Research Square

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

SAF-RTDETR: Spatially Aligned High-Resolution Feature Fusion for Small Object Detection in UAV Imagery

Ming Zhang
Northwest Normal University
Jiahui He
Northwest Normal University
Dongmei Ma
1220284219@qq.com

Northwest Normal University

Research Article

Keywords:

Posted Date: October 30th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11390432/v1

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

Additional Declarations: No competing interests reported.

SAF-RTDETR: Spatially Aligned High-Resolution Feature Fusion for Small Object Detection in UAV Imagery

Ming Zhang[1], Jiahui He[1], Dongmei Ma[1*]

[1*]College of Physics and Electronic Engineering, Northwest Normal University, Lanzhou, Gansu, 730070, China.

*Corresponding author(s). E-mail(s): 1220284219@qq.com;

Abstract

Backbone downsampling weakens small objects in unmanned aerial vehicle imagery, whereas direct use of shallow features also amplifies structured background. We propose SAF-RTDETR, a four-scale RT-DETR extension that treats high-resolution detection as selective feature fusion. A grouped sampling unit aligns stride-8 shallow and semantic features; the aligned feature then guides a directionally weighted stride-4 branch. Partial-channel single-head attention supplies high-level context. Under the same 100-epoch VisDrone2019-DET protocol, three runs obtain $\mathbf{31.14 \pm 0.51}\%$ $\mathbf{AP_{50:95}}$ versus $\mathbf{28.46 \pm 0.73}\%$ for RT-DETR-R50, with gains in every paired seed and a 2.77-point increase in $\mathbf{AP_S}$. The model also leads all 18 tested corruption conditions and raises mean corrupted AP from 24.85% to 27.45%. Cross-dataset AP changes are +4.54, +1.00, and -0.34 points on AI-TOD-v2, HIT-UAV, and UAVDT. SAF-RTDETR reaches 57.62 FPS on an RTX 5090 but requires 288.23 GFLOPs. The results support selective high-resolution fusion while exposing its computational and transfer limits. A controlled ablation further shows that a plain stride-4 level adds only 0.63 AP, whereas selective enhancement accounts for most of the gain.

Keywords: UAV object detection, small object detection, RT-DETR, high-resolution representation, cross-scale feature alignment, feature fusion

1 Introduction

UAV imagery contains many small, crowded, and partially occluded objects whose appearance varies with altitude and view. VisDrone captures these conditions across diverse cities, scenes, weather, and platforms [11]. Weak targets may occupy only a few pixels, while roads, roofs, shadows, and vegetation produce stronger local contrast.

FPN combines deep semantics with shallow detail [1], and aerial detectors further exploit tiling, high-resolution features, and scale-aware assignment [12, 13, 18]. The recurring trade-off is that finer maps preserve targets but also increase computation and background activation. RT-DETR uses end-to-end set prediction and an efficient hybrid encoder, but its ResNet-50 variant sends only stride-8, 16, and 32 features to the decoder [10]. Our plain-P_2 experiment gains only 0.63 AP, suggesting that resolution alone is insufficient.

SAF-RTDETR therefore calibrates shallow and top-down features at stride 8 before using

the result both for decoding and as semantic guidance for a new stride-4 branch. Directional weighting suppresses indiscriminate fine-scale responses, while partial-channel single-head attention supplies high-level context. The decoder receives strides 4–32; matching, losses, queries, and decoder depth remain unchanged.

The contributions of this work are threefold:

- We develop a selective high-resolution path for RT-DETR in which a calibrated stride-8 feature guides directional weighting at stride 4, distinguishing selective detail from a direct shallow connection.
- We place grouped spatial calibration before both stride-8 decoding and stride-4 construction, linking cross-scale alignment and high-resolution enhancement in one feature flow, with partial-channel SHSA supplying high-level semantics.
- Three paired 100-epoch runs show a consistent AP advantage over RT-DETR-R50; component, scale, corruption, transfer, qualitative, and efficiency studies expose both benefits and costs.

2 Related Work

2.1 Small Object Detection in UAV Imagery

VisDrone covers pedestrians and vehicles under varied views and densities [11]; UAVDT emphasizes vehicles in video and scene attributes [23]. ClusDet crops object clusters [12], UFPMP-Det packs foreground regions [14], QueryDet sparsely activates fine features [13], and CEASC adapts sparse-head activation using global context [15]. ESOD filters background feature patches [16], while Dome-DETR sparsifies features and allocates queries by density [17]. Normalized Wasserstein distance addresses the positional sensitivity of tiny boxes [18]. These methods motivate selective high-resolution processing; our study changes the internal semantic–detail path while retaining RT-DETR's objective.

2.2 High-Resolution and Multi-scale Feature Representation

FPN introduced top-down lateral fusion [1]; PANet added a bottom-up path [2], and recursive pyramids and Dynamic Head further refined multi-scale interaction [19, 20]. Fine pyramid levels improve small-object sampling but retain structured background and redundant computation. We fuse C_2 with semantic guidance before applying ELA, whose directional pooling produces spatial weights [21]. ELA itself is not claimed as novel; its placement in the four-scale RT-DETR path is the point studied here.

2.3 Feature Alignment in Cross-scale Fusion

Resizing equalizes tensor dimensions but cannot model content-dependent displacement. Deformable convolution and attention learn sparse sampling locations [4, 5]; CSFCN uses grouped sampling for spatial calibration [22]. Our SFC_{G2} follows this mechanism at the stride-8 fusion point, where both branches receive predicted offsets before their output guides the new stride-4 feature.

2.4 Transformer-based Object Detection

DETR uses bipartite matching and fixed queries [3]. Deformable attention, DAB-DETR, DN-DETR, DINO, and Co-DETR improve attention, queries, denoising, or supervision [5–9]. RT-DETR targets real-time inference through a hybrid encoder and query selection [10]. SHViT applies single-head attention to only a channel subset [25]; we adapt this operation to RT-DETR while retaining its decoder, matching, and losses.

3 Proposed Method

3.1 Overall Architecture

SAF-RTDETR retains the ResNet-50 backbone and RT-DETR head and modifies only the hybrid encoder and decoder inputs. The backbone produces

$$\mathcal{C} = \{C_2, C_3, C_4, C_5\}, \qquad C_i \in \mathbb{R}^{B\times c_i\times H_i\times W_i}, \tag{1}$$

at strides $\{4, 8, 16, 32\}$; encoder outputs use 256 channels.

SHSA encodes projected C_5, whose top-down fusion with C_4 gives Y_4. Projected C_3 and Y_4 then

pass through two-group spatial calibration:

$$P_3^{\text{aligned}} = \text{RepC3}\left(\text{SFC}_{G=2}\left(\phi_3(C_3), Y_4\right)\right), \quad (2)$$

where ϕ_3 is a 1×1 projection. The aligned result enters the decoder and guides the stride-4 branch:

$$X_2 = \text{RepC3}\left[\phi_2(C_2), \mathcal{U}_2\left(P_3^{\text{aligned}}\right)\right], \quad (3)$$

$$P_2^{\text{enh}} = \text{ELA}(X_2). \quad (4)$$

Here, $[\cdot,\cdot]$ is channel concatenation and $\mathcal{U}_2$ denotes $2\times$ nearest-neighbor upsampling. The original bottom-up path produces P_4 and P_5, so the decoder receives

$$\mathcal{F}_{\text{dec}} = \left\{P_2^{\text{enh}}, P_3^{\text{aligned}}, P_4, P_5\right\}. \quad (5)$$

Calibration precedes semantic guidance; the design is therefore a connected feature path rather than three independent modules.

3.2 Selective High-Resolution Feature Enhancement

Let $X_2 \in \mathbb{R}^{B\times C\times H\times W}$ be the fused feature in Eq. (3). ELA summarizes it separately along the horizontal and vertical axes:

$$z_h(b,c,i) = \frac{1}{W}\sum_{j=1}^{W} X_2(b,c,i,j), \quad (6)$$

$$z_w(b,c,j) = \frac{1}{H}\sum_{i=1}^{H} X_2(b,c,i,j). \quad (7)$$

The two descriptors are processed by the same channel-preserving one-dimensional convolution. In the repository implementation, its kernel size is 7. Group normalization with 16 groups and a sigmoid activation then produce two directional weights:

$$A_h = \sigma\left(\text{GN}_{16}\left(g_7(z_h)\right)\right), \quad (8)$$

$$A_w = \sigma\left(\text{GN}_{16}\left(g_7(z_w)\right)\right). \quad (9)$$

After their singleton dimensions are restored, the output is

$$P_2^{\text{enh}} = X_2 \odot A_h \odot A_w, \quad (10)$$

where $\odot$ denotes element-wise multiplication with broadcasting.

The operation does not remove background responses explicitly. Instead, it learns height- and width-dependent weights on a feature that already contains both C_2 detail and P_3 semantics. Its purpose is to make the added high-resolution level more selective than a plain lateral P_2 connection. Whether this selection is responsible for the observed gain is examined by the plain-P_2 and selective-P_2 ablations in Section 4.4.

3.3 Spatially Calibrated Cross-Scale Feature Fusion

Because C_3 and Y_4 have different sampling histories, SFC_{G2} applies learnable sampling before stride-8 RepC3 refinement.

Let $X_l = \phi_3(C_3)$ and $X_h = Y_4$. After 3×3 transformations, the semantic branch is bilinearly resized:

$$\widehat{X}_l = f_l(X_l), \quad (11)$$

$$\widehat{X}_h = \mathcal{U}_{\text{bil}}\left(f_h(X_h); H_l, W_l\right). \quad (12)$$

An offset head reduces their concatenation to 64 channels and predicts $R = [\Delta_l, \Delta_h, m_h, m_l]$. Each offset has $2G$ channels for $G = 2$; m_h and m_l are branch-wise spatial modulation maps.

For channel group g, offsets are added to normalized grid $\mathcal{G}$:

$$\widetilde{X}_l^{(g)} = \text{GridSample}\left(\widehat{X}_l^{(g)}, \mathcal{G} + \frac{\Delta_l^{(g)}}{(W_l, H_l)}\right), \quad (13)$$

$$\widetilde{X}_h^{(g)} = \text{GridSample}\left(\widehat{X}_h^{(g)}, \mathcal{G} + \frac{\Delta_h^{(g)}}{(W_l, H_l)}\right). \quad (14)$$

After bilinear grid sampling, the branches are fused as

$$\begin{aligned} X_{\text{SFC}} = {} & (1 + \tanh(m_h)) \odot \widetilde{X}_h \\ & + (1 + \tanh(m_l)) \odot \widetilde{X}_l. \end{aligned} \quad (15)$$

The maps are shared within each branch. Zero initialization of the final offset convolution starts training from the regular grid with unit modulation.

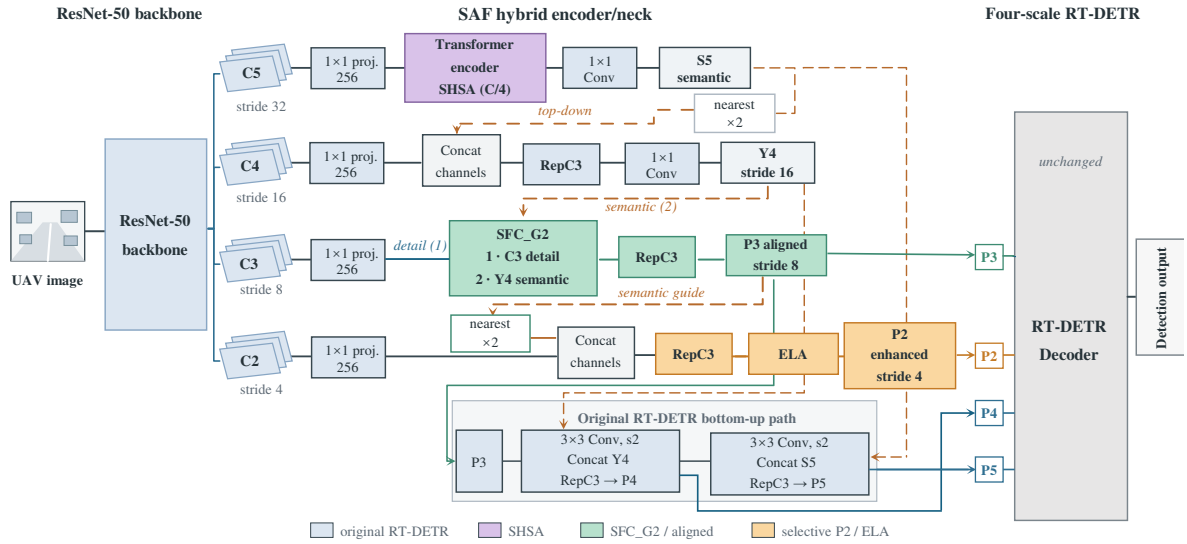


Fig. 1 Overall architecture of SAF-RTDETR. The stride-8 feature is spatially calibrated before being used both as a decoder input and as semantic guidance for constructing the selective stride-4 feature.

The calibrated feature is refined by RepC3 to obtain

$$P_3^{\text{aligned}} = \text{RepC3}(X_{\text{SFC}}). \tag{16}$$

This feature also supplies the semantic input used to construct P_2^{enh}.

3.4 Lightweight Semantic Encoding and Multi-scale Decoding

We replace AIFI with partial-channel SHSA [25]. For projected C_5 feature $X \in \mathbb{R}^{B\times C\times H\times W}$ and $N = HW$,

$$X = [X_a, X_b], \qquad C_a = \frac{C}{4}, \qquad C_b = \frac{3C}{4}. \tag{17}$$

only X_a enters attention:

$$[Q, K, V] = \text{ConvBN}_{1\times1}\left(\text{GN}_1(X_a)\right), \tag{18}$$

where $C_Q = C_K = C/2$ and $C_V = C/4$. After spatial flattening,

$$A = \text{Softmax}\left(\frac{Q^{\mathsf{T}}K}{\sqrt{C_Q}}\right) \in \mathbb{R}^{B\times N\times N}, \tag{19}$$

and $X'_a = VA^{\mathsf{T}}$. Concatenating the bypass X_b gives

$$\text{SHSA}(X) = \text{ConvBN}_{1\times1}\left(\text{SiLU}\left([X'_a, X_b]\right)\right). \tag{20}$$

The final BN scale is zero-initialized. A post-normalized residual encoder and 1×1 feed-forward block follow. Here $C = 256$; attention uses 64 input/value channels, 128-channel Q and K, and a hidden width of 1024.

The four levels enter the unchanged RT-DETR decoder [10]:

$$N_q = 300, \qquad L = 6, \qquad N_h = 8, \qquad N_p = 4, \tag{21}$$

for queries, layers, heads, and sampling points. Matching, losses, auxiliary supervision, and loss weights remain unchanged.

4 Experiments

4.1 Experimental Setup

4.1.1 Datasets

The main experiments use VisDrone2019-DET [11], with 6,471 training images, 548 validation images, and 1,610 test-dev images covering ten object categories. Generalization is examined by retraining the baseline and the proposed model on three additional benchmarks rather

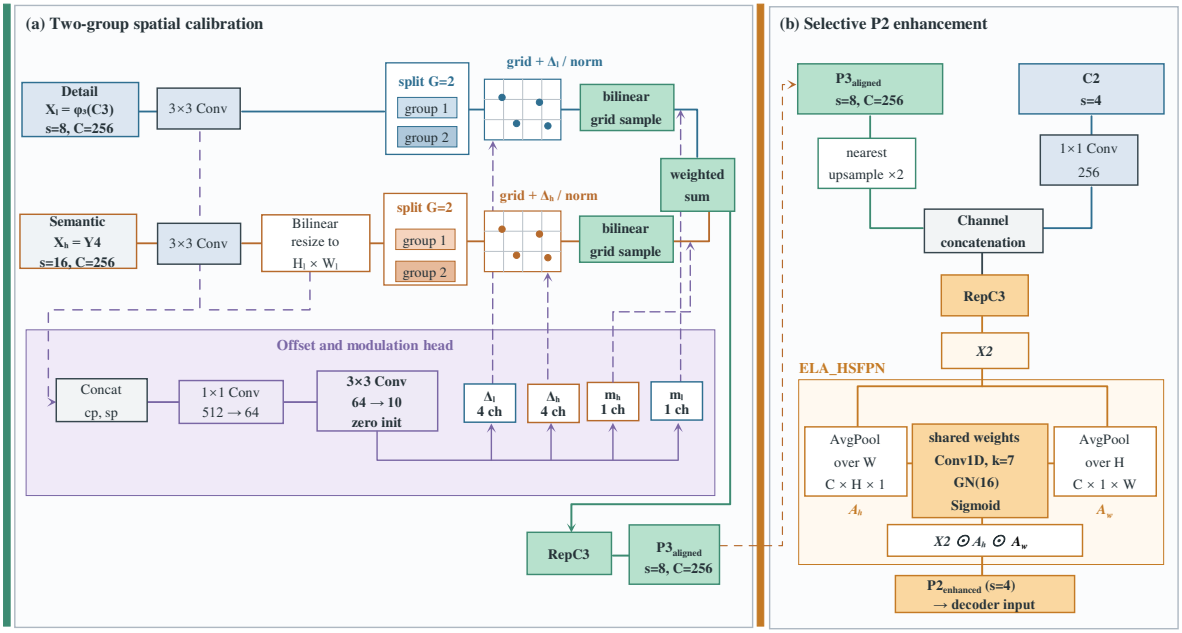


Fig. 2 Construction of the calibrated stride-8 feature and the selective stride-4 representation. SFC_{G2} calibrates both cross-scale branches, and the aligned feature subsequently guides the high-resolution pathway.

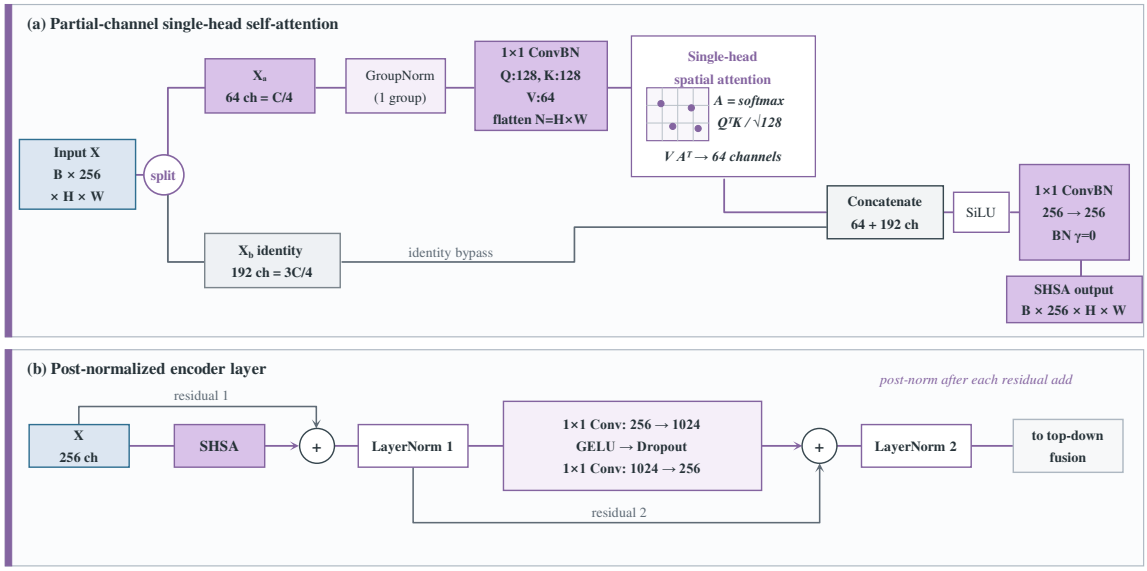


Fig. 3 Structure of the partial-channel SHSA encoder. One quarter of the channels participates in single-head spatial attention, while the remaining channels follow a bypass path before channel fusion.

than transferring a VisDrone checkpoint directly. AI-TOD-v2 [18] contains 11,213/2,804/14,018 train/validation/test images and eight categories; UAVDT contains 19,049/5,094/16,592 images and three vehicle categories [23]; HIT-UAV contains 2,029/290/579 thermal images and four scored categories after excluding DontCare [24]. The datasets differ in category definition, object scale, scene content, and sensing modality. Consequently, we compare the two models only within

a dataset and do not interpret absolute AP differences between datasets.

4.1.2 Metrics and reporting rule

The primary metric is AP averaged over IoU thresholds from 0.50 to 0.95 in steps of 0.05, denoted by $AP_{50:95}$. Precision, recall, and AP_{50} are also reported. Component ablations and controlled detector comparisons use the best validation checkpoint from their seed-0 100-epoch run. The central baseline comparison is repeated with seeds 1, 2, and 3 and reported as mean ± sample standard deviation; paired differences are formed using equal seed indices. The scale-wise results come from a separate COCO-style evaluator with maxDets = 300 and are kept in their own table.

For corruption robustness, the same clean-data `best.pt` checkpoint is evaluated, without retraining, on the clean validation set and on 18 derived sets. Following common-corruption evaluation practice [26, 27], mean performance under corruption is $\mathrm{mPC} = \frac{1}{18}\sum_{c=1}^{6}\sum_{s=1}^{3} \mathrm{AP}_{c,s}$, and relative performance is $\mathrm{rPC} = \mathrm{mPC}/\mathrm{AP}_{clean}$. These sets are derived from the VisDrone validation images and are not an official "VisDrone-C" benchmark.

4.1.3 Implementation details

Both RT-DETR variants are trained for 100 epochs at 640 × 640 resolution with batch size 8, AdamW, an initial learning rate of 10^{-4}, weight decay of 10^{-4}, deterministic execution, and no mixed precision. They use the same ResNet-50 backbone, data split, augmentations, 300 object queries, six decoder layers, eight attention heads, and four sampling points. Apart from the seed in the repeated experiment, the training settings are unchanged. The matching rule and classification and localization losses are also retained. YOLO runs use the same image size, batch size, epoch budget, split, seed, and optimizer, while keeping the learning-rate and augmentation settings of their own configurations. In the cross-dataset experiments, AI-TOD-v2 uses batch size 4 because of memory demand, whereas HIT-UAV and UAVDT use batch size 8; each baseline/proposed pair otherwise shares the same dataset-specific setup.

4.2 Comparison on VisDrone2019-DET

Table 1 gives the seed-0 controlled context. The proposed model improves $AP_{50:95}$ by 3.15 points and recall by 4.06 points over RT-DETR-R50. The YOLO entries are practical references rather than component controls because their optimization and post-processing differ.

The YOLOv10m run reaches its best AP at epoch 11 and then degrades, unlike the other controlled runs. We retain the observed best value for transparency but avoid interpreting it as a stable ranking of the detector family.

The repeated rows in Table 1 provide the stronger baseline comparison. SAF-RTDETR improves AP in every paired seed by 1.55, 2.99, and 3.51 points. Its mean AP is 2.68 points higher and its standard deviation is lower (0.51 versus 0.73), so the seed-0 result is not an isolated favorable run. With only three repetitions, however, these values describe stability rather than establish a formal significance claim.

4.3 Comparison with Published Aerial-Detection Methods

Table 2 complements the controlled study with representative non-YOLO aerial detectors spanning cropping, packing, sparse querying, adaptive convolution, feature filtering, and density-aware queries. We do not retrain them; the table reproduces their published values and is not a ranking under our environment.

These values provide context rather than a strict ranking: the cited methods differ in backbone, resolution, training, crop/packing pipeline, and detection limit. UFPMP-Det and enlarged ESOD use higher-resolution inputs, while Dome-DETR allows 1500 detections per image. Our 32.54 AP uses 640 × 640 inputs and 300 queries; SAF-RTDETR instead modifies a fixed RT-DETR-R50 pipeline without crop-based or multi-pass inference.

4.4 Ablation Study

The ablation sequence follows the actual information path. A1 and A2 replace only the high-level encoder or stride-8 fusion, respectively. A3 adds a plain P_2 decoder input. A4 uses the semantic-guided selective P_2 path with ELA. A5 and A6 add

Table 1 Controlled validation results on VisDrone2019-DET. The final two rows are mean ± sample standard deviation over seeds 1–3; other rows are seed-0 runs. Base and SAF denote RT-DETR-R50 and SAF-RTDETR.

Method	P	R	AP_{50}	$AP_{50:95}$
YOLOv8m	48.90	38.13	37.73	21.76
YOLOv10m	41.02	32.71	30.73	17.27
YOLO11m	50.33	37.24	37.86	21.91
Base	62.12	46.37	47.95	29.39
SAF	**63.72**	**50.44**	**51.79**	**32.54**
Base (3 runs)	60.93 ± 1.06	45.27 ± 1.31	47.00 ± 1.22	28.46 ± 0.73
SAF (3 runs)	$\mathbf{63.81 \pm 0.50}$	$\mathbf{48.30 \pm 0.89}$	$\mathbf{50.08 \pm 0.85}$	$\mathbf{31.14 \pm 0.51}$

Table 2 Published results of representative non-YOLO methods on the VisDrone validation set. Values are copied from the cited papers and are not a controlled ranking. A dash means that the corresponding paper did not report the value in the cited comparison.

Method	Venue	Main strategy	Input	$AP_{50:95}$	AP_{50}	AP_{75}
ClusDet [12]	ICCV 2019	Cluster-region detection	–	26.7	50.6	–
QueryDet [13]	CVPR 2022	Cascaded sparse query	–	28.3	48.1	–
UFPMP-Det-R50 [14]	AAAI 2022	Foreground packing	1333×800	36.6	62.4	36.7
CEASC [15]	CVPR 2023	Adaptive sparse head	1333×800	28.7	50.7	–
ESOD† [16]	IEEE TIP 2025	Feature-level filtering	long side 1920	37.9	62.3	–
Dome-DETR-L [17]	ACM MM 2025	Density-aware DETR	800×800	39.0	61.1	40.8

†The enlarged ESOD setting scales both input dimensions by 1.25 from a 1536-pixel long side. Dome-DETR evaluates with at most 1500 detections per image, whereas our controlled protocol uses 300.

SFC_{G2} or SHSA to that path, while A7 combines SHSA and SFC_{G2} without P_2. The full model contains all three changes.

SHSA and SFC_{G2} alone change AP by -0.04 and -0.36 points, and plain P_2 adds only 0.63 points. Semantic guidance plus ELA (A4) instead adds 2.43 points. SHSA contributes another 0.31 points in this selective path, and the full model exceeds A6 by 0.42 points, although SFC_{G2} is negative when added to A4 alone. Thus, the gain is mainly tied to selective P_2 and its interaction with the complete feature flow, not independently additive blocks.

4.5 Robustness to Image Degradation

The clean validation images are transformed with Gaussian noise ($\sigma = 10, 20, 30$), motion blur (kernel size 5, 9, 13), low-light mapping ($\gamma = 1.5, 2.0, 2.5$), fog (severity 1, 3, 5), JPEG compression (quality 70, 40, 20), and downsample–restore scaling (0.75, 0.50, 0.33). Table 4 averages the three severities of each corruption. Besides the baseline and full model, A3 is included to test whether resolution alone explains the result.

SAF-RTDETR obtains the highest absolute AP in every one of the 18 corruption conditions. Its mPC exceeds the baseline by 2.60 points and plain P2 by 2.06 points. The largest type-level gain occurs under fog (+3.35), whereas motion blur remains the most difficult degradation. The rPC values are nearly equal, and SAF-RTDETR is 0.21 points lower than the baseline because its clean AP is also substantially higher. We therefore conclude that the proposed feature path preserves its clean-data advantage under these degradations, not that it reduces the proportional degradation rate.

4.6 Scale- and Class-wise Analysis

Table 5 uses the separate COCO-style evaluator. SAF-RTDETR improves AP_S by 2.77 points and AR_S by 3.70 points. Improvements also occur for medium and large objects in AP, so the method is not exclusively beneficial to small instances. The only negative entry is AR_L, which drops by 0.92 points. Because VisDrone contains far fewer large objects than small ones, this value should be interpreted cautiously, but it rules out a claim of uniform improvement at every scale.

All ten VisDrone categories improve in class-wise AP, as detailed in Table 6. The largest gains occur for pedestrian (+4.39), tricycle (+4.30), van (+3.21), motor (+3.14), and truck (+3.00); the

Table 3 Component ablation on the VisDrone2019-DET validation set.

Exp.	High-resolution path	ELA	SFC_{G2}	SHSA	P (%)	R (%)	AP_{50} (%)	$AP_{50:95}$ (%)
A0	None				62.12	46.37	47.95	29.39
A1	None			✓	61.32	46.15	47.84	29.34
A2	None		✓		62.06	45.69	47.78	29.03
A3	Plain P_2				62.49	47.83	49.21	30.02
A4	Selective P_2	✓			62.87	49.68	50.94	31.81
A5	Selective P_2	✓	✓		**65.05**	49.01	50.80	31.74
A6	Selective P_2	✓		✓	63.97	49.74	51.31	32.12
A7	None		✓	✓	61.54	46.34	47.91	29.13
Full	Selective P_2	✓	✓	✓	63.72	**50.44**	**51.79**	**32.54**

Table 4 Corruption robustness in $AP_{50:95}$ (%). Each corruption row averages three severity levels. P2 denotes A3 with a plain stride-4 input.

Condition	Base	P2	SAF	Δ Base
Gaussian noise	20.71	21.16	**22.95**	+2.24
Motion blur	19.02	19.16	**20.64**	+1.62
Low light	27.80	28.31	**30.77**	+2.98
Fog	26.63	27.51	**29.98**	+3.35
JPEG	27.88	28.60	**30.83**	+2.95
Downsample	27.09	27.63	**29.56**	+2.47
Clean AP	29.38	30.03	**32.53**	+3.16
mPC	24.85	25.39	**27.45**	+2.60
rPC (%)	**84.59**	84.57	84.38	-0.21

smallest is still positive for bus (+1.73). Thus, the overall improvement is not produced by one dominant class.

4.7 Cross-dataset Generalization

Table 7 reports independent 100-epoch training on each external dataset. On AI-TOD-v2, SAF-RTDETR improves AP by 4.54 points and recall by 6.37 points, consistent with the high proportion of tiny objects. On thermal HIT-UAV, AP improves by 1.00 point despite a 4.48-point reduction in recall; the higher precision compensates under the AP integral, illustrating that a single operating-point recall value should not replace AP. On UAVDT, AP decreases by 0.34 points. Thus, the architecture transfers well to AI-TOD-v2 and modestly to HIT-UAV, but it does not provide a universal gain.

4.8 Feature-response and Qualitative Analysis

For $F \in \mathbb{R}^{C\times H\times W}$, the displayed response is

$$M(i,j) = \frac{1}{C}\sum_{c=1}^{C} |F(c,i,j)|. \tag{22}$$

Maps are independently normalized between their 1st and 99th percentiles, so colors show within-panel spatial variation rather than cross-panel magnitude. Figure 4 suggests that calibrated and ELA-enhanced features concentrate more strongly on object-bearing regions. The lower-row SHSA, SFC_{G2}, and ELA maps come from different branches and scales, not a sequential chain; they are diagnostic rather than causal evidence.

Figure 5 uses confidence 0.25 for both models. A red dashed box is a verified baseline miss: a same-class SAF detection matches the ground truth at IoU $\geq$ 0.50, while no baseline detection does; numbers link the same object across columns. SAF recovers additional small objects in all three scenes, although false positives, severe occlusion, and very low contrast remain difficult.

4.9 Complexity and Inference Efficiency

Efficiency uses batch 1, FP32, a fused model, 200 warm-up and 1,000 timed pure-forward iterations on an RTX 5090 (PyTorch 2.8.0, CUDA 12.8), excluding loading, decoding, NMS, and drawing. SAF adds 3.88 M parameters; GFLOPs and peak memory rise by 114.3% and 87.9%. Latency increases by 1.44 ms and FPS falls from 62.83 to

Table 5 Scale-wise results under the unified COCO-style evaluator (maxDets = 300). Values are percentages.

Method	AP_S	AP_M	AP_L	AR_S	AR_M	AR_L
RT-DETR-R50	18.85	37.18	38.92	27.70	48.60	**54.53**
SAF-RTDETR	**21.63**	**40.55**	**43.11**	**31.39**	**51.52**	53.61

Table 6 Class-wise $AP_{50:95}$ on VisDrone2019-DET under the unified evaluator. Values and changes are percentage points.

Class	Base	SAF	Δ	Class	Base	SAF	Δ
Pedestrian	25.29	**29.67**	+4.39	Truck	25.93	**28.93**	+3.00
People	19.43	**22.03**	+2.60	Tricycle	16.91	**21.21**	+4.30
Bicycle	9.21	**12.02**	+2.82	Awning-tricycle	8.69	**11.24**	+2.55
Car	59.57	**62.29**	+2.72	Bus	45.32	**47.05**	+1.73
Van	30.86	**34.07**	+3.21	Motor	27.58	**30.72**	+3.14

Table 7 Generalization results from the best validation checkpoint of each 100-epoch run. Comparisons are meaningful within, not across, datasets.

Dataset	Method	P	R	AP_{50}	$AP_{50:95}$
AI-TOD-v2	Base	56.07	48.86	47.75	20.71
	SAF	**65.88**	**55.23**	**56.51**	**25.25**
HIT-UAV	Base	85.94	**83.80**	85.92	56.14
	SAF	**89.19**	79.32	**86.36**	**57.14**
UAVDT	Base	**48.64**	**40.60**	**39.98**	**25.32**
	SAF	47.80	40.29	39.54	24.98

57.62. It remains real-time on this device, but is not a lightweight modification.

Table 8 Complexity and pure-forward inference efficiency at 640×640.

Metric	RT-DETR-R50	SAF-RTDETR
Parameters (M)	**42.80**	46.69
GFLOPs	**134.48**	288.23
Weight size (MB)	**163.89**	178.77
Peak memory (MB)	**283.19**	531.95
Latency (ms)	**15.92**	17.35
FPS	**62.83**	57.62

5 Discussion

Selective stride-4 fusion is the main gain source: plain P_2 adds 0.63 AP, versus 2.43 after semantic guidance and ELA. SHSA and SFC_{G2} are neutral or negative alone, confirming that the blocks are not independently additive. The full model leads all 18 corruptions, but its nearly unchanged rPC indicates higher absolute accuracy rather than reduced proportional degradation.

The main limits are transfer and cost: UAVDT drops by 0.34 AP, while GFLOPs and peak memory rise by 114.3% and 87.9%. Evidence is limited to three VisDrone seeds, one validation split, single external runs, and synthetic corruptions; A4–A6 were not repeated. Future work should study sparse or channel-reduced P_2, test-server evaluation, and repeated real-weather experiments.

6 Conclusion

SAF-RTDETR combines grouped stride-8 calibration, semantic-guided stride-4 selection, and partial-channel attention in a four-scale RT-DETR encoder. It raises three-run AP from $28.46 \pm 0.73\%$ to $31.14 \pm 0.51\%$, improves every paired seed, adds 2.77 small-object AP, and leads all 18 corruptions. Ablations attribute most of the gain to selective high-resolution fusion. Mixed transfer and 114.3% more GFLOPs prevent universal or lightweight claims; the method is an accuracy-oriented option for tiny-object-heavy UAV scenes.

Statements and Declarations. **Funding** No funding was received for this work. **Competing interests** The authors have no relevant financial or non-financial interests to disclose.

Author contributions Ming Zhang performed conceptualization, methodology, software, validation, formal analysis, investigation, data curation, visualization, drafting, review, and project administration. Jiahui He contributed to

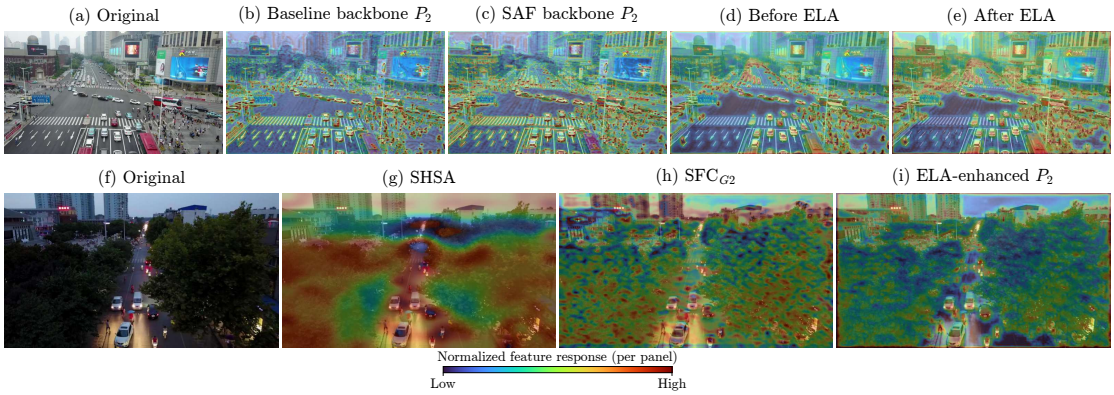


Fig. 4 Feature responses: (a)–(e) original, baseline/SAF P_2, and pre-/post-ELA responses; (f)–(i) SHSA, SFC$_{G2}$, and ELA-enhanced P_2 from different low-light branches/scales. Maps use independent 1st–99th percentile normalization.



Fig. 5 Qualitative results for dense, crowded-pedestrian, and low-light scenes. Columns are overview, ground truth, RT-DETR-R50, and SAF-RTDETR. Red dashed boxes are verified baseline misses recovered by SAF at confidence 0.25 and same-class IoU ≥ 0.50.

formal analysis, review, and project administration. Dongmei Ma supervised the study and contributed to review and project administration. All authors approved the final manuscript.

Data availability The datasets analyzed are public benchmarks cited herein. Derived evaluation records, trained weights, and corruption-generation resources are available from the corresponding author on reasonable request.

Code availability Code and final configurations will be released upon acceptance.

Ethics approval and consent Ethics approval, consent to participate, and consent for publication are not applicable.

Use of AI tools OpenAI ChatGPT/Codex assisted with language editing, LaTeX formatting, and visualization scripts. All content was author-verified; the authors take full responsibility.

References

[1] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2117–2125.

[2] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8759–8768.

[3] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proceedings of the European Conference on Computer Vision*, 2020, pp. 213–229.

[4] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 764–773.

[5] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," in *Proceedings of the International Conference on Learning Representations*, 2021.

[6] F. Li, H. Zhang, S. Liu, J. Guo, L. M. Ni, and L. Zhang, "DN-DETR: Accelerate DETR training by introducing query denoising," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 13619–13627.

[7] S. Liu, F. Li, H. Zhang, X. Yang, X. Qi, H. Su, J. Zhu, and L. Zhang, "DAB-DETR: Dynamic anchor boxes are better queries for DETR," in *Proceedings of the International Conference on Learning Representations*, 2022.

[8] H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, L. M. Ni, and H.-Y. Shum, "DINO: DETR with improved denoising anchor boxes for end-to-end object detection," in *Proceedings of the International Conference on Learning Representations*, 2023.

[9] Z. Zong, G. Song, and Y. Liu, "DETRs with collaborative hybrid assignments training," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 6748–6758.

[10] Y. Zhao, W. Lv, S. Xu, J. Wei, G. Wang, Q. Dang, Y. Liu, and J. Chen, "DETRs beat YOLOs on real-time object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 16965–16974.

[11] P. Zhu, L. Wen, D. Du, X. Bian, H. Fan, Q. Hu, and H. Ling, "Detection and tracking meet drones challenge," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 11, pp. 7380–7399, 2022.

[12] F. Yang, H. Fan, P. Chu, E. Blasch, and H. Ling, "Clustered object detection in aerial images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 8311–8320.

[13] C. Yang, Z. Huang, and N. Wang, "QueryDet: Cascaded sparse query for accelerating high-resolution small object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 13668–13677.

[14] Y. Huang, J. Chen, and D. Huang, "UFPMP-Det: Toward accurate and efficient object detection on drone imagery," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 1, 2022, pp. 1026–1033, doi: 10.1609/aaai.v36i1.19986.

[15] B. Du, Y. Huang, J. Chen, and D. Huang, "Adaptive sparse convolutional networks with global context enhancement for faster object detection on drone images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 13435–13444.

[16] K. Liu, Z. Fu, S. Jin, Z. Chen, F. Zhou, R. Jiang, Y. Chen, and J. Ye, "ESOD:

Efficient small object detection on high-resolution images," *IEEE Transactions on Image Processing*, vol. 34, pp. 183–195, 2025, doi: 10.1109/TIP.2024.3501853.

[17] Z. Hu, P. Wu, J. Chen, H. Zhu, Y. Wang, Y. Peng, H. Li, and X. Sun, "Dome-DETR: DETR with density-oriented feature-query manipulation for efficient tiny object detection," in *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025, pp. 101–110, doi: 10.1145/3746027.3754861.

[18] C. Xu, J. Wang, W. Yang, H. Yu, L. Yu, and G.-S. Xia, "Detecting tiny objects in aerial images: A normalized Wasserstein distance and a new benchmark," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 190, pp. 79–93, 2022.

[19] S. Qiao, L.-C. Chen, and A. Yuille, "DetectoRS: Detecting objects with recursive feature pyramid and switchable atrous convolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 10213–10224.

[20] X. Dai, Y. Chen, B. Xiao, D. Chen, M. Liu, L. Yuan, and L. Zhang, "Dynamic Head: Unifying object detection heads with attentions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 7373–7382.

[21] W. Xu and Y. Wan, "ELA: Efficient local attention for deep convolutional neural networks," *arXiv preprint arXiv:2403.01123*, 2024.

[22] K. Li, Q. Geng, M. Wan, X. Cao, and Z. Zhou, "Context and spatial feature calibration for real-time semantic segmentation," *IEEE Transactions on Image Processing*, vol. 32, pp. 5465–5477, 2023, doi: 10.1109/TIP.2023.3318967.

[23] D. Du, Y. Qi, H. Yu, Y. Yang, K. Duan, G. Li, W. Zhang, Q. Huang, and Q. Tian, "The unmanned aerial vehicle benchmark: Object detection and tracking," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 370–386.

[24] J. Suo, T. Wang, X. Zhang, H. Chen, W. Zhou, and W. Shi, "HIT-UAV: A high-altitude infrared thermal dataset for unmanned aerial vehicle-based object detection," *Scientific Data*, vol. 10, Art. no. 227, 2023, doi: 10.1038/s41597-023-02066-6.

[25] S. Yun and Y. Ro, "SHViT: Single-head vision transformer with memory efficient macro design," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 5756–5767.

[26] D. Hendrycks and T. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," in *Proceedings of the International Conference on Learning Representations*, 2019.

[27] C. Michaelis, B. Mitzkus, R. Geirhos, E. Rusak, O. Bringmann, A. S. Ecker, M. Bethge, and W. Brendel, "Benchmarking robustness in object detection: Autonomous driving when winter is coming," *arXiv preprint arXiv:1907.07484*, 2019.