ART: FROM COMMON SEMANTICS TO UNIQUE DETAILS VIA DECOUPLED FEATURE INTERACTION FOR RGBT DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

RGB–thermal (RGBT) object detection is essential for around-the-clock perception. However, many existing detectors implicitly couple fusion with interaction—pushing aggressive cross-modal mixing early—which can *prematurely collapse* modality diversity and suppress complementary cues. We argue that effective interaction should be built on the *right* factors: *shared* semantics should be aligned, while *modality-specific* details should be preserved. Based on this insight, we present **ART**, a decoupled feature interaction framework that follows Decouple $\rightarrow$ Align $\rightarrow$ Interact. ART first disentangles features into cross-modal *common* semantics and modality-specific *unique* details using a modality-agnostic shared projection head and modality-dependent heads. It then stabilizes the shared semantic subspace with a correlation-based alignment objective that encourages an identity-like cross-modal channel correlation—strengthening matched-channel semantic consistency while reducing off-diagonal leakage and redundancy. Finally, ART performs cross-modal interaction *separately* in the common and unique spaces, enabling semantic agreement without erasing fine-grained complementary information. Experiments on public RGBT benchmarks demonstrate that ART achieves state-of-the-art performance.

1 INTRODUCTION

RGB–thermal (RGBT) object detection has become a key ingredient for around-the-clock perception in safety-critical scenarios such as autonomous driving and surveillance, where RGB offers rich texture and color cues while thermal infrared remains reliable under poor illumination and adverse conditions (Liu et al., 2021; Yuan et al., 2025; Zhou et al., 2021). By exploiting modality complementarity, RGBT detectors aim to be both accurate and robust across day, night and weather changes (Zhang et al., 2019; Shen et al., 2023).

However, effective cross-modal modeling requires distinguishing what should be shared across modalities from what should remain modality-specific. Paired RGB and thermal features can be conceptually characterized by an overlapping *common space*, which captures semantically consistent object information, and two *unique spaces*, which retain complementary modality-specific cues. This perspective exposes three closely related requirements: separating common semantics from unique details (*decouple*), establishing reliable correspondence between the common representations (*align*), and exchanging complementary information without erasing modality diversity (*interact*). The key lies not merely in performing these operations, but in organizing them so that semantic agreement and modality complementarity can be jointly preserved.

As summarized in Table 1, existing methods explore cross-modal relationships mainly through two paradigms. The first follows a decouple-then-interact pipeline. MANet [41] employs multiple adapters to separate modality-shared and modality-specific representations, while APFNet [46] decomposes the fusion process into attribute-specific branches. Beyond Duality [XX] further separates shared and private features for RGB-event detection before adaptive feature encoding and injection. By preserving heterogeneous information before interaction, these methods mitigate the premature assimilation of modality-specific cues. However, they generally lack an explicit mechanism for aligning the shared semantics across modalities. Consequently, semantically corresponding features

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.

Table 1: Comparison of cross-modal relation modeling paradigms.

Method	Decouple	Align	Interact
QFDet (Zhang et al., 2023)	✗	✗	✓
SAMS-YOLO (Wang et al., 2024)	✗	✗	✓
ProtoHGF-Net (Chen et al., 2026)	✗	✗	✓
Pei et al. (Pei et al., 2026)	✗	✗	✓
MANet (Wang et al., 2018)	✓	✗	✓
APFNet (Xiao et al., 2022)	✓	✗	✓
RDCRNet (Li et al., 2025b)	✗	✓	✓
OAFA (Chen et al., 2024a)	✗	✓	✓
MRTTrack (Lai et al., 2025)	✗	✓	✓
ART (Ours)	✓	✓	✓

may remain inconsistent, causing subsequent interaction to operate on mismatched representations and limiting effective cross-modal complementarity.

The second paradigm follows an align-then-interact pipeline. RDCRNet [XX] reduces cross-modal distribution discrepancies through feature remapping before bidirectional interaction, OAFA [4] estimates spatial offsets for adaptive feature alignment and fusion, and MRTTrack [18] constrains background feature distributions before representative-token interaction. These methods establish more reliable cross-modal correspondences and thereby facilitate information exchange. However, without first disentangling common semantics from modality-specific details, their alignment is applied to entangled representations and cannot explicitly distinguish what should be consistent from what should remain unique. Such indiscriminate consistency regularization may suppress fine-grained complementary cues, introduce inter-modal interference, and reduce the discriminability of fused representations [16].

Can we disentangle what should be shared from what should remain unique before interaction, and achieve synergy without collapsing modality diversity?

To address these challenges, we propose ART, a decouple–align–interact framework for RGBT object detection. ART first factorizes RGB and thermal features into a shared *common* semantic space and two modality-specific *unique* spaces, allowing cross-modal consistency and modality-dependent details to be modeled separately. It then applies an identity-like cross-modal channel-similarity regularizer exclusively to the common representations. The diagonal and off-diagonal terms encourage agreement between corresponding channels and suppress similarities between non-corresponding channels, respectively, while leaving the unique representations unconstrained by alignment. Finally, ART performs cross-modal interaction separately in the common and unique spaces, enabling the detector to exploit complementary information without explicitly forcing the unique branches to agree. Rather than introducing common–unique decomposition or redundancy reduction as new principles, ART contributes a selective common-space alignment and factorized-interaction formulation tailored to spatially dense, multi-scale RGBT detection features.

In summary, Our main contributions are: Our main contributions are summarized as follows:

- We propose ART, a decoupled feature interaction framework for RGB–Thermal object detection. Unlike existing methods that implicitly couple fusion, alignment, and interaction, ART targets the key tension between semantic agreement and modality diversity in RGBT detection. We introduce a *Decouple* $\rightarrow$ *Align* $\rightarrow$ *Interact* paradigm to explicitly separate cross-modal common semantics from modality-specific details, mitigating feature collapse, modality bias, and complementary cue suppression caused by aggressive early fusion.
- We introduce factorized alignment and decoupled interaction.ART uses a shared common projector to model modality-invariant semantics and unique projectors to preserve modality-specific details. It applies identity-like channel-similarity regularization only in the common space and performs lightweight interaction separately in the common and unique spaces, balancing semantic consistency and complementarity.

- We achieve state-of-the-art performance on public RGBT benchmarks. ART reaches 51.6% and 62.5% mAP on FLIR and M^{3}FD, respectively, surpassing previous state-of-the-art methods; ablations verify the effectiveness of common-space alignment and decoupled interaction.

2 PRELIMINARY

2.1 PROBLEM DIFINITION

Given a synchronized pair of RGB image I_{RGB} and thermal infrared image I_{IR}, the goal is to predict object bounding boxes and class labels:

$$\mathcal{D}(I_{RGB}, I_{IR}) \rightarrow \{(b_i, c_i)\}_{i=1}^{N}. \tag{1}$$

where $b_i \in \mathbb{R}^4$ denotes the bounding box of the i-th instance, c_i is the corresponding class label, and N is the number of predicted objects.

We aim to explicitly model the *common* semantics shared across modalities and the modality-specific *unique* details, enforce cross-modal consistency in the shared semantic space, and preserve complementary cues from the *unique* branches. This design encourages more discriminative and robust detection features, ultimately improving object detection performance.

We use preprocessed RGBT datasets with verified RGB–IR pairing, such that each RGB image has a paired thermal infrared image captured at the same time, forming synchronized input pairs (I_{RGB}, I_{IR}) for training.

2.2 NOTATIONS

We denote the RGB and infrared backbones as Φ_{RGB} and Φ_{IR}, producing multi-scale feature maps $\{F^s_{RGB}\}_{s=1}^{S}$ and $\{F^s_{IR}\}_{s=1}^{S}$, where $F^s_* \in \mathbb{R}^{B\times C_s\times H_s\times W_s}$. For each scale s, we decompose features into a shared *common* space and two modality-specific *unique* spaces:

$$C^s_{RGB} = \psi^s_c(F^s_{RGB}), \;\; C^s_{IR} = \psi^s_c(F^s_{IR}), \tag{2}$$

$$U^s_{RGB} = \psi^s_{u,RGB}(F^s_{RGB}), \;\; U^s_{IR} = \psi^s_{u,IR}(F^s_{IR}). \tag{3}$$

where ψ^s_c is shared across modalities, C^s_N are common features, while $\psi^s_{u,RGB}$ and $\psi^s_{u,IR}$ are modality-specific. U^s_N are specific features. We use $\mathcal{G}^s(\cdot,\cdot)$ to denote the cross-modal interaction block applied to either common or unique features. Besides, λ_c controls the relative weight of the off-diagonal decorrelation term within the common-space alignment objective, while α controls the contribution of $\mathcal{L}_{align}$ to the overall training objective. Additionally, The superscript S denotes the *scale/level index* (multi-scale feature level), distinguishing the module or loss computed at different feature hierarchies. The interacted features are aggregated and fed into the detector heads to produce final predictions.

3 METHODOLOGY

3.1 OVERVIEW

The Preliminary could be found in Appendix Sec 2. We propose **ART**, a *decoupled* feature interaction framework for RGB–thermal (RGBT) object detection. Different from prior methods that couple fusion and interaction implicitly Shen et al. (2023; 2025), ART follows a **Decouple** $\rightarrow$ **Align** $\rightarrow$ **Interact** paradigm: (i) we *explicitly decompose* features into a cross-modal *common* semantic space and two modality-specific *unique* spaces using shared and non-shared projection heads; (ii) we *stabilize* the common space by a correlation-based alignment objective that encourages semantic agreement while reducing channel redundancy; and (iii) we perform *lightweight* cross-modal interaction *independently* in the common and unique spaces. The interacted features are aggregated and fed into a standard detector head for final predictions. An overview is shown in Fig. 1, and the pipeline is summarized in Appendix Sec 8 Algorithm 1.

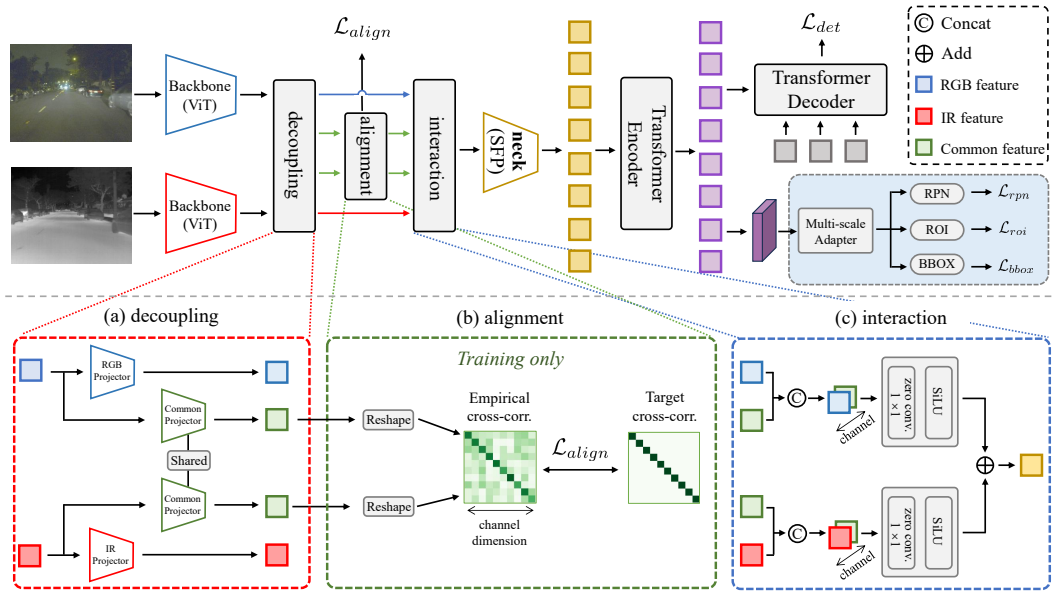


Figure 1: **Overview of ART**. (a) **Decoupling stage**. Dual backbones extract multi-scale features from RGB and IR inputs which are explicitly decomposed into a cross-modal *common semantic space* and modality-specific *unique spaces* via shared and modality-dependent projection heads. This structured decomposition separates modality-invariant semantics from heterogeneous details before interaction. (b) **Common-space alignment**. During training only, we reshape and ℓ_2-normalize the common features. An identity-like regularization is imposed on the correlation matrix to maximize matched-channel agreement while suppressing cross-channel redundancy, yielding the alignment loss $\mathcal{L}_{align}$. (c) **Interaction modules**. After stabilization, the cross-modal interaction blocks $G^s(\cdot,\cdot)$ are independently applied in the common and unique spaces through channel-wise concatenation and pointwise mixing.

3.2 Can we explicitly decouple cross-modal common semantics from modality-specific details?

A key failure mode in RGBT detection is *premature feature collapse* Tian et al. (2026): aggressively mixing RGB and thermal features too early may blur modality boundaries and suppress complementary cues. Instead of treating fusion and interaction as a single entangled process, we ask:

*Can we disentangle what should be shared across modalities from what should remain modality-dependent **before** cross-modal interaction?*

The detail of subspace

For each scale s, we implement the common and unique projections using lightweight 1×1 conv heads. The common head ψ_c^s is shared for both modalities, while the unique heads $\psi_{u,RGB}^s$ and $\psi_{u,IR}^s$ are modality-specific.

$$\psi(F) = \mathrm{GN}\Big(\mathrm{Conv}_{1\times1}\big(\mathrm{SiLU}(\mathrm{GN}(\mathrm{Conv}_{1\times1}(F)))\big)\Big). \tag{4}$$

Each head adopts the same structure.

Given synchronized inputs (I_{RGB}, I_{IR}), dual-stream backbones Φ_{RGB} and Φ_{IR} extract multi-scale features $\{F_{RGB}^s\}_{s=1}^S$ and $\{F_{IR}^s\}_{s=1}^S$. For each scale s, we project features into one *shared common* subspace and two *modality-specific unique* subspaces:

$$C_{RGB}^s = \psi_c^s(F_{RGB}^s), \quad C_{IR}^s = \psi_c^s(F_{IR}^s), \tag{5}$$
$$U_{RGB}^s = \psi_{u,RGB}^s(F_{RGB}^s), \quad U_{IR}^s = \psi_{u,IR}^s(F_{IR}^s). \tag{6}$$

Here ψ_c^s is *shared across modalities* to encourage modality-invariant semantics, while $\psi_{u,RGB}^s$ and $\psi_{u,IR}^s$ are *not shared* to preserve heterogeneous modality cues. This structured decomposition makes subsequent learning *well-posed*: we **align** the common space for semantic agreement, while keeping the unique spaces unconstrained to preserve complementarity. Additional details of the subspaces are provided above 3.2.

3.3 Can we align semantics without destroying complementary information?

Alignment is beneficial only when applied to the *right* factors. Enforcing strong agreement everywhere may over-regularize and reduce diversity Fang et al. (2025); Wang & Isola (2022).

we ask: *Can we enforce cross-modal consistency **only** in the common semantic space, while keeping unique branches free to retain complementary details?*

Theory

Let $\hat{C}_{RGB}^s, \hat{C}_{IR}^s \in \mathbb{R}^{B\times C_s\times N_s}$ denote the spatially ℓ_2-normalized common features, where $N_s = H_sW_s$. For each non-degenerate channel, $\|\hat{C}_*^s(b,c,:)\|_2 = 1$.
For each batch element b, we define the cross-modal channel similarity matrix as

$$\mathbf{R}^s(b) = \hat{C}_{RGB}^s(b)\hat{C}_{IR}^s(b)^\top \in \mathbb{R}^{C_s\times C_s}, \tag{7}$$

where

$$R_{ij}^s(b) = \left\langle \hat{C}_{RGB}^s(b,i,:), \hat{C}_{IR}^s(b,j,:) \right\rangle \in [-1,1]. \tag{8}$$

Consider the common alignment loss

$$\mathcal{L}_{com}^s = \mathbb{E}_b\left[\|\mathbf{1} - \mathrm{diag}(\mathbf{R}^s(b))\|_2^2\right] + \lambda_c\mathbb{E}_b\left[\|\mathbf{R}^s(b) - \mathrm{diag}(\mathbf{R}^s(b))\|_F^2\right]. \tag{9}$$

Since both terms are non-negative, $\mathcal{L}_{com}^s \geq 0$. Moreover,

$$\mathcal{L}_{com}^s = 0 \quad\Longleftrightarrow\quad \mathbf{R}^s(b) = \mathbf{I}_{C_s}, \quad \forall b, \tag{10}$$

i.e., for every sample, corresponding RGB–thermal channels achieve maximal cosine agreement, $R_{ii}^s(b) = 1$, while non-corresponding channels satisfy $R_{ij}^s(b) = 0$ for $i \neq j$.
This zero-loss solution is attainable only when the identity-like correlation structure is geometrically feasible. In particular, because $\mathrm{rank}(\mathbf{R}^s(b)) \leq N_s$, an exact identity matrix requires $C_s \leq N_s$. When $C_s > N_s$, $\mathbf{R}^s(b) = \mathbf{I}_{C_s}$ is infeasible, and the objective instead encourages the closest feasible identity-like correlation structure.

Correlation-based alignment in the common space. For each scale s, we reshape and normalize common features $C_{RGB}^s, C_{IR}^s \in \mathbb{R}^{B\times C_s\times H_s\times W_s}$ into $\tilde{C}_{RGB}^s, \tilde{C}_{IR}^s \in \mathbb{R}^{B\times C_s\times N_s}$ with $N_s = H_sW_s$, and apply ℓ_2 normalization along the spatial dimension:

$$\hat{C}_*^s(b,c,:) = \frac{\tilde{C}_*^s(b,c,:)}{\max\left(\|\tilde{C}_*^s(b,c,:)\|_2, \epsilon\right)}, \tag{11}$$

where ϵ is a small constant used only to avoid numerical instability for degenerate channels. For non-degenerate channels, each spatial channel vector therefore has unit ℓ_2 norm.

We then compute the cross-modal channel correlation matrix $\mathbf{R}^s \in \mathbb{R}^{B\times C_s\times C_s}$ as

$$\mathbf{R}^s = \hat{C}_{RGB}^s(\hat{C}_{IR}^s)^\top. \tag{12}$$

Each entry $R_{ij}^s = \langle \hat{C}_{RGB}^s(i,:), \hat{C}_{IR}^s(j,:)\rangle$ is thus a cosine similarity satisfying $R_{ij}^s \in [-1,1]$. To promote semantic agreement while mitigating redundancy, we penalize deviations from an identity-like correlation:

$$\mathcal{L}_{com}^s = \underbrace{\mathbb{E}\left[\|\mathbf{1} - \mathrm{diag}(\mathbf{R}^s)\|_2^2\right]}_{\text{maximize matched-channel agreement}} + \lambda_c \underbrace{\mathbb{E}\left[\|\mathbf{R}^s - \mathrm{diag}(\mathbf{R}^s)\|_F^2\right]}_{\text{reduce off-diagonal redundancy}}. \tag{13}$$

The overall alignment loss sums across scales:

$$\mathcal{L}_{align} = \frac{1}{S}\sum_{s=1}^{S}\mathcal{L}_{com}^{s}. \tag{14}$$

Importantly, $\mathcal{L}_{align}$ is applied *only* to $(C_{RGB}^{s}, C_{IR}^{s})$, leaving the unique branches unconstrained to preserve complementary cues. Enforcing alignment on the unique features would inevitably wash out modality-specific details (e.g., fine-grained textures), undermining their role in retaining complementary information. The optimization procedure is provided in Algorithm **??**. About the Proof of Theory 3.3 can be found below 3.3. The full detail about 3.3 can be found in our supplemental materials.

Proof

Proof. Fix a scale s and batch index b. Let $u_c = \hat{C}_{RGB}^{s}(b, c, :)$ and $v_c = \hat{C}_{IR}^{s}(b, c, :) \in \mathbb{R}^{N_s}$. Spatial ℓ_2 normalization gives $\|u_c\|_2 = \|v_c\|_2 = 1$, and

$$r_{ij} = u_i^{\top} v_j, \qquad -1 \le r_{ij} \le 1. \tag{15}$$

(i) Diagonal term controls matched-channel discrepancy. For each channel c,

$$\|u_c - v_c\|_2^2 = 2 - 2u_c^{\top} v_c = 2(1 - r_{cc}). \tag{16}$$

Thus, driving r_{cc} toward 1 directly minimizes the discrepancy between corresponding RGB and thermal common channels.

(ii) Off-diagonal term controls cross-channel leakage. Let $U = [u_1, \ldots, u_{C_s}]$ and $V = [v_1, \ldots, v_{C_s}]$. Then

$$\|\mathbf{R}^s(b) - \mathrm{diag}(\mathbf{R}^s(b))\|_F^2 = \sum_{i \ne j} r_{ij}^2 = \sum_{i \ne j} (u_i^{\top} v_j)^2. \tag{17}$$

Hence, minimizing the off-diagonal term suppresses similarities between non-corresponding channels and reduces cross-channel leakage.

Combining (i) and (ii), $\mathcal{L}_{com}^{s}$ encourages high agreement between matched channels while suppressing off-diagonal correlations, yielding an identity-like cross-modal correlation structure whenever geometrically feasible.

3.4 Decoupled Interaction and Training Objective.

Heavy attention-based fusion can be computationally expensive and may re-entangle modalities, undoing the benefits of explicit decomposition Wang et al. (2025); El Ahmar et al. (2023).

Decoupled interaction. We use lightweight cross-modal interaction block $G^s(\cdot, \cdot)$ independently in each space:

$$X_{com}^{s} = G^{s}(C_{RGB}^{s}, C_{IR}^{s}), \tag{18}$$

$$X_{uni}^{s} = G^{s}(U_{RGB}^{s}, U_{IR}^{s}). \tag{19}$$

The final multi-scale features are aggregated (simply added) and fed into standard detection heads to predict $\{(b_i, c_i)\}_{i=1}^{N}$:

$$X = \mathrm{Agg}\Big(\{X_{com}^{s}\}_{s=1}^{S}, \{X_{uni}^{s}\}_{s=1}^{S}\Big), \quad \{(b_i, c_i)\}_{i=1}^{N} = \mathcal{H}(X). \tag{20}$$

Training objective. We jointly optimize the detection objective and the common-space alignment regularizer:

$$\mathcal{L} = \mathcal{L}_{det} + \alpha\, \mathcal{L}_{align}, \tag{21}$$

where α controls the contribution of the alignment regularizer. The alignment term encourages matched RGB–thermal common channels to achieve high cosine similarity while suppressing off-diagonal cross-channel correlations, subject to the feasible correlation structure at each feature scale. With the common representations thus regularized, the interaction module can remain lightweight and primarily facilitates cross-modal information exchange in the decoupled feature spaces. The implementation details of the interaction block are provided in Algorithm **??**.

Table 2: Comparison on the FLIR dataset. The red text represents the best previous results and the blue text represents the improvement of our model compared to the best previous results. '-' indicates unreported results.

Methods	Year	Source	mAP50↑	mAP↑
GAFF (Zhang et al., 2021)	2021	CVPR	72.9	36.6
CFT (Qingyun et al., 2021)	2021	arXiv	78.7	40.2
AMSF-net (Bao et al., 2022)	2022	PRCV	78.9	40.9
MFPT (Zhu et al., 2023)	2023	TITS	80.0	-
LRAF-Net (Fu et al., 2023)	2023	TNNLS	80.5	42.8
ICAFusion (Shen et al., 2023)	2024	PR	79.2	41.4
MMFN (Yang et al., 2024)	2024	TCSVT	80.8	-
RSDet (Acharjee et al., 2025)	2024	arXiv	81.1	41.4
UniRGB-IR (Yuan et al., 2025)	2024	ACM MM	81.4	44.1
YOLOXCPCF (Hu et al., 2024)	2024	TITS	82.1	44.6
SCFR (Sun et al., 2024)	2024	TITS	82.3	-
GM_DETR (Xiao et al., 2024)	2024	CVPR	83.9	45.8
MMPedestron (Zhang et al., 2024)	2024	ECCV	86.4	-
DAMSDet (Guo et al., 2024)	2024	ECCV	86.6	49.3
Fusion-Mamba (Dong et al., 2025)	2025	TMM	84.9	47.0
IRDFusion (Shen et al., 2025)	2025	PR	88.3	50.7
Ours	2026	-	**88.6**(↑**0.3**)	**51.6**(↑**0.9**)

Table 3: Comparison on the M^{3}FD dataset. The red text represents the best previous results and the blue text represents the improvement of our model compared to the best previous results. '-' indicates unreported results.

Methods	Year	Source	mAP50↑	mAP↑
DIDFuse (Zhao et al., 2020)	2020	arXiv	78.9	52.6
SDNet (Zhang & Ma, 2021)	2021	IJCV	79.0	52.9
RFNet (Xu et al., 2022)	2022	CVPR	79.4	53.2
TarDAL (Liu et al., 2022a)	2022	CVPR	80.5	54.1
SuperFusion (Tang et al., 2022)	2022	JAS	83.5	56.0
CDDFusion (Zhao et al., 2023)	2023	CVPR	81.1	54.3
IGNet (Li et al., 2023)	2023	MM	81.5	54.5
DAMSDet (Guo et al., 2024)	2024	ECCV	80.2	52.9
MMFN (Yang et al., 2024)	2024	TCSVT	86.2	–
ICAFusion (Shen et al., 2023)	2024	PR	90.8	60.9
Fusion-Mamba (Dong et al., 2025)	2025	TMM	88.0	61.9
IRDFusion (Shen et al., 2025)	2025	PR	90.8	61.9
Ours	2026	-	**91.2**(↑**0.4**)	**62.5**(↑**0.6**)

4 EXPERIMENT

4.1 DATASETS AND EVALUATION METRICS.

We evaluate our proposed ART on four public RGBT datasets as follows: FLIR (FLIR ADA, 2021) is a high-quality dataset for multispectral object detection, consisting of paired IR and RGB images. It is primarily used in autonomous driving and surveillance applications. FLIR includes object categories such as 'person', 'car', and 'bicycle'. With detailed annotations, it is suitable for research on crossmodal fusion/alignment between infrared and visible light modalities. M^{3}FD (Liu et al., 2022b) includes 4,200 infrared and visible light aligned image pairs collected under various environments such as different lighting, seasons, and weather scenarios. It encompasses six typical categories in automated driving and road surveillance. M^{3}FD is partitioned according to an 8:2 training/testing split as provided in (Liang et al., 2023). SeaDronesSee (Varga et al., 2022) is a maritime UAV dataset collected from aerial drone views under diverse viewpoints and flight attitudes. We use its object detection split, containing 8,930 training, 1,547 validation, and 3,750 test images, with five categories: swimmer, boat, jetski, life-saving appliance, and buoy. Folowing ICAFusion (Shen et al., 2023), LLVIP (Jia et al., 2021) is a paired RGB–IR dataset for pedestrian detection under low-light conditions. It is widely used to evaluate multispectral fusion methods in challenging illumination environments. *Mean Average Precision (mAP).* The COCO (Lin et al., 2015) mAP (mean average precision) evaluation metric is a standard for assessing object detection models. The mean AP is calculated across all classes and IoU thresholds (from 0.5 to 0.95, in increments of 0.05), offering a comprehensive evaluation of detection accuracy and localization precision. mAP50 measures AP at IoU = 0.5, focusing on sufficient overlap, while mAP75 evaluates AP at IoU = 0.75, requiring stricter localization.

4.2 IMPLEMENTATION DETAILS.

We use the double-co-detr framework from (Zong et al., 2023) for our experiments. All experiments are carried out on 8 $\times$ NVIDIA A800 GPUs (80GB). The image input size is set to 640 $\times$ 640, and the data augmentation is followed by the v1 version from Zhou et al. (2024), with all other settings remaining consistent with those in the original paper. Our experimental setup uses PyTorch 1.13.1 with CUDA 11.8, and a total batch size of 64. The λ_c is 0.01 while α is 0.4. The dimension of feature extracted from backbone(ViT) is 1024, and the s is 5. In the final experiments, the FLIR (FLIR ADA, 2021) and SeaDronesSee (Varga et al., 2022) datasets are trained for 12 epochs, while the M^{3}FD Liu et al. (2022b) and LLVIP (Jia et al., 2021) dataset are trained for 36 epochs.

4.3 MAIN RESULTS.

Table 4: Comparison on the LLVIP dataset. The red text represents the best previous results and the blue text represents the improvement of our model compared to best previous results. '-' is used to indicate papers that have not been reported the results.

Methods	Year	Source	mAP50↑	mAP↑
CFT (Qingyun et al., 2021)	2021	arXiv	97.5	63.6
LRAF-Net (Fu et al., 2023)	2023	TNNLS	97.9	66.3
UniRGB-IR (Yuan et al., 2025)	2024	ACM MM	96.1	63.2
YOLOXCPCF (Hu et al., 2024)	2024	TITS	96.4	65.2
AMSF-Net (Bao et al., 2022)	2022	PRCV	97.0	64.5
MMFN (Yang et al., 2024)	2024	TCSVT	97.2	–
GM_DETR (Xiao et al., 2024)	2024	CVPR	97.4	70.2
SCFR (Sun et al., 2024)	2024	TITS	97.5	–
DAMSDet (Guo et al., 2024)	2024	ECCV	97.9	69.6
ICAFusion (Shen et al., 2023)	2025	PR	98.4	64.5
Fusion-Mamba (Dong et al., 2025)	2025	TMM	97.0	64.3
Ours	2026	–	**98.6**(↑0.2)	**71.2**(↑1.0)

Table 5: Comparison on the SeaDronesSee dataset. The red text represents the best previous results and the blue text represents the improvement of our model compared to best previous results. '-' is used to indicate papers that have not been reported the results.

Methods	Year	Source	mAP50↑	mAP↑
Saliency (Recasens et al., 2018)	–	ECCV	73.9	45.1
FOVEA (Thavamani et al., 2021)	2021	ICCV	74.5	44.9
LZU (Thavamani et al., 2023)	2023	CVPR	75.0	45.0
ZoomDet (Wang et al., 2026)	2026	ISPRS	75.0	46.2
DEYOLO (Chen et al., 2024b)	2024	ICPR	49.2	27.8
CFT (Qingyun et al., 2021)	2021	arXiv	64.5	27.6
ICAFusion (Shen et al., 2023)	2025	PR	61.2	30.5
MCAFNet (Zheng et al., 2025)	2025	DSP	58.7	32.1
Ours	2026	–	**82.5**(↑7.5)	**48.0**(↑1.8)

Table 7: Ablation Study on key modules over FLIR Dataset. When there is no interaction module, we simply add the outputs of the neck as the input to the head.

Key Modules			FLIR		
Decouple	**Alignment**	**Interaction**	**mAP50**	**mAP75**	**mAP**
✓	✓	✓	88.6	49.2	51.6
✓	✓	✗	87.1(-1.5)	46.6(-2.6)	48.9(-2.7)
✓	✗	✗	85.1(-3.5)	44.4(-4.8)	47.1(-4.5)
✗	✗	✗	84.8(-3.8)	44.0(-5.2)	46.9(-4.7)

Table 6: Efficiency comparison on FLIR.

Model	FPS ↑	Params (M) ↓	GFLOPs ↓	mAP ↑
Double-Co-DETR	3.5	485.3	510.5	46.9
ART	2.9	677.6	1192.1	**51.6**
Absolute change	−0.6	+192.3	+681.6	**+4.7**
Relative change	−17.1%	+39.6%	+133.5%	–

We conduct comprehensive comparisons between ART and recent state-of-the-art RGBT object detectors on the **FLIR** and **M^{3}FD** benchmarks, as summarized in Tab. 2 and Tab. 3. Our ART framework achieves **51.6**% and **62.5**% mAP on FLIR and M3FD, respectively. Specifically, On the FLIR dataset, compared with the previous state-of-the-art model, we achieve improvements of **0.3** and **0.9** in mAP50 and mAP, respectively. On the M^{3}FD dataset, the corresponding gains are **0.4** and **0.6**, respectively. These results underscore the effectiveness of our ART framework design in handling diverse RGBT detection settings. In particular, by aligning only the common semantic subspace while preserving modality-specific unique cues, ART competes with but often surpasses existing state-of-the-art multi-modal detectors. The superior performance across both datasets highlights the strong adaptability ability of the proposed framework. As shown in Tab. 4 and Tab. 5, we further compare ART with recent state-of-the-art methods on **LLVIP** and **SeaDronesSee**. On LLVIP, ART achieves **98.6%** mAP50 and **71.2%** mAP, outperforming the previous best results by **0.2** and **1.0** points, respectively. On the more challenging SeaDronesSee dataset, ART reaches **82.5%** mAP50 and **48.0%** mAP, yielding substantial gains of **7.5** and **1.8** points over the previous best methods. In particular, the remarkable improvement in mAP50 on SeaDronesSee demonstrates that ART effectively learns discriminative complementary representations and generalizes well across diverse detection scenarios.

4.4 Ablation Study.

As shown in Tab. 7, we also conduct ablation experiments to quantify the contribution of each key module in ART. Starting from the baseline without any of our designs, we obtain 84.8 mAP50, 44.0 mAP75, 46.9 mAP. Enabling *Decouple* alone yields a consistent yet modest gain (84.8→85.1 mAP50, 44.0→44.4 mAP75, 46.9→47.1 mAP), indicating that explicitly separating common semantics from modality-specific details provides a cleaner representation for subsequent cross-modal learning. On top of Decouple, adding the *Alignment* module leads to a much larger improvement (85.1→87.1

mAP50, 44.4→46.6 mAP75, 47.1→48.9 mAP). This confirms our hypothesis that alignment becomes a key driver under true cross-modal settings, as it stabilizes the shared semantic subspace and prevents over-entanglement. Introducing *Interaction* further boosts performance to 88.6 mAP50, 49.2 mAP75, 51.6 mAP, suggesting that interaction is crucial for refining complementary cues and improving localization quality beyond pure alignment.

We further analyze the trade-off between computational complexity and detection performance. As shown in Tab. 6, compared with the Double-Co-DETR baseline, ART substantially improves the FLIR mAP from 46.9 to 51.6, yielding a **+4.7**-point gain in detection accuracy. This improvement comes with increased model complexity: the number of parameters increases from 485.3M to 677.6M (**+39.6%**), GFLOPs increase from 510.5 to 1192.1 (**+133.5%**), and the inference speed decreases from 3.5 to 2.9 FPS (-17.1%). Considering the substantial improvement in detection performance, the additional computational cost represents a reasonable accuracy–complexity trade-off. Notably, the common-space alignment objective is applied only during training and therefore does not introduce additional inference-time computation; the extra cost mainly arises from the common/unique feature decomposition and decoupled cross-modal interaction.

Table 8: Effect of common-space alignment across different backbone scales.

Backbone	Alignment	mAP50 ↑	mAP ↑	ΔAP ↑
ViT-Tiny	×	65.3	29.6	–
ViT-Tiny	✓	66.1	31.7	**+2.1**
ViT-Small	×	73.2	36.9	–
ViT-Small	✓	74.6	38.2	**+1.3**
ViT-Large	×	81.4	46.4	–
ViT-Large	✓	**82.5**	**48.0**	**+1.6**

To verify that the improvement of ART is not simply attributed to a large backbone, we evaluate common-space alignment with ViT-Tiny, ViT-Small, and ViT-Large. As shown in Tab. 8, Alignment consistently improves all backbone scales by **+2.1**, **+1.3**, and **+1.6** mAP, respectively. Notably, ViT-Tiny obtains the largest gain, indicating that the effectiveness of ART does not rely on increased model capacity. Instead, explicitly aligning cross-modal common semantics provides consistent benefits across different backbone scales.The Visualization and Limitation could be found in Appendix Sec 10, 7. In addition, the more ablation studies can be found in Appendix Sec 6.

4.5 Ablation study on the Hyper-parameter α.

To investigate the effect of the alignment weight α in Eq. 21, we evaluate ART using a predefined set of α values. As shown in Fig. 2, ART achieves the best mAP when $\alpha = 0.4$. A small α provides insufficient regularization of the common space, whereas an excessively large α overemphasizes the alignment objective relative to the detection objective and leads to degraded detection performance. These results indicate that a moderate alignment weight provides the best trade-off between cross-modal semantic consistency and task-specific feature learning.

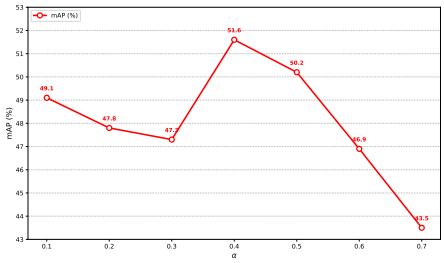


Figure 2: Ablation on the hyper-parameter α in Eq. 21.

5 Conclusion

In this work, we presented **ART**, a decoupled framework for RGB–Thermal object detection that explicitly separates *alignment* from *interaction* to mitigate premature cross-modal feature collapse. Instead of enforcing indiscriminate agreement, we proposed a correlation-based common-space alignment that encourages matched-channel semantic consistency while suppressing cross-channel leakage, leading to a better-conditioned shared subspace for subsequent fusion. Extensive experiments on public RGBT benchmarks demonstrate that ART consistently improves detection performance and achieves strong competitiveness against state-of-the-art methods. Qualitative visualizations further verify that our model learns more structured and reliable cross-modal correspondences.

AI USE STATEMENT

Generative AI tools were used to assist with manuscript editing, figure preparation, and discussion of mathematical arguments. All AI-assisted content was reviewed and verified by the authors against the relevant sources and experimental artifacts. The authors take full responsibility for the final manuscript.

REFERENCES

Suvojit Acharjee, Singhan Ganguly, Asfak Ali, and Jaroslav Frnda. Comprehensive performance analysis of object detection algorithms across diverse sensor modalities. *IEEE Sensors Reviews*, 2025.

Wei Bao, Meiyu Huang, Jingjing Hu, and Xueshuang Xiang. Attention-guided multi-modal and multi-scale fusion for multispectral pedestrian detection. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pp. 382–393. Springer, 2022.

Chen Chen, Jiahao Qi, Xingyue Liu, Kangcheng Bin, Ruigang Fu, Xikun Hu, and Ping Zhong. Weakly misalignment-free adaptive feature alignment for uavs-based multimodal object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 26826–26835. IEEE, 2024a. doi: 10.1109/CVPR52733.2024.02534. URL https://doi.org/10.1109/CVPR52733.2024.02534.

Xiangqi Chen, Xiuling Zhang, Chengzhuan Yang, Li Zhao, Dawei Zhang, Yanchao Wang, Liyuan Chen, Hua Wang, Hao Peng, and Zhonglong Zheng. Protohgf-net: Prototype hypergraph fusion with intra-modal calibration for rgbt object detection, 2026. URL https://arxiv.org/abs/2608.11595.

Yishuo Chen, Boran Wang, Xinyu Guo, Wenbin Zhu, Jiasheng He, Xiaobin Liu, and Jing Yuan. Deyolo: Dual-feature-enhancement yolo for cross-modality object detection. In *International conference on pattern recognition*, pp. 236–252. Springer, 2024b.

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*, 2025.

Wassim El Ahmar, Yahya Massoud, Dhanvin Kolhatkar, Hamzah AlGhamdi, Mohammad Alja'Afreh, Riad Hammoud, and Robert Laganiere. Enhanced thermal-rgb fusion for robust object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 365–374, 2023.

Wanlong Fang, Tianle Zhang, and Alvin Chan. To align or not to align: Strategic multimodal representation alignment for optimal performance, 2025. URL https://arxiv.org/abs/2511.12121.

Team FLIR ADA. Free teledyne flir thermal dataset for algorithm training, 2021. URL https://www.flir.com/oem/adas/adas-dataset-form/.

Haolong Fu, Shixun Wang, Puhong Duan, Changyan Xiao, Renwei Dian, Shutao Li, and Zhiyong Li. Lraf-net: Long-range attention fusion network for visible–infrared object detection. *IEEE Transactions on Neural Networks and Learning Systems*, 35(10):13232–13245, 2023.

Junjie Guo, Chenqiang Gao, Fangcen Liu, Deyu Meng, and Xinbo Gao. Damsdet: Dynamic adaptive multispectral detection transformer with competitive query selection and adaptive feature fusion. In *European Conference on Computer Vision*, pp. 464–481. Springer, 2024.

Sijie Hu, Fabien Bonardi, Samia Bouchafa, Helmut Prendinger, and Désiré Sidibé. Rethinking self-attention for multispectral object detection. *IEEE Transactions on Intelligent Transportation Systems*, 25(11):16300–16311, 2024.

Xinyu Jia, Chuang Zhu, Minzhen Li, Wenqi Tang, and Wenli Zhou. Llvip: A visible-infrared paired dataset for low-light vision. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 3489–3497. IEEE, 2021.

Pujian Lai, Dong Gao, Shilei Wang, and Gong Cheng. Mining representative tokens via transformer-based multi-modal interaction for rgb-t tracking. *Pattern Recognition*, pp. 112162, 2025.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. *arXiv preprint arXiv:2308.00692*, 2023.

Jiawei Li, Jiansheng Chen, Jinyuan Liu, and Huimin Ma. Learning a graph neural network with cross modality interaction for image fusion. In *Proceedings of the 31st ACM international conference on multimedia*, pp. 4471–4479, 2023.

Yanshu Li, Jianjiang Yang, Ziteng Yang, Bozheng Li, Ligong Han, Hongyang He, Zhengtao Yao, Yingjie Victor Chen, Songlin Fei, Dongfang Liu, and Ruixiang Tang. Make lvlms focus: Context-aware attention modulation for better multimodal in-context learning, 2025a. URL https://arxiv.org/abs/2505.17097.

Yubin Li, Weida Zhan, Yichun Jiang, and Jinxin Guo. Rdcrnet: Rgb-t object detection network based on cross-modal representation model. *Entropy*, 27, 2025b. URL https://api.semanticscholar.org/CorpusID:277988699.

Zhengdao Li, Siheng Wang, Zeyu Zhang, and Hao Tang. Remomask: Retrieval-augmented masked motion generation. *arXiv preprint arXiv:2508.02605*, 2025c.

Mingjian Liang, Junjie Hu, Chenyu Bao, Hua Feng, Fuqin Deng, and Tin Lun Lam. Explicit attention-enhanced fusion for rgb-thermal perception tasks, 2023. URL https://arxiv.org/abs/2303.15710.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. URL https://arxiv.org/abs/1405.0312.

Jinyuan Liu, Xin Fan, Zhanbo Huang, Guanyao Wu, Risheng Liu, Wei Zhong, and Zhongxuan Luo. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5802–5811, 2022a.

Jinyuan Liu, Xin Fan, Zhanbo Huang, Guanyao Wu, Risheng Liu, Wei Zhong, and Zhongxuan Luo. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5802–5811, 2022b.

Zhengyi Liu, Yacheng Tan, Qian He, and Yun Xiao. Swinnet: Swin transformer drives edge-aware rgb-d and rgb-t salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(7):4486–4497, 2021.

Ruoyan Pei, Pengge Ma, Jinwang Qian, Hai Yang, and Junling Sun. Learning to recover weak signals: A two-stage graph-based framework for all-weather rgb-t object detection. *PloS one*, 21 (8):e0354466, 2026.

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

Adria Recasens, Petr Kellnhofer, Simon Stent, Wojciech Matusik, and Antonio Torralba. Learning to zoom: a saliency-based sampling layer for neural networks. In *european conference on computer vision*, pp. 52–67. Springer, 2018.

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*, pp. 109913, 2023. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2023.109913.

Jifeng Shen, Haibo Zhan, Xin Zuo, Heng Fan, Xiaohui Yuan, Jun Li, and Wankou Yang. Irdfusion: Iterative relation-map difference guided feature fusion for multispectral object detection, 2025. URL https://arxiv.org/abs/2509.09085.

Xiaoyu Sun, Yaohui Zhu, and Hua Huang. Specificity-guided cross-modal feature reconstruction for rgb-infrared object detection. *IEEE Transactions on Intelligent Transportation Systems*, 26(1): 950–961, 2024.

Linfeng Tang, Yuxin Deng, Yong Ma, Jun Huang, and Jiayi Ma. Superfusion: A versatile image registration and fusion network with semantic awareness. *IEEE/CAA Journal of Automatica Sinica*, 9(12):2121–2137, 2022.

Chittesh Thavamani, Mengtian Li, Nicolas Cebron, and Deva Ramanan. Fovea: Foveated image magnification for autonomous navigation. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15519–15528. IEEE, 2021.

Chittesh Thavamani, Mengtian Li, Francesco Ferroni, and Deva Ramanan. Learning to zoom and unzoom. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5086–5095. IEEE, 2023.

Chao Tian, Zikun Zhou, Chao Yang, Guoqing Zhu, Fu'an Zhong, and Zhenyu He. Modality-decoupled rgb-thermal object detector via query fusion, 2026. URL https://arxiv.org/abs/2601.08458.

Leon Amadeus Varga, Benjamin Kiefer, Martin Messmer, and Andreas Zell. Seadronessee: A maritime benchmark for detecting humans in open water. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 3686–3696. IEEE, 2022.

Jinzhong Wang, Xuetao Tian, Shun Dai, Tao Zhuo, Haorui Zeng, Hongjuan Liu, Jiaqi Liu, Xiuwei Zhang, and Yanning Zhang. Rgb-t object detection via group shuffled multi-receptive attention and multi-modal supervision, 2024. URL https://arxiv.org/abs/2405.18955.

Shiyao Wang, Yucong Zhou, Junjie Yan, and Zhidong Deng. Fully motion-aware network for video object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 542–557, 2018.

Sirui Wang, Guiling Sun, Liang Dong, and Bowen Zheng. Carnet: Cross-attention guided feature reconstruction for rgb-t object detection. *Expert Systems with Applications*, pp. 129865, 2025.

Tao Wang, Chenyu Lin, Chenwei Tang, Jizhe Zhou, Deng Xiong, Jianan Li, Jian Zhao, and Jiancheng Lv. Adaptive image zoom-in with bounding box transformation for uav object detection. *ISPRS Journal of Photogrammetry and Remote Sensing*, 233:452–466, 2026.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere, 2022. URL https://arxiv.org/abs/2005.10242.

Yiming Xiao, Fanman Meng, Qingbo Wu, Linfeng Xu, Mingzhou He, and Hongliang Li. Gm-detr: Generalized muiltispectral detection transformer with efficient fusion encoder for visible-infrared detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5541–5549, 2024.

Yun Xiao, Mengmeng Yang, Chenglong Li, Lei Liu, and Jin Tang. Attribute-based progressive fusion network for rgbt tracking. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 2831–2838, 2022.

Han Xu, Jiayi Ma, Jiteng Yuan, Zhuliang Le, and Wei Liu. Rfnet: Unsupervised network for mutually reinforcing multi-modal image registration and fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19679–19688, 2022.

Fan Yang, Binbin Liang, Wei Li, and Jianwei Zhang. Multidimensional fusion network for multi-spectral object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(1): 547–560, 2024.

Senqiao Yang, Tianyuan Qu, Xin Lai, Zhuotao Tian, Bohao Peng, Shu Liu, and Jiaya Jia. An improved baseline for reasoning segmentation with large language model. *arXiv preprint arXiv:2312.17240*, 2023.

Maoxun Yuan, Bo Cui, Tianyi Zhao, Jiayi Wang, Shan Fu, Xue Yang, and Xingxing Wei. Unirgb-ir: A unified framework for visible-infrared semantic tasks via adapter tuning. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 2409–2418, 2025.

Hao Zhang and Jiayi Ma. Sdnet: A versatile squeeze-and-decomposition network for real-time image fusion. *International Journal of Computer Vision*, 129(10):2761–2785, 2021.

Heng Zhang, Elisa Fromont, Sébastien Lefèvre, and Bruno Avignon. Guided attentive feature fusion for multispectral pedestrian detection. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 72–80, 2021.

Qiang Zhang, Nianchang Huang, Lin Yao, Dingwen Zhang, Caifeng Shan, and Jungong Han. Rgb-t salient object detection via fusing multi-level cnn features. *IEEE Transactions on Image Processing*, 29:3321–3335, 2019.

Yan Zhang, Chang Xu, Wen Yang, Guangjun He, Huai Yu, Lei Yu, and Gui-Song Xia. Drone-based rgbt tiny person detection. *ISPRS Journal of Photogrammetry and Remote Sensing*, 204:61–76, 2023. doi: https://doi.org/10.1016/j.isprsjprs.2023.08.016. URL https://www.sciencedirect.com/science/article/pii/S0924271623002319.

Yi Zhang, Wang Zeng, Sheng Jin, Chen Qian, Ping Luo, and Wentao Liu. When pedestrian detection meets multi-modal learning: Generalist model and benchmark dataset. In *European Conference on Computer Vision*, pp. 430–448. Springer, 2024.

Zixiang Zhao, Shuang Xu, Chunxia Zhang, Junmin Liu, Pengfei Li, and Jiangshe Zhang. Didfuse: Deep image decomposition for infrared and visible image fusion. *arXiv preprint arXiv:2003.09210*, 2020.

Zixiang Zhao, Haowen Bai, Jiangshe Zhang, Yulun Zhang, Shuang Xu, Zudi Lin, Radu Timofte, and Luc Van Gool. Cddfuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5906–5916, 2023.

Shangpo Zheng, Liu Junfeng, and Jun Zeng. Mcafnet: Multiscale cross-modality adaptive fusion network for multispectral object detection. *Digital Signal Processing*, 159:104996, 2025.

Chen Zhou, Peng Cheng, Junfeng Fang, Yifan Zhang, Yibo Yan, Xiaojun Jia, Yanyan Xu, Kun Wang, and Xiaochun Cao. Optimizing multispectral object detection: A bag of tricks and comprehensive benchmarks. *arXiv preprint arXiv:2411.18288*, 2024.

Wujie Zhou, Qinling Guo, Jingsheng Lei, Lu Yu, and Jenq-Neng Hwang. Ecffnet: Effective and consistent feature fusion network for rgb-t salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(3):1224–1235, 2021.

Yaohui Zhu, Xiaoyu Sun, Miao Wang, and Hua Huang. Multi-modal feature pyramid transformer for rgb-infrared object detection. *IEEE Transactions on Intelligent Transportation Systems*, 24(9): 9984–9995, 2023.

Zhuofan Zong, Guanglu Song, and Yu Liu. Detrs with collaborative hybrid assignments training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6748–6758, 2023.

Table 9: Ablation Study on Two Modalities(FLIR). Blue text indicates our model's performance change on the RGB modality relative to the baseline, green text indicates the change on the IR modality, and red text indicates the change on both modalities.

Method	Modality	mAP50				mAP75				mAP			
		person	car	bicycle	all	person	car	bicycle	all	person	car	bicycle	all
Baseline	RGB	76.8	86.3	61.0	74.7	25.4	51.9	10.7	29.3	34.3	50.9	22.7	36.0
	IR	88.9	93.0	70.4	84.1	53.0	68.6	15.1	45.6	51.8	62.9	28.2	47.6
	RGB+IR	89.1	92.4	72.9	84.8	46.2	70.4	15.5	44.0	48.5	62.5	28.7	46.9
Ours	RGB	76.9 (+0.1)	86.5 (+0.2)	61.0 (+0.0)	74.8 (+0.1)	25.5 (+0.1)	51.9 (+0.0)	10.7 (+0.0)	29.3 (+0.0)	34.4 (+0.1)	50.9 (+0.0)	22.7 (+0.0)	36.0 (+0.0)
	IR	88.9 (+0.0)	93.1 (+0.1)	70.5 (+0.1)	84.1 (+0.0)	53.0 (+0.0)	68.7 (+0.1)	15.2 (+0.1)	45.7 (+0.1)	51.8 (+0.0)	62.9 (+0.0)	28.2 (+0.0)	47.6 (+0.0)
	RGB+IR	89.7 (+0.6)	92.9 (+0.5)	83.1 (+10.2)	88.6 (+3.8)	54.8 (+8.6)	71.2 (+0.8)	21.7 (+6.2)	49.2 (+5.2)	52.6 (+4.1)	64.7 (+2.2)	37.5 (+8.8)	51.6 (+4.7)

APPENDIX

6 ADDITIONAL EXPERIMENTS

6.1 ABLATION STUDY ON THE MODALITIES.

We conduct a modality ablation study to quantify how each input modality contributes to ART, as shown in Tab. 9. When using **a single modality** as input, i.e., **RGB-only** or **IR-only**, ART yields *almost no improvement* over the corresponding baseline: the overall mAP remains unchanged (RGB: 36.0→36.0; IR: 47.6→47.6), and the per-class metrics only show marginal gains (blue and green texts). This behavior is expected because, under single-modality input, the two streams carry the same information, and are thus *naturally aligned*; consequently, the alignment module brings little benefit, and the tiny changes mainly come from the *interaction* design. In contrast, when enabling **dual-modality** input (RGB+IR, red texts), ART exhibits a *substantial* performance boost: the overall scores improve by +**3.8** mAP50, +**5.2** mAP75, and +**4.7** mAP. Notably, the improvements are most pronounced for the **bicycle** category (mAP50: +**10.2**, mAP: +**8.8**), suggesting that bicycles can be visually indistinct in either modality alone (e.g., low contrast in RGB or weak thermal response), and only by *aligning* and *interacting* the complementary cues across modalities can the detector recover salient object evidence. These results confirm our key claim: the advantage of **alignment** becomes evident *only* under true cross-modal settings, where it facilitates effective fusion without collapsing complementary modality-specific cues.

6.2 ROBUSTNESS TO TRAINING-TIME MISREGISTRATION.

To further evaluate the robustness of ART to imperfect cross-modal registration, we introduce spatial shifts into the RGB–thermal pairs throughout training and evaluate all models on the original aligned validation set. Different from test-time perturbation, this setting directly exposes the alignment objective to persistently misregistered feature pairs, thereby isolating the effect of optimization-time correspondence errors on the learned representation.

As shown in Tab. **??**, ART remains stable under mild and moderate training-time misregistration. With persistent 10- and 20-pixel shifts, ART retains **97.7%** and **96.0%** of its aligned performance, respectively, while still outperforming the aligned same-detector baseline by **6.1** and **4.7** mAP50 points. Even with a 40-pixel shift, ART preserves **92.1%** of its aligned performance and remains 1.5 points above the baseline. These results indicate that imperfect cross-modal correspondence during optimization does not immediately invalidate the representation learned by ART.

This tolerance is consistent with the design of ART. The alignment objective operates on multi-scale common feature representations rather than directly enforcing pixel-level RGB–IR correspondence. Moreover, alignment is applied only to the common semantic space, while the modality-specific unique branches remain unconstrained, allowing them to preserve complementary unimodal evidence when cross-modal correspondence becomes inaccurate. However, under the severe 80-pixel training shift, mAP50 drops to 65.3, showing that persistent large displacement introduces substantial incorrect correspondence into the alignment supervision. Therefore, ART is robust to mild and mod-

Table 10: Results over 10 independent random seeds.

Dataset	Metric	Previous Best	ART (Mean ± Std.)	95% CI	Improvement
FLIR	mAP50	88.3	$\mathbf{88.60 \pm 0.08}$	$[88.54, 88.66]$	$\mathbf{+0.3}$
FLIR	mAP	50.7	$\mathbf{51.60 \pm 0.04}$	$[51.57, 51.63]$	$\mathbf{+0.9}$
M^3FD	mAP50	90.8	$\mathbf{91.20 \pm 0.06}$	$[91.16, 91.24]$	$\mathbf{+0.4}$
M^3FD	mAP	61.9	$\mathbf{62.50 \pm 0.09}$	$[62.44, 62.56]$	$\mathbf{+0.6}$

erate training-time misregistration, but should not be interpreted as invariant to arbitrary geometric misalignment.

6.3 Stability across Random Seeds.

To verify that the improvement of ART is not caused by a favorable random initialization, we repeat the complete training and evaluation procedure using **10 independent random seeds**, while keeping all other training settings unchanged. As shown in Tab. 10, ART exhibits highly stable performance on both FLIR and M^3FD, with standard deviations ranging from only **0.04 to 0.09**. Moreover, the improvements over the previous best results (**0.3–0.9** points) are consistently larger than the observed run-to-run variation, indicating that the gains are not driven by a single favorable seed.

The 95% confidence intervals are also very narrow, and their lower bounds remain above the corresponding reported point estimates of the previous best methods. For example, ART achieves 51.60 ± 0.04 mAP on FLIR with a 95% CI of $[51.57, 51.63]$, compared with the previous best result of 50.7. The same trend is observed on M^3FD. These results provide further evidence that the improvements of ART are stable and reproducible across different random initializations and data orders.

For statistical precision, we note that a formal paired or two-sample significance test against each previous method cannot be performed, because their per-seed results and variances are not publicly available. Therefore, we do not interpret these results as a formal cross-method significance test. Instead, the low variance and narrow confidence intervals across our 10 runs support the conclusion that the observed improvements are reproducible rather than attributable to a single lucky run.

7 Limitations

As shown in Fig. **??**, the red region at the bottom of the figure is not due to dispersed attention. Instead, it contains a bicycle icon, which leads the model to treat this area as evidence of a bicycle. However, in reality, there is no actual bicycle in that region. This observation suggests that, without explicit semantic guidance, our model can be distracted by icon-like spurious cues. It also points to a promising future direction: integrating vision–language models (VLMs) Li et al. (2025a;c); Lai et al. (2023); Yang et al. (2023) to provide stronger semantic priors and improve the model's understanding of the real world.

8 Algorithm

9 Proof and explanation

This subsection provides a more comprehensive theoretical proof for the method in Section 3.3.

Proof and explanation

Proof. We provide a functional interpretation of $\mathcal{L}^s_{com}$ by showing that its two terms control (i) *matched-channel semantic consistency* and (ii) *cross-channel leakage* across modalities.

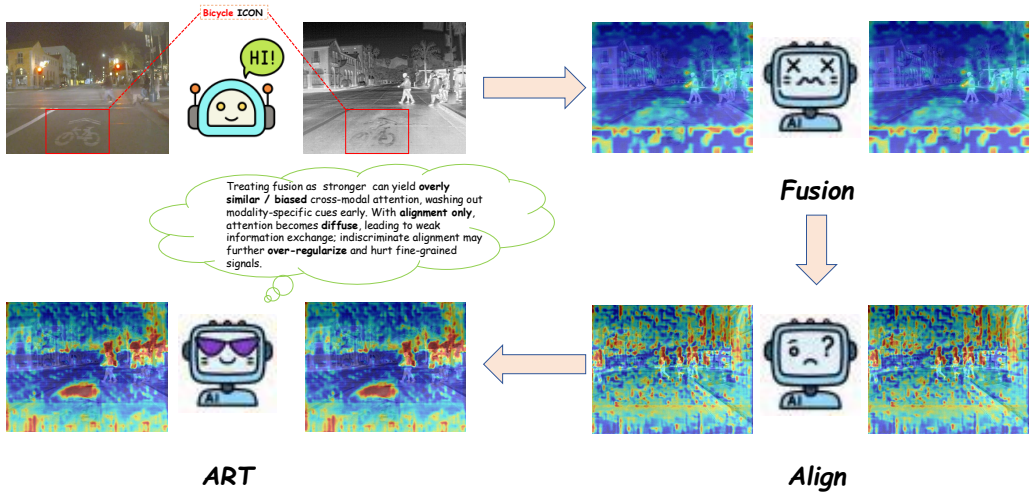


Figure 3: We visualize the attention maps of different methods during inference. Strong fusion tends to produce overly similar / biased cross-modal attention, while only alignment yields more diffuse attention; ART balances semantic agreement and complementary interaction. All attention heatmaps are taken from the same transformer layer, and redder regions indicate stronger attention.

Algorithm 1 ART Overview

Require: Synchronized pair (I_{RGB}, I_{IR}), detector $\mathcal{D}$ with parameters Θ
Ensure: Detection results $\{(b_i, c_i)\}_{i=1}^{N}$ or training loss $\mathcal{L}$

Dual-stream feature extraction:
$\{F_{RGB}^s\}_{s=1}^{S} \leftarrow \Phi_{RGB}(I_{RGB})$, $\{F_{IR}^s\}_{s=1}^{S} \leftarrow \Phi_{IR}(I_{IR})$.
Explicit decoupling:
$\{C_{RGB}^s, U_{RGB}^s\}_{s=1}^{S} \leftarrow \text{Decouple}(\{F_{RGB}^s\})$, $\{C_{IR}^s, U_{IR}^s\}_{s=1}^{S} \leftarrow \text{Decouple}(\{F_{IR}^s\})$.
Common-space stabilization:
$\mathcal{L}_{align} \leftarrow \text{CommonAlign}(\{C_{RGB}^s\}, \{C_{IR}^s\})$.
Decoupled interaction:
$\{X_{com}^s\}_{s=1}^{S} \leftarrow \text{Interact}(\{C_{RGB}^s\}, \{C_{IR}^s\})$, $\{X_{uni}^s\}_{s=1}^{S} \leftarrow \text{Interact}(\{U_{RGB}^s\}, \{U_{IR}^s\})$.
Aggregation and detection:
$X \leftarrow \text{Agg}(\{X_{com}^s\}, \{X_{uni}^s\})$, $\{(b_i, c_i)\}_{i=1}^{N} \leftarrow \mathcal{H}(X)$.
if training **then**
  $\mathcal{L} \leftarrow \mathcal{L}_{det} + \alpha\, \mathcal{L}_{align}$
  return $\mathcal{L}$
else
  return $\{(b_i, c_i)\}_{i=1}^{N}$
end if

Fix a scale s and a batch index b. Let

$$u_c = \hat{C}_{RGB}^s(b, c, :), \qquad v_c = \hat{C}_{IR}^s(b, c, :) \in \mathbb{R}^{N_s} \tag{22}$$

denote the c-th spatial channel vectors. For non-degenerate channels, spatial ℓ_2 normalization gives

$$\|u_c\|_2 = \|v_c\|_2 = 1. \tag{23}$$

Accordingly, the entries of the cross-modal correlation matrix are

$$r_{ij} = u_i^\top v_j, \tag{24}$$

which are cosine similarities satisfying

$$-1 \le r_{ij} \le 1. \tag{25}$$

(i) Diagonal term ⇒ matched-channel consistency. For a corresponding RGB–thermal channel pair, we have

$$\|u_c - v_c\|_2^2 = \|u_c\|_2^2 + \|v_c\|_2^2 - 2u_c^\top v_c \tag{26}$$

$$= 2 - 2r_{cc} = 2(1 - r_{cc}). \tag{27}$$

Hence,

$$1 - r_{cc} = \frac{1}{2}\|u_c - v_c\|_2^2. \tag{28}$$

Therefore, minimizing the diagonal penalty $(1 - r_{cc})^2$ directly encourages the corresponding normalized RGB and thermal channels to approach each other. Moreover, $r_{cc} = 1$ if and only if $u_c = v_c$, by the equality condition of the Cauchy–Schwarz inequality.

(ii) Off-diagonal term ⇒ cross-channel leakage control. Let

$$U = [u_1, \ldots, u_{C_s}] \in \mathbb{R}^{N_s \times C_s}, \qquad V = [v_1, \ldots, v_{C_s}] \in \mathbb{R}^{N_s \times C_s}. \tag{29}$$

Then

$$\mathbf{R}^s(b) = U^\top V. \tag{30}$$

The off-diagonal component of the objective is

$$\|\mathbf{R}^s(b) - \mathrm{diag}(\mathbf{R}^s(b))\|_F^2 = \sum_{i \neq j} r_{ij}^2 \tag{31}$$

$$= \sum_{i \neq j} \left(u_i^\top v_j\right)^2. \tag{32}$$

For each RGB channel i, define its cross-channel leakage energy as

$$\ell_i := \sum_{j \neq i} r_{ij}^2 = \sum_{j \neq i} \left(u_i^\top v_j\right)^2. \tag{33}$$

It follows that

$$\sum_{i \neq j} r_{ij}^2 = \sum_{i=1}^{C_s} \ell_i. \tag{34}$$

Thus, minimizing the off-diagonal term suppresses undesired similarities between non-corresponding channels and discourages redundant cross-channel responses in the common representation. We emphasize that this term encourages decorrelation rather than guaranteeing statistical independence between channels.

Optimality. Since both components of $\mathcal{L}_{com}^s$ are non-negative,

$$\mathcal{L}_{com}^s \geq 0. \tag{35}$$

For each sample, zero loss requires simultaneously

$$r_{ii} = 1, \qquad r_{ij} = 0 \ \ \forall i \neq j, \tag{36}$$

or equivalently,

$$\mathbf{R}^s = \mathbf{I}_{C_s}. \tag{37}$$

Because $r_{ii} = 1$ and $\|u_i\|_2 = \|v_i\|_2 = 1$ imply $u_i = v_i$, the zero-loss condition further implies

$$u_i^\top u_j = 0, \qquad i \neq j. \tag{38}$$

Hence, an exact zero-loss solution requires the matched channel vectors to form an orthonormal set in $\mathbb{R}^{N_s}$.

Consequently, such a solution is geometrically feasible only when

$$C_s \leq N_s. \tag{39}$$

Equivalently, since

$$\mathrm{rank}(\mathbf{R}^s) = \mathrm{rank}(U^\top V) \leq N_s, \tag{40}$$

$\mathbf{R}^s = \mathbf{I}_{C_s}$ cannot be attained when $C_s > N_s$. In this case, minimizing $\mathcal{L}^s_{com}$ should be interpreted as encouraging the closest feasible identity-like correlation structure rather than attaining an exact identity matrix.

Explanation. The diagonal term encourages corresponding RGB–thermal common channels to achieve high cosine similarity, promoting cross-modal semantic consistency. The off-diagonal term suppresses cosine similarity between non-corresponding channels, thereby reducing cross-channel leakage and redundant responses. Together, these two terms encourage an identity-like correlation structure whenever permitted by the feature dimensionality. Importantly, this regularization is applied only to the common features (C^s_{RGB}, C^s_{IR}), while the unique branches remain unconstrained so that modality-specific complementary cues can be preserved.

10 VISUALIZATION

To better understand how ART stabilizes the shared semantic subspace, we visualize the cross-modal channel similarity (correlation) matrix between the RGB and IR *common* features in Fig. 4. As can be observed, the similarity concentrates along the diagonal, forming a more identity-like pattern: matched channels exhibit higher agreement, while off-diagonal responses are notably suppressed. This qualitatively validates our correlation-based alignment objective, which explicitly encourages matched-channel consistency and penalizes cross-channel leakage/redundancy in the common space. Together with the quantitative ablations where adding alignment yields a clear gain, the visualization suggests that ART learns a better shared semantic space for subsequent interaction, enabling cross-modal synergy without collapsing modality-specific details.

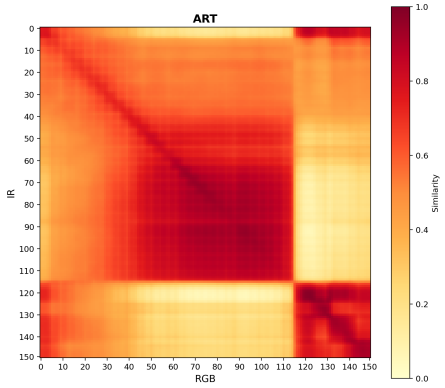


Figure 4: We visualize the cross-modal channel similarity matrix between RGB and IR common features.