This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2026.3750106

An Asymmetric Multi-Task Framework with Spatial Alignment and Gaussian Soft Detection-Oriented Fusion for Visible-Infrared Object Detection

Yu Wan, Jie Li, *Member, IEEE*, Liupeng Lin, Yiliang Li, Yuzeng Chen, Zaiyan Zhang, Qiangqiang Yuan, *Member, IEEE*, Liangpei Zhang, *Fellow, IEEE*

***Abstract*—Visible-infrared object detection is important for UAV perception under challenging conditions such as low illumination, fog, and cross-modal spatial offsets. Existing RGB-IR detectors still face three coupled issues. First, symmetric dual-stream backbones fuse visible and infrared features at all depths, although the deepest visible features are often unreliable under adverse imaging conditions. Second, sensor differences and UAV viewpoint changes introduce residual spatial misalignment between modalities. Third, common image fusion losses often suppress infrared-visible differences as reconstruction errors, rather than exploiting the thermal-appearance discrepancy to enhance detection-oriented target saliency while preserving context. To address these issues, we propose an Asymmetric Multi-Task Framework for Visible-Infrared Detection (VID-AMTF). The visible branch is truncated at res4, while the infrared branch preserves res5 as a reliable high-level semantic prior. A bilateral spatial aware module (BSAM) then predicts cross-guided deformable offsets and modulation masks for both modalities, performs depthwise feature alignment, and fuses features through an infrared-primary residual rule. In addition, a multi-modal alignment loss (MAL) regularizes pre-fusion RGB/IR feature distributions, and a Gaussian Soft Detection-Oriented Fusion Loss (S-DOFL) supervises an auxiliary decoder with Gaussian weights generated from ground-truth bounding boxes, selectively converting infrared-visible modality discrepancy into target-centric infrared saliency and visible-context preservation. Experiments on DroneVehicle and RGBT-Tiny show that VID-AMTF achieves strong detection performance and improves robustness under illumination degradation and spatial misalignment. The code will be released at https://github.com/KIKYOUWY/VID-AMTF.**

***Index Terms*—Visible-infrared object detection, UAV imagery, asymmetric fusion, spatial alignment, multi-task learning.**

I. Introduction

UAV-based object detection supports many remote sensing applications, including intelligent transportation, agricultural monitoring, and disaster response. In these scenarios, visible images provide rich texture and color information, but their target cues can be weakened by low illumination, fog, or other adverse imaging conditions. Infrared images capture thermal radiation and are less dependent on ambient light, offering complementary target saliency for all-weather perception. Therefore, visible-infrared (RGB-IR) detection has become an important direction for robust UAV object detection.

Although RGB-IR fusion has achieved notable progress, existing detection frameworks still face three coupled problems. The first problem is semantic reliability across network depth. Recent uncertainty-aware and illumination-aware detectors adaptively adjust modality contributions under degraded imaging conditions [1], [2], and cross-modal detection networks further enhance semantic interaction through adaptive fusion, semantic conflict modeling, or decision-level integration [3]–[5]. However, many frameworks still rely on largely symmetric dual-stream feature extraction and fuse visible and infrared features at comparable semantic depths. When visible images are degraded, the deepest visible stages tend to encode illumination artifacts, background interference, and weak object semantics [6]. Directly fusing these unreliable high-level visible features with infrared features may contaminate the infrared semantic prior and reduce detection accuracy.

The second problem is cross-modal spatial correspondence. Even after image-level registration, RGB and IR pairs usually retain residual offsets caused by sensor differences, platform motion, and viewpoint changes. Recent disparity-guided frameworks have started to improve spatial correspondence between modalities [7]. Nevertheless, pixel-level concatenation, rigid feature aggregation, purely channel-wise reweighting, or limited alignment cues cannot fully adapt to nonlinear local shifts in UAV scenes. As a result, unaligned fusion may mix object responses from different spatial locations and weaken bounding-box localization.

The third problem is the objective gap between image fusion and object detection. Recent RGB-IR studies demonstrate the value of task-driven multi-modal fusion, but conventional fusion losses are usually designed to optimize global visual quality and reduce modal differences uniformly. For detection, however, the infrared-visible discrepancy is not merely reconstruction noise. Target-centric regions can benefit from stronger infrared thermal contrast, while surrounding contextual regions often require visible texture and scene layout. A global fusion constraint may suppress this target-relevant discrepancy and provide suboptimal supervision for the detector.

Y. Wan, Y. Chen, Zaiyan Zhang and Q. Yuan are with the School of Geodesy and Geomatics, Wuhan University, Wuhan 430079, China (2024282140091@whu.edu.cn, yzchen1006@163.com, zzaiyan@whu.edu.cn, qqyuan@sgg.whu.edu.cn).

L. Lin is with the School of Resource and Environmental Sciences, Wuhan University, Wuhan 430079, China (e-mail: linliupeng@whu.edu.cn).

J. Li is with the School of Geodesy and Geomatics, Wuhan University, Wuhan 430079, China, and is also with Hubei Luojia Laboratory, Wuhan 430079, China (e-mail: jli89@sgg.whu.edu.cn).

Y. Li is with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China (e-mail: 3234902193@qq.com).

L. Zhang is with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China (e-mail: zlp62@whu.edu.cn).

© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2026.3750106

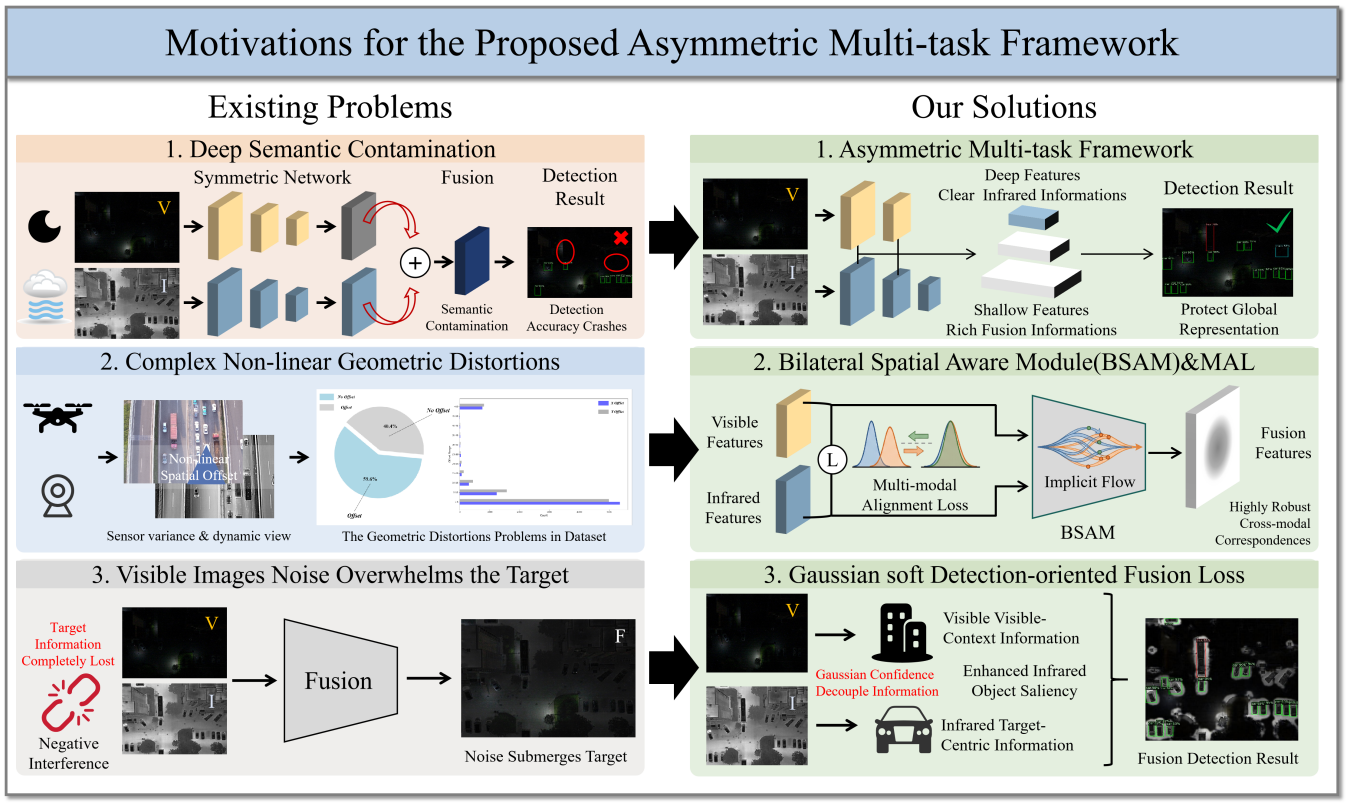


Fig. 1. Motivation of this work. VID-AMTF addresses deep visible semantic contamination, nonlinear cross-modal misalignment, and task-agnostic fusion supervision through asymmetric feature extraction, spatial alignment, and difference-aware Gaussian soft detection-oriented fusion.

Overall, asymmetric semantic reliability, feature-level spatial alignment, and detection-oriented fusion supervision are rarely considered jointly in one framework.

Motivated by these observations, we propose an Asymmetric Multi-Task Framework for Visible-Infrared Detection (VID-AMTF). The framework follows a simple design principle: preserve the most reliable infrared semantics, use visible information only at trustworthy shallow and middle levels, align modalities before fusion, and exploit infrared-visible differences for detection-aware saliency supervision. Specifically, the visible branch is truncated at res4 to avoid introducing degraded res5 semantics, while the infrared branch preserves res5 as a stable high-level semantic prior. A bilateral spatial aware module (BSAM) and a multi-modal alignment loss (MAL) are then used to reduce feature-level geometric and statistical discrepancies. BSAM predicts cross-guided deformable sampling fields for both modalities, aligns them in feature space, and performs infrared-primary residual fusion so that visible cues are injected as complementary information rather than treated as an equally reliable deep semantic source. Finally, a Gaussian Soft Detection-Oriented Fusion Loss (S-DOFL) is introduced on an auxiliary training decoder to softly decouple infrared saliency and visible context. Instead of eliminating modality discrepancy globally, S-DOFL uses it as a target-saliency cue: ground-truth bounding boxes are converted into continuous Gaussian foreground confidence maps, encouraging target centers to inherit infrared saliency while surrounding regions preserve visible contextual fidelity. The overall motivation is summarized in Fig. 1. The main contributions of this work are as follows:

(1) We propose an asymmetric multi-task framework (VID-AMTF) for RGB-IR UAV object detection. By truncating the visible branch at res4 and preserving the infrared res5 feature, the framework reduces deep visible semantic contamination while retaining reliable thermal high-level semantics.

(2) We develop a bilateral spatial aware module (BSAM) with a multi-modal alignment loss (MAL) to handle cross-modal misalignment. BSAM performs cross-guided offset prediction, depthwise deformable alignment, and infrared-primary residual fusion, while MAL constrains pre-fusion RGB/IR feature distributions.

(3) We introduce a Gaussian Soft Detection-Oriented Fusion Loss (S-DOFL) to bridge image fusion and object detection. By using ground-truth bounding boxes to build continuous target-centric confidence maps, S-DOFL transforms infrared-visible discrepancy into target-centric thermal saliency and visible-context preservation without increasing inference complexity.

Extensive experiments on DroneVehicle and RGBT-Tiny show that the proposed framework achieves leading detection performance and consistent robustness under illumination degradation, fog, and spatial offsets.

II. Related Work

A. Visible and Infrared Image Fusion

Single-modality sensing is highly susceptible to environmental influences, which can degrade detection and tracking

© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2026.3750106

performance [8]. For instance, visible imagery may lose target cues under nighttime or foggy conditions, while infrared imagery lacks color and fine texture information. Visible-infrared image fusion therefore aims to combine complementary modality information and generate more informative visual representations. Early traditional fusion algorithms mainly focused on handcrafted feature extraction and fusion rules. By decomposing images at different scales, wavelet transforms can capture detail and approximation information, while Laplacian pyramid methods [9] fuse images through hierarchical multi-scale decomposition. Wang et al. [10] presented a sparse-representation-based fusion method, which extracts image features through sparse coding and introduces an adaptive fusion rule.

Deep learning-based methods have become the mainstream approach for image fusion. YDTR [11] is based on a Y-shaped dynamic transformer with two branches for modality-specific feature extraction. A multi-level refinement registration network (MRRN) [12] is an unsupervised method for fusing infrared and visible images through generation and registration. MHTNet [13] is a multilevel hybrid transmission network that generates fused images with prominent target details. For zero-illumination environments, SLFusion [14] introduces a structure-aware network that separates degraded illumination from scene information and injects edge priors into low-light fusion. From a spatial perspective, recent region-fusion paradigms generate multi-modal semantic-sensitive masks to establish bidirectional collaboration, enhancing reliable target regions while suppressing noisy modality responses [15]. In the frequency domain, dual-branch auto-encoders based on invertible neural networks (INN) [16] preserve fine-grained texture through high-frequency encoding.

Although these fusion methods are effective for visual quality, most of them are task-agnostic and are not explicitly optimized for object localization. They usually apply global visual constraints, which may preserve background texture but cannot guarantee target-centric saliency for detection. This motivates a detection-oriented fusion strategy that treats infrared-visible discrepancy as a saliency cue and softly coordinates infrared target responses with visible contextual fidelity to serve the detector rather than only the visual appearance.

B. Visible and Infrared Object Detection

Multi-modal detection tasks differ from fusion tasks in that they require higher-level semantic information and salient target features to achieve better detection performance. Existing multi-modal object detection methods are primarily categorized into three fusion strategies: pixel-level fusion, feature-level fusion, and decision-level fusion. Pixel-level fusion directly adds or concatenates images and uses them as the detector input. YOLOrs [17] presents a two-stage fusion approach that uses a deep convolutional network to separately extract features from different modalities. Subsequently, the extracted features are combined via concatenation and element-wise cross-multiplication operations. SuperYOLO [18] introduces a multi-modal fusion method to improve small target detection.

Feature-level fusion maps multi-modal images into high-dimensional modality features and performs interaction inside the detection network. UA-CMDet [1] is a drone-based cross-modality detection method that improves robustness through uncertainty-aware learning. SLBAF-Net [3] is designed for multi-modal UAV detection under challenging recognition conditions. He et al. presented CALNet [4] for multispectral object detection, addressing semantic conflicts between visible and infrared images. To dynamically balance cross-modal contributions under changing illumination, the edge-guided illumination-aware interactive learning framework (EI2Det) [2] predicts illumination intensity and adaptively adjusts RGB/IR fusion weights. To address spatial misalignments in aerial imagery with efficient sequence modeling, Zhou et al. [7] proposed the disparity-guided multispectral Mamba (DMM) architecture. Within decision-level fusion, individual detectors are trained separately for each modality, and their detection outputs are combined through voting or weighted averaging. MFPT [5] strengthens single-modal features through intra- and cross-modal transformations and implements an enhanced decision-level fusion strategy.

Despite these advances, most RGB-IR detectors still follow a largely symmetric fusion philosophy, assuming that both modalities are equally reliable at all feature depths. This assumption becomes weak when visible deep semantics are degraded by illumination or fog. In addition, residual cross-modal misalignment remains difficult to handle through rigid aggregation or channel reweighting alone. Few studies jointly consider asymmetric semantic reliability, feature-level spatial alignment, and detection-oriented fusion supervision in a unified multi-task framework. These gaps directly motivate the proposed VID-AMTF.

III. Details of the Proposed Framework

A. Architecture of the Proposed Method

The proposed VID-AMTF follows the problem-motivation-design chain introduced above. As illustrated in Fig. 2, the architecture first adopts an asymmetric feature extraction strategy to address depth-dependent semantic reliability. The infrared stream retains a complete hierarchy to capture stable target semantics, while the visible stream is truncated before the deepest stage to avoid introducing unreliable high-level visible responses under adverse imaging conditions. After feature extraction, the heterogeneous visual signals are projected into a unified latent space through the redesigned bilateral spatial aware module (BSAM). Rather than computing only attention weights, BSAM explicitly performs cross-guided deformable alignment: the offset field of each modality is predicted under the guidance of the opposite modality, depthwise deformable sampling corrects spatial shifts, and an infrared-primary residual fusion unit injects visible cues into an infrared-dominant representation. In this way, the multi-modal interaction is aligned before fusion and remains consistent with the asymmetric backbone.

The BSAM outputs at stem through res4 are subsequently aggregated together with the standalone infrared res5 feature by a Feature Pyramid Module (FPM) to construct the detection pyramid. For the detection task, a diffusion-based detection head is employed, whose iterative refinement mechanism is

© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2026.3750106

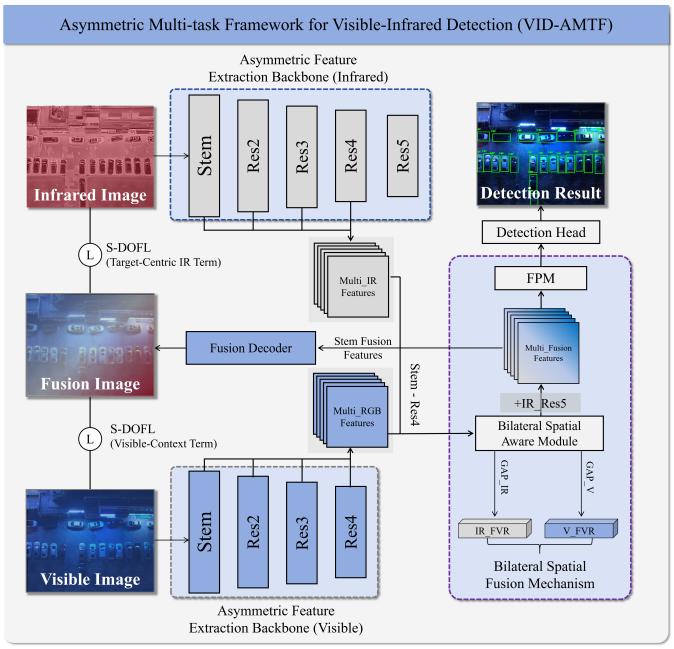


Fig. 2. Architecture of VID-AMTF. The infrared branch preserves the complete feature hierarchy to provide a stable res5 semantic prior, while the visible branch is truncated at res4 to avoid unreliable deep visible semantics. BSAM aligns and fuses stem-to-res4 RGB-IR features before FPM aggregation, and the auxiliary fusion decoder is supervised by S-DOFL only during training.

well suited to the ambiguous object boundaries commonly observed in misaligned multi-modal imagery. In parallel, an auxiliary fusion decoder is retained only during training. Supervised by the Gaussian Soft Detection-Oriented Fusion Loss (S-DOFL), this decoder regularizes the fused representation in pixel space without altering the infrared-primary design of BSAM. The entire architecture is therefore jointly optimized by detection loss, multi-modal alignment loss (MAL), and S-DOFL.

B. Asymmetric Feature Extraction Backbone

Traditional multi-modal detection networks commonly use symmetric dual-stream backbones, implicitly assuming that both modalities provide equally reliable information across all feature depths. This assumption is not always valid for UAV detection in adverse environments. Deep convolutional stages are designed to encode abstract object-level semantics, but visible images under low illumination or fog may contain weak target responses and strong background artifacts. Propagating such visible signals into the deepest stage can amplify unreliable high-level responses, and fusing them with infrared features may contaminate the infrared semantic prior.

To reduce this contamination, an Asymmetric Feature Extraction Backbone is formulated. Both modality branches are instantiated based on the standard ResNet-50 architecture. The infrared backbone preserves the complete hierarchy, including stem and res2 through res5. This full-depth infrared extraction retains high-level thermal semantics that are less dependent on ambient lighting.

In contrast, the visible feature extractor is truncated at res4 and omits the deepest res5 stage. This design excludes the most degradation-sensitive visible semantics while preserving local textures, edge details, and contextual cues from shallow and middle layers, which remain useful for background interpretation and object boundaries.

This asymmetric strategy makes the highest-level representation governed by the infrared modality. It also matches the deployment pattern of BSAM: fusion is performed only at stem, res2, res3, and res4, whereas the res5 feature is kept as a pure infrared semantic prior. As a result, visible features contribute where texture, contour, and context are reliable, while the detector avoids unnecessary fusion at the most degradation-sensitive stage.

C. Bilateral Spatial Aware Module

Residual sensor disparities and UAV viewpoint changes can introduce nonlinear spatial offsets between RGB and IR features. To reduce this geometric conflict without explicit image-level warping, BSAM is formulated as a three-stage cross-guided deformable alignment pipeline, as shown in Fig. 3: (1) cross-guided offset prediction, (2) depthwise deformable alignment, and (3) infrared-primary residual fusion.

For each modality, the offset field is not predicted from the modality itself alone, but under the guidance of the opposite modality. Given a main feature F_{main} and a guide feature F_{guide} at the same pyramid level, the two are concatenated and processed by a depthwise-separable predictor:

$$H = \mathrm{PW}_{2C\to C}(\mathrm{DW}_{3\times3}([F_{main}; F_{guide}])), \tag{1}$$

$$\begin{aligned} \Delta &= \mathrm{Conv}_{1\times1}^{C\to 2K}(\mathrm{ReLU}(H)), \\ M &= \sigma\big(\mathrm{Conv}_{1\times1}^{C\to K}(\mathrm{ReLU}(H))\big), \end{aligned} \tag{2}$$

where $K = 9$ denotes the number of sampling points in a 3×3 neighborhood, $\Delta \in \mathbb{R}^{B\times 2K\times H\times W}$ is the deformable sampling offset field, and $M \in \mathbb{R}^{B\times K\times H\times W}$ is the modulation mask. For the infrared stream, (Δ_{IR}, M_{IR}) is generated from (F_{IR}, F_{RGB}); for the visible stream, (Δ_{RGB}, M_{RGB}) is generated from (F_{RGB}, F_{IR}). The offset and mask heads are zero-initialized so that training starts from a stable near-identity alignment regime and progressively learns geometry-aware sampling.

The predicted offsets are then used to align the main feature through a depthwise deformable convolution:

$$F'_{main} = \mathrm{DeformConv}_{dw}(F_{main}, \Delta, M), \tag{3}$$

where groups are set to C, so each channel performs deformable sampling independently. This design substantially reduces parameter cost while still enabling each spatial location to adapt its sampling pattern according to cross-modal geometric cues. The aligned infrared and visible features are therefore written as:

$$\begin{aligned} F'_{IR} &= \mathcal{A}(F_{IR}, \Delta_{IR}, M_{IR}), \\ F'_{RGB} &= \mathcal{A}(F_{RGB}, \Delta_{RGB}, M_{RGB}), \end{aligned} \tag{4}$$

with $\mathcal{A}(\cdot)$ denoting the deformable alignment operator.

After alignment, BSAM performs infrared-primary residual fusion instead of symmetric averaging. The aligned infrared feature is treated as the semantic anchor, while the visible feature contributes only through a gated residual branch:

$$\begin{aligned} \Delta F &= \mathrm{Conv}_{1\times1}([F'_{IR}; F'_{RGB}]), \\ \alpha &= \sigma(\mathrm{Conv}_{1\times1}([F'_{IR}; F'_{RGB}])), \end{aligned} \tag{5}$$

© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2026.3750106

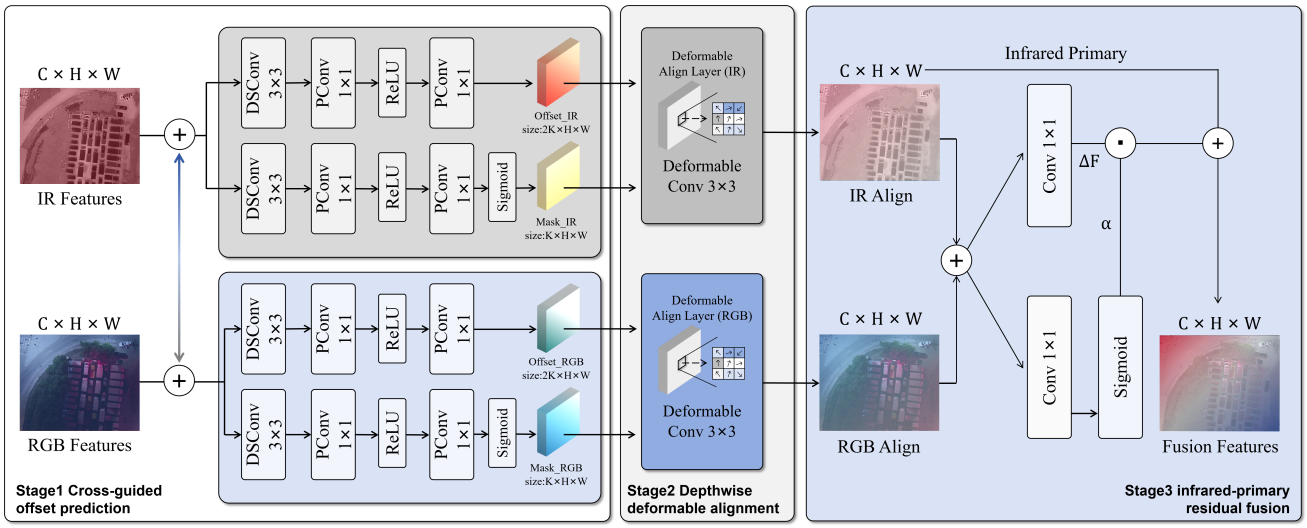


Fig. 3. Structure of the bilateral spatial aware module (BSAM). Cross-guided offset prediction, depthwise deformable alignment, and infrared-primary residual fusion are sequentially performed to obtain spatially calibrated multi-modal features.

$$F_{Fuse} = F'_{IR} + \alpha \odot \Delta F. \tag{6}$$

Here $\alpha \in \mathbb{R}^{B\times C\times H\times W}$ is a channel-wise gate. Its bias is initialized to -2.0, which yields $\alpha \approx 0.12$ at the beginning of training and prevents unstable visible compensation from disturbing the infrared trunk too early. When visible cues are severely degraded, both the residual term ΔF and the gate α are naturally suppressed, so the output tends toward the aligned infrared feature. This makes the fusion rule fully consistent with the asymmetric backbone design.

BSAM is deployed at four pyramid levels, namely stem, res2, res3, and res4, with channel dimensions $\{64, 256, 512, 1024\}$. Owing to the depthwise-separable offset predictors and depthwise deformable alignment, the total parameter count of the four BSAM instances is about 16.84M, which is markedly smaller than a naive full-convolution counterpart while retaining explicit geometry-aware alignment capability.

D. Fusion Decoder

While contemporary multi-modal detectors typically relegate image fusion to an isolated preprocessing step, VIDAMTF embeds it as a training-only auxiliary branch. The fusion decoder projects fused latent representations back into image space, providing the canvas on which S-DOFL imposes Gaussian soft target-aware reconstruction supervision. It therefore complements BSAM and MAL with detection-guided pixel-space regularization without altering the inference pipeline.

As shown in Fig. 4, the fusion decoder takes the fused stem feature $F_{\text{StemFusion}}$ as input and progressively restores the image scale through two stages. Each stage consists of a Residual Block (RSB), a depthwise separable convolution, a pointwise convolution, and an upsampling operation. A 3×3 convolution f_{Conv4} adjusts the channels and yields the final fusion result I_{Fusion}:

$$I_{\text{Fusion}} = f_{\text{Conv4}}(f_{\text{stage2}}(f_{\text{stage1}}(F_{\text{StemFusion}}))) \tag{7}$$

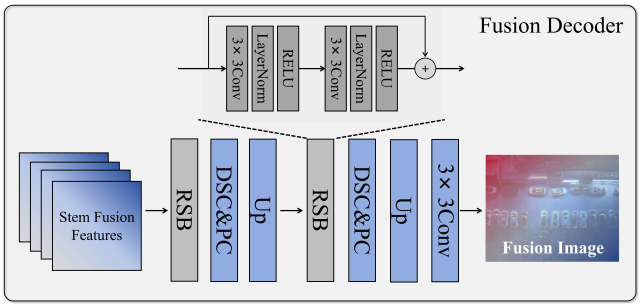


Fig. 4. Structure of the auxiliary fusion decoder. The fused stem feature is progressively upsampled through residual and depthwise separable convolution stages to reconstruct the fusion image for training-time S-DOFL supervision.

where $f_{\text{stage1}}(\cdot)$ and $f_{\text{stage2}}(\cdot)$ represent the two stages of the fusion decoder. In each stage, the RSB is composed of two 3×3 convolutions, normalization layers, and ReLU activation functions. Finally, the original input F_{in} is added back to the output, thereby preserving details and maintaining gradient flow. The operation of RSB is represented as follows:

$$F_{\text{out}} = f_{\text{relu}}(f_{\text{LN2}}(f_{\text{Conv6}}(f_{\text{relu}}(f_{\text{LN1}}(f_{\text{Conv5}}(F_{\text{in}})))))) + F_{\text{in}} \tag{8}$$

where $f_{\text{Conv5}}(\cdot)$ and $f_{\text{Conv6}}(\cdot)$ represent the 3×3 convolutions, and $f_{\text{LN1}}(\cdot)$ and $f_{\text{LN2}}(\cdot)$ represent LayerNorm operations. By introducing the fusion decoder, the network obtains an auxiliary reconstruction branch whose output is used only for S-DOFL supervision. This preserves the infrared-primary detection pipeline at inference time while still allowing training-time pixel-space regularization.

E. Loss Functions

To jointly optimize detection and auxiliary fusion, we use a multi-task loss consisting of detection loss, multi-modal alignment loss, and Gaussian soft detection-oriented fusion loss. The overall loss function can be expressed as:

$$L_{Total} = L_D + L_{MAL} + L_{S\text{-}DOFL} \tag{9}$$

© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2026.3750106

VID-AMTF optimizes the fusion and detection tasks cooperatively through this joint objective. The detection loss is composed of three terms and can be expressed as:

$$L_D = \beta_{focal} L_{focal} + \beta_{L1} L_{L1} + \beta_{giou} L_{giou}, \quad (10)$$

where L_{focal} denotes the focal loss associated with the discrepancy between predicted and actual class labels. L_{L1} represents the L1 loss between the estimated and actual bounding boxes. L_{giou} denotes the loss based on GIoU metric [19]. β_{focal}, β_{L1} and β_{giou} are the weight coefficients that balance the overall multiple losses. In this work, β_{focal}, β_{L1} and β_{giou} are set to 2, 5, and 2, respectively.

Multi-Modal Alignment Loss: Optimizing the implicit manifold alignment requires more than terminal detection gradients. To enforce explicit statistical consistency before BSAM fusion, the multi-modal alignment loss (MAL) is applied to the raw RGB and IR features at the four BSAM insertion levels, namely $\mathcal{S} = \{\text{stem}, \text{res2}, \text{res3}, \text{res4}\}$. For each level l, global average pooling first compresses the modality features into channel descriptors:

$$z^l_{IR} = \text{GAP}(F^l_{IR}), \quad z^l_{RGB} = \text{GAP}(F^l_{RGB}). \quad (11)$$

Since these descriptors are not probability distributions, a channel-wise Softmax operation is further applied to obtain valid non-negative and normalized distributions:

$$P^l_{IR} = \text{Softmax}(z^l_{IR}), \quad P^l_{RGB} = \text{Softmax}(z^l_{RGB}). \quad (12)$$

The Jensen-Shannon (JS) divergence is then computed on P^l_{IR} and P^l_{RGB} rather than on the raw pooled features. Specifically, MAL is accumulated as:

$$L_{MAL} = \sum_{l \in \mathcal{S}} \left[\frac{1}{2} \sum_i P^l_{IR}(i) \log \frac{2P^l_{IR}(i)}{P^l_{IR}(i) + P^l_{RGB}(i)} + \frac{1}{2} \sum_i P^l_{RGB}(i) \log \frac{2P^l_{RGB}(i)}{P^l_{IR}(i) + P^l_{RGB}(i)} \right]. \quad (13)$$

In implementation, a small numerical constant is added inside the logarithm for numerical stability. Minimizing the JS divergence compels the network to align the channel-wise feature probability distributions of the two modalities before offset prediction and deformable fusion take place. MAL therefore complements BSAM: MAL penalizes distribution mismatch in feature statistics, whereas BSAM corrects geometry directly in feature space.

Gaussian Soft Detection-Oriented Fusion Loss: While MAL and BSAM operate in feature space, S-DOFL regularizes the auxiliary fusion decoder in pixel space. Traditional image fusion losses usually apply global constraints, such as structural similarity or global L1 distance, to synthesize visually pleasing composites. However, detection requires spatially varying modality preferences: target-centric regions should emphasize the thermal-appearance discrepancy that makes infrared targets salient, while surrounding context-dominant regions should preserve visible texture and scene layout. Therefore, S-DOFL uses ground-truth bounding boxes to generate Gaussian soft foreground confidence maps, enabling continuous difference-aware target supervision as shown in Fig. 5. S-DOFL adopts a Gaussian soft target-aware formulation. For

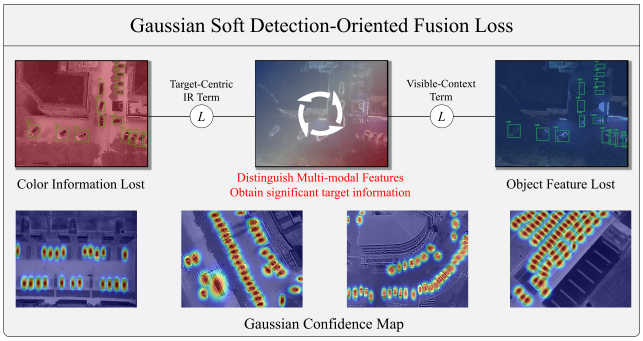


Fig. 5. Illustration of the Gaussian Soft Detection-Oriented Fusion Loss (S-DOFL). Ground-truth bounding boxes are converted into Gaussian soft foreground confidence maps, transforming infrared-visible discrepancy into target-centric infrared saliency while preserving visible contextual fidelity in surrounding regions.

the i-th ground-truth bounding box $B_i = (x^i_1, y^i_1, x^i_2, y^i_2)$, its center location and size are computed as:

$$c^i_x = \frac{x^i_1 + x^i_2}{2}, \quad c^i_y = \frac{y^i_1 + y^i_2}{2}, \quad w_i = x^i_2 - x^i_1, \quad h_i = y^i_2 - y^i_1. \quad (14)$$

The Gaussian bandwidth is adaptively determined by the box scale:

$$\sigma^i_x = \lambda w_i, \quad \sigma^i_y = \lambda h_i, \quad (15)$$

where $\lambda = 0.35$ is the scale coefficient. It controls the spatial bandwidth of the Gaussian confidence map relative to the object size. A small λ makes the infrared-dominant supervision concentrate only around the object center, whereas a large λ expands it into surrounding background regions. We set $\lambda = 0.35$ to provide a moderate soft transition that covers the target body and boundary while retaining sufficient visible-context supervision outside the target region. The Gaussian confidence map generated by this box is formulated as

$$G_i(x, y) = \exp\left(-\frac{1}{2} \left[\frac{(x - c^i_x)^2}{(\sigma^i_x)^2} + \frac{(y - c^i_y)^2}{(\sigma^i_y)^2} \right] \right). \quad (16)$$

When multiple objects appear in one image, the infrared supervision weight is obtained by pixel-wise maximum aggregation:

$$W_{IR}(x, y) = \max_i G_i(x, y), \quad W_{RGB}(x, y) = 1 - W_{IR}(x, y). \quad (17)$$

Here W_{IR} highlights target-centric regions where infrared thermal responses differ from weak or degraded visible appearances, while W_{RGB} assigns complementary emphasis to visible contextual information. To avoid ambiguity in the image-domain supervision, all images used in S-DOFL are represented as three-channel tensors. When the raw infrared image is single-channel, it is converted into a three-channel pseudo-RGB tensor by channel replication to match the visible image and the fusion decoder output. Thus, $I_{\text{Fusion}}, I_{IR}, I_{RGB} \in \mathbb{R}^{3 \times H \times W}$ and $C = 3$. Before computing S-DOFL, the three tensors are normalized by the same channel-wise affine transformation used in the detector:

$$\tilde{I}^c = \frac{I^c - \mu_c}{\sigma_c}, \quad (18)$$

where μ_c and σ_c are the channel-wise mean and standard deviation used in detector input preprocessing. S-DOFL is then

© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2026.3750106

defined on the normalized tensors as

$$L_{S\text{-}DOFL} = \frac{\sum_{c,x,y} W_{IR}(x,y)\left|\tilde{I}^{c}_{\text{Fusion}}(x,y) - \tilde{I}^{c}_{IR}(x,y)\right|}{C\sum_{x,y} W_{IR}(x,y) + \epsilon} + \frac{\sum_{c,x,y} W_{RGB}(x,y)\left|\tilde{I}^{c}_{\text{Fusion}}(x,y) - \tilde{I}^{c}_{RGB}(x,y)\right|}{C\sum_{x,y} W_{RGB}(x,y) + \epsilon}, \tag{19}$$

where $\tilde{I}_{\text{Fusion}}$, $\tilde{I}_{IR}$, and $\tilde{I}_{RGB}$ denote the normalized reconstructed fusion image, infrared image, and visible image, respectively, and ϵ is a small constant for numerical stability.

This soft dual-weight strategy directly follows the detection requirement. The target-center regions receive stronger infrared supervision to preserve high-contrast thermal signatures, which is useful when visible targets are weak. The surrounding context-dominant regions receive stronger visible supervision to preserve texture distribution and scene layout. In this way, S-DOFL does not suppress cross-modal differences as reconstruction noise, but selectively converts the infrared-visible discrepancy into detection-oriented saliency cues. The continuous Gaussian weighting provides smooth target-aware regularization for the multi-task architecture and reduces the objective gap between visual fusion and target localization.

IV. Experiments

To comprehensively evaluate the robustness and accuracy of the proposed VID-AMTF architecture, this section details the datasets, implementation protocols, and evaluation metrics. The detection performance is systematically benchmarked against state-of-the-art single-modality and multi-modality paradigms. Furthermore, qualitative fusion evaluations are conducted to verify the effectiveness of the auxiliary S-DOFL regularization branch.

A. Datasets

To thoroughly evaluate the proposed VID-AMTF, two multi-modal object detection datasets were selected for the experiments: DroneVehicle and RGBT-Tiny.

DroneVehicle [1]: DroneVehicle is a drone-based visible-infrared detection dataset. In our experiments, the training set contains 17,990 RGB-IR image pairs, while the testing set contains 1,469 pairs. The original oriented bounding boxes are resampled into horizontal boxes to accommodate the detector.

RGBT-Tiny [20]: RGBT-Tiny is a visible-infrared tiny object detection dataset comprising 115 paired RGB-infrared video sequences, totaling 93,000 frames and providing 1.2 million manual annotations. The large annotation demand in dense remote sensing tasks has also motivated annotation-efficient learning research [21]. The objects in this dataset are extremely small and subject to cross-modal semantic conflicts, spatial misalignment, and low-light conditions, thus offering a challenging benchmark for algorithm development.

Both datasets provide registered image pairs from different modalities. Nevertheless, perfect alignment is difficult to guarantee, and residual misalignment can still exist between the RGB and infrared data.

B. Experimental Settings

Experiments were conducted using an NVIDIA GeForce RTX 4080 GPU with 16 GB memory. The model was optimized by the AdamW optimizer [22] with an initial learning rate of 0.000025. The baseline of the proposed VID-AMTF is DiffusionDet with ResNet-50-FPM, where the number of diffusion detection boxes was set to 500.

In the detection comparison experiments, the compared methods are divided into three groups: single RGB detectors, single IR detectors, and RGB-IR multi-modal detectors. On DroneVehicle, Faster R-CNN [23], Sparse R-CNN [24], RetinaNet [25], and DiffusionDet [26] were selected as single-modality baselines, and each model was trained using the corresponding modality labels. For pixel-level multi-modal fusion, UMF [12] and BDLFusion [27] were used to generate fused images, which were then fed into the baseline detector for training and inference. Based on the original baseline, we also constructed DiffusionDet*, which concatenates infrared and visible images at the input stage and modifies the first convolutional layer to accept the increased number of channels. Finally, we selected several advanced RGB-IR detection algorithms, including SLBAF-Net [3], CMA-Det [28], C2Former [29], M2FNet [30], DaFF [31], E2E-MFD [32], FMPENet [33], DMM [7], DDCINet [34], and CCSFuse [35]. These methods perform feature-level fusion inside the network to exploit complementary information from both modalities.

On RGBT-Tiny, which provides a relatively comprehensive benchmark, Faster R-CNN, Sparse R-CNN, RetinaNet, and DiffusionDet were selected for the RGB modality, while Faster R-CNN, Sparse R-CNN, RetinaNet, DiffusionDet, CenterNet [36], ACM [37], FCOS [38], ALCNet [39], and DNA-Net [40] were selected for the IR modality. For RGB-IR fusion, DiffusionDet*, ProbEn [41], CMA-Det [28], UA-CMDet [1], C2Former [29], M2FNet [30], and DMM [7] were chosen as comparative methods.

In the multi-modal fusion comparison experiment, UMF [12], BDLFusion [27], TGFuse [42], YDTR [11], and VID-AMTF* were selected as comparative methods. Among them, UMF, BDLFusion, and VID-AMTF* output three-channel visible-like fusion images, while TGFuse and YDTR output single-channel grayscale fusion images. VID-AMTF* denotes the variant trained with separate L1 losses to visible and infrared images.

C. Evaluation Metrics

The model's effectiveness was mainly evaluated using mean average precision (mAP) at the IoU threshold of 0.5. The mAP is computed by averaging the average precision (AP) over all classes. In the experiments on DroneVehicle, mAP50 was selected as the evaluation metric. On RGBT-Tiny, in addition to mAP50, mAPs, mAPm, and mAPl were also used. These three metrics evaluate mAP50 for small, medium, and large objects, respectively. Small objects are defined as those with an area smaller than 32^2 pixels, medium objects have an area from 32^2 to 96^2 pixels, and large objects have an area larger than 96^2 pixels. In the experimental tables, the best values are

© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2026.3750106

TABLE I
DETECTION PERFORMANCE COMPARISON ON DRONEVEHICLE

Modality	Method	Venue	mAP50	bus	car	freight	truck	van
RGB	RetinaNet	ICCV 2017	18.6	19.7	62.5	2.8	5.9	2.0
	Faster R-CNN	TPAMI 2017	56.3	83.1	81.9	33.2	47.2	36.0
	Sparse R-CNN	CVPR 2021	64.4	87.4	84.4	45.3	54.5	50.5
	DiffusionDet	ICCV 2023	65.4	88.6	85.4	45.1	54.9	52.8
IR	RetinaNet	ICCV 2017	36.7	53.9	88.4	16.1	18.0	7.2
	Faster R-CNN	TPAMI 2017	62.1	88.7	93.3	46.5	44.9	37.3
	Sparse R-CNN	CVPR 2021	75.1	93.3	95.6	61.9	65.8	59.0
	DiffusionDet	ICCV 2023	75.2	93.0	94.0	64.8	65.6	58.6
RGB+IR	UMF	IJCAI 2022	68.5	88.6	94.0	49.2	58.7	52.1
	BDLFusion	IJCAI 2023	68.2	88.3	94.2	47.7	57.9	53.0
	DiffusionDet*	ICCV 2023	77.1	93.1	94.5	66.4	71.2	60.4
	SLBAF-Net	Multimed.T.A 2023	60.4	88.5	97.4	32.3	46.9	37.0
	DaFF	CVPR 2024	76.1	89.9	89.6	62.6	73.9	58.8
	E2E-MFD	NeurIPS 2024	77.4	89.8	90.3	64.6	**79.3**	63.1
	CMA-Det	TIV 2024	75.1	**95.8**	**98.0**	60.6	71.8	49.2
	C2Former	TGRS 2024	77.2	94.1	97.2	66.1	68.5	60.1
	M2FNet	JAG 2024	76.1	94.1	94.2	65.1	65.8	61.4
	FMPENet	Sensors 2025	76.8	94.5	95.2	60.5	76.8	57.1
	DMM	TGRS 2025	77.6	89.4	90.8	65.5	76.2	65.9
	DDCINet	TGRS 2025	78.4	90.7	91.0	66.1	78.9	65.5
	CCSFuse	TGRS 2026	79.4	94.3	97.1	67.8	72.1	64.1
	VID-AMTF	-	**82.2**	95.4	95.4	**73.8**	77.1	**69.4**

indicated in bold, while the second-best values are underlined. For parameter count and inference time, lower values are considered better.

D. Detection Performance Comparison

To evaluate the detection performance of VID-AMTF, comparisons are conducted against both single-modality baselines and recent multi-modal detection methods. The quantitative results on the DroneVehicle dataset are reported in Table I. VID-AMTF achieves the best overall mAP50 of 82.2, exceeding the second-best CCSFuse by 2.8 percentage points.

The comparison also reflects the motivation of the proposed asymmetric design. The infrared DiffusionDet baseline is much stronger than the visible one, indicating that infrared features provide more reliable target semantics in the DroneVehicle scenarios. DiffusionDet* improves the infrared baseline from 75.2 to 77.1 mAP50 through simple input concatenation, but it still treats both modalities symmetrically and does not explicitly address unreliable deep visible features or residual spatial offsets.

Compared with pixel-level fusion methods such as UMF and BDLFusion, VID-AMTF does not rely on a task-agnostic fused image as detector input. Instead, it uses the asymmetric backbone to preserve reliable infrared depth, BSAM and MAL to align RGB/IR features before fusion, and S-DOFL to provide Gaussian soft detection-oriented auxiliary supervision. As a result, VID-AMTF improves over DiffusionDet* by 5.1 percentage points in mAP50 and obtains clear gains on difficult categories such as freight and van, supporting the effectiveness of the proposed design rationale: asymmetric semantic selection, geometry-aware alignment, and target-aware fusion regularization.

The qualitative detection results in Fig. 6 further support this interpretation. Under nighttime or low-visibility conditions, visible single-modality detectors tend to miss weak targets, while symmetric fusion baselines may produce false positives or ambiguous boxes when spatial shifts are not calibrated. By combining BSAM-based geometric correction with infrared-primary residual fusion, VID-AMTF preserves reliable thermal target cues while suppressing degraded visible interference. The auxiliary S-DOFL branch further regularizes the fused representation during training, helping the detector maintain stable localization in challenging scenes.

To further evaluate robustness against spatial misalignment and semantic conflicts, experiments are conducted on the RGBT-Tiny benchmark. The results in Table II show that VID-AMTF achieves the best mAP50, mAPs, and mAPm among the compared methods, while reaching the same best mAPl value as several infrared and RGB-IR baselines.

Tiny objects are particularly sensitive to cross-modal spatial offsets, because small localization errors can cover a large portion of the target region. The integration of BSAM within the multi-scale fusion hierarchy directly addresses this issue. Through cross-guided offset prediction and depthwise deformable alignment, the network adaptively corrects local spatial shifts before fusion.

This geometry-aware fusion strategy brings clear perfor-

© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2026.3750106

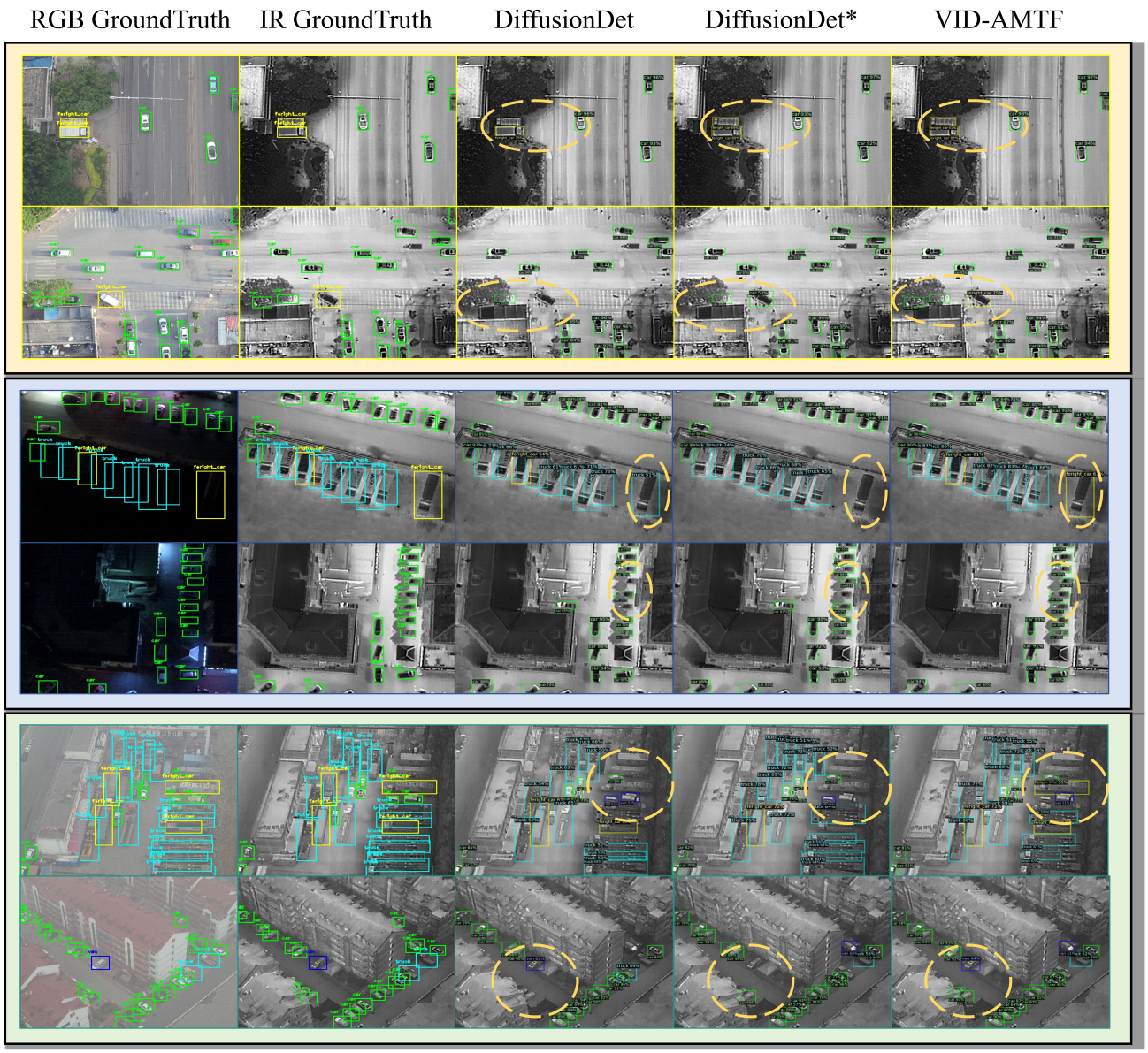


Fig. 6. Qualitative detection results on the DroneVehicle dataset. The proposed VID-AMTF preserves reliable infrared target cues and suppresses degraded visible interference under severe illumination deficits. The yellow, blue, and green boxes denote detection results under sunny, nighttime, and foggy conditions, respectively.

mance gains, outperforming the strong infrared DiffusionDet baseline and the simple fusion variant DiffusionDet* by 4.3 and 6.1 percentage points in mAP50, respectively. Together with the Gaussian soft target-aware regularization imposed by S-DOFL during training, the network better localizes tiny objects in complex backgrounds. The scale-specific metrics further support this observation: VID-AMTF achieves 46.5 mAPs, 58.2 mAPm, and 100.0 mAPl, obtaining the best small- and medium-object performance among all compared methods. The improvement is particularly clear for small objects, where VID-AMTF exceeds the second-best mAPs by 7.9 percentage points, indicating stronger robustness for tiny targets that are more sensitive to weak responses and cross-modal offsets.

E. Fusion Performance Comparison

Beyond bounding-box evaluation, fusion visualization is used to analyze whether the auxiliary decoder learns the intended target-aware regularization. Conventional fusion algorithms usually pursue a global visual compromise, preserving visible background texture but weakening infrared target contrast. In contrast, the reconstructed representations generated by VID-AMTF reflect the constraints of S-DOFL. As shown in Fig. 7, several baselines struggle to separate target cues from degraded contexts, whereas VID-AMTF better preserves the infrared-visible contrast around objects.

The proposed architecture reduces this ambiguity through Gaussian soft detection-oriented supervision. The target-centric infrared term retains high-contrast thermal signatures,

© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2026.3750106

TABLE II
DETECTION PERFORMANCE COMPARISON ON RGBT-TINY

Modality	Method	mAP50	mAPs	mAPm	mAPl
RGB	Faster R-CNN	25.2	25.5	16.4	0.0
	Sparse R-CNN	33.9	33.9	38.9	64.6
	RetinaNet	8.0	2.3	8.7	0.0
	DiffusionDet	34.1	33.1	31.8	68.9
IR	Faster R-CNN	27.7	27.6	38.5	**100.0**
	Sparse R-CNN	36.2	35.2	54.0	**100.0**
	RetinaNet	15.5	5.3	27.5	90.0
	DiffusionDet	37.8	38.3	56.1	**100.0**
	CenterNet	27.3	25.7	28.5	30.0
	FCOS	27.7	38.6	29.0	90.0
	ACM	32.0	22.1	11.4	40.0
	ALCNet	22.0	17.8	7.2	19.3
	DNA-Net	22.0	31.3	14.3	0.0
RGB + IR	DiffusionDet*	36.0	35.2	54.3	**100.0**
	ProbEn	26.8	32.3	22.2	76.7
	CMA-Det	36.0	23.1	24.6	70.0
	UA-CMDet	35.0	35.0	38.5	9.7
	C2Former	37.1	36.9	53.5	19.4
	M2FNet	34.6	33.2	37.9	38.7
	DMM	38.6	36.0	54.9	**100.0**
	VID-AMTF	**42.1**	**46.5**	**58.2**	**100.0**

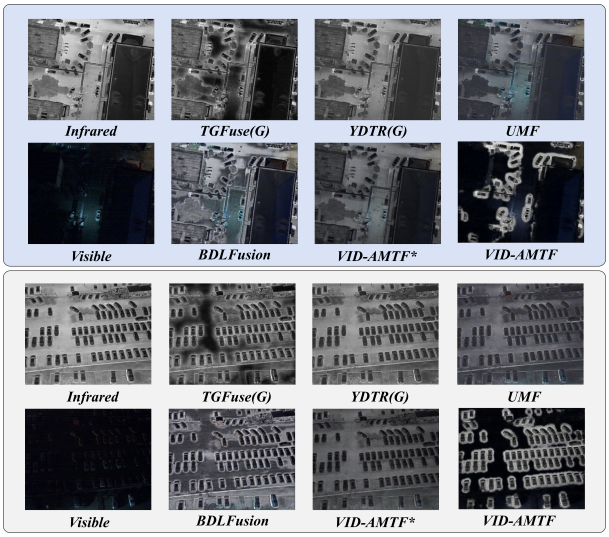


Fig. 7. Qualitative fusion comparison among representative methods. VID-AMTF better preserves infrared target saliency while retaining visible scene context.

while the visible-context term preserves texture and layout in context-dominant regions. Thus, infrared-visible differences are transformed into supervised saliency cues. Although the fusion decoder is used only during training, this target-aware supervision complements the asymmetric backbone and BSAM, explaining why detection improves together with the fusion visualization.

V. DISCUSSION

To analyze the effectiveness of VID-AMTF, ablation studies and efficiency analyses are conducted. This section evaluates whether each component supports the unified design rationale: asymmetric semantic selection, geometry-aware alignment, and detection-oriented fusion regularization.

A. Architectural Component Analysis

To quantify the contribution of each component, a progressive ablation study is conducted on the DroneVehicle dataset. Table III reports the performance changes obtained by adding the asymmetric multi-scale structure, BSAM, MAL, and S-DOFL. The visible-only result is lower than the infrared-only result, again showing that visible semantics are less reliable in challenging UAV scenes. DiffusionDet* improves the infrared baseline from 75.2 to 77.1 mAP50 through simple concatenation, but the remaining gap indicates that symmetric early fusion still lacks explicit semantic selection and spatial alignment.

TABLE III
COMPONENT ABLATION STUDY

Components	mAP50	bus	car	freight	truck	van
Visible	65.4	88.6	85.4	45.1	54.9	52.8
Infrared	75.2	93.0	94.0	64.8	65.6	58.6
DiffusionDet*	77.1	93.1	94.5	66.4	71.2	60.4
+Asymmetric Multi-Scale	79.2	94.3	95.3	68.2	72.9	65.5
+BSAM	80.3	94.8	**95.6**	69.1	74.1	67.9
+MAL	81.0	94.5	94.6	71.3	76.0	68.7
+S-DOFL	**82.2**	**95.4**	95.4	**73.8**	**77.1**	**69.4**

Adding the asymmetric multi-scale structure raises mAP50 from 77.1 to 79.2, confirming the value of limiting visible participation to reliable shallow and middle levels while preserving infrared depth. Adding BSAM further improves mAP50 to 80.3, showing the benefit of cross-guided deformable alignment before fusion. MAL increases the score to 81.0 by regularizing RGB/IR feature distributions, which complements the geometric alignment of BSAM and further stabilizes cross-modal feature fusion. Finally, S-DOFL reaches 82.2 mAP50 by adding Gaussian soft target-aware fusion supervision. The progressive gains verify that the three design motivations are complementary rather than isolated.

B. Asymmetric Multi-scale Feature Aggregation

The asymmetric backbone is built on the hypothesis that visible features are useful mainly at shallow and middle levels, whereas the deepest semantic prior should remain infrared-dominant. To validate this design choice, an ablation study evaluating different visible-infrared fusion depths is conducted. The isolated settings test single-level fusion, while the cumulative settings test whether progressively extending the visible fusion path benefits or harms the infrared-primary detection pyramid. Quantitative outcomes are detailed in Table IV.

TABLE IV
ABLATION STUDY OF MULTI-SCALE FUSION LEVELS

Level	mAP50	bus	car	freight	truck	van
Stem	77.0	92.1	93.1	66.0	70.6	63.2
Res2	77.3	92.4	93.4	66.3	70.9	63.5
Res3	77.7	92.8	93.8	66.7	71.3	63.9
Res4	77.6	92.7	93.7	66.6	71.2	63.8
Res5	77.2	92.3	93.3	66.2	70.8	63.4
Stem-Res2	77.5	92.6	93.6	66.5	71.1	63.8
Stem-Res3	78.3	93.4	94.4	67.3	71.9	64.5
Stem-Res4	**79.2**	**94.3**	**95.3**	**68.2**	**72.9**	**65.5**
Stem-Res5	78.1	93.2	94.2	67.1	71.7	64.3

© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2026.3750106

The isolated results show that no single fusion level is sufficient to support robust RGB-IR detection, because shallow layers provide local texture and boundary cues while middle layers encode stronger object structure. The cumulative results further reveal the central principle of the proposed asymmetric framework: fusing visible and infrared features from stem to res4 achieves the optimal balance, reaching 79.2 mAP50 and the best AP across all object categories. This configuration allows the visible branch to contribute reliable low- and mid-level geometric cues, while the deepest res5 semantic representation remains governed by the infrared branch as an uncorrupted semantic prior.

Crucially, extending the fusion pipeline to encompass the deepest visible stage (res5) reduces mAP50 from 79.2 to 78.1. Compared with Stem-Res3, Stem-Res5 also drops across all categories, indicating that simply accumulating more visible fusion levels does not guarantee better detection. Instead, adding deep visible semantics begins to contaminate the infrared-primary representation. This observation directly supports the asymmetric truncation strategy rather than a conventional symmetric dual-stream fusion design.

Under severe environmental conditions, the deepest convolutional stages of the visible extractor often fail to encode meaningful object semantics. Instead, they capture and amplify illumination noise, fog artifacts, and background interference. Consequently, the proposed VID-AMTF does not treat the two modalities symmetrically: it preserves infrared depth for stable target semantics, but limits visible participation to the levels where texture, contour, and contextual cues remain trustworthy. The Stem-Res4 result therefore validates the architectural premise that asymmetric fusion is not merely a lightweight implementation choice, but the key mechanism for preventing deep semantic contamination.

To further inspect this phenomenon, Fig. 8 provides Grad-CAM visualizations of high-level features under challenging imaging conditions. From left to right, the five columns correspond to the visible image, infrared image, visible baseline Res5 feature, VID-AMTF infrared Res5 feature, and VID-AMTF P5 feature. Under nighttime or foggy conditions, the Res5 feature of the visible-only baseline becomes diffuse and is easily activated by degraded background regions, indicating that the deepest visible semantics are contaminated by adverse appearance cues and lose reliable target information.

In contrast, the infrared Res5 feature in VID-AMTF preserves clearer thermal target responses and provides a more stable high-level semantic prior. More importantly, compared with the infrared Res5 feature alone, the final VID-AMTF P5 feature exhibits stronger and more complete target-centered activations. This indicates that VID-AMTF does not simply rely on infrared information, but improves the final high-level detection representation by injecting useful visible cues from shallow and middle layers while avoiding unreliable deep visible semantics.

C. Gaussian Soft Detection-Oriented Fusion Loss Evaluation

To examine the effectiveness of S-DOFL, its two decomposed terms, *Target-Centric IR Term* and *Visible-Context Term*, are evaluated in Table V. Applying a global constraint biased toward infrared (*L1_IR*) or visible (*L1_V*) produces competitive but limited detection scores. Combining both global constraints (*L1_VIR*) raises mAP50 to 81.3, indicating that visible and infrared reconstruction cues are complementary. S-DOFL further introduces detection-guided Gaussian weights so that the modality discrepancy is emphasized near targets as infrared saliency and relaxed in surrounding regions to retain visible context.

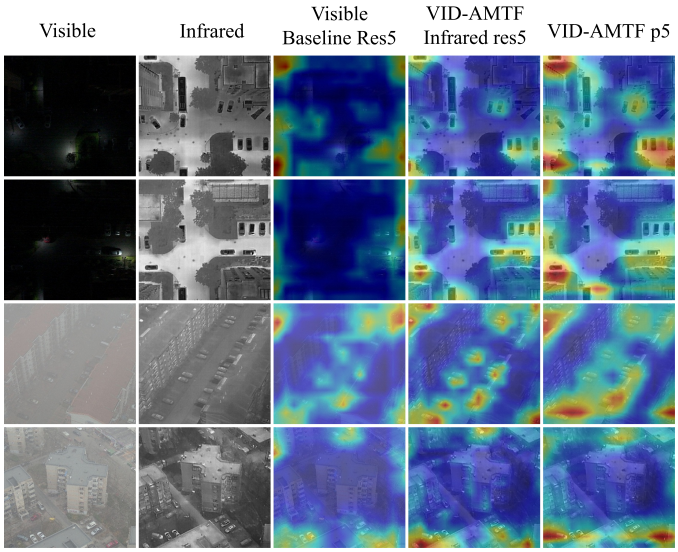


Fig. 8. Grad-CAM visualization of high-level features. From left to right: visible image, infrared image, visible baseline Res5 feature, VID-AMTF infrared Res5 feature, and VID-AMTF P5 feature. Under nighttime and foggy conditions, the visible baseline Res5 feature is easily contaminated by degraded appearance cues, whereas VID-AMTF preserves stable infrared semantics and produces higher-quality target-centered high-level detection features.

TABLE V
ABLATION RESULTS OF FUSION LOSS

Loss	mAP50	bus	car	freight	truck	van
L1_IR	81.1	94.3	95.1	71.8	76.1	68.3
L1_V	81.0	94.4	94.8	72.5	75.8	67.7
L1_VIR	81.3	94.8	95.1	72.9	75.6	68.1
Target-Centric IR Term	81.9	95.0	**95.6**	73.4	76.7	68.6
Visible-Context Term	82.0	**95.8**	95.3	73.0	76.5	69.2
S-DOFL	**82.2**	95.4	95.4	**73.8**	**77.1**	**69.4**

Using the target-centric infrared term alone improves target-oriented supervision and obtains the best car AP as well as strong freight and truck AP. Using the visible-context term alone emphasizes visible contextual fidelity, yielding the best bus AP and the second-best van AP. The complete S-DOFL achieves the best mAP50, freight AP, truck AP, and van AP, showing that Gaussian soft infrared saliency guidance and visible context regularization are complementary. The qualitative results in Fig. 9 are consistent with this conclusion: S-DOFL preserves visible scene context while using infrared-visible differences to make targets more distinguishable under fog and nighttime conditions.

To evaluate the sensitivity of S-DOFL to the Gaussian scale coefficient, we further compare three representative values of λ in Table VI. A small value of $\lambda = 0.20$ produces a narrow target-centered supervision region and achieves 81.3 mAP50. Increasing λ to 0.35 improves mAP50 to 82.2 by providing

© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2026.3750106

a more balanced coverage of the target body and boundary. When λ is further increased to 0.50, the infrared-dominant region extends into more surrounding background areas, and mAP50 decreases to 81.8. The setting $\lambda = 0.35$ also achieves the best AP for bus, car, truck, and van, while $\lambda = 0.20$ is only marginally higher for freight. These results validate $\lambda = 0.35$ as a moderate setting that best balances target saliency and visible-context preservation across categories.

TABLE VI
SENSITIVITY ANALYSIS OF GAUSSIAN SCALE COEFFICIENT λ

λ	mAP50	bus	car	freight	truck	van
0.20	81.3	94.3	94.8	**73.9**	76.2	67.3
0.35	**82.2**	**95.4**	**95.4**	73.8	**77.1**	**69.4**
0.50	81.8	95.1	95.0	73.2	76.9	68.8

S-DOFL does not assume a globally fixed reliability order between modalities. Instead, it uses a detection-guided spatial prior: infrared supervision is emphasized near object centers, where thermal contrast is generally more stable under nighttime, foggy, and low-contrast conditions, while visible supervision preserves surrounding texture and context, especially under good illumination. Thus, its modality preference is local and target-aware, with the Gaussian transition avoiding abrupt switching between target and context regions.

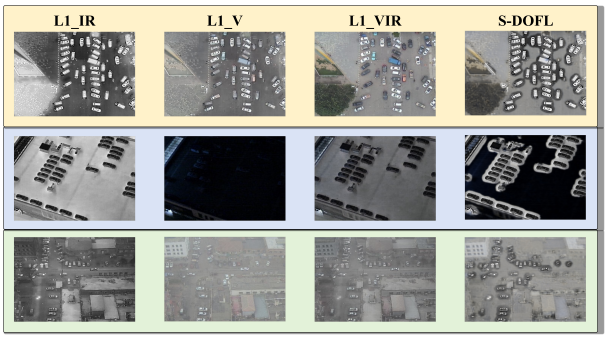


Fig. 9. Qualitative evaluation of different fusion-loss settings. The complete S-DOFL formulation balances target-centric infrared saliency and visible contextual fidelity under severe fog and nighttime conditions. The yellow, blue, and green boxes denote detection results under sunny, nighttime, and foggy conditions, respectively.

D. Offset Robustness Analysis

To systematically quantify the robustness against cross-modal spatial misalignments, the detection performance is evaluated across coupled offsets in the X and Y directions. The offset range is derived from the original DroneVehicle annotations by using the infrared image as the reference and computing the signed pixel displacement between the center of each infrared bounding box and the center of its corresponding visible bounding box. Based on this annotation-derived offset distribution, Fig. 10 visualizes the 55 unique coupled (x, y) offset pairs recorded in the original evaluation files. Across these pairs, the observed horizontal coordinates are in $\{-20, -10, -5, 0, 5, 10, 15, 20, 25\}$ pixels, and the observed vertical coordinates are in $\{-20, -15, -10, -5, 0, 5, 10, 15, 20, 25\}$ pixels.

Fig. 10 visualizes the resulting mAP50 distribution via a three-dimensional surface plot. Conventional baselines are sensitive to modality offsets, and their performance decreases as the spatial shift increases. In contrast, VID-AMTF maintains a higher and more stable performance surface across the tested offsets. Based on the original evaluation records, VID-AMTF improves the average mAP50 across these 55 evaluated offset pairs from 60.0 to 61.7, indicating that the proposed method maintains more stable detection performance under different levels of modality displacement. As the annotation-measured displacement becomes larger, the overlapping object evidence between paired modalities gradually decreases, and the performance gain correspondingly becomes more moderate. Overall, this trend supports the role of BSAM and MAL: BSAM predicts cross-modal offsets and performs deformable alignment before fusion, while MAL regularizes feature distributions so that alignment is easier to optimize.

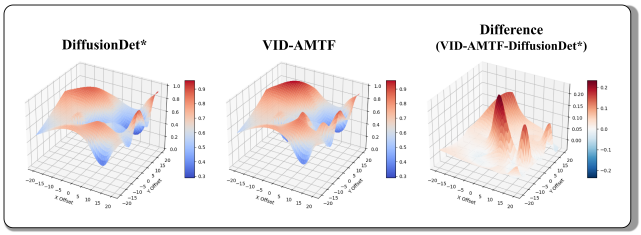


Fig. 10. Offset robustness analysis under coupled X-Y modality shifts. The mAP50 surface and differential performance map demonstrate the stability of VID-AMTF against severe cross-modal spatial misalignment.

E. Comprehensive Performance Analysis

Finally, to verify the comprehensive performance of VID-AMTF, we compared it with recent RGB-IR fusion methods in terms of parameter count and inference speed. To ensure a fair efficiency comparison, all inference-time measurements were conducted on the same NVIDIA GeForce RTX 4080 GPU with a batch size of 1 and an input size of 512×512 for each RGB-IR image pair. For methods with publicly available code, we used their official implementations. For methods without released code, we reproduced their network architectures according to the model descriptions and experimental settings reported in the original papers and evaluated them under the same inference configuration. The experimental results are shown in Table VII.

TABLE VII
COMPREHENSIVE PERFORMANCE COMPARISON

Method	Params (M)	Speed (ms)	mAP50
DiffusionDet	110.6	**29.1**	75.2
DiffusionDet*	110.6	<u>29.3</u>	77.1
C2Former	118.5	32.7	77.2
DMM	**88.0**	34.1	77.6
DDCINet	122.5	39.4	78.4
CCSFuse	126.1	37.3	<u>79.4</u>
VID-AMTF	135.9	42.4	**82.2**

As shown in Table VII, VID-AMTF achieves the highest mAP50 of 82.2 with 135.9 M parameters and a single-image inference time of 42.4 ms. Compared with the infrared DiffusionDet baseline and the simple fusion variant DiffusionDet*,

© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2026.3750106

VID-AMTF brings 7.0 and 5.1 percentage-point mAP50 improvements, respectively. Among recent RGB-IR fusion methods, VID-AMTF improves over the second-best CCSFuse by 2.8 percentage points in mAP50, with 9.8 M additional parameters and 5.1 ms longer inference time. Compared with DDCINet, VID-AMTF further improves mAP50 by 3.8 percentage points with a 3.0 ms increase in inference time. These results indicate that VID-AMTF is not the lightest model, but it provides the strongest detection accuracy while keeping the inference time within a competitive range. Moreover, the auxiliary fusion decoder is used only as a training-time regularization branch and is discarded during inference, so the deployed detector remains centered on the asymmetric backbone, BSAM, and diffusion detection head.

F. Failure Case Analysis

Although VID-AMTF improves RGB-IR detection under challenging conditions, it still has limitations in extremely degraded foggy scenes. As shown in Fig. 11, the first column gives the ground-truth labels, while the second and third columns show the detection results of DiffusionDet* and VID-AMTF, respectively. The first row presents an extreme foggy failure case, and the second row provides a nighttime reference case. In the extreme foggy scene, dense atmospheric interference weakens both visible appearance cues and infrared target cues, so several vehicles remain difficult to detect with high confidence.

Nevertheless, VID-AMTF still performs more reliably than the baseline. Compared with DiffusionDet*, the proposed model recalls more target instances in the degraded foggy scene and assigns higher confidence scores to several detected vehicles. The nighttime reference further shows that when infrared target contrast is clearer, VID-AMTF achieves higher confidence scores. These observations indicate that VID-AMTF alleviates, but cannot fully eliminate, the information loss caused by extremely poor visibility. Future work will further explore degradation-aware enhancement and uncertainty modeling for extremely dense fog and other severe weather conditions.

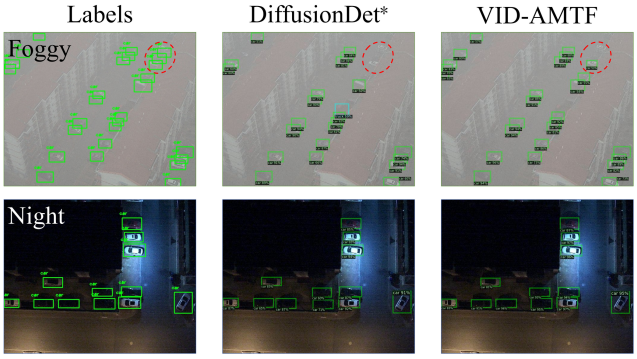


Fig. 11. Failure-case visualization under extreme foggy conditions with a nighttime reference. The first column shows the ground-truth labels, while the second and third columns show the detection results of DiffusionDet* and VID-AMTF, respectively.

VI. Conclusion

This paper addresses three coupled challenges in RGB-IR UAV object detection: unreliable deep visible semantics, residual cross-modal spatial misalignment, and the objective gap between image fusion and detection. We propose VID-AMTF, which truncates the visible branch at res4 while preserving infrared res5 as a stable semantic prior. BSAM and MAL align modalities before fusion through cross-guided deformable sampling, infrared-primary residual fusion, and feature-distribution regularization. S-DOFL further supervises the training-only fusion decoder with Gaussian soft confidence maps generated from ground-truth bounding boxes, converting infrared-visible modality discrepancy into target-centric infrared saliency while preserving visible contextual fidelity. Experiments on DroneVehicle and RGBT-Tiny validate the accuracy gains and the contributions of asymmetric feature extraction, spatial alignment, and difference-aware Gaussian soft detection-oriented fusion supervision.

Acknowledgment

This work is supported in part by the National Natural Science Foundation of China under Grants 42471504, 42301417, and 424B2009, and by the Hubei Provincial Key Research and Development Program under Grant 2023BAB066. (Corresponding author: Jie Li.)

References

[1] Y. M. Sun, B. Cao, P. F. Zhu, and Q. H. Hu, "Drone-based RGB-infrared cross-modality vehicle detection via uncertainty-aware learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, pp. 6700–6713, 2022.

[2] K. Hu, Y. He, Y. Li, J. Zhao, S. Chen, and Y. Kang, "EI^{2}Det: Edge-guided illumination-aware interactive learning for visible-infrared object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 35, pp. 7101–7115, 2025.

[3] X. L. Cheng, K. K. Geng, Z. W. Wang, J. H. Wang, Y. X. Sun, and P. B. Ding, "SLBAF-Net: Super-lightweight bimodal adaptive fusion network for UAV detection in low recognition environment," *Multimedia Tools Appl.*, vol. 82, pp. 47773–47792, 2023.

[4] X. He, C. Tang, X. Zou, and W. Zhang, "Multispectral object detection via cross-modal conflict-aware learning," in *Proc. ACM Int. Conf. Multimedia (MM)*, 2023, pp. 1465–1474.

[5] Y. H. Zhu, X. Y. Sun, M. Wang, and H. Huang, "Multi-modal feature pyramid transformer for RGB-infrared object detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, pp. 9984–9995, 2023.

[6] Y. Wan, J. Li, L. Lin, Z. Lv, Z. Zhang, Q. Yuan, and H. Shen, "Bridging the domain gap: A transferable dynamic routing enhancer for robust aerial detection under adverse weather," *ISPRS J. Photogramm. Remote Sens.*, pp. 1–18, 2026, doi: 10.1016/j.isprsjprs.2026.08.041.

[7] M. Zhou *et al.*, "DMM: Disparity-guided multispectral Mamba for oriented object detection in remote sensing," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, pp. 1–13, 2025.

[8] C. J. Zhao, R. W. Liu, J. X. Qu, and R. B. Gao, "Deep learning-based object detection in maritime unmanned aerial vehicle imagery: Review and experimental comparisons," *Eng. Appl. Artif. Intell.*, vol. 128, 2024.

[9] A. Sahu, V. Bhateja, A. Krishn, and Himanshi, "Medical image fusion with Laplacian pyramids," in *Proc. Int. Conf. Medical Imaging, m-Health and Emerging Communication Systems (MedCom)*, 2014, pp. 448–453.

[10] J. Wang, J. Peng, X. Feng, G. He, and J. Fan, "Fusion method for infrared and visible images by using non-negative sparse representation," *Infrared Phys. & Technol.*, vol. 67, pp. 477–489, 2014.

[11] W. Tang, F. Z. He, and Y. Liu, "YDTR: Infrared and visible image fusion via Y-shape dynamic transformer," *IEEE Trans. Multimedia*, vol. 25, pp. 5413–5428, 2023.

© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2026.3750106

[12] D. Wang, J. Y. Liu, X. Fan, and R. S. Liu, "Unsupervised misaligned infrared and visible image fusion via cross-modality image generation and registration," in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, 2022, pp. 3508–3515.
[13] Q. Q. Li *et al.*, "A multilevel hybrid transmission network for infrared and visible image fusion," *IEEE Trans. Instrum. Meas.*, vol. 71, 2022.
[14] G. Lv, X. Gao, A. Dong, Z. Wei, and J. Cheng, "SLFusion: A structure-aware infrared and visible image fusion network for low-light scenes," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 35, pp. 11877–11892, 2025.
[15] X. Li, G. Zhang, W. Chen, L. Cheng, Y. Xie, and J. Ma, "An infrared and visible image fusion method based on semantic-sensitive mask selection and bidirectional-collaboration region fusion," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 35, pp. 5518–5532, 2025.
[16] H. Liu, Q. Mao, M. Dong, and Y. Zhan, "Infrared-visible image fusion using dual-branch auto-encoder with invertible high-frequency encoding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 35, pp. 2675–2688, 2025.
[17] M. Sharma *et al.*, "YOLOrs: Object detection in multimodal remote sensing imagery," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 1497–1508, 2021.
[18] J. Q. Zhang, J. Lei, W. Y. Xie, Z. M. Fang, Y. S. Li, and Q. Du, "SuperYOLO: Super resolution assisted object detection in multimodal remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023.
[19] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," arXiv:1902.09630, 2019.
[20] X. Y. Ying *et al.*, "Visible-thermal tiny object detection: A benchmark dataset and baselines," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, pp. 6088–6096, 2025.
[21] R. Liu, T. Luo, S. Huang, Y. Wu, Z. Jiang, and H. Zhang, "CrossMatch: Cross-view matching for semi-supervised remote sensing image segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, Art. no. 5650515, 2024, doi: 10.1109/TGRS.2024.3507050.
[22] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," arXiv:1711.05101, 2017.
[23] S. Q. Ren, K. M. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, pp. 1137–1149, 2017.
[24] P. Z. Sun *et al.*, "Sparse R-CNN: End-to-end object detection with learnable proposals," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 14449–14458.
[25] T. Y. Lin, P. Goyal, R. Girshick, K. M. He, and P. Dollar, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 2999–3007.
[26] S. Chen, P. Sun, Y. Song, and P. Luo, "DiffusionDet: Diffusion model for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2023, pp. 19773–19786.
[27] Z. Liu, J. Y. Liu, G. Y. Wu, L. Ma, X. Fan, and R. S. Liu, "Bi-level dynamic learning for jointly multi-modality image fusion and beyond," in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, 2023, pp. 1240–1248.
[28] K. Song, X. Xue, H. Wen, Y. Ji, Y. Yan, and Q. Meng, "Misaligned visible-thermal object detection: A drone-based benchmark and baseline," *IEEE Trans. Intell. Veh.*, vol. 9, pp. 7449–7460, 2024.
[29] M. Yuan and X. Wei, "C^{2}Former: Calibrated and complementary transformer for RGB-infrared object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024.
[30] C. C. Jiang *et al.*, "M2FNet: Multi-modal fusion network for object detection from visible and thermal infrared images," *Int. J. Appl. Earth Obs. Geoinf.*, vol. 130, 2024.
[31] A. Althoupety, L. Y. Wang, W. C. Feng, and B. Rekabdar, "DaFF: Dual attentive feature fusion for multispectral pedestrian detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2024, pp. 2997–3006.
[32] J. Zhang *et al.*, "E2E-MFD: Towards end-to-end synchronous multimodal fusion detection," in *Advances in Neural Information Processing Systems*, vol. 37, 2024.
[33] Z. Wang and Q. Zhang, "Real-time aerial multispectral object detection with dynamic modality-balanced pixel-level fusion," *Sensors*, vol. 25, Art. no. 3039, 2025.
[34] W. Bao, M. Huang, J. Hu, and X. Xiang, "Dual-dynamic cross-modal interaction network for multimodal remote sensing object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, pp. 1–13, 2025.
[35] T. Zhang, R. Lu, X. Yang, D. Zhang, Y. Li, X. Xie, and Y. Li, "CCSFuse: Collaborative compensation and selective fusion for UAV-based RGB-IR object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 64, pp. 1–14, 2026, doi: 10.1109/TGRS.2025.3647293.
[36] X. Zhou, D. Wang, and P. Krahenbuhl, "Objects as points," arXiv:1904.07850, 2019.
[37] Y. M. Dai, Y. Q. Wu, F. Zhou, and K. Barnard, "Asymmetric contextual modulation for infrared small target detection," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, 2021, pp. 949–958.
[38] Z. Tian, C. H. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 9626–9635.
[39] Y. M. Dai, Y. Q. Wu, F. Zhou, and K. Barnard, "Attentional local contrast networks for infrared small target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, pp. 9813–9824, 2021.
[40] B. Li *et al.*, "Dense nested attention network for infrared small target detection," arXiv preprint, 2022.
[41] Y. T. Chen, J. H. Shi, Z. L. Ye, C. Mertz, D. Ramanan, and S. Kong, "Multimodal object detection via probabilistic ensembling," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 139–158.
[42] D. Y. Rao, T. Y. Xu, and X. J. Wu, "TGFuse: An infrared and visible image fusion approach based on transformer and generative adversarial network," *IEEE Trans. Image Process.*, 2024.



Yu Wan received the M.S. degree from the School of Geodesy and Geomatics, Wuhan University, Wuhan, China, in 2026. He is currently pursuing the Ph.D. degree with the School of Geodesy and Geomatics, Wuhan University. His research interests include remote sensing object detection, image quality improvement, remote sensing image processing, and deep learning.



Jie Li (Member, IEEE) received the B.S. degree in sciences and techniques of remote sensing and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2011 and 2016, respectively. He is currently an Associate Professor with the School of Geodesy and Geomatics, Wuhan University. His research interests include image quality improvement, image super-resolution reconstruction, data fusion, remote sensing image processing, sparse representation, and deep learning.



Liupeng Lin received the B.S. degree in geographical information science and the Ph.D. degree in cartography and geographic information engineering from Wuhan University, Wuhan, China, in 2017 and 2022, respectively. He is currently a Hongyi Postdoctoral Researcher with the School of Resource and Environmental Sciences (SRES), Wuhan University. His research interests include SAR image reconstruction, remote sensing image processing and applications, and data fusion.



Yiliang Li received the B.E. degree from the College of Information and Electrical Engineering, China Agricultural University, Beijing, China, in 2023. He is currently pursuing the M.S. degree in pattern recognition and intelligent systems with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China. His research interests include deep learning and remote sensing object detection. He has published multiple papers in leading interdisciplinary journals related to computational sciences.



Yuzeng Chen received the B.S. degree in geographic information science from Southwest University of Science and Technology, Mianyang, China, in 2020, and the M.S. degree from Central South University, Changsha, China, in 2023. He is currently pursuing the Ph.D. degree with the School of Geodesy and Geomatics, Wuhan University, Wuhan, China. His research interests include remote sensing image and video processing and computer vision. More information can be found at https://yzcu.github.io/.

© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2026.3750106



Zaiyan Zhang is currently pursuing the M.E. degree in photogrammetry and remote sensing with Wuhan University, Wuhan, China. He received the B.E. degree in data science and big data technology from China University of Geosciences, Wuhan, China, in 2025. His research interests include remote sensing image processing, multimodal fusion, and representation learning. More information can be found at https://zzaiyan.github.io/.



Qiangqiang Yuan (Member, IEEE) received the B.S. degree in surveying and mapping engineering and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2006 and 2012, respectively. In 2012, he joined the School of Geodesy and Geomatics, Wuhan University, where he is currently a Professor. He has published more than 90 research articles, including more than 70 peer-reviewed articles in international journals, such as IEEE TRANSACTIONS ON IMAGE PROCESSING and IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING. His research interests include image reconstruction, remote sensing image processing and applications, and data fusion.

Dr. Yuan was a recipient of the Top-Ten Academic Star of Wuhan University in 2011 and the Youth Talent Support Program of China in 2019. In 2014, he received the Hong Kong Scholar Award from the Society of Hong Kong Scholars and the China National Postdoctoral Council. He serves on the editorial boards of nine international journals and has frequently served as a referee for more than 50 international journals in remote sensing and image processing.



Liangpei Zhang (Fellow, IEEE) is currently a "Chang-Jiang Scholar" Chair Professor appointed by the Ministry of Education of China with the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing (LIESMARS), Wuhan University, Wuhan, China, and the Aerospace Information Research Institute, Henan Academy of Sciences. He was a Principal Scientist with the China State Key Basic Research Project, appointed by the Ministry of Science and Technology of China to lead the remote sensing program from 2011 to 2016. He has published more than 700 research papers and five books. He is an Institute for Scientific Information (ISI) Highly Cited Author and holds 30 patents. His research interests include hyperspectral remote sensing, high-resolution remote sensing, image processing, and artificial intelligence.

Prof. Zhang is a Fellow of the Institution of Engineering and Technology (IET). He is currently serving as an Associate Editor for IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING.

© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.