M2C: MATCHABILITY-AWARE MUTUAL-CONSISTENCY FUSION FOR RGB–INFRARED OBJECT DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

We propose M2C, a matchability-aware mutual-consistency fusion method for RGB–infrared object detection. For each infrared query, M2C augments the local RGB candidate set with an explicit no-match state, measures ambiguity through the concentration of the conditional candidate distribution, and verifies the forward correspondence using reverse matchability and forward–reverse soft-offset agreement. The resulting confidence serves as a continuous evidence-admission signal that regulates RGB routing and gates target-aware cross-stage updates to the infrared stream. This formulation distinguishes finding a relative local match from determining whether its RGB evidence is reliable enough for fusion, while requiring neither cross-modal correspondence annotations nor an additional cycle-consistency loss. Experiments on DroneVehicle and LLVIP demonstrate the effectiveness of M2C across aerial and pedestrian scenes. On DroneVehicle, M2C achieves 63.67% $\mathrm{mAP}_{50:95}$, outperforming a controlled LocalMatch baseline by 0.62 percentage points. Under RGB replacement, the improvement increases to 5.45 points, while the clean-to-degraded performance drop decreases from 14.98 to 10.14 points. These results show that structured correspondence reliability enables more selective and robust RGB–infrared interaction.

1 INTRODUCTION

RGB and infrared (RGB–IR) sensors provide complementary appearance and thermal cues, yet paired acquisition does not ensure reliable correspondence at every location. Parallax, residual misregistration, occlusion, and modality-specific degradation can make a nearby RGB feature misleading for an infrared query. The challenge is therefore not only to locate cross-modal evidence, but also to decide whether it should affect detection.

Alignment and cross-modal attention can retrieve displaced features (Yuan & Wei, 2024), while recent detectors explicitly account for spatial reliability or modality quality (Hou et al., 2026; Feng et al., 2026; Fan et al., 2026). Nevertheless, a Softmax restricted to K real candidates produces a unit-sum assignment even when every candidate is weak or ambiguous. Finding a relative local match is not equivalent to establishing that its evidence is trustworthy. This motivates separating *correspondence estimation* from *evidence admission*.

We introduce Matchability-Aware Mutual-Consistency Fusion (M2C). For each infrared query, an explicit no-match state estimates whether a plausible RGB counterpart exists; conditional candidate concentration measures ambiguity among real candidates; and reverse matchability with forward–reverse soft-displacement agreement provides independent verification. Their combined confidence modulates current-scale RGB routing and gates target-conditioned updates to the infrared stream. Infrared serves as a spatial reference with a retained direct pathway, while RGB is admitted conditionally; we do not assume that infrared is superior under every degradation. The method uses existing detection annotations without additional correspondence labels or a separate cycle-consistency loss.

On the DroneVehicle validation split (Sun et al., 2022), M2C-E1 obtains 84.38% mAP_{50} and 63.72% $\mathrm{mAP}_{50:95}$, exceeding a controlled LocalMatch variant by 0.70 points on the latter metric. When RGB is replaced by a constant image, it obtains 53.53% $\mathrm{mAP}_{50:95}$, compared with 48.07% for LocalMatch.

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

Results on LLVIP (Jia et al., 2021) provide supplementary evidence in low-light pedestrian detection. Our contributions are:

- We distinguish forced local correspondence from reliable evidence admission in RGB–IR detection.
- We combine no-match, conditional concentration, and reverse verification into a continuous confidence signal that controls RGB routing and infrared feedback.
- We evaluate this design using controlled architecture and factor ablations, RGB-only perturbations, and a second detection dataset, while reporting the additional inference cost.

2 Related Work

2.1 Alignment and Reliable Fusion in RGB–IR Detection

RGB–infrared detection exploits complementary appearance and thermal cues through cross-modal interaction. Existing methods have evolved from direct feature aggregation toward explicit alignment and reliability-aware fusion. C^{2}Former (Yuan & Wei, 2024) combines adaptive feature sampling with cross-modal attention to associate spatially displaced features, although attention strength does not necessarily reflect correspondence reliability.

Recent methods further account for modality-dependent quality. LER-YOLO (Hou et al., 2026) incorporates alignment-aware reliability into sparse expert routing, InfraNet (Feng et al., 2026) adopts an infrared-primary architecture with quality-gated RGB guidance, and EGM-Det (Fan et al., 2026) uses input-derived modality priors for adaptive alignment and fusion.

M2C instead evaluates reliability within each local correspondence set. It jointly considers whether a valid RGB counterpart exists, whether candidate assignments are ambiguous, and whether the estimated correspondence receives reverse support. The resulting confidence directly controls RGB evidence admission and subsequent infrared feature updates.

2.2 Rejectable Matching and Reverse Verification

Reliable correspondence estimation has been extensively studied in feature matching. LoFTR (Sun et al., 2021) establishes dense correspondences, while SuperGlue (Sarlin et al., 2020) introduces partial assignment with an unmatched state. Probabilistic Warp Consistency (Truong et al., 2022) explicitly models unmatched regions, LightGlue (Lindenberger et al., 2023) separates similarity from feature matchability, and forward–backward consistency provides additional verification of correspondence assignments (Zhou et al., 2016).

Unlike geometric matching, RGB–IR detection uses correspondences as intermediate evidence for feature fusion. A locally dominant candidate may still be unreliable when all nearby matches are weak or ambiguous, and a plausible forward assignment may lack reverse support.

M2C therefore adapts rejectable matching and bidirectional verification to detection by combining an explicit no-match state, conditional candidate concentration, and reverse soft-correspondence verification. Rather than producing hard matches or introducing an additional cycle-consistency loss, the resulting confidence is used as a continuous signal for regulating cross-modal evidence admission.

3 Method

3.1 Overall Architecture

Given a nominally registered RGB–IR pair (X^R, X^I), M2C uses the infrared feature grid as the spatial reference and regulates the contribution of supplementary RGB evidence according to local correspondence reliability. As shown in Figure 1, a dual-stream backbone produces RGB and infrared features R_ℓ, I_ℓ at P3–P5 (strides $8, 16, 32$). At each level, M2C yields a fused feature F_ℓ and an updated infrared feature I_ℓ^+, while leaving the RGB stream unchanged ($R_\ell^+ = R_\ell$). At P3 and P4, F_ℓ feeds the PAN–FPN neck and I_ℓ^+ enters the next infrared backbone stage. At P5, pooled context refinement processes F_5 before the neck; I_5^+ has no downstream backbone consumer.

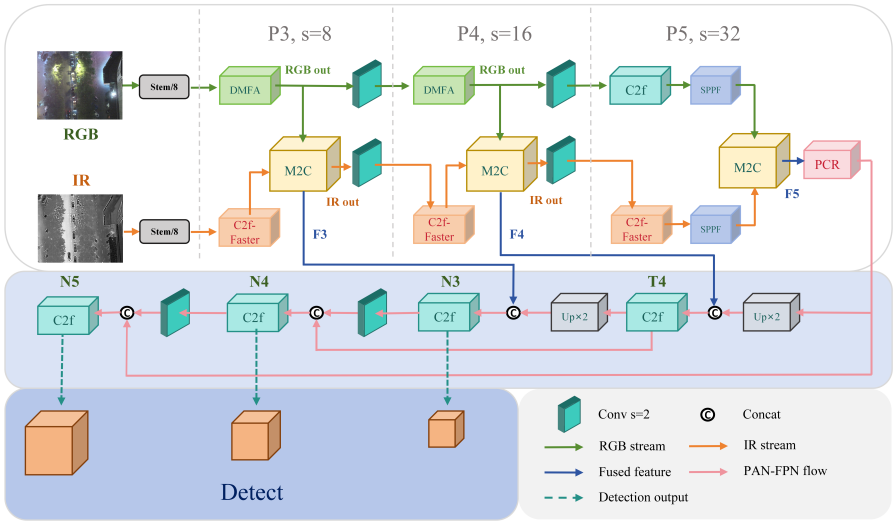

Figure 1: Dual-stream detector with reliability-conditioned M2C fusion at P3–P5. Fused features enter the PAN–FPN neck, with pooled context refinement at P5.

3.2 Matchability-Aware Correspondence Reliability

Figure 2 details the M2C module. Modality-specific adapters followed by a stage-wise shared encoder produce $R_s = E(A_R(R))$ and $I_s = E(A_I(I))$. A shared projection P and channel normalization define the matching features. For each infrared query at position x, we consider $K = 9$ RGB candidates at offsets δ_k in a 3×3 neighborhood ($r = 1$), together with an explicit no-match logit $b_\varnothing$:

$$\ell_k(x) = \frac{\left\langle \widehat{P(I_s(x))}, \widehat{P(R_s(x+\delta_k))} \right\rangle}{\tau_{\text{corr}}} - \lambda_d \|\delta_k\|_2^2, \qquad (p_1, \ldots, p_K, p_\varnothing)(x) = \text{Softmax}\big(\ell_1(x), \ldots, \ell_K(x), b_\varnothing\big). \tag{1}$$

Forward matchability m_f measures the mass assigned to real RGB candidates. We renormalize their probabilities to obtain conditional weights A_k and compute their normalized concentration q_f:

$$\begin{aligned} m_f(x) &= 1 - p_\varnothing(x), \\ A_k(x) &= \frac{p_k(x)}{\sum_{j=1}^{K} p_j(x)}, \\ q_f(x) &= 1 - \frac{-\sum_{k=1}^{K} A_k(x) \log A_k(x)}{\log K}. \end{aligned} \tag{2}$$

Here, m_f measures support for a real candidate relative to no-match, whereas q_f measures ambiguity *among* real candidates. A peaked conditional distribution need not indicate a valid match. The weights A_k produce aligned RGB features and the expected forward displacement:

$$\begin{aligned} R_a(x) &= \sum_k A_k(x) R(x + \delta_k), \\ R_{s,a}(x) &= \sum_k A_k(x) R_s(x + \delta_k), \\ d_f(x) &= \sum_k A_k(x) \delta_k. \end{aligned} \tag{3}$$

The no-match state does not hard-reject the aligned feature; its influence is controlled by the confidence defined below. Applying the same matching operator in the reverse direction (RGB queries, infrared candidates) yields reverse matchability m_r and expected displacement d_r. We transport them to the infrared grid using the forward weights: $\bar{m}_r(x) = \sum_k A_k(x) m_r(x + \delta_k)$

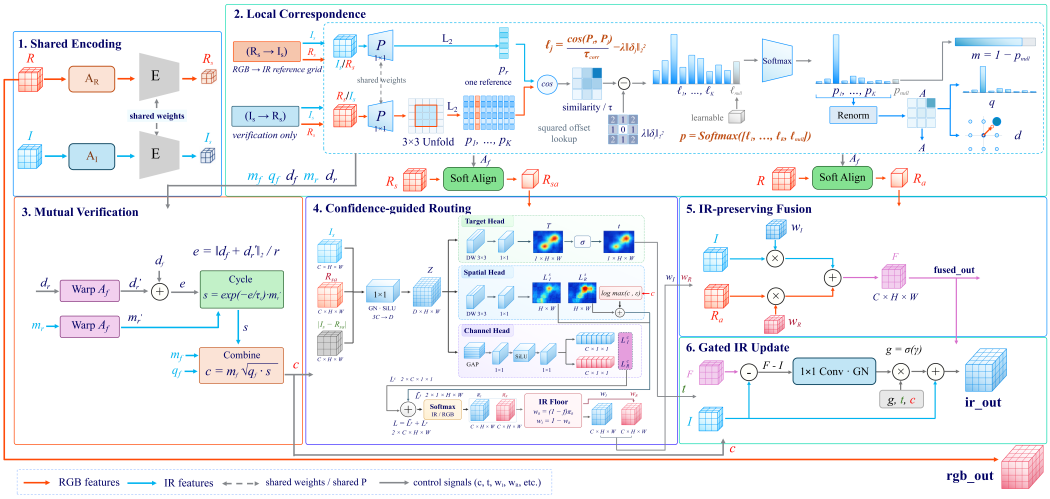

Figure 2: M2C estimates local correspondence confidence from no-match, conditional concentration, and reverse verification, then uses it to control RGB routing and infrared updates.

and $\bar{d}_r(x) = \sum_k A_k(x) d_r(x + \delta_k)$. The round-trip error e, reverse-verification score s, and final correspondence confidence c are

$$\begin{aligned} e(x) &= \|d_f(x) + \bar{d}_r(x)\|_2 / r, \\ s(x) &= \bar{m}_r(x) \exp[-e(x)/\tau_c], \\ c(x) &= m_f(x) \sqrt{q_f(x) s(x)}. \end{aligned} \tag{4}$$

The three factors capture candidate support, conditional ambiguity, and reverse support, respectively. In particular, s combines reverse matchability with soft-displacement agreement rather than representing a purely geometric cycle score. The resulting c is a continuous fusion-control signal, not a calibrated matching probability; reverse matching does not introduce a second fusion pathway.

3.3 Reliability-Conditioned Evidence Admission

Spatial and channel routing heads operate on the semantic representation $Z = [I_s, R_{s,a}, |I_s - R_{s,a}|]$ to produce modality logits L_I, L_R. Correspondence confidence modifies only the RGB logit:

$$\begin{aligned} \widetilde{c} &= \max(c, \epsilon), \qquad \widetilde{L}_R = L_R + \log \widetilde{c}, \\ (P_I, P_R) &= \mathrm{Softmax}_{\mathrm{mod}}(L_I, \widetilde{L}_R), \\ w_R &= (1 - \rho) P_R, \qquad w_I = 1 - w_R. \end{aligned} \tag{5}$$

For fixed semantic logits, decreasing c reduces the RGB routing weight; $w_I \geq \rho$ retains a direct infrared pathway without assuming that infrared is always more informative. A separate target head predicts $t = \sigma(h_t(Z))$. This prior does not enter the modality Softmax; it modulates cross-stage infrared feedback:

$$\begin{aligned} F &= w_I \odot I + w_R \odot R_a, \\ I^+ &= I + g\, t\, c \odot \phi(F - I), \qquad R^+ = R. \end{aligned} \tag{6}$$

Here $g = \sigma(\gamma)$ is a stage-specific bounded gain and ϕ is a pointwise projection with GroupNorm. Thus, confidence controls both the current-scale RGB contribution and infrared updates consumed by subsequent stages at P3–P4. The P5 update is computed but does not feed the main detection path.

3.4 Supervision and Instantiation

The target head is supervised with scale-specific binary maps derived from existing detection boxes: a grid location is positive when its center lies inside a ground-truth box. The training objective is

$$\mathcal{L} = \mathcal{L}_{\text{det}} + \lambda_p \sum_{\ell \in \{3,4,5\}} \left(\mathcal{L}^{\ell}_{\text{focal}} + \mathcal{L}^{\ell}_{\text{Dice}} \right), \tag{7}$$

where the auxiliary terms are focal binary cross-entropy (Lin et al., 2017) and soft Dice loss (**?**). The P5 target head remains supervised during training even though I_5^+ is not consumed by the main detection path. At inference, t is predicted from features; no correspondence annotations or separate cycle-consistency loss are required. C2f-DMFA prepares RGB features at P3–P4, C2f-Faster prepares infrared features, and pooled context refinement acts on fused P5. These are supporting detector components rather than separate contributions of M2C.

4 Experiments

4.1 Experimental Setup

We evaluate M2C on DroneVehicle (Sun et al., 2022) and LLVIP (Jia et al., 2021). DroneVehicle contains 28,439 RGB–IR pairs with 953,087 vehicle instances across five classes, split into 17,990/1,469/8,980 pairs for training/validation/testing. LLVIP provides 16,836 pairs for supplementary nighttime pedestrian detection. We use IR as the reference modality, horizontal bounding boxes (HBBs), and mAP_{50}/$\text{mAP}_{50:95}$ as evaluation metrics. The dual-stream YOLOv8 (Jocher et al., 2023) employs 640×640 inputs and M2C fusion at P3–P5. Models are trained with SGD (learning rate 0.01, momentum 0.937, weight decay 5×10^{-4}), batch size 16, and 200/120 epochs on DroneVehicle/LLVIP, respectively, using Mosaic, scaling, and horizontal flipping. Training uses PyTorch 2.8.0, CUDA 12.8, and an RTX 4070 Ti; efficiency is measured on an RTX 4070 Ti. Controlled ablations and robustness tests use the DroneVehicle validation set. Architectural ablations report training-log scores, whereas robustness tests re-evaluate fixed checkpoints, yielding clean $\text{mAP}_{50:95}$ scores of 63.72 and 63.67 for M2C-E1, respectively. Performance changes are computed within each pipeline, and shift standard deviations reflect perturbation realizations rather than independent training runs.

4.2 Comparison with Existing Detectors

DroneVehicle. Table 1 reports the complete comparison, including category-level AP for methods with available results. The single-modal references include RetinaNet, S^2A-Net, R^3Det, Oriented R-CNN, and YOLOv8 (Lin et al., 2017; Han et al., 2022; Yang et al., 2021; Xie et al., 2021; Jocher et al., 2023). The multimodal comparison includes GLFNet, TSFADet, C^2Former, DMM, IV-YOLO, WaveMamba, Fusion-Mamba, CFM-YOLO, and MDCFT-HSM (Kang et al., 2024; Yuan et al., 2022; Yuan & Wei, 2024; Zhou et al., 2025; Tian et al., 2024; Zhu et al., 2025; Dong et al., 2025; Huang et al., 2026; Long et al., 2026). M2C-E1 obtains 84.4% mAP_{50} and 63.7% $\text{mAP}_{50:95}$ in the source comparison, versus 78.7%/58.5% for the listed dual-stream YOLOv8 reference. The matched architectural comparison is reported in Table 3.

LLVIP. As shown in Table 2, the single-modal references include Faster R-CNN, DDQ DETR, SSD, and YOLOv8 (Ren et al., 2015; Zhang et al., 2023; Liu et al., 2016; Jocher et al., 2023), while the multimodal references include CSAA, RSDet, DIVFusion, ICAFusion, MetaFusion, Fusion-Mamba, and MFYOLO (Cao et al., 2023; Zhao et al., 2026; Tang et al., 2023; Shen et al., 2024; Zhao et al., 2023; Dong et al., 2025; Ning et al., 2025). M2C-E1 achieves 96.7% mAP_{50} and 64.4% $\text{mAP}_{50:95}$ on low-light pedestrian detection, compared with 95.4%/59.6% for the listed dual-stream YOLOv8 reference. Compared with Fusion-Mamba, its mAP_{50} is 0.3 points lower and its $\text{mAP}_{50:95}$ is 0.1 points higher; this small cross-paper difference does not establish a consistent advantage under a matched protocol. LLVIP provides supplementary validation in a different scene and target category, not a zero-shot cross-domain transfer experiment.

Table 1: Comparison on DroneVehicle (%). RGB/IR denotes single-modal input; RGB+IR denotes multimodal input.

Method	Backbone	Modality	Car	Truck	Freight car	Bus	Van	mAP_{50}	$mAP_{50:95}$
RetinaNet	ResNet50	RGB	78.5	34.4	24.1	69.8	28.8	47.1	23.4
S^2A-Net	ResNet50	RGB	80.0	54.2	42.2	84.9	43.8	61.0	–
R^3Det	ResNet50	RGB	80.3	56.1	42.7	80.2	44.4	60.8	27.4
Oriented R-CNN	ResNet50	RGB	80.1	53.8	41.6	85.4	43.3	60.8	32.7
YOLOv8	YOLOv8	RGB	92.5	68.8	47.8	91.7	50.2	70.2	48.6
RetinaNet	ResNet50	IR	88.8	35.4	39.5	76.5	32.1	54.5	30.4
S^2A-Net	ResNet50	IR	89.9	54.5	55.8	88.9	48.4	67.5	–
R^3Det	ResNet50	IR	89.5	48.3	16.6	87.1	39.9	62.3	40.4
Oriented R-CNN	ResNet50	IR	89.8	57.4	53.1	89.3	45.4	67.0	–
YOLOv8	YOLOv8	IR	93.4	69.3	53.7	91.9	51.1	71.9	49.1
GLFNet	ResNet50	RGB+IR	90.3	72.7	53.6	88.0	52.6	71.4	42.9
TSFADet	ResNet50	RGB+IR	89.9	67.9	63.7	89.8	54.0	73.1	44.1
C^2Former	ResNet50	RGB+IR	90.2	68.3	64.4	89.8	58.5	74.2	47.5
DMM	ResNet50	RGB+IR	90.4	78.9	66.2	89.5	68.3	78.6	50.6
IV-YOLO	YOLOv8	RGB+IR	97.2	65.4	63.1	94.3	53.0	74.6	56.8
WaveMamba	YOLOv8	RGB+IR	95.0	80.4	68.5	90.6	64.5	79.8	60.5
Fusion-Mamba	YOLOv8	RGB+IR	96.7	80.2	59.8	95.3	64.0	79.2	56.0
CFM-YOLO	YOLOv11n	RGB+IR	–	–	–	–	–	81.6	60.3
MDCFT-HSM	Custom	RGB+IR	98.1	81.4	70.0	96.7	**71.8**	83.6	63.3
Dual-stream YOLOv8	YOLOv8	RGB+IR	98.3	78.3	64.3	95.6	57.0	78.7	58.5
M2C (Ours)	YOLOv8	RGB+IR	**98.6**	80.1	**76.5**	**97.3**	69.4	**84.4**	**63.7**

Table 2: Comparison on LLVIP (%) for low-light pedestrian detection. Best results are in bold.

Method	Backbone	Modality	mAP_{50}	$mAP_{50:95}$
Faster R-CNN	ResNet50	RGB	88.8	47.5
DDQ DETR	ResNet50	RGB	86.1	46.7
SSD	VGG-16	RGB	82.6	39.8
YOLOv8	YOLOv8	RGB	88.0	49.2
Faster R-CNN	ResNet50	IR	92.6	50.7
DDQ DETR	ResNet50	IR	93.9	58.6
SSD	VGG-16	IR	90.2	53.5
YOLOv8	YOLOv8	IR	94.5	61.5
CSAA	ResNet50	RGB+IR	94.3	59.2
RSDet	ResNet50	RGB+IR	95.8	61.3
DIVFusion	YOLOv5	RGB+IR	89.8	52.0
ICAFusion	YOLOv5	RGB+IR	96.3	62.3
Metafusion	YOLOv5	RGB+IR	91.0	56.9
Fusion-Mamba	YOLOv8	RGB+IR	**97.0**	64.3
MFYOLO	YOLOv8	RGB+IR	96.4	63.8
Dual-stream YOLOv8	YOLOv8	RGB+IR	95.4	59.6
M2C (Ours)	YOLOv8	RGB+IR	96.7	**64.4**

4.3 Qualitative Analysis

Qualitative comparison on DroneVehicle. Figure 3 compares M2C with DMM and C^2Former on DroneVehicle under occlusion, dense targets, and complex backgrounds. Red and yellow ellipses indicate missed detections and false positives, respectively. Due to different bounding-box formats, we focus on detection errors rather than localization geometry.

In the nighttime occlusion example, M2C detects a partially visible vehicle missed by C^2Former. In densely packed parking areas, both DMM and C^2Former exhibit missed detections, while M2C preserves more complete detection results. Under complex backgrounds with small-scale and partially occluded vehicles, M2C further reduces the highlighted missed detections and false positives. These observations are consistent with the quantitative results in Section 4.2, providing additional qualitative evidence of M2C's effectiveness in challenging UAV scenes.

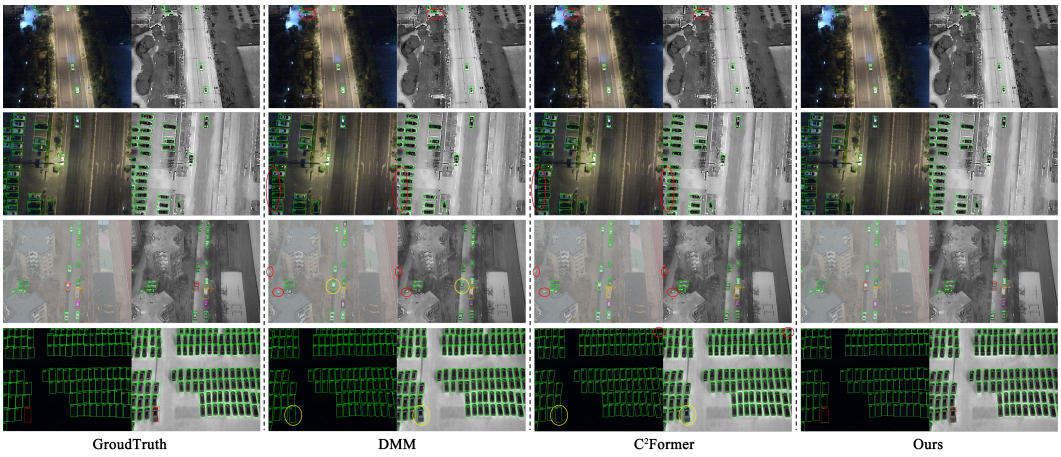

Figure 3: Qualitative coQualitative comparisononeVehicle. Columns show Ground Truth, DMM, C^{2}Former, and M2C (Ours), each with paired RGB and IR images. Red and yellow ellipses indicate missed detections and false positives, respectively.

Qualitative comparison on LLVIP. Figure 4 compares M2C with ICAFusion on LLVIP under challenging nighttime conditions, including closely spaced pedestrians, uneven illumination, and partial occlusion. In the first example, ICAFusion fails to distinguish two overlapping pedestrians, whereas M2C detects both individuals. In the second example, weak RGB appearance responses under uneven illumination lead to a missed detection by ICAFusion, while M2C successfully detects the corresponding pedestrian. In the third example, M2C detects a partially occluded pedestrian missed by ICAFusion.

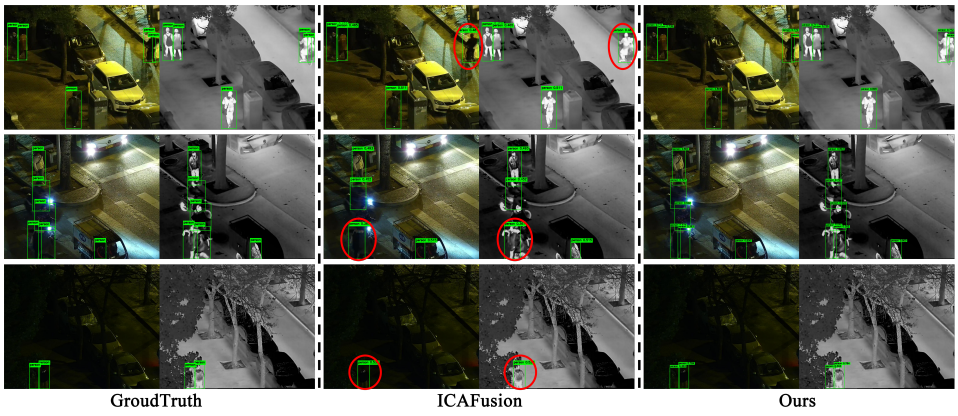

Figure 4: Qualitative comparison on LLVIP. Columns show Ground Truth, ICAFusion, and M2C (Ours), each with paired RGB and IR images. Red ellipses indicate missed pedestrian detections.

These observations, together with the quantitative results in Section 4.2, provide supplementary evidence of M2C's applicability to low-light pedestrian detection. However, these visualizations do not directly establish the internal responses of matchability, concentration, or reverse verification.

4.4 Ablation of the Overall Architecture

We compare four variants on DroneVehicle: the strict dual-stream baseline, E1-ADD with direct fusion, E1-LocalMatch with local correspondence estimation, and the complete M2C-E1 with reliability-conditioned fusion. The controlled strict baseline is distinct from the external dual-stream YOLOv8 reference in Table 1.

Table 3: Architecture ablation on DroneVehicle validation (%). Prep.: E1 feature preparation; Local: correspondence estimation; Verify: reliability verification. One checkpoint per variant.

Method	Prep.	Local	Verify	mAP_{50}	$mAP_{50:95}$
Baseline	–	–	–	78.72	58.50
E1-ADD	•	–	–	81.42	61.63
E1-LocalMatch	•	•	–	83.11	63.02
M2C-E1	•	•	•	**84.38**	**63.72**

As shown in Table 3, M2C-E1 achieves 84.38% mAP_{50} and 63.72% $mAP_{50:95}$, improving the strict baseline by 5.66 and 5.22 points, respectively. With the same E1 feature representation, local matching raises $mAP_{50:95}$ from 61.63% to 63.02% and mAP_{50} from 81.42% to 83.11%. Adding the complete reliability-conditioned fusion design improves mAP_{50} and $mAP_{50:95}$ by 1.27 and 0.70 points over E1-LocalMatch, respectively. These results support assessing candidate reliability in addition to estimating correspondence; the full baseline-to-M2C gain also includes changes to feature preparation.

4.5 Robustness to Unreliable Cross-Modal Evidence

To evaluate robustness to unreliable RGB evidence, we apply spatial misalignment, photometric degradation, and complete RGB replacement to the DroneVehicle validation set. For spatial misalignment, RGB images are translated by independently sampled integer offsets from $[-s, s]^2$ ($s = 4, 8, 12$ pixels), using bilinear interpolation and constant-114 padding. Each shift is evaluated over three perturbation realizations per fixed checkpoint. Photometric degradation follows $X_\gamma^R = 255(X^R/255)^\gamma$ with $\gamma \in \{1, 1.5, 2, 3\}$, while RGB drop replaces all RGB pixels with 114. All perturbations are applied at inference time without modifying IR inputs, annotations, or model weights.

Table 4: RGB spatial misalignment on DroneVehicle ($mAP_{50:95}$, %). Shift-s: offsets in $[-s, s]^2$; mean $\pm$ std. over three perturbations per fixed checkpoint.

Method	Shift-0	Shift-4	Shift-8	Shift-12
E1-ADD	61.63	60.46 ± 0.08	59.26 ± 0.14	58.40 ± 0.17
E1-LocalMatch	63.03	62.02 ± 0.08	61.39 ± 0.09	60.93 ± 0.11
M2C-E1	**63.67**	$\mathbf{62.68 \pm 0.18}$	$\mathbf{62.08 \pm 0.13}$	$\mathbf{61.64 \pm 0.17}$

Table 5: RGB photometric degradation and replacement on DroneVehicle ($mAP_{50:95}$, %). Higher γ indicates stronger darkening; RGB drop replaces RGB pixels with 114 while preserving IR.

Method	$\gamma = 1$	$\gamma = 1.5$	$\gamma = 2$	$\gamma = 3$	RGB drop
E1-ADD	61.63	61.62	60.37	57.88	51.35
E1-LocalMatch	63.04	63.31	62.62	59.82	48.07
M2C-E1	**63.67**	**63.86**	**62.77**	**60.36**	**53.53**

As shown in Tables 4 and 5, M2C-E1 consistently achieves higher AP than E1-ADD and E1-LocalMatch under the evaluated RGB perturbations. At Shift-12, M2C-E1 achieves 61.64 AP, exceeding E1-LocalMatch by 0.71 points. This relatively modest gain suggests that local correspondence estimation already compensates for part of the spatial misalignment, while reliability-conditioned fusion provides additional benefits. Under photometric degradation, M2C-E1 achieves 60.36 AP at $\gamma = 3$, outperforming E1-LocalMatch by 0.54 points. Its performance is not strictly monotonic, increasing from 63.67 AP at $\gamma = 1$ to 63.86 AP at $\gamma = 1.5$ before declining under stronger darkening.

The largest performance difference occurs under complete RGB replacement. E1-LocalMatch decreases to 48.07 AP, below direct addition (51.35), whereas M2C-E1 maintains 53.53 AP, exceeding LocalMatch by 5.46 points. Relative to their respective clean scores, the performance drops are 14.97 and 10.14 points for E1-LocalMatch and M2C-E1, respectively. These results highlight the limitation of forced local matching: even when RGB contains no informative appearance, normalized matching still assigns weights to local candidates. In contrast, M2C regulates RGB evidence

admission according to correspondence reliability, consistent with its improved performance under RGB replacement.

These results support M2C under the tested RGB perturbations, particularly RGB dropout, but do not establish its internal routing behavior or robustness to IR degradation.

4.6 Ablation of Correspondence Reliability

We individually remove the no-match mechanism, conditional concentration, and reverse verification from $c = m_f\sqrt{q_f s}$ (Eq. 4). The strict reverse-verification ablation uses $c = m_f\sqrt{q_f}$, retaining the original concentration exponent.

Table 6: Reliability-factor ablations (DroneVehicle validation; $\text{mAP}_{50:95}$, %). Shift-12 reports mean $\pm$ std. over three sampled perturbations of each fixed checkpoint, not three training runs.

Variant	Clean	Shift-12	RGB drop
w/o No-match	63.16	61.16 ± 0.19	52.33
w/o Concentration	**63.84**	61.66 ± 0.24	49.49
w/o Reverse Verification	62.88	61.08 ± 0.06	50.51
Full M2C	63.67	61.64 ± 0.17	**53.53**

Table 6 shows that removing no-match or reverse verification lowers clean $\text{mAP}_{50:95}$ by 0.51 and 0.79 points, respectively. Removing concentration instead slightly improves clean AP (63.84% versus 63.67%) and yields nearly identical Shift-12 AP (61.66% versus 61.64%). Under complete RGB replacement, however, removing concentration, reverse verification, and no-match reduces AP by 4.04, 3.02, and 1.20 points, respectively. These results show different component effects across the tested conditions rather than uniform gains: concentration is particularly useful under RGB replacement, while the small Shift-12 differences do not establish training-seed stability. The reported standard deviations reflect sampled perturbations of fixed checkpoints, not independent training runs.

4.7 Computational Complexity and Efficiency

All models are evaluated on an NVIDIA RTX 4070Ti using 640×640 six-channel inputs, batch size 1, and FP32 precision. Latency is averaged over 500 synchronized forward passes after 50 warm-up iterations, excluding preprocessing and post-processing. GFLOPs are measured using a consistent THOP-based protocol. Because operation counts do not uniquely determine wall-clock speed (Chen et al., 2023), we report both GFLOPs and synchronized forward latency/FPS.

Table 7: Accuracy and efficiency on DroneVehicle. Timings are forward-only on an RTX 4070 Ti; AP is measured on the validation set.

Method	Params (M)	GFLOPs	ms	FPS	$\text{mAP}_{50:95}$
Strict Baseline	8.25	14.27	11.49	87.01	58.50
E1-ADD	9.04	15.65	16.16	61.87	61.63
E1-LocalMatch	9.53	16.67	21.70	46.09	63.02
M2C-E1	9.53	16.75	25.37	39.42	**63.72**

As shown in Table 7, M2C-E1 improves $\text{mAP}_{50:95}$ by 0.70 points over E1-LocalMatch with 0.08 additional GFLOPs. Its forward latency increases by 16.9% (21.70 to 25.37 ms), reflecting the runtime cost of correspondence verification.

5 Conclusion

We proposed M2C for RGB–infrared object detection, integrating no-match modeling, conditional concentration, and reverse verification to regulate cross-modal evidence admission. M2C-E1 achieves 63.72% $\text{mAP}_{50:95}$ on DroneVehicle, outperforming LocalMatch by 0.70 points under nominal conditions and 5.46 points under complete RGB replacement. LLVIP results further support its applicability to low-light pedestrian detection. Future work will address inference efficiency and robustness to infrared degradation and simultaneous modality failure.

AI USE STATEMENT

Generative AI tools were used to assist with English translation, manuscript drafting and refinement, and code implementation. The authors independently reviewed and verified the AI-assisted content, experimental results, and source code, and take full responsibility for the final manuscript.

REFERENCES

Yue Cao, Junchi Bin, Jozsef Hamari, Erik Blasch, and Zheng Liu. Multimodal object detection by channel switching and spatial attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 403–411, 2023. URL https://openaccess.thecvf.com/content/CVPR2023W/PBVS/html/Cao_Multimodal_Object_Detection_by_Channel_Switching_and_Spatial_Attention_CVPRW_2023_paper.html.

Jierun Chen, Shiu hong Kao, Hao He, Weipeng Zhuo, Song Wen, Chul-Ho Lee, and S.-H. Gary Chan. Run, don't walk: Chasing higher FLOPS for faster neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12021–12031, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Chen_Run_Dont_Walk_Chasing_Higher_FLOPS_for_Faster_Neural_Networks_CVPR_2023_paper.html.

Wenhao Dong, Haodong Zhu, Shaohui Lin, Xiaoyan Luo, Yunhang Shen, Guodong Guo, and Baochang Zhang. Fusion-Mamba for cross-modality object detection. *IEEE Transactions on Multimedia*, 27:7392–7406, 2025. doi: 10.1109/TMM.2025.3599020.

Cunzheng Fan, Dawei Yan, Guanlin Wang, Xingshuo Yang, Yupeng Jia, Jing Yang, and Haokui Zhang. EGM-Det: Entropy-guided multimodal adaptive fusion for UAV RGB-IR object detection, 2026. URL https://arxiv.org/abs/2608.11685.

Zichao Feng, Haodong Zhu, Jingying Yang, Sheng Xu, Yangyang Ren, Yuguang Yang, Xuhui Liu, Juan Zhang, Tian Wang, Linlin Yang, and Baochang Zhang. InfraNet: Quality-aware RGB guidance for efficient infrared object detection, 2026. URL https://arxiv.org/abs/2607.03795.

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

Liming Hou, Yueping Peng, Hexiang Hao, Ji Wang, Xuekai Zhang, Wei Tang, Zecong Ye, Xin Ying, and Yubo He. LER-YOLO: Reliability-aware expert routing for misaligned RGB-Infrared UAV detection, 2026. URL https://arxiv.org/abs/2605.20667.

Zedong Huang, Kangkang Du, Xiaohuang Hu, and Jianfeng Li. Object detection algorithm based on infrared–visible dual-modality feature fusion. *Scientific Reports*, 16:27378, 2026. doi: 10.1038/s41598-026-63114-w.

Xinyu Jia, Chuang Zhu, Minzhen Li, Wenqi Tang, and Wenli Zhou. LLVIP: A visible-infrared paired dataset for low-light vision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pp. 3496–3504, 2021. URL https://arxiv.org/abs/2108.10831.

Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics YOLOv8. Software, 2023. URL https://github.com/ultralytics/ultralytics.

Xudong Kang, Hui Yin, and Puhong Duan. Global–local feature fusion network for visible–infrared vehicle detection. *IEEE Geoscience and Remote Sensing Letters*, 21:1–5, 2024. doi: 10.1109/LGRS.2024.3375634.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2980–2988, 2017. doi: 10.1109/ICCV.2017.324.

Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. LightGlue: Local feature matching at light speed. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17627–17638, 2023. URL https://arxiv.org/abs/2306.13643.

Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. SSD: Single shot MultiBox detector. In *Computer Vision – ECCV 2016*, pp. 21–37, 2016. doi: 10.1007/978-3-319-46448-0_2.

Tianchen Long, Yanwen Wang, Hezhuo Yuan, Qian Zhang, Hongqing Ma, Rijin Zhou, Zhen Wang, and Feng Wang. MDCFT-HSM: Modality-shift-aware RGB-T small-object detection for low-altitude UAV remote sensing. *Sensors*, 26(16):5146, 2026. doi: 10.3390/s26165146.

Liyang Ning, Xuxun Liu, Luoyu Zhou, and Xueyu Zou. Cross-modal object detection based on content-guided feature fusion and self-calibration. *Sensors*, 25(11):3392, 2025. doi: 10.3390/s25113392.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, volume 28, 2015. URL https://papers.nips.cc/paper/2015/hash/14bfa6bb14875e45bba028a21ed38046-Abstract.html.

Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperGlue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4938–4947, 2020. URL https://arxiv.org/abs/1911.11763.

Jifeng Shen, Yifei Chen, Yue Liu, Xin Zuo, Heng Fan, and Wankou Yang. ICAFusion: Iterative cross-attention guided feature fusion for multispectral object detection. *Pattern Recognition*, 145:109913, 2024. doi: 10.1016/j.patcog.2023.109913.

Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. LoFTR: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8922–8931, 2021. URL https://arxiv.org/abs/2104.00680.

Yiming Sun, Bing Cao, Pengfei Zhu, and Qinghua Hu. Drone-based RGB-Infrared cross-modality vehicle detection via uncertainty-aware learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(10):6700–6713, 2022. doi: 10.1109/TCSVT.2022.3168279. URL https://arxiv.org/abs/2003.02437.

Linfeng Tang, Xinyu Xiang, Hao Zhang, Meiqi Gong, and Jiayi Ma. DIVFusion: Darkness-free infrared and visible image fusion. *Information Fusion*, 91:477–493, 2023. doi: 10.1016/j.inffus.2022.10.034.

Dan Tian, Xin Yan, Dong Zhou, Chen Wang, and Wenshuai Zhang. IV-YOLO: A lightweight dual-branch object detection network. *Sensors*, 24(19):6181, 2024. doi: 10.3390/s24196181.

Prune Truong, Martin Danelljan, Fisher Yu, and Luc Van Gool. Probabilistic warp consistency for weakly-supervised semantic correspondences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8698–8708, 2022. doi: 10.1109/CVPR52688.2022.00851. URL https://arxiv.org/abs/2203.04279.

Xingxing Xie, Gong Cheng, Jiabao Wang, Xiwen Yao, and Junwei Han. Oriented R-CNN for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3520–3529, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Xie_Oriented_R-CNN_for_Object_Detection_ICCV_2021_paper.html.

Xue Yang, Junchi Yan, Ziming Feng, and Tao He. R3Det: Refined single-stage detector with feature refinement for rotating object. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(4):3163–3171, 2021. doi: 10.1609/aaai.v35i4.16426.

Maoxun Yuan and Xingxing Wei. C2Former: Calibrated and complementary transformer for RGB-Infrared object detection. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–12, 2024. doi: 10.1109/TGRS.2024.3376819. URL https://arxiv.org/abs/2306.16175.

Maoxun Yuan, Yinyan Wang, and Xingxing Wei. Translation, scale and rotation: Cross-modal alignment meets RGB-Infrared vehicle detection. In *Computer Vision – ECCV 2022*, pp. 509–525, 2022. doi: 10.1007/978-3-031-20077-9_30.

Shilong Zhang, Xinjiang Wang, Jiaqi Wang, Jiangmiao Pang, Chengqi Lyu, Wenwei Zhang, Ping Luo, and Kai Chen. Dense distinct query for end-to-end object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7329–7338, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Zhang_Dense_Distinct_Query_for_End-to-End_Object_Detection_CVPR_2023_paper.html.

Tianyi Zhao, Maoxun Yuan, Feng Jiang, Nan Wang, and Xingxing Wei. Removal then selection: A coarse-to-fine fusion perspective for RGB-Infrared object detection. *IEEE Transactions on Intelligent Transportation Systems*, 27(2):2504–2519, 2026. doi: 10.1109/TITS.2025.3638627.

Wenda Zhao, Shigeng Xie, Fan Zhao, You He, and Huchuan Lu. MetaFusion: Infrared and visible image fusion via meta-feature embedding from object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13955–13965, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Zhao_MetaFusion_Infrared_and_Visible_Image_Fusion_via_Meta-Feature_Embedding_From_CVPR_2023_paper.html.

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

Tinghui Zhou, Philipp Krähenbühl, Mathieu Aubry, Qixing Huang, and Alexei A. Efros. Learning dense correspondence via 3d-guided cycle consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 117–126, 2016. URL https://arxiv.org/abs/1604.05383.

Haodong Zhu, Wenhao Dong, Linlin Yang, Hong Li, Yuguang Yang, Yangyang Ren, Qingcheng Zhu, Zichao Feng, Changbai Li, Shaohui Lin, Runqi Wang, Xiaoyan Luo, and Baochang Zhang. WaveMamba: Wavelet-driven mamba fusion for RGB-Infrared object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11219–11229, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Zhu_WaveMamba_Wavelet-Driven_Mamba_Fusion_for_RGB-Infrared_Object_Detection_ICCV_2025_paper.html.

A Additional Details of the M2C Module

A.1 Stage-wise data flow

Table 8 makes explicit where the two outputs of each M2C block are consumed. The matching and routing operations are applied at P3, P4, and P5. The RGB stream remains unchanged by these blocks; the updated infrared stream is carried to the next backbone stage only at P3 and P4. At P5, the fused feature is refined before entering the neck, while the computed I_5^+ has no downstream backbone consumer. Consequently, P5 target supervision does not imply a P5 infrared-feedback path to the detection head.

Table 8: Consumers of the M2C outputs at the three fusion stages. The feature preparation and output pathways follow the architecture in Section 3.1.

Stage	Input preparation	Fused output F_ℓ	Updated output I_ℓ^+
P3	RGB: C2f-DMFA; IR: C2f-Faster	PAN–FPN neck	Next IR stage
P4	RGB: C2f-DMFA; IR: C2f-Faster	PAN–FPN neck	Next IR stage
P5	IR: C2f-Faster	Context refinement, then neck	No later IR stage

A.2 Conditional alignment and evidence admission

The local assignment in Eq. 1 contains $K = 9$ real RGB candidates and one no-match state. Let $S(x) = \sum_{j=1}^{K} \exp \ell_j(x)$. The two probabilities used by M2C can equivalently be written as

$$A_k(x) = \frac{\exp \ell_k(x)}{S(x)}, \qquad m_f(x) = \frac{S(x)}{S(x) + \exp b_\varnothing}. \tag{8}$$

Thus, the conditional alignment weights A_k select among real local candidates, whereas m_f compares their aggregate support with the no-match state. Changing the no-match logit while holding the real candidate logits fixed changes m_f but not A_k. A sharp A_k can therefore coexist with low matchability; conditional alignment alone does not determine whether the aligned RGB feature should be admitted.

The normalized concentration is $q_f = 1 - H(A)/\log K$, where $H(A) = -\sum_k A_k \log A_k$. It approaches zero for a uniform conditional assignment and approaches one for a distribution concentrated on a single real candidate. Reverse verification transports reverse matchability and the reverse expected displacement to the infrared query before computing $s = \bar{m}_r \exp(-e/\tau_c)$. With positive temperatures and $0 \le m_f, \bar{m}_r \le 1$, the factors q_f, s, and $c = m_f\sqrt{q_f s}$ lie in $[0, 1]$. Accordingly, either strong ambiguity, low support for a real candidate, or weak reverse verification can lower the final confidence. This is an algebraic property of the module, not a measurement of its learned internal responses.

For a fixed pair of semantic routing logits, the modality Softmax in Eq. 5 has the explicit form

$$P_R = \frac{\widetilde{c}\exp L_R}{\exp L_I + \widetilde{c}\exp L_R}, \qquad w_R = (1-\rho)P_R, \qquad w_I = 1 - w_R. \tag{9}$$

These expressions apply elementwise to the spatial and channel routing outputs. For $c > \epsilon$ and fixed L_I, L_R,

$$\frac{\partial P_R}{\partial c} = \frac{\exp(L_I + L_R)}{(\exp L_I + c\exp L_R)^2} > 0. \tag{10}$$

Consequently, lower confidence reduces RGB routing when the semantic logits are held fixed, and $w_I \ge \rho$ retains the direct infrared path. The ϵ clipping avoids an undefined logarithm and means that the RGB weight need not be exactly zero when $c = 0$. The infrared feedback in Eq. 6 is separately multiplied by $t\,c$ and a bounded stage gain g. Under an input perturbation the semantic logits can also change, so Eq. 10 is a conditional property rather than a claim that measured RGB routing must decrease monotonically for every corrupted image.

A.3 Target supervision and information requirements

For a ground-truth HBB set $\mathcal{B}$ and a cell u at stage ℓ, the auxiliary target label can be expressed as

$$Y_\ell(u) = \mathbf{1}[\exists B \in \mathcal{B} :\ \mathrm{center}_\ell(u) \in B]\,. \tag{11}$$

The focal and soft Dice terms in Eq. 7 use these scale-specific maps at P3–P5. Correspondence logits, no-match probabilities, concentration, and reverse verification are learned through the detector and target objectives without annotated RGB–IR pixel correspondences. At inference, the target head predicts t from the current features; ground-truth boxes are not inputs to the fusion module.

B Additional Details of the Robustness Protocol

The three RGB perturbation families in Section 4.5 act only at inference. In every case the IR image, detection labels, and checkpoint remain fixed. Spatial shift samples an integer offset on each image axis from $[-s, s]$; the reported standard deviation covers three realizations of this input perturbation for each fixed model. Gamma degradation and RGB replacement are single conditions in the reported tables, so they do not carry a perturbation-realization standard deviation. None of these standard deviations estimates variation across independent training seeds.

The M2C search radius is one cell at each fusion level. To relate this radius to the input-space perturbation, Table 9 reports s/stride, the largest nominal displacement along one axis in feature-cell units for a Shift-s input transformation. For Shift-12, this ratio is 1.5 at P3 and at most 0.75 at P4. These ratios describe the input transformation and feature-grid spacing; they are not measured offsets of actual backbone features, which have spatial receptive fields and interpolation effects. In particular, the P3 local window need not contain the corresponding RGB evidence for every Shift-12 realization.

Table 9: Maximum nominal per-axis RGB input shift expressed in cells at each fusion level. Values equal s/stride; M2C examines a 3×3 feature-cell neighborhood at every level.

Level	Stride	Shift-4	Shift-8	Shift-12
P3	8	0.50	1.00	1.50
P4	16	0.25	0.50	0.75
P5	32	0.125	0.25	0.375

C Supplementary Analysis of the Reported Ablations

C.1 Accuracy loss relative to each clean evaluation

Table 10 derives the absolute AP loss for each model from Tables 4 and 5. For a perturbation family p, the reported quantity is $D_p = \text{AP}^{(p)}_{\text{clean}} - \text{AP}^{(p)}_{\text{perturbed}}$ in percentage points. Each column uses the clean evaluation from its corresponding main-text table; for example, E1-LocalMatch has 63.03 clean AP in the shift table and 63.04 at $\gamma = 1$ in the photometric table. The values below are arithmetic transformations of existing results, not new evaluation runs.

Table 10: Clean-to-perturbed loss in $\text{mAP}_{50:95}$ percentage points; lower is better within a perturbation family. Derived from the main-text robustness tables.

Method	Shift-12	$\gamma = 3$	RGB replacement
E1-ADD	3.23	3.75	10.28
E1-LocalMatch	2.10	**3.22**	14.97
M2C-E1	**2.03**	3.31	**10.14**

The direct-addition and LocalMatch variants respond differently to complete RGB replacement: LocalMatch achieves higher clean AP but loses 14.97 points, compared with 10.28 for E1-ADD. Full M2C reduces that loss to 10.14 points and has the highest absolute AP under RGB replacement (Table 5). At $\gamma = 3$, LocalMatch has a slightly smaller within-model drop than M2C-E1, although M2C-E1 still has the higher absolute degraded AP. These distinct comparisons should not be conflated.

C.2 Condition-specific effects of the reliability factors

Table 11 subtracts the Full M2C score from each ablated score in Table 6. Negative values indicate lower AP than Full M2C in the corresponding condition. Shift-12 differences compare the reported means only; the standard deviation of a paired difference cannot be recovered from the individual standard deviations without the per-realization scores.

Removing concentration slightly raises clean AP and leaves Shift-12 mean AP almost unchanged, while substantially lowering AP when the RGB image is replaced. The reverse-verification and no-match removals lower AP in all three reported conditions. This evidence supports condition-dependent

Table 11: Difference from Full M2C in $\text{mAP}_{50:95}$ percentage points, computed from the main-text factor ablation.

Variant	Clean Δ	Shift-12 Δ	RGB drop Δ
w/o No-match	-0.51	-0.48	-1.20
w/o Concentration	$+0.17$	$+0.02$	-4.04
w/o Reverse Verification	-0.79	-0.56	-3.02

contributions of the reliability factors; it does not establish their responses on IR degradation or simultaneous failure of both modalities.