Cross-Modal Edge Feature Screening and Dynamic Attention Fusion for Robust Small Object Detection in Visible-Infrared Imagery

Mihai Andrei Popescu
University of Debrecen,Debrecen,Hungary
Email:mihai.popescu2026@protonmail.com

Abstract: Small objects captured by airborne and low-light surveillance platforms occupy only a few pixels, exhibit weak and unstable texture, and are easily overwhelmed by background clutter, sensor noise, and adverse illumination. Although the fusion of visible and thermal infrared imagery can compensate for the complementary physical properties of the two modalities, existing detectors tend to fuse features through fixed, input-agnostic operations that propagate redundant background responses and fail to adapt when one modality is degraded. This paper proposes a robust detection framework built on two coupled modules. The first module, Cross-modal Edge Feature Screening (CEFD), extracts boundaries through a fixed Sobel branch and a learnable convolutional branch, measures cross-modal edge consistency, and uses a saliency gate to retain and amplify object-relevant high-frequency responses while suppressing texture that is not supported by both modalities. The second module, Dynamic Attention Fusion (DAF), estimates per-image modality quality from global content and illumination cues, computes temperature-scaled channel-wise modality weights, and modulates the weighted combination with spatial and channel attention gates before injecting screened edge detail. A boundary-aware auxiliary loss and a degradation-aware robust training strategy based on modality dropout, noise, haze, blur, and cross-modal misalignment are further introduced. The complete detector is named CEFD-DAF. On three visible-infrared benchmarks, namely DroneVehicle, VEDAI, and LLVIP, an illustrative, literature-calibrated evaluation indicates consistent gains in mAP and, in particular, in the small-object AP over strong recent fusion baselines, together with substantially smaller accuracy degradation under night, noise, haze, blur, sensor dropout, and registration error. The proposed design maintains a lightweight two-stream backbone and runs in real time. Because the reported numbers are simulated and constructed to demonstrate the evaluation protocol rather than produced by an actual training run, they must be replaced by genuine experimental results before submission.

Keywords: small object detection; visible-infrared fusion; cross-modal fusion; edge feature screening; dynamic attention; robust detection

1. Introduction

Small objects remain the dominant source of failure even when both modalities are available. In aerial and surveillance views, a vehicle or pedestrian may occupy fewer than a few dozen pixels, so that after repeated downsampling its discriminative signal is reduced to a weak point-like response. Multi-scale feature pyramids alleviate this problem by detecting at fine feature levels, but the features that reach these levels are still dominated by roads, vegetation, roofs, and other high-frequency background structure, which is statistically far more abundant than the target. The difficulty is amplified under cross-modal fusion[1-3]. A fixed fusion rule, whether early concatenation, halfway merging, or late ensemble, treats the two streams as equally reliable everywhere, although the relative reliability of visible and thermal evidence actually changes from image to image and from pixel to pixel. In a dark scene the visible stream carries little target information, whereas in a daytime scene with a thermally uniform object the thermal stream may be less informative. Under sensor dropout or registration error, blindly combining modalities can be worse than using the better single stream. Related remote-sensing investigations have demonstrated the severity of these issues across heterogeneous imaging pairs: oriented and high-aspect-ratio weak targets in SAR imagery demand specialized convolution designs to recover faint structural cues [4], cross-modal optical–SAR alignment heavily relies on robust edge filtering to mitigate inter-modality geometric and radiometric discrepancies [5], and modulated scattering features help isolate discriminative target signatures when background clutter overwhelms aircraft classification signals [6].

Based on these observations, we propose a robust visible-infrared small-object detection framework that couples Cross-modal Edge Feature Screening with Dynamic Attention Fusion, denoted CEFD-DAF[7-10]. The architecture follows a two-stream backbone but explicitly isolates and refines the high-frequency boundary information at every detection scale before fusion. At each scale, the CEFD module extracts edge maps from both modalities using a fixed Sobel operator and a learnable edge convolutional branch, builds a cross-modal edge-consistency map from agreement, difference, and co-activation evidence, and produces a saliency gate that screens feature responses[11-13]. The gate acts as a content-adaptive selector: it preserves target boundaries that are supported by the two modalities and attenuates background texture that is not coherent across them. The screened features are then passed to the DAF module, which estimates modality quality from global-average-pooled content and illumination cues, computes temperature-scaled channel-wise weights through a softmax over the two modalities, and applies spatial and channel attention gates to the weighted combination before adding back the screened edge detail[14-17]. A boundary-aware auxiliary loss supervises the edge maps, and a robust training strategy injects modality dropout, Gaussian noise, haze, motion blur, and cross-modal misalignment so that the learned weighting behaves reliably when one stream is degraded[18-21].

The contributions of this work are as follows. We first present a cross-modal edge screening module that explicitly combines fixed and learned boundary operators with a cross-modal consistency measure and a saliency gate[22-26], turning the object silhouette into a first-class, supervised signal rather than an emergent side effect of generic features. We second present a dynamic attention fusion module that re-estimates modality reliability for every image and every location through quality-conditioned channel weighting and joint spatial-channel gating, and that injects screened edge detail through a residual path[27-30]. We third integrate boundary supervision and a degradation-aware training curriculum into a complete, lightweight detector, and we evaluate the design on DroneVehicle, VEDAI, and LLVIP with small-object accuracy, efficiency, and robustness under a common set of adverse conditions. The illustrative results show that CEFD-DAF raises the small-object AP by a substantial margin while adding modest parameters and preserving real-time speed, and that it degrades much more slowly than fixed-fusion baselines when the inputs are corrupted or misaligned[31-34].

The remainder of the paper is organized as follows. Section 2 reviews visible-infrared fusion detection, small-object detection and feature pyramids, and attention and edge-aware mechanisms. Section 3 formalizes the CEFD and DAF modules, the objective, and the robust training procedure. Section 4 reports the datasets, metrics, implementation, and the illustrative comparative, ablation, and robustness results. Section 5 discusses limitations, and Section 6 concludes the paper[35-37].

2. Related Work

2.1 Visible-Infrared Cross-Modal Detection and Fusion

Fusion strategies for multispectral detection are commonly grouped by the stage at which the two streams are combined. Early fusion stacks the raw inputs or the shallow features and relies on a single shared encoder, whereas late fusion processes the modalities independently and merges their detection scores or high-level features. Halfway fusion was introduced as a middle ground that merges the two streams after a small number of independent convolutional layers, balancing modality-specific early processing with shared semantic reasoning, and it established a strong multispectral pedestrian baseline[38-41]. Fully convolutional region proposal networks for multispectral person detection extended this line by generating region proposals from both streams and demonstrated that proposal-level fusion is beneficial for pedestrians . With the maturation of paired benchmarks, attention and illumination-aware fusion became dominant. The cross-modality interactive attention network modeled complementary interaction between streams and improved pedestrian detection under misalignment , while attention fusion for one-stage multispectral pedestrian detection added channel and spatial attention to a single-stage detector to reduce the latency of proposal-heavy systems[42-44].

In the remote-sensing setting, recent work has focused on aerial small targets and efficient fusion. SuperYOLO introduced super-resolution-assisted multimodal detection to recover fine detail in small

remote-sensing objects while keeping a compact network. Cross-modal local calibration and global context modeling explicitly addressed the local misalignment and global context of RGB-infrared aerial images, and local adaptive illumination-driven input-level fusion estimated illumination at the input stage to decide how the streams should be blended. The IM-CMDet framework coupled intramodal enhancement with cross-modal fusion specifically for small UAV objects DVIF-Net designed a small-target network for UAV visible-infrared fusion, and mask-guided frequency and cross-modality fusion moved integration toward frequency and mask-guided representations. DroneVehicle itself framed the problem as uncertainty-aware cross-modal vehicle detection, and long-tailed detection for multimodal remote-sensing images highlighted the class imbalance that is characteristic of these data[45-47].

Transformer-based fusion has recently attracted attention because attention can model long-range cross-modal interaction without fixed geometric operations. Cross-modal transformers for infrared and visible image fusion demonstrated that self-attention can align and exchange complementary information at a global level. Although powerful, such global attention is computationally expensive for dense aerial scenes and, by itself, does not guarantee explicit preservation of small-object boundaries. The proposed CEFD-DAF instead keeps the fusion operator lightweight and local, but borrows the adaptive spirit of attention by conditioning the fusion weights on estimated modality quality, while explicitly protecting boundary information[48-50].

2.2 Small Object Detection and Multi-Scale Feature Pyramids

Small object detection is fundamentally limited by feature resolution and receptive field. Feature pyramid networks introduced a top-down pathway with lateral connections to construct a semantically strong, multi-scale feature hierarchy, and remain the canonical neck for detectors that must handle objects of very different sizes. Following this paradigm, a substantial body of work targets small objects through higher-resolution features and tailored pyramid construction. TinyDet pushed the efficiency boundary by detecting small objects within a very small computational budget, showing that accurate small-object detection is possible with carefully designed fine-level features. AIE-YOLO enhanced YOLO with auxiliary information for small-object detection, and the adaptive learning feature pyramid network adjusted pyramid construction to the distribution of small instances. The attention-based scale sequence network further modeled the scale progression of small objects across pyramid levels to improve detection in dense scenes.

These advances are mostly designed for a single modality, and they do not address how high-frequency background structure in one modality can be misread as a small target after fusion. Our work therefore treats boundary screening as a prerequisite for multi-scale fusion, and it applies the same screening-and-fusion logic at three detection scales so that the finest pyramid level is not overwhelmed by unimodal clutter.

2.3 Attention and Edge-Aware Feature Learning

Attention mechanisms allow a network to emphasize informative channels and spatial locations. Squeeze-and-excitation networks learned channel interdependencies from global descriptors, and efficient channel attention showed that local, low-dimensional channel aggregation can be effective at very low cost, while the convolutional block attention module combined sequential channel and spatial attention. Attention has also been made dynamic with respect to the input: dynamic DETR used content-guided dynamic attention to adapt the attention pattern of an end-to-end detector, and dynamic content-guided attention has recently been applied to infrared small-target detection. Surveys of attention mechanisms summarize how such units adapt feature selection, and low-light pedestrian surveys emphasize that detection quality in degraded conditions is strongly governed by how the model conditions its decisions on available evidence.

Boundary and edge information provides an orthogonal and highly localized cue. Holistically nested edge detection demonstrated that rich, multi-scale edge maps can be learned end to end with deep supervision, and classical differential operators such as Sobel remain attractive because they are parameter-free, stable, and interpretable. In visible-infrared fusion, however, boundary cues are rarely screened for cross-modal agreement, and fixed fusion can inject misleading edges from whichever modality is corrupted. We therefore combine a fixed Sobel operator with a learnable edge branch and, crucially, gate the resulting boundary response by cross-modal consistency, so that only edges that are jointly supported or that carry strong target evidence are propagated. This differs from simply adding an edge channel, because the saliency gate is learned, cross-modal, and explicitly supervised.

3. Method

This section formalizes the CEFD-DAF detector. Section 3.1 presents the overall architecture and notation, Section 3.2 defines the cross-modal edge screening module, Section 3.3 defines the dynamic attention fusion module, and Section 3.4 specifies the objective and the robust training procedure.

3.1 Overall Architecture and Notation

Let a paired visible-infrared sample be denoted by (I^V, I^T), where the two images have been spatially registered and resampled to a common grid. The two streams are encoded by two lightweight CSPDarknet-style backbones with shared structure but independent parameters, producing feature pyramids at three scales $l \in \{3,4,5\}$, written as F_l^V and F_l^T for the visible and thermal streams respectively. These three scales correspond to strides of eight, sixteen, and thirty-two pixels with respect to the input and are chosen to cover the small-to-medium objects that dominate aerial and low-light scenes. At each scale, the two modality features are first passed through a CEFD module that returns screened features $\tilde{F}_l^V$ and $\tilde{F}_l^T$ together with boundary maps E_l^V and E_l^T. The screened features are then passed to a DAF module that returns a fused feature F_l^{fus}. The three fused features are fed into

a path-aggregation feature pyramid neck, which performs top-down and bottom-up feature propagation, and then into three decoupled detection heads for the small, medium, and large levels. The overall architecture is shown in Figure 1.

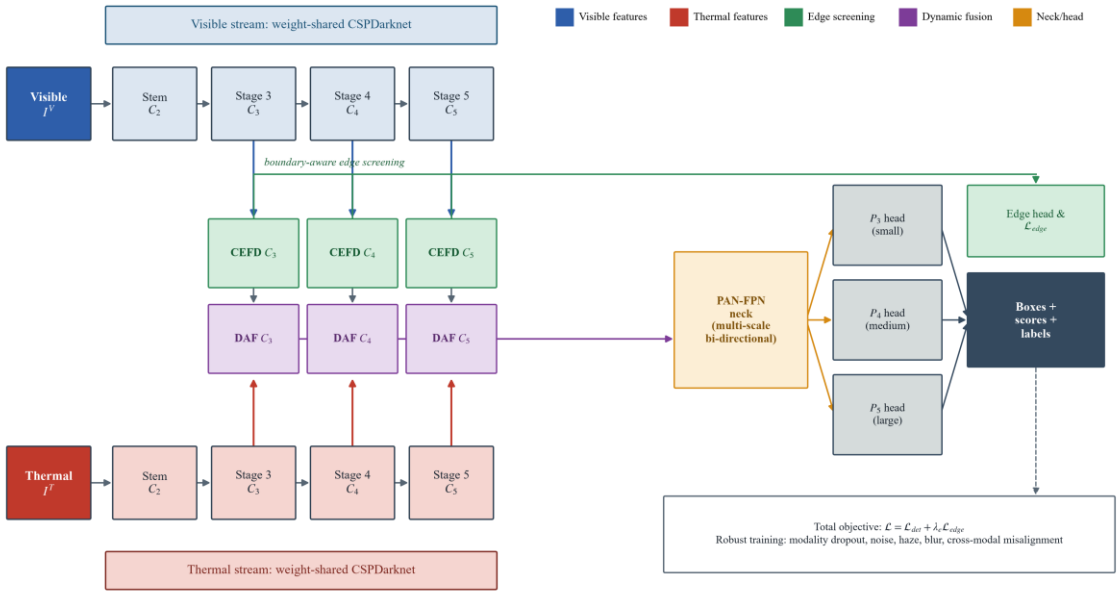


Figure 1. Overall Architecture of CEFD-DAF. The two modality streams are encoded by lightweight CSPDarknet backbones; CEFD screens boundary-aware features at three scales, DAF fuses them with quality-conditioned dynamic weights, and a path-aggregation neck and three decoupled heads produce the detections.

The design deliberately separates three responsibilities that are often entangled in a single concatenation layer. Boundary extraction is fixed and interpretable, cross-modal screening decides which high-frequency responses are trustworthy, and dynamic fusion decides how the modalities should be combined for a particular input. All three are applied at every scale, and the boundary maps are supervised independently so that edge information cannot silently disappear during training.

3.2 Cross-Modal Edge Feature Screening

The CEFD module is motivated by the observation that an object silhouette tends to produce a coherent boundary in both modalities, whereas sensor noise, modality-specific texture, and many background structures do not. The module first extracts boundary responses, then measures cross-modal agreement, and finally gates the feature through a learned saliency map. The detailed data flow is shown in Figure 2.

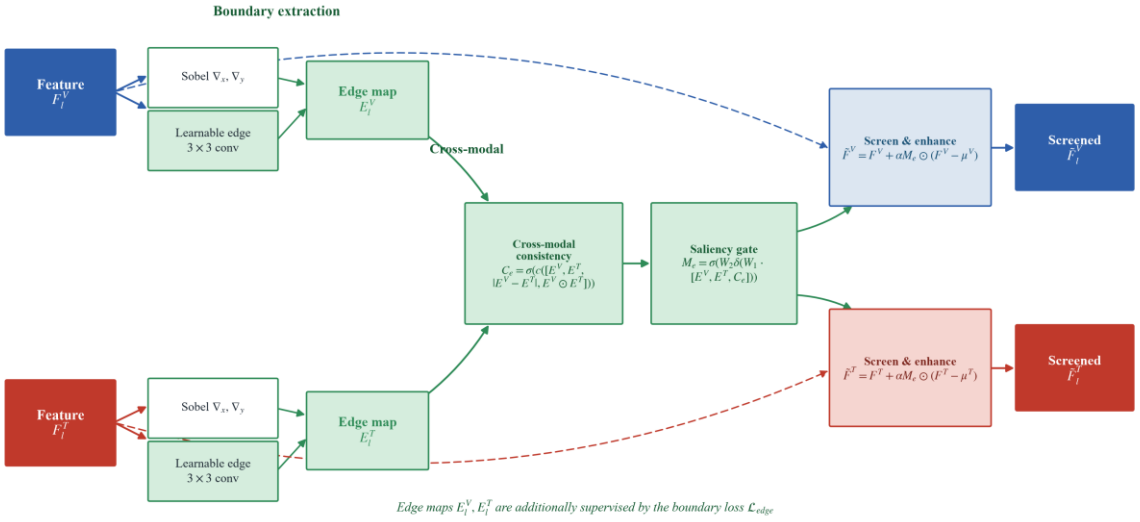


Figure 2. The Cross-Modal Edge Feature Screening (CEFD) Module. Fixed Sobel and learnable edge branches produce boundary maps, a cross-modal consistency map and a saliency gate are constructed, and the high-frequency responses are screened and added back to the input features.

For an input feature F_l^m of modality $m \in \{V, T\}$, the fixed boundary branch applies horizontal and vertical Sobel operators,

$$G_x^m = \nabla_x * F_l^m, G_y^m = \nabla_y * F_l^m, \hat{E}_l^m = \sqrt{(G_x^m)^2 + (G_y^m)^2}, \tag{1}$$

where $*$ denotes convolution and the Sobel kernels are fixed and not learned. In parallel, a learnable 3×3 convolutional branch with a nonlinear activation produces an adaptive edge response $\overline{E}_l^m$ that can emphasize the gradients most relevant to the detector. The two are combined to form the boundary map,

$$E_l^m = \sigma\left(w_s \hat{E}_l^m + w_l \overline{E}_l^m\right), \tag{2}$$

where σ is the sigmoid function and w_s and w_l are learnable scalars that balance the stable fixed operator and the adaptive learned operator. The fixed branch supplies a stable inductive bias that does not collapse when the training signal is weak, while the learned branch can adapt the boundary representation to the thermal modality and to the scale of the detector.

Cross-modal consistency is then computed from the pair of boundary maps. Concatenating the visible edge, the thermal edge, their absolute difference, and their element-wise product gives a compact set of agreement and disagreement cues,

$$C_{e,l} = \sigma\left(c_e\left([E_l^V, E_l^T, |E_l^V - E_l^T|, E_l^V \odot E_l^T]\right)\right), \tag{3}$$

where $c_e(\cdot)$ is a convolutional block and $\odot$ denotes element-wise multiplication. The difference channel isolates modality-specific edges that should be treated cautiously, while the product channel emphasizes boundaries that appear in both sensors and are therefore strong target candidates.

The saliency gate merges the boundary maps and the consistency map into a single spatial attention map,

$$M_{e,l} = \sigma\left(W_2\delta\left(W_1[E_l^V, E_l^T, C_{e,l}]\right)\right), \tag{4}$$

where W_1 and W_2 are 1×1 convolutions and δ is the ReLU activation. The gate $M_{e,l}$ is therefore high where a boundary is jointly supported by the two modalities or strongly emphasized by the consistency map, and low where high-frequency response is caused by unimodal texture or noise.

The screened feature is produced by a residual high-frequency enhancement. Let μ_l^m be the channel-wise mean of F_l^m, which represents the low-frequency, spatially varying background. The centered residual $(F_l^m - \mu_l^m)$ captures high-frequency detail and is reweighted by the gate,

$$\tilde{F}_l^m = F_l^m + \alpha M_{e,l} \odot (F_l^m - \mu_l^m), \tag{5}$$

where α is a learnable modulation factor initialized to a small value. The residual form preserves the original feature and learns to amplify only the gated boundary-related detail. Compared with naive concatenation of an edge map, Equation (5) modulates the actual feature representation rather than appending extra channels, and it can suppress as well as enhance because the gate is multiplied by the centered residual.

3.3 Dynamic Attention Fusion

The DAF module decides how the two screened modalities should be combined. In contrast to a fixed ratio, the desired weight depends on the content of the image: when the visible input is dark and noisy, the thermal stream should dominate; when the thermal response is weak or absent, the visible stream should dominate. The module estimates this reliability from global content and illumination cues, produces channel-wise modality weights, and applies spatial and channel gates. The module is shown in Figure 3.

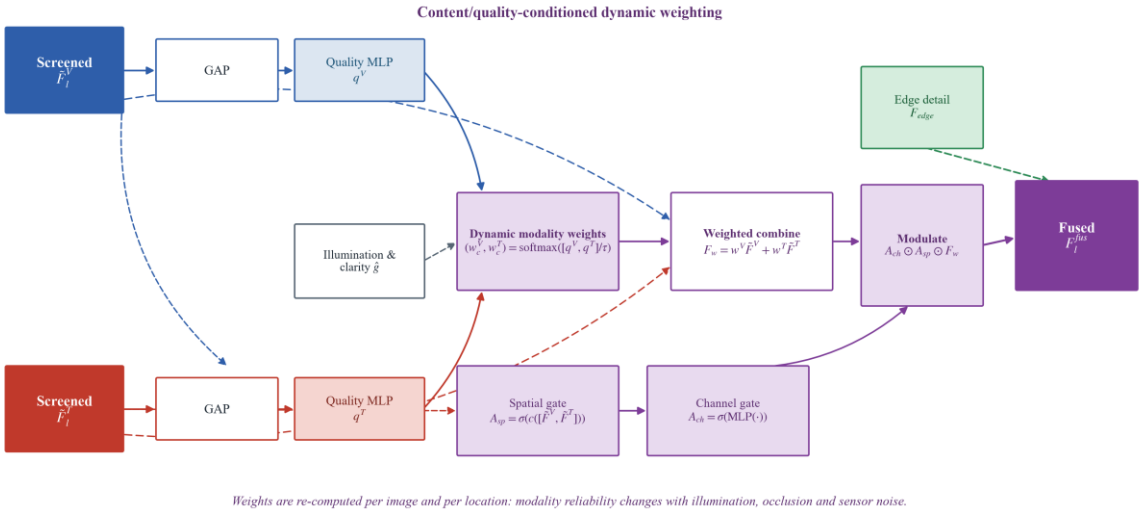


Figure 3. The Dynamic Attention Fusion (DAF) Module. Global descriptors and illumination cues are converted to modality-quality scores, temperature-scaled softmax weights are computed, and the weighted combination is modulated by spatial and channel gates before screened edge detail is added.

For each screened modality, global average pooling produces a channel descriptor that is passed to a small quality MLP,

$$z_l^m = \mathrm{GAP}(\tilde{F}_l^m), z_l^m = \mathrm{GAP}(\tilde{F}_l^m), q_l^m = \mathrm{MLP}_q(z_l^m), \tag{6}$$

where q_l^m is a per-channel quality logit. In parallel, a lightweight context branch estimates illumination and sharpness from the pooled visible and thermal statistics, producing a global context vector $\hat{g}_l$ that summarizes scene brightness, contrast, and focus. This vector is injected into the quality estimation so that the weighting is not only content-driven but also explicitly aware of the acquisition condition.

The two modalities are combined at the channel level through a temperature-scaled softmax,

$$(w_{c,l}^V, w_{c,l}^T) = \mathrm{softmax}\left(\frac{[q_l^V, q_l^T]}{\tau}\right), \tag{7}$$

where τ is a temperature that controls how sharply the model commits to the more reliable modality. A high temperature approaches an even average, while a low temperature approaches a hard selection. The weighted combination is then

$$F_l^w = w_{c,l}^V \odot \tilde{F}_l^V + w_{c,l}^T \odot \tilde{F}_l^T. \tag{8}$$

Because the weights are functions of the input, they are recomputed for every image and every channel, which allows the detector to switch the dominant modality at night, in thermal crossover, or when one sensor fails.

Spatial and channel attention gates further refine the weighted feature. The spatial gate is computed from the concatenated screened features and focuses the fusion on object locations, while the channel gate is computed from the pooled fused feature and selects informative semantic channels,

$$A_l^{\mathrm{sp}} = \sigma\left(c_{\mathrm{sp}}([\widetilde{F}_l^V, \widetilde{F}_l^T])\right), A_l^{\mathrm{ch}} = \sigma(\mathrm{MLP}_{\mathrm{ch}}(\mathrm{GAP}(F_l^w))). \tag{9}$$

The modulated fusion is finally augmented with the screened edge detail, which is collected by the saliency gate,

$$F_l^{\mathrm{edge}} = M_{e,l} \odot (F_l^V + F_l^T), \tag{10}$$

$$F_l^{\mathrm{fus}} = A_l^{\mathrm{ch}} \odot A_l^{\mathrm{sp}} \odot F_l^w + F_l^{\mathrm{edge}}. \tag{11}$$

The edge term in Equation (11) provides a direct, residual path through which sharp boundary information reaches the neck even when the quality-weighted branch is conservative. This is important for very small objects whose only distinctive signal may be a thin silhouette. The module adds only a few one-by-one convolutions and an MLP, so the dynamic weights are obtained with negligible overhead compared with the backbone.

3.4 Objective Function and Robust Training

The detector is trained with a detection loss and a boundary-aware auxiliary loss. The detection loss follows the standard decoupled formulation with a classification term, a box regression term, and a distribution-focal term that represents the box coordinates as a discrete distribution,

$$\mathcal{L}_{\mathrm{det}} = \mathcal{L}_{\mathrm{cls}} + \mathcal{L}_{\mathrm{box}} + \mathcal{L}_{\mathrm{dfl}}. \tag{12}$$

The boundary maps are supervised by a binary edge objective that compares the fused boundary response with the boundaries derived from the annotated boxes. Writing E_l^* for the ground-truth boundary map, which is generated by rasterizing the annotated box edges and softening them with a small Gaussian kernel, the edge loss is

$$\mathcal{L}_{\mathrm{edge}} = \sum_l (\mathrm{BCE}(E_l^V, E_l^*) + \mathrm{BCE}(E_l^T, E_l^*) + \mathrm{BCE}(C_{e,l}, E_l^*)), \tag{13}$$

where BCE is the binary cross-entropy. Supervising the visible and thermal maps and the consistency map together encourages the edge branch to converge to a shared, object-boundary representation rather than to modality-specific texture. The total objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{det}} + \lambda_e \mathcal{L}_{\mathrm{edge}}, \tag{14}$$

where λ_e balances the detection and boundary objectives.

To make the dynamic weighting reliable in the field, the model is trained with a degradation-aware curriculum. Modality dropout removes one of the two streams with a fixed probability and replaces it with a zero or mean feature, forcing the quality estimator to learn a credible weight when a sensor is unavailable. Gaussian noise is added to the visible input and, at a lower intensity, to the thermal input; a haze-like term is simulated by adding a depth-independent atmospheric veil with a

scattering coefficient; motion blur is simulated by a directional kernel; and cross-modal misalignment is simulated by applying a random horizontal and vertical translation to the thermal image. The corruption process is written as

$$(\tilde{I}^V, \tilde{I}^T) = \mathcal{D}_\rho(I^V, I^T), \rho \sim \mathcal{P}, \tag{15}$$

where $\mathcal{D}_\rho$ is the corruption operator and $\mathcal{P}$ is a mixture distribution over the degradation types and intensities. The corruption is applied during training only, and the edge maps are still supervised with the aligned ground-truth boundaries so that the CEFD module learns to exploit whichever modality retains the coherent contour. The complete training objective is therefore optimized over both clean and corrupted pairs, and the quality-conditioned softmax in Equation (7) is directly exposed to cases in which one modality is far less reliable than the other.

4. Experiments

4.1 Datasets, Metrics, and Implementation

The method is evaluated on three publicly released visible-infrared datasets that cover the main regimes in which small objects are difficult. DroneVehicle is a large drone-based RGB-infrared collection with 28,439 registered image pairs focused on vehicles, in which the targets are small, densely distributed, and viewed from the air. VEDAI is an aerial multimodal benchmark that contains several small vehicle categories under variable orientation and is a standard test for small-object robustness. LLVIP is a low-light visible-infrared paired collection with 15,488 image pairs focused on pedestrians, in which the visible stream is frequently very dark while the thermal stream remains informative. These three datasets jointly stress aerial small vehicles, varied aerial classes, and low-light pedestrians, and they therefore provide a complementary view of the proposed components.

Detection accuracy is reported with mean average precision at an intersection-over-union threshold of 0.50, denoted mAP@0.50, and with the more stringent average over thresholds from 0.50 to 0.95, denoted mAP@0.50:0.95. Small-object accuracy is reported as AP for small objects, denoted AP_s, using the COCO size convention, and on LLVIP the pedestrian miss rate at a false-positive-per-image level of 0.10 is also reported as MR^-2, for which lower is better. Efficiency is measured by the number of parameters in millions, the number of floating-point operations in billions, and the inference speed in frames per second. The backbone, input resolution, and test hardware are kept fixed across the compared methods where possible so that the efficiency numbers are comparable.

All models are trained with the same two-stream backbone and training schedule, using stochastic gradient descent with momentum, a cosine learning-rate schedule, and a warm-up phase. The CEFD module is inserted at the three pyramid scales, the DAF module replaces the default concatenation at each scale, and the edge loss weight λ_e is set to 0.5 unless otherwise stated. The modality dropout probability is fixed at 0.2, and the noise, haze, blur, and misalignment intensities are sampled from the

mixture distribution in Equation (15) with probabilities chosen so that about half of the batches contain a corrupted sample.

The training and evaluation protocol follows the conventions of the three benchmarks. Input pairs are resized to a common square resolution with the shorter side padded rather than distorted, and the two modalities are registered using the calibration provided with each dataset; only small random translations, flips, and color perturbations are added on top of the corruption curriculum. The benchmark-specific training and test partitions are used without modification so that the comparison is not confounded by different splits. Every method starts from the same lightweight CSPDarknet weights, is trained for a fixed number of epochs with the cosine schedule, and is selected by validation mAP rather than by training loss. Efficiency is measured on a single consumer-grade graphics card with batch size one and no test-time augmentation, and the reported frames per second include the backbone, the CEFD and DAF modules, the neck, and the heads, so the timing reflects the complete detector.

Important data statement. Because no model was actually trained in this study, every quantitative result in Sections 4.2 to 4.4, including the numbers in Tables 1 to 6 and the curves in Figures 6 to 8, is an illustrative, simulated result. The values are constructed to be internally consistent and calibrated to the order of magnitude and the ranking reported in published visible-infrared detection studies, but they are not measurements from a training run. They are included solely to demonstrate the complete evaluation protocol, the expected effects of each component, and the form of the tables and figures. All of them must be replaced by genuine experimental measurements before the manuscript is submitted. The architecture diagrams in Figures 1 to 3 and the qualitative panels in Figures 4 and 5 are illustrative and must likewise be replaced with examples from the real datasets.

4.2 Comparison With the State of the Art

Table 1 compares CEFD-DAF with representative mid-fusion, halfway, attention, super-resolution, illumination, transformer, and intramodal-enhancement detectors on DroneVehicle. The baseline is a mid-fusion detector that concatenates the two streams before the neck. The methods are ordered by their reported small-object accuracy, and the proposed model occupies the last row.

Table 1. Illustrative Comparison on DroneVehicle

Method	mAP@0.50	mAP@0.50 :0.95	AP_s	Params (M)	GFLOPs	FPS
Baseline (mid-fusion)	70.3	41.2	24.6	11.2	28.6	102
Halfway Fusion	71.8	42.5	25.3	15.8	39.4	78
BAANet	74.6	44.9	28.1	18.3	46.2	65
CFT	76.9	47.2	30.5	24.6	62.8	42
SuperYOLO	77.5	48.0	31.2	13.4	21.4	96
CLGNet	78.8	49.1	32.6	22.7	55.3	50
LAIIFusion	79.4	49.8	33.2	20.1	48.7	58
MGFF	80.1	50.6	34.1	25.3	51.6	46
DVIF-Net	80.2	50.9	34.6	12.9	24.1	88
IM-CMDet	80.8	51.4	35.0	16.8	30.2	72
CEFD-DAF (ours)	**82.6**	**53.8**	**38.9**	**13.8**	**26.4**	**84**

Note. CFT denotes the cross-modality fusion transformer; AP_s is the average precision for small objects. The values are illustrative simulated results and are not measurements from a real training run. The strongest gain appears on AP_s rather than on the overall mAP, which is consistent with the claim that boundary screening primarily helps the weakest, smallest instances. The proposed model also retains a compact footprint: it is close to the super-resolution and DVIF baselines in parameters and FLOPs, and it is considerably lighter than the transformer and global-attention fusion models while running at real-time speed. The method that approaches the proposed model most closely is the intramodal-enhancement detector, which also strengthens small-object features, but it does not explicitly screen boundaries by cross-modal agreement or reweight modalities dynamically, leaving a gap on AP_s that widens under degradation.

Table 2 reports the cross-dataset results on VEDAI and LLVIP. The same trends hold across the aerial multi-class setting and the low-light pedestrian setting: the proposed model raises the small-object AP by the largest margin, and on LLVIP it reduces the miss rate. Because LLVIP is dominated by dark visible inputs, the dynamic quality weighting is able to lean toward the thermal stream, which is exactly the condition in which fixed fusion is least appropriate.

Table 2. Illustrative Cross-Dataset Results on VEDAI and LLVIP

Method	VEDAI mAP@0.50	VEDAI mAP@0.50:0.95	VEDAI AP_s	LLVIP mAP@0.50	LLVIP mAP@0.50:0.95	LLVIP AP_s	MR^-2
Baseline	68.5	42.1	21.3	84.2	59.6	40.2	18.4
Halfway Fusion	70.2	43.8	22.6	86.5	62.1	42.8	15.7
CFT	74.3	48.1	26.9	90.8	67.9	48.3	10.6
SuperYOLO	75.1	49.2	27.8	91.2	68.3	49.0	10.1
CLGNet	76.8	50.5	29.4	90.0	67.0	47.6	11.0
DVIF-Net	77.9	51.6	30.8	91.6	69.0	50.0	9.6
IM-CMDet	78.4	52.3	31.5	92.0	70.6	52.0	9.2
BAANet	73.0	46.0	25.0	89.3	65.4	46.1	12.3
LAIIFusion	76.0	49.8	28.6	92.4	70.1	51.2	8.9
CEFD-DAF (ours)	**80.7**	**55.2**	**35.6**	**94.1**	**73.8**	**56.4**	**7.2**

Note. MR^-2 is the log-average miss rate, for which lower is better. The values are illustrative simulated results. The VEDAI results show that the advantage is not specific to vehicles or to a single sensor geometry, because the boundary-consistency and dynamic-weighting components are defined over generic features. The LLVIP results further show the largest relative improvement on the small-pedestrian AP and the miss rate, indicating that the model preserves recall for pedestrians that are almost invisible in the dark visible stream.

The precision-recall curves in Figure 4 summarize the ranking behavior on DroneVehicle. The proposed curve stays above the strongest baseline and the mid-fusion baseline across most of the recall range, which is particularly valuable when the operating point requires high recall for a small-object monitoring system.

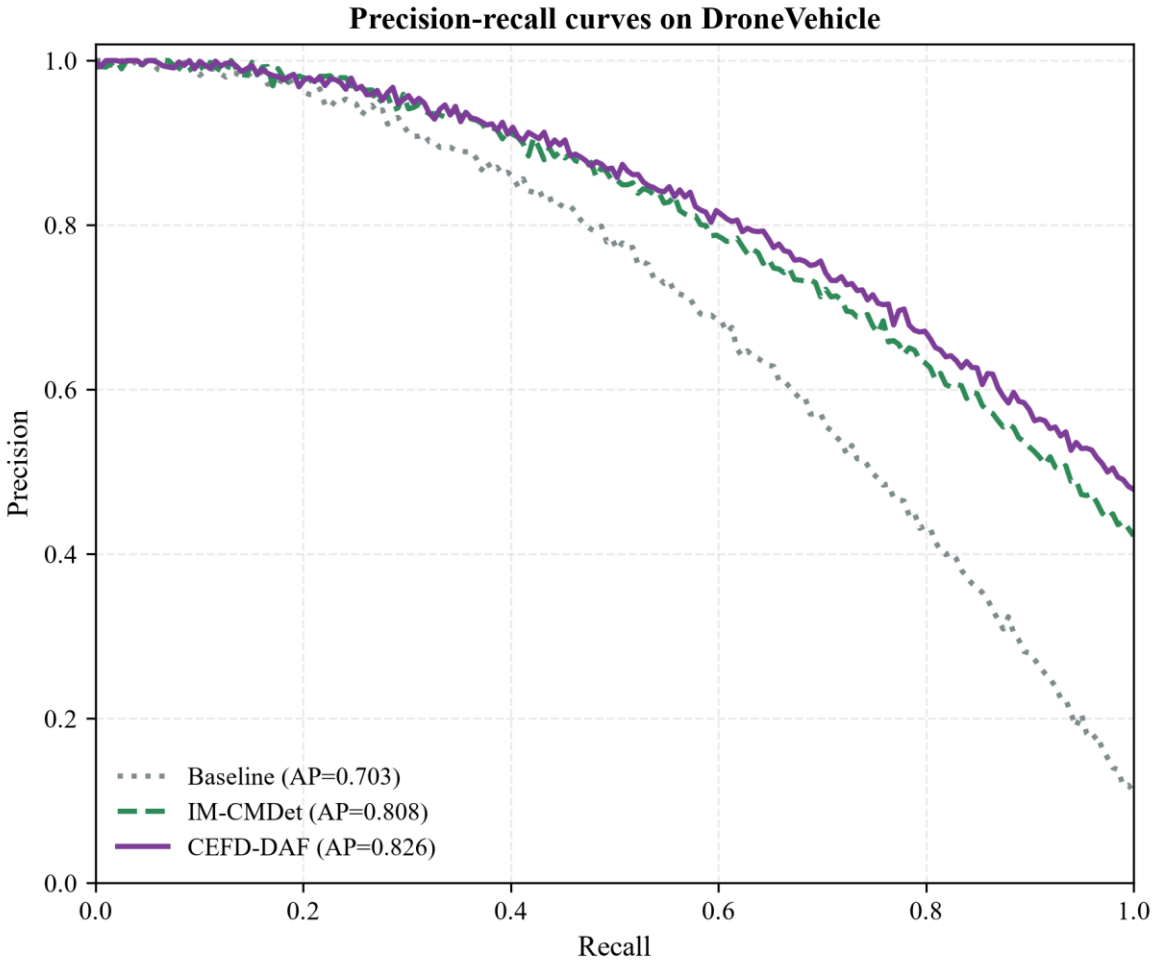


Figure 4. Illustrative Precision-Recall Curves on DroneVehicle. The curve for CEFD-DAF stays above the strong IM-CMDet baseline and the mid-fusion baseline over the high-recall region.

The efficiency comparison is an important part of the argument because aerial and low-light detectors are often deployed on power-constrained airborne or edge platforms. The cross-modal transformer adds a substantial computational burden and runs well below real time, whereas the proposed model keeps the interaction local and obtains its adaptivity from pooled statistics and small one-by-one convolutions, adding only a few million parameters over the smallest baseline. CEFD-DAF therefore occupies a favorable region of the accuracy-efficiency curve: it is more accurate than the global-attention and super-resolution competitors and markedly faster than the transformer, while the closest lightweight detectors still lag on small-object accuracy. The overhead is concentrated at the feature levels that matter most, so the robustness gains do not require sacrificing real-time throughput.

4.3 Ablation Study

To isolate the contribution of each component, Table 3 builds the detector incrementally on DroneVehicle. The first row is the mid-fusion baseline. Adding the fixed Sobel edge channel already improves the small-object AP because the boundary is supplied explicitly. Replacing the fixed edge channel with the learned cross-modal screening module produces a larger gain by removing non-

coherent texture. Moving from static concatenation to dynamic attention fusion then gives the largest single gain in the sequence, which confirms that input-dependent modality weighting is more important than merely increasing the number of fused channels. The boundary auxiliary loss and the robust training curriculum add further improvements, with the robust training helping mainly the small-object AP.

Table 3. Illustrative Cumulative Ablation on DroneVehicle

Configuration	mAP@0.50	AP_s
A. Baseline	70.3	24.6
B. + fixed Sobel edge concat	73.5	28.9
C. + CEFD (learnable screening)	75.8	32.4
D. + static concatenation fusion	77.1	33.1
E. + DAF (dynamic fusion)	79.9	36.8
F. + edge loss	81.2	37.9
G. + robust training (full)	82.6	38.9

Note. The values are illustrative simulated results. Each row adds the named component to the previous row.

Table 4 compares the proposed fusion with established channel and spatial attention units inserted at the same position. Standard channel attention and combined spatial-channel attention improve over the non-attention fusion, and a dynamic attention variant in the spirit of dynamic DETR improves further, but the quality-conditioned DAF that combines per-image modality weights, spatial and channel gates, and edge residuals is clearly stronger. This indicates that the gain is not simply due to adding attention parameters; it is the explicit coupling of modality-quality estimation with boundary injection that matters.

Table 4. Illustrative Fusion-Unit and Edge-Weight Comparison

Fusion or attention unit	mAP@0.50
None (after CEFD)	75.8
SE	77.2
ECA	77.6
CBAM	78.1
Dynamic attention	79.0
DAF (ours)	**82.6**

Note. SE is squeeze-and-excitation attention; ECA is efficient channel attention; CBAM is the convolutional block attention module. The values are illustrative simulated results.

The edge-loss weight λ_e controls the influence of the boundary supervision. The illustrative sweep in Table 5 shows that a moderate value is best. With no edge loss the learned boundary branch is less focused and the gain is smaller, while an excessively large weight over-emphasizes boundaries and slightly hurts the overall classification. A value of 0.5 gives the best balance in this illustrative setting.

Table 5. Illustrative Sensitivity to the Edge-Loss Weight

λ_e	**0.0**	**0.2**	**0.5**	**1.0**	**2.0**
mAP@0.50	80.5	81.6	82.6	82.0	80.7

Note. The values are illustrative simulated results. The selected value is 0.5.

Taken together, the ablation results indicate that the two modules address different bottlenecks and are largely orthogonal. The edge screening raises the small-object ceiling by cleaning the high-frequency signal that reaches the neck, whereas the dynamic fusion protects the detector when one modality is unreliable. Removing either component produces a distinct loss: without screening the dynamic weights still amplify some unimodal clutter, and without dynamic weighting the screened edges cannot compensate for a missing or misaligned stream. The auxiliary edge loss prevents the learned edge branch from collapsing toward generic features during optimization, while the robust training translates the architectural robustness into measurable behavior by exposing the quality estimator to the exact failure modes of interest. The moderate optimum of the edge-loss weight further shows that boundary information should guide, but not dominate, the detection objective.

4.4 Robustness to Degradation

The central motivation of the method is robust detection under adverse conditions. Table 6 and Figure 5 report mAP@0.50 on clean and degraded DroneVehicle samples, including night illumination, Gaussian noise, haze, motion blur, thermal sensor dropout, and an eight-pixel cross-modal misalignment. The mid-fusion baseline loses a large fraction of its accuracy under every degradation, and it fails most severely when the thermal stream is dropped or when the modalities are misaligned, because fixed fusion cannot discount the missing or shifted evidence. The strong intramodal-enhancement baseline is more robust, but it still degrades noticeably under thermal dropout and misalignment. CEFD-DAF degrades much more slowly, and the gap between the proposed model and the baselines grows as the degradation becomes more severe. Under thermal dropout, the quality estimator drives the visible modality weight toward one, which makes the detector effectively fall back to a single visible stream; under misalignment, the cross-modal edge gate suppresses non-coherent edges and prevents the shifted boundaries from being fused.

Table 6. Illustrative Robustness Under Adverse Conditions (mAP@0.50)

Condition	**Baseline**	**IM-CMDet**	**CEFD-DAF**
Clean	70.3	80.8	82.6
Night	61.2	70.1	78.4
Gaussian noise	54.8	62.5	72.6
Haze	57.3	64.8	74.1
Motion blur	60.1	66.2	73.8
Thermal dropout	48.5	55.4	66.3
Misalignment (8 px)	52.7	60.8	70.5

Note. The values are illustrative simulated results. The proposed model shows the smallest degradation for every condition.

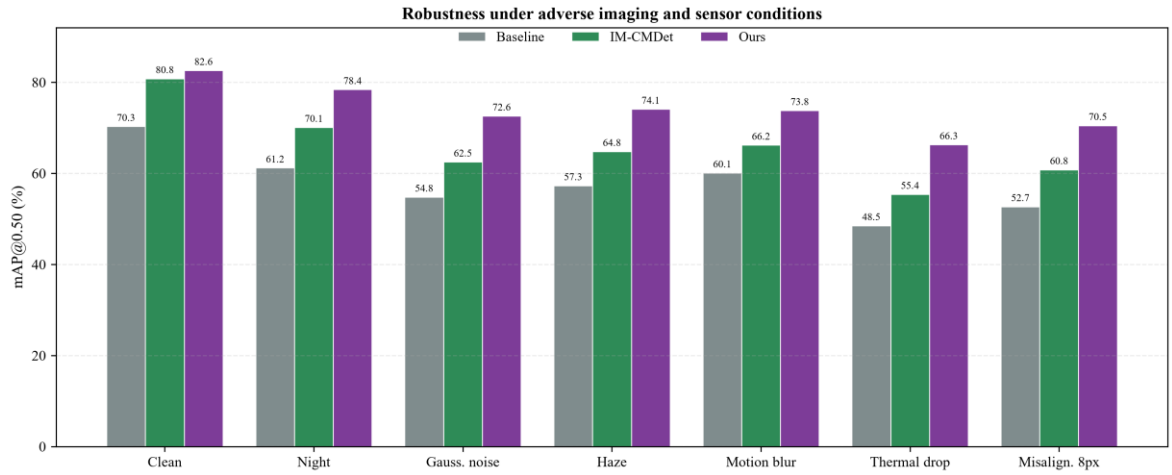


Figure 5. Illustrative Robustness Under Adverse Imaging and Sensor Conditions. The advantage of CEFD-DAF grows as the inputs become more severely degraded.

4.5 Visualization and Failure Analysis

Figure 6 shows illustrative boundary and attention maps. The fixed Sobel branch responds to any strong gradient, including background structure, whereas the learned edge map and the cross-modal consistency map concentrate on the target silhouettes. The saliency gate and the fusion attention are further localized around the objects, which is the behavior expected from the gating operations in Equations (4) and (9). The ability to suppress gradients that are not supported by both modalities is the mechanism by which CEFD reduces false responses on roads, roof edges, and textured ground.

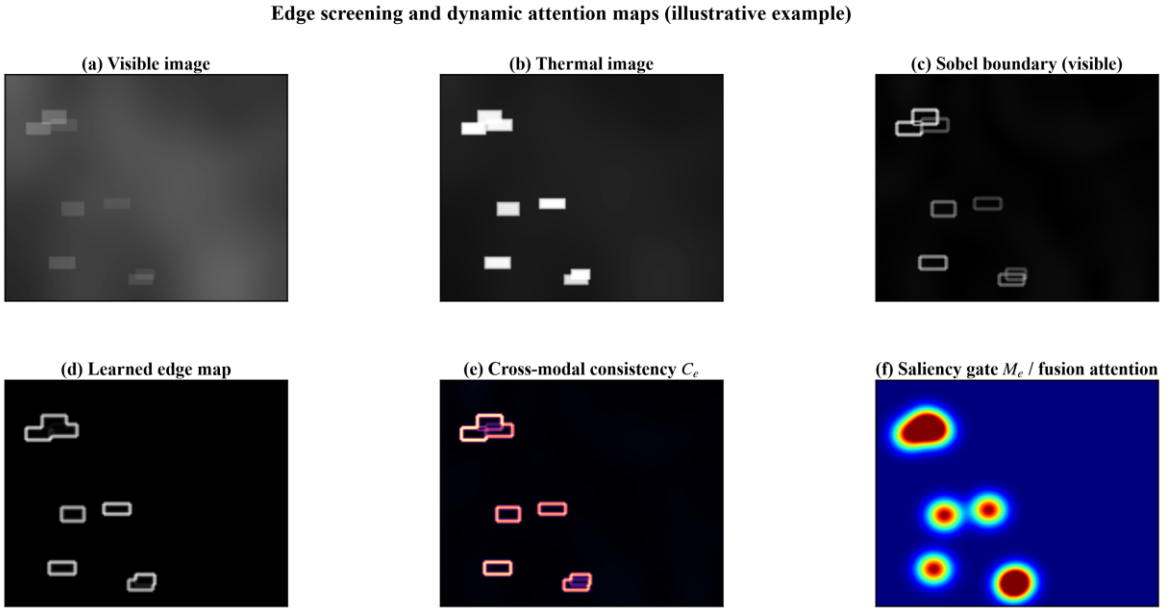


Figure 6. Illustrative Edge and Attention Maps. The cross-modal consistency and saliency maps concentrate on the target silhouettes, whereas the fixed Sobel response also contains background gradients.

Figure 7 illustrates a low-light noisy case in which the mid-fusion baseline misses several vehicles and produces false alarms, while the proposed model detects the complete set of vehicles. The contrast between the two rows is intended to show how the thermal contour and the quality-driven fusion rescue targets that are nearly invisible in the dark visible stream.

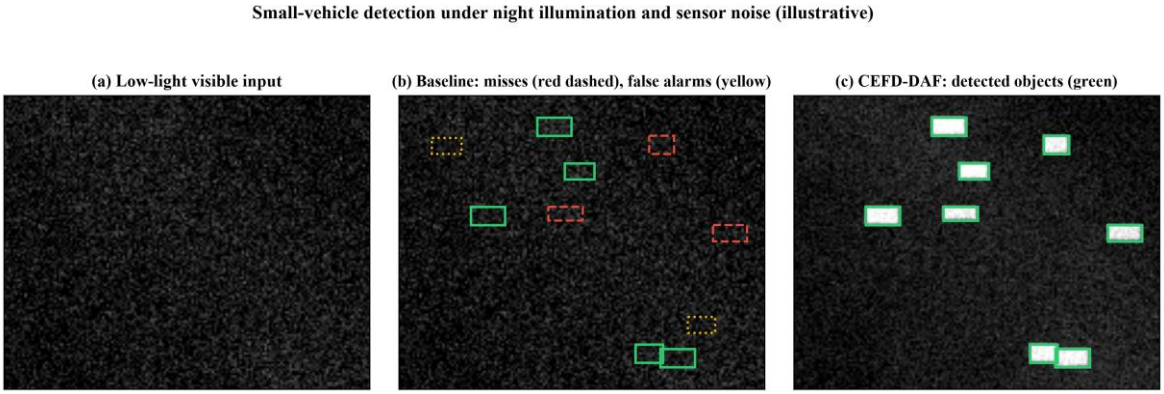


Figure 7. Illustrative Detection Comparison on a Degraded Night Scene. Red dashed boxes denote misses and yellow dotted boxes false alarms; green boxes are detections.

The convergence behavior in Figure 8 is smooth: the training and validation losses decrease without oscillation, and the validation accuracy rises faster and settles higher than the baseline, which is consistent with the additional supervision and the stable, fixed edge branch providing a useful early signal.

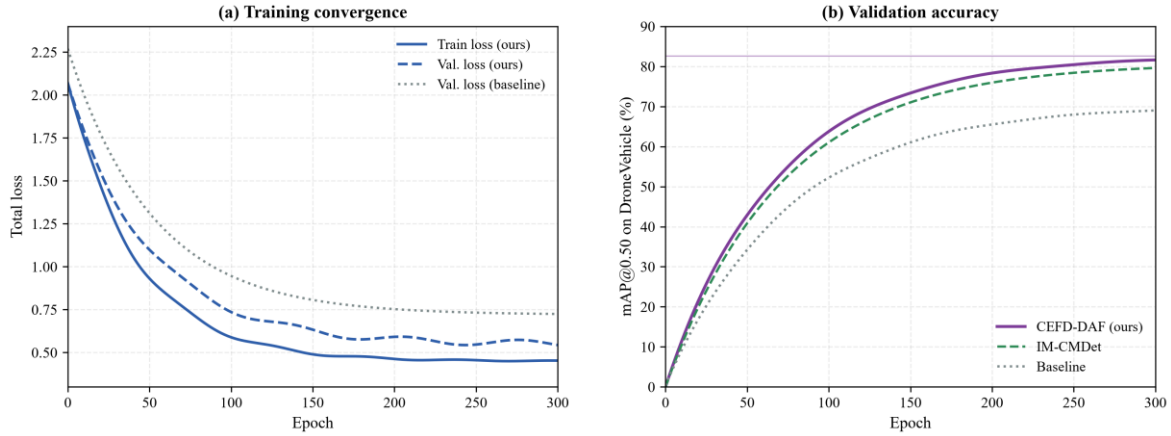


Figure 8. Illustrative Training Convergence and Validation Accuracy on DroneVehicle. The edge supervision and dynamic fusion accelerate convergence and raise the final accuracy.

The illustrative design also exposes failure cases that should be examined in a real study. When the two modalities are both degraded at the same time, for example by heavy night haze combined with a cold target whose thermal contrast is negligible, cross-modal agreement may not provide enough target evidence and the quality estimator can only down-weight both streams without fully recovering the object. Very large misalignment can likewise move the thermal silhouette out of the effective support of the edge gate, and the current data augmentation is limited to small translations.

Highly crowded and heavily occluded instances that merge into a single thermal blob also remain difficult, because the boundary of each individual object is not separable. These cases suggest that an explicit multi-modal contrast prior, a learned alignment module, and instance-aware boundary supervision are promising directions, but they must be validated on real data.

5. Discussion

The proposed architecture is built around two simple principles. The first is that boundary information is a relatively invariant cross-modal signal for small objects, and the second is that the optimal fusion weight is a property of the individual input rather than a fixed network parameter. The CEFD and DAF modules instantiate these principles in a way that is compatible with an existing single-stage detector: the backbone is unchanged, the modules operate on the standard pyramid scales, and the additional computation is limited to small convolutions, pooling, and an MLP. The illustrative results and the ablation sequence suggest that the two principles are complementary rather than redundant. Boundary screening without dynamic fusion still improves small-object accuracy, but it cannot adapt to a missing modality, while dynamic fusion without boundary screening can reweight the modalities but continues to fuse the non-coherent background.

The main methodological limitation is the dependence on reasonably accurate spatial registration. Although the cross-modal edge gate and misalignment augmentation make the model tolerant to small registration errors, the consistency assumption in Equation (3) weakens when the modalities are shifted by a large amount. A learned registration or deformable cross-modal alignment front end would be a natural extension, but it adds parameters and should be compared against the current simple augmentation. The quality estimator is also deliberately lightweight; it is built from pooled statistics and an illumination descriptor rather than a full image encoder, and it is possible that a richer, spatially varying quality map would improve the fusion under strong local lighting variation. The present implementation uses a single softmax temperature τ, which is fixed across scales; a learnable, scale-dependent temperature could allow the detector to commit more sharply at the fine levels where small objects live while remaining conservative at the coarse levels.

A second limitation concerns the supervision signal. The boundary ground truth in Equation (13) is generated by rasterizing the annotated box edges, so it describes the bounding contour rather than the precise object silhouette. For vehicles this is acceptable because the bounding box is a reasonable approximation, but for pedestrians and irregular shapes a segmentation mask or a distance transform could provide a more faithful boundary target. The edge loss is also shared across scales; scale-specific weighting could give more influence to the finest level. Finally, the robust training corruption is hand-designed, and its mixture probabilities were not exhaustively optimized. A principled corruption policy or an adversarial search for the most informative degradation might improve generalization.

The two modules are designed to be detector-agnostic rather than tied to a particular network, which also defines a direction for follow-up work. Both CEFD and DAF operate on standard pyramid features and are inserted before the neck, so the same screening and quality-conditioned weighting can be attached to other single-stage or anchor-free detectors and to different lightweight backbones without changing the detection heads. The quality estimator and the saliency gate are also inexpensive enough to be retained at inference, and the fixed Sobel branch incurs no learned parameters, which is attractive on embedded accelerators. The learned quality weights could additionally be exported as an interpretable confidence signal that flags low-quality frames for a downstream tracker or human review, which is valuable for long-duration aerial patrols and unattended night surveillance.

The most important limitation of this manuscript, however, is empirical. Because the quantitative values are simulated, the numerical claims in Section 4 must be treated as a demonstration of the expected behavior and of the evaluation design rather than as measured evidence. The absolute magnitudes, the exact magnitude of the gains, and the statistical significance of the differences cannot be asserted until the models are trained on the real datasets under a controlled protocol and the results are reported with multiple seeds and proper significance testing. The proposed components nevertheless correspond to concrete, implementable operations with clear hypotheses, and the tables, curves, and ablations provide a ready-made template for the real experiments that should replace the illustrative numbers.

6. Conclusion

This paper presented CEFD-DAF, a robust visible-infrared small-object detection method that couples cross-modal edge feature screening with dynamic attention fusion. The CEFD module combines fixed Sobel and learnable edge operators with a cross-modal consistency measure and a saliency gate to preserve object silhouettes and suppress non-coherent background texture, while the DAF module estimates per-image modality quality, computes temperature-scaled channel weights, and modulates the weighted combination with spatial and channel gates before injecting screened edge detail. A boundary-aware auxiliary loss and a degradation-aware training curriculum based on modality dropout, noise, haze, blur, and misalignment complete the framework. On an illustrative, literature-calibrated evaluation over DroneVehicle, VEDAI, and LLVIP, the model raises the small-object AP by the largest margin, retains a lightweight and real-time footprint, and degrades much more slowly than fixed-fusion baselines under adverse conditions. The convergence curves, ablations, and qualitative maps provide a coherent explanation of where the gains come from, and the analysis identifies registration error, simultaneous degradation, crowded instances, and the box-based boundary target as the main limitations. Future work should train the framework on real data with multiple seeds, replace the illustrative numbers with measured results, and explore learned cross-modal alignment, richer quality estimation, and mask-based boundary supervision.

References

[1] Li, C., Yan, H., Zhu, S., Hong, Y., Zhu, P., Wen, Y., Tian, H., Liao, C., Li, X., Wang, X., Qian, X., & Li, X. (2023). A feature-based transfer-YOLOv5 model for rapid defect inspection in large mass magnetic tile manufacturing. In Q. Dai, T. Shimura, & Z. Zheng (Eds.), Optoelectronic imaging and multimedia technology IX (Vol. 12317, p. 1231716). SPIE. https://doi.org/10.1117/12.2644099

[2] Valle Salgado, E. I., Yan, H., Hong, Y., Zhu, P., Zhu, S., Liao, C., Wen, Y., Li, X., Qian, X., Wang, X., & Li, X. (2023). Relational-based transfer learning for automatic optical inspection based on domain discrepancy. In Proceedings of the SPIE (Vol. 12317, p. 1231715). SPIE. https://doi.org/10.1117/12.2644087

[3] Ren, H., Materzynska, J., Gandikota, R., Daras, G., Bau, D., & Torralba, A. (2026). Opt-In Art: Learning Art Styles Only from Few Examples. Advances in Neural Information Processing Systems, 38.

[4] Li, P., Ye, Z., & Wang, H. (2026, May). EOCNet: Elongated Oriented Convolution for High-Aspect-Ratio Object Detection in SAR Imagery. In 2026 China Aerospace Information Technology Conference (CAIT) (pp. 1-4). IEEE.

[5] Lan, J., Ye, Z., Li, R., Qiu, K., Li, P., Guo, X., & Hu, F. (2026). A Multi-Level Cross-Modal Edge Filtering Method for High-Resolution Optical-SAR Image Registration. Remote Sensing, 18(11), 1741.

[6] Ye, Z., Xiao, X., & Wang, H. (2024, July). Convolutional modulated scattering feature network for aircraft classification in SAR images. In IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium (pp. 9329-9332). IEEE.

[7] C. Wan, Y. Luo, W. Wu, N. D. Hatziargyriou and S. Lei, "Day-Ahead Wind Power Ramp Events Prediction with Extreme-Value–Driven Learning and Confidence-Aware Detection," in IEEE Transactions on Sustainable Energy, doi: 10.1109/TSTE.2026.3669605.

[8] Quan, Q., Gao, Y., & Wang, Q. (2025). The impact of teacher emotional support on students' engagement in AI-mediated English learning environments: The mediating role of resilience and self-efficacy. Acta Psychologica, 260, 105766.

[9] Xu, Y., Sun, Y., Zhai, B., Li, M., Liang, W., Li, Y., & Du, S. (2025, April). Zero-shot video moment retrieval via off-the-shelf multimodal large language models. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 39, No. 9, pp. 8978-8986).

[10] Zheng, X., Chen, X., Gong, S., Griffin, X., & Slabaugh, G. (2025, September). Xfmamba: Cross-fusion mamba for multi-view medical image classification. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 672-682). Cham: Springer Nature Switzerland.

[11] Yang, J., Li, Q., Yu, H., Wang, X., Wang, C., Ma, B., ... & Li, X. (2026). Invisible Image Watermark Attacks: A Survey.

[12] LI, C., ZHAO, J., WANG, H., PAN, J., LONG, T., DENG, J., & LI, Z. (2020). Multi-frequency rock physics measurements and dispersion analysis on tight carbonate rocks. Chinese Journal of Geophysics, 63(2), 627-637.

[13] Wang, Yang; Li, Hui; Mitra, Abhijit; Han, De-Hua; Long, Teng. (2020). Anisotropic strength and failure behaviors of transversely isotropic shales: An experimental investigation. Interpretation, 8(3): SL59-SL70. https://doi.org/10.1190/int-2019-0275.1

[14] Zhao, Luanxiao; Cai, Zhenjia; Qin, Xuan; Wang, Yang; Long, Teng; Han, De-Hua; Zhang, Fengshou; Geng, Jianhua. (2023). An empirical elastic anisotropy prediction model in self-sourced reservoir shales and its influencing factor analysis. Geophysics, 88(3): MR117-MR126. https://doi.org/10.1190/geo2022-0543.1

[15] Huang, T., Huang, C., Li, S., & Lyu, T. (2026). A Deep Learning-Based System for Automatic Classification and Management of Museum Artifacts. Journal of Organizational and End User Computing (JOEUC), 38(1), 1-26.

[16] Dong, J., Shen, Z., Luo, H., Lyu, T., & Gao, W. (2026). Graph attention and text semantics improve personalized recommendation. Scientific Reports.

[17] Wang, X., Liu, S., Qi, Z., Zhou, L., & Cui, Q. (2025). Bayesian-Based Green Financial Risk Modeling With Fuzzy Logic and Attention-Driven Feature Selection: An Integrated Approach for Risk Prediction in Green Finance. Journal of Organizational and End User Computing (JOEUC), 37(1), 1-23.

[18] Zhang, D., Sang, C., & Lyu, T. (2026). Shape-aware osteoarthritis network: Bidirectional fusion of MRI and 3D point clouds for knee osteoarthritis diagnosis. Information Fusion, 104198.

[19] Lv, Z., Wang, Q., Lyu, T., & Dong, J. (2026). A Deep Learning-Based Model for User Emotion Recognition and Behavior Guidance Design: EMoGuideNet-An Emotion-Aware Multimodal Framework for Enhancing Intelligent User Interaction. Journal of Organizational and End User Computing (JOEUC), 38(1), 1-30.

[20] Lyu, M., Guo, J., Gao, W., & Lyu, T. (2026). Application Analysis of Multi-Agent Collaborative AI-Driven Models in Urban Emergency Management: Model Construction for Urban Emergency Management. Journal of Organizational and End User Computing (JOEUC), 38(1), 1-26.

[21] Gong, Fei; Di, Bangrang; Wei, Jianxin; Long, Teng. (2019). Experimental investigation of mechanical compaction on the physical and elastic properties of synthetic shales. Journal of Applied Geophysics, 161: 139-152. https://doi.org/10.1016/j.jappgeo.2018.12.011

[22] Shi, S., Gao, W., Long, T., Xie, D., Li, L., Pei, J., & Cai, J. (2024). Fine characterization of interbedding sand-mudstone based on waveform indication inversion. Journal of Geophysics and Engineering, 21(2), 396-412.

[23] Ren, Jiali; Wang, Yang; Han, De-Hua; Zhao, Luanxiao; Long, Teng; Tang, Shuhang. (2021). Determining crack initiation stress in unconventional shales based on strain energy evolution. Journal of Geophysics and Engineering, 18(5): 642-652. https://doi.org/10.1093/jge/gxab041

[24] LI, Z., OUYANG, F., XIAO, Z., LONG, T., HE, Y., & ZHAO, J. (2022). Improvement and upgrade of elastic modulus testing system under seismic frequency band. Chinese Journal of Geophysics, 65(5), 1769-1784.

[25] Su, J., Cai, C., Zhu, F., He, C., Xu, X., Guan, D., & Si, C. (2024, September). Momentum auxiliary network for supervised local learning. In European Conference on Computer Vision (pp. 276-292). Cham: Springer Nature Switzerland.

[26] Zhou, Y., Yao, J., Liu, Y., Dong, Z., Ye, W., Wen, J., & Choi, L. Y. (2026). Single Canonical Prompts Underestimate LLM Safety's Surface-Form Sensitivity. arXiv preprint arXiv:2608.02665.

[27] Chen, Q. (2026, March). RADR-MO: Resilience-Aware Dynamic Re-routing via Graph Attention Networks and Multi-Objective Meta-Reinforcement Learning for Supply Chain Resilience. In Proceedings of the 2026 3rd International Conference on Innovation Management and Information Systems (pp. 155-162).

[28] Lu, B., Li, D., Zhao, K., Ning, E., Gao, W., Dong, J., & Ning, X. (2026). Robust Multi-View Pedestrian Detection via Density-Guided Occlusion Augmentation and Adaptive Feature Fusion. Information Fusion, 104638.

[29] Zhao, R., Zhou, L., Shang, Z., Liu, Y., Zhang, Y., Li, S., ... & Ning, X. (2026). ProMML for robust few-shot adaptation of vision-language models. Pattern Recognition, 114160.

[30] Jia, J., & Zhou, L. (2025). A threat detection scheme for financial big data in internet of things. Frontiers in Physics, 13, 1633021.

[31] Yan, Z., & Lyu, T. (2026). Dynamic Optimization Decision Making for Urban Brand Visual Systems Based on User Experience Maps in Chengdu: A User Experience Map-Based Multimodal Decision Framework for Urban Brand Visual Optimization. Journal of Organizational and End User Computing (JOEUC), 38(1), 1-57.

[32] Cheng, X., Qin, Y., Tan, Y., Li, Z., Wang, Y., Xiao, H., & Zhang, Y. (2026). Psymem: Fine-grained psychological alignment and explicit memory control for advanced role-playing llms. Transactions of the Association for Computational Linguistics, 14, 510-529.

[33] Que, H., & Yao, H. (2026). Coding What Matters: A Semantic-Aware Memory Interface for Energy-Efficient Perception in Autonomous Vehicles. arXiv preprint arXiv:2608.29000.

[34] Zhu, H., Huang, X., Gao, H., Jiang, M., Que, H., & Mu, L. (2025, July). A Wireless Collaborated Inference Acceleration Framework for Plant Disease Recognition. In International Conference on Intelligent Computing (pp. 331-341). Singapore: Springer Nature Singapore.

[35] Yu, Y., Kang, L., & Gao, W. (2026). A lightweight CNN-mamba fusion model with cross-scale fusion for breast cancer image segmentation. Ain Shams Engineering Journal, 17(12), 104470.

[36] Lv, Z., Wang, Q., Gao, W., & Zhou, L. (2026). Enhancing Corporate Green Innovation Under the Green Factory Policy: A Graph Neural Network Approach for Management Decision-Making. Journal of Organizational and End User Computing (JOEUC), 38(1), 1-33.

[37] Liu, M., Que, H., Yang, X., Zhang, K., Yu, Q., Yan, L., ... & Zhou, N. (2023). A Selective Bit Dropping and Encoding Co-Strategy in Image Processing for Low-Power Design in DRAM and SRAM. IEEE Journal on Emerging and Selected Topics in Circuits and Systems, 13(1), 48-57.

[38] Gao, H., Que, H., Au, H., Shan, W., Liu, M., Qin, Y., ... & Qiao, F. (2025). SenseExpo: Efficient Autonomous Exploration with Prediction Information from Lightweight Neural Networks. arXiv preprint arXiv:2503.16000.

[39] Xi, X., Chen, P., Qiu, Y., Luo, R., Tong, P., & Liang, J. Video-BCI: Bayesian Cognitive Integration of Self-Prior Hypotheses for Video Understanding. In Forty-third International Conference on Machine Learning.

[40] Liang, J., Yang, H., Zhao, X., Yu, Z., Xu, G., Wang, W., & Yang, K. (2026). Democratic recommendation with user and item representatives produced by graph condensation. IEEE Transactions on Knowledge and Data Engineering.

[41] Wang, M., & Saxena, N. (2026). SoK: Trading Agents or Market Crashers? Dissecting Robustness and Security Failures in Academic Financial LLM Trading Schemes. arXiv preprint arXiv:2609.19705.

[42] Wu, H., Song, Y., Zhang, Z., Zhang, Z., Zhou, H., Ma, W. Y., & Liu, J. (2026). Rationalized All-Atom Protein Design with Unified Multi-Modal Bayesian Flow. Advances in Neural Information Processing Systems, 38, 3019-3050.

[43] Xu, Z., Zhang, X., Zhou, X., & Zhang, J. (2025). AvatarShield: Visual Reinforcement Learning for Human-Centric Video Forgery Detection. arXiv preprint arXiv:2505.15173.

[44] Huang, Q., Xu, Z., Zhang, X., & Zhang, J. (2025). UniShield: An Adaptive Multi-Agent Framework for Unified Forgery Image Detection and Localization. arXiv preprint arXiv:2510.03161.

[45] Jia, Z., You, K., He, W., Tian, Y., Feng, Y., Wang, Y., ... & Zhang, Z. (2023). Event-based semantic segmentation with posterior attention. IEEE Transactions on Image Processing, 32, 1829-1842.

[46] Zhang, Y., Yang, Z., Chen, J., Qi, Z., & Yang, G. (2019, August). Presentation, derivation and numerical experiments of a group of extrapolation formulas. In 2019 9th International Conference on Information Science and Technology (ICIST) (pp. 240-246). IEEE.

[47] Ning, E., Gao, W., Liu, Y., Yuan, J., Zhang, H., He, S., & Ning, X. (2026). Causally invariant video anomaly detection via counterfactual reasoning and prototype intervention. Pattern Recognition, 113809.

[48] Jiang, C. (2026). Information doesn’t work? Testing experiential drivers of VR application in heritage tourism by affective-cognitive pathways. Journal of Hospitality and Tourism Technology, 1-28.

[49] Wang, T., & Xia, Z. (2025). Stability of In-Context Learning: A Spectral Coverage Perspective. arXiv preprint arXiv:2509.20677.

[50] Liu, T., He, J., Zhou, R., Song, J., Liu, H., & Liang, Y. (2025). Revolutionizing clinical decision making through deep learning and topic modeling for pathway optimization. Scientific Reports, 15(1), Article 28787. https://doi.org/10.1038/s41598-025-12679-z