BASFR-NET: LEARNING STAGEWISE RESIDUAL INTERACTIONS FOR RGB-INFRARED ORIENTED OBJECT DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

RGB–infrared oriented object detection in aerial remote sensing must exploit complementary appearance and thermal cues while preserving the spatial detail of small, arbitrarily oriented objects. Existing multimodal object detection methods often optimize a particular cross-modal fusion mechanism, although the role of interaction changes across feature extraction, multiscale aggregation, and prediction. We present BASFR-Net, a compact detector that formulates cross-modal fusion as stagewise residual interaction. It preserves base representations while progressively calibrating modality responses, reconciling semantic and spatial information, and routing local evidence to prediction. BASFR-Net achieves 84.30% and 76.80% mAP@0.5 on DroneVehicle and VEDAI, outperforming a feature-concatenation baseline by 2.00 and 7.60 points, respectively. On DroneVehicle, it reaches 69.64% mAP@0.5:0.95 with 5.06M parameters and 13.38 GFLOPs. Ablations show consistent gains from each interaction stage. These results establish stagewise residual interaction as an effective and compact design principle for RGB–infrared oriented object detection.

1 INTRODUCTION

RGB and infrared provide complementary appearance and thermal cues, but their value varies with illumination, background, and object category. Aerial vehicles are often small, densely packed, and arbitrarily oriented, so detection must preserve spatial detail while combining both streams. The central challenge is to make cross-modal evidence useful throughout the detector, from feature extraction to prediction.

Our paired experiments expose this challenge across datasets (Figure 1). We use direct feature concatenation as the common comparison model, denoted **Baseline**. On DroneVehicle, infrared exceeds RGB by 9.20 mAP@0.5 points, and Baseline adds 2.20 points over infrared. On VEDAI, RGB exceeds infrared by 7.20 points, while Baseline adds only 0.30 points over RGB. The stronger single modality therefore reverses, and concatenating the streams yields markedly different gains. A detector must organize interaction so complementary evidence can contribute beyond the stronger stream.

We formulate this problem as stagewise residual interaction: the evidence used to condition an update and the representation retained after it are specified at each stage. BASFR-Net applies Bidirectional Affine Fusion Guidance (BAFG) to calibrate paired backbone responses, the Spatial-Frequency Adaptive Bridge (SFAB) to reconcile semantic and spatial detail in the neck, and the Cross-Modal Routing Alignment Head (CRAH) to route local evidence before oriented prediction. The modules preserve a base path and use distinct residual updates for their different roles.

Our contribution is a stagewise formulation that assigns response calibration, semantic–detail reassembly, and prediction routing to distinct detector stages, with an explicit retained path and update at each stage. On DroneVehicle and VEDAI, BASFR-Net improves Baseline by 2.00 and 7.60 mAP@0.5 points, respectively; component and sequential ablations show gains from all three stages in a 5.06M-parameter detector.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

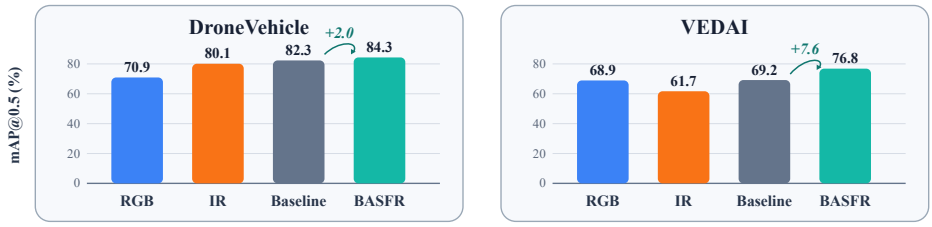


Figure 1: Dataset-level modality preferences. The stronger single modality reverses between DroneVehicle and VEDAI, while Baseline provides unequal gains. BASFR-Net improves Baseline by 2.00 and 7.60 mAP@0.5 points, respectively.

2 Related Work

Multimodal evidence and interaction. Multispectral benchmarks have enabled systematic study of illumination changes, modality imbalance, and cross-modal alignment (Hwang et al., 2015; Sun et al., 2022; Liu et al., 2022; Li et al., 2019; Guan et al., 2019; Zhang et al., 2019b; Zhou et al., 2020). Subsequent methods estimate modality confidence or uncertainty (Li et al., 2023; 2024), learn cross-modal correspondence through attention (Zhang et al., 2019a; Fang and Wang, 2022; Fu et al., 2024; Shen et al., 2024; Yuan and Wei, 2024; Guo et al., 2024), or explicitly address geometric and appearance differences in aerial imagery (Yuan et al., 2022; Chen et al., 2024a; Yuan et al., 2024). Task-level fusion further couples image fusion and detection objectives (Zhang et al., 2024). Recent aerial RGB–infrared work explores modality-specific feature learning, reflectance guidance, and progressive alignment (Zhao et al., 2025a;b; Wu et al., 2025); related designs use disparity-guided state-space fusion or adaptive cross-modal interaction (Zhou et al., 2024; Yu and Shin, 2026). These approaches improve particular aspects of fusion and alignment. BASFR-Net instead assigns different residual interactions to backbone response calibration, neck aggregation, and prediction-stage routing.

Multiscale aggregation and frequency-aware fusion. Feature Pyramid Networks and Path Aggregation Networks establish top-down and bottom-up routes for combining representations across scales (Lin et al., 2017; Liu et al., 2018); multimodal pyramids extend this idea to RGB–infrared features (Zhu et al., 2023). Content-aware reassembly adapts upsampling to local structure (Wang et al., 2019), while frequency-aware fusion and wavelet or state-space designs refine complementary spatial and spectral information (Chen et al., 2024b; Zhu et al., 2025; Wang et al., 2024). Spatially aware fusion also targets efficient aerial feature aggregation (Xiao et al., 2026). SFAB shares the goal of reconciling coarse semantics with fine detail, but makes the two roles explicit through low-pass semantic reassembly and high-pass detail residuals, followed by channel balancing and scene-conditioned modulation.

Oriented detection and prediction heads. Aerial oriented detectors address rotation through region transformation, rotation-equivariant features, and alignment of deep features with rotated boxes (Xia et al., 2018; Ding et al., 2019; Han et al., 2021; 2022; Xie et al., 2021). Task-aligned and dynamic heads further adapt classification and localization features to their prediction objectives (Feng et al., 2021; Dai et al., 2021). These methods primarily address geometric representation or task alignment. CRAH addresses a complementary question: how modality cues should route already fused local evidence immediately before oriented prediction. This stage-specific routing separates cross-modal evidence selection from box geometry and preserves a direct path for the fused features.

3 METHOD

3.1 ARCHITECTURE AND INTERACTION INTERFACES

Given paired RGB and infrared images I_v, I_r, separate YOLO11n (Ultralytics, 2024) backbone streams produce $X_v^l, X_r^l \in \mathbb{R}^{C_l \times H_l \times W_l}$ at strides 2^l, where $l \in \{3, 4, 5\}$. The output consists of categories and oriented boxes $b = (x, y, w, h, \theta)$. Figure 2 follows the feature flow from backbone to prediction. BAFG produces the paired-scale representations F3–F5. The deepest representation passes through SPPF and C2PSA; SFAB then combines coarse semantics with lateral detail at the P4 and P3 top-down junctions. Each bridge is followed by C3k2 refinement. Stride-2 convolution, concatenation, and C3k2 blocks rebuild the bottom-up pyramid. CRAH routes the resulting P3 and P4 features using RGB–infrared cues, while P5 bypasses local routing before cross-scale head processing.

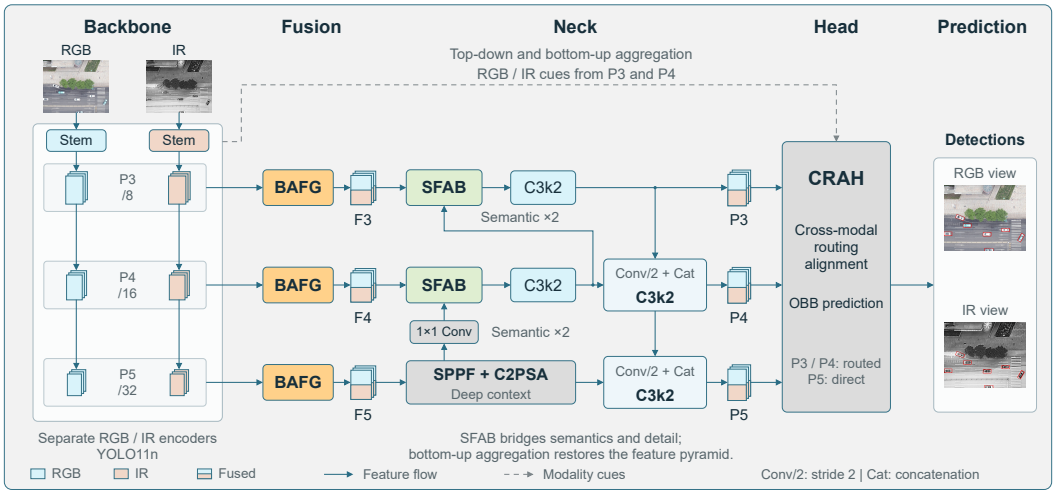


Figure 2: Overall architecture of BASFR-Net. BAFG fuses paired backbone outputs at three scales; SFAB bridges semantics and detail in the top-down neck; CRAH refines the fused pyramid before oriented prediction. Solid arrows denote feature flow and dashed arrows denote modality cues. F3–F5 are backbone fusion outputs; P3–P5 are the prediction pyramid.

The shared design makes the retained representation and adaptive update explicit: BAFG retains both sensor inputs, SFAB retains the reconstructed semantic and detail branches, and CRAH retains the projected fused feature. Sigmoid or tanh parameterizations bound selected coefficients, while softmax normalization constrains local kernel weights. Throughout this paper, *bounded* refers to these coefficients rather than a global feature-norm constraint. The complete architecture is trained with ordinary oriented-detection supervision.

3.2 STAGE I: RESPONSE CALIBRATION WITH BAFG

BAFG addresses response-scale and offset differences at corresponding pyramid levels (Figure 3). Its bidirectional cross-modal affine calibration (CMAC) uses each modality as a reference for the other. Let X_1, X_2 denote the visible and infrared features. Modality j proposes an affine calibration of modality i:

$$\widetilde{X}_i = X_i \odot (1 + \gamma_{j\to i}) + \beta_{j\to i}, \qquad i \neq j. \tag{1}$$

Global average pooling and a two-layer 1×1 convolutional predictor produce the channel-wise parameters. We use a bounded scale $\gamma = 0.5 \tanh(\widehat{\gamma})$ and a scaled shift $\beta = 0.25\widehat{\beta}$. The final convolution is zero-initialized, so the affine proposal initially equals X_i.

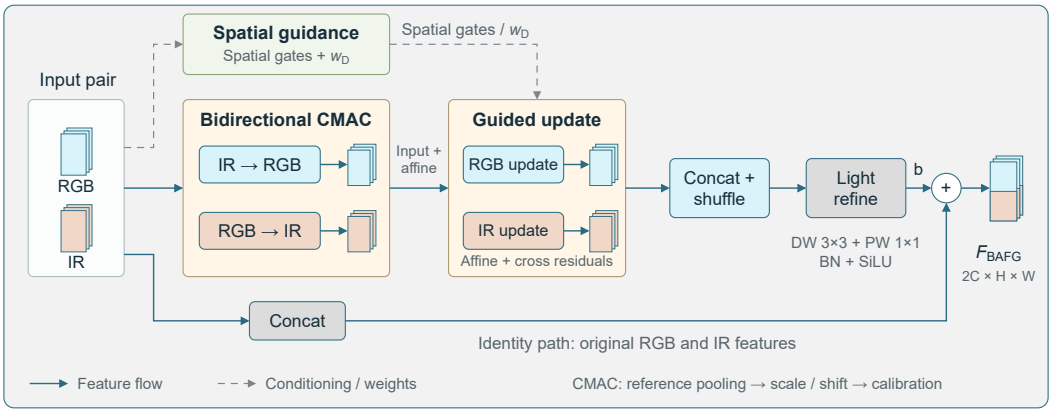


Figure 3: BAFG response calibration. Bidirectional CMAC provides affine proposals; spatial guidance supplies gates and disagreement weights for the guided update. Updated features are concatenated, channel-shuffled, and lightly refined. The original input concatenation remains an additive identity path. Solid and dashed arrows distinguish feature flow from conditioning signals.

To condition the update spatially, a depthwise 5×5 block processes $D = |X_1 - X_2|$. Its output is concatenated with the channel-mean maps of X_1, X_2, D and passed to a sigmoid head, yielding s_1, s_2. The updates and output are

$$\begin{aligned} X_i' &= X_i + a\, s_i \odot \mathcal{D}(\widetilde{X}_i - X_i) + c\, w_D \odot (X_j - X_i), \\ F_{\mathrm{BAFG}} &= \mathrm{Cat}(X_1, X_2) + b\ \mathrm{Ref}(\mathrm{Shuf}(\mathrm{Cat}(X_1', X_2'))), \end{aligned} \tag{2}$$

where $\mathcal{D}$ is channel dropout with rate 0.05 during training and identity at inference. The learned scalars are $a = 0.7\sigma(\eta_a)$, $b = 0.5\sigma(\eta_b)$, and $c = 0.15\sigma(\eta_c)$, initialized with $\eta_a = \eta_b = -4$ and $\eta_c = -6$. Let d be the channel mean of D. The disagreement weight is

$$w_D = 0.5 + \frac{1}{3}\,\mathrm{clip}\left(\frac{d}{\mathrm{mean}_{h,w}(d) + 10^{-6}}, 0, 3\right). \tag{3}$$

Thus $w_D \in [0.5, 1.5]$ and the exchange coefficient cw_D is at most 0.225, with larger exchange weights assigned to regions of stronger disagreement. Shuf interleaves channels; Ref uses a depthwise 3×3 convolution, a pointwise convolution, batch normalization, and SiLU. The unchanged input concatenation remains an additive path at the output. Small initial gates allow calibration and refinement to develop around this retained representation.

3.3 Stage II: Semantic–Detail Reassembly with SFAB

At a top-down junction, the semantic input has lower spatial resolution than the lateral detail input. SFAB uses content-aware reassembly and residual filtering to reconcile them (Figure 4), building on CARAFE and FreqFusion (Wang et al., 2019; Chen et al., 2024b). The figure separates joint kernel encoding, low-pass semantic reassembly, high-pass detail residuals, channel balance, and scene affine modulation. Let D_i and S_i have matching channel counts and a spatial scale ratio of two, as arranged by the neck. Separate 1×1 convolutions compress them to 64-channel kernel-prediction features.

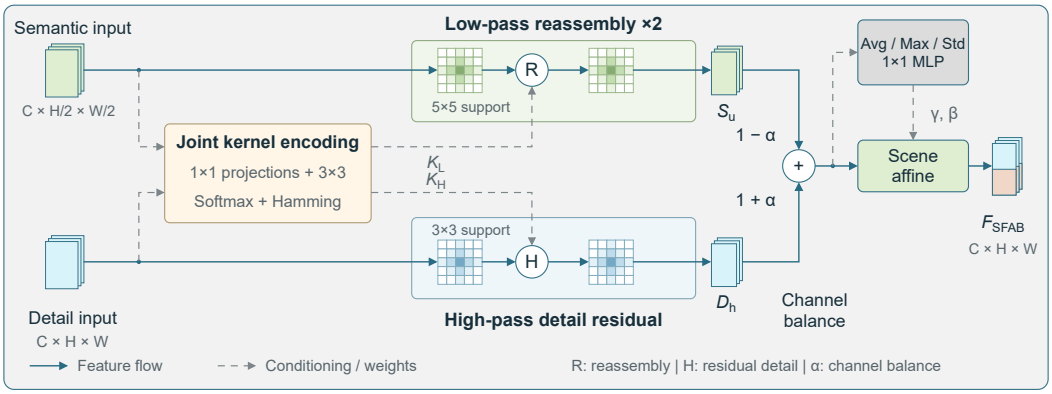


Figure 4: SFAB semantic–detail bridge. Joint kernel encoding conditions $2\times$ semantic reassembly (R) and detail-residual extraction (H). The reconstructed branches S_u, D_h are weighted by $1-\alpha, 1+\alpha$, respectively, and summed. Global average, maximum, and standard-deviation descriptors parameterize the final scene affine modulation.

Two 3×3 generators predict low-pass and detail-smoothing masks. The detail feature first supplies a smoothing residual and an initial semantic reassembly kernel. That kernel upsamples the semantic mask response, which is added to the detail-derived response to obtain the final low-pass mask. The resulting semantic context also updates the detail-smoothing mask. Each mask is locally softmax-normalized, multiplied by a Hamming window, and renormalized. The low-pass kernel K_L has 5×5 support; the smoothing kernel K_H has 3×3 support.

Writing $\mathcal{R}_s$ for content-aware reassembly with scale s, the branches are

$$S_u = \mathcal{R}_2(S_i; K_L), \tag{4}$$

$$L_D = \mathcal{R}_1(D_i; K_H), \qquad D_h = D_i + (D_i - L_D). \tag{5}$$

Both kernels are positive and normalized. The high-frequency residual arises from subtraction, not from a signed high-pass kernel. In particular, D_h retains the original detail and adds its deviation from a learned local average; it is not an explicit Fourier or wavelet decomposition.

The two branches are combined by

$$F_b = (1+\alpha) \odot D_h + (1-\alpha) \odot S_u, \qquad \alpha = 0.1 \tanh(\widehat{\alpha}), \tag{6}$$

$$F_i = (1+\gamma) \odot F_b + \beta, \qquad \gamma = 0.08 \tanh(\widehat{\gamma}). \tag{7}$$

Here α is a learned, input-independent channel parameter, initialized to zero. Its update can be written as $F_b = D_h + S_u + \alpha \odot (D_h - S_u)$: it adjusts the difference between branches around their sum. A descriptor formed from global average, maximum, and standard-deviation pooling of F_b predicts the scene-dependent γ, β. We set $\beta = 0.03 \tanh(\widehat{\beta}) \odot \mathrm{sg}(\mathrm{mean}_{h,w} |F_b|)$, with sg stopping the gradient through the magnitude reference. Zero initialization of the final affine predictor makes the initial output $D_h + S_u$. The branch coefficients remain in $[0.9, 1.1]$ and the scene scale in $[0.92, 1.08]$, giving the bridge explicit control over branch balance and scene modulation.

3.4 Stage III: Local Prediction Routing with CRAH

CRAH separates *where to aggregate* from *what to aggregate* (Figure 5). Its cross-modal routing aggregation (CMRA) stem uses sensor-specific cues to weight a local neighborhood, but samples values exclusively from the fused feature. After channel projection, RGB, infrared, and fused features are centered with 3×3, 7×7, and 5×5 local-average windows, respectively. Five spatially standardized maps encode RGB residual magnitude, positive and negative infrared contrast, fused residual magnitude, and absolute RGB–infrared disagreement. Depthwise context encoders and a prior encoder predict a softmax distribution $A^l(p, q)$ over the 3×3 neighborhood $\mathcal{N}(p)$.

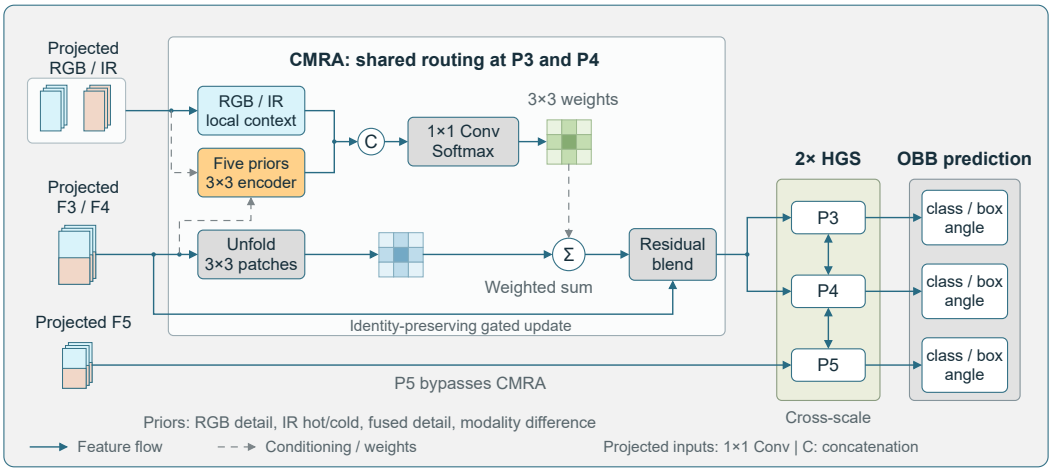


Figure 5: CRAH prediction head. CMRA combines RGB–infrared context and five spatial priors to predict normalized 3×3 routing weights. It aggregates unfolded fused patches and blends the result with the input. P3 and P4 share CMRA; P5 bypasses it. Two hierarchical geometric synergy (HGS) blocks exchange cross-scale features before class, box, and angle prediction.

At P3 and P4, the routing stem computes

$$\widehat{F}^l(p) = (1-g)F^l(p) + g \sum_{q \in \mathcal{N}(p)} A^l(p,q)F^l(q), \qquad g = 0.28\sigma(\eta_g). \tag{8}$$

The same routing stem is shared between these levels. Since A^l is nonnegative and sums to one, the routed value lies in the convex hull of the local fused values, including zero padding at image boundaries. In particular,

$$\|\widehat{F}^l(p) - F^l(p)\|_\infty \le g \max_{q \in \mathcal{N}(p)} \|F^l(q) - F^l(p)\|_\infty. \tag{9}$$

By the triangle inequality and kernel normalization, the routing displacement is controlled by the gate and the local variation of the fused representation.

The gate is initialized to 0.03. A center-logit bias of 4 gives approximately 87% initial weight to the center, making routing initially close to identity. P5 bypasses CMRA, restricting local routing to the two finer levels. Two hierarchical geometric synergy (HGS) blocks then combine same-level and neighboring-level features through residual, task-conditioned updates. Standard classification, box-distribution, and angle branches produce the oriented predictions. CRAH does not estimate a geometric registration transform, and modality features do not bypass the fused values into the predictor.

3.5 Training Objective

The detector uses the YOLO11n-OBB objective

$$\mathcal{L} = \lambda_{\text{box}}\mathcal{L}_{\text{rbox}} + \lambda_{\text{cls}}\mathcal{L}_{\text{cls}} + \lambda_{\text{dfl}}\mathcal{L}_{\text{dfl}}. \tag{10}$$

The predicted angle and distance distribution jointly define the rotated box. The angle is optimized through rotated-box regression rather than a separate angle loss. The interaction modules add no labels or module-specific objective. This keeps the architectural contribution separate from changes to the detection criterion.

4 Experiments

We examine three questions: Does stagewise interaction improve Baseline on two datasets with different single-modal rankings? How do the individual stages and their combination affect accuracy? What computational cost accompanies the improvement?

4.1 Datasets and Evaluation

Datasets. DroneVehicle contains 28,439 paired RGB–infrared images and five oriented vehicle categories (Sun et al., 2022). The data preparation uses the crop100 version, removes the image border, and follows the 17,990/1,469/8,980 train/validation/test split. VEDAI contains 1,246 paired aerial views (Razakarivony and Jurie, 2016); we use the 512-pixel data and report eight categories, excluding plane. Both benchmarks use 640×640 detector inputs and are trained independently.

Metrics. We report class-wise AP and mean AP at IoU 0.5. The DroneVehicle experiments also report mean AP averaged from IoU 0.5 to 0.95. Scores are percentages and differences are percentage points. Within each benchmark, all local configurations use the same paired data, split, detector input, OBB representation, and metric implementation. The component study uses the same DroneVehicle test set throughout.

Implementation. The implementation uses PyTorch 2.2.2 on an NVIDIA RTX 4090. The dataset-specific training parameters are listed in Appendix A. Mosaic and horizontal flipping are applied synchronously to each image pair. BAFG, SFAB, and CRAH are optimized jointly using the objective in Section 3.5.

4.2 Performance across Different Modality Regimes

Table 1 shows that stagewise interaction improves on Baseline in both modality regimes. Baseline reaches 82.30 mAP@0.5 on DroneVehicle and 69.20 on VEDAI, whereas BASFR-Net reaches 84.30 and 76.80. The improvement therefore extends beyond the benefit of supplying both sensor streams.

Table 1: Comparison on DroneVehicle and VEDAI (mAP_{50}, %). RGB+IR is used unless a single modality is indicated; – denotes an unavailable published score. Bold: best in each column.

Method	DroneVehicle	VEDAI
Published methods		
YOLOrs (Sharma et al., 2021)	–	59.73
UA-CMDet (Sun et al., 2022)	64.00	–
CMAFF (Fang and Wang, 2022)	–	68.74
CALNet (He et al., 2023)	75.40	68.09
C2Former (Yuan and Wei, 2024)	72.80	60.82
OAFA (Chen et al., 2024a)	79.40	–
CAGTDet (Yuan et al., 2024)	74.57	–
DMM (S2A) (Zhou et al., 2024)	79.40	64.25
DMM (O-RCNN) (Zhou et al., 2024)	–	72.63
WaveMamba (Zhu et al., 2025)	79.80	–
M2D-LIF (Zhao et al., 2025a)	81.40	–
LPANet (Wu et al., 2025)	77.70	73.69
CMFADet (Yu and Shin, 2026)	83.42	76.28
BDDS (Tang et al., 2026)	83.81	–
Our models		
YOLO11n (RGB)	70.90	68.90
YOLO11n (IR)	80.10	61.70
Baseline	82.30	69.20
BASFR-Net	**84.30**	**76.80**

DroneVehicle. The full model improves Baseline by 2.00 mAP@0.5 points and 2.77 mAP@0.5:0.95 points, reaching 84.30/69.64. Improvement is therefore present both at IoU 0.5 and when averaging over stricter thresholds. Every class improves over Baseline: freight car, truck, and van gain 3.30, 3.10, and 3.20 points, respectively, while the already strong car and bus classes remain stable and increase by 0.10 and 0.60 points.

VEDAI. In Table 1, YOLO11n abbreviates YOLO11n-OBB. The full model improves Baseline by 7.60 points, RGB-only by 7.90 points, and IR-only by 15.10 points. Class-wise results in Appendix Table 5 show that truck, tractor, van, and boat gain 13.10, 8.60, 21.10, and 12.00 points over Baseline, while car and pickup remain within 0.70 points. The gains are concentrated in these four

categories rather than spread uniformly across all eight. Together, the two benchmarks demonstrate that the architecture handles different single-modal rankings without imposing a fixed RGB- or infrared-dominant design.

Published comparisons. For VEDAI, the published eight-class mAP values in Table 1 follow the comparison in Yu and Shin (2026); S2A and O-RCNN denote the S2A-Net and Oriented R-CNN variants of DMM. The four rows under Our models are evaluated under a shared protocol. BASFR-Net reaches 84.30 mAP@0.5 on DroneVehicle and 76.80 on VEDAI, numerically above the highest published entries in Table 1 by 0.49 and 0.52 points, respectively. Appendix B reports the full class-wise results.

4.3 Component Contributions and Sequential Composition

Table 2 evaluates the three intervention points by enabling each stage individually and then composing them sequentially. This distinguishes the contribution of an individual component from the performance of the complete pipeline.

Table 2: Component comparisons on DroneVehicle under the fixed local test protocol. B, S, and C denote BAFG, SFAB, and CRAH. AP_{50} and $AP_{50:95}$ denote class-averaged AP. Complexity is measured for the five-class detector at 640×640.

Configuration	B	S	C	Params (M)	GFLOPs	AP_{50}	$AP_{50:95}$
Baseline	–	–	–	3.87	9.82	82.30	66.87
+ BAFG	✓	–	–	4.44	11.24	83.20	68.12
+ SFAB	–	✓	–	4.32	11.13	83.10	67.95
+ CRAH	–	–	✓	4.04	10.64	83.20	67.85
+ BAFG+SFAB	✓	✓	–	4.89	12.55	83.70	68.48
BASFR-Net	✓	✓	✓	5.06	13.38	**84.30**	**69.64**

Each individual component improves both metrics. BAFG yields gains of 0.90/1.25 points in mAP@0.5/mAP@0.5:0.95, SFAB yields 0.80/1.08, and CRAH yields 0.90/0.98. Their class-wise effects differ: BAFG increases truck and van AP by 1.10 and 2.30 points, SFAB obtains the highest truck AP among the single-component variants (84.10), and CRAH increases freight car AP by 2.20 points. Thus the benefits extend across categories rather than being confined to one isolated gain.

The BAFG+SFAB configuration reaches 83.70/68.48, above every single-component variant. Adding CRAH increases the scores by a further 0.60/1.16 points to 84.30/69.64. The monotonic sequence shows that prediction-stage routing remains beneficial after backbone calibration and neck reassembly.

4.4 Accuracy and Computational Cost

We measure computational cost in GFLOPs at 640×640 input resolution. BASFR-Net uses 5.06M parameters and 13.38 GFLOPs on DroneVehicle. Compared with Baseline, it adds 1.19M parameters (30.7%) and 3.56 GFLOPs (36.3%) for gains of 2.00/2.77 points. CRAH is the least expensive isolated component, adding 0.17M parameters and 0.82 GFLOPs. BAFG and SFAB add 0.57M/1.42 GFLOPs and 0.45M/1.31 GFLOPs, respectively. The complete model remains compact while allocating additional computation to three distinct interaction interfaces.

4.5 Qualitative Detection Examples

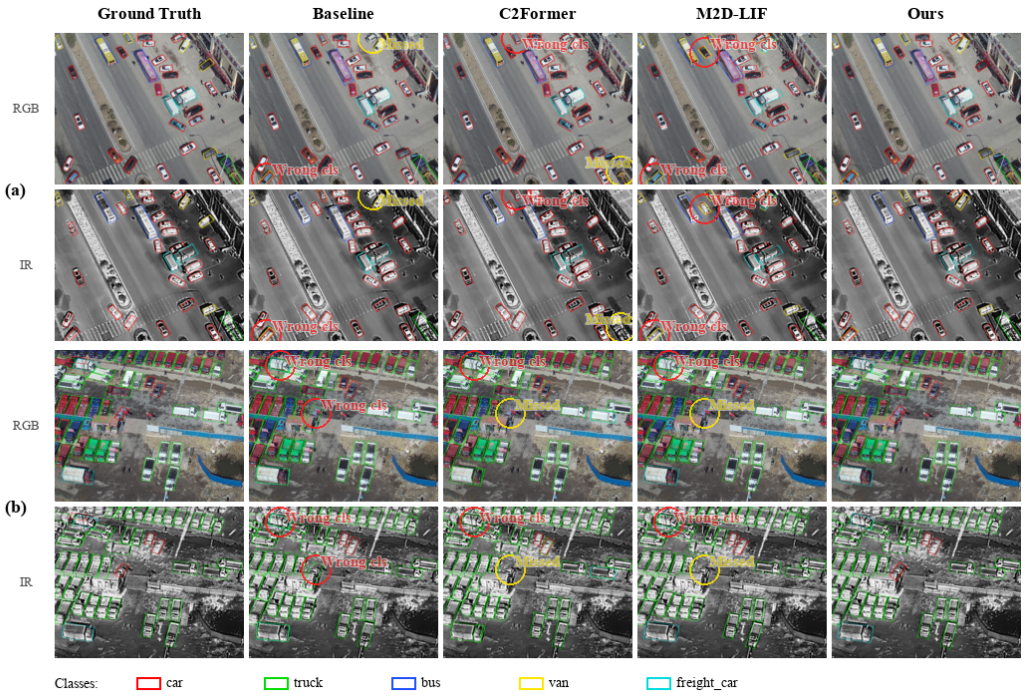


Figure 6: DroneVehicle detection examples: ground truth, Baseline, C2Former, M2D-LIF, and BASFR-Net. RGB/IR rows share predictions; yellow/red mark misses/class errors.

Figure 6 compares two DroneVehicle scenes. Each detector's predictions are overlaid on both sensor views, so the paired RGB and infrared rows show one prediction set rather than separate evaluations. In scene (a), Baseline misses a highlighted vehicle and assigns the wrong class to another. C2Former also misses a marked target, while M2D-LIF makes a class error. BASFR-Net recovers the highlighted instances with their reference categories. In scene (b), tightly parked vehicles with similar appearance lead Baseline to two marked class errors. C2Former and M2D-LIF miss a central vehicle, whereas BASFR-Net detects the highlighted targets with the corresponding classes. These examples complement the DroneVehicle class-wise AP gains for truck, van, and freight car in Appendix Table 4.

5 Discussion and Conclusion

Across its three interfaces, BASFR-Net conditions residual updates on complementary evidence while retaining the representation needed downstream. BAFG calibrates paired responses without discarding their concatenation, SFAB balances semantic and detail features, and CRAH uses modality cues to route the fused stream. The sequential ablation gains show that these roles remain complementary under joint optimization.

On DroneVehicle and VEDAI, whose stronger single modalities differ, BASFR-Net improves Baseline and reaches 84.30 mAP@0.5 and 69.64 mAP@0.5:0.95 on DroneVehicle and 76.80 mAP@0.5 on VEDAI with 5.06M parameters. This supports stagewise residual interaction as a compact design for RGB–infrared oriented detection. Future work will test the retained-path and controlled-update formulation on additional sensor pairs, detectors, and imperfectly aligned observations.

AI use statement. Generative AI tools assisted with manuscript translation, language polishing, and literature search. The authors take responsibility for the final paper.

REFERENCES

Chen, C., Qi, J., Liu, X., Bin, K., Fu, R., Hu, X., and Zhong, P. Weakly misalignment-free adaptive feature alignment for UAVs-based multimodal object detection. In *CVPR*, 2024.

Chen, L., Fu, Y., Gu, L., Yan, C., Harada, T., and Huang, G. Frequency-aware feature fusion for dense image prediction. *IEEE TPAMI*, 46(12):10763–10780, 2024.

Dai, X., Chen, Y., Xiao, B., Chen, D., Liu, M., Yuan, L., and Zhang, L. Dynamic Head: Unifying object detection heads with attentions. In *CVPR*, 2021.

Ding, J., Xue, N., Long, Y., Xia, G.-S., and Lu, Q. Learning RoI Transformer for oriented object detection in aerial images. In *CVPR*, 2019.

Fang, Q. and Wang, Z. Cross-modality attentive feature fusion for object detection in multispectral remote sensing imagery. *Pattern Recognition*, 130:108786, 2022.

Feng, C., Zhong, Y., Gao, Y., Scott, M. R., and Huang, W. TOOD: Task-aligned one-stage object detection. In *ICCV*, 2021.

Fu, H., Wang, S., Duan, P., Xiao, C., Dian, R., Li, S., and Li, Z. LRAF-Net: Long-range attention fusion network for visible-infrared object detection. *IEEE TNNLS*, 35(10):13232–13245, 2024.

Guan, D., Cao, Y., Yang, J., Cao, Y., and Yang, M. Y. Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection. *Information Fusion*, 50:148–157, 2019.

Guo, J., Gao, C., Liu, F., Meng, D., and Gao, X. DAMSDet: Dynamic adaptive multispectral detection Transformer with competitive query selection and adaptive feature fusion. In *ECCV*, 2024.

Han, J., Ding, J., Li, J., and Xia, G.-S. Align deep features for oriented object detection. *IEEE TGRS*, 60:1–11, 2022.

Han, J., Ding, J., Xue, N., and Xia, G.-S. ReDet: A rotation-equivariant detector for aerial object detection. In *CVPR*, 2021.

He, X., Tang, C., Zou, X., and Zhang, W. Multispectral object detection via cross-modal conflict-aware learning. In *ACM MM*, 2023.

Hwang, S., Park, J., Kim, N., Choi, Y., and Kweon, I. S. Multispectral pedestrian detection: Benchmark dataset and baseline. In *CVPR*, 2015.

Li, C., Song, D., Tong, R., and Tang, M. Illumination-aware Faster R-CNN for robust multispectral pedestrian detection. *Pattern Recognition*, 85:161–171, 2019.

Li, Q., Zhang, C., Hu, Q., Fu, H., and Zhu, P. Confidence-aware fusion using Dempster–Shafer theory for multispectral pedestrian detection. *IEEE TMM*, 25:3420–3431, 2023.

Li, Q., Zhang, C., Hu, Q., Zhu, P., Fu, H., and Chen, L. Stabilizing multispectral pedestrian detection with evidential hybrid fusion. *IEEE TCSVT*, 34(4):3017–3029, 2024.

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., and Belongie, S. Feature pyramid networks for object detection. In *CVPR*, 2017.

Liu, J., Fan, X., Huang, Z., Wu, G., Liu, R., Zhong, W., and Luo, Z. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In *CVPR*, 2022.

Liu, S., Qi, L., Qin, H., Shi, J., and Jia, J. Path aggregation network for instance segmentation. In *CVPR*, 2018.

Razakarivony, S. and Jurie, F. Vehicle detection in aerial imagery: A small target detection benchmark. *Journal of Visual Communication and Image Representation*, 34:187–203, 2016.

Sharma, M., Dhanaraj, M., Karnam, S., Chachlakis, D. G., Ptucha, R., Markopoulos, P. P., and Saber, E. YOLOrs: Object detection in multimodal remote sensing imagery. *IEEE JSTARS*, 14:1497–1508, 2021. doi:10.1109/JSTARS.2020.3041316.

Shen, J., Chen, Y., Liu, Y., Zuo, X., Fan, H., and Yang, W. ICAFusion: Iterative cross-attention guided feature fusion for multispectral object detection. *Pattern Recognition*, 145:109913, 2024.

Sun, Y., Cao, B., Zhu, P., and Hu, Q. Drone-based RGB-infrared cross-modality vehicle detection via uncertainty-aware learning. *IEEE TCSVT*, 32(10):6700–6713, 2022.

Tang, Z., Xie, Y., Xu, T., Wu, X.-J., and Kittler, J. Learning bi-directional fusion and deformation-sensitive loss for RGB-T tiny object detection. *Information Fusion*, 128:103985, 2026. doi:10.1016/j.inffus.2025.103985.

Ultralytics. Ultralytics YOLO11. `https://docs.ultralytics.com/models/yolo11/`, 2024.

Wang, J., Chen, K., Xu, R., Liu, Z., Loy, C. C., and Lin, D. CARAFE: Content-aware reassembly of features. In *ICCV*, 2019.

Wang, S., Wang, C., Shi, C., Liu, Y., and Lu, M. Mask-guided Mamba fusion for drone-based visible-infrared vehicle detection. *IEEE TGRS*, 62:1–12, 2024.

Wu, W., Li, C., Wang, X., Luo, B., and Liu, Q. Large language model guided progressive feature alignment for multimodal UAV object detection. arXiv:2503.06948, 2025.

Xia, G.-S., Bai, X., Ding, J., Zhu, Z., Belongie, S., Luo, J., Datcu, M., Pelillo, M., and Zhang, L. DOTA: A large-scale dataset for object detection in aerial images. In *CVPR*, 2018.

Xiao, H., Zhuang, J., Yang, B., Hu, J., Zhu, J., Lian, Z., Zhao, S., and Zhou, Y. SAFF: A spatially aware fusion framework for effective and efficient aerial object detection. *IEEE TGRS*, 2026. doi:10.1109/TGRS.2026.3674467.

Xie, X., Cheng, G., Wang, J., Yao, X., and Han, J. Oriented R-CNN for object detection. In *ICCV*, 2021.

Yu, C. and Shin, Y. A cross-modality feature adaptive interaction approach for RGB-infrared object detection in aerial imagery. *IEEE TGRS*, 2026. doi:10.1109/TGRS.2026.3657379.

Yuan, M., Shi, X., Wang, N., Wang, Y., and Wei, X. Improving RGB-infrared object detection with cascade alignment-guided Transformer. *Information Fusion*, 105:102246, 2024.

Yuan, M., Wang, Y., and Wei, X. Translation, scale and rotation: Cross-modal alignment meets RGB-infrared vehicle detection. In *ECCV*, 2022.

Yuan, M. and Wei, X. C^2Former: Calibrated and complementary Transformer for RGB-infrared object detection. *IEEE TGRS*, 62:1–12, 2024.

Zhang, J., Cao, M., Xie, W., Lei, J., Li, D., Huang, W., Li, Y., and Yang, X. E2E-MFD: Towards end-to-end synchronous multimodal fusion detection. In *NeurIPS*, 2024.

Zhang, L., Liu, Z., Zhang, S., Yang, X., Qiao, H., Huang, K., Hussain, A. Cross-modality interactive attention network for multispectral pedestrian detection. *Information Fusion*, 50:20–29, 2019.

Zhang, L., Zhu, X., Chen, X., Yang, X., Lei, Z., and Liu, Z. Weakly aligned cross-modal learning for multispectral pedestrian detection. In *ICCV*, 2019.

Zhao, T., Liu, B., Gao, Y., Sun, Y., Yuan, M., and Wei, X. Rethinking multi-modal object detection from the perspective of mono-modality feature learning. In *ICCV*, 2025.

Zhao, Z., Zhang, W., Xiao, Y., Li, C., and Tang, J. Reflectance-guided progressive feature alignment network for all-day UAV object detection. *IEEE TGRS*, 63:1–15, 2025.

Zhou, K., Chen, L., and Cao, X. Improving multispectral pedestrian detection by addressing modality imbalance problems. In *ECCV*, 2020.

Zhou, M., Li, T., Qiao, C., Xie, D., Wang, G., Ruan, N., Mei, L., and Yang, Y. DMM: Disparity-guided multispectral Mamba for oriented object detection in remote sensing. arXiv:2407.08132, 2024.

Zhu, H., Dong, W., Yang, L., Li, H., Yang, Y., Ren, Y., Zhu, Q., Feng, Z., Li, C., Lin, S., Wang, R., Luo, X., and Zhang, B. WaveMamba: Wavelet-driven Mamba fusion for RGB-infrared object detection. In *ICCV*, 2025.

Zhu, Y., Sun, X., Wang, M., and Huang, H. Multi-modal feature pyramid Transformer for RGB-infrared object detection. *IEEE TITS*, 24(9):9984–9995, 2023.

A Training Parameters

Table 3: Training parameters on the two datasets.

Parameter	DroneVehicle	VEDAI
Input size	640×640	640×640
Training epochs	300	300
Optimizer	SGD	SGD
Initial learning rate	0.01	0.01
Momentum	0.937	0.937
Weight decay	0.001	0.001
Batch size	16	4

B Complete Quantitative Results

Table 4: Class-wise AP@0.5 on DroneVehicle. A dash denotes an unreported value, and bold marks the best score in each column.

Method	Modality	Car	Freight	Truck	Bus	Van	mAP
RetinaNet	RGB	67.50	13.72	28.24	62.05	19.26	38.16
Faster R-CNN	RGB	67.88	26.31	38.59	66.98	23.20	44.59
Oriented R-CNN	RGB	80.20	41.57	53.83	85.43	43.32	60.85
S2A-Net	RGB	80.03	42.24	54.17	84.92	43.81	61.03
ReDet	RGB	69.48	31.46	47.87	77.37	29.03	51.04
YOLO11n-OBB	RGB	93.90	51.80	65.00	91.20	52.50	70.90
RetinaNet	IR	79.86	28.05	32.84	67.32	16.44	44.90
Faster R-CNN	IR	88.63	35.16	42.51	77.92	28.52	54.55
Oriented R-CNN	IR	89.82	53.12	57.37	89.25	45.32	66.98
S2A-Net	IR	89.93	55.81	54.52	88.87	48.42	67.51
ReDet	IR	89.47	42.82	53.95	79.89	36.56	60.54
YOLO11n-OBB	IR	98.30	66.30	78.90	95.30	61.70	80.10
UA-CMDet	RGB+IR	87.50	46.80	60.70	87.10	38.00	64.00
C2Former	RGB+IR	90.00	57.60	72.10	88.70	55.40	72.80
CALNet	RGB+IR	90.30	63.00	76.20	89.10	58.50	75.40
OAFA	RGB+IR	90.30	**73.30**	76.80	90.30	66.00	79.40
CAGTDet	RGB+IR	–	–	–	–	–	74.57
DMM	RGB+IR	90.40	68.20	79.80	89.90	68.60	79.40
LPANet	RGB+IR	–	–	–	–	–	77.70
WaveMamba	RGB+IR	95.00	68.50	80.40	90.60	64.50	79.80
M2D-LIF	RGB+IR	97.80	67.90	81.00	96.00	64.60	81.40
CMFADet	RGB+IR	**98.60**	71.60	82.90	96.20	67.80	83.42
Baseline	RGB+IR	98.50	68.90	82.20	96.10	65.60	82.30
BASFR-Net	RGB+IR	**98.60**	72.20	**85.30**	**96.70**	**68.80**	**84.30**

Table 5: Class-wise AP@0.5 on VEDAI (eight categories). Bold marks the best score in each column.

Method	Mod.	Car	Truck	Tractor	Camping	Van	Pickup	Boat	Others	mAP
RetinaNet	RGB	70.87	54.74	50.47	59.36	47.45	62.98	17.66	19.78	47.91
Faster R-CNN	RGB	71.40	54.20	61.00	70.50	59.50	67.60	52.30	40.10	59.58
RoI Transformer	RGB	77.30	56.10	64.70	73.60	60.20	71.50	56.70	42.80	62.86
S2A-Net	RGB	74.50	47.30	55.60	61.70	32.50	65.10	16.70	39.80	49.15
YOLO11n-OBB	RGB	85.70	75.30	76.20	75.90	58.00	70.60	54.50	54.80	68.90
RetinaNet	IR	59.66	52.81	35.30	53.38	42.29	54.09	9.57	24.89	41.50
Faster R-CNN	IR	71.60	49.10	49.20	68.10	57.00	66.50	35.60	29.50	53.33
RoI Transformer	IR	76.10	51.70	51.90	71.20	64.30	70.70	46.90	28.30	57.64
S2A-Net	IR	73.00	39.20	41.90	59.20	32.30	65.60	13.90	23.10	43.52
YOLO11n-OBB	IR	83.80	73.50	47.30	63.40	73.70	74.90	44.00	33.00	61.70
C2Former	RGB+IR	81.20	59.40	37.70	77.60	53.70	78.30	52.60	46.10	60.82
YOLOrs	RGB+IR	84.15	52.60	67.88	68.81	57.91	78.27	21.47	46.75	59.73
CMAFF	RGB+IR	81.70	58.80	68.70	78.40	68.50	76.30	66.00	51.50	68.74
CALNet	RGB+IR	82.30	66.70	73.00	75.40	65.50	78.20	56.60	47.00	68.09
DMM (S2A-Net)	RGB+IR	77.90	59.30	68.10	70.80	57.40	75.80	61.20	43.50	64.25
DMM (Oriented R-CNN)	RGB+IR	84.20	65.70	72.30	**79.00**	72.50	78.80	**72.30**	56.20	72.63
LPANet	RGB+IR	82.40	70.60	77.00	78.40	75.40	**80.00**	69.00	56.70	73.69
CMFADet	RGB+IR	**88.20**	**85.30**	79.60	77.60	77.10	78.20	62.80	61.40	76.28
Baseline	RGB+IR	85.90	67.10	73.60	66.80	67.70	75.20	59.20	58.50	69.20
BASFR-Net	RGB+IR	85.20	80.20	**82.20**	69.90	**88.80**	75.10	71.20	**61.80**	**76.80**

Table 6: Class-wise DroneVehicle ablation.

Configuration	Car	Freight	Truck	Bus	Van	mAP@0.5	mAP@0.5:0.95
Baseline	98.50	68.90	82.20	96.10	65.60	82.30	66.87
+ BAFG	98.60	69.90	83.30	96.40	67.90	83.20	68.12
+ SFAB	98.50	69.30	84.10	96.50	67.00	83.10	67.95
+ CRAH	98.50	71.10	83.50	96.50	66.50	83.20	67.85
+ BAFG+SFAB	98.60	70.80	84.70	96.50	67.80	83.70	68.48
BASFR-Net	**98.60**	**72.20**	**85.30**	**96.70**	**68.80**	**84.30**	**69.64**