


# RINGCO-DETR: INFERENCE-NEUTRAL STABILIZATION OF ORIENTED TRANSFORMER DETECTION FOR AERIAL SENSING

Jianqin Gao[1], Tianqi Wang[2], Hongyi Gao[3], Yujie Ding[4] & Zihao Wang[5]

[1 3 4 5]The University of Sydney
[2]Taiyuan University of Technology

## Abstract

Oriented object detection in aerial imagery supports safety-critical aerospace perception, including UAV-based low-altitude patrol and airfield surface monitoring. Without oriented localization, axis-aligned boxes can overestimate occupied regions and confuse association in dense, rotated scenes, increasing missed hazards and false alarms. Yet DETR-style set prediction can be fragile when matching costs drift across classification, box, angle, and IoU. We propose RingCo-DETR, a training-only, inference-neutral framework with: (i) Adaptive Hungarian Matching, which self-balances heterogeneous costs via EMA scaling and quantile normalization and adds bounded IoU-ring auxiliaries for denser supervision; and (ii) F-RIPE, a frequency-ring consistency loss that regularizes features under platform attitude perturbations. On DOTA-v1.5 and HIT-UAV, RingCo-DETR reaches 81.7/83.1 mAP@50 and improves training stability and recall over representative NMS-free and NMS-based baselines.

**Keywords:** Aerial vision, Oriented object detection, DETR, Adaptive Hungarian matching, Frequency-domain consistency, UAV safety

## 1. Introduction

Oriented object detection in aerial imagery underpins safety-critical aerospace operations, including unmanned aerial vehicle (UAV) autonomy and airfield surface monitoring [1, 2]. Accurate orientation estimation is essential for determining the true occupied footprint and separation margins of rotated targets such as aircraft, ground-service vehicles, and infrastructure, thereby improving situational awareness for small or elongated hazards that axis-aligned bounding boxes would otherwise inflate or miss [3].

The safety risks are particularly acute in airport surface operations. Runway incursions and foreign object debris (FOD) remain leading contributors to aviation accidents and near-miss events. The U.S. Department of Transportation Office of Inspector General recently emphasized that improved surface situational awareness, achieved through advanced sensor systems and automated detection, is a key component in mitigating runway-related incidents. Traditional surface surveillance technologies such as Airport Surface Detection Equipment Model X (ASDE-X) and Automatic Dependent Surveillance-Broadcast (ADS-B) provide controller-centric traffic displays, yet they are costly to install, limited in coverage, and do not directly resolve small-scale hazards such as debris, wildlife, or unauthorized personnel at the pixel level [4]. UAV-mounted high-resolution imaging platforms offer a complementary, rapidly deployable means of dynamic runway and apron surveillance, but their operational value hinges on robust computer vision algorithms capable of detecting and localizing arbitrarily oriented objects in cluttered, rotated, and densely packed scenes [4, 5].

Early convolutional neural network (CNN) detectors such as R-CNN [6], Faster R-CNN [7], and RetinaNet [8], along with the You Only Look Once (YOLO) family, remain widely used baselines. Starting from YOLOv8, the YOLO series incorporated ProbIoU to enhance oriented detection, and YOLOv11

further strengthened the approach with more thorough engineering. As transformers [9] were adapted to vision tasks [10], the Detection Transformer (DETR) [11] emerged, reformulating detection as direct set prediction, removing non-maximum suppression (NMS) via end-to-end Hungarian matching, and enabling learnable queries. Subsequent variants improved horizontal detection through better feature sampling, query initialization, and training dynamics, e.g., Deformable DETR [12], Conditional DETR [13], and DETR with Improved DeNoising Anchor Boxes (DINO) [14].

Frequency-domain learning has proven effective [15]: frequency-domain representations can capture cues that spatial features may miss under rotation and other invariances [16]. Extending to oriented detection raises specific challenges: matching must respect angular periodicity and spatial crowding; the costs for classification, box, angle, and IoU are heterogeneous, drift during training, and are often assigned fixed weights; and strict one-to-one assignment under-supervises dense scenes, while queries depicting similar appearances at different orientations are insufficiently coordinated.

Transformer-based detectors, notably DETR and its variants, offer an end-to-end set-prediction formulation that eliminates hand-crafted post-processing such as NMS [11]. This NMS-free pipeline is especially attractive for safety-relevant airborne embedded systems where deterministic latency and simplified deployment are paramount. However, oriented extensions of DETR are sensitive to heterogeneous cost scales—classification, box regression, angle, and IoU terms evolve at disparate magnitudes—so early-training outliers may cause one term to dominate Hungarian assignments, leading to unstable matching and under-supervision in crowded aerial scenes [17, 18, 14]. Moreover, rotation-induced feature drift can degrade recall for small or densely packed targets, a failure mode that is unacceptable when missed detections correspond to FOD or runway incursions.

We introduce RingCo-DETR, an *inference-neutral* training framework that stabilizes oriented DETR for aerospace perception tasks. Our contributions are threefold:

1. Adaptive Hungarian Matching (AHM), a training-only matcher that self-balances heterogeneous matching costs via EMA-tracked statistics and rank-preserving quantile normalization, and injects bounded auxiliary supervision to remedy the insufficient supervision that strict one-to-one assignment provides in dense scenes, all without altering the inference pipeline.

2. F-RIPE (Frequency-domain Ring-Invariant Profile Enhancement), a training-only spectral consistency loss that aligns equal-energy frequency-ring profiles between an image and its randomly rotated augmentation, suppressing high-frequency feature drift induced by platform attitude perturbations.

3. Evaluation on safety-relevant benchmarks. On DOTA-v1.5 and HIT-UAV, RingCo-DETR achieves 81.7/83.1 mAP@50 with no inference-time modification, validating its effectiveness across optical and infrared aerial sensing modalities relevant to airfield monitoring and UAV-based low-altitude patrol.

We adopt an inference-neutral design: all regularization is confined to the training phase and removed from the forward graph at inference. This matters for safety-critical airborne applications, where extra computation or non-deterministic behavior in auxiliary branches could violate latency budgets and certification requirements for embedded perception systems [19]. As illustrated in Figure 1, RingCo-DETR has three training-time stages: (i) a spectral consistency branch (F-RIPE, upper-left) that operates on augmented views in parallel to the main pipeline; (ii) an adaptive training module (upper-right) that adjusts Hungarian matching via EMA-guided cost balancing and bounded IoU-ring auxiliary assignments under bipartite and cross-query constraints; and (iii) a standard DETR encoder–decoder detector (lower part) whose architecture is unchanged at deployment. The inference graph is therefore identical to a vanilla oriented DETR, while the training graph carries the multi-objective stabilization.

## 2. Methodology

Figure 1 shows the overall architecture of RingCo-DETR. The framework ingests aerial images together with their augmented counterparts—flipped and rotated to mimic the platform attitude perturbations typical of UAV and airborne sensing—through a shared ResNet backbone and a multi-scale

transformer encoder that produces multi-scale feature maps. Parallel to the main detection stream, the F-RIPE module performs spectral decomposition on backbone features and extracts equal-energy frequency rings to enforce rotation-invariant ring-profile consistency (Figure 3a). Within the RingCo detector, dual query streams, namely cooperative queries $Q_{\text{coop}}$ and competing positional queries $Q_{\text{comp}}$, interact through cooperative self-attention and competing positional self-attention layers, followed by cross-attention to the multi-scale feature maps, so that the cooperative branch yields a query-alignment loss $L_{\text{coop}}$ and the competing branch a positional regularization loss $L_{\text{comp}}$. Query predictions (boxes and logits) are then passed to the adaptive training module, where Adaptive Hungarian Matching (Figure 3b) establishes bipartite correspondences between predictions and ground truth under EMA-scaled, quantile-normalized composite costs with cross-query constraints, supplemented by bounded IoU-ring auxiliary matches. The total objective $L = L_{\text{match}} + L_{\text{fripe}} + L_{\text{comp}}$ aggregates the primary matching loss, the competing-query loss, and the spectral consistency loss; all auxiliary branches are detached during inference, leaving only the encoder–decoder pathway active so as to meet the determinism requirements of airborne embedded deployment.

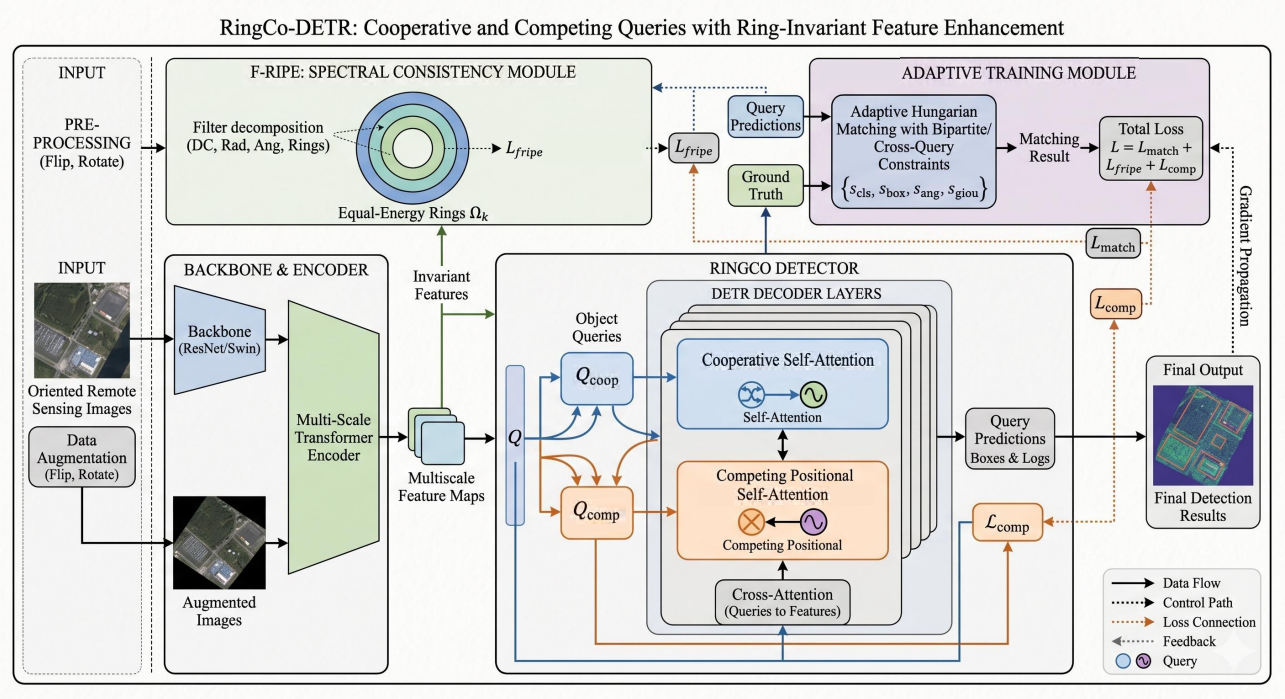



Figure 1 – Overall architecture of RingCo-DETR. The training graph carries F-RIPE spectral consistency, cooperative/competing query coordination, and adaptive matching to address attitude perturbations and dense aerial layouts; the inference graph reduces to a standard oriented DETR encoder–decoder, preserving deterministic latency for airborne embedded deployment.

## 2.1 Adaptive Hungarian Matching

Oriented detection in aerospace monitoring faces two geometrically distinct regimes that dominate our target applications. The first is airport surface monitoring, where aircraft, ground-service vehicles, and infrastructure are densely arranged in near-parallel formations on aprons and taxiways, with object scales spanning from large airliners down to small FOD. The second is UAV-based low-altitude patrol, in both visible-light and infrared conditions, where safety-critical targets such as personnel, vehicles, and intruders occupy small image footprints under varying platform attitudes. The two regimes induce markedly different distributions over the four heterogeneous matching costs: classification $c_{\text{cls}}$, box regression $c_{\text{box}}$, angle $c_{\text{angle}}$, and IoU $c_{\text{IoU}}$. surface monitoring is dominated by $c_{\text{box}}$ and $c_{\text{IoU}}$ because of dense layouts, with small targets producing low but high-variance $c_{\text{IoU}}$, whereas low-altitude patrol elevates $c_{\text{angle}}$, rendering the angle cost more discriminative because attitude perturbations rotate target appearances. Under fixed-weight Hungarian matching, either regime can let a single cost dominate the bipartite assignment, causing early-training drift and systematic missed detec-

tions, an outcome that directly compromises operational safety when the missed objects correspond to FOD or runway incursions [3]. In contrast to one-to-many assignment and denoising-based DETR training [17, 18, 14], we retain the standard one-to-one Hungarian formulation to preserve inference efficiency and inject only bounded auxiliary supervision during training.

The matching pipeline is visualized in Figure 3(b). Given predicted boxes and classification logits from the decoder, a cost-computation block first evaluates the four heterogeneous terms: classification confidence, generalized IoU (GIoU), axis-aligned $L_1$ box distance, and oriented angle difference. These raw costs pass through a normalization stage and an adaptive EMA scaling stage that maintains running estimates of per-component magnitudes and derives bounded scale factors $\{s_{\text{cls}}, s_{\text{box}}, s_{\text{ang}}, s_{\text{giou}}\}$, preventing any single term from dominating the composite cost matrix $\mathbf{C}$. The combined cost is then routed through two parallel pathways: Hungarian assignment, which produces final one-to-one primary matches $\mathscr{P}$, and auxiliary selection, which retrieves IoU-ring candidates whose IoU with each target $t_j$ falls within an adaptive ring $[r_{\min}(j), r_{\max,t}(j)]$ and keeps the top-$K$ auxiliary matches $\mathscr{A}_j$. The final result fuses primary and auxiliary pairs, with auxiliary losses temperature-scaled and capped at $\lambda_{\text{aux}} = 0.3$ so that primary assignments remain dominant. At inference, the auxiliary branch is discarded and only the primary Hungarian pathway is retained, so the computational graph reverts to the standard DETR matching protocol with no added latency.

For each component $k \in \{\text{cls}, \text{box}, \text{angle}, \text{IoU}\}$ we track the mini-batch mean $\bar{c}_{k,t}$ and derive a bounded scale $s_{k,t}$,

$$\bar{c}_{k,t} = \alpha \bar{c}_{k,t-1} + (1-\alpha)\mathbb{E}_b[c_{k,t}], \quad s_{k,t} = \text{clip}\left(\frac{\tau}{\bar{c}_{k,t}+\varepsilon}, 0.25, 4.0\right), \tag{1}$$

where $\mathbb{E}_b[\cdot]$ averages over the current mini-batch. The clip range $[0.25, 4.0]$ reflects the practical range of scene transitions encountered in aerial monitoring—from open runways and high-altitude cruise views with sparse targets and small $c_{\text{IoU}}$, to crowded apron and close-range UAV views with dense targets and large $c_{\text{box}}$, and from optical to thermal-infrared modalities. Across these scenarios, cost magnitudes can vary by an order of magnitude. The EMA momentum $\alpha = 0.95$ interpolates smoothly between regimes, allowing the matcher to operate across mission phases and sensor modalities without per-condition retuning, an important property for deployed systems whose field of view shifts continuously during flight.

To prevent rare high-cost targets from being suppressed as outliers, we normalize each component by its high quantile $Q_q(\cdot)$ over the mini-batch,

$$\mathbf{C} = \sum_k s_{k,t} \, \frac{\mathbf{C}_k}{Q_q(\mathbf{C}_k) + \varepsilon}. \tag{2}$$

In airfield surveillance, safety consequence is inversely correlated with target frequency: a single piece of FOD or a single unauthorized intruder on an otherwise empty runway is a statistical outlier, yet missing it can be catastrophic. Conventional mean/variance normalization would dampen the gradient contribution of such targets, whereas the rank-preserving, outlier-capped quantile normalization with $q = 0.95$ compresses only the extreme magnitudes while preserving the relative ordering of rare samples, so the Hungarian solver still assigns them to appropriate queries. This directly supports the low-probability, high-consequence detection regime central to aerospace safety monitoring.

Strict one-to-one matching under-supervises crowded scenes, so we increase supervision density with bounded IoU-ring auxiliaries. For each target $t_j$ we define an IoU-ring pool

$$\hat{\mathscr{R}}_j = \left\{ i \mid r_{\min}(j) \le \text{IoU}(b_i, t_j) \le r_{\max,t}(j) \right\}, \tag{3}$$

where $r_{\min}(j) = \max(0.3, Q_j^{\text{IoU}}(0.5))$ and $r_{\max,t}(j)$ is an EMA-smoothed high IoU quantile with conservative clipping. The cross-query constraint (see Bipartite/Cross-Query Constraints in Figure 1) further prevents auxiliary candidates from a single target cluster from redundantly covering the same predicted query, encoding a geometric prior of dense aerial formations such as aircraft parked in herringbone or parallel patterns, ground-service vehicles queued along taxiways, equally spaced runway lighting, and linear roadway convoys. The capacity cap $K = 3$ ensures that even in the most congested apron scenes the primary matching remains the dominant loss component. The final loss

keeps primary matches dominant while auxiliary pairs are temperature-scaled ($T_0 = 0.5$) and capped by $\lambda_{\text{aux}} = 0.3$,

$$\mathscr{L}_{\text{match}} = \sum_{(i,j)\in\mathscr{P}} \ell(i,j) + \lambda_{\text{aux}} \sum_{j} \sum_{i\in\mathscr{A}_j} w_{ij}\,\ell(i,j), \tag{4}$$

with $w_{ij} = \text{softmax}_{i\in\mathscr{A}_j}(-\mathbf{C}_{ij}/T)$. All auxiliary pathways are stripped at inference, leaving the original encoder–decoder latency unchanged.

## 2.2 F-RIPE: Frequency-Ring Consistency Under Platform Attitude Perturbations

Attitude perturbations are inherent to airborne monitoring platforms. Fixed-wing UAVs roll and pitch under gusts, multirotor platforms undergo yaw drift during hovering, and gimbal-mounted sensors continuously reorient to track ground targets. These motions induce global image rotations and cause the same physical object to appear at different orientations across frames and missions, although such angular variation is largely irrelevant to its semantic category. This effect is particularly harmful for small and elongated targets in aerial sensing, such as pedestrians, vehicles, aircraft parts, and debris, whose discriminative cues are often carried by high-frequency contours and fine structures. Conventional rotation augmentation encourages pixel-level consistency, but it does not explicitly constrain the frequency-domain redistribution caused by platform-level rotations. As visualized in Figure 2, global rotations produce substantially stronger high-frequency drift across multiple frequency rings than local object rotations, indicating that the dominant disturbance is a scene-level spectral shift rather than a purely local spatial deformation. F-RIPE addresses this failure mode by enforcing consistency over equal-energy frequency rings $\Omega_k$. Since 2D rotation mainly redistributes Fourier responses along angular directions while preserving the radial energy profile, ring-wise consistency provides a rotation-tolerant constraint that suppresses attitude-induced high-frequency drift without sacrificing the spatial discriminability required for oriented detection.

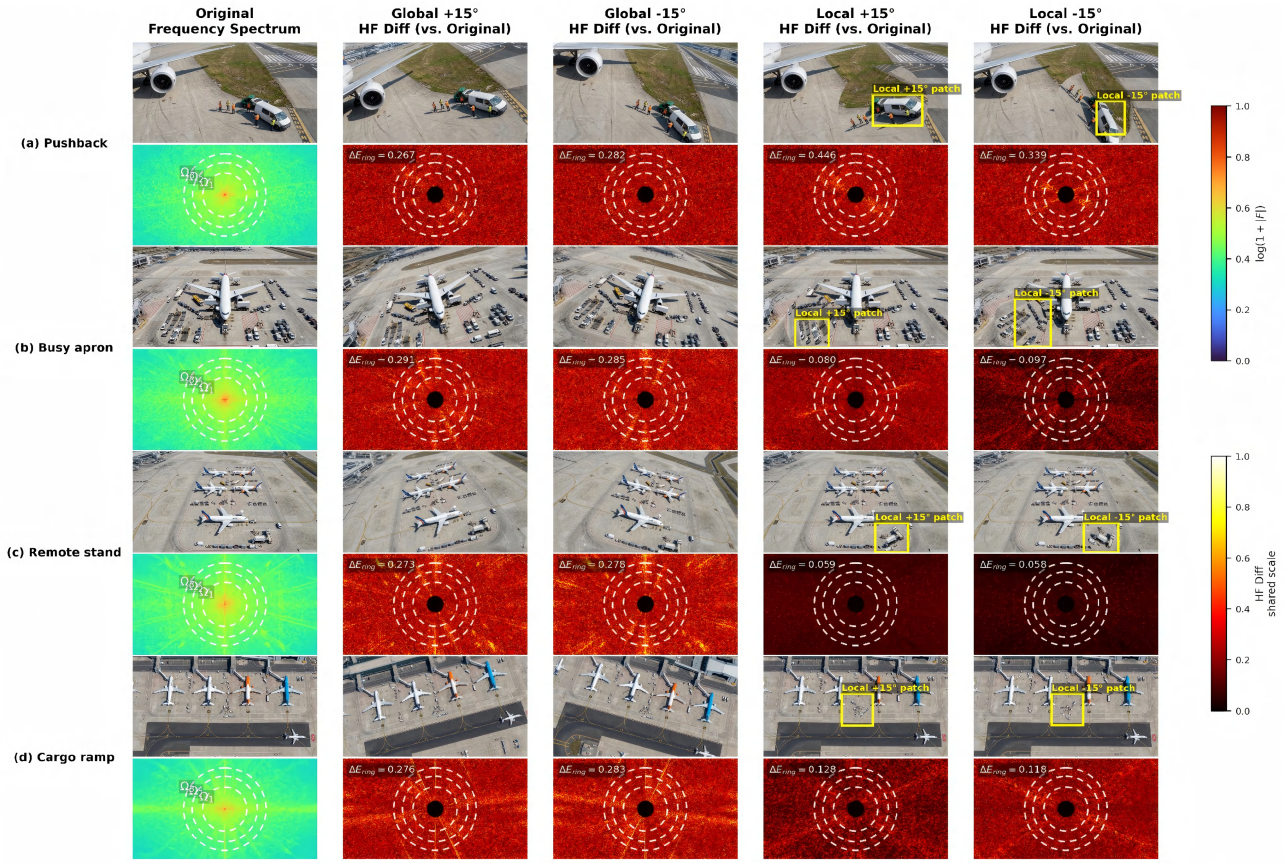



Figure 2 – Frequency-domain motivation for F-RIPE.

The module is detailed in Figure 3(a). For a backbone feature map $X \in \mathbb{R}^{C\times H\times W}$ extracted from an aerial image and its randomly rotated counterpart $X_{\text{aug}}$, the module first computes centered discrete

Fourier transform (DFT) magnitudes $A$ and $A_{\text{aug}}$. A spectral decomposition stage partitions the frequency domain via multi-scale filter components $\Phi_k^j \in \{1\times 1, 3\times 3, 5\times 5\}$ encompassing DC, radial (Rad), angular (Ang), and ring (Rings) channels. From the pooled average spectrum $\bar{A} = \frac{1}{2}(A + A_{\text{aug}})$, equal-energy rings $\{\Omega_k\}$ are constructed by accumulating spectral energy until at least $95\%$ of the cumulative energy is captured, yielding the smallest effective ring set $K_{\text{eff}}$. Per-filter ring-energy profiles $E_k^j$ and $E_k^{\text{aug},j}$ are then obtained by averaging log-compressed magnitudes within each ring across all channels, and the F-RIPE loss penalizes the squared discrepancy between these profiles, suppressing rotation-induced high-frequency drift without imposing pixel-level alignment. A parallel query-processing branch (right side of Figure 3a) discretizes decoder queries $q_i$ into angular bins $f_i$ parameterized by polar angle $\phi$, providing the angular coordination signals $L_{\text{coop}}$ and $L_{\text{comp}}$ for the cooperative/competing query streams. Both pathways are disabled during inference, so the DFT computation and angular binning incur no forward-pass overhead.

Formally, given the centered DFT magnitudes [20] $A$ and $A_{\text{aug}}$ and their pooled average $\bar{A} = \frac{1}{2}(A + A_{\text{aug}})$, we construct equal-energy rings $\{\Omega_k\}$ on $\bar{A}$ and keep the smallest $K_{\text{eff}}$ covering at least $95\%$ cumulative energy. For each filter component $j$ we form the ring-energy profiles

$$E_k^j = \frac{1}{C\,|\Omega_k|} \sum_{c=1}^{C} \sum_{(u,v)\in\Omega_k} \log\bigl(1 + A_c^j(u,v)\bigr), \tag{5}$$

$$E_k^{\text{aug},j} = \frac{1}{C\,|\Omega_k|} \sum_{c=1}^{C} \sum_{(u,v)\in\Omega_k} \log\bigl(1 + A_{\text{aug},c}^j(u,v)\bigr), \tag{6}$$

and define

$$L_{\text{fripe}} = \frac{1}{K_{\text{eff}}} \sum_{j,k\in\Omega_k} \bigl(E_k^j - E_k^{\text{aug},j}\bigr)^2. \tag{7}$$

F-RIPE is applied during training only; enforcing ring consistency suppresses rotation-induced high-frequency drift [21].

### 2.2.1 Final Objective and Airborne Deployment Considerations

Let $L_{\text{match}}$ denote the matching loss from Adaptive Hungarian Matching and $L_{\text{comp}}$ the competing-query positional loss. The total training objective combines all three components, with $L_{\text{fripe}}$ balanced by batch-median normalization (denominator detached),

$$L_{\text{total}} = L_{\text{match}} + L_{\text{comp}} + \lambda_{\text{fripe}} \frac{L_{\text{fripe}}}{\text{median}(L_{\text{fripe}}) + \varepsilon}, \tag{8}$$

where $\varepsilon$ is a stabilizing constant for numerical stability; this corresponds to the loss $L = L_{\text{match}} + L_{\text{fripe}} + L_{\text{comp}}$ shown in Figure 1.

Because F-RIPE, the auxiliary matching pathway, and the cooperative/competing query losses are all stripped at inference, the deployed forward graph is identical to a standard oriented DETR, which matters for airborne perception in three ways. First, deterministic latency: airworthiness software standards such as DO-178C require bounded inference-time variance, and any training-time branch that remained in the inference graph would inject non-determinism into per-frame latency. Second, model auditability: certification review must verify the deployed model layer by layer, and an inference-neutral approach keeps the auditable structure identical to the DETR baseline, introducing no additional certification burden. Third, resource constraints: UAV-grade computing platforms such as Jetson-class embedded GPUs operates under tight power and memory budgets, and any extra inference-time operator would compete for these scarce resources. The inference-neutrality of RingCo-DETR therefore ensures that training-stage stabilization gains transfer to deployment without incurring any of these costs.

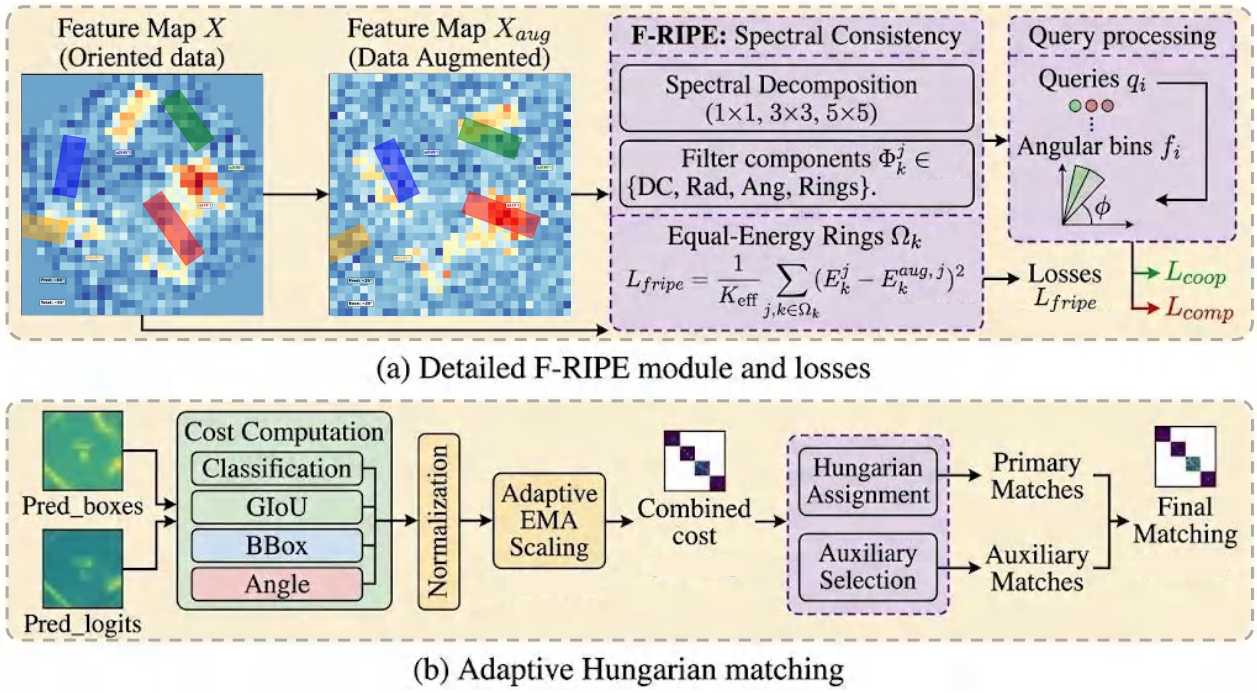



Figure 3 – Main modules of RingCo-DETR. (a) F-RIPE enforces frequency-ring consistency between an aerial feature map and its rotated counterpart and provides angular query coordination signals $L_{\text{coop}}$ and $L_{\text{comp}}$. (b) Adaptive Hungarian Matching balances heterogeneous costs and increases supervision density for dense aerial formations.

## 3. Experiments

### 3.1 Dataset Preparation and Experimental Setup

We perform all of our experiments on two widely used datasets: DOTA v1.5 [1] and HIT-UAV [2], which together cover the two operational regimes targeted by our method, namely airfield surface monitoring and UAV-based low-altitude patrol. DOTA-v1.5 is an enhanced version of the original DOTA-v1.0 dataset, which consists of 2,806 aerial images annotated for object detection, covering 15 categories and providing 188,282 instances. In DOTA-v1.5, the dataset retains the same number of images and splits as v1.0 but introduces annotations for extremely small instances (less than 10 pixels) and adds a new category, “container crane.” The HIT-UAV dataset includes 2,898 infrared thermal images extracted from 43,470 frames of various scenes such as schools, parking lots, roads, and playgrounds. Additionally, it records key flight data such as altitude, camera angle, date, and sunlight intensity. Each image is manually annotated with both oriented and standard bounding boxes to address significant object overlap in aerial images.

We evaluate on DOTA v1.5 and HIT-UAV with a ResNet-50 backbone and a standard DETR head; the encoder–decoder and inference stay unchanged. Models train for 300 epochs with AdamW (lr $1e{-4}$), batch size 16. We report mAP@50 and $\Delta$Time (%/epoch vs. Default; negative is faster). All experiments are performed on Tesla V100 GPUs.

### 3.2 Safety-Critical Benchmarks and Metrics

DOTA-v1.5 contains densely packed, arbitrarily oriented objects such as aircraft, vehicles, and infrastructure, representing the congestion typical of airport aprons and taxiways and thus serving as a proxy for airfield surface monitoring. HIT-UAV provides high-altitude infrared thermal imagery acquired from a UAV platform, simulating low-visibility and nighttime patrol conditions where FOD, intruders, or wildlife incursions are hardest to detect [3]. We report mAP@50 as the primary metric, alongside per-epoch $\Delta$Time to verify that our training-only approach does not degrade inference throughput, which is required for deployed airborne perception systems.

### 3.3 Ablation Studies

Table 3 reports the incremental effects of each component. We first evaluate Adaptive Hungarian Matching alone, which improves mAP@50 to 76.3 on DOTA and 77.6 on HIT-UAV by rebalancing heterogeneous matching costs and stabilizing early assignments. Adding F-RIPE further increases performance to 78.2 and 79.7, respectively, indicating that frequency-ring consistency improves robustness to rotation-like attitude perturbations. Finally, integrating the CCQA (Cooperative–Competing Query Alignment) module, which exploits the dual query streams $Q_{\text{coop}}$ and $Q_{\text{comp}}$ shown in Figure 1, yields the full RingCo-DETR with 81.7/83.1 mAP@50. All components are applied during training only, and the encoder–decoder and inference graph remain unchanged.

To better understand the gain from F-RIPE, Figure 4 visualizes image-domain differences before and after simulated inverse-rotation compensation. Before compensation, global attitude perturbation produces widespread residual responses over aircraft boundaries, runway markings, service vehicles, and background structures. After inverse compensation, the residuals become substantially weaker and more localized. This suggests that a large portion of the apparent frame-level change is induced by platform rotation rather than object motion. Such scene-wide perturbation can disturb high-frequency feature responses and reduce recall for small or elongated targets, explaining why the ring-wise spectral stabilization introduced by F-RIPE improves the ablation results.

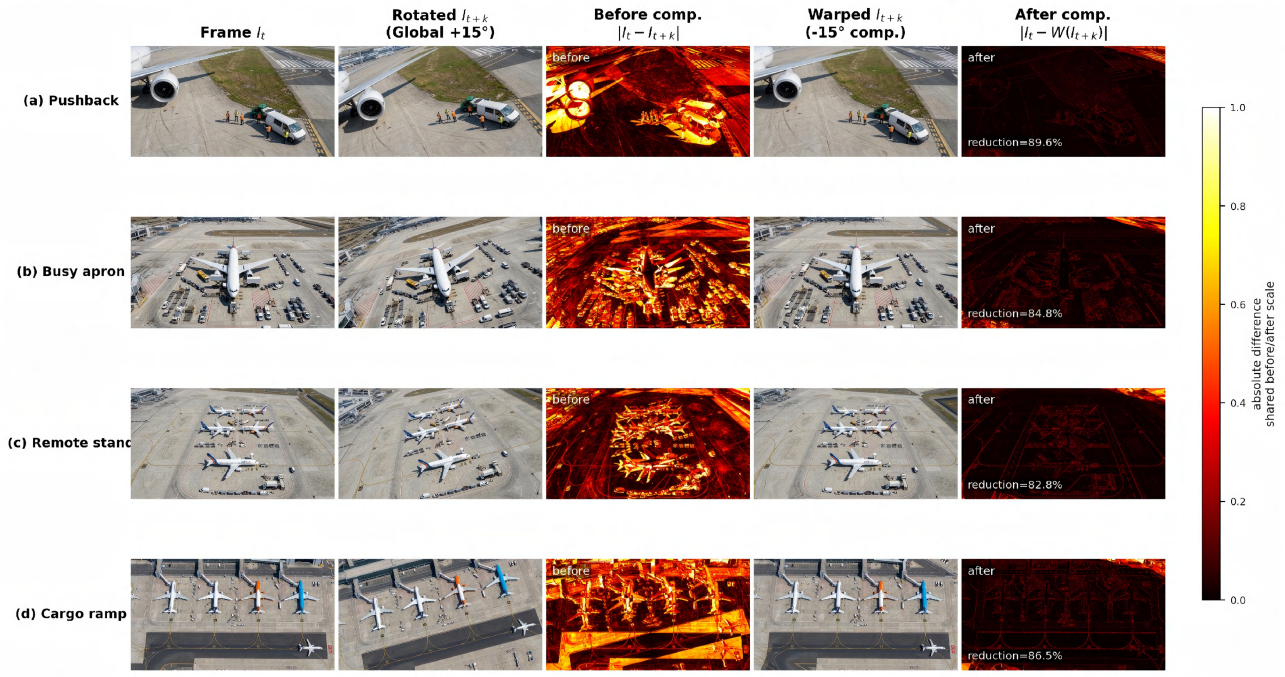



Figure 4 – Image-domain evidence of platform-induced change.

Table 1 – Hyperparameter sensitivity (mAP@50 and ΔTime). Default: $\alpha$=0.95, $q$=0.95, $K$=3.

| Config | DOTA-v1.5 | | HIT-UAV | |
|---|---|---|---|---|
| | mAP | ΔTime | mAP | ΔTime |
| Default | 81.7 | – | 83.1 | – |
| $\alpha = 0.90$ | 79.3 | ↓2.3% | 80.7 | ↓2.1% |
| $\alpha = 0.98$ | 79.1 | ↑1.8% | 81.3 | ↑1.6% |
| $q = 0.90$ | 80.2 | ↓1.2% | 80.3 | ↓1.4% |
| $q = 0.98$ | 79.7 | ↑0.9% | 79.7 | ↑1.2% |
| $K = 1$ | 80.1 | ↓18.5% | 79.4 | ↓17.8% |
| $K = 5$ | 82.3 | ↑20.7% | 84.4 | ↑19.3% |

Table 2 – Comparison on DOTA-v1.5 and HIT-UAV (mAP@50).

| Model | Param.(M) | DOTA | HIT-UAV |
|---|---|---|---|
| YOLOv8-S | 11.4 | 76.1 | 77.4 |
| YOLOv8-M | 26.4 | 78.7 | 79.1 |
| O-RCNN | 41.1 | 73.3 | 75.5 |
| O-RepPoints | 35.4 | 62.7 | 59.1 |
| ReDet | 32.4 | 76.8 | 77.3 |
| RingCo-DETR | 43.9 | 81.7 | 83.1 |

Table 3 – Ablation study on DOTA and UAV validation sets.

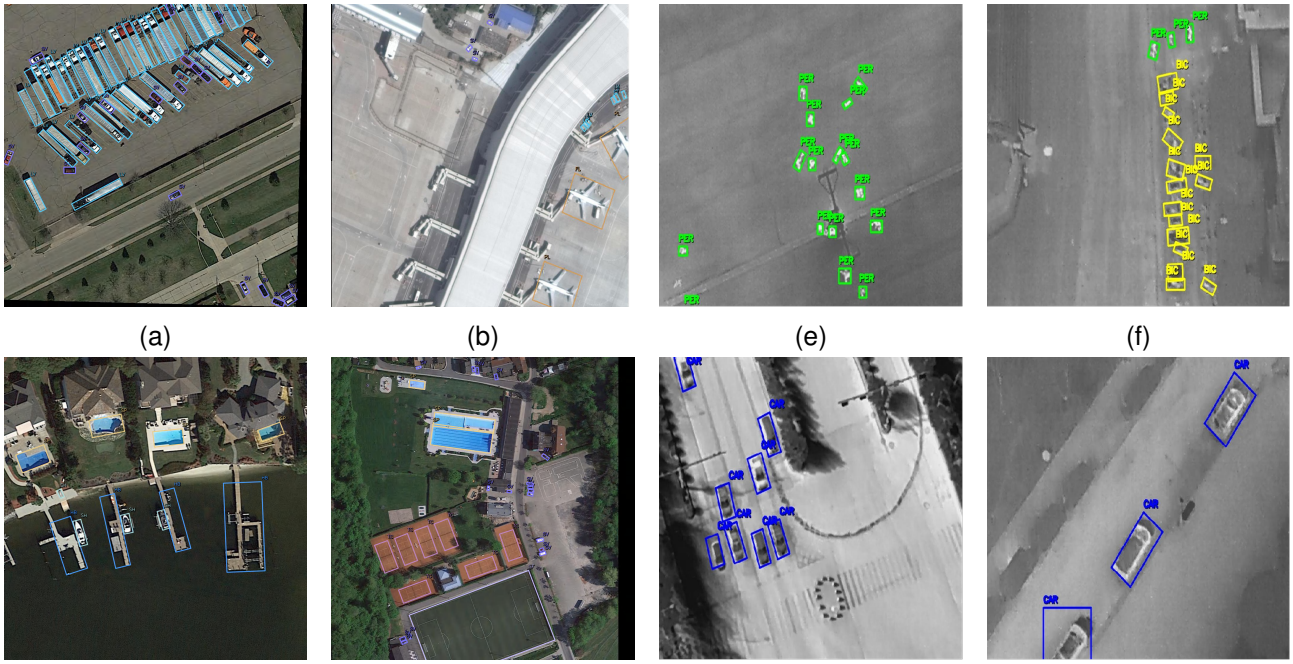
| Adaptive Matcher | F-RIPE | CCQA | mAP@50 (%) | |
|---|---|---|---|---|
| | | | DOTA | UAV |
| ✓ | | | 76.3 | 77.6 |
| ✓ | ✓ | | 78.2 | 79.7 |
| ✓ | ✓ | ✓ | 81.7 | 83.1 |

## 3.4 Results and Analysis

Table 1 presents a systematic sensitivity analysis of the three hyperparameters governing Adaptive Hungarian Matching. The default configuration ($\alpha = 0.95, q = 0.95, K = 3$) achieves 81.7 mAP@50 on DOTA-v1.5 and 83.1 mAP@50 on HIT-UAV. Reducing the EMA momentum to $\alpha = 0.90$ degrades performance by $-2.4/-2.4$ points, indicating that insufficient temporal smoothing allows transient cost outliers to destabilize early assignments, which is a particularly damaging effect when view content changes rapidly during flight. Conversely, excessive smoothing ($\alpha = 0.98$) also hurts accuracy ($-2.6/-1.8$ points), suggesting that overly sluggish adaptation fails to track the evolving cost landscape across training epochs. The quantile threshold $q$ exhibits a similar inverted-U pattern: $q = 0.90$ under-normalizes extreme costs, while $q = 0.98$ over-caps legitimate outliers, both yielding suboptimal mAP. The auxiliary ring capacity $K$ shows a clear accuracy–efficiency trade-off. Setting $K = 1$ reduces per-epoch time by $\sim$18% but sacrifices $-1.6/-3.7$ mAP points, confirming that under-supervision of crowded scenes degrades generalization to dense aerial layouts. Increasing to $K = 5$ boosts mAP to 82.3/84.4 at the expense of $\sim$20% longer training time, validating that denser positive assignment improves recall for densely packed aerial targets such as aircraft rows on aprons. The default $K = 3$ strikes a pragmatic balance, delivering near-optimal accuracy with no training-time penalty relative to the heavier configuration.

Table 2 compares RingCo-DETR against representative oriented detectors. Among NMS-based methods, O-RCNN and O-RepPoints achieve 73.3/75.5 and 62.7/59.1 mAP@50, respectively, lagging behind our approach by substantial margins. ReDet, which explicitly encodes rotation equivariance in its architecture, attains 76.8/77.3, yet remains below RingCo-DETR by $-4.9/-5.8$ points. This demonstrates that training-time stabilization via adaptive matching and spectral regularization can outperform architectural rotation priors without inflating inference complexity, which is a desirable property for airborne deployment where every additional inference operator carries a power and latency cost. YOLOv8-S and YOLOv8-M, though efficient in parameter count (11.4M and 26.4M), top out at 76.1/77.4 and 78.7/79.1, respectively. RingCo-DETR surpasses YOLOv8-M by $+3.0/+4.0$ points with a larger ResNet-50 backbone (43.9M parameters); the gain is attributable to the training approach rather than backbone scale. All competing methods require NMS post-processing, whereas RingCo-DETR preserves the end-to-end NMS-free property of DETR, which matters for deterministic embedded deployment on airborne platforms.

Figure 5 provides qualitative evidence of the training stabilization benefits. In the leftmost panel, a densely packed airport apron scene contains dozens of aircraft and service vehicles arranged in tight, near-parallel rows. RingCo-DETR produces dense yet non-redundant oriented boxes with accurate heading angles, avoiding the spurious duplicate detections typical of unstable matching. The second panel depicts an airfield taxiway with small elongated objects; our method maintains consistent orientations even for partially occluded targets. The third and fourth panels illustrate HIT-UAV infrared scenarios over water and residential areas, respectively, where low-visibility conditions and thermal clutter challenge detection. RingCo-DETR exhibits robust recall for small boats and swimming pools, with bounding-box orientations closely aligned to the true object axes. These visual results are consistent with the quantitative gains: adaptive matching increases supervision density in crowded regions, while F-RIPE suppresses attitude-induced feature drift, together yielding more reliable oriented localization across diverse aerial conditions.



(a) (b) (e) (f)

(c) (d) (g) (h)

DOTA-v1.5 UAV

Figure 5 – Qualitative results on DOTA-v1.5 and HIT-UAV.

## 4. Conclusion

We present RingCo-DETR, a training-only framework that stabilizes oriented DETR for safety-relevant aerial sensing. Adaptive Hungarian Matching mitigates cost-scale drift across airfield surface monitoring and UAV low-altitude patrol regimes and increases supervision density with bounded IoU-ring auxiliaries under cross-query constraints, while F-RIPE enforces rotation-consistent spectral regularization to counteract platform attitude perturbations and provides angular coordination signals to the cooperative/competing query streams. On DOTA-v1.5 and HIT-UAV, RingCo-DETR improves training stability, recall, and mAP@50 with no inference-time changes. The inference-neutral design preserves deterministic latency and an unchanged auditable structure for airborne embedded deployment. Beyond airfield and UAV monitoring, the same training approach can potentially extend to other orientation-sensitive aerial sensing settings, such as maritime patrol and post-disaster aerial assessment.

## Contact Author Email Address

zihao.wang@sydney.edu.au

## Copyright Statement

The authors confirm that they, and/or their company or organization, hold copyright on all of the original material included in this paper. The authors also confirm that they have obtained permission, from the copyright holder of any third party material included in this paper, to publish it as part of their paper. The authors confirm that they give permission, or have obtained permission from the copyright holder of this paper, for the publication and distribution of this paper as part of the ICAS proceedings or as individual off-prints from the proceedings.

## References


[1] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. Dota: A large-scale dataset for object detection in aerial images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3974–3983, 2018.

[2] Jiashun Suo, Tianyi Wang, Xingzhou Zhang, Haiyang Chen, Wei Zhou, and Weisong Shi. Hit-uav: A high-altitude infrared thermal dataset for unmanned aerial vehicle-based object detection. *Scientific Data*, 10:227, 2023.

[3] Travis Munyer, Daniel Brinkman, Xin Zhong, Chenyu Huang, and Iason Konstantzos. Foreign object debris detection for airport pavement images based on self-supervised localization and vision transformer. In *2022 International conference on computational science and computational intelligence (CSCI)*, pages 1388–1394. IEEE, 2022.

[4] Liang Wang, Jianliang Ai, Li Zhang, and Zhenlin Xing. Design of airport obstacle-free zone monitoring uav system based on computer vision. *Sensors*, 20(9):2475, 2020.

[5] Jingfeng Shan, Lapo Miccinesi, Alessandra Beni, Lorenzo Pagnini, Andrea Cioncolini, and Massimiliano Pieraccini. A review of foreign object debris detection on airport runways: sensors and algorithms. *Remote Sensing*, 17(2):225, 2025.

[6] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[7] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[8] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[10] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12179–12188, 2021.

[11] 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*, pages 213–229. Springer, 2020.

[12] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.

[13] Depu Meng, Xiaokang Chen, Zejia Fan, Gang Zeng, Houqiang Li, Yuhui Yuan, Lei Sun, and Jingdong Wang. Conditional detr for fast training convergence. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3651–3660, 2021.

[14] 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. *arXiv preprint arXiv:2203.03605*, 2022.

[15] Kai Xu, Minghai Qin, Fei Sun, Yuhao Wang, Yen-Kuang Chen, and Fengbo Ren. Learning in the frequency domain. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1740–1749, 2020.

[16] Ian Goodfellow, Honglak Lee, Quoc Le, Andrew Saxe, and Andrew Ng. Measuring invariances in deep networks. *Advances in neural information processing systems*, 22, 2009.

[17] 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*, pages 6633–6642, 2023.

[18] Feng Li, Hao Zhang, Shilong Liu, Jian Guo, Lionel M Ni, and Lei Zhang. Dn-detr: Accelerate detr training by introducing query denoising. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13619–13627, 2022.

[19] Pierre Ancey, Andrew Price, Saqib Javed, and Mathieu Salzmann. Fastpose-vit: a vision transformer for real-time spacecraft pose estimation. *arXiv preprint arXiv:2512.09792*, 2025. Accepted to IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) 2026.

[20] Rafael C Gonzalez and Richard E Woods. *Processamento de imagens digitais*. Editora Blucher, 2000.

[21] Xiaodi Hou and Liqing Zhang. Saliency detection: A spectral residual approach. In *2007 IEEE Conference on computer vision and pattern recognition*, pages 1–8. Ieee, 2007.