BRIDGE: BOOSTING REPRESENTATION FIDELITY VIA SHAPE-INFORMED DISTANCE-FIELD GEOMETRIC ENCODING FOR MULTIMODAL OBJECT DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal object detection seeks to exploit complementary information across heterogeneous sensing modalities for robust object perception. Different modalities reveal distinct structural and geometric cues of the same object through different imaging principles. However, heterogeneous imaging mechanisms complicate cross-modal geometric interaction, while complex object geometries remain difficult to capture with shapes parameterized by only a few parameters. In this paper, we introduce BRIDGE, a novel geometric representation framework for object detection that moves beyond predefined shapes toward continuous geometry modeling. Specifically, we observe that Signed Distance Fields (SDF) map heterogeneous modalities into a shared continuous space while preserving fine-grained geometry. We therefore introduce Continuous Geometry Encoding (CGE) to represent objects beyond predefined parametric shapes. We further propose Fourier-Conditioned Geometry Fusion (FCGF) to preserve informative geometric cues while exploiting cross-modal complementarity through frequency-conditioned residual refinement. Finally, we dive into SDF optimization during training and introduce Gradient-Modulated Loss (GML). GML suppresses gradients from well-fit regions while strengthening corrections for severe geometric deviations. Our analysis characterizes the bounded and spatially distributed optimization behavior of SDF supervision, while experiments validate the effectiveness of the resulting geometric framework. BRIDGE achieves 85.0% mAP_{50} on DroneVehicle, outperforming existing multimodal detectors and demonstrating the effectiveness of continuous geometry modeling. The code and models will be available on GitHub: to facilitate future work.

1 INTRODUCTION

Object detection aims to localize and recognize objects in complex visual scenes, where geometric representation determines how object location, extent, and shape are described. The oriented bounding box (OBB) is typically parameterized as $B = (c_x, c_y, w, h, \theta)$, where (c_x, c_y) denotes the box center, w and h denote its width and height, and θ denotes its orientation angle. This representation is widely adopted for objects with arbitrary orientations due to its compactness and efficiency (Ding et al., 2019; Xie et al., 2021). OBB-based detection is particularly important in remote sensing, where objects frequently exhibit arbitrary orientations, dense distributions, and large scale variations. Multimodal oriented object detection further exploits complementary observations from heterogeneous sensors to improve robustness under challenging imaging conditions. Recent studies have advanced from adaptive RGB–infrared feature fusion (Shen et al., 2024; Zhu et al., 2025) to condition-aware multimodal interaction that dynamically adjusts modality contributions under diverse sensing environments (Chen et al., 2025a; Hao et al., 2026). Meanwhile, emerging studies increasingly emphasize that effective multimodal learning requires not only cross-modal fusion but also preservation of modality-specific information (Zhao et al., 2025b; Wang et al., 2026). This exposes a more fundamental representation problem: how to preserve the modality-invariant geometry of the same object while retaining modality-specific geometric evidence induced by different sensing mechanisms.

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.

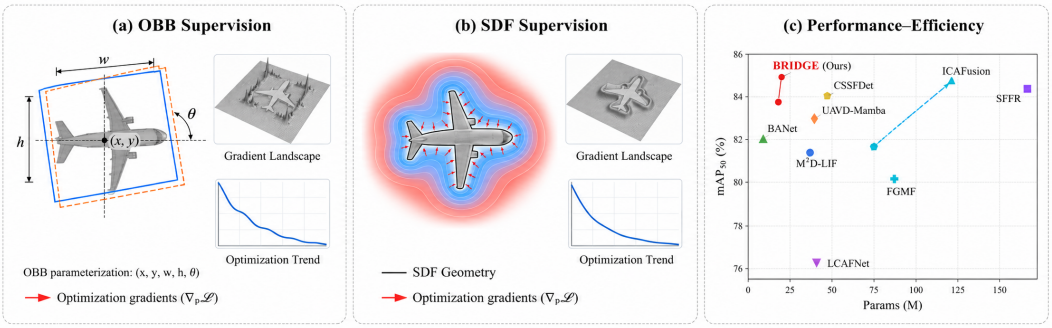

Figure 1: Comparison of geometric representations and their performance–efficiency trade-offs. (a) OBB supervision compresses geometry into a few parameters, coupling local errors into parameter-level updates and potentially causing conflicting gradients under spatial misalignment. (b) SDF supervision represents geometry as a continuous spatial field, preserving fine-grained boundaries while distributing residuals and gradients spatially. (c) BRIDGE achieves a favorable accuracy–complexity trade-off among multimodal detectors.

Geometric representation determines not only which object structures are preserved, but also how localization errors are formulated and propagated during optimization. Conventional oriented object detection predominantly localizes objects through compact OBB parameters, while more recent work has increasingly revisited the representation itself. COBB (Xiao et al., 2024) explicitly studies the continuity of OBB encoding and develops a continuous box representation, while Gaussian-based formulations further reformulate OBB geometry as differentiable probability distributions (Marques et al., 2025; Zhao et al., 2025a). Recent approaches further exploit frequency-domain rotation properties (Gu et al., 2026) or intrinsic geometric priors such as rotation invariance and structural symmetry (Zhan et al., 2026). Despite their differences, these methods still rely on finite-dimensional or predefined geometric forms, which only indirectly capture boundaries and local structure. In multimodal settings, this makes it difficult to preserve shared object geometry while retaining modality-specific variations.

These limitations motivate us to move beyond sparse parametric geometry toward a shared continuous space for multimodal geometric modeling. In this paper, we introduce **BRIDGE**, a continuous geometric framework that unifies representation, cross-modal interaction, and optimization through SDFs. This formulation provides a flexible shared space for preserving fine-grained geometry and modality-specific evidence, and naturally extends from OBBs to irregular contours generated by SAM3 (Carion et al., 2026). Based on this representation, FCGF estimates modality reliability from frequency-domain cues for adaptive geometric refinement (Sun et al., 2025), while GML reallocates optimization emphasis according to SDF residual severity beyond conventional box-level objectives (Yao et al., 2024). Theoretical analysis characterizes the bounded and spatially distributed optimization behavior, while experiments demonstrate the effectiveness of continuous geometry modeling. Our contributions are summarized as follows:

- We introduce CGE to map modality-specific object geometry from predefined parametric shapes into a shared continuous field, enabling spatially distributed modeling and natural extension to fine-grained, non-parametric shapes.
- To exploit modality complementarity in a shared geometric space, we propose FCGF, which estimates modality reliability from frequency-domain cues and performs confidence-aware residual refinement.
- We analyze the optimization behavior of continuous SDF supervision and develop GML, which reallocates gradients across multimodal geometry branches according to residual severity; the analysis further reveals bounded and spatially distributed optimization signals.

2 Related Work

2.1 Geometric Object Representation

Oriented bounding boxes (OBBs), typically parameterized by (c_x, c_y, w, h, θ), are widely used in oriented detection due to their compact and interpretable formulation. However, explicit angle parameterization suffers from boundary discontinuity and representation ambiguity, motivating alternative geometric representations. Early approaches alleviate these issues through alternative parameterizations or angle encodings, such as the Gliding Vertex (Xu et al., 2021) and DCL (Yang et al., 2021a). Point-based methods further relax rigid box constraints; Oriented RepPoints (Li et al., 2022), for example, represent objects with adaptive point sets to capture more flexible geometry. Another line of work reformulates rotated objects as probability distributions. GWD (Yang et al., 2021c), KLD (Yang et al., 2021d), and G-Rep (Hou et al., 2023) exploit Gaussian formulations to obtain smoother and more geometry-aware optimization. More recently, COBB (Xiao et al., 2024) and GauCho (Marques et al., 2025) further explore continuous and distribution-based representations, reflecting a broader trend toward moving beyond conventional angle-based box parameterization.

2.2 Oriented Object Detection

Oriented object detection aims to localize objects with arbitrary orientations and has evolved from rotation-aware feature alignment toward more flexible localization and optimization strategies. R3Det (Yang et al., 2021b) progressively refines rotated predictions and realigns features according to updated boxes, while S^2A-Net (Han et al., 2022) combines anchor refinement with alignment convolution to reduce the mismatch between object orientation and sampled features. Beyond feature alignment, another line of work focuses on improving rotated box optimization. KFIoU (Yang et al., 2023), for example, introduces a geometry-aware regression objective to provide more stable localization signals. More recently, Transformer-based detectors have extended end-to-end modeling to oriented scenarios. OrientedFormer (Zhao et al., 2024) incorporates orientation-aware geometric modeling into Transformer detection, while ReDiffDet (Zhao et al., 2025a) further formulates oriented detection as a rotation-equivariant diffusion process. Overall, recent progress has gradually shifted from rotation-specific alignment and refinement toward geometry-aware optimization and end-to-end oriented detection.

2.3 Multimodal Object Detection

Multimodal object detection exploits complementary information from heterogeneous sensors to improve robustness under challenging imaging conditions. In RGB–infrared detection, visible images provide richer texture cues, whereas infrared images remain informative under poor illumination. UA-CMDet (Sun et al., 2022) introduces uncertainty- and illumination-aware fusion, while CMAFF (Fang & Wang, 2022) models shared and modality-specific cues via cross-modal attention.

With advances in Transformer architectures, recent methods increasingly model explicit cross-modal dependencies. CFT (Fang et al., 2021) captures intra- and inter-modal relationships, while ICAFusion (Shen et al., 2024) and C^2Former (Yuan & Wei, 2024) enhance feature exchange through iterative or calibrated attention. In aerial imagery, spatial misalignment poses an additional challenge. CAGTDet (Yuan et al., 2024) combines geometric alignment with complementary fusion, while CMAI-Det (Wen et al., 2025) explores alignment and interaction in drone scenes. Frequency-domain approaches have also emerged: MGFF (Chen et al., 2025b) uses wavelet decomposition to reduce modality discrepancies, while WaveMamba (Zhu et al., 2025) fuses complementary frequency subbands. CrossWeaver (Yang et al., 2026) addresses weakly aligned multispectral detection through shared semantics, modality-specific deformation, and cross-hierarchical interaction.

3 Method

3.1 Overview

We propose BRIDGE, a multimodal oriented detection framework consisting of three components: Continuous Geometry Encoding (CGE), Fourier-Conditioned Geometry Fusion (FCGF), and

Gradient-Modulated Loss (GML), as illustrated in Fig. 2. CGE constructs a shared continuous geometry space for multimodal objects. FCGF performs modality-conditioned geometric fusion in this space, while GML modulates optimization according to geometric residual severity.

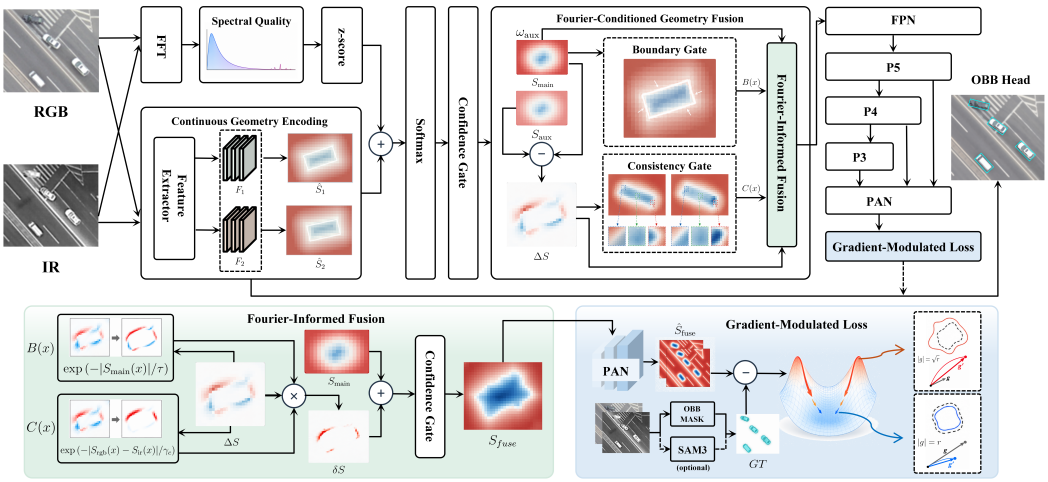

Figure 2: Overview of BRIDGE. CGE constructs a shared geometric space, FCGF performs frequency-conditioned geometric fusion, and GML modulates optimization through geometric residuals.

3.2 Continuous Geometry Encoding

CGE maps modality-specific object geometry into a shared continuous field, supporting both parametric oriented boxes and fine-grained object contours. Beyond representation construction, we further analyze how this continuous formulation responds to spatial perturbations and propagates geometric errors during optimization.

Continuous Geometry Construction For an object region $\mathcal{E} \subset \mathbb{R}^2$ with boundary $\partial\mathcal{E}$, its SDF is defined as

$$\phi_{\mathcal{E}}(\mathbf{q}) = \mathrm{sgn}_{\mathcal{E}}(\mathbf{q}) d(\mathbf{q}, \partial\mathcal{E}), \tag{1}$$

where $\mathbf{q} \in \mathbb{R}^2$ is a spatial location, $d(\mathbf{q}, \partial\mathcal{E})$ denotes its Euclidean distance to the boundary, and $\mathrm{sgn}_{\mathcal{E}}(\mathbf{q}) = -1$ inside $\mathcal{E}$ and $+1$ outside. Accordingly,

$$\partial\mathcal{E} = \{\mathbf{q} \in \mathbb{R}^2 \mid \phi_{\mathcal{E}}(\mathbf{q}) = 0\}. \tag{2}$$

CGE therefore represents object geometry as a continuous scalar field rather than a finite set of geometric parameters.

CGE instantiates this representation through two complementary routes. For standard oriented annotations, consider a predicted OBB $B^p = (x_p, y_p, w_p, h_p, \theta_p)$, with center $\mathbf{c}_p = (x_p, y_p)^\top$ and half-size $\mathbf{a}_p = (w_p/2, h_p/2)^\top$. Its continuous field at spatial sample $\mathbf{q}_i$ is

$$\phi_{B^p}(\mathbf{q}_i) = \Psi\left(\left|R(\theta_p)^\top(\mathbf{q}_i - \mathbf{c}_p)\right| - \mathbf{a}_p\right), \tag{3}$$

where $R(\theta_p)$ is the 2D rotation matrix and $|\cdot|$ denotes element-wise absolute value. The function $\Psi(\cdot)$ is the SDF of an axis-aligned rectangle:

$$\Psi(\mathbf{u}) = \|[\mathbf{u}]_+\|_2 + \min\left(\max(u_x, u_y), 0\right), \tag{4}$$

where $\mathbf{u} = (u_x, u_y)^\top$ and $[\mathbf{u}]_+$ denotes the element-wise positive part. The same transformation is applied to the ground-truth OBB B^t, so predicted and target boxes share the same continuous geometric space. This common representation provides a shared geometric basis across modalities.

The same representation is not restricted to rectangular geometry. Given a ground-truth box B^t, we additionally use it as a box prompt for SAM3 (Carion et al., 2026) to obtain an object mask $\mathcal{M}^t$. Its corresponding continuous field is

$$\phi_{\mathcal{M}^t}(\mathbf{q}) = \mathrm{sgn}_{\mathcal{M}^t}(\mathbf{q}) d(\mathbf{q}, \partial\mathcal{M}^t). \tag{5}$$

While the box-induced field follows the oriented rectangular boundary, the mask-induced field follows the finer object contour and can therefore encode non-rectangular geometry without introducing an additional parametric shape model. SAM3 is used only for offline supervision generation and introduces no additional computation during inference.

Compared with Gaussian-based representations such as GWD (Yang et al., 2021c), CGE retains explicit spatial information after encoding. GWD measures box discrepancy through Gaussian distributions, whereas CGE assigns each spatial location a signed distance to the object boundary. This allows geometric errors to be localized and optimized spatially, while the same formulation also applies to irregular object contours.

Stability under Spatial Perturbations We next analyze how spatial misalignment affects both the continuous representation and its induced optimization signal. Consider two modality observations of the same target differing by a spatial displacement $\boldsymbol{\delta}$. Let $\phi_t(\mathbf{q})$ denote the reference field and $\phi_m(\mathbf{q}) = \phi_t(\mathbf{q}-\boldsymbol{\delta})$ the shifted one. Since an SDF is 1-Lipschitz with respect to spatial coordinates,

$$|\phi_m(\mathbf{q}) - \phi_t(\mathbf{q})| = |\phi_t(\mathbf{q}-\boldsymbol{\delta}) - \phi_t(\mathbf{q})| \le \|\boldsymbol{\delta}\|_2. \tag{6}$$

Thus, spatial displacement produces a bounded change in the continuous geometric representation.

For a truncation threshold $T > 0$, we further define the normalized truncated field $\widetilde{\phi}(\mathbf{q}) = \mathrm{clip}(\phi(\mathbf{q}), -T, T)/T$. Its perturbation satisfies

$$\left|\tilde{\phi}_m(\mathbf{q}) - \tilde{\phi}_t(\mathbf{q})\right| \le \min\left(\|\delta\|_2/T,\, 2\right). \tag{7}$$

The representation error therefore grows with displacement only within a bounded range and saturates after truncation.

This bounded field perturbation further propagates to the optimization signal. Let g_α^t and g_α^m denote the gradients induced by the reference and shifted targets for a box parameter α. Then

$$\left|g_\alpha^m - g_\alpha^t\right| \le C_\alpha \min\left(\|\delta\|_2/T,\, 2\right). \tag{8}$$

where C_α depends on the spatial sampling weights and the local Jacobian of the box-induced field. Hence, cross-modal spatial misalignment induces bounded variations at both the representation and gradient levels, providing a stable basis for continuous geometric supervision. The complete derivations are provided in Appendix A.3 and Appendix A.4.

Spatially Distributed Optimization CGE also changes how geometric errors are propagated during training. For clarity, consider a canonical weighted squared objective over N spatial samples. For an arbitrary box parameter $\alpha \in \{x_p, y_p, w_p, h_p, \theta_p\}$, the induced gradient is

$$\frac{\partial \mathcal{L}_{\mathrm{SDF}}}{\partial \alpha} = \frac{2}{N}\sum_{i=1}^{N} w_i \left(\phi^p(\mathbf{q}_i) - \phi^t(\mathbf{q}_i)\right) \frac{\partial \phi^p(\mathbf{q}_i)}{\partial \alpha}. \tag{9}$$

Each spatial sample therefore contributes through two factors: its local geometric residual $\phi^p(\mathbf{q}_i) - \phi^t(\mathbf{q}_i)$ and its local response to the box parameter $\partial\phi^p(\mathbf{q}_i)/\partial\alpha$. Instead of assigning the entire localization error to a small set of box parameters, CGE decomposes the optimization signal over spatial locations, allowing different geometric regions to contribute independently.

This behavior contrasts with direct parametric regression, where inconsistent annotations may impose immediately opposing update directions. Consider two center annotations separated by $\boldsymbol{\delta}$ and a predicted center lying between them, parameterized by $s \in (0, 1)$. Under squared center regression, the two induced gradients satisfy

$$\mathbf{g}_1^\top \mathbf{g}_2 = -4s(1-s)\|\boldsymbol{\delta}\|_2^2 < 0, \qquad 0 < s < 1. \tag{10}$$

This simple case illustrates that inconsistent parametric targets can produce directly conflicting optimization directions.

CGE does not eliminate all such conflicts. Instead, it decomposes geometric supervision into spatially distributed residual contributions that can be modulated independently. This property provides the basis for multimodal learning in the shared space: FCGF selectively controls cross-modal geometric interaction in the shared continuous space, while GML adjusts the optimization contribution of each geometric residual according to its severity. Detailed Jacobian and gradient analyses are provided in Appendix A.2 and Appendix A.5, together with a visualization of the geometric optimization process in the continuous field (Fig. A.1).

3.3 Fourier-Conditioned Geometry Fusion

Once different modalities are mapped by CGE into a common continuous geometric space, FCGF estimates their sample-wise reliability from the original observations and performs reliability-guided geometric interaction. The more reliable modality preserves the dominant geometry, while the other modality provides spatially constrained residual refinement.

Fourier-Based Modality Reliability Estimation For each RGB or infrared input I_m, we compute its spectral magnitude as $A_m = |\mathcal{F}(I_m)|$, where $\mathcal{F}(\cdot)$ denotes the two-dimensional Fourier transform. We focus on the mid- and high-frequency bands, denoted by Ω_{mid} and Ω_{high}, respectively. For either band Ω_b, the average spectral response is computed as $E_m^b = |\Omega_b|^{-1} \sum_{(u,v)\in\Omega_b} A_m(u,v)$. These statistics are summarized by a Fourier quality function $q_m = \mathcal{Q}_{\text{Fourier}}(I_m)$, which characterizes the structural quality of the current observation from its mid- and high-frequency responses.

Because RGB and infrared images exhibit different spectral distributions, directly comparing their raw scores is inappropriate. We therefore normalize each modality using training-set statistics, $z_m = (\log q_m - \mu_m)/\sigma_m$, where μ_m and σ_m are the mean and standard deviation of $\log q_m$ for that modality. The two normalized scores are converted into relative modality weights by $[w_{\text{rgb}}, w_{\text{ir}}] = \text{Softmax}([z_{\text{rgb}}, z_{\text{ir}}]/T_f)$, where $T_f > 0$ controls the sharpness of the weighting.

Let S_{rgb} and S_{ir} denote the corresponding continuous geometric fields. The modality with the larger weight is treated as the main modality and the other as the auxiliary modality, yielding S_{main}, S_{aux}, and w_{aux} for the subsequent residual fusion.

Quality-Conditioned Residual Geometric Fusion FCGF preserves S_{main} as the dominant geometric representation and uses S_{aux} only for residual refinement. For a spatial location $\mathbf{x}$ in the continuous geometric field,

$$S_{\text{res}}(\mathbf{x}) = S_{\text{main}}(\mathbf{x}) + \lambda B(\mathbf{x}) C(\mathbf{x}) w_{\text{aux}} \left(S_{\text{aux}}(\mathbf{x}) - S_{\text{main}}(\mathbf{x})\right), \tag{11}$$

where λ controls the refinement strength, while $B(\mathbf{x})$ and $C(\mathbf{x})$ denote the boundary-proximity and cross-modal consistency gates, respectively.

The two gates are defined as

$$B(\mathbf{x}) = \exp\left(-|S_{\text{main}}(\mathbf{x})|/\tau\right), \qquad C(\mathbf{x}) = \exp\left(-|S_{\text{rgb}}(\mathbf{x}) - S_{\text{ir}}(\mathbf{x})|/\gamma_c\right). \tag{12}$$

where $\tau > 0$ and $\gamma_c > 0$ control boundary proximity and cross-modal consistency, respectively. Thus, $B(\mathbf{x})$ emphasizes refinement near the geometric boundary encoded by the zero level set, while $C(\mathbf{x})$ suppresses corrections under large cross-modal disagreement.

When the two modality weights are close, selecting a dominant modality becomes unreliable. We therefore introduce a confidence threshold $\delta_c \in [0, 1)$ and define

$$G = \text{clip}\left((|w_{\text{rgb}} - w_{\text{ir}}| - \delta_c)/(1 - \delta_c),\, 0,\, 1\right). \tag{13}$$

where $G \in [0, 1]$ denotes the confidence of the dominant-modality selection. The final fused geometric field is

$$S_{\text{fuse}} = (1 - G) S_{\text{avg}} + G S_{\text{res}}, \qquad S_{\text{avg}} = (S_{\text{rgb}} + S_{\text{ir}})/2. \tag{14}$$

where S_{avg} is the equal-weight average of the two modality-specific geometric fields.

Thus, FCGF uses frequency-domain reliability to find dominant geometry, while boundary proximity and cross-modal consistency control where complementary geometric information is introduced.

3.4 GRADIENT-MODULATED LOSS

In multimodal geometric learning, SDF residuals reflect local geometric deviations. Since a signed distance field satisfies $\|\nabla_{\mathbf{q}}\phi(\mathbf{q})\|_2 = 1$ almost everywhere, a small displacement $\Delta\mathbf{q}$ along the local normal gives $\Delta\phi \approx \nabla_{\mathbf{q}}\phi(\mathbf{q})^\top \Delta\mathbf{q}$, and hence $|\Delta\phi| \approx |\Delta d|$, where Δd denotes normal displacement. Thus, the residual magnitude provides a direct measure of local geometric deviation.

Let $\widehat{\phi}_i$ and ϕ_i denote the predicted and target geometric values at sample i, and define the residual $e_i = \widehat{\phi}_i - \phi_i$ with magnitude $r_i = |e_i|$. For a standard ℓ_1 objective, the gradient magnitude is constant, $|\partial|e_i|/\partial e_i| = 1$ for $e_i \neq 0$. Consequently, small and severe deviations across modality-specific geometry branches receive the same optimization strength, despite representing different levels of geometric error.

More generally, for a point-wise objective $\rho(e_i)$ over N samples, its gradient with respect to an arbitrary box or network parameter α is

$$\frac{\partial \mathcal{L}_{\text{geo}}}{\partial \alpha} = \frac{1}{N}\sum_{i=1}^{N} \underbrace{\rho'(e_i)}_{\text{residual influence}} \frac{\partial \widehat{\phi}_i}{\partial \alpha}. \tag{15}$$

Thus, $\rho'(e_i)$ directly determines the optimization emphasis assigned to each geometric residual. For simplicity, let e denote a generic residual and $r = |e|$ its magnitude. GML is designed to provide vanishing influence as $r \to 0$, while preserving a magnitude-sensitive but sub-linear response for severe geometric deviations.

Let $p \in [1, 2]$ denote the tail-response exponent. We define the gradient modulation function as

$$\psi_p(r) = \begin{cases} r, & r \leq 1, \\ r^{p-1}, & r > 1, \end{cases} \tag{16}$$

which attenuates small residuals. For $1 < p < 2$, the tail response is strictly increasing and sub-linear; $p = 1$ gives a constant tail response, while $p = 2$ gives a linear response. Integrating $\psi_p(r)$ and enforcing continuity at $r = 1$ gives

$$\rho_p(e) = \begin{cases} \dfrac{1}{2}r^2, & r \leq 1, \\ \dfrac{1}{2} + \dfrac{r^p - 1}{p}, & r > 1. \end{cases} \tag{17}$$

Its gradient is

$$\frac{\partial \rho_p}{\partial e} = \operatorname{sgn}(e) \begin{cases} r, & r \leq 1, \\ r^{p-1}, & r > 1. \end{cases} \tag{18}$$

For the final setting $p = 3/2$, the tail response becomes $r^{p-1} = \sqrt{r}$.

For $r \leq 1$, the gradient magnitude decreases linearly toward zero as the prediction approaches the target, reducing unnecessary updates on well-fitted geometric regions. For $r > 1$, the gradient magnitude $\sqrt{r}$ remains strictly increasing with residual severity, so larger geometric deviations receive progressively stronger correction. Meanwhile, its sub-linear growth prevents severe residuals from dominating optimization excessively. The loss is continuously differentiable at $r = 1$, providing a smooth transition between the two regimes. Detailed properties are provided in Appendix B.3.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets and metrics. Our main experiments are conducted on DroneVehicle, a paired RGB–IR aerial dataset containing five vehicle categories. We use its oriented bounding-box annotations for the main comparison and ablation studies, and additionally consider the horizontal detection setting to evaluate the generality of our formulation. We further conduct experiments on M4-SAR, which contains paired optical and SAR imagery, to examine generalization to a substantially different sensing modality. We report mAP_{50} and $\mathrm{mAP}_{50:95}$, together with class-wise AP_{50} on DroneVehicle.

Table 1: Comparison with recent multimodal object detectors on DroneVehicle. Class-wise results are reported in AP_{50} (%). The best and second-best results are highlighted in red and blue.

Method	Venue	Car	Truck	Bus	Van	Freight Car	mAP_{50} (%)	Params (M)
M^2D-LIF (Zhao et al., 2025b)	ICCV 2025	97.8	81.0	96.0	64.6	67.9	81.4	37.10
MSDF-Mamba (Shen et al., 2025)	Remote Sens. 2025	98.6	82.5	96.9	69.8	64.5	82.5	–
UAVD-Mamba (Li et al., 2025b)	IEEE IV 2025	98.6	83.9	96.9	66.1	69.8	83.0	39.70
SFFR (Zuo et al., 2025)	IEEE TGRS 2025	97.9	82.6	96.6	71.1	74.0	84.4	167.10
RGFNet (Zhao et al., 2025c)	IEEE TGRS 2025	98.7	80.2	95.5	67.9	76.6	83.8	–
MDPMFN (Liu et al., 2025)	IEEE TGRS 2025	98.0	80.9	95.5	64.3	67.7	81.3	–
FGMF (Lan et al., 2025)	Remote Sens. 2025	90.5	78.5	90.3	67.1	74.7	80.2	87.20
AFFNet (Chen et al., 2025c)	Neurocomput. 2025	90.2	66.0	88.1	48.8	50.9	68.8	63.02
IGIANet (Chen et al., 2026)	AAAI 2026	90.5	78.2	90.3	69.4	76.2	80.9	–
CMFADet (Yu & Shin, 2026)	IEEE TGRS 2026	98.6	82.9	96.2	67.8	71.6	83.4	–
BANet (Shi et al., 2026)	IEEE TGRS 2026	97.7	82.3	95.6	65.1	69.9	82.1	8.80
CMTDet (Li et al., 2026)	IEEE JSTARS 2026	98.5	83.0	96.4	66.2	70.0	82.8	–
ISGM (Liu et al., 2026)	IEEE JSTARS 2026	90.4	80.7	90.1	68.9	70.4	80.1	–
CTFLS (Qi et al., 2026)	Neural Netw. 2026	97.6	77.4	96.5	67.8	73.4	82.5	–
SIA (Cui et al., 2026)	Neural Netw. 2026	90.4	78.4	89.3	62.4	64.6	77.0	–
DEAF-Net (Weng et al., 2026)	Remote Sens. 2026	90.5	83.2	90.0	73.0	73.0	81.9	–
DMM (Zhou et al., 2025)	IEEE TGRS 2025	90.4	79.8	89.9	68.6	68.2	79.4	–
LMDENet (Weng & Niu, 2026)	Sensors 2026	98.6	59.8	95.9	72.6	67.6	78.9	–
DFDNet (Zhou et al., 2026)	ESWA 2026	97.5	83.0	96.5	69.3	65.1	82.3	–
LCAFNet (Wu et al., 2026)	PR 2026	90.3	69.7	89.7	63.4	67.9	76.2	41.10
SemFusion (Li et al., 2025a)	ACM MM 2025	90.4	80.0	89.9	68.5	68.9	79.6	–
CSSFDet (Jin et al., 2025)	ACM MM 2025	98.6	84.1	96.6	69.4	71.8	84.1	47.06
CrossWeaver* (Yang et al., 2026)	CVPR 2026	95.6	81.3	94.3	69.8	77.5	83.7	18.41
BRIDGE (Ours)	–	96.7	83.9	96.2	74.2	74.0	85.0	18.96

* denotes results reproduced using the experimental configuration reported in the original paper.

Implementation details. All experiments use one NVIDIA GeForce RTX 4090 GPU and, unless otherwise specified, the same training and evaluation protocol. BRIDGE is trained from scratch for 200 epochs with batch size 8, 800×800 inputs, and SGD (initial learning rate 0.01, cosine decay, and 5 warm-up epochs). Patience is 60 and the random seed is 1; AdaLoRA uses $r_{\text{init}} = 9$ and $r_{\text{target}} = 6$. The overall objective is $\mathcal{L} = \mathcal{L}_{\text{OBB}} + g_{\text{box}}\mathcal{L}_1^{\text{box-SDF}} + \sum_{k\in\{\text{rgb,ir,fuse}\}} g_k \mathcal{L}_{\text{GML}}^k$, where $\mathcal{L}_{\text{OBB}} = \mathcal{L}_{\text{box}} + \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{dfl}}$ and $p = 3/2$.

4.2 Main Results

Comparison on DroneVehicle. Table 1 compares BRIDGE with representative multimodal detectors on DroneVehicle. BRIDGE achieves the best overall mAP_{50} of 85.0%, exceeding the previous best of 84.4% by 0.60 percentage points. Class-wise AP_{50} is 96.7%, 83.9%, 96.2%, 74.2%, and 74.0% for car, truck, bus, van, and freight car, respectively. With only 18.96M parameters, substantially fewer than several competitive multimodal detectors, BRIDGE offers a favorable accuracy–complexity trade-off for multimodal detection.

4.3 Ablation Studies

Main Component Ablation. Table 2 evaluates the contributions of the core components. Introducing CGE improves the baseline from 69.3% to 72.5% mAP_{50}. Adding FCGF and GML further improves detection performance, indicating that both fusion and loss modulation contribute to the result.

Table 2: Main component ablations.

CGE	FCGF	GML	mAP_{50} (%)	Improve (%)
			69.3	–
✓			72.5	↑ 3.2
✓	✓		74.3	↑ 5.0
✓		✓	74.0	↑ 4.7

Continuous Geometry Encoding. We further compare CGE with representative geometric formulations. As shown in Table 3, CGE outperforms GWD, supporting shared continuous geometry modeling across modalities. Incorporating SAM3-derived object contours further improves performance to 75.8% mAP_{50}, showing that CGE can use fine-grained object contours.

Fourier-Conditioned Geometry Fusion. Table 4 examines the effects of the boundary-aware term $B(x)$ and the cross-modal consistency term $C(x)$. Without either term, the model obtains 68.9% mAP_{50}. Introducing $B(x)$ alone improves the performance to 73.3%, while using $C(x)$ alone achieves 72.8%. Combining both terms further increases mAP_{50} to 74.3%, showing they supply complementary cues for geometric fusion.

Gradient-Modulated Loss. Table 5 studies the effect of the modulation exponent p in GML. Increasing p from 1.00 to 1.50 improves mAP_{50} from 71.1% to 74.0%, whereas a further increase to $p = 2.00$ reduces the performance to 73.3%. The best result is obtained at $p = 1.50$, suggesting that moderate modulation provides a better balance between large-residual correction and stable optimization.

Table 3: Geometric representation ablations.

Setting	mAP_{50} (%)	Improve (%)
Baseline	69.3	–
GWD	71.3	↑ 2.0
CGE (Ours)	**72.5**	↑ **3.2**
CGE (SAM3-derived)	**75.8**	↑ **6.5**

Table 4: FCGF gating-term ablations.

$B(x)$	$C(x)$	mAP_{50} (%)	Improve (%)
		68.9	–
✓		73.3	↑ 4.4
	✓	72.8	↑ 3.9
✓	✓	**74.3**	↑ **5.4**

Table 5: GML exponent ablations.

p	mAP_{50} (%)	Improve (%)
1.00	71.1	–
1.25	72.5	↑ 1.4
1.50	**74.0**	↑ **2.9**
2.00	73.3	↑ 2.2

4.4 Generalization Analysis

We further examine whether BRIDGE is specific to RGB–IR oriented detection. Results across sensing modalities and detection representations are reported in Table 6.

Table 6: Generalization of BRIDGE across sensing modalities and detection representations. Results are reported in mAP_{50} (%).

Dataset	Modalities	Representation	Baseline	BRIDGE (Ours)	Improve (%)
M4-SAR (Wang et al., 2025)	RGB–SAR	OBB	86.8	**87.7**	↑ 0.9
DroneVehicle (Sun et al., 2022)	RGB–IR	HBB	85.2	**85.6**	↑ 0.4

On M4-SAR, BRIDGE improves the corresponding RGB–SAR OBB baseline, demonstrating that the proposed geometric formulation remains effective under a substantially different sensing pair. We further evaluate BRIDGE under the HBB setting on DroneVehicle, where it also improves the corresponding baseline. These results indicate that BRIDGE is not restricted to either RGB–IR sensing or orientation-specific bounding-box regression.

5 Conclusion

In this work, we introduced BRIDGE, a continuous geometry framework that unifies geometric representation, cross-modal interaction, and optimization through SDFs. BRIDGE achieves 85.0% mAP_{50} on DroneVehicle and generalizes across sensing modalities and detection representations, demonstrating the effectiveness and flexibility of continuous geometry modeling. These results suggest that continuous spatial fields can serve as a practical alternative to predefined parametric geometry for multimodal and geometry-aware visual perception.

STATEMENT ON THE USE OF AI

We used AI tools to assist with language polishing and visual refinement of figures during manuscript preparation. AI tools also supported repetitive formatting tasks, including table layout and the typesetting of mathematical expressions in LaTeX.

REFERENCES

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. Sam 3: Segment anything with concepts. In *International Conference on Learning Representations*, 2026.

Chen Chen, Kangcheng Bin, Ting Hu, Jiahao Qi, Xingyue Liu, Tianpeng Liu, Zhen Liu, Yongxiang Liu, and Ping Zhong. Fusion meets diverse conditions: A high-diversity benchmark and baseline for UAV-based multimodal object detection with condition cues. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 27958–27967, October 2025a.

Xiangqi Chen, Shan Jin, Li Zhao, Chengzhuan Yang, Dawei Zhang, Xiao Wang, Xiaowei He, Hua Wang, Zhongyu Chen, and Zhonglong Zheng. Mask-guided frequency feature fusion for visible–infrared remote sensing object detection. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–15, 2025b. doi: 10.1109/TGRS.2025.3612495.

Xiangqi Chen, Dawei Zhang, Li Zhao, Chengzhuan Yang, Zhongyu Chen, Jungang Lou, Zhonglong Zheng, Sang-Woon Jeon, and Hua Wang. Igianet: Illumination guided implicit alignment network for infrared–visible uav detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(4):3056–3064, 2026. doi: 10.1609/aaai.v40i4.37298.

Zhenshuai Chen, Wei Xiang, Zhiyuan Lin, Kaixuan Yang, Yunpeng Liu, and Zelin Shi. Alignment-assisted frequency fusion network for rgb-infrared vehicle detection. *Neurocomputing*, 647: 130505, 2025c. doi: 10.1016/j.neucom.2025.130505.

Minghao Cui, Jing Nie, Hanqing Sun, Jin Xie, Jiale Cao, Yanwei Pang, and Xuelong Li. Multispectral remote sensing object detection via selective cross-modal interaction and aggregation. *Neural Networks*, 197:108533, 2026. doi: 10.1016/j.neunet.2025.108533.

Jian Ding, Nan Xue, Yang Long, Gui-Song Xia, and Qikai Lu. Learning RoI transformer for oriented object detection in aerial images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2849–2858, 2019.

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

Qingyun Fang, Dapeng Han, and Zhaokui Wang. Cross-modality fusion transformer for multispectral object detection. *arXiv preprint arXiv:2111.00273*, 2021.

Changyu Gu, Linwei Chen, Lin Gu, and Ying Fu. Fourier angle alignment for oriented object detection in remote sensing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 42225–42235, June 2026.

Jiaming Han, Jian Ding, Jie Li, and Gui-Song Xia. Align deep features for oriented object detection. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–11, 2022. doi: 10.1109/TGRS.2021.3062048.

Xiaohui Hao, Yanglin Pu, Yongjun Wang, and Rui She. Distribution-aligned multimodal fusion for robust object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25494–25503, June 2026.

Liping Hou, Ke Lu, Xue Yang, Yuqiu Li, and Jian Xue. G-Rep: Gaussian representation for arbitrary-oriented object detection. *Remote Sensing*, 15(3):757, 2023. doi: 10.3390/rs15030757.

Guyue Jin, Tianming Zhao, Jiacan Yan, and Tian Tian. Contextually-guided state space fusion for misaligned multi-spectral object detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 2526–2535, 2025. doi: 10.1145/3746027.3754550.

Xin Lan, Shaolin Zhang, Yuhao Bai, and Xiaolin Qin. Fine-grained multispectral fusion for oriented object detection in remote sensing. *Remote Sensing*, 17(22):3769, 2025. doi: 10.3390/rs17223769.

Jianfeng Li, Siyu Cheng, Xiang Li, Longqing Tu, Yingjie Mei, Guangjiao Zhou, and Chenxu Wang. Cmtdet: Complementary mask-enhanced transformer for multimodal remote sensing object detection. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 19: 9620–9635, 2026. doi: 10.1109/JSTARS.2026.3669610.

Ting Li, Songtao Li, Shuaifeng Li, Xiaolin Qin, Maoyuan Zhao, Luping Ji, and Mao Ye. Sam-guided semantic knowledge fusion for visible-infrared object detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 8835–8844, 2025a. doi: 10.1145/3746027.3755718.

Wei Li, Jiaman Tang, Yang Li, Beihao Xia, Ligang Tan, and Hongmao Qin. Uavd-mamba: Deformable token fusion vision mamba for multimodal uav detection. In *2025 IEEE Intelligent Vehicles Symposium (IV)*, pp. 561–568, 2025b. doi: 10.1109/IV64158.2025.11097442.

Wentong Li, Yijie Chen, Kaixuan Hu, and Jianke Zhu. Oriented RepPoints for aerial object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1829–1838, 2022.

Jinyu Liu, Ming Li, Yuli Sun, Lin Lei, and Gangyao Kuang. Isgm: An illumination-aware semantic-guided mamba network for rgb-infrared vehicle detection in uav imagery. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2026. doi: 10.1109/JSTARS.2026.3723465.

Kewei Liu, Dongliang Peng, and Tao Li. Multimodal remote sensing object detection based on prior-enhanced mixture-of-experts fusion network. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–14, 2025. doi: 10.1109/TGRS.2025.3585634.

José Henrique Lima Marques, Jeffri Murrugarra-Llerena, and Claudio R. Jung. GauCho: Gaussian distributions with cholesky decomposition for oriented object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3593–3602, June 2025.

Herong Qi, Xuanyu Xiang, Hui Qin, Yuan Tai, and Yihua Tan. Cross-modal transformer fusion via local sampling for drone rgb-infrared object detection. *Neural Networks*, 202:108997, 2026. doi: 10.1016/j.neunet.2026.108997.

Jiashuo Shen, Jun He, Qiuyu Liu, Zhilong Zhang, Guoyan Wang, and Dawei Lu. Msdf-mamba: Mutual-spectrum perception deformable fusion mamba for drone-based visible–infrared cross-modality vehicle detection. *Remote Sensing*, 17(24):4037, 2025. doi: 10.3390/rs17244037.

Jifeng Shen, Yifei Chen, Yue Liu, Xin Zuo, Heng Fan, and Wankou Yang. ICAFusion: Iterative cross-attention guided feature fusion for multispectral object detection. *Pattern Recognition*, 145: 109913, 2024.

Yutian Shi, Guoquan Li, Zhilong Shen, Hongying Meng, and Yu Pang. Banet: Enhancing weakly aligned multimodal object detection via balanced bidirectional alignment network. *IEEE Transactions on Geoscience and Remote Sensing*, 64:1–13, 2026. doi: 10.1109/TGRS.2026.3674946.

Hui Sun, Long Lv, Pingping Zhang, Tongdan Tang, Feng Tian, Weibing Sun, and Huchuan Lu. Spatial-frequency enhanced mamba for multi-modal image fusion. *IEEE Transactions on Image Processing*, 34:7684–7696, 2025. doi: 10.1109/TIP.2025.3632221.

Yiming Sun, Bing Cao, Pengfei Zhu, and Qinghua Hu. Drone-based RGB-Infrared cross-modality vehicle detection via uncertainty-aware learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(10):6700–6713, 2022.

Chao Wang, Wei Lu, Xiang Li, Jian Yang, and Lei Luo. M4-sar: A multi-resolution, multi-polarization, multi-scene, multi-source dataset and benchmark for optical-sar fusion object detection. *arXiv preprint arXiv:2505.10931*, 2025.

Keyao Wang, Shuai Liu, Hengda Shi, Lukui Shi, and Haiyong Chen. Beyond duality: A hybrid framework of leveraging shared and private features for RGB-event object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4415–4424, June 2026.

Xin Wen, Haixu Yin, Kai Li, Wanying Nie, Jianxun Zhao, and Kechen Song. CMAI-Det: Cross-modal alignment and interaction for rgb-t object detection in drone scenes. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–11, 2025. doi: 10.1109/TGRS.2025.3645820.

Qian Weng, Yu Zhang, Xiansheng Huang, Liming Deng, and Jiawen Lin. Deaf-net: Dual-domain enhanced adaptive fusion network for uav visible–infrared object detection. *Remote Sensing*, 18 (13):2241, 2026. doi: 10.3390/rs18132241.

Tianhang Weng and Xiaopeng Niu. Lmdenet: A lightweight rgb-ir object detection network for low-light remote sensing images. *Sensors*, 26(4):1130, 2026. doi: 10.3390/s26041130.

Wencong Wu, Hongxi Zhang, Xiuwei Zhang, Hanlin Yin, and Yanning Zhang. Lightweight modal-guided cross-attention fusion network for visible-infrared object detection. *Pattern Recognition*, 177:113350, 2026. doi: 10.1016/j.patcog.2026.113350.

Zikai Xiao, Guoye Yang, Xue Yang, Taijiang Mu, Junchi Yan, and Shimin Hu. Theoretically achieving continuous representation of oriented bounding boxes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16912–16922, 2024.

Xingxing Xie, Gong Cheng, Jiabao Wang, Xiwen Yao, and Junwei Han. Oriented R-CNN for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3520–3529, 2021.

Yongchao Xu, Mingtao Fu, Qimeng Wang, Yukang Wang, Kai Chen, Gui-Song Xia, and Xiang Bai. Gliding vertex on the horizontal bounding box for multi-oriented object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(4):1452–1459, 2021. doi: 10.1109/TPAMI.2020.2974745.

Haitian Yang, Juan Fang, Yiren Zhu, Xudong Zhao, Yufei Guo, Xiaohan Zhang, Xiaoxing Hu, Xue Yang, and Qi Ming. CrossWeaver: Towards efficient cross-modal interweaving and decoupling for weakly-aligned multispectral object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6361–6370, 2026.

Xue Yang, Liping Hou, Yue Zhou, Wentao Wang, and Junchi Yan. Dense label encoding for boundary discontinuity free rotation detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15819–15829, 2021a.

Xue Yang, Junchi Yan, Ziming Feng, and Tao He. R3Det: Refined single-stage detector with feature refinement for rotating object. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35 (4):3163–3171, 2021b.

Xue Yang, Junchi Yan, Qi Ming, Wentao Wang, Xiaopeng Zhang, and Qi Tian. Rethinking rotated object detection with Gaussian Wasserstein Distance loss. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 11830–11841. PMLR, 2021c.

Xue Yang, Xiaojiang Yang, Jirui Yang, Qi Ming, Wentao Wang, Qi Tian, and Junchi Yan. Learning high-precision bounding box for rotated object detection via Kullback-Leibler divergence. In *Advances in Neural Information Processing Systems*, volume 34, pp. 18381–18394, 2021d.

Xue Yang, Yue Zhou, Gefan Zhang, Jirui Yang, Wentao Wang, Junchi Yan, Xiaopeng Zhang, and Qi Tian. The kfiou loss for rotated object detection. In *International Conference on Learning Representations*, 2023.

Yanqing Yao, Gong Cheng, Chunbo Lang, Xiang Yuan, Xingxing Xie, and Junwei Han. Hierarchical mask prompting and robust integrated regression for oriented object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(12):13071–13084, 2024. doi: 10.1109/TCSVT.2024.3444795.

Chushi Yu and Yoan Shin. A cross-modality feature adaptive interaction approach for rgb-infrared object detection in aerial imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 64: 1–15, 2026. doi: 10.1109/TGRS.2026.3657379.

Maoxun Yuan and Xingxing Wei. C^2Former: Calibrated and complementary transformer for rgb-infrared object detection. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–12, 2024. doi: 10.1109/TGRS.2024.3376819.

Maoxun Yuan, Xiaorong Shi, Nan Wang, Yinyan Wang, and Xingxing Wei. Improving RGB-Infrared object detection with cascade alignment-guided transformer. *Information Fusion*, 105: 102246, 2024.

Jialei Zhan, Li Liu, Jiehua Zhang, Yuhang Xie, Yongxiang Liu, Jiangming Chen, and Ming-Ming Cheng. Rotation invariant and symmetry aware pixel difference network for remote sensing object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13264–13274, June 2026.

Jiaqi Zhao, Zeyu Ding, Yong Zhou, Hancheng Zhu, Wen-Liang Du, Rui Yao, and Abdulmotaleb El Saddik. OrientedFormer: An end-to-end transformer-based oriented object detector in remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–16, 2024.

Jiaqi Zhao, Zeyu Ding, Yong Zhou, Hancheng Zhu, Wen-Liang Du, and Rui Yao. ReDiffDet: Rotation-equivariant diffusion model for oriented object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24429–24439, June 2025a.

Tianyi Zhao, Boyang Liu, Yanglei Gao, Yiming Sun, Maoxun Yuan, and Xingxing Wei. Rethinking multi-modal object detection from the perspective of mono-modality feature learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6364–6373, October 2025b.

Zhicheng Zhao, Wei Zhang, Yun Xiao, Chenglong Li, and Jin Tang. Reflectance-guided progressive feature alignment network for all-day uav object detection. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–15, 2025c. doi: 10.1109/TGRS.2025.3574963.

Liming Zhou, Yuanfei Xie, Jiakang Yang, Guochong Zhang, Yi Xie, Yadi Wang, and Xianyu Zuo. Dfdnet: Differential fusion detection network for rgb-ir objects. *Expert Systems with Applications*, 324:132494, 2026. doi: 10.1016/j.eswa.2026.132494.

Minghang Zhou, Tianyu Li, Chaofan Qiao, Dongyu Xie, Guoqing Wang, Ningjuan Ruan, Lin Mei, Yang Yang, and Heng Tao Shen. Dmm: Disparity-guided multispectral mamba for oriented object detection in remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–13, 2025. doi: 10.1109/TGRS.2025.3578309.

Haodong Zhu, Wenhao Dong, Linlin Yang, Hong Li, Yuguang Yang, Yangyang Ren, Qingcheng Zhu, Zichao Feng, Changbai Li, Shaohui Lin, Runqi Wang, Xiaoyan Luo, and Baochang Zhang. WaveMamba: Wavelet-driven mamba fusion for RGB-Infrared object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11219–11229, October 2025.

Xin Zuo, Chenyu Qu, Haibo Zhan, Jifeng Shen, and Wankou Yang. Sffr: Spatial–frequency feature reconstruction for multispectral aerial object detection. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–11, 2025. doi: 10.1109/TGRS.2025.3631708.

2027 ICLR Appendix

Contents

A Theoretical Analysis of Continuous Geometry Encoding **2**

A.1 Detailed Construction of Box-Induced Continuous Geometry 2

A.2 Local Jacobian of Box-Induced SDF 2

A.3 Bounded Perturbation of Continuous Geometry 3

A.4 Bounded Gradient Perturbation 4

A.5 Comparison with Direct Box Regression 4

A.6 Extension to Fine-Grained SAM3-Derived Geometry 5

B Derivation and Properties of GML **5**

B.1 Geometric Interpretation of Continuous-Field Residuals 5

B.2 From Gradient Modulation to GML 6

B.3 Properties of the Gradient-Modulated Response 7

C Additional Analysis of GML **8**

C.1 Residual Distribution and Design Motivation 8

C.2 Core GML Design Ablations 8

C.3 Residual Redistribution and Gradient Allocation 8

C.4 Exploratory Optimization Alternatives 9

D Additional Analysis of FCGF **10**

D.1 Fourier Quality Estimation 10

D.2 Modality-Wise Score Calibration 10

D.3 Boundary and Consistency Gates 11

D.4 Residual Fusion Mechanism 12

E Implementation Details **13**

E.1 SDF Construction and Normalization 13

E.2 Fourier Preprocessing and Training-Set Statistics 13

E.3 FCGF Hyperparameters 13

E.4 GML and Loss Weights 13

E.5 SAM3-Derived Supervision Generation 14

F Additional Generalization Analysis **14**

F.1 RGB–SAR Generalization 14

F.2 HBB Generalization 14

A Theoretical Analysis of Continuous Geometry Encoding

This section provides the detailed derivations underlying CGE. We first derive the continuous geometric field induced by an oriented bounding box and analyze its local Jacobian with respect to box parameters. Based on this formulation, we establish the representation- and gradient-stability properties discussed in the main paper, followed by a comparison with direct parametric box regression and an extension to fine-grained geometry derived from SAM3 masks.

A.1 Detailed Construction of Box-Induced Continuous Geometry

CGE maps each oriented box to a continuous geometric field represented by an SDF. Consider a predicted oriented bounding box

$$B^p = (x_p, y_p, w_p, h_p, \theta_p), \tag{A.1}$$

where (x_p, y_p) denotes the predicted box center, w_p and h_p denote its width and height, and θ_p denotes its orientation angle. Its center and half-size are respectively

$$\mathbf{c}_p = \begin{bmatrix} x_p \\ y_p \end{bmatrix}, \qquad \mathbf{a}_p = \begin{bmatrix} w_p/2 \\ h_p/2 \end{bmatrix}. \tag{A.2}$$

The corresponding 2D rotation matrix is

$$R(\theta_p) = \begin{bmatrix} \cos\theta_p & -\sin\theta_p \\ \sin\theta_p & \cos\theta_p \end{bmatrix}. \tag{A.3}$$

For the i-th spatial sample $\mathbf{q}_i$, we transform it into the local coordinate system of the predicted box:

$$\boldsymbol{\xi}_i = R(\theta_p)^\top (\mathbf{q}_i - \mathbf{c}_p). \tag{A.4}$$

Let

$$\mathbf{u}_i = |\boldsymbol{\xi}_i| - \mathbf{a}_p, \tag{A.5}$$

where the absolute value is applied element-wise. For an axis-aligned rectangle, the signed distance function is

$$\Psi(\mathbf{u}) = \|[\mathbf{u}]_+\|_2 + \min(\max(u_x, u_y), 0)\,, \tag{A.6}$$

where $\mathbf{u} = (u_x, u_y)^\top$ and

$$[\mathbf{u}]_+ = \begin{bmatrix} \max(u_x, 0) \\ \max(u_y, 0) \end{bmatrix}. \tag{A.7}$$

Substituting equation A.4–equation A.5 into equation A.6, the oriented box induces the continuous geometric field

$$\phi^p(\mathbf{q}_i) = \Psi\big(\big|R(\theta_p)^\top (\mathbf{q}_i - \mathbf{c}_p)\big| - \mathbf{a}_p\big)\,. \tag{A.8}$$

Thus, the parametric prediction admits the explicit mapping

$$B^p \longmapsto \phi_{B^p}. \tag{A.9}$$

A.2 Local Jacobian of Box-Induced SDF

We next derive how continuous geometric supervision propagates to the five oriented-box parameters. For analysis, consider the weighted squared SDF objective

$$\mathcal{L}_{\mathrm{SDF}} = \frac{1}{N} \sum_{i=1}^{N} w_i \left(\phi^p(\mathbf{q}_i) - \phi^t(\mathbf{q}_i)\right)^2, \tag{A.10}$$

where N is the number of sampled spatial locations, $w_i \geq 0$ is a spatial weight treated as constant with respect to the predicted box, and ϕ^p and ϕ^t are the predicted and target fields.

Define

$$e_i = \phi^p(\mathbf{q}_i) - \phi^t(\mathbf{q}_i), \qquad \lambda_i = \frac{2}{N} w_i e_i. \tag{A.11}$$

For any box parameter $\alpha \in \{x_p, y_p, w_p, h_p, \theta_p\}$, the chain rule gives

$$\frac{\partial \mathcal{L}_{\text{SDF}}}{\partial \alpha} = \sum_{i=1}^{N} \lambda_i \frac{\partial \phi^p(\mathbf{q}_i)}{\partial \alpha}. \tag{A.12}$$

To derive the local SDF Jacobian, define

$$\mathbf{m}_i = \nabla_{\mathbf{u}_i} \Psi(\mathbf{u}_i), \qquad S_i = \operatorname{diag}(\operatorname{sgn}(\boldsymbol{\xi}_i)), \tag{A.13}$$

where $\mathbf{m}_i = (m_{x,i}, m_{y,i})^\top$. At non-differentiable locations, any valid subgradient can be used. Since $\mathbf{u}_i = |\boldsymbol{\xi}_i| - \mathbf{a}_p$,

$$\frac{\partial \phi^p(\mathbf{q}_i)}{\partial \boldsymbol{\xi}_i} = S_i \mathbf{m}_i. \tag{A.14}$$

For convenience, denote $\mathbf{n}_i = S_i \mathbf{m}_i$.

Center parameters. From equation A.4,

$$\frac{\partial \boldsymbol{\xi}_i}{\partial x_p} = -R(\theta_p)^\top \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \qquad \frac{\partial \boldsymbol{\xi}_i}{\partial y_p} = -R(\theta_p)^\top \begin{bmatrix} 0 \\ 1 \end{bmatrix}. \tag{A.15}$$

Hence,

$$\frac{\partial \phi^p(\mathbf{q}_i)}{\partial x_p} = \mathbf{n}_i^\top \frac{\partial \boldsymbol{\xi}_i}{\partial x_p}, \qquad \frac{\partial \phi^p(\mathbf{q}_i)}{\partial y_p} = \mathbf{n}_i^\top \frac{\partial \boldsymbol{\xi}_i}{\partial y_p}. \tag{A.16}$$

Scale parameters. Since

$$\frac{\partial \mathbf{u}_i}{\partial w_p} = \begin{bmatrix} -1/2 \\ 0 \end{bmatrix}, \qquad \frac{\partial \mathbf{u}_i}{\partial h_p} = \begin{bmatrix} 0 \\ -1/2 \end{bmatrix}, \tag{A.17}$$

we obtain

$$\frac{\partial \phi^p(\mathbf{q}_i)}{\partial w_p} = -\frac{1}{2} m_{x,i}, \qquad \frac{\partial \phi^p(\mathbf{q}_i)}{\partial h_p} = -\frac{1}{2} m_{y,i}. \tag{A.18}$$

Orientation parameter. For θ_p,

$$\frac{\partial \boldsymbol{\xi}_i}{\partial \theta_p} = \frac{\partial R(\theta_p)^\top}{\partial \theta_p} (\mathbf{q}_i - \mathbf{c}_p), \tag{A.19}$$

where

$$\frac{\partial R(\theta_p)^\top}{\partial \theta_p} = \begin{bmatrix} -\sin\theta_p & \cos\theta_p \\ -\cos\theta_p & -\sin\theta_p \end{bmatrix}. \tag{A.20}$$

Therefore,

$$\frac{\partial \phi^p(\mathbf{q}_i)}{\partial \theta_p} = \mathbf{n}_i^\top \frac{\partial \boldsymbol{\xi}_i}{\partial \theta_p}. \tag{A.21}$$

If the box prediction is produced by network parameters Θ, the complete backward path becomes

$$\frac{\partial \mathcal{L}_{\text{SDF}}}{\partial \Theta} = \sum_{\alpha} \left[\sum_{i=1}^{N} \lambda_i \frac{\partial \phi^p(\mathbf{q}_i)}{\partial \alpha} \right] \frac{\partial \alpha}{\partial \Theta}. \tag{A.22}$$

The resulting differentiable path is $\Theta \to B^p \to \phi_{B^p} \to \mathcal{L}_{\text{SDF}}$.

A.3 Bounded Perturbation of Continuous Geometry

Consider a reference field $\phi_t(\mathbf{q})$ and another observation translated by a spatial displacement $\boldsymbol{\delta}$. Its SDF is

$$\phi_m(\mathbf{q}) = \phi_t(\mathbf{q} - \boldsymbol{\delta}). \tag{A.23}$$

An SDF is 1-Lipschitz with respect to spatial coordinates:

$$|\phi_t(\mathbf{q}_1) - \phi_t(\mathbf{q}_2)| \le \|\mathbf{q}_1 - \mathbf{q}_2\|_2. \tag{A.24}$$

Setting $\mathbf{q}_1 = \mathbf{q} - \boldsymbol{\delta}$ and $\mathbf{q}_2 = \mathbf{q}$ gives

$$|\phi_m(\mathbf{q}) - \phi_t(\mathbf{q})| \leq \|\boldsymbol{\delta}\|_2. \tag{A.25}$$

Thus, displacement in the image plane produces a bounded change in the continuous geometric representation.

For a truncation threshold $T > 0$, define

$$\phi^T(\mathbf{q}) = \mathrm{clip}\left(\phi(\mathbf{q}), -T, T\right). \tag{A.26}$$

Since clipping is 1-Lipschitz and both truncated fields lie in $[-T, T]$,

$$|\phi_m^T(\mathbf{q}) - \phi_t^T(\mathbf{q})| \leq \min\left(\|\boldsymbol{\delta}\|_2, 2T\right). \tag{A.27}$$

For the normalized truncated field $\widetilde{\phi}^T(\mathbf{q}) = \phi^T(\mathbf{q})/T$,

$$|\widetilde{\phi}_m^T(\mathbf{q}) - \widetilde{\phi}_t^T(\mathbf{q})| \leq \min\left(\frac{\|\boldsymbol{\delta}\|_2}{T}, 2\right). \tag{A.28}$$

A.4 Bounded Gradient Perturbation

For a box parameter α, let g_α^t and g_α^m denote the gradients induced by the reference and shifted geometric targets, respectively:

$$g_\alpha^t = \frac{2}{N}\sum_{i=1}^{N} w_i \left(\phi^p(\mathbf{q}_i) - \phi_t(\mathbf{q}_i)\right) \frac{\partial \phi^p(\mathbf{q}_i)}{\partial \alpha}, \tag{A.29}$$

$$g_\alpha^m = \frac{2}{N}\sum_{i=1}^{N} w_i \left(\phi^p(\mathbf{q}_i) - \phi_m(\mathbf{q}_i)\right) \frac{\partial \phi^p(\mathbf{q}_i)}{\partial \alpha}. \tag{A.30}$$

Subtracting them and applying the triangle inequality gives

$$|g_\alpha^m - g_\alpha^t| \leq \frac{2}{N}\sum_{i=1}^{N} w_i |\phi_t(\mathbf{q}_i) - \phi_m(\mathbf{q}_i)| \left|\frac{\partial \phi^p(\mathbf{q}_i)}{\partial \alpha}\right|. \tag{A.31}$$

Define

$$C_\alpha = \frac{2}{N}\sum_{i=1}^{N} w_i \left|\frac{\partial \phi^p(\mathbf{q}_i)}{\partial \alpha}\right|, \tag{A.32}$$

where C_α collects the spatial sampling weights and local geometric Jacobian. By equation A.25,

$$|g_\alpha^m - g_\alpha^t| \leq C_\alpha \|\boldsymbol{\delta}\|_2. \tag{A.33}$$

For normalized truncated SDFs,

$$|g_\alpha^m - g_\alpha^t| \leq C_\alpha \min\left(\frac{\|\boldsymbol{\delta}\|_2}{T}, 2\right). \tag{A.34}$$

Hence, spatial perturbations induce bounded changes in the continuous geometric optimization signal.

A.5 Comparison with Direct Box Regression

Consider two center annotations $\mathbf{c}_t^1$ and $\mathbf{c}_t^2$ separated by $\boldsymbol{\delta} \neq 0$:

$$\mathbf{c}_t^1 = \mathbf{c}_t, \qquad \mathbf{c}_t^2 = \mathbf{c}_t + \boldsymbol{\delta}. \tag{A.35}$$

Using squared center regression,

$$\mathcal{L}_1 = \|\mathbf{c}_p - \mathbf{c}_t^1\|_2^2, \qquad \mathcal{L}_2 = \|\mathbf{c}_p - \mathbf{c}_t^2\|_2^2, \tag{A.36}$$

with gradients

$$\mathbf{g}_1 = 2(\mathbf{c}_p - \mathbf{c}_t), \qquad \mathbf{g}_2 = 2(\mathbf{c}_p - \mathbf{c}_t - \boldsymbol{\delta}). \tag{A.37}$$

If the predicted center lies between the two annotations, $\mathbf{c}_p = \mathbf{c}_t + s\boldsymbol{\delta}$ with $0 < s < 1$, then

$$\mathbf{g}_1^\top \mathbf{g}_2 = -4s(1-s)\|\boldsymbol{\delta}\|_2^2 < 0. \tag{A.38}$$

Hence, inconsistent parametric annotations can induce directly opposing optimization directions. CGE does not imply that all such conflicts disappear; instead, it decomposes geometric optimization into spatially distributed residual contributions that can be weighted or modulated locally.

Visualization of geometric optimization. Figure A.1 shows the evolution of the geometric contours in the same scene at epochs 1, 10, and 100. Early contours cover only parts of some vehicles and include responses on the background. At later epochs, the contours more closely follow the visible vehicle boundaries and the isolated background responses diminish. This example illustrates the progressive spatial refinement of the learned geometry during training.

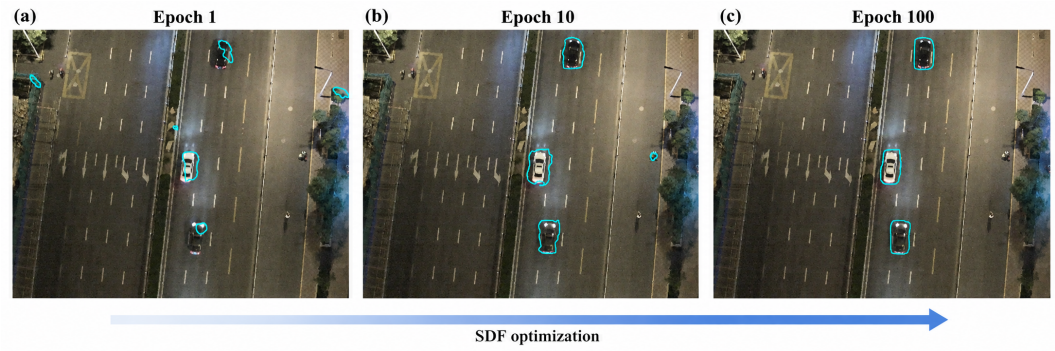

Figure A.1: Qualitative visualization of SDF optimization during training. Cyan curves show the geometric contours overlaid on the same scene at (a) epoch 1, (b) epoch 10, and (c) epoch 100. The contours become more complete around vehicles as training progresses, while isolated background responses diminish.

A.6 Extension to Fine-Grained SAM3-Derived Geometry

CGE can encode an irregular object region without changing the continuous-field representation. As described in Section 3.2, a ground-truth box B^t provides a prompt for SAM3 (Carion et al., 2026), yielding an object mask $\mathcal{M}^t$. The supervision-generation pipeline is

$$B^t \longrightarrow \mathcal{M}^t \longrightarrow \phi_{\mathcal{M}^t}, \qquad \phi_{\mathcal{M}^t}(\mathbf{q}) = \operatorname{sgn}_{\mathcal{M}^t}(\mathbf{q}) d(\mathbf{q}, \partial\mathcal{M}^t). \tag{A.39}$$

The sign is negative inside the mask and positive outside. Its zero level set follows the mask contour, whereas the box-induced field follows the rectangular annotation. Both targets provide a scalar geometric value at each sampled location, allowing the same residual formulation to accommodate different boundary shapes.

Table 3 reports 69.3 mAP_{50} for the baseline, 71.3 for GWD, 72.5 for box-induced CGE, and 75.8 for SAM3-derived CGE. The mask-derived target improves over box-induced CGE by 3.3 percentage points in this experiment. This result supports the usefulness of richer contour supervision within the continuous-field formulation; it does not establish that generated masks are uniformly accurate. The quality of the resulting distance field remains dependent on the mask boundary. SAM3 is used offline to generate supervision and adds no inference computation to the detector.

The local box-parameter Jacobians in Appendix A.2 apply specifically to box-induced fields. The mask extension changes the target geometry and should not be interpreted as differentiating through SAM3 or assigning box-parameter Jacobians to arbitrary masks.

B Derivation and Properties of GML

B.1 Geometric Interpretation of Continuous-Field Residuals

A signed distance field satisfies the Eikonal property

$$\|\nabla_{\mathbf{q}}\phi(\mathbf{q})\|_2 = 1 \tag{B.1}$$

almost everywhere away from non-differentiable medial-axis locations. For a small displacement $\Delta\mathbf{q}$, a first-order expansion gives

$$\phi(\mathbf{q}+\Delta\mathbf{q}) - \phi(\mathbf{q}) = \nabla_{\mathbf{q}}\phi(\mathbf{q})^\top \Delta\mathbf{q} + \mathcal{O}\left(\|\Delta\mathbf{q}\|_2^2\right). \tag{B.2}$$

When the displacement follows the local unit normal

$$\mathbf{n} = \frac{\nabla \phi}{\|\nabla \phi\|_2}, \qquad \Delta \mathbf{q} = \Delta d\, \mathbf{n}, \tag{B.3}$$

the corresponding field variation satisfies

$$|\Delta \phi| = |\Delta d| + \mathcal{O}\left(|\Delta d|^2\right), \tag{B.4}$$

where $\Delta \phi = \phi(\mathbf{q} + \Delta \mathbf{q}) - \phi(\mathbf{q})$. Hence, the residual magnitude locally reflects the severity of geometric displacement.

B.2 From Gradient Modulation to GML

Consider a generic point-wise continuous-field objective

$$\mathcal{L}_\rho = \frac{1}{N} \sum_{i=1}^{N} \rho(e_i), \qquad e_i = \widehat{\phi}_i - \phi_i, \tag{B.5}$$

where $\widehat{\phi}_i$ and ϕ_i denote the predicted and target geometric values at sample i. For an arbitrary network or geometric parameter α,

$$\frac{\partial \mathcal{L}_\rho}{\partial \alpha} = \frac{1}{N} \sum_{i=1}^{N} \rho'(e_i) \frac{\partial \widehat{\phi}_i}{\partial \alpha}. \tag{B.6}$$

Therefore, the influence function

$$\psi(e) = \rho'(e) \tag{B.7}$$

directly controls the optimization contribution of each spatial geometric residual.

For ℓ_1 regression,

$$\rho_{\ell_1}(e) = |e|, \qquad |\rho'_{\ell_1}(e)| = 1, \qquad e \neq 0, \tag{B.8}$$

so all non-zero residuals receive the same gradient magnitude, regardless of their geometric severity.

Let $r = |e|$ denote the residual magnitude and $p \in (1, 2)$ the modulation exponent. We define the gradient-modulation function as

$$\psi_p(r) = \begin{cases} r, & r \leq 1, \\ r^{p-1}, & r > 1. \end{cases} \tag{B.9}$$

For $r \leq 1$, integrating $\psi_p(r) = r$ gives

$$\rho_p(r) = \frac{1}{2} r^2. \tag{B.10}$$

For $r > 1$, continuity at $r = 1$ gives

$$\begin{aligned} \rho_p(r) &= \rho_p(1) + \int_1^r s^{p-1}\, ds \\ &= \frac{1}{2} + \frac{r^p - 1}{p}. \end{aligned} \tag{B.11}$$

Thus,

$$\rho_p(e) = \begin{cases} \dfrac{1}{2} r^2, & r \leq 1, \\ \dfrac{1}{2} + \dfrac{r^p - 1}{p}, & r > 1, \end{cases} \qquad r = |e|. \tag{B.12}$$

Differentiating with respect to e yields

$$\frac{\partial \rho_p}{\partial e} = \operatorname{sgn}(e) \begin{cases} r, & r \leq 1, \\ r^{p-1}, & r > 1. \end{cases} \tag{B.13}$$

B.3 Properties of the Gradient-Modulated Response

For $1 < p < 2$, the exponent satisfies

$$0 < p - 1 < 1. \tag{B.14}$$

For $r > 1$,

$$\frac{d}{dr} r^{p-1} = (p-1) r^{p-2} > 0, \tag{B.15}$$

so the gradient magnitude remains strictly increasing with residual severity. Since $p - 2 < 0$,

$$\frac{d}{dr} r^{p-1} \to 0 \qquad \text{as } r \to \infty, \tag{B.16}$$

which yields a sub-linear modulation response.

At the transition point,

$$\rho_p(1^-) = \rho_p(1^+) = \frac{1}{2}, \qquad \psi_p(1^-) = \psi_p(1^+) = 1, \tag{B.17}$$

so ρ_p is continuously differentiable at $r = 1$.

The final GML uses $p = 3/2$:

$$\rho_{\mathrm{GML}}(e) = \begin{cases} \dfrac{1}{2} r^2, & r \le 1, \\ \dfrac{1}{2} + \dfrac{r^{3/2} - 1}{3/2}, & r > 1, \end{cases} \qquad r = |e|, \tag{B.18}$$

with gradient magnitude

$$\left| \frac{\partial \rho_{\mathrm{GML}}}{\partial e} \right| = \begin{cases} r, & r \le 1, \\ \sqrt{r}, & r > 1. \end{cases} \tag{B.19}$$

GML therefore produces vanishing updates for well-fitted geometric regions while retaining an increasing, sub-linear response to severe geometric deviations.

Analytic comparison of residual responses. Figure B.1 plots the loss and gradient magnitude from the definitions above. All GML curves share the same small-residual response; their tail behavior differs with p. The $p = 2$ endpoint is included as the quadratic reference. These are analytic curves, not measured training trajectories.

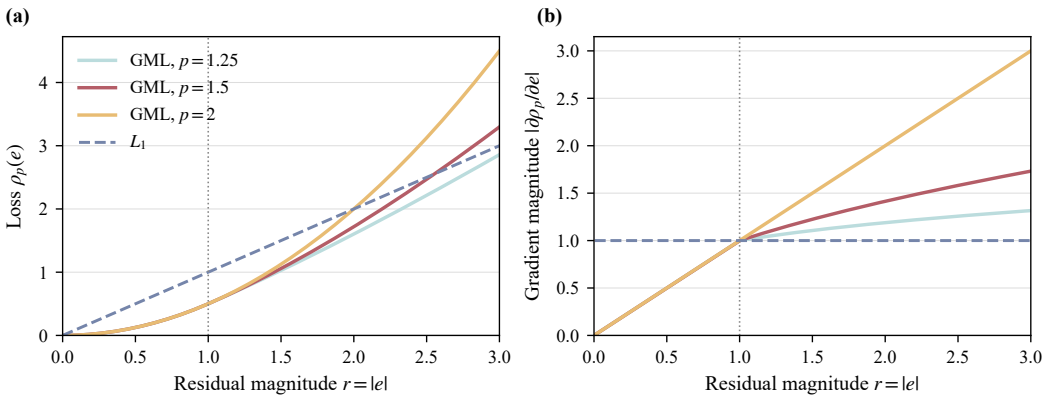

Figure B.1: Analytic GML responses compared with L1. (a) Loss value. (b) Gradient magnitude for positive residual magnitude. The vertical dotted line marks $r = 1$. For $p = 1.5$, the tail response grows as $\sqrt{r}$, between constant L1 influence and linear growth.

C Additional Analysis of GML

C.1 Residual Distribution and Design Motivation

Before applying GML, the mean absolute residuals of the RGB, IR, and fused dense geometric branches are 0.180, 0.230, and 0.198, respectively. Most residuals lie in the well-fitted regime: only 2.72%, 2.74%, and 3.11% of RGB, IR, and fused residuals exceed $r = 1$, while the proportions above $r = 0.5$ are 6.69%, 12.39%, and 8.00%, respectively. Thus, roughly 97% of dense residuals remain below $r = 1$. This motivates a modulation strategy that de-emphasizes the well-fitted residual body while preserving magnitude sensitivity for the severe tail. Relative to L1, the residual influence is reduced from 1 to r for $0 < r < 1$, and increases as r^{p-1} for $r > 1$.

The modality-specific distributions also support continuous residual-dependent modulation. IR contains more medium residuals than RGB, while the fused branch contains the largest fraction above $r = 1$. A binary hard/easy partition would discard this structure, whereas GML retains a magnitude-sensitive gradient response.

C.2 Core GML Design Ablations

Table C.1 groups the experiments by four design questions: which residual range to emphasize, how strongly to modulate the tail, whether to clip extreme residuals, and how to allocate weights across dense branches. Values are reported on the 0–100 scale. Exploratory objectives are reported separately in Appendix C.4.

Table C.1: Core GML design ablations. Best mAP_{50} denotes the best value within each reported training run.

Objective	Main modification	Best mAP_{50} (%)
L1	Constant residual influence	72.53
Middle Boost	Enhance $0.5 < r < 1$	72.05
GML, $p = 1.25$	Weak tail modulation	72.50
Capped GML (Dense)	Clip dense tail only	70.57
Capped GML (Full)	Clip box + dense tail	71.99
Fuse-emphasis GML	$0.75/0.75/1.50$ branch weights	72.22
GML, $p = 1.5$	**Sub-linear gradient modulation**	**73.97**

Residual range and modulation strength. Middle Boost reaches 72.05 mAP_{50}, compared with 72.53 for L1, so emphasizing the larger middle-residual population does not improve this run. GML with $p = 1.25$ reaches 72.50, while $p = 1.5$ reaches 73.97. The larger exponent gives a stronger response for $r > 1$; these results favor $p = 1.5$ in this setting.

Extreme-tail clipping. Capping the dense tail gives 70.57 mAP_{50}, and capping both the box and dense tails gives 71.99, compared with 73.97 for uncapped GML. These results are consistent with severe geometric residuals retaining useful correction signals. They do not imply that every large residual is informative. Automatic-differentiation checks were also used to verify the capped-loss implementation and gradient sign.

Branch allocation. Assigning weights of 0.75, 0.75, and 1.50 to the RGB, IR, and fused dense branches, respectively, gives 72.22 mAP_{50} in the reported run. The final configuration retains equal coefficients for the three dense branches.

C.3 Residual Redistribution and Gradient Allocation

GML changes the learned residual distribution beyond the small severe-tail population. In the fused branch, the 90th-percentile residual decreases from 0.4314 to 0.3635, a relative reduction of about 15.8%, while the fraction above 0.5 decreases from 8.00% to 6.50%. In the IR branch, the 90th percentile decreases from 0.5775 to 0.4780 and the fraction above 0.5 decreases from 12.39% to 9.41%.

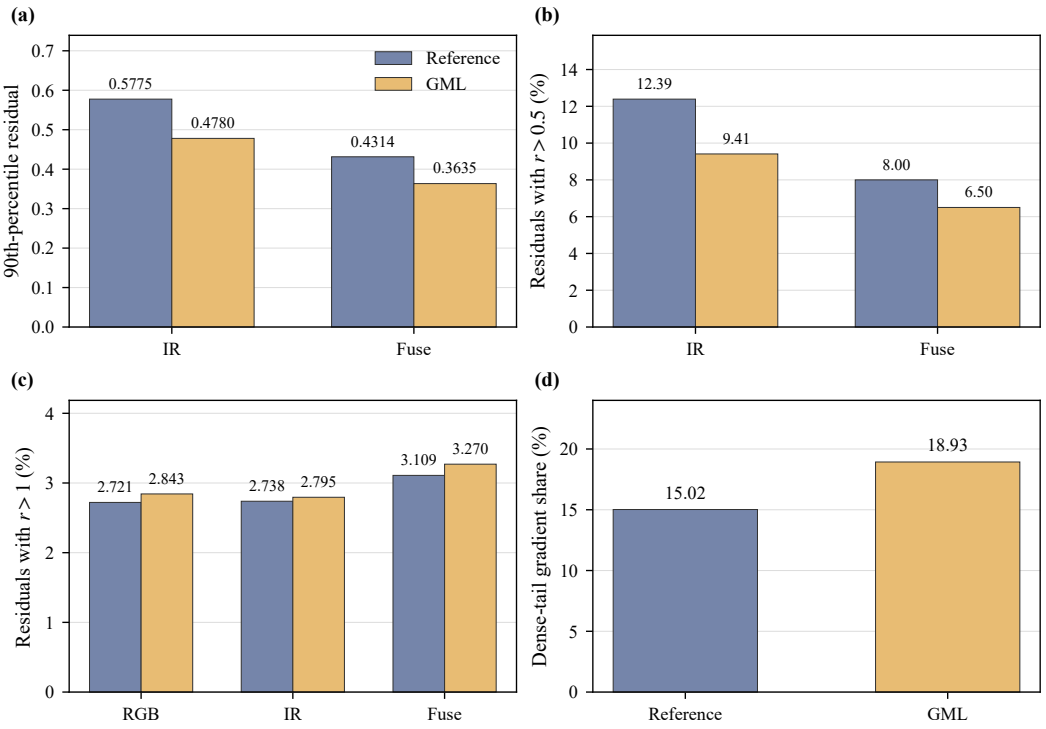

Figure C.1: Residual and gradient redistribution under GML. (a) 90th-percentile residual. (b) Fraction of residuals above 0.5. (c) Fraction above 1 across RGB, IR, and fused branches. (d) Share of dense-branch gradient magnitude contributed by residuals above 1. Bars reproduce the summary statistics reported in this section. Reference denotes the L1 geometric objective.

The number of the most extreme residuals does not decrease monotonically: the fraction above 1 changes from 2.721% to 2.843% for RGB, from 2.738% to 2.795% for IR, and from 3.109% to 3.270% for Fuse. Thus, the detection improvement is not accompanied by a lower count of extreme residuals in these measurements. This is consistent with the objective redistributing optimization influence rather than directly minimizing the number of residuals above a threshold.

The gradient-budget analysis gives the same interpretation. Residuals above 1 account for about 15.02% of the dense-branch gradient magnitude under L1 and 18.93% under GML, corresponding to an approximately 26% relative increase even though only about 3% of dense residuals lie in this region. In contrast, the corresponding extreme-tail contribution of the box-induced SDF branch is only about 0.06%. These observations support applying GML to the dense RGB, IR, and fused geometric branches while retaining L1 supervision for the box-induced SDF term.

These statistics show how GML redistributes gradient magnitude, although they do not isolate the cause of the detection gain. The final objective uses L1 for the box-induced geometric term and equally weighted GML terms for the RGB, IR, and fused dense branches.

Figure C.1 summarizes these measurements. Panels (a) and (b) show the IR and fused branches, for which both before-and-after statistics are available. Panels (c) and (d) distinguish the frequency of severe residuals from their share of the dense-branch gradient magnitude.

C.4 Exploratory Optimization Alternatives

Table C.2 summarizes exploratory SDF objectives that complement the core GML ablations.

Boundary/reliability weighting performs similarly to L1, while the tested sign-and-normal constraints, stage transition, and annealed tolerance variants yield lower mAP_{50}. These results are specific to the tested implementations and configurations.

Table C.2: Exploratory SDF optimization alternatives. The L1 reference is shared with Table C.1.

Objective	Main modification	Best mAP_{50} (%)
L1	Constant residual influence	72.53
RAB-SDF-Lite	Boundary/reliability weighting	72.49
Geometry Consistency	Sign + normal constraints	64.46
Stage Transition	Epoch-dependent response	68.26
Annealed Tolerance	Dynamic tolerance region	69.55

D Additional Analysis of FCGF

FCGF uses calibrated Fourier scores to select the main modality and spatial gates to control auxiliary corrections. The training and validation statistics below characterize the score calibration and fusion weights.

D.1 Fourier Quality Estimation

For modality $m \in \{\text{rgb}, \text{ir}\}$, we first convert the input to grayscale, remove its spatial mean, and compute the magnitude spectrum

$$A_m = |\mathcal{F}(I_m)| \,. \tag{D.1}$$

Let E_m^{mid} and E_m^{high} denote the average spectral magnitudes within the predefined radial mid- and high-frequency bands. The Fourier quality score used in our implementation is

$$q_m = \frac{\log\left(1 + E_m^{\text{mid}}\right)}{1 + \alpha \dfrac{E_m^{\text{high}}}{E_m^{\text{mid}} + \epsilon}}, \tag{D.2}$$

where α controls the penalty on excessive high-frequency responses and ϵ is a small numerical constant.

The score favors informative mid-frequency structure while suppressing cases in which the high-frequency response becomes disproportionately large. The score q_m is used only as a proxy for the structural quality of the input observation. It does not directly measure the error of the predicted SDF, since high-frequency responses may originate from either fine image structure or noise.

D.2 Modality-Wise Score Calibration

Direct comparison of the raw Fourier scores produces a clear modality-scale bias. On the training set, the raw scores select RGB as the dominant modality for only approximately 11.3% of the samples and IR for 88.7%. A similar distribution is observed on the validation set, where the corresponding proportions are 9.9% and 90.1%.

We therefore calibrate each modality independently using the training-set statistics of the log quality score:

$$z_m = \frac{\log q_m - \mu_m}{\sigma_m}. \tag{D.3}$$

The statistics are computed over $17{,}990$ training pairs and are summarized in Table D.1.

After calibration, the dominant-modality ratio becomes substantially more balanced: approximately $55.0\%/45.0\%$ for RGB/IR on the training set and $54.0\%/46.0\%$ on the validation set. Table D.2 summarizes this change.

These statistics show that calibration substantially reduces the observed dominant-modality imbalance associated with raw-score comparison. The calibrated scores should therefore be interpreted

Table D.1: Training-set statistics used for modality-wise Fourier score calibration.

Modality	μ_m	σ_m
RGB	-2.81929	0.24782
IR	-2.60568	0.22796

Table D.2: Dominant-modality statistics before and after modality-wise calibration.

Setting	RGB main (%)	IR main (%)
Raw Fourier, train	11.3	88.7
Raw Fourier, val	9.9	90.1
z-score, train	55.0	45.0
z-score, val	54.0	46.0

as relative deviations from the typical score distribution of each modality, rather than direct comparisons of raw spectral magnitude.

The normalized scores are converted into modality weights using

$$[w_{\mathrm{rgb}}, w_{\mathrm{ir}}] = \mathrm{Softmax}\left(\frac{[z_{\mathrm{rgb}}, z_{\mathrm{ir}}]}{T_f}\right). \tag{D.4}$$

The temperature T_f controls how strongly a given score difference is translated into a modality preference. Table D.3 summarizes the validation-set weight distributions at different temperatures.

Table D.3: Effect of temperature on the sharpness of modality weights. The two columns report the proportions of samples with small and large RGB–IR weight differences, respectively.

T_f	$\lvert w_{\mathrm{rgb}} - w_{\mathrm{ir}}\rvert < 0.1$ (%)	$\lvert w_{\mathrm{rgb}} - w_{\mathrm{ir}}\rvert > 0.2$ (%)
1.0	24.2	58.3
1.5	32.5	41.4
2.0	41.3	28.4
2.5	49.9	19.7
3.0	58.3	13.8

As expected from temperature scaling, increasing T_f produces smoother modality weights, whereas a smaller temperature yields more decisive selection. The final implementation uses $T_f = 2.56$, close to the moderate regime observed in this diagnostic analysis. This analysis characterizes the behavior of the weighting mechanism and should not be interpreted as an independent performance ablation over T_f.

Figure D.1 visualizes the statistics in Tables D.2 and D.3. Calibration reduces the observed main-modality imbalance, while higher temperatures yield more samples with similar modality weights.

D.3 Boundary and Consistency Gates

The boundary gate

$$B(x) = \exp\left(-\frac{|S_{\mathrm{main}}(x)|}{\tau}\right) \tag{D.5}$$

restricts auxiliary refinement primarily to the neighborhood of the zero-level set, while the consistency gate

$$C(x) = \exp\left(-\frac{|S_{\mathrm{rgb}}(x) - S_{\mathrm{ir}}(x)|}{\gamma_c}\right) \tag{D.6}$$

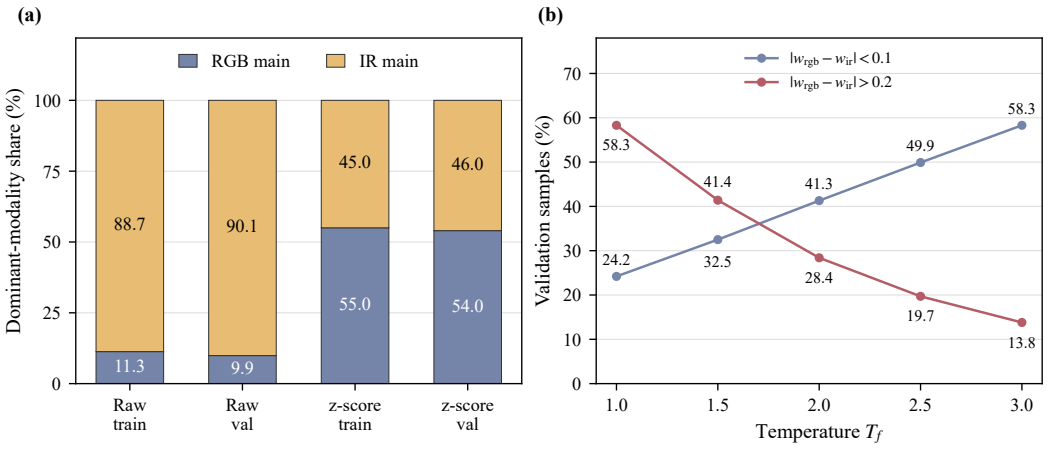

Figure D.1: FCGF calibration and weight-sharpness diagnostics. (a) RGB/IR dominant-modality proportions before and after z-score calibration on training and validation sets. (b) Validation proportions with small (< 0.1) or large (> 0.2) absolute modality-weight differences. Lines connect the measured temperature settings as a visual guide; these are weight diagnostics, not detection-performance comparisons.

suppresses residual correction where the two modality-specific fields strongly disagree.

Table 4 evaluates the two gates independently and jointly. Without either gate, the model obtains 68.9% mAP_{50}. Using only $B(x)$ increases performance to 73.3%, while using only $C(x)$ reaches 72.8%. Combining both gates gives 74.3%. The joint configuration therefore improves over the corresponding single-gate settings by 1.0 and 1.5 percentage points, respectively.

These results support the complementary roles of the two gates: $B(x)$ determines where auxiliary geometry is introduced, whereas $C(x)$ controls how strongly geometrically inconsistent regions are allowed to participate in the refinement.

D.4 Residual Fusion Mechanism

After the modality weights are obtained, the modality with the larger weight is treated as the main geometry and the other as the auxiliary geometry. Their geometric residual is

$$\Delta S(x) = S_{\text{aux}}(x) - S_{\text{main}}(x). \tag{D.7}$$

FCGF does not add this residual directly. Instead, it first constructs the modulated correction

$$\delta S(x) = \lambda B(x) C(x) w_{\text{aux}} \Delta S(x), \tag{D.8}$$

and updates the main field as

$$S_{\text{res}}(x) = S_{\text{main}}(x) + \delta S(x). \tag{D.9}$$

Thus, the auxiliary modality acts as a constrained geometric correction rather than replacing the dominant geometry.

The confidence term

$$G = \text{clip}\left(\frac{|w_{\text{rgb}} - w_{\text{ir}}| - \delta_c}{1 - \delta_c}, 0, 1\right) \tag{D.10}$$

provides an additional sample-wise control. When the modality weights are close, G approaches zero and the method falls back toward the equal-weight average

$$S_{\text{avg}} = \frac{S_{\text{rgb}} + S_{\text{ir}}}{2}. \tag{D.11}$$

When the modality preference is more pronounced, the residual-refined field receives a larger contribution:

$$S_{\text{fuse}} = (1 - G)S_{\text{avg}} + GS_{\text{res}}. \tag{D.12}$$

The complete mechanism therefore contains three distinct controls: sample-wise Fourier weighting identifies the dominant modality, $B(x)$ and $C(x)$ constrain the spatial residual correction, and G controls how strongly the dominant-modality decision affects the final fused field.

E Implementation Details

E.1 SDF Construction and Normalization

Box-induced fields use the local-coordinate construction in Appendix A.1, with negative distances inside the object. Predicted and target values must use the same coordinate system and distance units when forming point-wise residuals. For truncation scale $T > 0$, the normalized construction analyzed in Appendix A.3 is

$$\widetilde{\phi}^T(\mathbf{q}) = \text{clip}(\phi(\mathbf{q}), -T, T)/T. \tag{E.1}$$

The resulting field lies in $[-1, 1]$. The threshold $r = 1$ in GML is expressed in the units of the fields used by the objective; it should not be interpreted as a one-pixel displacement without accounting for normalization. The truncation bounds are theoretical properties of this construction and do not determine a numerical truncation scale.

E.2 Fourier Preprocessing and Training-Set Statistics

Inputs are converted to grayscale and spatially mean-centered before Fourier magnitude extraction and radial-band aggregation, as described in Appendix D.1. Calibration statistics are computed from 17,990 training pairs: $(\mu_{\text{rgb}}, \sigma_{\text{rgb}}) = (-2.81929, 0.24782)$ and $(\mu_{\text{ir}}, \sigma_{\text{ir}}) = (-2.60568, 0.22796)$. These training-set statistics are also used for validation score calibration. The score definition, input intensity scale, and frequency bands must match those used to estimate the reference statistics.

E.3 FCGF Hyperparameters

We use $T_f = 2.56$ to control modality-weight sharpness. The coefficient λ scales the auxiliary residual, τ controls the boundary neighborhood, γ_c controls tolerance to cross-modal field disagreement, and δ_c controls when dominant-modality selection is used. The gate scales operate in field-value units, so their interpretation depends on SDF normalization. Table 4 evaluates gate inclusion rather than sensitivity to the gate scales.

E.4 GML and Loss Weights

The final objective retains L1 supervision for the box-induced SDF term and applies GML with $p = 3/2$ to the RGB, IR, and fused dense branches. The three dense branches use equal coefficients. This specifies their relative allocation, not the absolute scale of the geometric loss relative to the detection objective.

As reported in Section 4.1, BRIDGE is trained from scratch for 200 epochs with batch size 8 and input resolution 800×800 on one NVIDIA GeForce RTX 4090. The optimizer is SGD with initial learning rate 0.01, cosine decay, and five warm-up epochs. The reported patience is 60 and the random seed is 1. AdaLoRA uses $r_{\text{init}} = 9$ and $r_{\text{target}} = 6$.

E.5 SAM3-Derived Supervision Generation

Ground-truth boxes prompt SAM3 to generate object masks offline. The mask boundary then defines the signed-distance target according to Eq. (A.39). The masks supply boundary supervision during training; inference uses the same scalar-field representation without SAM3.

F Additional Generalization Analysis

This section extends the RGB–IR oriented-detection analysis to optical–SAR inputs and horizontal boxes, combining the results in Table 6 with selected published comparisons.

F.1 RGB–SAR Generalization

Table F.1: Selected published optical–SAR OBB detection results on M4-SAR. Class-wise AP and mAP_{50} are in percent. Red and blue indicate the highest and second-highest available values in each accuracy column.

Method	Venue	BD	HB	OT	PG	AP	WT	mAP_{50}	Params (M)
CFT	PR 2022	75.8	92.5	61.3	91.6	90.3	96.3	84.6	53.80
CLANet	ACM MM 2023	74.8	92.2	60.7	91.3	91.6	97.2	84.6	48.20
CSSA	CVPR 2023	73.3	91.7	59.3	88.9	91.6	95.8	83.4	13.50
CMADet	IEEE TIV 2024	70.9	90.7	52.0	86.4	91.7	97.1	81.5	41.50
ICAFusion	PR 2024	74.7	91.9	60.9	91.0	91.8	96.7	84.5	29.00
MMIDet	IEEE TCSVT 2024	74.9	92.6	61.1	91.7	91.4	97.0	84.8	53.80
E2E-MFD	NeurIPS 2024	76.1	91.9	61.1	91.8	91.3	97.2	84.9	31.30
E2E-OSDet	arXiv 2025	77.7	90.7	64.3	91.8	92.1	97.8	85.7	27.50
BRIDGE (Ours)	–	82.0	91.9	78.2	91.2	84.9	98.2	87.7	18.97

BD: bridge; HB: harbor; OT: oil tank; PG: playground; AP: airport; WT: wind turbine. –: unavailable.

Table F.1 compares BRIDGE with selected published optical–SAR OBB detectors. Evaluation protocols differ across sources.

On M4-SAR (Wang et al., 2025), BRIDGE improves mAP_{50} from 86.8 to 87.7, approximately 0.9 percentage points above the baseline. This experiment evaluates optical–SAR fusion on a different dataset, so the change in sensor modality cannot be isolated from the dataset change.

F.2 HBB Generalization

Table F.2 summarizes the published multimodal methods in the OAFA HBB comparison and the SIM-PCSR validation result alongside BRIDGE. The comparison covers selected methods under their reported protocols.

On DroneVehicle (Sun et al., 2022), BRIDGE improves HBB detection from 85.2 to 85.6 mAP_{50}, a gain of approximately 0.4 percentage points. Here the modality pair remains RGB–IR while the detection representation changes from oriented to horizontal boxes. A horizontal rectangle can be encoded by the box-induced SDF with zero orientation, so the continuous-field construction remains applicable.

Table F.2: Published RGB–IR HBB detection results on DroneVehicle. Reference results are collected from the OAFA HBB comparison and the SIM-PCSR validation study. Class-wise AP and mAP_{50} are in percent; red and blue mark the highest and second-highest available values.

Method	Venue	Car	Truck	Bus	Van	Freight Car	mAP_{50}	Params (M)
PIAFusion	Inf. Fusion 2022	–	–	–	–	–	60.5	–
LAIIFusion	Remote Sens. 2023	94.5	54.4	90.5	33.9	57.9	66.2	–
D-ViTDet	ICICML 2022	–	–	–	–	–	66.3	–
RISNet	Remote Sens. 2022	–	–	–	–	–	66.4	–
GFD-SSD	arXiv 2019	90.1	71.7	89.3	57.6	61.4	74.0	–
AFFCM	IEEE JSTARS 2023	90.1	73.4	89.9	64.9	64.9	76.6	–
ICAFusion	PR 2024	98.4	76.3	96.4	65.7	75.1	82.4	–
SLBAF-Net	Multimed. Tools Appl. 2023	98.5	76.9	96.4	67.6	77.3	83.3	–
OAFA	CVPR 2024	98.5	77.8	96.8	69.6	77.2	84.0	–
SIM-PCSR†	Sensors 2026	98.1	83.8	97.3	73.9	73.5	85.3	18.19
BRIDGE (Ours)	–	98.6	78.7	96.8	82.5	71.6	85.6	18.68

–: unavailable. †SIM-PCSR reports results on the official validation split with 640×640 inputs.