13] amplifies residual foreground signals through image-difference maps; TinyDef-DETR [9] integrates edge enhancement with dual-domain attention; and DETR-ORD [14] combines feature reconstruction with adaptive sample selection for oriented detection. Channel and spatial attention mechanisms underpin most of these feature calibration pipelines: SENet [6] recalibrates channel-wise responses via Squeeze-and-Excitation blocks, CBAM [7] extends this to spatial dimensions, and FPN [15] provides the canonical multi-scale fusion backbone.

What this line of work leaves open is not the existence of feature calibration mechanisms, but their coupling to allocation: every method above improves the representation *in isolation*, with no mechanism by which improved features inform how detection capacity (e.g., query budget) is subsequently distributed across an image. Moreover, as analyzed in Section 1, the Sigmoid gating shared by SENet and CBAM-style designs imposes a structural ceiling specific to tiny objects: it maps neutral activations to exactly 0.5, eroding the very residual cues that representation-side methods above aim to preserve. This is the gap our feature branch, ZFC, targets directly—replacing the Sigmoid gate with a Tanh-centered formulation that preserves neutral channel responses under the proposed gate

$(1 + \tanh(0) = 1)$ —while, unlike prior calibration modules, additionally consuming a density signal from the allocation side (Section 3) rather than operating as a self-contained block.

2.2 Assignment-Side Gap: Heuristic and Discrete, Decoupled from Feature Quality

The fixed-query design of vanilla DETR [1] and its variants is fundamentally mismatched with tiny object detection, where target counts vary by orders of magnitude. DDQ [16] enhanced recall in crowded scenes with dense distinct queries, while DQ-DETR [8] explicitly modulates query count via density-inferred visual features. DQ-DETR is the closest prior work to ours, but it relies on hard thresholds derived from dataset-level mean and variance, implicitly assuming a unimodal Gaussian distribution—a prior that breaks down under the long-tailed distributions of AI-TOD-v2 [5], and it discretizes density into categorical levels, introducing discontinuous allocation jumps at threshold boundaries.

The deeper gap, however, is not merely that the threshold is hard rather than soft: DQ-DETR’s density estimate is read out from features subject to the same Sigmoid-gated suppression discussed above, so allocation quality inherits feature-side noise without any explicit correction. This is the gap our assignment branch, SBA, targets: it replaces hard statistical thresholds with continuous, per-image log-space boundaries optimized under a CDF-based True Adaptive Boundary Loss, and—critically—is designed to consume the calibrated density representation shared with ZFC rather than an uncorrected feature read-out.

2.3 DETR Architectures: Necessary Backbone, Insufficient Unification

Carion et al. introduced DETR [1], recasting detection as direct set prediction via bipartite matching and eliminating NMS and handcrafted anchors. To address its slow convergence and quadratic attention cost on high-resolution maps, Zhu et al. proposed Deformable DETR [2], replacing dense self-attention with sparse deformable attention across multiple scales and reducing complexity from $O(H^2W^2)$ to $O(HWK)$ —this serves

as our architectural backbone. Subsequent variants refined the paradigm: SMCA [17] accelerates convergence through spatially modulated co-attention; Dynamic DETR [18] introduces dynamic attention; DAB-DETR [3] uses dynamic anchor boxes as queries; Conditional DETR [19] decouples spatial and content queries; DINO [4] advances convergence via contrastive denoising; RT-DETR [20] enables real-time inference; and Sparse R-CNN [21] extends set prediction beyond Transformer architectures.

None of these architectures, including DQ-DETR, treats feature quality and query density as coupled quantities to be optimized jointly: each either fixes the query budget irrespective of feature content, or, where density-awareness exists, derives it from features that have not been corrected for the Sigmoid suppression pathology endemic to the field. We integrate ZFC into the deformable encoder stage and couple it with SBA through a shared density representation (Section 3), so that feature calibration and query allocation inform each other rather than operating as two independently bolted-on components.

3 Method

3.1 Problem Reformulation

We formalize the coupled failure mode motivated in Section 1 so that ZFC and SBA can be understood as two coordinated solutions to one problem. Given an input image $I \in \mathbb{R}^{3 \times H_0 \times W_0}$ and backbone-extracted multi-scale features $\mathcal{F} = \{F_1, F_2, \dots, F_L\}$, conventional detectors treat feature calibration and query allocation as largely independent functions of the same raw representation. This independence is precisely what allows the two errors to reinforce each other: feature erosion biases the density estimate used for allocation, while a mismatched query budget leaves weak but recoverable regions under-served by the decoder. We therefore introduce an explicit shared density representation D that is consumed by both the assignment branch and the feature-calibration branch. Sections 3.3–3.4 instantiate this design as SBA and ZFC, and Section 3.5 details how the two branches exchange information through D .

3.2 Unified Framework Overview

Fig. 2 illustrates the resulting architecture. ZFC and SBA are two coupled components of one adaptive pipeline sharing the density representation D defined above. SBA computes D from front-end dense response feature maps, uses it to regress the query budget N , and exposes D as a spatial prior. ZFC consumes the same D to decide *where* in the feature pyramid to apply spatial recalibration, in addition to its channel-wise Tanh-centered gating. Concretely:

- **Statistical Boundary Allocator (Section 3.3):** From front-end dense response feature maps, SBA regresses continuous logarithmic-space statistical boundaries for dynamic soft query assignment and produces the spatial density prior map D consumed by ZFC.
- **Zero-Harm Feature Calibration (Section 3.4):** The multi-scale feature set $\mathcal{F}$ passes through ZFC before the Transformer encoder. Using the density prior D from SBA, ZFC applies Tanh-centered channel gating and hierarchical adaptive spatial residuals to perform neutral-preserving calibration of weak foreground features.
- **Joint decoding:** Calibrated features and SBA-generated queries are fed into a deformable Transformer decoder for bounding box regression and classification, with no NMS required.

Section 3.5 returns to this coupling explicitly after both branches have been defined.

3.3 Statistical Boundary Allocator: Assignment as Continuous Density Optimization

SBA treats query assignment as a continuous density optimization problem: the allocation boundary is a learned function of per-image features, not a fixed scalar derived from dataset statistics. In micro-target detection, object density can vary from a few to several thousand instances per image, yet existing dynamic query methods such as DQ-DETR rely on hard thresholds assuming unimodal Gaussian count distributions—a prior that fails under the long-tailed statistics characteristic of AI-TOD-v2, causing severe query misallocation, and one whose

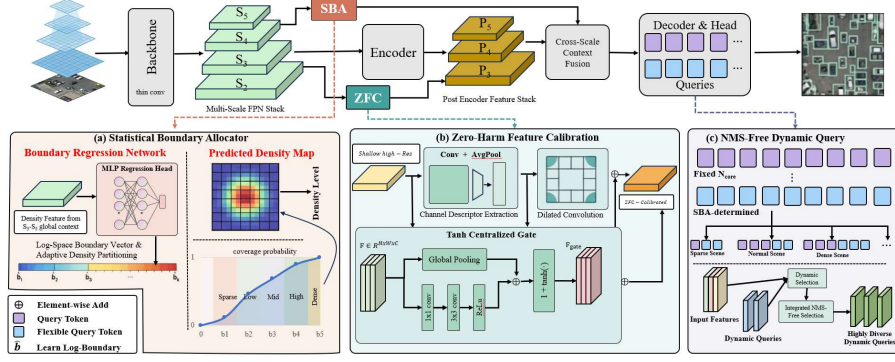


Fig. 2 Overall architecture of the proposed end-to-end framework. The Statistical Boundary Allocator (SBA) performs dynamic query assignment via continuous log-boundary regression, while the Zero-Harm Feature Calibration (ZFC) module supports neutral-preserving foreground calibration, jointly facilitating robust NMS-free main inference for tiny object detection.

density estimate is computed without correction for the upstream Sigmoid-suppression pathology discussed in Section 1. The Statistical Boundary Allocator (SBA) instantiates the assignment branch g_{alloc} and the shared density representation h from Section 3.1: it treats density boundaries not as fixed dataset-level scalars but as continuous functions of per-image features, optimized end-to-end alongside detection, and it exposes its internal density representation D as the prior consumed by ZFC (Section 3.5).

Logarithmic space regression of dynamic boundaries. Instead of relying on preset statistical scalars, SBA treats density boundaries as continuous variables inferred dynamically from image features $F_{\text{den}} \in \mathbb{R}^{C \times H \times W}$. A boundary prediction head composed of adaptive global average pooling followed by a multi-layer perceptron directly regresses three consecutive logarithmic boundaries $B_{\log} = [\log b_1, \log b_2, \log b_3]$ that partition the object count range into four density levels. To enforce strict monotonicity among the hierarchical intervals, we adopt a cascaded residual prediction scheme: the first boundary is clamped to a valid range, and each subsequent boundary is obtained by adding a strictly positive offset via Softplus:

$$\log b_1 = \text{Clamp}(H_{\text{base}}(F_{\text{den}})). \quad (1)$$

$$\log b_i = \log b_{i-1} + \text{Softplus}\left(H_{\text{base}}(F_{\text{den}}) \cdot \gamma_{\text{warmup}} + \theta_{\text{gap}}\right), \quad i \in \{2, 3\}. \quad (2)$$

where H_{base} denotes a learnable network branch, θ_{gap} is a safety margin (set to 0.3) that enforces a minimum separation between levels, and γ_{warmup} is a cosine-annealing warmup factor that prevents boundary collapse caused by unconverted features early in training.

Soft Query Assignment. SBA formulates query allocation as a smooth, differentiable expectation over a probability distribution, avoiding the discontinuous jumps of hard step functions at fixed boundaries. The network additionally regresses a coarse global count N_{pred} . We compute the absolute difference between $\log N_{\text{pred}}$ and the center C_k of each interval in log space, and apply a temperature-scaled Softmax with $\tau = 1.0$ to obtain soft weights $W \in \mathbb{R}^4$. Specifically, the four interval centers are defined as $C_0 = \log b_1 - 0.5$, $C_1 = (\log b_1 + \log b_2)/2$, $C_2 = (\log b_2 + \log b_3)/2$, and $C_3 = \log b_3 + 0.5$, where b_1, b_2, b_3 are the boundary values that partition the count range.

$$W_k = \frac{\exp(-|\log N_{\text{pred}} - C_k|/\tau)}{\sum_{i=0}^3 \exp(-|\log N_{\text{pred}} - C_i|/\tau)} \quad (3)$$

Intuitively, the negative exponential assigns the highest weight to the query level whose center C_k is nearest to $\log N_{\text{pred}}$, implementing a differentiable nearest-center assignment. The final number of queries supplied to the decoder, N_{eval} , is the weighted expectation over the preset candidate

query levels $Q = [K_0, K_1, K_2, K_3]$:

$$N_{\text{eval}} = \sum_{k=0}^3 W_k \cdot Q_k. \quad (4)$$

This soft allocation mechanism not only preserves end-to-end differentiability but also substantially smooths allocation fluctuations caused by count jitter, which frequently occurs when tiny objects are partially occluded.

True Adaptive Boundary Loss. To drive the boundaries to adapt to the true object statistics, we introduce the True Adaptive Boundary Loss L_{TABL} , which simultaneously constrains the boundary positions and imposes coverage-aware penalties from a probabilistic perspective. Given the ground-truth object count N_{gt} and its logarithm $\log C = \log N_{\text{gt}}$, the cumulative distribution function response of each predicted boundary $\log b_i$ is:

$$\text{CDF}_i = \sigma\left(\frac{\log b_i - \log C}{\tau}\right), \quad i = \{1, 2, 3\} \quad (5)$$

where $\sigma(\cdot)$ denotes the sigmoid function. This maps the signed log-space distance between a boundary and the ground-truth count to a smooth CDF value, making the boundary target explicit while avoiding unstable gradients from direct distribution matching. The full loss combines coverage, spacing, and count terms:

$$L_{\text{TABL}} = \lambda_1 L_{\text{coverage}} + \lambda_2 L_{\text{spacing}} + \lambda_3 L_{\text{count}} \quad (6)$$

The coverage term penalizes deviations from the target CDF coverage matrix with clipped per-sample weights to suppress outliers. The spacing term enforces a minimum log-space separation between adjacent boundaries and a physical lower bound on the first boundary, preventing collapse under long-tailed imbalance. The count term is a Smooth L1 penalty between the predicted log-count and $\log N_{\text{gt}}$.

3.4 Zero-Harm Feature Calibration: Stabilizing Weak Foreground Signals

We first address representation collapse by stabilizing weak foreground signals before they reach the Transformer encoder. The encoder enhances

semantic interactions among multi-scale features, but for tiny objects whose local responses are easily overwhelmed by larger-region signals, conventional Sigmoid-based attention introduces a critical flaw: the gating value $\sigma(0) = 0.5$ halves neutral channel energy, destroying weak boundary and texture cues essential for sub-pixel localization. Zero-Harm Feature Calibration (ZFC) instantiates the feature branch g_{feat} from Section 3.1, and—unlike a self-contained attention module—additionally consumes the density representation D produced by SBA (Section 3.3) to decide where spatial recalibration should be applied, as detailed below and revisited in Section 3.5.

Tanh-Centered Channel Gate. To overcome this, we propose a neutral-preserving gating mechanism centered on the Tanh function:

$$\vartheta_{\text{chan}}(F) = 1.0 + \tanh(S_f \cdot M_{\text{mlp}}(P(F))), \quad (7)$$

$$F'_{\text{out}} = F_{\text{in}} \odot \vartheta_{\text{chan}}(F). \quad (8)$$

Here, $P(\cdot)$ denotes a mixed pooling operation, M_{mlp} is a channel-mapping network, and $S_f \in [0, 0.2]$ is a constrained learnable scale factor. This design exhibits three key properties: neutral responses are transmitted with coefficient 1.0 because $\tanh(0) = 0$, foreground-responsive channels are smoothly amplified, and noise-dominated channels are softly suppressed without destructive hard pruning.

Level-Adaptive Spatial Residual. Beyond channel-wise calibration, ZFC leverages the density map D_{sba} provided by SBA to reinforce fine-grained spatial information. Because receptive fields vary substantially across pyramid levels, blindly applying coarse density guidance to the finest Level 0 feature map can inject background interference.

To address this, ZFC introduces a level-adaptive spatial residual mechanism. Let the input feature at layer l be F_l , and let $A_l = \text{Interpolate}(D_{\text{sba}})$ be the density-assistive feature after resolution alignment. The spatial attention weight at each pixel, derived from the density map via pooling and convolution, is computed as:

$$S_{\text{attn}}^l = \sigma(\text{Conv2D}([\text{MaxPool}(A_l); \text{AvgPool}(A_l)])), \quad (9)$$

and the final calibrated feature is obtained by residual modulation, where α_l controls how

strongly the spatial density bias is applied at level l :

$$F_{\text{final}}^l = F_{\text{out}}^l \odot (1.0 + \alpha_l \cdot S_{\text{attn}}^l), \quad (10)$$

where F_{out}^l is the channel-calibrated feature from the preceding Tanh gate. We set $\alpha_0 = 0$ for the finest level to avoid coarse density guidance on extremely fine-grained pixels, while applying mild spatial correction at higher levels. This preserves local raw information for weak targets and aligns calibration with the multi-scale pyramid.

3.5 Integration: A Coupled System, Not Two Independent Modules

Sections 3.3 and 3.4 introduced SBA and ZFC as instantiations of g_{alloc} and g_{feat} from Section 3.1; here we make explicit the coupling pathway through the shared density representation D that distinguishes this design from a simple concatenation of two independently motivated modules.

Coupling pathway. The dependency is bidirectional and operates through two concrete interfaces. First, $D = D_{\text{sba}}$, the spatial density prior produced by SBA’s front-end response maps, is interpolated and injected directly into ZFC’s level-adaptive spatial residual (Eq. 9–10). Thus ZFC does not compute an independent saliency map; it reuses SBA’s density estimate to decide where spatial recalibration is applied and at what level-specific strength α_l . Second, because ZFC’s Tanh-centered gate (Eq. 7–8) operates upstream of SBA’s boundary-prediction head (Eq. 1–2) within the shared feature hierarchy, the density estimate is computed from cleaner, neutral-preserving calibrated features. SBA therefore guides where ZFC corrects, while ZFC improves what SBA measures.

Why this is not redundant. A natural question is whether jointly modeling feature calibration and query allocation introduces redundant capacity relative to addressing either problem alone. The ablation in Section 4.6 answers this empirically: each branch in isolation (ZFC-only or SBA-only) yields a measurable but partial gain, and—more importantly for the coupling claim—each isolated branch also *degrades* a metric the other branch is responsible for (ZFC-only increases LRP-FN as enriched features draw queries toward already-dense regions; SBA-only slightly lowers AP_{vt} as count-driven allocation alone cannot compensate for unrecovered feature

noise on the smallest instances). The full coupled configuration not only avoids both regressions but exceeds the sum of the two branches’ individual AP gains, which would not be expected if the branches were functionally independent. We interpret this synergy, together with the explicit shared-representation interfaces described above, as evidence that ZFC and SBA constitute one coupled adaptive system rather than two separable engineering tricks.

4 Experiments

4.1 Datasets and Evaluation Metrics

AI-TOD-V2 [5] is the most challenging publicly available benchmark for aerial tiny object detection, containing 28,036 aerial images with 752,745 annotated instances across eight semantic categories, partitioned into 11,214 training, 2,804 validation, and 14,018 test images. The object size distribution is severely skewed: the average footprint is 12.7 pixels, 86% of instances are smaller than 16 pixels, and the largest instance does not exceed 64 pixels. The per-image object count ranges from 1 to 2,667 (mean 24.64, standard deviation 63.94), imposing a severe instance-count imbalance. All AI-TOD-V2 results are reported on the test split.

VisDrone-DET2019 [22] comprises 14,018 drone-captured surveillance images partitioned into 6,471 training, 548 validation, and 3,190 test images, spanning 10 semantic categories. The dataset exhibits diverse object densities (mean 40.7 objects per image, standard deviation 46.41) and a broad range of imaging conditions. All VisDrone results are reported on the validation split following established protocols [8].

Evaluation Metrics. We employ the standard COCO-style Average Precision (AP) computed as the mean of AP values from $\text{IoU} = 0.50$ to 0.95 with step 0.05 , allowing a maximum of 1,500 detections per image. Scale-specific metrics are additionally reported: AP_{vt} (very tiny), AP_{t} (tiny), AP_{s} (small), and AP_{m} (medium). For error diagnosis, we also report the Localization Recall Precision (LRP) metric and its decomposed components LRP-Loc, LRP-FP, and LRP-FN, which quantify localization, false positive, and false negative errors respectively; lower LRP values indicate superior performance.

4.2 Implementation Details

We build our model upon DQ-DETR [8] and retain its ResNet-50 backbone, a 6-layer deformable transformer encoder, and a 6-layer deformable transformer decoder with a hidden dimension of 256. The ZFC and SBA modules are integrated by directly replacing the standard attention-based feature enhancement and query-count classification stages, respectively. All models are trained on a single NVIDIA A6000 GPU (48 GB). We train all modules for 24 epochs with an initial learning rate of 1×10^{-4} . The warm-up factor for the cosine annealing schedule in the SBA logarithmic boundary prediction is linearly increased from 0 to 1 during the first 2 epochs. The safety margin in the spacing loss L_{spacing} is set to 0.3 in logarithmic space. The learnable scale factor of the ZFC Tanh gate is clipped to the range $[0, 0.2]$. We apply random cropping and multi-scale jitter augmentation following the default DETR training protocol. For the optional test-time augmentation (TTA) setting only, we employ horizontal flipping and three-scale multi-crop inference, with augmented predictions ensembled via Soft-NMS. These Soft-NMS operations are part of TTA ensembling and are not used in the main single-scale inference results, which remain fully NMS-free.

Validation strategy. Given the coupled failure mode motivated in Section 1, evaluating only aggregate accuracy would not establish that ZFC and SBA address the mechanisms they claim to. We therefore organize the remaining experiments into three tiers of increasing specificity: (i) *overall performance* against strong representative detectors on standard benchmarks (Sections 4.3–4.5), establishing that the framework is competitive in aggregate and computationally efficient; (ii) a *small-object focus* that isolates performance on the finest-scale instances most affected by representation collapse (AP_{vt} , AP_{75} , throughout Sections 4.3 and 4.6); and (iii) *density-variation robustness*, in which the ablation study (Section 4.6) is read not as a simple module-removal exercise but as an experimental test of the two failure-mode hypotheses from Section 1—what happens when representation collapse is left unaddressed, and what happens when assignment remains non-continuous.

4.3 Overall Performance on AI-TOD-V2

Table 1 presents a quantitative comparison of our method with representative CNN-based and DETR-like detectors on the AI-TOD-V2 benchmark. All DETR-like baselines marked with * were re-implemented using our unified training protocol, which adopts 5-scale feature maps and deformable attention, ensuring a fair comparison. CNN-based detectors employ ResNet-50 with a Feature Pyramid Network, following their original publications.

Without TTA, our full model attains 32.0% AP, improving over the strongest DETR-like baseline in Table 1, DQ-DETR, by 1.8 percentage points and over the strongest CNN-based baseline, RFLA, by 6.3 points. Consistent with the small-object focus of tier (ii), the gains are most pronounced on the finest scale metrics: AP_{vt} rises from 15.3% to 16.9% (+1.6 percentage points), and AP_{75} improves from 22.3% to 25.7% (+3.4 points), indicating improved localization precision from ZFC’s neutral-preserving feature calibration and SBA’s soft query allocation. With the optional TTA setting, the model reaches 33.4% AP and 18.9% AP_{vt} ; because this row uses Soft-NMS ensembling, we report it separately from the main NMS-free result.

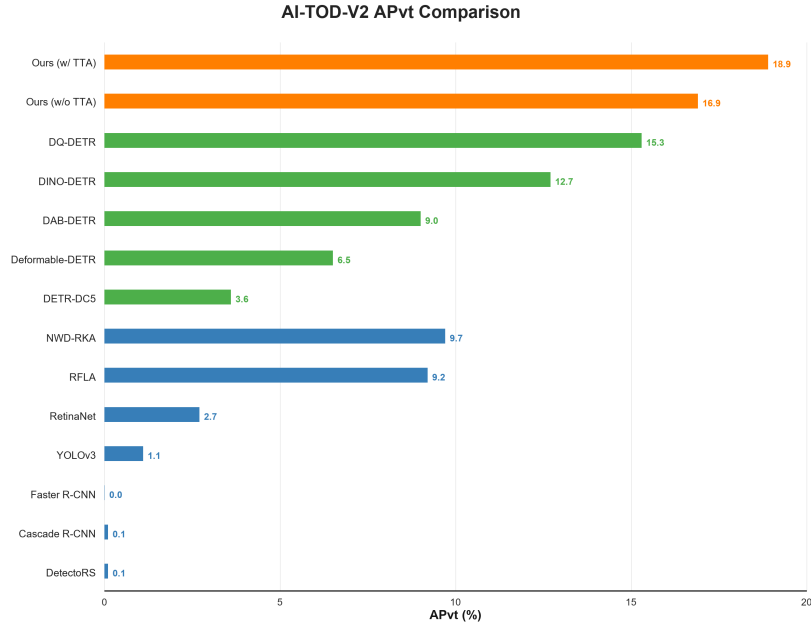
The AP_{75} gain is particularly noteworthy: this strict metric is sensitive to bounding-box localization accuracy for tiny objects, indicating that SBA provides better-matched query resources for dense clusters while ZFC preserves fine-grained spatial detail for precise regression. The AP_{vt} comparison in Fig. 3 highlights the same trend at the finest scale. All non-TTA DETR-like methods listed above operate without NMS at inference; the optional TTA row is reported separately and uses Soft-NMS ensembling, as indicated in Table 1.

4.4 Generalization on VisDrone-DET2019

To further validate the generalization capability of the proposed method beyond AI-TOD-V2, we evaluate our full model (ZFC + SBA) on the VisDrone-DET2019 dataset [22]. Table 2 compares our method against strong representative detectors.

Table 1 Comparison with representative methods on AI-TOD-V2.

Method	Backbone	NMS-Free	AP	AP ₅₀	AP ₇₅	AP _{vt}	AP _t	AP _s	AP _m
CNN-based Methods									
YOLOv3 [23]	DarkNet-53	✗	4.1	14.6	0.9	1.1	4.8	7.7	8.0
RetinaNet [24]	R-50-FPN	✗	8.9	24.2	4.6	2.7	8.4	13.1	20.2
Faster R-CNN [25]	R-50-FPN	✗	12.8	29.9	9.4	0.0	9.2	24.6	37.0
Cascade R-CNN [26]	R-50-FPN	✗	15.1	34.2	11.2	0.1	11.5	26.7	38.5
DetectoRS [27]	R-50-FPN	✗	16.1	35.5	12.5	0.1	12.6	28.3	40.0
NWD-RKA [10]	R-50-FPN	✗	24.7	57.4	17.1	9.7	24.2	29.8	39.3
RFLA [12]	R-50-FPN	✗	25.7	58.9	18.8	9.2	25.5	30.2	40.2
End-to-End DETR-like Methods (ResNet-50 Backbone)									
DETR-DC5 [1]	ResNet-50	✓	10.4	32.5	3.9	3.6	9.3	13.2	24.6
Deformable-DETR [2]	ResNet-50	✓	18.9	50.0	10.5	6.5	17.6	25.3	34.4
DAB-DETR [3]	ResNet-50	✓	22.4	55.6	14.3	9.0	21.7	28.3	38.7
DINO-DETR [4]	ResNet-50	✓	25.9	61.3	17.5	12.7	25.3	32.0	39.7
DQ-DETR [8]	ResNet-50	✓	30.2	68.6	22.3	15.3	30.5	36.5	44.6
Ours (w/o TTA)	ResNet-50	✓	32.0	69.2	25.7	16.9	32.0	37.7	47.2
Ours (w/ TTA)	ResNet-50	✗(Soft-NMS)	33.4	69.6	27.5	18.9	34.1	38.4	47.6

**Fig. 3** Very-tiny AP (AP_{vt}, <8px) comparison. Our model reaches 18.9% AP_{vt} with optional TTA, a +3.6-point gain over DQ-DETR, supporting the complementary benefit of ZFC neutral-preserving calibration and SBA soft allocation on the finest-scale targets.

Our model achieves 38.2% AP, 61.5% AP₅₀, and 39.5% AP₇₅, outperforming DQ-DETR by 1.2 pp, 0.6 pp, and 1.6 pp, respectively.

The AP gain over DQ-DETR is slightly smaller than on AI-TOD-V2, consistent with VisDrone’s higher proportion of medium and large objects; our optimizations specifically target tiny-object

challenges. Nevertheless, our method achieves the best AP₇₅ among the methods compared in Table 2, indicating high-precision localization across mixed scales and supporting the cross-dataset generalization of ZFC and SBA.

Table 2 Comparison on VisDrone-DET2019. Best in bold.

Model	AP	AP ₅₀	AP ₇₅
Faster R-CNN [25]	21.4	40.7	19.9
Cascade R-CNN [26]	22.6	38.8	23.2
YOLOv5 [28]	24.1	44.1	22.3
CEASC [29]	28.7	50.7	24.7
SDP [30]	30.2	52.5	28.4
DNTR [31]	33.1	53.8	34.8
DINO-DETR [4]	35.8	58.3	36.8
DQ-DETR [8]	37.0	60.9	37.9
Ours (ZFC + SBA)	38.2	61.5	39.5

4.5 Model Complexity Analysis

As detailed in Table 3, the proposed method achieves a significant reduction in computational cost while improving detection accuracy.

Our method retains exactly the same parameter count as the baseline DQ-DETR (58.7 M), yet reduces the computational cost from 1805.4 GFLOPs to 921.0 GFLOPs—a 49% reduction. This efficiency gain stems primarily from SBA’s dynamic query allocation: unlike DQ-DETR’s fixed over-provisioned query count set to the maximum required for the densest scenes, SBA adaptively reduces the query count in sparse and moderate-density images, substantially lowering the cross-attention cost in the decoder. The ZFC module introduces negligible additional overhead, as the Tanh-centered gating shares pooling operations with the existing attention and the level-adaptive spatial residual is applied only on middle and high-level FPN features.

Table 3 Model complexity comparison. Best in bold.

Model	Params (M)	GFLOPs
DQ-DETR [8]	58.7	1805.4
Ours (ZFC + SBA)	58.7	921.0

4.6 Ablation Study as Failure-Hypothesis Verification

Rather than a simple module-removal exercise, this ablation tests the two failure-mode hypotheses introduced in Section 1: what happens when representation collapse remains active, and what happens when query assignment remains non-continuous. All variants use the same training schedule and hyperparameters on the AI-TOD-V2 validation set. Table 4 reports AP metrics and LRP error decomposition; Fig. 4 and Fig. 5 visualize the results.

Hypothesis 1: Leaving Representation Collapse Unaddressed. Comparing DQ-DETR (Baseline) to + ZFC only isolates the cost of representation collapse: enabling ZFC alone raises AP from 30.2% to 31.0%. The LRP decomposition attributes this gain primarily to a reduction in LRP-FP from 25.0 to 23.2—the largest single-branch FP reduction in the ablation—supporting the hypothesis that the Sigmoid gate’s neutral-activation suppression ($\sigma(0) = 0.5$) contributes to false alarms rather than being an incidental design choice. ZFC’s Tanh gate assigns a multiplicative weight of exactly 1.0 to neutral inputs, avoiding the 0.5 attenuation imposed by Sigmoid while gently amplifying foreground-responsive channels (weight $\in (1.0, 1.2]$) and softly suppressing noise channels (weight $\in [0.8, 1.0]$). AP_{vt} improves to 15.7%, suggesting that neutral-feature preservation particularly benefits the very tiniest instances—the regime in which representation collapse, left unaddressed, is most damaging.

Fig. 6 directly compares Sigmoid and Tanh variants of the ZFC gate. Tanh consistently outperforms Sigmoid on all AP sub-metrics (Fig. 6(a)) and achieves the largest LRP improvement (Fig. 6(b)), supporting the benefit of neutral-feature preservation. Notably, ZFC alone increases LRP-FN from 31.2 to 32.4, which is consistent with the coupling claim: correcting representation without correcting allocation can still leave sparse regions under-queried.

Hypothesis 2: Leaving Assignment Non-Continuous. Comparing DQ-DETR (Baseline) to + SBA only isolates the cost of assignment misallocation: enabling SBA alone yields 31.2% AP with LRP-FN improving slightly to 31.7, supporting the hypothesis that continuous, per-image density boundaries—rather than

Table 4 Ablation study on AI-TOD-V2 validation set. $\checkmark$ / $\times$ indicates module enabled / disabled. LRP and its components are reported as percentages ($\times 100$); lower is better.

Model	ZFC	SBA	AP	AP ₅₀	AP ₇₅	AP _{vt}	AP _t	LRP	LRP-FP	LRP-FN
DQ-DETR (Baseline)	$\times$	$\times$	30.2	68.6	22.3	15.3	30.5	72.7	25.0	31.2
+ ZFC only	$\checkmark$	$\times$	31.0	68.5	23.3	15.7	30.9	72.0	23.2	32.4
+ SBA only	$\times$	$\checkmark$	31.2	68.3	23.5	15.0	31.2	72.0	23.5	31.7
Full (ZFC+SBA)	$\checkmark$	$\checkmark$	32.0	69.2	25.7	16.9	32.0	70.8	23.1	31.2
Full + TTA	$\checkmark$	$\checkmark$	33.4	69.6	27.5	18.9	34.1	70.3	23.0	32.7

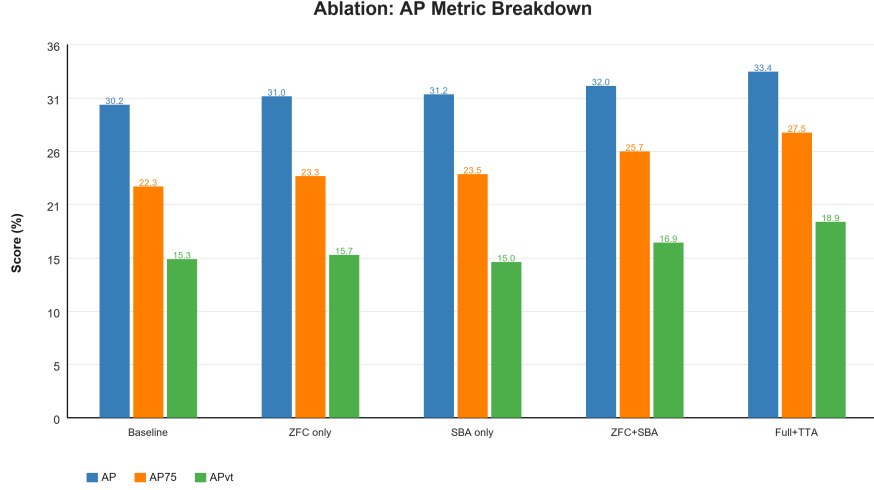


Fig. 4 Ablation study: AP metric breakdown across configurations. Grouped bars report AP, AP₇₅, and AP_{vt}. The full ZFC+SBA model achieves 32.0% AP (+1.8% over DQ-DETR baseline), substantially exceeding the sum of individual contributions (+0.5%+0.5%=1.0%).

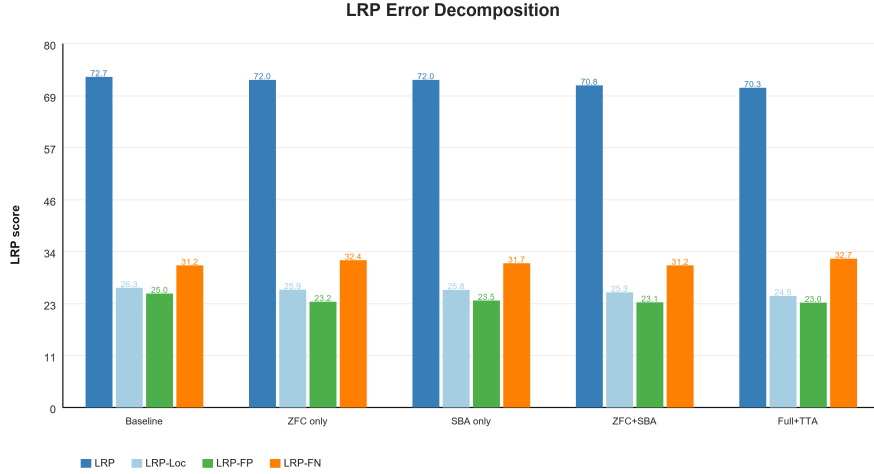


Fig. 5 LRP error decomposition (grouped bars) showing complementary module roles: ZFC primarily reduces LRP-FP (false positives) via Tanh-centered neutral-preserving feature calibration; SBA primarily reduces LRP-FN (missed detections) via CDF-based soft query allocation. The full model achieves overall LRP=70.8.

DQ-DETR’s hard, dataset-level thresholds—help reduce missed detections through better-matched query coverage. By regressing continuous logarithmic boundaries conditioned on per-image density features and optimizing with the CDF-based True Adaptive Boundary Loss, SBA eliminates the discontinuous allocation jumps caused by hard statistical thresholds, improving recall in moderate-to-dense scenes. The AP_{75} gain indicates that more accurately matched query budgets also improve localization precision. However, SBA alone causes AP_{vt} to slip slightly to 15.0%: mirroring Hypothesis 1’s signature in reverse, *leaving representation collapse unaddressed while only correcting assignment* means count-driven allocation can redirect query resources toward regions whose feature evidence is still Sigmoid-suppressed, away from the most challenging sub-8px targets. LRP-FP is also marginally higher than under ZFC alone. Together, the two single-branch variants reproduce, in both directions, the coupling mechanism argued for in Section 1: correcting either axis alone leaves a residual error signature attributable specifically to the uncorrected axis.

Both Hypotheses Addressed Jointly. The full model corrects both axes of the coupled failure mode simultaneously and achieves 32.0% AP and LRP = 70.8, with LRP-FP = 23.1 and LRP-FN = 31.2—improving on *both* the LRP-FP regression observed under SBA-only and the LRP-FN regression observed under ZFC-only, rather than merely averaging the two single-branch outcomes. This is the joint-correction signature predicted in Section 3.5: ZFC produces higher-quality, less noisy features; SBA then allocates a better-matched query budget across density regimes using those calibrated features, recovering the sparse-region coverage lost by ZFC alone, while ZFC’s cleaner representation in turn gives SBA’s count-driven allocation a more reliable signal to serve the hardest sub-8px samples. The synergy is most pronounced on AP_{75} and AP_{vt} —metrics sensitive to the finest-scale instances where both failure modes are expected to matter most. The optional TTA setting further raises performance to 33.4% AP ($AP_{75} = 27.5\%$, $AP_{vt} = 18.9\%$, LRP = 70.3); this TTA result uses multi-crop/flip ensembling with Soft-NMS and is therefore separate from the main 32.0% NMS-free inference result. The marginally higher LRP-FN under TTA

may reflect the conservative Soft-NMS ensemble strategy.

Comparison of Alternative Gating Activations: Swish and GLU. To further validate the design choice of Tanh as the ZFC gating function, we additionally fine-tune the full ZFC+SBA model with Swish [32] and a Gated Linear Unit (GLU) [33] under identical conditions. For Swish, the reported value is the maximum of the prediction interval; for GLU, it is the prediction median. Fig. 7 summarises the results.

As shown in Fig. 7(a), Tanh achieves the best AP (32.0%), AP_{75} (25.6%), and AP_{vt} (16.8%). The LRP decomposition in Fig. 7(b) further shows that Tanh achieves the lowest total LRP (70.8 vs. 72.8 for Swish and 71.4 for GLU), driven primarily by lower false-negative error. These results support the zero-centered gate design in Section 3.4.

Fig. 8 reveals a consistent inverse relationship between gain magnitude and object size: the largest improvements are in AP_{vt} (+10.5% relative) and AP_{75} (+15.2% relative), benefiting the dominant 86% of AI-TOD-V2 instances smaller than 16 pixels.

4.7 Sensitivity Analysis of LTABL Loss Coefficients

To validate the robustness of the proposed LTABL loss design, we conduct a systematic sensitivity analysis of its three weighting coefficients defined in Eq. (6). Each coefficient is swept over $\{0.05, 0.10, 0.20\}$ while the other two are held fixed at their baseline values ($\lambda_1 = 0.10$, $\lambda_2 = 0.05$, $\lambda_3 = 0.20$). Fig. 9 reports both AP metrics and LRP decomposition on the AI-TOD-V2 validation set.

Coverage loss coefficient λ_1 (Fig. 9a, d).

The coverage loss enforces distinct CDF expectations across density levels. At $\lambda_1 = 0.05$, AP drops to 31.0% (vs. 32.0% at the baseline 0.10), with AP_{vt} and AP_{75} declining to 15.8% and 23.6%, respectively, and LRP rising to 72.0, indicating that insufficient coverage supervision allows boundary positions to drift and degrades localization of the finest instances. Increasing λ_1 to 0.20 maintains AP at 32.0% and further improves AP_{75} to 26.7% and LRP to 70.4, primarily through

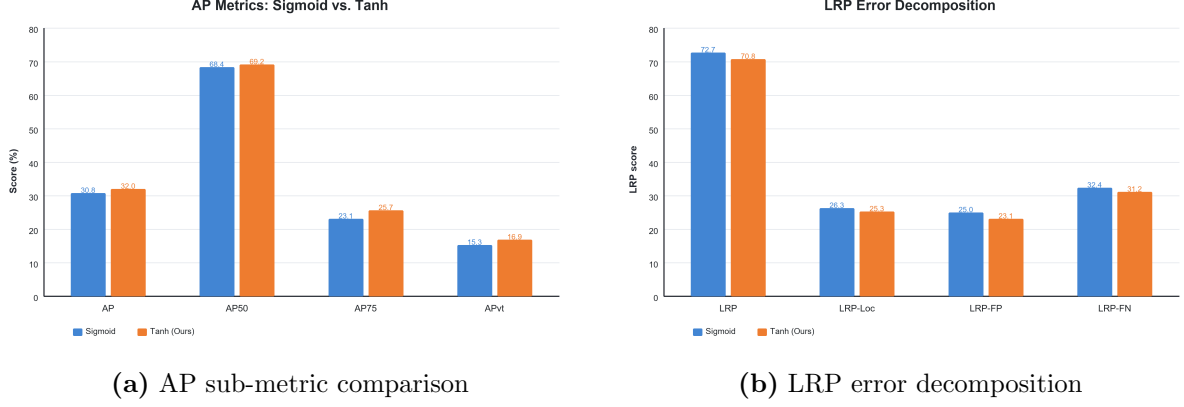
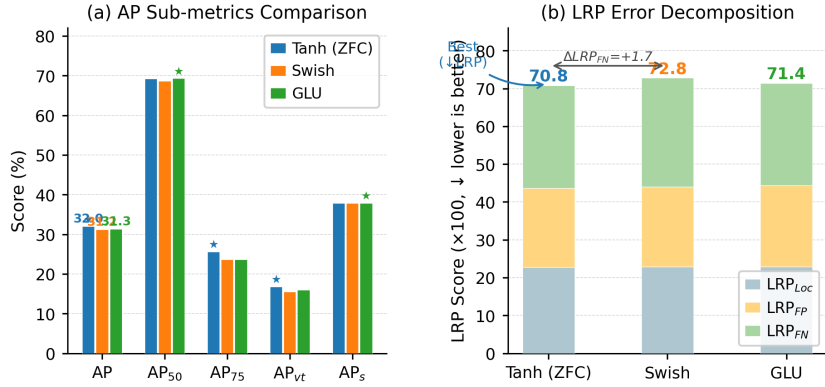


Fig. 6 Quantitative comparison of Sigmoid and Tanh activation functions in the ZFC module. (a) AP sub-metric comparison: Tanh outperforms Sigmoid on AP, AP₅₀, AP₇₅ and AP_{vt}. (b) LRP error decomposition: Tanh achieves the largest improvement on LRP, indicating that the Tanh-centered neutral-preserving gate reduces false-positive detections relative to Sigmoid; lower values are better.



* = best per metric. Fine-tuning: DQ_5scale\finetuneV2, lr=1e-5, mosaic_p=0.3, copy\paste\p=0.3, 18 epochs. Swish: max of prediction interval; GLU: prediction median.

Fig. 7 Comparison of Tanh, Swish, and GLU gating activations in the ZFC module under identical fine-tuning conditions on AI-TOD-V2. (a) AP sub-metrics: Tanh (ZFC) consistently outperforms Swish and GLU, particularly on AP₇₅ and AP_{vt} (★ = best per metric). (b) LRP error decomposition: Tanh achieves the lowest total LRP (70.8), with the principal advantage in LRP_{FN} ($\Delta = +1.7$ over Swish), indicating that the zero-centered $\tanh(0) = 0$ property helps reduce missed detections for weak tiny-object signals. Lower LRP is better.

Table 5 LRP error decomposition analysis ($\times 100$, lower is better). $\downarrow$ and $\uparrow$ arrows indicate improvement and degradation relative to the preceding row.

Method	Situation	LRP $\downarrow$	LRP-Loc $\downarrow$	LRP-FP $\downarrow$	LRP-FN $\downarrow$
DQ-DETR (Baseline)	Overall	72.7	26.3	25.0	31.2
+ ZFC only	Overall	72.0	25.9	23.2 $\downarrow$	32.4 $\uparrow$
+ SBA only	Overall	72.0	25.8	23.5 $\downarrow$	31.7 $\downarrow$
Full (ZFC+SBA)	Overall	70.8	25.3	23.1	31.2
Full + TTA	Overall	70.3	24.5	23.0	32.7

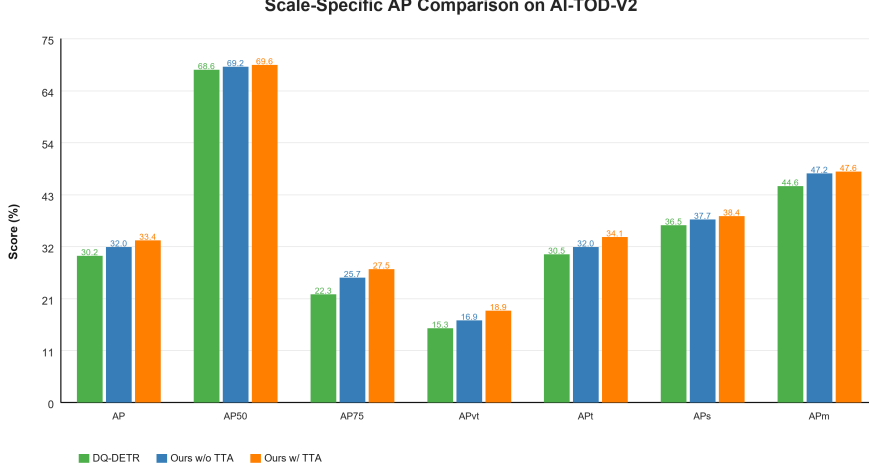


Fig. 8 Scale-specific AP improvement over DQ-DETR baseline. the grouped bars correspond to DQ-DETR, Ours w/o TTA, and Ours w/ TTA. Gains are largest for AP_{vt} (+1.6%/+3.6%) and AP_{75} (+3.4%/+5.2%), supporting that ZFC and SBA jointly improve localization precision and recall for the finest-scale targets.

reduced false positives (LRP-FP = 22.8), indicating that the coverage penalty remains well-behaved across a range of weights. We adopt $\lambda_1 = 0.10$ as the baseline for balanced sensitivity.

Spacing loss coefficient λ_2 (Fig. 9b, e).

The spacing loss prevents boundary collapse via ReLU-based repulsive regularization. AP exhibits a clear monotonic degradation as λ_2 increases: 32.0% (0.05) $\rightarrow$ 31.7% (0.10) $\rightarrow$ 30.9% (0.20). AP_{75} degrades sharply from 25.6% to 22.6%, indicating that over-regularizing boundary separation sacrifices localization precision by forcing boundaries apart even when the data favors finer granularity. LRP correspondingly rises from 70.8 to 72.0, with LRP-FN increasing from 31.3 to 32.7, as rigid spacing constraints cause miss-allocation of queries in regions where natural density transitions are subtle. The selected value is $\lambda_2 = 0.05$, where spacing regularization is sufficient to prevent collapse without constraining flexibility.

Count loss coefficient λ_3 (Fig. 9c, f).

The count loss provides global supervision on the predicted object count N_{pred} . A pronounced non-monotonic trend is observed: AP = 31.5% at $\lambda_3 = 0.05$, drops to 30.6% at 0.10, and peaks at 32.0% at 0.20. The dip at 0.10 is most severe on AP_{75} (22.5%, a drop of 3.4 points from 0.05) and accompanied by the highest LRP (72.8) across all configurations, driven by elevated false positives (LRP-FP = 25.8) and missed detections (LRP-FN = 34.5). This pattern indicates that intermediate

count supervision provides misleading gradients: the signal is strong enough to perturb boundary estimation but too weak to enforce coherent count-boundary alignment, resulting in unstable boundary estimates under the long-tailed distribution of AI-TOD-V2. At $\lambda_3 = 0.20$, strong count supervision anchors the boundaries, restoring the main metrics to their selected-setting values. We therefore select $\lambda_3 = 0.20$ as the baseline.

Summary. Across all three coefficients, the baseline configuration ($\lambda_1 = 0.10$, $\lambda_2 = 0.05$, $\lambda_3 = 0.20$) achieves the best joint AP of 32.0% and competitive LRP of 70.8. Notably, neighboring values (e.g., $\lambda_1 = 0.20$ with AP = 32.0%, or $\lambda_2 = 0.05$ adjacent to 0.10) yield comparable or only marginally degraded performance, suggesting that the selected configuration is not a sharp local optimum and endowing LTABL with reliable practical robustness.

4.8 Evidence of Density-Variation Robustness

Complementing the failure-hypothesis analysis in Section 4.6, Fig. 10 summarizes LRP, LRP-FP, and LRP-FN across DQ-DETR, the proposed model without TTA, and the optional TTA setting. The lower LRP and LRP-FP values for our model are consistent with the module-level ablation in Tables 4 and 5. Fig. 11 then provides qualitative examples under diverse object densities and scene conditions.

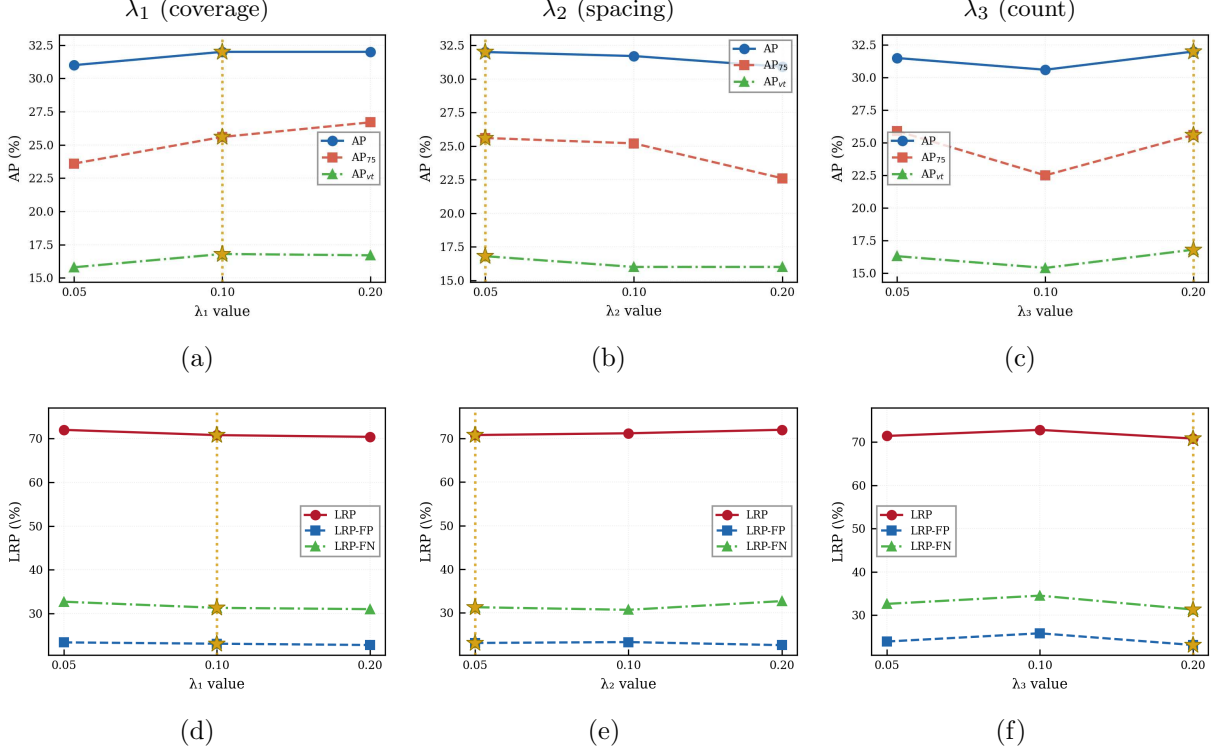


Fig. 9 Sensitivity analysis of the LTABL loss coefficients on the AI-TOD-V2 validation set. Top row (a-c): AP sub-metrics ($\uparrow$ higher is better). Bottom row (d-f): LRP error decomposition ($\downarrow$ lower is better). Each column varies one coefficient while fixing the other two at their baseline values ($\lambda_1 = 0.10$, $\lambda_2 = 0.05$, $\lambda_3 = 0.20$). Gold dotted vertical lines and star markers (★) indicate the selected baseline configuration.

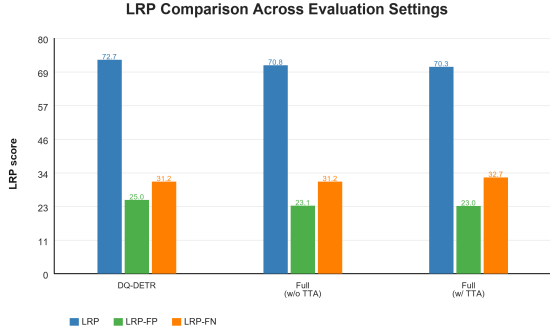


Fig. 10 LRP comparison across evaluation settings on AI-TOD-V2. Bars report LRP, LRP-FP, and LRP-FN for DQ-DETR, the proposed full model without TTA, and the optional TTA setting; lower values are better.

Dense scenes. The qualitative examples in Fig. 11 show that SBA’s soft allocation helps recover instances that are likely to be missed when query allocation is too rigid, consistent with the LRP-FN trends in Table 5.

Sparse scenes. The same examples indicate that ZFC’s Tanh-centered calibration reduces texture-induced false positives, consistent with the LRP-FP reduction in Table 5.

Moderate-density scenes. Under mixed density and clutter, combining ZFC and SBA addresses both false positives and missed detections, supporting the joint-design interpretation rather than a single-module explanation.

4.9 Discussion

Why Joint Modeling Works. The ablation results support a bidirectional error-propagation interpretation. When representation collapse is left uncorrected, the allocator reads density from Sigmoid-attenuated features; when allocation remains hard-thresholded, improved features

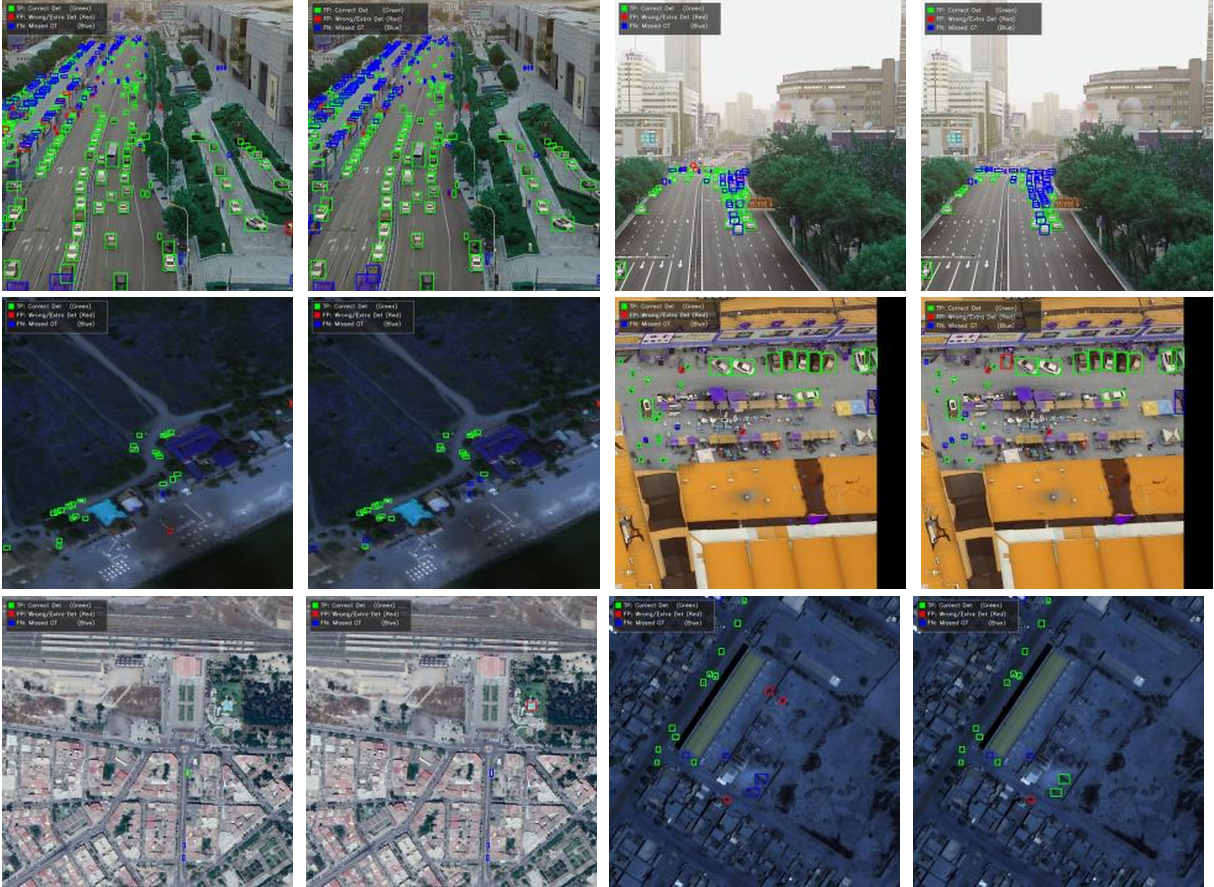


Fig. 11 Visualization of the proposed method under diverse scenes. Detection examples across various object densities, scene types, and lighting conditions illustrate the robustness and generalization of ZFC and SBA modules. The boxes denote true positive detections.

still may not receive enough decoder capacity in sparse or transitional regions. The full model addresses both error sources jointly, which is consistent with the gain exceeding the sum of the single-branch gains in Fig. 4.

Fair-Comparison Conditions and Source of Efficiency Gains. Both our model and DQ-DETR are trained under identical conditions (ResNet-50, 24 epochs, 5-scale features, deformable attention) for a fair comparison. The improvements over DQ-DETR—replacing CBAM-style Sigmoid attention with the ZFC Tanh gate, and replacing hard statistical thresholds with SBA’s CDF-constrained soft allocation—are achieved while preserving the NMS-free end-to-end paradigm in the main non-NTA inference setting and identical parameter count (58.7 M). The 49% GFLOPs reduction

(1805.4 $\rightarrow$ 921.0) arises purely from SBA dynamically scaling query counts, rather than over-provisioning for worst-case density in every image. Integrating the proposed modules with stronger backbones and longer training schedules is a promising direction for future work.

5 Conclusion

This paper identifies representation collapse and density-variant assignment misallocation as a coupled failure mode in tiny object detection, rather than two independent limitations of existing DETR-based paradigms, and reformulates detection as joint optimization over a feature-calibration branch and a density-allocation branch sharing a common representation.

We demonstrate that the proposed Zero-Harm Feature Calibration (ZFC) module effectively mitigates the structural suppression of weak foreground features caused by conventional activation functions. Through a Tanh-centered gating mechanism, ZFC softly attenuates background noise while preserving neutral feature integrity, incurring negligible additional inference latency. In parallel, the Statistical Boundary Allocator (SBA) replaces rigid heuristic thresholds with a novel soft-constraint mechanism grounded in cumulative distribution functions. This enables the query budget to adaptively track the continuous density profile of each image, substantially alleviating the recall collapse observed in extremely crowded scenes. Critically, ZFC and SBA are coupled rather than independent: they share a density representation through which SBA’s spatial prior informs ZFC’s level-adaptive correction and ZFC’s calibrated features in turn sharpen the density signal SBA allocates against (Section 3.5), and the ablation study supports that addressing either failure mode alone leaves a residual error signature attributable to the other (Section 4.6).

This coupled design empirically yields a synergistic accuracy gain that exceeds the sum of each branch’s individual contribution. More importantly, the proposed framework avoids NMS in the main inference setting while reserving Soft-NMS only for optional TTA ensembling, preserving the fully end-to-end DETR-style design philosophy for the primary result and advancing tiny object detection from coarse, independently patched heuristics toward a unified statistical adaptation paradigm.

Statements and Declarations

Funding. No funding was received to assist with the preparation of this manuscript.

Competing Interests. The authors have no competing interests to declare that are relevant to the content of this article.

Ethics Approval. Not applicable. This article does not contain studies with human participants or animals performed by any of the authors.

Data Availability. The AI-TOD-V2 and VisDrone-DET2019 datasets analyzed in this study are publicly available from their original dataset providers and are cited in the manuscript.

Data generated during the current study, including trained model checkpoints and evaluation logs, are available from the corresponding author on reasonable request.

Code Availability. The custom code used in this study is available from the corresponding author on reasonable request.

Author Contributions. All authors contributed to the study conception and design. Method development, experiments, and analysis were performed by Yingying Zhai, Chang He, and Zhi Wang. The first draft of the manuscript was written by Yingying Zhai. Zezheng Feng supervised the work and served as the corresponding author. Bin Wang and Xiaochun Yang contributed to supervision, review, and editing. All authors read and approved the final manuscript.

References

- [1] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: Computer Vision – ECCV 2020, pp. 213–229. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58452-8_13
- [2] Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable DETR: Deformable transformers for end-to-end object detection. In: International Conference on Learning Representations (2021)
- [3] Liu, S., Li, F., Zhang, H., Yang, X., Qi, X., Su, H., Zhu, J., Zhang, L.: DAB-DETR: Dynamic anchor boxes are better queries for DETR. In: International Conference on Learning Representations (2022)
- [4] Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L.M., Shum, H.-Y.: DINO: DETR with improved denoising anchor boxes for end-to-end object detection. In: International Conference on Learning Representations (2023)
- [5] Wang, J., Yang, W., Guo, H., Zhang, R., Xia, G.-S.: Tiny object detection in aerial images. In: 2020 25th International Conference on Pattern Recognition, pp. 3791–3798 (2021). <https://doi.org/10.1109/ICPR48806>

- [6] Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7132–7141 (2018). <https://doi.org/10.1109/CVPR.2018.00745>
- [7] Woo, S., Park, J., Lee, J.-Y., Kweon, I.S.: CBAM: Convolutional block attention module. In: Computer Vision – ECCV 2018, pp. 3–19. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01234-2_1
- [8] Huang, Y.-X., Liu, H.-I., Shuai, H.-H., Cheng, W.-H.: DQ-DETR: DETR with dynamic query for tiny object detection. In: Computer Vision – ECCV 2024, pp. 290–305. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-73116-7_17
- [9] Shen, F., Cui, J., Li, W., Zhou, S.: TinyDef-DETR: A transformer-based framework for defect detection in transmission lines from UAV imagery. Remote Sensing **17**(23), 3789 (2025) <https://doi.org/10.3390/rs17233789>
- [10] Xu, C., Wang, J., Yang, W., Yu, H., Yu, L., Xia, G.-S.: Detecting tiny objects in aerial images: A normalized wasserstein distance and a new benchmark. ISPRS Journal of Photogrammetry and Remote Sensing **190**, 79–93 (2022) <https://doi.org/10.1016/j.isprsjprs.2022.06.002>
- [11] Xu, C., Wang, J., Yang, W., Yu, L.: Dot distance for tiny object detection in aerial images. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 1192–1201 (2021). <https://doi.org/10.1109/CVPRW53098.2021.00130>
- [12] Xu, C., Wang, J., Yang, W., Yu, L., Xia, G.-S.: RFLA: Gaussian Receptive Field Based Label Assignment for Tiny Object Detection (2022)
- [13] Cao, B., Yao, H., Zhu, P., Hu, Q.: Visible and clear: Finding tiny objects in difference map. In: Computer Vision – ECCV 2024, pp. 1–18. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-72643-9_1
- [14] He, X., Liang, K., Zhang, W., Li, F., Jiang, Z., Zuo, Z., Tan, X.: DETR-ORD: An improved DETR detector for oriented remote sensing object detection with feature reconstruction and dynamic query. Remote Sensing **16**(18), 3516 (2024) <https://doi.org/10.3390/rs16183516>
- [15] Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, pp. 936–944 (2017). <https://doi.org/10.1109/CVPR.2017.106>
- [16] Zhang, S., Wang, X., Wang, J., Pang, J., Lyu, C., Zhang, W., Luo, P., Chen, K.: Dense distinct query for end-to-end object detection. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7329–7338 (2023). <https://doi.org/10.1109/CVPR52729.2023.00708>
- [17] Gao, P., Zheng, M., Wang, X., Dai, J., Li, H.: Fast convergence of DETR with spatially modulated co-attention. In: 2021 IEEE/CVF International Conference on Computer Vision, pp. 3601–3610 (2021). <https://doi.org/10.1109/ICCV48922.2021.00360>
- [18] Dai, X., Chen, Y., Yang, J., Zhang, P., Yuan, L., Zhang, L.: Dynamic DETR: End-to-end object detection with dynamic attention. In: 2021 IEEE/CVF International Conference on Computer Vision, pp. 2968–2977 (2021). <https://doi.org/10.1109/ICCV48922.2021.00298>
- [19] Meng, D., Chen, X., Fan, Z., Zeng, G., Li, H., Yuan, Y., Sun, L., Wang, J.: Conditional DETR for fast training convergence. In: 2021 IEEE/CVF International Conference on Computer Vision, pp. 3631–3640 (2021). <https://doi.org/10.1109/ICCV48922.2021.00363>
- [20] Zhao, Y., Lv, W., Xu, S., Wei, J., Wang, G., Dang, Q., Liu, Y., Chen, J.: DETRs beat YOLOs on real-time object detection. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16965–16974 (2024). <https://doi.org/10.1109/CVPR.2024.1017714>

- [21] Sun, P., Zhang, R., Jiang, Y., Kong, T., Xu, C., Zhan, W., Tomizuka, M., Li, L., Yuan, Z., Wang, C., Luo, P.: Sparse R-CNN: End-to-end object detection with learnable proposals. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14449–14458 (2021). <https://doi.org/10.1109/CVPR46437.2021.01422>
- [22] Zhu, P., Wen, L., Du, D., Bian, X., Fan, H., Hu, Q., Ling, H.: Detection and tracking meet drones challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **44**(11), 7380–7399 (2022) <https://doi.org/10.1109/TPAMI.2021.3119563>
- [23] Redmon, J., Farhadi, A.: YOLOv3: An Incremental Improvement (2018)
- [24] Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: 2017 IEEE International Conference on Computer Vision, pp. 2999–3007 (2017). <https://doi.org/10.1109/ICCV.2017.324>
- [25] Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **39**(6), 1137–1149 (2017) <https://doi.org/10.1109/TPAMI.2016.2577031>
- [26] Cai, Z., Vasconcelos, N.: Cascade R-CNN: Delving into high quality object detection. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6154–6162 (2018). <https://doi.org/10.1109/CVPR.2018.00644>
- [27] Qiao, S., Chen, L.-C., Yuille, A.: DetectoRS: Detecting objects with recursive feature pyramid and switchable atrous convolution. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10208–10219 (2021). <https://doi.org/10.1109/CVPR46437.2021.01008>
- [28] Jocher, G., Chaurasia, A., Stoken, A., et al.: ultralytics/yolov5: v7.0 – YOLOv5 SOTA Realtime Instance Segmentation. Zenodo (2022). <https://doi.org/10.5281/zenodo.7347926>. <https://zenodo.org/record/7347926>
- [29] Du, B., Huang, Y., Chen, J., Huang, D.: Adaptive sparse convolutional networks with global context enhancement for faster object detection on drone images. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13435–13444 (2023). <https://doi.org/10.1109/CVPR52729.2023.01291>
- [30] Ma, Y., Chai, L., Jin, L.: Scale decoupled pyramid for object detection in aerial images. *IEEE Transactions on Geoscience and Remote Sensing* **61**, 1–14 (2023) <https://doi.org/10.1109/TGRS.2023.3298852>
- [31] Liu, H.-I., Tseng, Y.-W., Chang, K.-C., Wang, P.-J., Shuai, H.-H., Cheng, W.-H.: A denoising FPN with transformer R-CNN for tiny object detection. *IEEE Transactions on Geoscience and Remote Sensing* **62**, 1–15 (2024) <https://doi.org/10.1109/TGRS.2024.3396489>
- [32] Ramachandran, P., Zoph, B., Le, Q.V.: Searching for Activation Functions (2017)
- [33] Dauphin, Y.N., Fan, A., Auli, M., Grangier, D.: Language modeling with gated convolutional networks. In: Proceedings of the 34th International Conference on Machine Learning, pp. 933–941 (2017). <https://proceedings.mlr.press/v70/dauphin17a.html>