SMALL OBJECTS INSIDE THE GRID: ACTIVATING LATENT SCALE KNOWLEDGE IN FROZEN DETECTORS VIA TRAINING-FREE GRID REPARAMETERIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Small-object detection does not always improve when a frozen Transformer detector receives a higher-resolution image: its internal spatial tensors may no longer align with the resulting feature grids. We hypothesize that pretrained detectors already encode useful scale-related knowledge, but fixed positional embeddings, decoder anchors, and validity masks constrain its use. We propose Adaptive Causal Grid Intervention (ACGI), a training-free framework that reconstructs these tensors from the actual feature shapes. ACGI measures cross-grid geometric consistency and uses annotation-free spatial-compression routing to select an intervention scale, without gradients, target-domain labels, or parameter updates. Across COCO, TinyPerson, and VisDrone2019, ACGI improves overall and small-object detection for multiple frozen DETR variants. Reliability ranking, scale diagnosis, and deletion analyses further link cross-grid consistency to the model's use of internal spatial representations. These findings suggest that scale adaptation can draw on spatial knowledge already present in pretrained detectors.

1 INTRODUCTION

Small object detection is fundamental to high-resolution vision tasks such as aerial inspection, long-range surveillance, and maritime rescue. Small objects occupy only a few pixels and feature units, so their semantic evidence and geometric boundaries vary substantially with the sampling scale. The DEtection TRansformer (DETR) formulates detection as set prediction (Carion et al., 2020). Subsequent real-time DETR variants improve efficiency and accuracy through multi-scale encoding, query selection, and refined regression (Zhu et al., 2021; Meng et al., 2021; Liu et al., 2022; Li et al., 2022; Zhang et al., 2023; Zhao et al., 2024; Peng et al., 2025). A more fundamental question, however, is not how to simply increase resolution. It is how a scale transformation in continuous image coordinates can preserve semantic-geometric correspondence across the discrete token grid, positional encoding, and query anchors of a frozen detector. When the input size changes, the multi-scale feature grids change accordingly, while preset positional embeddings, decoder anchors, and validity masks may remain tied to the original discrete coordinates. This coupling can violate the geometric consistency expected under scale changes. The problem is especially pronounced for small objects because a shift of only a few grid units can alter feature assignment and candidate initialization. This observation raises a key scientific question. Must scale adaptation be relearned through downstream training, or is it already encoded as latent spatial knowledge in a frozen detector and directly activatable by reconstructing its discrete coordinate system?

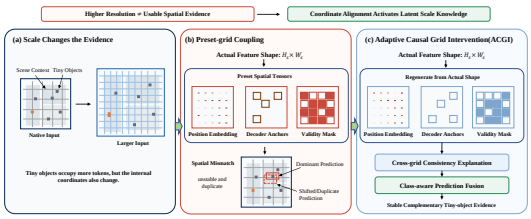

Figure 1: Motivation for ACGI. Input scaling changes both tiny-object resolution and feature coordinates, while misaligned spatial priors cause unstable predictions. ACGI restores grid alignment and fuses consistent, complementary evidence across scales.

Existing studies do not answer this question. Training-time methods for small objects allocate more learning capacity to low-resolution instances through feature pyramids, loss reweighting, query de-

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.

sign, and data augmentation (Zhu et al., 2021; Meng et al., 2021; Liu et al., 2022; Li et al., 2022; Zhang et al., 2023; Zhao et al., 2024; Peng et al., 2025; Huang et al., 2025; Lin et al., 2017; Singh & Davis, 2018; Singh et al., 2018; Yang et al., 2022; Xu et al., 2022). Although these methods improve representational capacity, they assume that scale knowledge requires external supervision and parameter updates. At test time, high-resolution inputs, multi-scale testing, candidate fusion, and image slicing can recover fine details (Najibi et al., 2019; Akyon et al., 2022; Bodla et al., 2017; Solovyev et al., 2021; Kisantal et al., 2019). These methods usually treat the detector as an input-output black box. They merge boxes at the prediction level without examining whether internal discrete coordinates remain synchronized with the actual feature shapes. Consequently, additional pixels do not necessarily become usable evidence for small objects, and dense duplicate candidates may exhaust the limited query budget. Gradient-, occlusion-, and attention-based explanations mainly identify what the model attends to (Selvaraju et al., 2017; Petsiuk et al., 2018; 2021; Abnar & Zuidema, 2020; Chefer et al., 2021a;b). They do not reveal how scale changes alter predictions through internal positional mechanisms, nor do they convert this mechanism into a training-free correction. The common gap is the absence of a mechanism-oriented paradigm for frozen detectors that isolates the relationships among input scale, internal coordinates, and output geometry through controlled variables.

As shown in Fig. 1, we hypothesize that pretrained detectors do not lack scale-related positional knowledge. Instead, this knowledge is embedded in frozen features and decoders but constrained by spatial tensors coupled to preset input shapes. To test this hypothesis, we introduce ACGI. The framework preserves image semantics and all model parameters while synchronously regenerating encoder positional embeddings, decoder anchors, and validity masks from the actual multi-scale feature shapes. This operation constructs a controlled scale-coordinate counterfactual. Cross-grid consistency then compares same-class geometric responses under the baseline and intervened coordinate systems, turning sensitivity to internal spatial representation into an annotation-free explanatory signal. Class-aware geometric deduplication retains stable predictions and complementary small objects. Finally, spatial-compression routing selects the intervention strength from the equivalent linear scale of each image so that scale expansion remains matched to grid reconstruction. The entire procedure requires no gradients, target-domain annotations, or parameter updates. ACGI therefore forms a closed loop of internal coordinate intervention, mechanism-level consistency explanation, and adaptive prediction correction. It transforms scale adaptation from an external engineering enhancement into a testable activation of internal spatial knowledge in a frozen model.

Our main contributions are as follows:

- We formulate scale adaptation in small object detection as a consistency problem between continuous image-scale transformations and the discrete spatial representations of frozen detectors. We further propose the mechanistic hypothesis that scale-related positional knowledge is encoded by pretrained models but suppressed by preset spatial tensors.
- We introduce training-free ACGI, which constructs controlled internal counterfactuals by regenerating spatial tensors from actual feature shapes. It explains spatial prediction stability through cross-grid geometric consistency and converts this signal into performance gains through annotation-free routing and class-aware fusion.
- We demonstrate the generality and small-object advantage of ACGI on multiple frozen DETR baselines across COCO, TinyPerson, and VisDrone2019. Reliability ranking, scale-response, and deletion-faithfulness experiments provide quantitative evidence for the proposed internal spatial mechanism.

2 Related Work

2.1 Small Object Detection

Small object detection methods mainly enhance representations or preserve image details. Representation-enhancement approaches improve low-resolution features through feature pyramids, contextual modeling, scale-aware queries, label assignment, and training objectives (Zhu et al., 2021; Meng et al., 2021; Liu et al., 2022; Li et al., 2022; Zhang et al., 2023; Zhao et al., 2024;

Peng et al., 2025; Huang et al., 2025; Lin et al., 2017; Liu et al., 2018; Kim et al., 2018; Ghiasi et al., 2019; Li et al., 2019; Tan et al., 2020; Wang et al., 2021; Singh & Davis, 2018; Singh et al., 2018; Yang et al., 2022; Xu et al., 2022). Although effective, they generally require architectural modification, retraining, or additional supervision. Detail-preservation approaches instead employ high-resolution inference, multi-scale testing, image slicing, or box fusion (Najibi et al., 2019; Akyon et al., 2022; Bodla et al., 2017; Solovyev et al., 2021; Kisantal et al., 2019). These methods recover fine visual details but mostly combine predictions at the output level without ensuring that internal spatial representations match the changed feature shapes. In contrast, our method directly intervenes in the spatial coordinates of a frozen detector to activate its latent scale-related knowledge without training.

2.2 Spatial Representation Adaptation

Spatial representation adaptation is primarily studied through positional modeling and geometric query priors. Positional methods encode spatial relations using absolute, relative, or feature-dependent embeddings (Carion et al., 2020; Zhu et al., 2021; Meng et al., 2021; Liu et al., 2022; Li et al., 2022; Zhang et al., 2023; Dosovitskiy et al., 2021; Chu et al., 2023; Wang et al., 2021; Liu et al., 2021; Lin et al., 2023). Geometric-prior methods introduce reference points, dynamic anchors, query selection, and iterative refinement to align decoder queries with object locations (Meng et al., 2021; Liu et al., 2022; Li et al., 2022; Zhang et al., 2023; Zhao et al., 2024; Peng et al., 2025; Huang et al., 2025; Gao et al., 2021; Dai et al., 2021; Wang et al., 2022; Chen et al., 2023; Zong et al., 2023; Lin et al., 2023). However, these representations are usually optimized for a preset training scale. Positional interpolation can adapt tensor dimensions, but it does not jointly align positional embeddings, anchors, and validity masks with actual feature shapes. Our method instead synchronously regenerates these coupled spatial tensors during frozen inference and compares predictions before and after intervention. It therefore corrects spatial-grid mismatch rather than learning a new representation.

3 Method

3.1 Problem Formulation and Overall Architecture

Given an input image $x \in \mathbb{R}^{H\times W\times C}$ and a detector F_θ with frozen parameters, let the preset baseline input side length be S_0. At this resolution, the model pre-generates and stores encoder positional embeddings, decoder anchors, and their validity mask. For an intervention side length s, the image is resized using $T_s(x)$. The encoder position set P_s and decoder anchor set A_s are then generated from the actual multi-scale feature shapes. The corresponding detections are

$$\mathcal{D}_s = F_\theta(T_s(x), P_s, A_s) = \left\{ \left(\boldsymbol{b}_i^{(s)}, c_i^{(s)}, p_i^{(s)} \right) \right\}_{i=1}^{Q}, \tag{1}$$

where $(\boldsymbol{b}_i^{(s)}, c_i^{(s)}, p_i^{(s)})$ denote the source-image box, class, and confidence of the i-th detection hypothesis, respectively, and Q is the output-candidate budget. We preserve the semantics of x and the model parameters θ and treat (P_s, A_s, M_s) as controlled internal variables. As illustrated in Fig. 2, ACGI comprises actual-shape positional-grid intervention, cross-grid explanatory consensus, and grid-pressure routing. Its final output is denoted by $\widehat{\mathcal{D}}$.

3.2 Actual-Shape Positional-Grid Intervention

Changing the input scale alters the multi-level feature shapes, causing cached positional embeddings, anchors, and validity masks to become misaligned with the actual grid. ACGI regenerates these coupled spatial tensors without modifying any learnable parameter.

At input scale s, let the l-th multi-scale feature map supplied to the encoder have spatial size $H_l^{(s)} \times W_l^{(s)}$, channel dimension d, and temperature constant t. For grid location (u, v), we regenerate the two-dimensional sinusoidal positional embedding

$$\boldsymbol{P}_l^{(s)}(u, v) = [\sin(u\omega_r), \cos(u\omega_r), \sin(v\omega_r), \cos(v\omega_r)]_{r=0}^{d/4-1}, \tag{2}$$

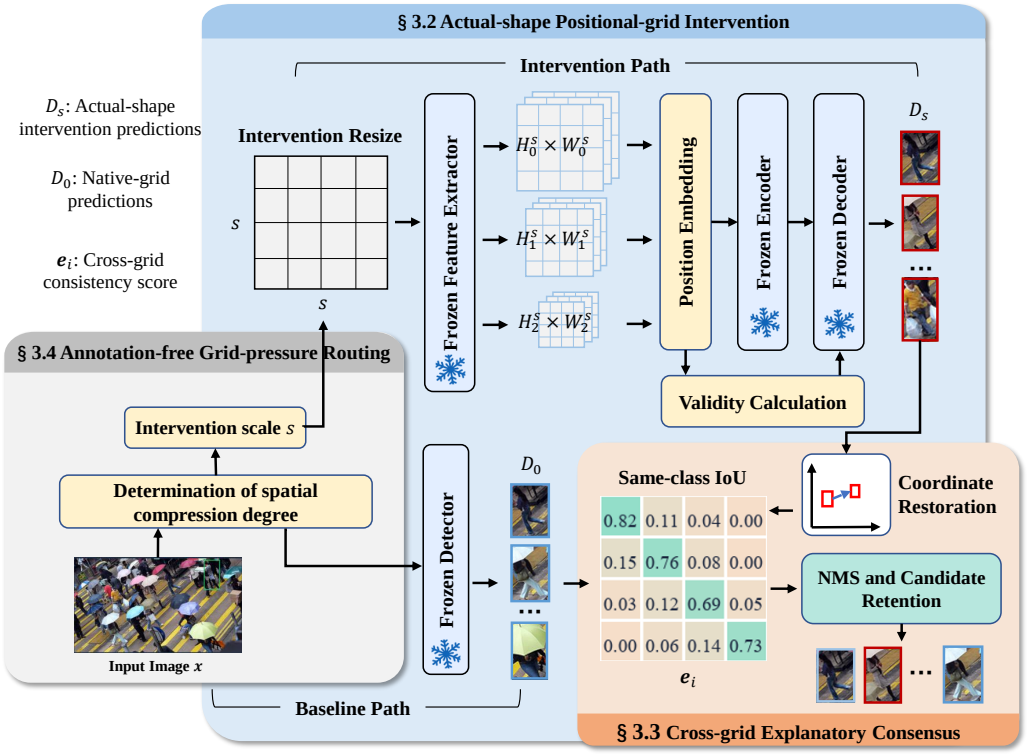

Figure 2: Overview of ACGI. Grid-pressure routing selects an intervention scale s. The baseline and actual-shape intervention paths produce $\mathcal{D}_0$ and $\mathcal{D}_s$, respectively, with the latter regenerating spatial tensors from the actual feature shapes. Their predictions are mapped to the source-image coordinates and fused through same-class IoU matching and class-aware NMS.

where $\omega_r = t^{-4r/d}$. Here, $r = 0, 1, \ldots, d/4 - 1$ is the frequency index. d is the positional-embedding dimension and matches the feature dimension of the Transformer encoder. Following the flattening order of the encoder positional sequence, we establish a normalized location at the center of every grid unit in each multi-scale feature level

$$\boldsymbol{r}_{l,u,v}^{(s)} = \left(\frac{u + \Delta u}{W_l^{(s)}}, \frac{v + \Delta v}{H_l^{(s)}} \right), \tag{3}$$

where $\Delta u = \Delta v = 0.5$ shifts the coordinate from the upper-left corner of a discrete unit to its center. Let κ be the base box scale at the feature level with the highest spatial resolution. The four-dimensional anchor at this location is $\boldsymbol{a}_{l,u,v}^{(s)} = \left[\boldsymbol{r}_{l,u,v}^{(s)}, \kappa 2^l, \kappa 2^l \right]$. To ensure valid anchor parameters and numerical stability under the inverse sigmoid transformation, we assess the normalized center coordinates, width, and height of each anchor componentwise. Its validity is defined as

$$m_{l,u,v}^{(s)} = \prod_{d_a=1}^{4} \mathbf{1}\left[\varepsilon < \boldsymbol{a}_{l,u,v,d_a}^{(s)} < 1 - \varepsilon \right], \tag{4}$$

where ε is the boundary tolerance. Valid anchors are transformed by the inverse sigmoid before encoder-candidate selection and decoder initialization. Because the positional sequence, anchors, and masks all depend on $(H_l^{(s)}, W_l^{(s)})$, they are regenerated synchronously and the cached native state is restored after each forward pass. The frozen decoder and detection head then produce $\mathcal{D}_s$.

3.3 Cross-Grid Explanatory Consensus

Increasing the input resolution allows small objects to occupy more feature units, but it can also alter contextual responses for large objects. The actual-shape branch should therefore not unconditionally replace the baseline branch. This module uses the response discrepancy of the same image under

two internal coordinate systems to measure detection stability and retain complementary information from both branches.

After mapping predictions from both branches back to the source-image coordinate system, we consider the i-th box in the baseline detections $\mathcal{D}_0$. We find all same-class boxes in the actual-shape detections $\mathcal{D}_s$, compute their intersection over union (IoU) with the baseline box, and define the maximum value as the consistency score e_i:

$$\boldsymbol{e}_i = \max_{j:\, c_j^{(s)} = c_i^{(0)}} \mathrm{IoU}\left(\boldsymbol{b}_i^{(0)}, \boldsymbol{b}_j^{(s)}\right). \tag{5}$$

Here, $\boldsymbol{b}_i^{(0)}$ denotes the i-th box in the baseline branch, and $c_i^{(0)}$ is its class. Likewise, $\boldsymbol{b}_j^{(s)}$ denotes the j-th box in the actual-shape branch, and $c_j^{(s)}$ is its class. A larger e_i indicates that the same-class geometric interpretation is more stable under grid intervention. Inference requires no annotations and does not assume one-to-one candidate indices across branches. If the actual-shape branch contains no prediction of the same class, the consistency of that detection is set to zero. This score explains cross-grid stability without ground-truth annotations. We then apply class-aware non-maximum suppression to $\mathcal{D}_0 \cup \mathcal{D}_s$: $\mathcal{C}_\eta(\mathcal{D}_0, \mathcal{D}_s) = \mathrm{TopQ}(\mathrm{ClassNMS}(\mathcal{D}_0 \cup \gamma\mathcal{D}_s, \eta))$, where η is the IoU threshold for same-class boxes, and γ is the score coefficient for the dynamic branch. Overlapping interpretations with high agreement are compressed into one candidate, while complementary small objects discovered by the dynamic branch are retained. $\mathrm{TopQ}(\cdot)$ sorts the candidates after non-maximum suppression (NMS) by confidence and retains only the top Q candidates to satisfy the detector's output budget. $\mathrm{ClassNMS}$ denotes class-aware NMS. This procedure suppresses duplicate predictions of the same object across branches while allowing complementary small-object detections from the actual-shape branch to enter the final output.

3.4 Annotation-Free Grid-Pressure Routing

Images with different source resolutions undergo different levels of spatial compression when mapped to the native input grid. For ordinary or low-resolution images, a large intervention scale mainly introduces interpolated redundancy and may amplify background noise. For dense, high-resolution scenes, a small intervention scale remains insufficient to preserve fine details of small objects. A fixed intervention scale therefore cannot accommodate all images. We address this issue with annotation-free grid-pressure routing, which determines whether to activate the actual-shape intervention and selects its strength from the compression of the source image relative to the native grid. Specifically, we use the square root of the source-image area as an equivalent linear scale and define its ratio to the baseline input side length as an annotation-free spatial-compression measure $\rho(x) = \frac{\sqrt{H_x W_x}}{S_0}$, where H_x, W_x are the original height and width of input image x before detector preprocessing. ρ measures the equivalent linear compression incurred when the source image is resized to the baseline input. Let τ_ρ denote the high-compression threshold. The unified ACGI output is

$$\widehat{\mathcal{D}} = \begin{cases} \mathcal{C}_\eta(\mathcal{D}_0, \mathcal{D}_{s_1}), & \rho \le 1, \\ \mathcal{D}_{s_1}, & 1 < \rho \le \tau_\rho, \\ \mathcal{D}_{s_2}, & \rho > \tau_\rho. \end{cases} \tag{6}$$

Where $\mathcal{D}_0$ is the baseline-branch output. $\mathcal{D}_{s_1}$ is the actual-shape output obtained with the lower intervention side length s_1, and $\mathcal{D}_{s_2}$ is the output obtained with the higher intervention side length s_2. The high spatial-compression threshold is τ_ρ. The compression ratio ρ reflects the linear compression incurred when the source image is mapped to the baseline input size. When $\rho \le 1$, the baseline branch remains well adapted to the image, so $\mathcal{D}_0$ is fused with $\mathcal{D}_{s_1}$. For an image in the medium-compression interval, $\mathcal{D}_{s_1}$ is used as the final result. When $\rho > \tau_\rho$, the higher-scale output $\mathcal{D}_{s_2}$ is selected to alleviate the loss of small-object information caused by resizing a high-resolution image.

4 Experiments

Datasets. We evaluate ACGI on COCO val2017 (Lin et al., 2014), TinyPerson test (Yu et al., 2020), and VisDrone2019-DET val (Zhu et al., 2022). COCO contains 5,000 images and 80 categories and

Table 1: Comparative results on COCO val2017.

Method	AP	AP_{50}	AP_{75}	AP_s
RT-DETR-R50 (Zhao et al., 2024)	53.08	71.23	57.69	34.73
RT-DETR-R50 + ACGI	54.17 (+1.10)	71.86 (+0.64)	59.21 (+1.52)	38.51 (+3.78)
RT-DETRv2-R50 (Lv et al., 2024)	53.40	71.56	57.48	36.06
RT-DETRv2-R50 + ACGI	54.31 (+0.91)	72.19 (+0.63)	58.83 (+1.35)	38.15 (+2.09)
RT-DETRv3-R50 (Wang et al., 2025)	52.72	70.95	57.26	35.15
RT-DETRv3-R50 + ACGI	53.73 (+1.01)	71.58 (+0.62)	58.69 (+1.42)	38.17 (+3.02)
D-FINE-S (Peng et al., 2025)	48.49	65.42	52.57	29.30
D-FINE-S + ACGI	49.59 (+1.11)	66.50 (+1.08)	54.13 (+1.55)	31.97 (+2.66)
DEIM-D-FINE-S (Huang et al., 2025)	48.97	65.88	53.11	30.38
DEIM-D-FINE-S + ACGI	49.76 (+0.79)	66.70 (+0.83)	54.34 (+1.22)	32.62 (+2.24)

Table 2: Cross-domain COCO-style results on TinyPerson test.

Method	AP	AP_{50}	AP_{75}	AP_s
RT-DETR-R50 (Zhao et al., 2024)	9.19	26.24	4.63	5.89
RT-DETR-R50 + ACGI	12.99 (+3.80)	34.78 (+8.54)	7.09 (+2.46)	9.41 (+3.52)
RT-DETRv2-R50 (Lv et al., 2024)	9.45	26.16	4.82	5.97
RT-DETRv2-R50 + ACGI	11.61 (+2.17)	32.10 (+5.94)	5.91 (+1.09)	7.79 (+1.82)
RT-DETRv3-R50 (Wang et al., 2025)	9.12	25.37	4.74	5.80
RT-DETRv3-R50 + ACGI	12.73 (+3.61)	33.88 (+8.51)	7.25 (+2.51)	9.13 (+3.33)
D-FINE-S (Peng et al., 2025)	7.23	20.72	3.67	4.37
D-FINE-S + ACGI	9.76 (+2.52)	27.35 (+6.63)	4.98 (+1.31)	6.77 (+2.40)
DEIM-D-FINE-S (Huang et al., 2025)	7.80	21.89	4.16	4.96
DEIM-D-FINE-S + ACGI	9.79 (+1.99)	27.22 (+5.34)	5.40 (+1.24)	6.71 (+1.76)

is used for in-domain evaluation, ablation, and interpretability analysis. TinyPerson contains 786 images; we merge `sea_person` and `earth_person` into person and exclude uncertain, ignore, and logo regions. VisDrone2019-DET val contains 548 images; we merge pedestrian/people into person, map van to car and motor to motorcycle, and exclude tricycle and awning-tricycle due to the lack of unambiguous COCO counterparts. For both cross-domain datasets, COCO-pretrained models are evaluated without target-domain annotations or fine-tuning.

Evaluation metrics. On COCO, we report the official bounding-box AP, AP_{50}, AP_{75}, and AP_s. TinyPerson and VisDrone2019 are evaluated with the same COCO-style protocol and `maxDets` $= 100$ for paired comparison with each frozen baseline. Thus, our TinyPerson results do not use the official miss-rate metric, and our VisDrone2019 results do not follow the challenge protocol. Cross-domain results should therefore be interpreted as relative gains under identical settings rather than absolute comparisons with target-domain-trained methods.

Implementation details. We use official COCO-pretrained RT-DETR-R50 (Zhao et al., 2024), RT-DETRv2-R50 (Lv et al., 2024), RT-DETRv3-R50 (Wang et al., 2025), D-FINE-S (Peng et al., 2025), and DEIM-D-FINE-S (Huang et al., 2025). All parameters are frozen, and inference is performed without gradients. Because RT-DETRv3 provides only Paddle weights, we convert its primary inference query branch to PyTorch while omitting training-only auxiliary branches; the converted RT-DETRv3 and official RT-DETRv2 checkpoints match all 736 and 744 state entries, respectively. All results are paired comparisons within the same frozen implementation. The native side length is $S_0 = 640$, with intervention scales $s_1 = 800$ and $s_2 = 960$. Positional embeddings, decoder anchors, and validity masks are regenerated from the actual multi-scale feature shapes. We use $d = 256$, temperature $t = 10{,}000$, anchor scale $\kappa = 0.05$, validity tolerance $\varepsilon = 0.01$, and grid-pressure threshold $\tau_\rho = 2.0$. Class-aware NMS uses $\eta = 0.8$ and $\gamma = 1.0$, retaining $Q = 300$ candidates. COCO fuses the 640 and 800 branches, whereas on TinyPerson and VisDrone2019, $\rho = \sqrt{H_x W_x}/S_0$ selects between 800 and 960, yielding selection counts of 76/710 and 529/19, respectively. The converted RT-DETRv3 achieves 52.72 AP on COCO, compared with the reported Paddle result of 53.4; we therefore report paired gains under the unified PyTorch implementation rather than native Paddle reproduction. All experiments use a single NVIDIA RTX 3060 GPU.

Table 3: Cross-domain COCO-style results on VisDrone2019-DET val.

Method	AP	AP_{50}	AP_{75}	AP_s
RT-DETR-R50 (Zhao et al., 2024)	15.11	27.12	14.57	6.16
RT-DETR-R50 + ACGI	17.13 (+2.02)	30.61 (+3.49)	16.37 (+1.80)	7.88 (+1.72)
RT-DETRv2-R50 (Lv et al., 2024)	15.06	27.16	14.43	6.02
RT-DETRv2-R50 + ACGI	17.37 (+2.30)	30.71 (+3.55)	16.73 (+2.30)	8.13 (+2.12)
RT-DETRv3-R50 (Wang et al., 2025)	15.07	26.83	14.59	6.22
RT-DETRv3-R50 + ACGI	17.27 (+2.20)	30.71 (+3.88)	16.70 (+2.11)	7.93 (+1.71)
D-FINE-S (Peng et al., 2025)	12.98	23.55	12.43	5.55
D-FINE-S + ACGI	15.27 (+2.29)	26.82 (+3.27)	14.87 (+2.43)	7.57 (+2.02)
DEIM-D-FINE-S (Huang et al., 2025)	14.49	26.23	13.68	6.22
DEIM-D-FINE-S + ACGI	17.11 (+2.62)	29.93 (+3.70)	16.80 (+3.12)	8.59 (+2.37)

4.1 Comparative Experiments

Tables 1–3 show that ACGI consistently improves all five frozen baselines across in-domain COCO and cross-domain TinyPerson and VisDrone2019. The AP gains are 0.79–1.11, 1.99–3.80, and 2.02–2.62 points on the three datasets, respectively, with particularly strong improvements on small objects. For example, RT-DETRv3-R50 gains 1.01 AP and 3.02 AP_s on COCO, and 3.61 AP and 3.33 AP_s on TinyPerson, while RT-DETRv2-R50 achieves the highest VisDrone2019 AP of 17.37. Improvements at multiple IoU thresholds further indicate better object discovery and localization. Moreover, routing favors the 960 grid on TinyPerson but mainly selects the 800 grid on VisDrone2019, demonstrating that ACGI adapts its spatial intervention to image-specific grid pressure rather than uniformly using the highest resolution. These consistent gains across datasets, training strategies, and decoder designs validate the transferability of the fine-grained spatial capability activated by ACGI.

4.2 Ablation Studies

Table 4: Ablation of ACGI components and parameters on COCO val2017 using RT-DETR-R50. DG and CGC denote dynamic grid and cross-grid consensus, respectively.

DG	Merge S_0	CGC	γ	η	AP	AP_{50}	AP_{75}	AP_s
			–	–	53.08	71.23	57.69	34.73
✓			1.0	–	54.14	72.22	58.69	38.50
✓	✓		1.0	1.0	33.01	43.15	35.91	31.89
✓	✓	✓	1.0	0.7	54.19	72.40	58.90	38.50
✓	✓	✓	0.8	0.8	53.90	71.82	58.99	37.42
✓	✓	✓	1.0	0.8	54.17	71.86	59.21	38.51

Components and parameters. Table 4 shows that the 800 dynamic grid improves AP from 53.08 to 54.14 and AP_s from 34.73 to 38.50. Direct candidate merging reduces AP to 33.01 because duplicates exhaust the output budget, whereas cross-grid consensus restores AP to 54.17 and achieves 38.51 AP_s. Performance is insensitive to η between 0.7 and 0.8, while reducing γ to 0.8 lowers AP and AP_s, supporting equal weighting of the two branches.

Dynamic grid scale. Figure 3 shows that AP and AP_s generally improve as the grid side length increases from the native 640. A side length of 800 achieves the best performance, with 54.14 AP and 38.50 AP_s, improving by 1.07 and 3.77 points, respectively. In contrast, AP_m increases only from 57.74 to 58.67, indicating that small objects drive most of the gain. Adjacent scales such as 736 and 768 also outperform the native setting, confirming that the improvement is not specific to a single scale. Together with Table 4, these results show that denser grids enhance small-object localization, while cross-grid consensus suppresses duplicates and preserves complementary native-grid evidence.

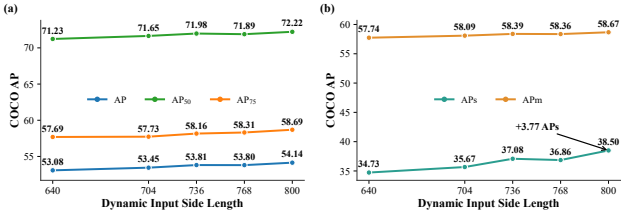

Figure 3: Dynamic-grid scale response of RT-DETR-R50 on the full COCO val2017 set. (a) Overall AP, AP_{50}, and AP_{75}. (b) AP_s and AP_m. Each point is annotated with its value.

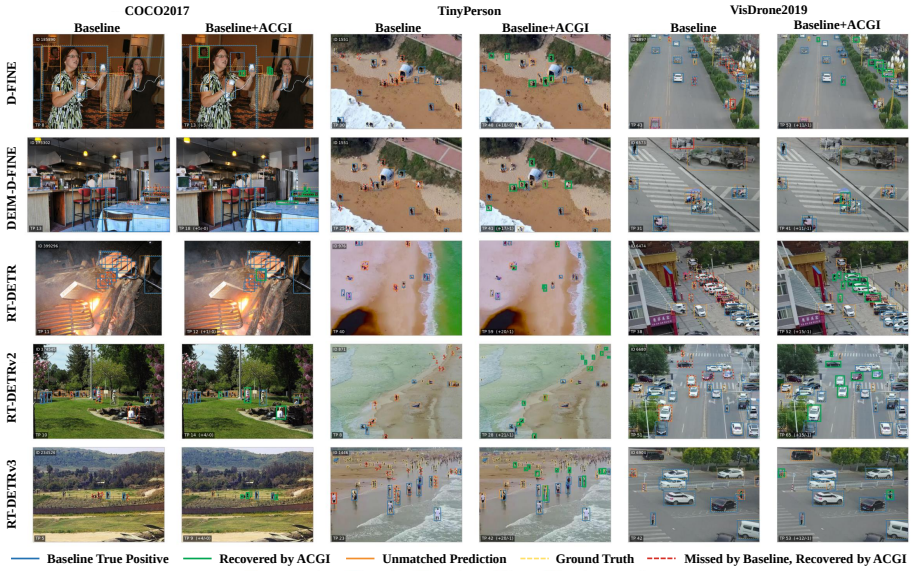

Figure 4: Qualitative comparison across detectors and datasets. ACGI consistently recovers missed small objects while introducing few background responses.

4.3 Qualitative Experiments

Figure 4 compares five baselines with ACGI on COCO, TinyPerson, and VisDrone2019. Across models and datasets, ACGI recovers small, dense, and weakly supported objects while preserving reliable baseline detections and introducing few background responses. The improvement is particularly evident on the tiny-object-dominated TinyPerson and VisDrone2019 datasets. These results suggest that grid alignment improves recall without unconstrained candidate proliferation, demonstrating model generality and cross-dataset robustness.

4.4 Mechanistic Interpretability and Analysis of Positional-Grid Intervention

Objective and protocol. We test whether preset positional grids constrain scale-related spatial knowledge in a frozen RT-DETR-R50 through five linked stages: mismatch diagnosis, controlled intervention, response observation, positive control, and evidence deletion. All experiments use the full COCO val2017 set. Cross-grid support and grid pressure are computed without annotations, which are used only for post hoc evaluation. Each stage converts the preceding observation into a testable variable, forming a unified mechanistic analysis.

Positional-mismatch diagnosis. We evaluate cross-grid consistency on 1.5 million native detection hypotheses. As shown in Fig. 5(a), higher consistency strongly correlates with greater GT overlap, yielding AUCs of 0.720 and 0.810 at IoU thresholds of 0.50 and 0.75, respectively. Fig. 5(b) shows that small objects have the lowest mean consistency (0.578 versus 0.693 and 0.828 for medium and large objects) but receive the largest GT-IoU gain from the dynamic grid (+0.0287 versus +0.0168 and +0.0089). Thus, small objects are the least spatially stable yet the most correctable.

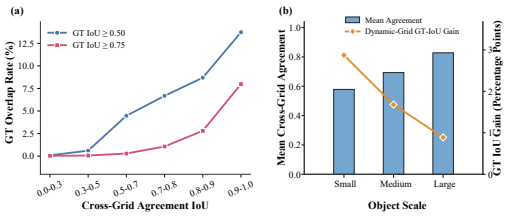

Figure 5: Positional-mismatch diagnosis on the full COCO val2017 set. (a) Cross-grid support and GT localization reliability. (b) Cross-grid stability and dynamic-grid localization gains at different object scales. GT is used only for post hoc validation.

Controlled positional intervention. We regenerate spatial tensors from the actual feature shapes while fixing image semantics, model weights, and candidate budget. Figure 3 shows consistent gains across adjacent scales, particularly for small objects, while Table 4 confirms that cross-grid consensus suppresses duplicates and preserves com-

plementary evidence. Thus, actual-shape alignment converts the diagnosed spatial instability into detection gains rather than merely exploiting a favorable scale.

Internal-response observation. Figure 6 visualizes channel-aggregated responses from the highest-resolution encoded feature. Increasing the grid from 80×80 to 100×100 adds 56.25% more spatial samples and redistributes responses around small objects and their boundaries. In the illustrated image, both the object-to-background response ratio and area-normalized response energy increase slightly. This example provides qualitative evidence of internal response redistribution rather than implying universally increased target activation; the mechanism is primarily supported by the full-set geometric and behavioral analyses.

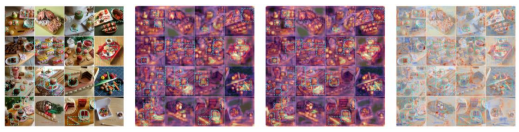

Figure 6: High-resolution feature responses in RT-DETR. From left to right: input with small-object annotations, responses on the native 640 and actual-shape 800 grids, and their difference (red: enhancement; blue: suppression).

Explanation-guided positive control. We test whether cross-grid diagnostics can guide intervention under fixed budgets. At 40% and 60% coverage, images are selected by cross-grid support, grid pressure, or random routing. As shown in Fig. 7, high-support-first routing achieves 53.48 and 53.83 AP and outperforms random routing in AP and AP_{75}, whereas high-pressure-first routing yields the strongest small-object performance, reaching 36.70 and 37.63 AP_{s}. These complementary effects show that support identifies reliable localization evidence, while pressure targets images requiring stronger scale compensation.

Reverse destruction of key evidence. We perform confidence-stratified deletion by removing high-consistency, low-consistency, or random candidates under identical budgets (375,000 candidates each). The removed candidates have closely matched score distributions (mean: 0.0720–0.0759; median: approximately 0.0395). As shown in Fig. 8, deleting high-consistency candidates reduces AP to 36.25, compared with 45.25 for low-consistency deletion and 41.41 for random deletion. This 8.99-point gap confirms that cross-grid consistency captures causal importance beyond classification confidence.

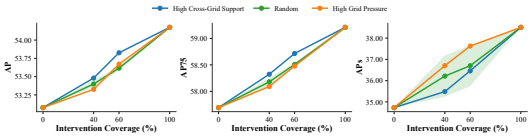

Figure 7: Explanation-guided intervention on COCO val2017. Prioritizing cross-grid support improves AP and AP_{75}, while grid pressure favors AP_{s}. Shaded regions indicate variation over two random-selection runs.

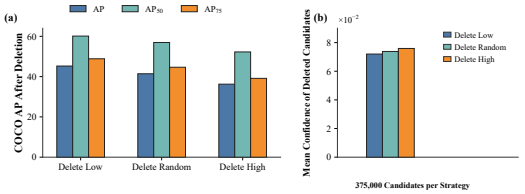

Figure 8: Confidence-matched reverse deletion on COCO val2017. AP is evaluated under an identical deletion budget, and each strategy removes 375,000 candidates.

Closed-loop conclusion. Together, the experiments establish a closed mechanistic chain: cross-grid consistency diagnoses small-object instability, actual-shape alignment provides corrective spatial evidence, support-guided routing predicts beneficial interventions, and removing high-support candidates causes the greatest degradation. At full coverage, ACGI improves AP from 53.08 to 54.17, AP_{75} from 57.69 to 59.21, and AP_{s} from 34.73 to 38.51. Rather than implying that higher resolution universally strengthens activation, these results reveal a latent scale-related positional mechanism in frozen detectors that can be identified and activated through grid alignment.

5 Conclusion

We introduced ACGI, a training-free framework that activates latent scale-related spatial knowledge in frozen detectors through actual-shape grid alignment, cross-grid consistency, and annotation-free routing. Experiments across multiple detectors and datasets demonstrate consistent improvements, particularly for small objects, while mechanistic analyses validate the role of positional-grid alignment.

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

Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4190–4197, 2020.

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

Navaneeth Bodla, Bharat Singh, Rama Chellappa, and Larry S. Davis. Soft-NMS—improving object detection with one line of code. In *IEEE International Conference on Computer Vision*, pp. 5561–5569, 2017.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pp. 213–229, 2020.

Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 782–791, 2021a.

Hila Chefer, Shir Gur, and Lior Wolf. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 397–406, 2021b.

Qiang Chen, Xiaokang Chen, Jian Wang, Shan Zhang, Kun Yao, Haocheng Feng, Junyu Han, Errui Ding, Gang Zeng, and Jingdong Wang. Group DETR: Fast DETR training with group-wise one-to-many assignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6633–6642, 2023.

Xiangxiang Chu, Zhi Tian, Bo Zhang, Xinlong Wang, and Chunhua Shen. Conditional positional encodings for vision transformers. In *International Conference on Learning Representations*, 2023.

Xiyang Dai, Yinpeng Chen, Jianwei Yang, Pengchuan Zhang, Lu Yuan, and Lei Zhang. Dynamic DETR: End-to-end object detection with dynamic attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2988–2997, 2021.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16×16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Peng Gao, Minghang Zheng, Xiaogang Wang, Jifeng Dai, and Hongsheng Li. Fast convergence of DETR with spatially modulated co-attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3621–3630, 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.

Shihua Huang, Zhichao Lu, Xiaodong Cun, Yongjun Yu, Xiao Zhou, and Xi Shen. DEIM: DETR with improved matching for fast convergence. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15162–15171, 2025.

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.

Mate Kisantal, Zbigniew Wojna, Jakub Murawski, Jacek Naruniec, and Kyunghyun Cho. Augmentation for small object detection. *arXiv preprint arXiv:1902.07296*, 2019.

Feng Li, Hao Zhang, Shilong Liu, Jian Guo, Lionel M. Ni, and Lei Zhang. DN-DETR: Accelerate DETR training by introducing query denoising. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13619–13627, 2022.

Yanghao Li, Yuntao Chen, Naiyan Wang, and Zhaoxiang Zhang. Scale-aware trident networks for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6054–6063, 2019.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, pp. 740–755, 2014.

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

Yutong Lin, Yuhui Yuan, Zheng Zhang, Chen Li, Nanning Zheng, and Han Hu. Detr doesn't need multi-scale or locality design. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6545–6554, 2023.

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

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.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022, 2021.

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

Depu Meng, Xiaokang Chen, Zejia Fan, Gang Zeng, Houqiang Li, Yuhui Yuan, Lei Sun, and Jingdong Wang. Conditional DETR for fast training convergence. In *IEEE/CVF International Conference on Computer Vision*, pp. 3651–3660, 2021.

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

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

Vitali Petsiuk, Abir Das, and Kate Saenko. RISE: Randomized input sampling for explanation of black-box models. In *British Machine Vision Conference*, 2018.

Vitali Petsiuk, Rajiv Jain, Varun Manjunatha, Vlad I. Morariu, Ashutosh Mehra, Vicente Ordonez, and Kate Saenko. Black-box explanation of object detectors via saliency maps. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11443–11452, 2021.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *IEEE International Conference on Computer Vision*, pp. 618–626, 2017.

Bharat Singh and Larry S. Davis. An analysis of scale invariance in object detection—SNIP. In *IEEE/CVF 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 *Advances in Neural Information Processing Systems*, volume 31, 2018.

Roman Solovyev, Weimin Wang, and Tatiana Gabruseva. Weighted boxes fusion: Ensembling boxes from different object detection models. *Image and Vision Computing*, 107:104117, 2021.

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. 10781–10790, 2020.

Shuo Wang, Chunlong Xia, Feng Lv, and Yifeng Shi. RT-DETRv3: Real-time end-to-end object detection with hierarchical dense positive supervision. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1628–1636, 2025.

Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 568–578, 2021.

Yingming Wang, Xiangyu Zhang, Tong Yang, and Jian Sun. Anchor DETR: Query design for transformer-based detector. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 2567–2575, 2022. doi: 10.1609/aaai.v36i3.20158.

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 *European Conference on Computer Vision*, pp. 526–543, 2022.

Chenhongyi Yang, Zehao Huang, and Naiyan Wang. QueryDet: Cascaded sparse query for accelerating high-resolution small object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13668–13677, 2022.

Xuehui Yu, Yuqi Gong, Nan Jiang, Qixiang Ye, and Zhenjun Han. Scale match for tiny person detection. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1257–1265, 2020.

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

Yian Zhao, Wenyu Lv, Shangliang Xu, Jinman Wei, Guanzhong Wang, Qingqing Dang, Yi Liu, and Jie Chen. DETRs beat YOLOs on real-time object detection. In *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. volume 44, pp. 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.

Zhuofan Zong, Guanglu Song, and Yu Liu. DETRs with collaborative hybrid assignments training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6748–6758, 2023.