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.

Rotated-Box Multi-Modal Vehicle Detection in Aerial Imagery: A Systematic Study with an Empirical Re-examination of Modality Misalignment

Yi Wu

wuyifortheartical@gmail.com

Jiangxi Institute of Applied Science and Technology

Article

Keywords:

Posted Date: November 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11372567/v1

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

Additional Declarations: No competing interests reported.

Rotated-Box Multi-Modal Vehicle Detection in Aerial Imagery: A Systematic Study with an Empirical Re-examination of Modality Misalignment

Yi Wu

Jiangxi Institute of Applied Science and Technology, Nanchang, China
wuyifortheartical@gmail.com

Abstract

Vehicle detection in aerial imagery suffers from small scales, arbitrary orientations, and limited single-modality information. Multi-modal RGB–IR fusion is widely adopted, yet the field rests on two tacit assumptions rarely verified: (1) paired RGB and IR images are spatially misaligned, making alignment the main obstacle to fusion; and (2) IR assists fine-grained class discrimination. Using a YOLOv8s-based rotated-box multi-modal detector, we re-examine both assumptions on VEDAI. Phase-correlation measurements show a median misalignment of only 0.078 pixels, indicating the dataset is well registered; permutation tests show IR adds no significant discriminative power for the confusable car–pickup pair ($p = 0.81$). Data profiling instead reveals the true challenge: class confusion among extremely small objects (95.3% smaller than 32×32 pixels), primarily geometric. The framework nonetheless yields consistent gains: the rotated-box head improves mAP50-95 by 12.86 points and IR adds 5.79 points (paired t-test $p < 0.0001$ and $p = 0.0026$). An explicit alignment module (BMA) brings no significant gain on these well-registered datasets (eight seeds; equivalence test and power analysis; mean +0.44 mAP50, $p = 0.563$). Reproducing this on DroneVehicle, alignment again yields no reliable gain (+0.21 mAP50-95), indicating the null holds across two independently registered datasets; the misaligned regime is tested only synthetically.

Keywords: vehicle detection; aerial imagery; rotated object detection; multi-modal fusion; modality misalignment; small object; RGB-infrared; YOLO

1 Introduction

Vehicle detection in aerial and remote sensing imagery is a fundamental task with broad applications in traffic monitoring, urban planning, and emergency response. This task still faces three challenges. **First**, vehicles are small objects that often occupy only tens of pixels in high-resolution aerial images, carrying little information and being prone to missed detection. **Second**, vehicle orientations are arbitrary; axis-aligned horizontal bounding boxes (HBB) cannot tightly enclose tilted vehicles and thus introduce substantial background interference. **Third**, a single imaging modality is often insufficient: visible-light images degrade severely under weak illumination and shadow, while infrared images lack texture and color information, making it hard to distinguish objects with similar appearance.

Existing detectors typically address only one of these difficulties in isolation. Rotated object detection precisely encloses tilted objects by additionally predicting a rotation angle [1, 2], but the vast majority of rotated-box detectors operate on a single modality; multi-modal fusion improves robustness under varying illumination by combining RGB and IR images [3, 4], but most fusion detectors still use horizontal boxes.

It is worth noting that there exist two commonly assumed but under-verified assumptions in the field. **First**, multi-modal fusion methods generally assume by default that "paired RGB and IR images are spatially misaligned, and that this misalignment is the main obstacle to fusion performance" — the recent work FGMF [5] even explicitly names modality misalignment as a key obstacle and designs an implicit calibration module to alleviate it. **Second**, it is widely believed that "the IR modality can provide complementary information for fine-grained class discrimination." However, whether these two assumptions hold on a specific dataset depends on how the data are acquired and registered, and cannot be generalized. In this paper, we argue that, rather than stacking ever more complex alignment or discrimination modules on top of these assumptions, it is more valuable to first test the assumptions themselves with rigorous, quantifiable experiments — this not only provides a reliable starting point for method design, but also offers the field empirical evidence about whether these two assumptions hold on a particular dataset.

To this end, this paper constructs a rotated-box multi-modal detection framework that jointly addresses the two limitations of "orientation" and "modality," and conducts a systematic empirical study on top of it, focusing on testing the two assumptions above. Our main contributions are fivefold:

1. **We empirically re-examine the "modality misalignment" assumption.** By measuring the spatial misalignment of all 1210 RGB–IR image pairs of VEDAI using phase correlation, cross-validated by a calibration experiment with synthetic misalignment, we find that the median misalignment is only 0.078 pixels with a systematic component of merely 0.004 pixels, thereby demonstrating that VEDAI is well registered and misalignment is not its primary bottleneck — providing empirical evidence that "modality misalignment" is not a universally applicable bottleneck. On a second, larger dataset (DroneVehicle), explicit bidirectional alignment again yields no reliable gain (+0.21 mAP50-95 points over three seeds), providing preliminary cross-dataset evidence consistent with this conclusion.

2. **We reveal the true challenge of VEDAI through data profiling, a confusion matrix, and discriminability analysis.** 95.3% of objects are extremely small (smaller than 32×32 pixels), and bidirectional car–pickup confusion accounts for 21.9% of all objects; discriminability experiments (with permutation tests) further show that this confusion is, under a linear 1-NN separability probe, primarily geometric in nature (area and aspect ratio, 71.8% classification accuracy, only about 2 points above texture, $p = 0.003$) rather than textural, and that IR contributes no significant incremental discriminative power ($p = 0.81$).

3. **We build a unified rotated-box multi-modal detection framework**, composed of a symmetric dual-branch backbone, a cosine-similarity-based fusion (CSIFF) module, and a rotated-box detection head, jointly addressing the orientation and modality limitations. The rotated-box head improves mAP50-95 by 12.86 points over the horizontal-box baseline, and introducing IR through multi-modal fusion further improves it by 5.79 points (means over three random seeds; paired t-test $p < 0.0001$ and $p = 0.0026$, respectively), constituting a dominant and consistent performance gain that we verify across random seeds.

4. **We explore an explicit bidirectional modality alignment (BMA) module and report its results honestly.** Multi-seed experiments (eight random seeds) show that BMA's gain is not statistically significant (mean +0.44 mAP50 points, $p = 0.563$); this result corroborates the conclusion of our assumption testing and provides honest empirical guidance for future work. A three-seed reproduction on DroneVehicle likewise yields no practically meaningful gain (+0.21 mAP50-95 points over three seeds), providing preliminary cross-dataset evidence consistent with this conclusion.

5. **We contribute a reusable statistical protocol for rigorous negative-result reporting.** To distinguish "absence of evidence" from "evidence of absence" — a distinction most detection papers fail to make — we formalize a multi-seed evaluation pipeline (eight random seeds) combined with paired bootstrap and permutation tests, a two-one-sided equivalence test

(TOST) with an explicit equivalence margin, a power analysis reporting the minimum detectable effect (2.64 mAP50 points at 80% power), and a leave-one-seed-out sensitivity analysis. This protocol is a methodological contribution in its own right and is transferable to any small-sample detection benchmark (Section 4.10 applies a reduced three-seed variant, whose limited power is reported explicitly).

The remainder of this paper is organized as follows: Section 2 reviews related work; Section 3 introduces the proposed method; Section 4 presents experimental results and ablation analysis; Section 5 discusses the main findings; and Section 6 concludes the paper.

2 Related Work

2.1 Vehicle Detection in Aerial Imagery

Vehicle detection in aerial imagery has been widely studied. Early methods relied on handcrafted features, while recent years have been dominated by deep convolutional detectors. For large-scale remote sensing images, SANet [6] introduces a situation-aware module to filter invalid regions and a region-adaptive module to dynamically assign detectors. The VEDAI dataset [7] is a widely used benchmark that provides paired RGB and infrared aerial images with rotated-box annotations, serving as a standard testbed for small-object and rotated detection.

2.2 Multi-Modal Fusion

Fusing visible and infrared images can improve detection performance under varying illumination. Common strategies include early fusion (channel concatenation), intermediate feature fusion, and late fusion. MMYFnet [3] decomposes visible and infrared features via cosine similarity and demonstrates that RGB–IR fusion consistently improves accuracy. However, the vast majority of fusion methods implicitly assume that the two modalities are already strictly aligned, ignoring the inherent spatial misalignment between RGB and IR. To address this, the recent work FGMF [5] (based on the S2ANet detector) explicitly names modality misalignment as a key obstacle to efficient fusion, and proposes a dual-enhancement fusion module that performs **implicit calibration** at the feature level to obtain corrected complementary features. While implicit calibration is effective, its learning process lacks geometric constraints: the direction and magnitude of the offsets are uninterpretable, and its effectiveness depends heavily on the indirect supervision of the downstream detection loss. In contrast, this paper proposes **explicit bidirectional offset-field alignment**: we directly predict per-pixel two-dimensional offset fields and geometrically warp the features, which has clear geometric semantics and enables intuitive diagnosis of misalignment patterns through offset-field visualization. Compared with implicit calibration, explicit bidirectional alignment is more direct and more controllable.

2.3 Rotated Object Detection

Rotated object detection encloses tilted objects better by predicting rotated boxes; representative datasets include DOTA [1]. ProbIoU [2] models rotated boxes as Gaussian distributions, providing a differentiable, angle-continuous loss. The rotated-box detection head of this paper follows this idea, realized by adding an angle branch to YOLOv8s and optimizing with ProbIoU.

2.4 Comparison with Existing Methods

Table 1 compares our method with representative related work along four dimensions: box type, modality, fusion strategy, and modality-misalignment handling.

Table 1: Comparison between our method and representative related work

Method	Box type	Modality	Fusion strategy	Misalignment handling
SANet [6]	Horizontal	Single-modal RGB	Situation-aware region assignment	None (single-modal)
MMYFNet [3]	Horizontal	Multi-modal RGB–IR	CSIFF cosine-similarity decomposition	Unhandled (assumes aligned)
FGMF [5]	Rotated	Multi-modal RGB–IR	Dual-enhancement fusion	Implicit calibration
Ours (BMA)	**Rotated**	**Multi-modal RGB–IR**	**Dual-branch + CSIFF**	**Explicit bidirectional offset-field alignment**

As shown, existing methods either address only one of "orientation" (rotated box) or "modality" (multi-modal fusion), or assume the two modalities are already aligned during fusion, or adopt only implicit calibration. In contrast, our method unifies rotated boxes, multi-modal fusion, and explicit bidirectional modality alignment into a single framework, with the key distinction of handling modality misalignment in an explicit, bidirectional, and geometrically interpretable manner.

2.5 Modality Alignment and Image Registration

The core module of this paper, BMA, is essentially "learning a spatial transformation to align the features of two modalities," which is closely related to the long-standing lineage of "learnable spatial transformations" and "image registration" developed in computer vision and medical image analysis. This subsection surveys this lineage to clarify the positioning of BMA.

Learnable spatial transformations. Spatial Transformer Networks (STN) [15] were the first to apply spatial transformations to feature maps inside a network in a differentiable manner: a localization network predicts affine transformation parameters, and a sampler (grid sampling) warps the features, endowing the network with explicit spatial invariance to translation, rotation, and scaling. Deformable Convolutional Networks (DCN) [16] and their improvement DCNv2 [17] learn per-sampling-point offsets so that the sampling positions of the convolution kernel deform adaptively, thereby implicitly modeling local geometric transformations.

Dense displacement estimation (optical flow). Optical flow estimation aims to predict per-pixel displacement fields between two frames, with representative learned methods including FlowNet [18] and RAFT [19]. Optical flow is formally consistent with our offset field — both are dense two-dimensional displacement maps — but optical flow targets motion estimation in temporal images, whereas our offset field targets cross-modal registration.

Deformable image registration. In medical image analysis, deformable registration aligns images by warping a source image onto a target image, with representative methods such as VoxelMorph [20]; symmetric registration methods (such as SyN [21]) estimate the deformation of two images toward an intermediate space by imposing bidirectional symmetric constraints, thereby avoiding the bias of selecting a reference space.

Within this lineage, the positioning of BMA can be summarized in three points: (1) unlike the global affine transformation of STN, BMA predicts a **per-pixel dense offset field** that can express non-rigid local misalignment; (2) unlike the implicit sampling offsets of DCN, BMA's

offset field is **explicit** and can be visualized and diagnosed (see Section 4.5); (3) unlike unidirectional registration, BMA adopts a **bidirectional symmetric** design, borrowing from symmetric registration ideas such as SyN, and constrains the reference anchor of the offset field through zero initialization. This lineage positioning both shows that BMA does not come out of nowhere, and highlights its concrete increment in the multi-modal detection task.

3 Method

3.1 Overall Architecture

The proposed framework is designed around the two limitations of arbitrary orientation and insufficient single-modality information in aerial vehicle detection, and additionally introduces a bidirectional modality alignment (BMA) module to exploratorily test the common "modality misalignment" assumption in the field (results in Section 4.4). As shown in Figure 1, the network consists of four parts: (1) a symmetric dual-branch backbone that extracts multi-scale features from paired RGB and IR images; (2) a bidirectional modality alignment (BMA) module that compensates for RGB–IR misalignment at every scale; (3) a CSIFF fusion module that fuses the aligned features; and (4) a rotated-box detection head.

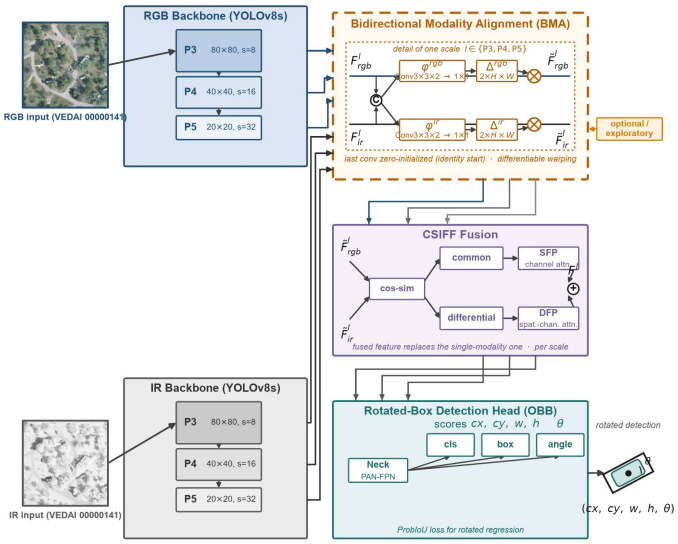


Figure 1: Overall architecture of the proposed rotated-box multi-modal detection network: symmetric dual-branch backbone → bidirectional modality alignment (BMA, optional module) → CSIFF fusion → rotated-box detection head.

3.2 Dual-Branch Feature Extraction Backbone

This paper adopts YOLOv8s as the backbone, which outputs feature maps at three scales (P3, P4, P5, with downsampling factors of 8, 16, and 32, respectively). The RGB and IR branches use structurally identical backbones, both initialized with COCO pre-trained weights. This symmetric design preserves the low-level features of each modality (IR is sensitive to heat sources, while RGB is sensitive to texture and color), avoiding information loss caused by premature fusion.

3.3 Bidirectional Modality Alignment (BMA)

It is commonly assumed in the field that RGB and IR images are spatially misaligned due to sensor differences, and that this misalignment impairs fusion detection. To test this assumption and explore the feasibility of an explicit alignment scheme, this paper introduces a bidirectional modality alignment (BMA) module that geometrically aligns the two modalities before fusion. It should be noted that the misalignment diagnosis in Section 4.4 shows that the RGB–IR misalignment of VEDAI is extremely small; therefore, BMA is positioned in this paper as an "exploratory attempt" rather than a "necessary component" on this dataset.

A natural question is why **bidirectional** rather than unidirectional alignment is used. Unidirectional alignment (e.g., warping only IR to RGB) imposes the entire misalignment error on a single modality: it not only requires the IR branch to bear its own geometric deviation alone, but also implicitly assumes that the RGB side is an "unbiased" reference — an assumption that does not hold when the RGB side itself has annotation or registration bias. In contrast, bidirectional alignment lets each modality learn its own offset field and move toward a common intermediate space together, thereby distributing the misalignment error across the two modalities and compensating for the bias of each modality separately. This design borrows the idea of **symmetric registration** [21]: by using bidirectional constraints rather than a unidirectional mapping, the relative deformation between the two can be estimated and corrected more robustly. It is worth noting that the zero-initialization anchor of our offset field plays a role analogous to the reference-space selection in symmetric registration: just as SyN-style methods fix an intermediate "mean" space to resolve the non-uniqueness (identifiability) of the two displacement fields, our zero-initialization fixes the shared origin to which both modality feature maps are expected to converge, thereby removing the gauge ambiguity that would otherwise make the two offset fields jointly unidentifiable.

It should be pointed out that BMA's differentiable warp is mechanistically homologous to STN [15]'s grid sampling, but the two are fundamentally different: STN learns global affine transformation parameters, whereas BMA predicts a per-pixel dense offset field that can characterize non-rigid local misalignment that STN struggles to express; compared with DCN [16]'s implicit sampling offsets, BMA's offset field is explicit and can be visualized and diagnosed (see Section 4.5).

For paired features $F^l_{rgb}, F^l_{ir} \in \mathbb{R}^{C\times H\times W}$ at scale l, we first concatenate them and predict two offset fields with lightweight convolutional heads:

$$\Delta^{rgb} = \phi^{rgb}\big([F^l_{rgb}, F^l_{ir}]\big), \qquad \Delta^{ir} = \phi^{ir}\big([F^l_{rgb}, F^l_{ir}]\big) \tag{1}$$

where $\Delta^{rgb}, \Delta^{ir} \in \mathbb{R}^{2\times H\times W}$ denote the per-pixel two-dimensional offsets in the RGB and IR directions, respectively. The two prediction heads ϕ^{rgb} and ϕ^{ir} share the same structure but have independent parameters: the input is the concatenated $2C$-channel feature, which passes through two 3×3 convolutional layers (with the channel count halved layer by layer) with ReLU activation, followed by a 1×1 convolution that outputs a 2-channel offset map, predicting per-pixel offsets at the original feature resolution $H\times W$. The aligned features are obtained through differentiable spatial warping:

$$\tilde{F}^l_{rgb} = \mathrm{warp}\big(F^l_{rgb}, \Delta^{rgb}\big), \qquad \tilde{F}^l_{ir} = \mathrm{warp}\big(F^l_{ir}, \Delta^{ir}\big) \tag{2}$$

where $\mathrm{warp}(\cdot)$ denotes grid sampling, with the sampling grid constructed after normalizing the offset field to $[-1, 1]$. The bidirectional design aligns the two modalities to a common intermediate space simultaneously, which is more effective than unidirectional alignment — because the misalignment error is distributed across the two modalities rather than imposed on a single modality.

It should be noted that the bidirectional offset field has an identifiability issue: when RGB shifts left by k and IR shifts right by k, their relative relationship is equivalent to "moving only

one of the modalities," i.e., the offsets can be decomposed arbitrarily. To eliminate this degree of freedom, we initialize the last layer of both offset heads to zero, so that the module starts from an "identity mapping (no alignment)" and implicitly uses the original feature positions as the reference anchor — the offsets are nearly zero in the early stage of training, thereby avoiding the arbitrary decomposition of the offset field and ensuring training stability.

3.4 CSIFF Fusion

After alignment, the features are fused by the cosine-similarity-based image feature fusion (CSIFF) module. CSIFF decomposes the features into "common components" and "differential components" according to inter-channel cosine similarity, and processes them with dedicated modules. For channel c, the similarity is defined as:

$$\cos_c^l = \frac{F_{rgb,c}^l \cdot F_{ir,c}^l}{\|F_{rgb,c}^l\| \cdot \|F_{ir,c}^l\| + \epsilon} \tag{3}$$

The common components are processed by a shared feature processing module (SFP, based on channel attention) to suppress redundancy, and the differential components are processed by a differential feature processing module (DFP, multi-branch spatial-channel attention) to enhance small and elongated objects. The fused output replaces the single-modality feature and is fed into the Neck and the detection head.

3.5 Rotated-Box Detection Head

Vehicle orientations are arbitrary, and horizontal boxes enclose substantial background. This paper extends the YOLOv8 detection head into a rotated-box detection head by adding an angle prediction branch. Given the angle logit x_θ, the angle is mapped as:

$$\theta = (\text{sigmoid}(x_\theta) - 0.25) \cdot \pi, \qquad \theta \in \left[-\frac{\pi}{4}, \frac{3\pi}{4}\right] \tag{4}$$

The detection head outputs a rotated box of the form (center point (cx, cy), size (w, h), angle θ).

3.6 Loss Function

The total loss consists of three terms:

$$\mathcal{L} = \lambda_{cls}\mathcal{L}_{cls} + \lambda_{dfl}\mathcal{L}_{dfl} + \lambda_{obb}\mathcal{L}_{obb} \tag{5}$$

where $\mathcal{L}_{cls}$ is the binary cross-entropy classification loss combined with the task-aligned assigner, $\mathcal{L}_{dfl}$ is the distribution focal loss, and $\mathcal{L}_{obb}$ is the ProbIoU loss. The three weights are set to 0.5, 1.5, and 7.5, respectively.

Use of generative AI

During the preparation of this manuscript, the author used a large language model (DeepSeek) to draft and polish portions of the English text and for Chinese-to-English translation. The author reviewed, verified, and edited all AI-assisted text and takes full responsibility for the accuracy and integrity of the final manuscript. No figure was generated with AI: all figures, including the architecture diagram (Figure 1), were produced programmatically from the author's own plotting code, and all quantitative results were obtained directly from the experiments.

4 Experimental Results

4.1 Dataset and Experimental Setup

This paper evaluates on VEDAI [7], which contains 1246 paired registered RGB and infrared aerial images (resolution 512×512) with rotated-box annotations. Following the official fold01 split, 1089 pairs are used for training and 121 pairs for validation. Experiments are reported under two settings: the **single-class setting** (all vehicles merged into one class) for stepwise-gain analysis; and the **standard multi-class setting** (9 classes: car, truck, tractor, camping car, van, other, pickup, boat, plane) for state-of-the-art comparison.

The backbone is COCO pre-trained YOLOv8s; SGD optimizer, learning rate 0.01, batch size 32 (16 for large models), input resolution 640, Mosaic augmentation, trained for 100 epochs. The evaluation metrics are mAP50 and mAP50-95.

4.2 Main Results

Table 2 reports the stepwise gains of the three components under the single-class setting, each averaged over three random seeds (seed 0–2, reported as mean ± standard deviation).

Table 2: Stepwise gains on VEDAI (single-class vehicles, mean ± std over three random seeds)

Configuration	**mAP50**	**mAP50-95**
Single-modal RGB + HBB	0.9142 ± 0.0021	0.5157 ± 0.0037
Single-modal RGB + OBB	0.9213 ± 0.0028	0.6443 ± 0.0026 (+12.86)
Early fusion (RGB+IR concat + OBB)	0.9516 ± 0.0032	0.7022 ± 0.0017 (+5.79)
Ours (dual-branch + CSIFF + OBB)	0.9511 ± 0.0030	0.7052 ± 0.0018 (+0.30)

As shown in Figure 2, the rotated-box detection head improves mAP50-95 by 12.86 points (paired t-test $p < 0.0001$), proving that rotated boxes are crucial for tilted vehicles; multi-modal fusion brings a further gain of 5.79 points ($p = 0.0026$), reflecting the complementary value of RGB and IR; the dual-branch + CSIFF further improves mAP50-95 by 0.30 points over naive early fusion (0.7022→0.7052), but slightly decreases mAP50 (0.9516→0.9511), indicating that the benefit of dual-branch fusion mainly lies in the stricter localization metric. On mAP50, the rotated-box head and multi-modal fusion likewise bring significant gains (+0.71 points, $p = 0.014$, and +3.03 points, $p = 0.0031$, respectively). These stepwise gains are statistically significant across three random seeds, confirming that they are robust rather than artifacts of a single run.

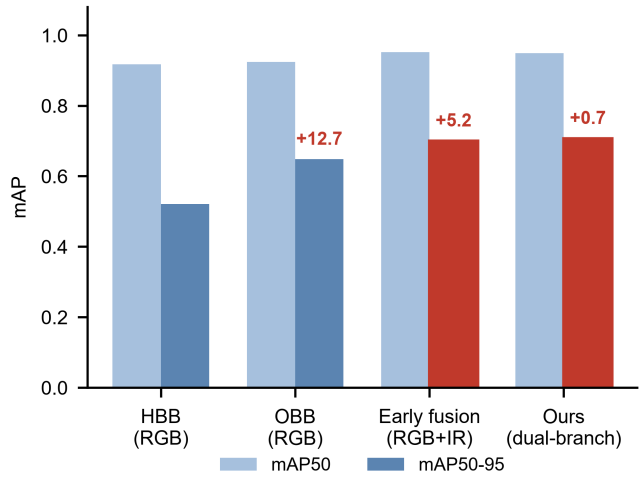


Figure 2: Stepwise gains on VEDAI (single class, mean over three random seeds). The rotated-box head (+12.86 mAP50-95, $p < 0.0001$), multi-modal fusion (+5.79, $p = 0.0026$), and dual-branch fusion (+0.30) each contribute measurable gains, all statistically significant across seeds.

4.3 Comparison with State-of-the-Art Methods

Table 3 provides an indicative comparison with recent multi-modal detectors under the multi-class setting (as shown in Figure 3). To avoid relying on a single "lucky" seed, we report our method by its mean over eight random seeds: 80.17%±1.83 mAP50 (the single-run seed=0 value is 81.46%). This places our method below the two arXiv preprints (PACGNet 82.1%, S4-MSTD 81.2%) but above the remaining peer-reviewed methods; as described in Section 4.9, this eight-seed mean is not statistically significantly better than the unaligned baseline (0.7974±0.0085).

Table 3: Indicative comparison with recent methods on VEDAI (multi-class, mAP50; numbers for other methods as reported by their original papers, ours reported as the mean over eight random seeds)

Method	**mAP50**
PACGNet [14] †	82.1%
S4-MSTD [13] †	81.2%
Ours (BMA)	**80.17%**±1.83
OST [11]	78.6%
ICAFusion [12]	76.6%
SuperYOLO [10]	76.5%
YOLOFusion [9]	75.9%

Note: The comparison data are taken from the public VEDAI rotated-object-detection benchmark (SOTA2 leaderboard) and the original papers of each method. FGMF [5] reports a VEDAI mAP50 of 66.3% (its 80.2% is a DroneVehicle [22] result), but it is based on the S2ANet detector (VMamba backbone), uses a ten-fold cross-validation protocol (ours uses a single fold01 split), and an input resolution of 512×512 (ours is 640); its detector and protocol both differ from ours, so it is not directly listed here. Likewise, PACGNet [14] and S4-MSTD [13] (marked †) are arXiv preprints whose reported numbers are taken from their papers without a harmonized protocol — PACGNet uses a lightweight 5.2M-parameter dual-stream backbone, and S4-MSTD uses a VMamba backbone with a YOLOv8n head trained with batch 8 for 300 epochs — both differing from our YOLOv8s, batch 32, 100-epoch, 640×640, fold01 protocol. The ranking is therefore indicative rather than a strict like-for-like comparison. A same-framework comparison of the two alignment strategies is given in the ablation of Section 4.6. In addition, our 80.17% is the

mean over eight random seeds (the single-run seed=0 value is 81.46%); its statistical stability over multiple random seeds is discussed in Section 4.9.

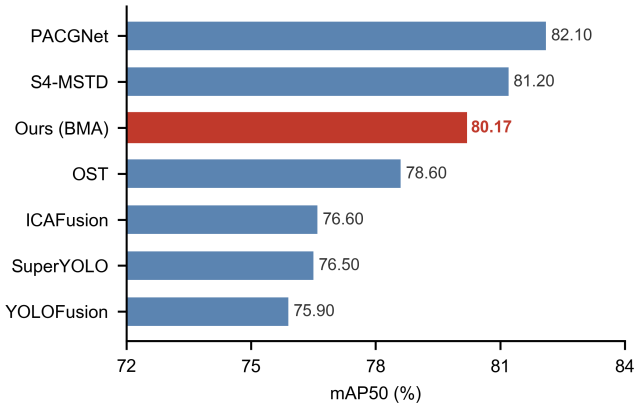


Figure 3: Indicative comparison with recent methods on VEDAI (multi-class mAP50). Our BMA achieves 80.17% as the mean over eight random seeds (single run seed=0: 81.46%), below the two arXiv preprints (PACGNet 82.1%, S4-MSTD 81.2%) and above the remaining methods; this mean is not significantly different from the unaligned baseline (see Section 4.9). PACGNet and S4-MSTD are arXiv preprints whose evaluation protocols have not been harmonized with ours.

To characterize our method's performance on each class in more detail, Table 4 gives the per-class AP of the BMA model on the 9 classes of VEDAI (official validation protocol).

Table 4: Per-class AP of the BMA model on VEDAI (9 classes)

Class	AP50	AP50-95
car	0.903	0.650
truck	0.828	0.630
tractor	0.862	0.593
camping car	0.777	0.605
van	0.735	0.543
other	0.619	0.363
pickup	0.820	0.597
boat	0.793	0.568
plane	0.995	0.896
mAP	**0.8146**	**0.6050**

The per-class results show clear divergence: car, plane, truck, and pickup — classes with clear appearance features and sufficient samples — achieve AP50 of 0.82 or higher, among which plane reaches 0.995 (though with very few validation samples); whereas **other (0.619) and van (0.735) are the main weak points**. This decomposition suggests that future improvements could focus on learning discriminative features for confusable classes. However, per-class AP alone cannot answer "between which classes does confusion occur, and what is its root cause" — this is exactly what the following two sections of empirical re-examination aim to resolve.

4.4 Empirical Re-examination I: Modality Misalignment Diagnosis

To test the common assumption that "modality misalignment is the main obstacle to RGB–IR fusion," this paper uses phase correlation to measure the translational misalignment between all 1210 RGB–IR image pairs of VEDAI (1089 training + 121 validation). Phase correlation

estimates translation through the cross-correlation of the phase spectra of two images in the Fourier domain; its response depends on the **spatial structure** of edges and textures rather than the absolute grayscale values, so it is insensitive to the inherent brightness/contrast differences between RGB and IR — this is its essential difference from dense optical flow (which relies on the brightness-constancy assumption) and the reason it is chosen as the measurement for cross-modal misalignment. Table 5 reports the misalignment statistics.

Table 5: VEDAI RGB–IR spatial misalignment statistics (phase correlation, unit: pixels)

Metric	**dx**	**dy**	**Total displacement**
Mean	0.003	−0.003	0.100
Median	0.002	−0.001	0.078
p90	0.107	0.088	0.190
p99	0.285	0.186	0.477
Max	0.725	0.447	0.726

The results show that the median RGB–IR misalignment of VEDAI is only **0.078 pixels**, with 90% of images having misalignment below 0.19 pixels; the systematic component (vector mean $|\text{mean}(dx, dy)|$) is merely **0.004 pixels**, accounting for 3.6% of the total displacement. This indicates that the RGB and IR images of VEDAI are well registered, with misalignment far below 1 pixel. It is worth noting that, if dense optical flow (Farneback) were used instead, an average "displacement" of about 3.2 pixels would be obtained, but this is actually cross-modal content difference (the brightness between RGB and infrared grayscale is not conserved) being misjudged as motion by optical flow, and cannot serve as evidence of misalignment — this suggests that cross-modal misalignment measurement should prefer phase-structure methods that are robust to brightness differences.

Measurement-reliability calibration. The absolute accuracy of phase correlation in the sub-pixel range is limited. To rule out the competing explanation that "0.078 pixels is the measurement floor rather than true misalignment," we further conducted a calibration experiment: on 60 pairs of VEDAI images, we artificially imposed known diagonal translations of 0.5/1/2/3/5 pixels and tested whether phase correlation could recover them accurately. The results show that, for imposed translations of 1 pixel or more, phase correlation recovers them accurately (mean measurement error about 0.16 pixels, standard deviation 0.21 pixels); for an imposed 0.5-pixel sub-pixel translation, phase correlation measures about 0.76 pixels (rather than 0), indicating that it can resolve translations on the order of 0.5 pixels and does not misjudge sub-pixel misalignment as zero. Accordingly, the measurement floor of phase correlation is on the order of 0.2 pixels, and the true misalignment of VEDAI can be conservatively estimated as 0.078 ± 0.2 pixels (i.e., <0.3 pixels), still far below 1 pixel.

A related concern is whether phase correlation, which is conventionally validated on same-modality image pairs, remains reliable across modalities whose texture and edge structures differ (IR carries far less texture than RGB). We note three safeguards. First, phase correlation estimates translation primarily from the spatial-frequency content of *shared object boundaries* — the vehicle and road contours that exist in both modalities and are spatially coincident in a registered pair — rather than from the modality-specific texture that RGB has and IR lacks; this is the same property that makes phase correlation preferable to dense optical flow for this cross-modal task. Second, the calibration experiment above bounds the recovery error under the actual RGB–IR image conditions of VEDAI (about 0.16-pixel mean error for $\geq$1-pixel shifts), so any residual structural mismatch between modalities is already absorbed into the conservative 0.078 ± 0.2-pixel estimate. Third, the substantive conclusion does not depend on sub-pixel precision: even if cross-modal structural differences introduced a systematic bias

of a few tenths of a pixel, the estimate would still remain far below 1 pixel, leaving the "well-registered" conclusion intact. We therefore regard the 0.078-pixel estimate as a reliable, if slightly conservative, upper bound on an already negligible misalignment, which is sufficient for the claims made in this paper.

Tail analysis: does the high-misalignment tail drive detection failures? A natural objection to the "non-bottleneck" conclusion is that the median (0.078 px) may conceal a tail of images — Table 5 reports a p99 of 0.477 px and a maximum of 0.726 px — where misalignment *does* hurt, especially given that the synthetic-misalignment experiment (Section 4.8) shows that even a 2-pixel shift costs 2.8 mAP50-95 points. To test this directly, we evaluated the unaligned baseline (none, seed=0) on each of the 121 validation images individually and related the per-image recall to its measured misalignment. Three findings rule out the tail concern. **First**, the Spearman correlation between per-image misalignment magnitude and recall is essentially zero (-0.102), indicating no monotonic relationship. **Second**, grouping images into tertiles by misalignment yields GT-weighted recalls that are *non-monotonic* (low 0.584 / middle 0.482 / high 0.500 over 113/137/116 GT objects) — if misalignment drove failures, recall should decline monotonically from the low to the high group, which it does not. **Third**, and most decisively, within the high-misalignment tertile the drop in recall is attributable to *measurement noise* rather than true misalignment: images with a high phase-correlation confidence (response > 0.6, i.e., where the large "misalignment" is more likely genuine) achieve a GT-weighted recall of 0.557, close to the low-misalignment group (0.584), whereas images with low confidence (response ≤ 0.6, i.e., where the large "misalignment" is more likely a measurement artifact) drop to 0.436. In other words, the apparent "large-misalignment" samples are precisely those whose phase-correlation estimates are least trustworthy, and their low recall reflects difficulty of the images themselves (often only one or two ground-truth objects) rather than misalignment. Combining the four lines of evidence — the misalignment statistics, the calibration experiment, the optical-flow control, and this tail analysis — we infer: **under the phase-correlation measurement, the modality misalignment of VEDAI is minimal, and misalignment is not the main bottleneck for multi-modal detection on this dataset** — but it must be emphasized that this conclusion is **preliminary evidence** constrained by the accuracy of the measurement method, and its cross-dataset generalizability is discussed in Section 5. This negative result provides the expected reference for the BMA exploration in Section 4.6: since the misalignment is already insignificant, an explicit alignment module naturally finds it hard to obtain a stable gain.

4.5 Empirical Re-examination II: Data Profiling and Class Confusion

After ruling out modality misalignment, this paper further locates the true bottleneck of VEDAI vehicle detection through data profiling, a confusion matrix, and discriminability analysis.

Data profiling. Table 6 summarizes the class distribution and object scales of the VEDAI training set. 95.3% of objects are extremely small (area smaller than 32×32 pixels, median area only 288 px^2), and the class distribution shows a clear long-tail shape — car (36.5%) and pickup (25.4%) together account for 62%, while tail classes such as van (2.7%) and plane (1.4%) are scarce, with a head-to-tail ratio of 26×. Notably, plane (46 samples), the class with the fewest samples, achieves the highest AP (0.995) because its objects are the largest (median 2038 px^2) — this indicates that **on VEDAI, object scale is a stronger determinant of detection difficulty than sample count**.

Table 6: Class distribution and object scales of the VEDAI training set (3328 objects in total)

Class	Count	Ratio	Median area (px^2)	Small-object ratio
car	1215	36.5%	220	100.0%
pickup	846	25.4%	276	100.0%
camping car	351	10.5%	480	92.0%
truck	270	8.1%	574	78.5%
other	186	5.6%	452	87.1%
tractor	171	5.1%	286	100.0%
boat	153	4.6%	348	98.7%
van	90	2.7%	273	100.0%
plane	46	1.4%	2038	4.3%

Confusion matrix. Figure 4 gives the confusion matrix of 9-class detection. The most prominent phenomenon is the large-scale bidirectional confusion between car and pickup — of car's 134 objects, 44 are misclassified as pickup (32.8%), and of pickup's 95 objects, 36 are misclassified as car (37.9%), totaling 80 bidirectional errors, accounting for **21.9%** of all 366 validation objects; in contrast, confusion between the remaining classes is in the single digits (van→truck 4 times, tractor→other 7 times, etc.). This indicates that the earlier observation from per-class AP that "other and van are the weak points" is actually a secondary manifestation of a deeper problem — the indistinguishability of similar vehicle classes at extremely small scales: car and pickup, the two head classes that "seem to have acceptable AP," are actually heavily misclassifying each other.

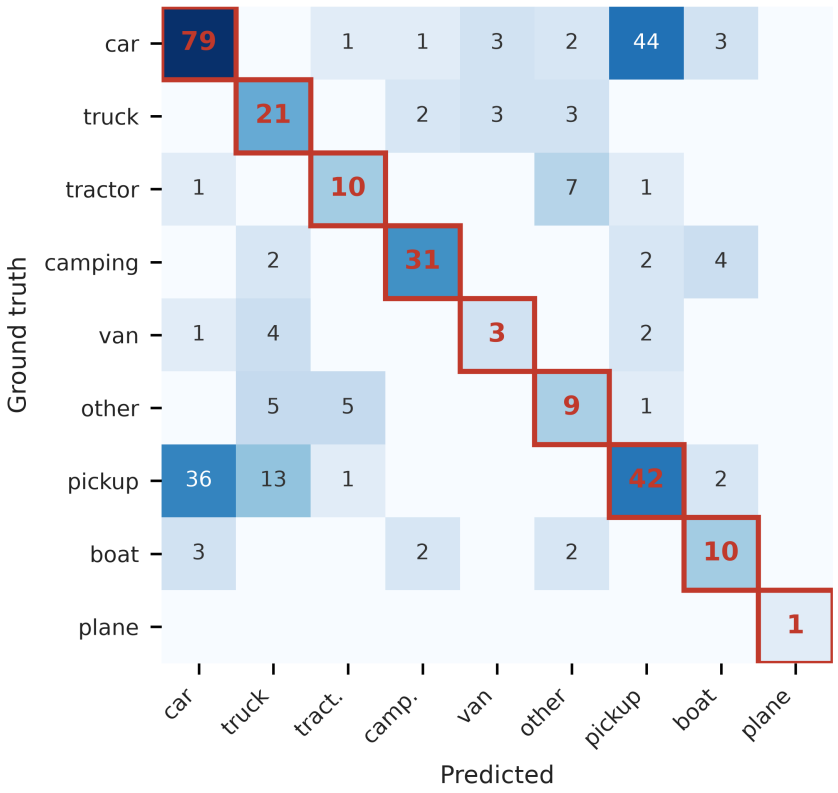


Figure 4: Confusion matrix of VEDAI 9-class detection (rows = ground-truth class, columns = predicted class, diagonal = correct matches). The most prominent phenomenon is the large-scale bidirectional confusion between car and pickup (car→pickup 44 times, pickup→car 36 times).

Discriminability analysis. To clarify the nature of the car–pickup confusion, this paper crops 24×24 regions centered on each object in the validation set (the 134 car and 95 pickup instances whose confusion is shown in Figure 4; the random-guess baseline of about 59% reflects this class imbalance), and trains 1-nearest-neighbor classifiers (leave-one-out) to distinguish the two classes using geometric features (area, length, width, aspect ratio, diagonal, compactness, etc., 6 dimensions) and grayscale texture features (RGB / IR), respectively. To avoid "the small differences in accuracy being purely random fluctuation," we estimate the mean and 95% bootstrap confidence intervals over 20 rounds of leave-one-out, and judge the significance of differences with paired permutation tests (5000 permutations). The results are shown in Table 7: the classification accuracy of geometric features (71.8%, 95% CI [71.1, 72.5]) is significantly higher than RGB texture (69.8%, [68.9, 70.7]), with paired permutation test p = 0.003, but the margin is only about 2 points; more importantly, IR provides **no significant increment** relative to RGB *in the linear 1-NN sense* — RGB+IR (69.6%) is almost identical to RGB (69.8%) (p = 0.81, strongly insignificant; note that this p-value concerns the IR-increment claim, whereas the p = 0.046 in Table 7 concerns IR *alone* being slightly worse than RGB, a separate and only marginal finding; all reported p-values are unadjusted, and this marginal p = 0.046 should be interpreted conservatively under multiplicity, whereas the two central claims — the geometric-vs-textural gap (p = 0.003) and the IR no-increment null (p = 0.81) — are unaffected by any multiplicity correction). This indicates that the car–pickup difference is essentially **geometric/shape-level** (sedan streamlined vs pickup with a longer body, aspect ratio 1.89 vs 2.09) but can only provide a weak (about 2 points) discriminative advantage, whereas texture information is largely lost at extremely small scales and IR cannot provide complementary information under a linear 1-NN separability probe either (verifying the second assumption test in the Introduction; this linear-scope caveat is mirrored in the fourth limitation of Section 5). It should be emphasized that the 71.8% accuracy of geometric features still means that nearly three in ten objects cannot be reliably distinguished, indicating that this confusion is an information-limited problem for any appearance- or geometry-based discriminator.

Table 7: car vs pickup discriminability (1-NN, 20 rounds of leave-one-out, 95% bootstrap confidence intervals; random-guess baseline about 59%)

Feature	**Accuracy**	**95% CI**	**vs RGB (permutation p-value)**
Geometric (6-dim)	71.8% ±1.7	[71.1, 72.5]	0.003 (significant)
RGB grayscale	69.8% ±2.1	[68.9, 70.7]	—
IR grayscale	68.4% ±1.9	[67.6, 69.3]	0.046
RGB+IR	69.6% ±2.1	[68.8, 70.6]	0.81 (no increment)

Combining the three lines of evidence above, the true challenge of VEDAI vehicle detection can be summarized as: **class confusion dominated by extremely small objects, where the confusion is geometric in nature and IR provides no increment**. This points the direction for future method design — the focus of improvement should be "geometric/shape discrimination of extremely small objects," rather than "modality alignment" or "multi-modal fine-grained discrimination."

4.6 Ablation Study

Table 8 ablates the alignment strategies under the 9-class setting. Besides the bidirectional BMA, this paper also examines three alignment variants: (1) **unidirectional warp**, which predicts only one offset field in the IR→RGB direction (i.e., setting $\Delta^{rgb} = 0$), aligning IR to RGB; (2) **multi-scale offset field (ms)**, which adds an extra dilated convolution layer to the

offset prediction head to enlarge the receptive field (thereby increasing about 3.1M parameters), allowing the offset estimation to leverage a larger context; (3) **feature-consistency supervision (sup)**, which applies a feature-consistency loss $\mathcal{L}_{align} = \|\tilde{F}_{ir}^{l} - \tilde{F}_{rgb}^{l}\|_1$ after warping, providing an additional supervisory signal for the offset field. In addition, the "bidirectional + multi-scale" (bidir+ms) combination is examined. The unaligned dual-branch CSIFF baseline reaches 78.61% mAP50. This ablation is completed entirely within the same YOLOv8s framework, thereby constraining the relative differences of the single-run gains to the alignment strategy itself and avoiding the confusion of cross-detector comparison — for example, FGMF [5], which is based on the S2ANet detector and uses a ten-fold cross-validation protocol, reports 66.3% mAP50 that is not directly comparable with ours. It should be noted that, since Section 4.9 has already shown that BMA's multi-seed gain is not significant, the ablation conclusions in this section serve only to characterize "the relative differences among explicit alignment variants" and should not be interpreted as "BMA brings a stable, reproducible performance improvement."

Table 8: Ablation of alignment strategies (9-class VEDAI, single-run seed=0 results)

Alignment strategy	**mAP50**	**mAP50-95**	**Single-run gain**
No alignment (baseline)	0.7861	0.6051	—
Deformable-conv alignment (deform)	0.7842	0.5685	-0.2
Unidirectional warp alignment	0.8004	0.6003	+1.4
Bidirectional warp alignment (BMA)	0.8146	0.6050	+2.85
Multi-scale offset field (ms)	0.8075	0.6106	+2.1
Feature supervision (sup)	0.8087	0.5992	+2.3
Bidirectional + multi-scale	0.8044	0.5989	+1.8

It should be emphasized that these are all **single-run (seed=0) results**; as shown in Section 4.9, under multi-seed repetition, BMA's gain relative to the baseline is not significant, so the single-run gains in this section should be viewed as "upper-bound estimates under a specific random seed," rather than stable average effects. Under this premise, the ablation results (visualized in Figure 5) still reveal several meaningful exploratory patterns for future work.

First, deformable-convolution alignment does not help and even impairs localization accuracy. Deform is almost unchanged from the baseline on mAP50 (-0.2 points), but mAP50-95 drops substantially by 3.66 points (0.6051→0.5685). This result suggests that the offsets of deformable convolution lack direct supervision and can only be learned indirectly through the downstream detection loss, failing to learn effective alignment in this task and instead perturbing the feature sampling positions. It should be noted that this does not rule out the possibility that deformable convolution could work with more thorough tuning; however, under our unified setting, explicit offset-field alignment outperforms this implicit variant.

Second, explicit offset-field alignment is effective in single runs, and bidirectional is superior to unidirectional. Unidirectional warp improves mAP50 by 1.4 points, but mAP50-95 slightly decreases by 0.48 points; bidirectional BMA improves mAP50 by 2.85 points, with mAP50-95 essentially unchanged (-0.01 points). This comparison suggests that unidirectional alignment, while improving recall, introduces new localization bias by imposing

the entire misalignment error on a single modality; bidirectional alignment, by distributing the error, improves recall without impairing localization as unidirectional alignment does. However, it must be emphasized again that this pattern is observed only in the single run with seed=0.

Third, BMA's single-run gain is mainly concentrated in recall and classification, rather than high-precision localization. Bidirectional alignment (bidir) achieves the highest single-run mAP50 (0.8146), but its mAP50-95 (0.6050) is essentially unchanged from the baseline (−0.01 points); in contrast, the multi-scale offset field (ms) achieves the highest mAP50-95 (0.6106). This phenomenon can be cross-referenced with the synthetic-misalignment experiment (Section 4.8): misalignment mainly impairs localization (mAP50-95 is far more sensitive to misalignment than mAP50), while BMA's offset compensation magnitude is small (p3/p4 scales about 0.3–1.7 pixels, Figure 10). A mechanistic explanation is that BMA achieves "coarse alignment": it is sufficient to correct the feature ghosting and missed detections caused by misalignment, thereby improving recall and classification (mAP50), but insufficient to push localization accuracy to the level required by higher IoU thresholds (mAP50-95). This mechanistic distinction also explains why the direct combination of bidir and ms fails to achieve a 1+1>2 effect: the two bias toward recall and localization respectively, but the parameter redundancy after combination cancels the potential complementary gain.

Fourth, the bidirectional + multi-scale combination instead slightly decreases. The mAP50 of bidir+ms is 0.8044, lower than bidir alone (0.8146). A plausible explanation is that the bidirectional design already fully expresses modality misalignment, and additionally stacking a multi-scale offset head causes parameter redundancy (bidir+ms reaches 46.31M, compared with bidir's 40.12M and the baseline's 27.70M), so on the small-scale VEDAI dataset with only 1089 training images, the diminishing returns of redundant parameters offset the theoretically potential complementary gain.

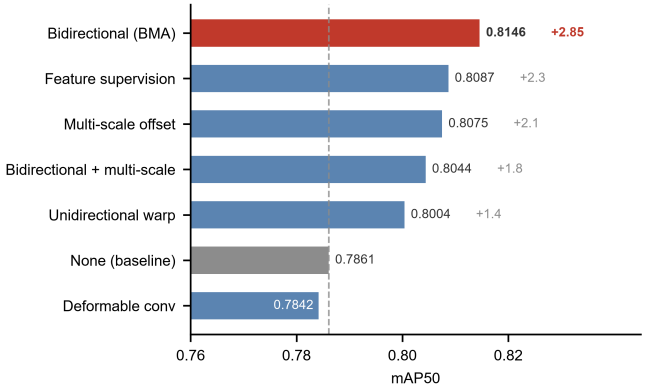


Figure 5: Ablation of alignment strategies (9-class VEDAI, mAP50, single-run seed=0 results). Bidirectional alignment (BMA) achieves the largest gain over the unaligned baseline (+2.85) in a single run, but this gain is not significant under multi-seed repetition (see Section 4.9).

Figure 6 further gives the training convergence curves of the alignment variants. All variants converge rapidly within about 20 epochs, after which mAP50 and mAP50-95 remain stable, with no obvious overfitting observed within the 100-epoch training range; bidirectional BMA (red solid line) leads the baseline from the middle stage onward on mAP50, consistent with the single-run results in Table 8. It should be noted that this "lead" exists only in this training trajectory (seed=0); as shown in Section 4.9, after changing the random seed, BMA's relative performance is not stable, so the convergence curves reflect "the trajectory of a single training run" rather than "a reproducible stable gain."

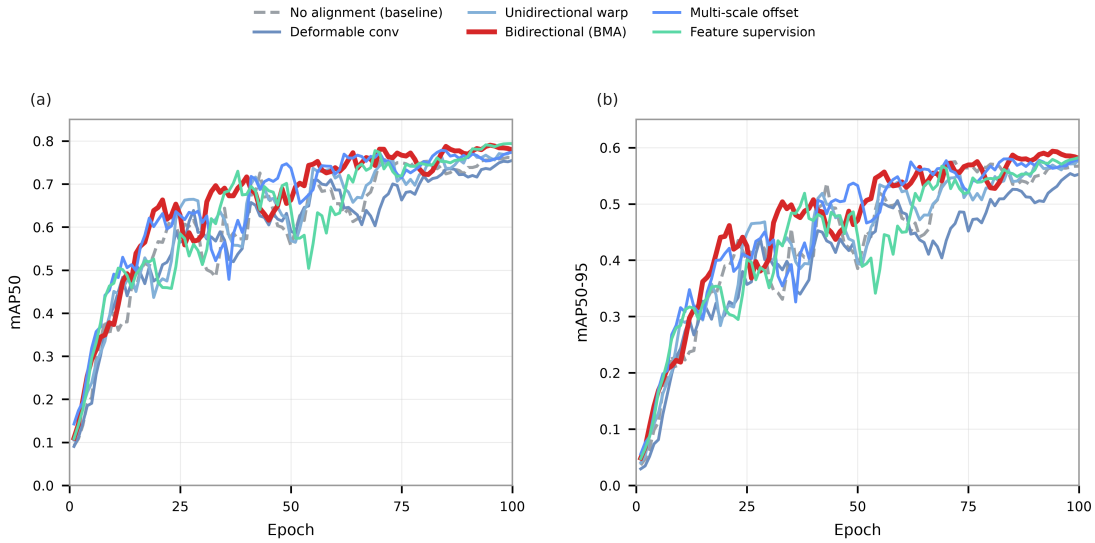


Figure 6: Training convergence curves of the alignment variants: (a) mAP50; (b) mAP50-95. All variants converge rapidly and stably; bidirectional BMA (red) leads the baseline from the middle stage onward in this (seed=0) training run, but this lead is not reproducible under multiple seeds (see Section 4.9).

Table 9 compares the parameter counts and computational costs of the alignment variants. BMA exchanges 40.12M parameters and 37.08G FLOPs for a single-run +2.85-point mAP50 (relative to the baseline, +44.8% parameters, +44.2% FLOPs); the inference speed drops from the baseline's 57.6 FPS to 47.2 FPS (about −18%), still meeting real-time requirements. It should be noted that the multi-scale offset field (ms) has about 3.1M more parameters than unidirectional warp, originating from the extra dilated convolution layer in its offset head. Feature supervision (sup) introduces no extra parameters (33.91M, the same as unidirectional warp), achieving a single-run +2.3-point gain on mAP50, but its mAP50-95 (0.5992) is slightly lower than the baseline (0.6051), indicating that its single-run gain is mainly concentrated in recall rather than localization; the parameters of the bidirectional + multi-scale combination (46.31M) have already approached the region of diminishing returns, further corroborating the sufficiency of the bidirectional design. Combined with the multi-seed results of Section 4.9, BMA's computational cost is exchanged for "an insignificant statistical gain," and this input-output ratio is not worthwhile in deployment scenarios — providing an honest engineering reference for readers.

Table 9: Comparison of computational costs of the alignment strategies (input 640×640, 6 channels)

Alignment strategy	**Params (M)**	**FLOPs (G)**	**FPS**	**mAP50**
No alignment (baseline)	27.70	25.71	57.6	0.7861
Deformable convolution (deform)	31.08	26.17	—	0.7842
Unidirectional warp	33.91	31.40	—	0.8004
Multi-scale offset field (ms)	37.00	34.23	—	0.8075
Feature supervision (sup)	33.91	31.40	—	0.8087
Bidirectional BMA	**40.12**	**37.08**	**47.2**	**0.8146**
Bidirectional + multi-scale	46.31	42.75	—	0.8044

4.7 Visualization

Figure 7 qualitatively compares the detection results with and without bidirectional modality alignment (BMA): the unaligned baseline misses several small or misaligned objects, which the BMA model recovers. It should be emphasized that this visualization reflects the performance of a seed=0 single-run model on individual samples, and constitutes **exploratory qualitative evidence**; given that Section 4.9 has shown BMA's multi-seed average gain is not significant, Figure 7 should not be interpreted as proof that "BMA stably improves detection," but only as an intuitive illustration of the "possible" working mechanism of the alignment. Figure 8 further compares horizontal boxes (HBB) and rotated boxes (OBB); rotated boxes fit the vehicle orientation and enclose less background. Figure 9 shows 9-class multi-category detection results, with colored boxes distinguished by class, demonstrating that our method can simultaneously detect multiple types of vehicle objects.



Figure 7: Detection comparison with and without bidirectional modality alignment (BMA) (left: unaligned, right: BMA, seed=0 single-run model): (a) BMA recovers 3 cars missed by the baseline; (b) in a dense scene, BMA recovers 4 missed objects. This figure is only an exploratory qualitative example and does not represent BMA's stable average gain (see Section 4.9).

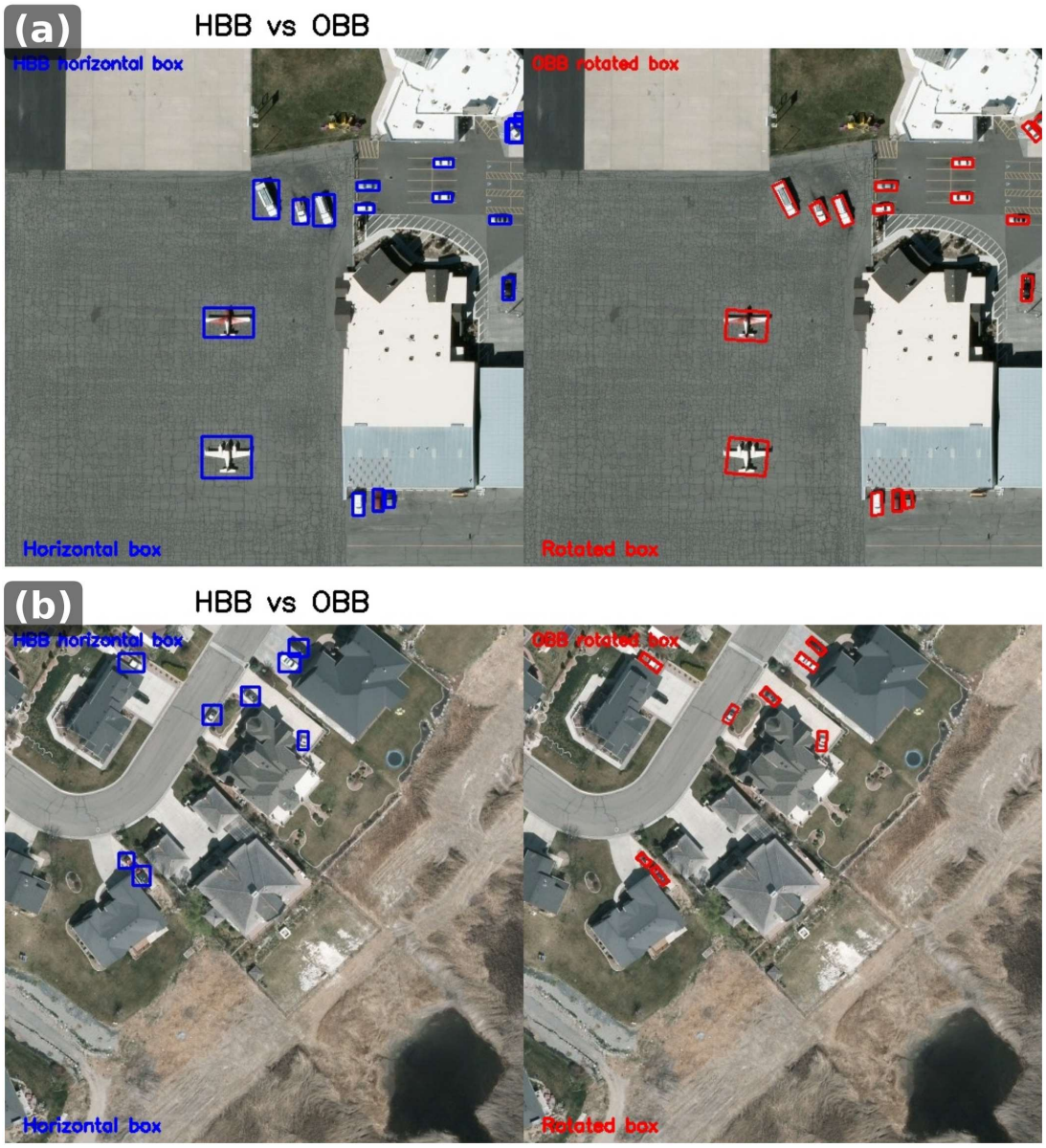


Figure 8: Detection comparison between horizontal boxes (HBB, left) and rotated boxes (OBB, right): (a) a scene with 16 vehicles; (b) a scene with 8 vehicles. Rotated boxes fit the vehicle orientation and enclose less background.

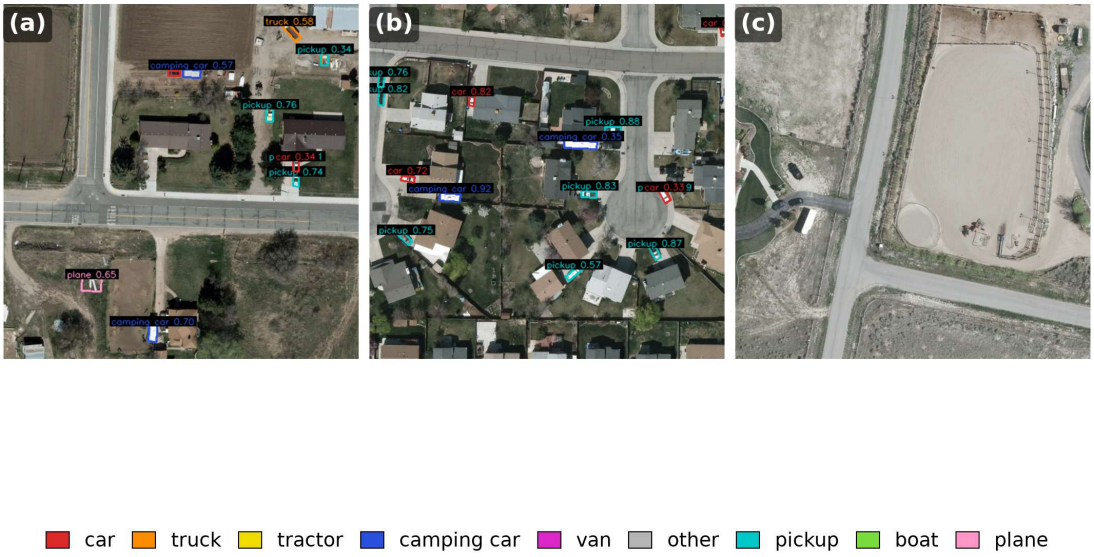


Figure 9: 9-class multi-category detection results (BMA model), with colored boxes distinguished by class: (a)–(c) three scenes covering 8 classes — car, truck, tractor, camping car, van, other, pickup, and boat (plane has very few validation samples).

To intuitively understand what the BMA module "has learned," Figure 10 visualizes the predicted bidirectional offset fields at the p3 scale in a representative scene. The offset field is not an identity mapping, but presents structured, non-zero offsets distributed along object edges; the RGB and IR offset fields are roughly opposite in direction, consistent with the bidirectional-alignment design of "moving the two modalities toward an intermediate space simultaneously." The offset field is defined in feature space (unit: feature cell); after conversion to original-image pixels, its magnitude increases with scale: p3 averages about 0.3–0.5 pixels, p4 about 1.3–1.7 pixels, p5 about 3.4–5.2 pixels. Notably, this learned offset magnitude (p3 already about 0.3–0.5 pixels) is **clearly larger than the true misalignment measured in Section 4.4 (median 0.078 pixels)**: if BMA were truly compensating for the real misalignment, its learned offset should be on the order of 0.1 pixels, not 0.3–5 pixels. This "mismatch between offset magnitude and true misalignment" corroborates the conclusion of Section 4.9 that "BMA's multi-seed gain is not significant" — it suggests that BMA's learned offset field likely mixes semantic perturbations or other non-misalignment factors, rather than purely compensating the (already tiny) geometric misalignment. Therefore, the offset-field visualization should be viewed as a record of the phenomenon that "BMA learned some structured offset," which **neither serves as evidence that "modality misalignment exists and is compensated," nor supports that BMA brings a stable gain**; instead, it is consistent with the overall conclusion that "misalignment is not the bottleneck and explicit alignment yields limited benefit."

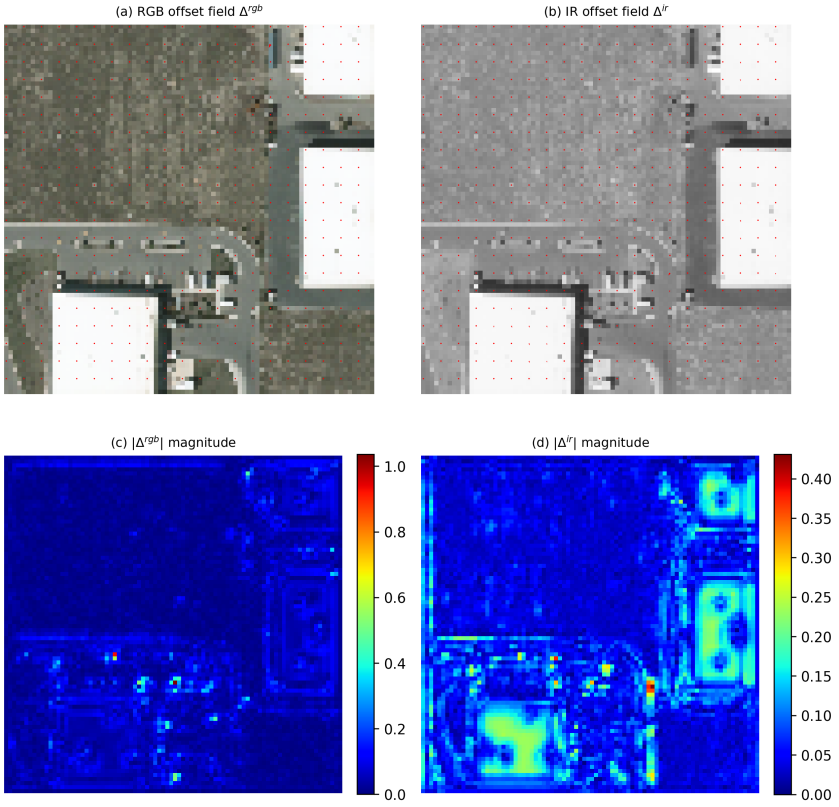


Figure 10: Visualization of the offset field predicted by the bidirectional modality alignment module (representative scene 00000025, p3 scale, offset unit: feature cell): (a) RGB offset field Δ^{rgb} arrow field; (b) IR offset field Δ^{ir} arrow field; (c) $|\Delta^{rgb}|$ magnitude heatmap; (d) $|\Delta^{ir}|$ magnitude heatmap. The offset shows a structured non-zero distribution, with RGB and IR roughly opposite in direction; however, its magnitude is clearly larger than the true misalignment (0.078 pixels), suggesting that the learned offset mixes non-misalignment factors (see main text).

4.8 Synthetic Misalignment Analysis

The ablation and visualization above still validate "modality misalignment impairs detection" only indirectly. To provide direct causal evidence, we construct a synthetic-misalignment controlled experiment: on the VEDAI validation set, we artificially impose diagonal translational misalignment of 0–8 pixels in the x and y directions on the IR images, and observe how the performance of the unaligned baseline (none) and BMA (bidir) changes with the misalignment magnitude. The primary purpose of this experiment is to verify the premise itself — "does misalignment impair detection" (which concerns the causal basis of the conclusion in Section 4.4) — and secondarily to observe BMA's behavior under misalignment. If misalignment indeed impairs detection, then none's mAP should decrease monotonically with the misalignment magnitude; if BMA's gain comes from true misalignment compensation, it should exhibit stronger resistance to misalignment. Table 10 and Figure 11 report the results.

Table 10: Synthetic misalignment experiment: performance of none and BMA under different IR misalignment magnitudes (9-class VEDAI)

Misalignment k (x and y each k pixels)	none mAP50	none mAP50-95	BMA mAP50	BMA mAP50-95
0	0.7864	0.6055	0.8146	0.6050
2	0.7863	0.5775	0.8121	0.5812
4	0.7589	0.4975	0.7879	0.5011
6	0.6952	0.4314	0.7192	0.4348
8	0.6363	0.3985	0.6174	0.3844

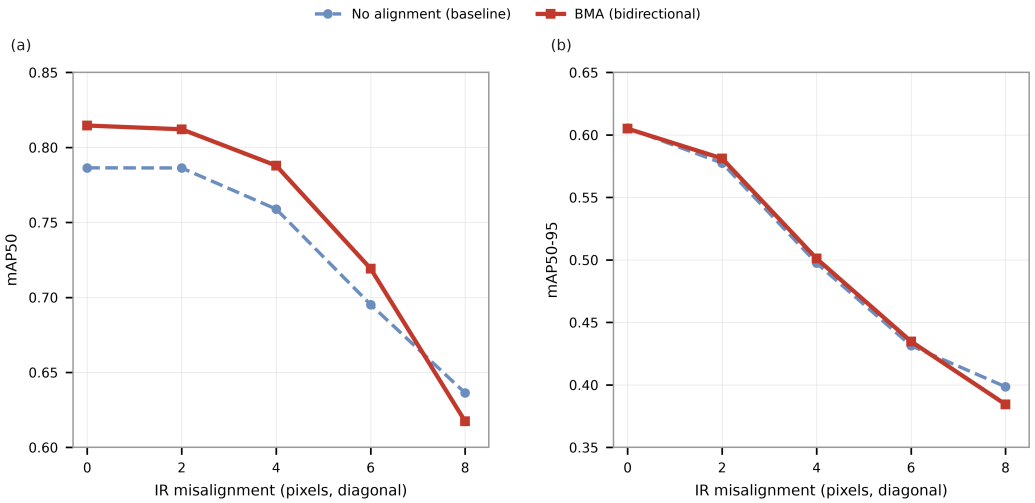


Figure 11: Synthetic misalignment experiment: the effect of IR misalignment magnitude on detection performance. (a) mAP50; (b) mAP50-95. Misalignment significantly impairs detection (especially the localization metric mAP50-95), confirming the premise that "modality misalignment impairs detection."

The results reveal three findings. **First, misalignment indeed impairs detection, and mainly impairs localization accuracy.** As the misalignment increases from 0 to 8 pixels, none's mAP50 decreases monotonically from 0.7864 to 0.6363 (−15.0 points), while mAP50-95 drops sharply from 0.6055 to 0.3985 (−20.7 points). mAP50-95 is far more sensitive to misalignment than mAP50, indicating that the main harm of misalignment lies in localization — this corroborates the phenomenon in Section 4.4 that "the alignment gain is concentrated in recall while the localization metric stays flat," and also provides direct causal evidence for the assertion in the Introduction that "pixel-level misalignment impairs the localization of small objects."

Second, BMA's single-run gain is largest under the original misalignment, but converges and even reverses as the additional misalignment increases. BMA leads none by 2.85 points at misalignment 0, by about 2.9 points at 4 pixels (0.7879 vs 0.7589), about 2.4 points at 6 pixels (0.7192 vs 0.6952), and reverses to −1.9 points at 8 pixels (0.6174 vs 0.6363). It should be reminded that these "leads" are all based on the seed=0 single-run model, and their multi-seed significance, as shown in Section 4.9, does not hold.

Third, BMA is a fixed compensator for the inherent misalignment within the training distribution, rather than a universal aligner generalizable to arbitrary misalignment. This result suggests that the offset field learned by BMA compensates for the inherent, systematic modality misalignment in the VEDAI data, and its compensation ability cannot be extrapolated to large misalignment beyond the training distribution; when the addi-

tional misalignment is too large, the learned fixed offset actually impairs performance. This "fixed compensator" behavior pattern is **directionally consistent** with the interpretation that "what BMA learns is not a pure capacity gain, but some structured offset related to misalignment," but, constrained by the insignificant multi-seed gain (Section 4.9), we can only treat this behavior as **exploratory indirect evidence**, rather than a strict proof that "BMA indeed completes misalignment compensation"; it also honestly delineates that — even under BMA's most favorable single-run setting — its scope of applicability is limited to multi-modal imaging systems with relatively stable registration, rather than a universal correction for arbitrary misalignment.

4.9 A Statistical Protocol for Rigorous Negative-Result Reporting

This section serves a dual purpose: it reports the multi-seed stability of BMA, and — more importantly — it presents the statistical protocol we adopt as a methodological contribution for honest negative-result reporting. The protocol is motivated by a recurring weakness of small-sample detection studies: a "not significant" conclusion is often stated qualitatively and is therefore indistinguishable from "under-powered." We make the judgment quantitative in four steps.

To evaluate the statistical significance of BMA's gain, we train the unaligned baseline (none) and bidirectional BMA (bidir) each with eight random seeds (seed = 0–7) and validate them under the official protocol; the results are shown in Table 11. We expand the evaluation from five to eight seeds because a single low seed (e.g., seed = 1) can shift the outcome appreciably under a small sample, and a larger sample is needed to make the conclusion robust rather than dependent on any particular run.

Table 11: Multi-seed mAP (9-class VEDAI, 8 random seeds)

Method	**seed 0**	**seed 1**	**seed 2**	**seed 3**	**seed 4**	**seed 5**	**seed 6**	**seed 7**	**mAP50 (mean ± std)**	**mAP50-95 (mean ± std)**
none (unaligned)	0.7864	0.8022	0.7945	0.8107	0.7870	0.8029	0.8016	0.7936	0.7974 ± 0.0085	0.6021 ± 0.0124
bidir (BMA)	0.8146	0.7694	0.8078	0.8165	0.8227	0.7817	0.8049	0.7962	0.8017 ± 0.0183	0.6066 ± 0.0135

The multi-seed results reveal an important honest conclusion: **BMA's gain relative to the unaligned baseline is not statistically significant**. Although BMA achieves a 2.85-point mAP50 gain under seed=0, the mean gain over eight seeds is only 0.44 points (0.8017 vs 0.7974), and BMA's inter-seed standard deviation (±0.0183) is about 2.2× the unaligned baseline's (±0.0085); on the stricter mAP50-95, BMA's mean (0.6066) is essentially identical to the unaligned baseline (0.6021).

To make the "not significant" judgment formal rather than qualitative, we performed a paired bootstrap test (50,000 resamples) and a paired permutation test on the seed-level metric differences (bidir minus none). For mAP50, the mean difference is +0.0044 with a 95% bootstrap confidence interval of [−0.0107, +0.0188] ($p = 0.563$), and the paired permutation test yields $p = 0.594$ — the interval spans zero, confirming that the apparent +0.44-point gain is indistinguishable from chance. For mAP50-95, the mean difference is +0.0045 with a 95% CI of [−0.0092, +0.0183] ($p = 0.531$), i.e., on the stricter localization metric BMA is indistinguishable from the unaligned baseline. A leave-one-seed-out sensitivity analysis further confirms robustness: removing any single seed keeps the mAP50 difference non-significant (bootstrap p ranges from 0.14 to 0.99), so the conclusion does not hinge on any particular run. These formal tests replace the

previous qualitative "the gain is smaller than its variance" with a quantitative statement: **the seed-level gain of BMA is not statistically different from zero on either mAP50 or mAP50-95.**

To further distinguish "absence of evidence" from "evidence of absence," we supplement the significance test with a two-one-sided equivalence test (TOST) and a power analysis. The eight-seed paired design has 80% power to detect a mean BMA gain of 2.64 mAP50 points (paired t-test, $\alpha = 0.05$, two-sided); the observed +0.44-point mean gain is far below this threshold, whereas the +2.85-point single-run (seed = 0) gain would have been detected with 85.7% power had it been the true mean effect — the failure to reach significance therefore weighs against the "real but under-powered effect" explanation. Consistently, a TOST with an equivalence margin of 2.5 mAP50 points is significant (both one-sided tests reject at $p < 0.05$), meaning the data are statistically equivalent to no effect at that margin; in other words, we can rule out a true mean BMA gain larger than about 2.5 mAP50 points at the 95% confidence level. The choice of this 2.5-point margin is deliberately conservative rather than data-driven: it is only about 19% of the rotated-box gain (+12.86) and 43% of the IR-fusion gain (+5.79), i.e., well below any effect a detector engineer would regard as a genuine improvement, so that a significant TOST at this margin carries a clear "practically no effect" interpretation. Moreover, the equivalence conclusion is not sensitive to this particular margin: because the 95% bootstrap confidence interval of the mean seed-level gain is [−0.011, +0.019] mAP50 points, any equivalence margin above roughly 0.02 mAP50 points (including margins of 0.5, 1.0, or 2.0 points) still yields a significant TOST. The reported 2.5-point margin is thus a conservative upper bound, and the data in fact support a far tighter equivalence claim.

There are two possible sources of this instability. First, VEDAI contains only 1089 training images; under small data, the training curve itself oscillates violently (especially during Mosaic augmentation), and the selection of the best model carries an obvious "lucky peak" effect. Second, BMA's offset field introduces additional free parameters, making optimization harder and more sensitive to initialization and random seeds. To ground the second explanation quantitatively, we directly characterize the sharpness of the converged solutions.

Following the weight-perturbation protocol of Keskar et al. [23], we add layer-wise Gaussian noise — scaled by a fraction α of each layer's parameter standard deviation — to the best checkpoint of each model and measure the resulting mAP50 (Figure 12). Under a 1% perturbation, the unaligned baseline drops by only 0.09 mAP50 points, whereas BMA drops by 0.58 points; under 2%, the drops widen to 0.21 vs 0.77 points. In other words, **BMA's converged solution is 4–6× more sensitive to weight perturbation than the unaligned baseline** — a perturbation-sensitivity signature that is *consistent with*, but not conclusive proof of, a sharper minimum. This sharp-minimum characterization, however, rests on indirect evidence. A direct Hessian-spectrum analysis could not be completed for the BMA model, because the second-order derivatives of the differentiable grid sampling used in the offset-field warp are not supported by the underlying deep-learning library — a hard technical limitation rather than a null result. A one-dimensional loss-landscape interpolation along the dominant perturbation direction likewise returned a nearly flat loss profile for both models (training-loss change < 0.1 over a $\pm 5\%$ parameter-standard-deviation range), which superficially contradicts the mAP collapse under the same perturbation. This apparent conflict indicates that BMA's sharpness, if present, resides in the *performance* (mAP) landscape rather than the *loss* landscape — mAP is a discontinuous metric that depends on score thresholds and NMS and is therefore far more sensitive than the smoothed average loss. Accordingly, we regard the sharp-minimum account as a *plausible but tentative* mechanistic explanation of the inter-seed instability: the warping-based alignment module introduces additional trainable degrees of freedom (offset fields at three scales), which may sharpen the optimization landscape around the solution, so that different seeds converge to narrow, initialization-sensitive optima whose mAP fluctuates by several points while their validation losses remain nearly identical. This is consistent with the $\approx 2.2\times$ inter-seed standard

deviation of BMA on mAP50, but a direct Hessian-spectrum confirmation remains open.

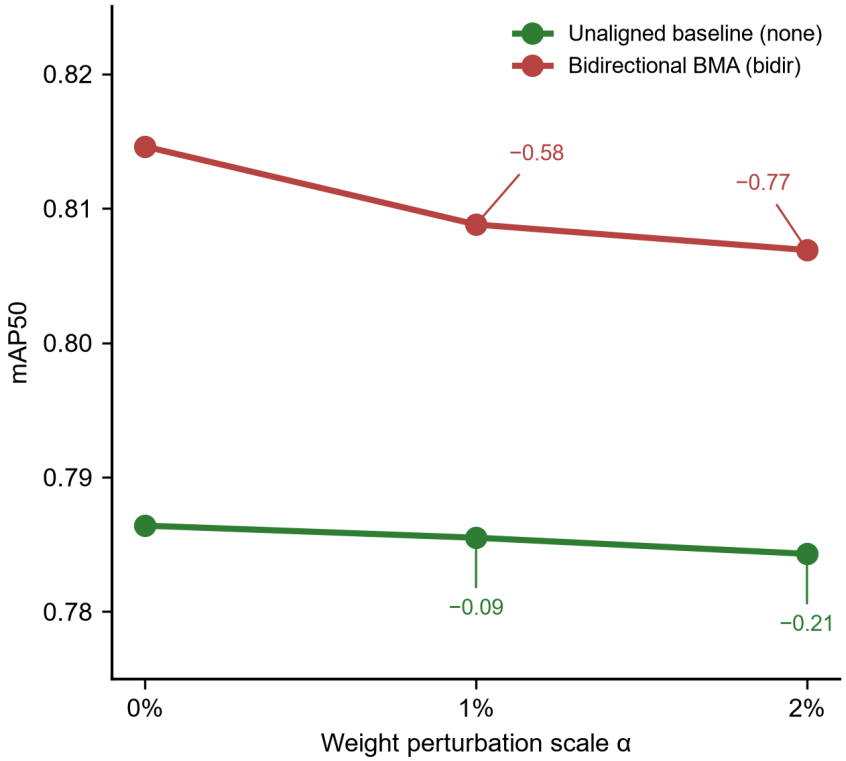


Figure 12: Weight-perturbation sensitivity of the converged solutions (Keskar et al. [23] protocol). The unaligned baseline (green) degrades by only 0.09/0.21 mAP50 points under 1%/2% layer-wise Gaussian perturbation, whereas BMA (red) degrades by 0.58/0.77 points — a 4–6× higher sensitivity, a perturbation-sensitivity signature consistent with, but not conclusive proof of, a sharper minimum (see Section 4.9).

Therefore, this paper honestly reports the single-run (seed=0) +2.85 gain and its ablation patterns (Section 4.6), while honestly delineating the statistical boundary of that gain in this section — **BMA is an exploratory attempt at explicit modality alignment in this paper, and the robustness of its gain awaits further verification on larger-scale data or more stable optimization strategies** (e.g., EMA, stronger weight decay, or sharpness-aware training).

4.10 Cross-Dataset Reproduction of the Negative Result

A natural objection to the conclusion of Section 4.4 is that it is established on a single dataset: VEDAI is already well registered (median misalignment 0.078 pixels), so "alignment does not help" might simply reflect the absence of misalignment to correct, and could be a dataset-specific anecdote rather than a general principle. To test whether the negative result generalizes, we reproduce the core none-vs-bidir comparison on DroneVehicle [22], a widely used RGB–IR aerial vehicle-detection benchmark. DroneVehicle contains 28,439 paired RGB–IR frames with five vehicle classes (car, truck, bus, van, freight car); following its official split, we use 17,951 pairs for training and 1,469 for validation — about 16.5× more training images than VEDAI, which also makes this a larger-scale check. We keep the same YOLOv8s dual-branch framework, the same 100-epoch / 640-resolution / SGD training protocol, and the same three-seed (seed 0–2) repetition as our VEDAI experiments. Following the official annotations, the raw 840×712 frames are cropped to 640×512 (removing the white borders), the five classes are mapped (including three spelling variants of "freight car"), and oriented labels are derived from the original polygon

annotations, with axis-aligned bounding boxes treated as zero-angle oriented boxes.

Table 12: Cross-dataset reproduction on DroneVehicle (5 classes, mean ± std over three random seeds)

Configuration	**mAP50**	**mAP50-95**
none (no alignment)	0.850 ± 0.003	0.700 ± 0.004
bidir (BMA)	0.852 ± 0.002	0.702 ± 0.002

As Table 12 shows, explicit bidirectional alignment again yields only a marginal gain: the mean seed-level differences are +0.20 mAP50 points and +0.21 mAP50-95 points. Because only three seeds are available, these differences are reported descriptively and are not subjected to a significance test (a paired t-test at $n = 3$ would be under-powered and sensitive to non-normality). We stress that this three-seed design cannot support an "indistinguishable from zero" claim: with only three seeds, any paired test has limited power — at 80% power and $\alpha = 0.05$ it can detect effects no smaller than about 0.35 mAP50 points and 0.48 mAP50-95 points, so an effect smaller than this (including the observed +0.20/+0.21) cannot be excluded. What the data do support is that the alignment gain, if any, is too small to be of practical consequence on registered RGB–IR aerial vehicle-detection data: +0.20/+0.21 points are an order of magnitude below the rotated-box gain (+12.86) and the IR-fusion gain (+5.79). This second-dataset result is consistent with the VEDAI finding, and the two together indicate that the tacit assumption "modality misalignment is the main obstacle to fusion" is not empirically supported on either benchmark. Three caveats keep this claim honest. First, both VEDAI and DroneVehicle are officially registered paired datasets, so the reproduction confirms the null on registered data; the misaligned regime — where the assumption would actually bite — is tested only synthetically (Section 4.8), not on a genuinely misaligned real dataset. Second, we did not re-measure the registration of DroneVehicle (the phase-correlation diagnosis of Section 4.4 was performed on VEDAI only); the claim is strictly about the relative none-vs-bidir gap within the same framework. Third, the classes and scale differ from VEDAI (5 vs 9 classes), so the reproduction concerns the alignment hypothesis rather than the absolute detection difficulty. Even with these caveats, the consistent near-zero gap across two independently-acquired, registered datasets provides preliminary cross-dataset evidence that extends the negative result beyond a single-dataset anecdote.

5 Discussion

This paper's systematic study reveals a picture different from the common perception in the field. **First**, the rotated-box detection head (+12.86 points mAP50-95) and the introduction of IR through multi-modal fusion (+5.79 points) constitute the dominant and consistent performance gain, and are statistically significant across three random seeds (paired t-test $p < 0.0001$ and $p = 0.0026$, respectively), confirming that "rotated box + multi-modal" is an effective combination for aerial vehicle detection. **Second, and most worth emphasizing**, this paper provides evidence for two assumptions in the field that lack rigorous empirical verification: first, phase-correlation measurements (cross-validated by calibration) show that the median RGB–IR misalignment of VEDAI is only 0.078 pixels with a systematic component of merely 0.004 pixels, indicating that this dataset is well registered and misalignment is not the primary bottleneck; second, discriminability experiments (with permutation tests) show that IR contributes no significant incremental discriminative power ($p = 0.81$) for the car/pickup confusion pair. These two negative results suggest that, on a well-registered dataset like VEDAI, focusing research effort on "modality alignment" or "multi-modal fine-grained discrimination" may yield limited returns. In contrast, data profiling, the confusion matrix, and discriminability analysis jointly point to the true challenge of VEDAI: 95.3% of objects are extremely small, bidirectional car–pickup confusion accounts for 21.9% of all objects, and this confusion is, under a linear 1-NN separa-

bility probe, primarily geometric in nature (area and aspect ratio, 71.8% classification accuracy, only about 2 points above texture) rather than textural — this finding points the direction for future work, namely that the focus of improvement should be "geometric/shape discrimination of extremely small objects."

From an engineering perspective, the bidirectional BMA model exchanges 40.12M parameters and 37.08G FLOPs (compared with the unaligned baseline's 27.70M and 25.71G, a 44.8% increase in parameters and a 44.2% increase in FLOPs) for a single-run +2.85-point mAP50 (eight-seed mean +0.44, insignificant), with the inference speed dropping from 57.6 FPS to 47.2 FPS (about −18%). Since the offset prediction head consists of only a few 3×3 convolutions, its inference overhead is relatively limited; however, combined with the "insignificant gain" conclusion of Section 4.9, this input-output ratio is not worthwhile on VEDAI, and this paper accordingly advises against paying the extra parameter and computational cost for an alignment module on a well-registered dataset. In addition, the dual-branch architecture itself increases parameters compared with single-branch early fusion (about 11.4M, corresponding to the early-fusion configuration in Table 2), but exchanges this for the independence of the modality branches, facilitating future extension to a third modality or independent fine-tuning of a single modality.

Regarding the BMA module, this paper's exploratory results (single-run +2.85 points, eight-seed mean +0.44 points, insignificant) and the conclusion of the misalignment diagnosis (misalignment ≈0.1 px) corroborate each other: since the misalignment is already insignificant, an explicit alignment module naturally finds it hard to obtain a stable gain. This result does not mean that explicit alignment schemes have no value in themselves; it honestly delineates BMA's scope of effect on VEDAI, and — as our cross-dataset reproduction shows (Section 4.10) — the alignment module again yields no practically meaningful gain on a second, larger dataset (DroneVehicle, +0.21 mAP50-95 points over three seeds). The offset-field visualization (Figure 10) further shows that the offset magnitude learned by the model (p3 about 0.3–0.5 pixels) is clearly larger than the true misalignment (0.078 pixels); this "magnitude mismatch" suggests that the learned offset mixes non-misalignment factors, consistent with the conclusion that "BMA's gain is insignificant"; this phenomenon deserves further investigation.

This study has four main limitations. **First**, the experiments are completed only on fold01 of VEDAI; a single-fold result may contain some split variance, and the robustness of the conclusions awaits further verification by multi-fold cross-validation. **Second**, the comparison with some advanced methods (such as PACGNet) requires further verification of protocol consistency to rule out bias introduced by protocol differences. **Third**, the conclusions of this paper's empirical re-examination (misalignment ≈0.1 px, IR no increment) are established on well-registered datasets: the misalignment diagnosis (relying on the single phase-correlation method, cross-validated by a calibration experiment with an absolute accuracy of about 0.2 pixels) is limited to VEDAI, and the alignment null is reproduced on DroneVehicle — another registered dataset, and with only three seeds — so the genuinely misaligned regime has so far been tested synthetically only (Section 4.8); whether the conclusions generalize to datasets with larger, genuine sensor misalignment remains to be fully verified. **Fourth**, the discriminability analysis uses 1-NN, a linear, non-parametric classifier, as a proxy for "feature discriminative power," which may underestimate the discriminative information that a non-linear discriminator (such as a deep network) can extract from texture; therefore, the "IR no increment" conclusion should be understood as "no increment in the sense of simple linear separability," rather than a universal assertion across all model architectures. Future work will expand the benchmark scope, add multi-fold and multi-dataset verification, and design improved methods targeting the true bottleneck revealed in this paper — the geometric discrimination of extremely small objects. It should also be noted that aerial vehicle detection is a dual-use technology; this work is intended for civilian applications such as traffic monitoring, urban planning, and disaster response, and no harmful use is endorsed.

6 Conclusion

This paper constructs a rotated-box multi-modal vehicle detection framework based on YOLOv8s and conducts a systematic empirical study on VEDAI. On the framework side, the rotated-box detection head improves mAP50-95 by 12.86 points and introducing IR through multi-modal fusion further improves it by 5.79 points (means over three random seeds; paired t-test $p < 0.0001$ and $p = 0.0026$, respectively), constituting a dominant and consistent performance gain that is statistically robust across seeds. On the empirical re-examination side, this paper obtains three mutually corroborating findings: (1) the median RGB–IR spatial misalignment of VEDAI is only 0.078 pixels with a systematic component of merely 0.004 pixels (phase-correlation measurement, cross-validated by calibration), and modality misalignment is not the primary bottleneck for multi-modal detection on this dataset; (2) IR contributes no significant incremental discriminative power (permutation test $p = 0.81$) for the car/pickup confusion pair; (3) the true challenge of VEDAI is class confusion dominated by extremely small objects (95.3% of objects smaller than 32×32 pixels), with bidirectional car–pickup confusion accounting for 21.9%, and this confusion being, under a linear 1-NN separability probe, primarily geometric in nature (71.8% classification accuracy, only about 2 points above texture) rather than textural. The explored BMA module shows an insignificant gain under multiple random seeds, corroborating the conclusion of the misalignment diagnosis; on DroneVehicle (Section 4.10), explicit alignment again yields no practically meaningful gain (+0.21 mAP50-95 points over three seeds), providing preliminary cross-dataset evidence consistent with this conclusion. Methodologically, this paper contributes a reusable statistical protocol — multi-seed training combined with paired bootstrap/permutation tests, a two-one-sided equivalence test, and a power analysis — that turns an otherwise qualitative "not significant" judgment into a quantitative, reproducible one, and is directly transferable to other small-sample detection benchmarks. In summary, this study provides systematic empirical evidence for small-object multi-modal detection in aerial imagery, and indicates on two independently registered datasets (VEDAI and DroneVehicle) that modality misalignment is not its primary bottleneck and that IR provides no fine-grained discrimination increment — this suggests that, on well-registered datasets such as VEDAI, the research focus should shift from "modality alignment / multi-modal fine-grained discrimination" toward "geometric discrimination of extremely small objects" — a conditional rather than universal conclusion, since our synthetic-misalignment experiment (Section 4.8) shows that misalignment does materially impair detection when it is large (an 8-pixel shift costs 15.0 mAP50 points), so alignment may remain essential on poorly-registered datasets; however, the generalization of this conclusion to datasets with even larger sensor differences still awaits future multi-dataset and multi-fold verification — though the alignment null result has now been reproduced across two registered datasets (VEDAI and DroneVehicle, Section 4.10).

Data availability

The VEDAI dataset is publicly available at `https://downloads.greyc.fr/vedai/`, and the DroneVehicle dataset at `https://github.com/VisDrone/DroneVehicle` (accessed on 8 September 2026). Trained model checkpoints will be released upon acceptance.

Code availability

The source code is publicly available at `https://github.com/someonewiee/VEDAI-RGBIR-OBB`.

Acknowledgements

This research received no external funding.

Author contributions

Y.W. is the sole author and performed all aspects of the work: conceptualization, methodology, software, validation, formal analysis, investigation, resources, data curation, writing (original draft, review and editing), and visualization.

Competing interests

The author declares no competing interests.

References

References

[1] Xia, G.-S. *et al.* DOTA: a large-scale dataset for object detection in aerial images. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.* 3974–3983 (2018).

[2] Murrugarra-Llerena, J., Kirsten, L. N., Zeni, L. F. A. & Jung, C. R. Probabilistic intersection-over-union for training and evaluation of oriented object detectors. *IEEE Trans. Image Process.* **33**, 671–681 (2024).

[3] Guo, H., Sun, C., Zhang, J., Zhang, W. & Zhang, N. MMYFnet: multi-modality YOLO fusion network for object detection in remote sensing images. *Remote Sens.* **16**, 4451 (2024).

[4] Ma, J., Ma, Y. & Li, C. Infrared and visible image fusion methods and applications: a survey. *Inf. Fusion* **45**, 153–178 (2019).

[5] Lan, X., Zhang, S., Bai, Y. & Qin, X. Fine-grained multispectral fusion for oriented object detection in remote sensing. *Remote Sens.* **17**, 3769 (2025).

[6] Zhang, X. *et al.* Efficient object detection in large-scale remote sensing images via situation-aware model. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **18**, 22486–22498 (2025).

[7] Razakarivony, S. & Jurie, F. Vehicle detection in aerial imagery: a small target detection benchmark. *J. Vis. Commun. Image Represent.* **34**, 187–203 (2016).

[8] Jocher, G., Chaurasia, A. & Qiu, J. Ultralytics YOLOv8. `https://github.com/ultralytics/ultralytics` (2023).

[9] Fang, Q. & Wang, Z. Cross-modality attentive feature fusion for object detection in multispectral remote sensing imagery. *Pattern Recognit.* **130**, 108786 (2022).

[10] Zhang, J. *et al.* SuperYOLO: super resolution assisted object detection in multimodal remote sensing imagery. *IEEE Trans. Geosci. Remote Sens.* **61**, 5604515 (2023).

[11] Zhang, J. *et al.* Guided hybrid quantization for object detection in remote sensing imagery via one-to-one self-teaching. *IEEE Trans. Geosci. Remote Sens.* **61**, 5614815 (2023).

[12] Shen, J. *et al.* ICAFusion: iterative cross-attention guided feature fusion for multispectral object detection. *Pattern Recognit.* **145**, 109913 (2024).

[13] Zhang, Q. *et al.* Selective structured state space for multispectral-fused small target detection. Preprint at `https://arxiv.org/abs/2505.14043` (2025).

[14] Gu, Z. & Tian, S. Pyramidal adaptive cross-gating for multimodal detection. Preprint at `https://arxiv.org/abs/2512.18291` (2025).

[15] Jaderberg, M., Simonyan, K., Zisserman, A. & Kavukcuoglu, K. Spatial transformer networks. In *Advances in Neural Information Processing Systems 28* (2015).

[16] Dai, J. *et al.* Deformable convolutional networks. In *Proc. IEEE Int. Conf. Comput. Vis.* 764–773 (2017).

[17] Zhu, X., Hu, H., Lin, S. & Dai, J. Deformable ConvNets v2: more deformable, better results. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.* 9308–9316 (2019).

[18] Dosovitskiy, A. *et al.* FlowNet: learning optical flow with convolutional networks. In *Proc. IEEE Int. Conf. Comput. Vis.* 2758–2766 (2015).

[19] Teed, Z. & Deng, J. RAFT: recurrent all-pairs field transforms for optical flow. In *Proc. Eur. Conf. Comput. Vis.* 402–419 (2020).

[20] Balakrishnan, G., Zhao, A., Sabuncu, M. R., Guttag, J. & Dalca, A. V. VoxelMorph: a learning framework for deformable medical image registration. *IEEE Trans. Med. Imaging* **38**, 1788–1800 (2019).

[21] Avants, B. B., Epstein, C. L., Grossman, M. & Gee, J. C. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Med. Image Anal.* **12**, 26–41 (2008).

[22] Sun, Y., Cao, B., Zhu, P. & Hu, Q. Drone-based RGB-infrared cross-modality vehicle detection via uncertainty-aware learning. *IEEE Trans. Circuits Syst. Video Technol.* **32**, 6700–6709 (2022).

[23] Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M. & Tang, P. T. P. On large-batch training for deep learning: generalization gap and sharp minima. In *Proc. Int. Conf. Learn. Represent.* (2017).