RETHINKING UNALIGNED RGB-T DENSE PREDICTION AS CROSS-MODAL INFORMATION TRANSPORT

Anonymous authors
Paper under double-blind review

ABSTRACT

RGB-T dense prediction remains challenging because visible and thermal observations are rarely perfectly aligned in practice. When the two modalities are spatially inconsistent, the thermal response at a given coordinate may originate from a different scene region, and reweighting such co-located responses can only attenuate the mismatch rather than recover the complementary information that lies elsewhere. We therefore propose unaligned cross-modal information transport (UCMIT), which reformulates RGB-T interaction as a structured cross-modal information transport process. Within UCMIT, local cross-modal information transport (LCIT) constructs a semantic-geometric transport plan through entropy-regularized optimal transport, retrieving complementary thermal evidence from plausible nearby locations and reconstructing it in the visible coordinate system without explicit global registration. Source-side marginal constraints couple evidence allocation across RGB targets and prevent unrestricted reuse of dominant thermal responses, while a fixed window-level null mass is redistributed across targets according to their relative matching preference. Reliability-controlled evidence fusion (REF) then arbitrates between the co-located and transported responses according to the local transport state and regulates residual injection according to evidence reliability and path-aware certainty. LCIT thus determines where complementary evidence is obtained, while REF determines which evidence is retained and how strongly it modifies the visible representation. Extensive experiments on RGB-T salient object detection and semantic segmentation demonstrate strong overall performance across unaligned benchmarks.

1 INTRODUCTION

RGB-T dense prediction aims to infer pixel-level scene structures from paired visible and thermal observations (Tang et al., 2025). Representative tasks include salient object detection (SOD) and semantic segmentation (Zhao et al., 2024). These tasks benefit from the complementary sensing characteristics of the two modalities (Zhao et al., 2025). Visible images provide rich texture, color, and boundary details, yet these cues often deteriorate under low illumination and adverse weather. Thermal images are less sensitive to illumination changes and can preserve stable object responses when visible appearance becomes unreliable (Wang et al., 2026). Their complementary properties make RGB-T interaction valuable for robust dense scene understanding.

Many RGB-T dense prediction methods implicitly assume spatially aligned cross-modal features during local interaction (Zhou et al., 2025). In practical systems, however, residual cross-modal misalignment frequently violates this assumption (Wang et al., 2025). The thermal response at a given coordinate may therefore originate from a different scene region, causing position-wise fusion to introduce spatially mismatched information into the visible representation. More importantly, reweighting co-located responses can only attenuate such mismatch rather than recover complementary information from the correct thermal location (Gao et al., 2026). The key problem shifts from how cross-modal features are fused to where complementary thermal evidence is retrieved.

Existing methods alleviate cross-modal misalignment through geometric transformation (Lai et al., 2024), deformable sampling (Liu et al., 2026), correlation matching (Hu et al., 2026) and local interaction (Wang et al., 2025). Explicit registration can compensate for global displacement, but its effectiveness degrades under parallax and locally varying misalignment. Registration-free methods

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

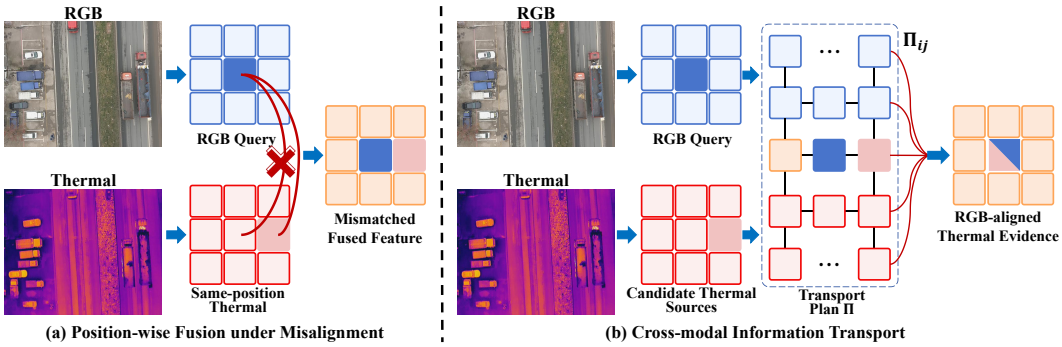

Figure 1: Motivation of UCMIT. Unlike position-wise fusion under cross-modal misalignment, UCMIT retrieves complementary thermal evidence from plausible nearby locations and reconstructs it in RGB coordinates before representation update.

instead search for related thermal responses in feature space, improving tolerance to spatial inconsistency. However, correspondence estimation is typically coupled with immediate feature aggregation, so matched thermal responses are directly introduced into the visible stream. This coupling prevents complementary evidence retrieval from being separated from the subsequent representation update. Local displacement may also move the true counterpart outside the search neighborhood, making forced correspondence potentially unreliable. **Therefore, reliably retrieving and reconstructing RGB-aligned thermal evidence while explicitly accounting for less plausible local correspondence constitutes the first key challenge of unaligned RGB-T dense prediction.**

Reliable correspondence alone does not guarantee reliable cross-modal fusion. Retrieved thermal evidence may remain degraded or ambiguous even when the spatial correspondence is plausible (Guo & Jin, 2025). Similar responses within a local region can further increase correspondence uncertainty. Meanwhile, reassignment becomes unnecessary when the co-located thermal response already agrees with the visible feature, and replacing such valid evidence may instead introduce additional errors. These cases indicate that the available thermal evidence requires explicit evaluation before it modifies the visible representation (Du et al., 2026). **Thus, deciding which thermal evidence to retain and how strongly it should update the RGB representation constitutes the second key challenge of unaligned RGB-T dense prediction.**

To address these challenges, we propose UCMIT for robust unaligned RGB-T dense prediction. UCMIT reformulates cross-modal interaction from fixed position-wise fusion into a structured information transport process. Specifically, local cross-modal information transport (LCIT) constructs a semantic-geometric transport plan, where source-side marginal constraints couple thermal evidence allocation across RGB targets and a fixed null mass is redistributed across targets according to their relative matching preference within each window. LCIT reconstructs RGB-aligned thermal evidence before reliability-controlled evidence fusion (REF) regulates the subsequent representation update. REF arbitrates between co-located and transported responses according to the local transport state and controls residual injection based on evidence reliability and path-aware certainty. LCIT therefore determines where complementary thermal evidence is retrieved, while REF determines which evidence is retained and how strongly it updates the RGB representation. The same multi-scale design is applied to separately trained RGB-T salient object detection and semantic segmentation networks. Our main contributions are summarized as follows:

- We propose UCMIT to reformulate unaligned RGB-T interaction as a two-stage process that separates evidence retrieval from representation update. LCIT implements the retrieval stage through structured local transport that reconstructs RGB-aligned thermal evidence while redistributing a fixed null mass according to relative matching preference within each local window.
- We further propose REF to implement the subsequent representation update stage by determining which thermal evidence should be retained and how strongly it should modify the RGB representation. REF balances co-located and transported responses according to

the local transport state before regulating residual injection using evidence reliability and path-aware certainty for stable cross-modal representation update.

- Extensive experiments on RGB-T salient object detection and semantic segmentation demonstrate the effectiveness and generality of UCMIT across different dense prediction tasks. UCMIT consistently achieves strong performance on both salient object detection and multi-class semantic segmentation benchmarks.

2 Related Work

2.1 RGB-T Dense Prediction

RGB-T salient object detection (Tu et al., 2021) and semantic segmentation (Ha et al., 2017) exploit complementary visible and thermal cues for pixel-level scene understanding. Early methods mainly aggregate multi-level features (Sun et al., 2019), while later approaches introduce cross-modal attention (Wang et al., 2021), semantic guidance (Li et al., 2023), and reliability-aware modeling (Cong et al., 2022) for more selective fusion. Transformer-based architectures (Liu et al., 2021) enhance global context modeling, while state-space models (He et al., 2025) strengthen long-range dependency propagation. Despite these advances, most methods implicitly assume spatially consistent RGB-T features. In practice, sensor parallax, viewpoint differences, and imperfect calibration often violate this assumption (Zhou et al., 2025). This motivates our formulation of RGB-T interaction as cross-modal information transport beyond fixed spatial correspondence.

2.2 Unaligned Cross-Modal Fusion

Existing methods address cross-modal misalignment mainly through explicit alignment and alignment-free interaction. The former estimates spatial correspondence using affine transformation (Tu et al., 2022), deformable sampling (Liu et al., 2026), correlation matching (Hu et al., 2026), and feature warping (Lai et al., 2024). The latter searches for related responses directly in feature space through local cross-modal interaction (Wang et al., 2024a), learned fusion of spatially inconsistent observations (Wang et al., 2025), and latent relational modeling (Wang et al., 2026). These methods improve robustness to spatial inconsistency, yet the estimated correspondence is typically consumed immediately during feature aggregation. As a result, the retrieved evidence cannot be independently examined before affecting the target representation. In contrast, UCMIT separates correspondence construction from representation update, enabling transported thermal evidence to be assessed before modifying the RGB stream.

2.3 Optimal Transport for Visual Correspondence

Entropy-regularized optimal transport provides differentiable structured matching under assignment constraints (Cuturi, 2013). In visual correspondence, Sinkhorn-based formulations have been used for optimal assignment with unmatched handling (Sarlin et al., 2020), dense matching (Sun et al., 2021), and semantic correspondence (Liu et al., 2020). UCMIT does not treat the resulting assignment as the final correspondence output. Instead, the transport plan is used to barycentrically reconstruct thermal evidence in RGB coordinates, after which REF explicitly arbitrates it against the co-located path before any representation update. Thus, the contribution lies in using structured matching as an intermediate evidence-transport operator within a reconstruction–arbitration–update pipeline, rather than in the assignment mechanism itself.

3 Method

3.1 Overview

As illustrated in Fig. 2, UCMIT takes a visible image I_v and a thermal image I_t as input and extracts multi-scale features using two modality-specific encoders with identical architectures and independent parameters. At the l-th stage, the visible and thermal branches produce X^l and Y^l, respectively. LCIT constructs local RGB-to-thermal correspondence using semantic compatibility and geometric proximity. The co-located thermal response is retained as a direct path, while a null source carries

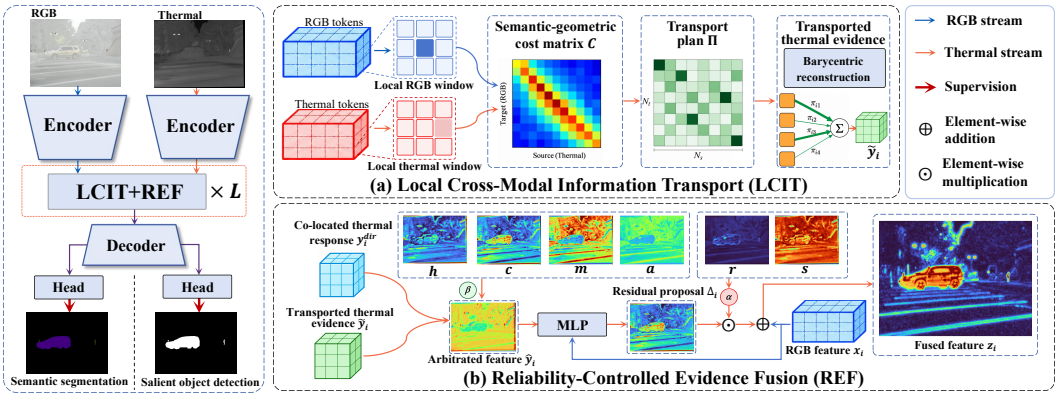

Figure 2: Overview of the proposed UCMIT framework. At each scale, LCIT reconstructs RGB-aligned thermal evidence from nearby thermal sources. REF then arbitrates between co-located and transported evidence and regulates its residual injection into the RGB stream. The resulting multi-scale representations are progressively aggregated for task-specific dense prediction.

a fixed window-level mass that is redistributed across targets according to their relative matching preference. Entropy-regularized optimal transport produces a structured plan Π^l, from which RGB-aligned thermal evidence $\widetilde{Y}^l$ is reconstructed. REF then arbitrates between the direct response Y^l_{dir} and transported evidence $\widetilde{Y}^l$, and regulates the resulting residual contribution to obtain the fused representation Z^l according to evidence reliability and path-aware certainty. The multi-scale representations $\{Z^l\}_{l=1}^L$ are progressively aggregated by a lightweight top-down decoder for task-specific prediction. The same framework is separately trained for RGB-T salient object detection and semantic segmentation with corresponding prediction heads.

3.2 Local Cross-Modal Information Transport (LCIT)

Cross-modal misalignment makes fixed co-located correspondence unreliable. LCIT instead retrieves complementary thermal evidence from nearby thermal locations within a local spatial window. We formulate this local correspondence as optimal transport, allowing source-side marginal constraints to jointly regulate evidence allocation across RGB targets.

Semantic-geometric cost. Consider a local window containing N RGB tokens $\{x_i\}_{i=1}^N$ and thermal tokens $\{y_j\}_{j=1}^N$. We use H transport heads indexed by $k = 1, \ldots, H$. For each head, the two modalities are projected into a shared metric space and normalized as $q_i^k = \text{Norm}(W_m^k x_i)$ and $k_j^k = \text{Norm}(W_m^k y_j)$. A single projection is applied to both modalities so that cosine similarity is measured in one common space rather than across two independently parameterized ones. Their matching cost combines semantic compatibility with spatial locality:

$$C_{ij}^k = 1 - (q_i^k)^\top k_j^k + \lambda_g d_{ij}, \tag{1}$$

where $d_{ij} = \|p_i - p_j\|_2^2 / d_{\max}$ is the squared distance between the window coordinates of x_i and y_j, normalized by the largest squared distance inside the window so that $d_{ij} \in [0, 1]$. The semantic term retrieves thermal responses according to feature compatibility, while the geometric term discourages unnecessarily distant assignments. We use a fixed geometric weight λ_g shared across all transport heads. We additionally mask the co-located entry by setting $C_{ii}^k = +\infty$. Leaving the co-located source inside the plan would subject it to the same supply competition as its neighbours, so a valid direct correspondence could be suppressed simply because nearby targets also demand that source. The co-located response is therefore excluded from the competition and retained as a separate path. We denote the semantic component of Eq. equation 1 as $C_{ij}^{\text{sem},k} = 1 - (q_i^k)^\top k_j^k$ and its head-averaged counterpart as $\overline{C}_{ij}^{\text{sem}} = \frac{1}{H} \sum_{k=1}^H C_{ij}^{\text{sem},k}$. The mask affects only the transport cost, leaving $\overline{C}_{ii}^{\text{sem}}$ available to measure co-located agreement.

Structured correspondence with fixed null allocation. Not every RGB location has a valid thermal counterpart inside the window, since a region may be occluded, may leave the neighbourhood under

displacement, or may lose its thermal signature entirely. We therefore augment the N thermal tokens with a null source and extend the cost to $C^k \in \mathbb{R}^{N\times(N+1)}$. Its rejection cost is predicted for each RGB target as $C^k_{i\varnothing} = f^k_\varnothing(q^k_i)$, where $f^k_\varnothing$ is a lightweight predictor. Together with the real-source matching costs, the target-dependent null cost determines how the fixed null budget is redistributed across RGB targets. Targets with relatively less favorable real-source matches can therefore receive a larger null allocation. The correspondence is obtained via entropy-regularized optimal transport:

$$\Pi^{\star,k} = \arg\min_{\Pi\in\mathcal{U}(a,b)} \langle \Pi, C^k\rangle - \tau_k \mathcal{H}(\Pi), \qquad \mathcal{U}(a,b) = \{\Pi \geq 0 \mid \Pi\mathbf{1} = a,\ \Pi^\top\mathbf{1} = b\}, \quad (2)$$

where $\mathcal{H}(\Pi) = -\sum_{i,j}\Pi_{ij}\log(\Pi_{ij}+\epsilon)$. The marginals couple target-side demand with source-side evidence allocation within each local window. We set $a = \frac{1}{N}\mathbf{1}_N$ so that every RGB target receives the same total transport mass. The source marginal is defined as $b = [\frac{1-\rho}{N}\mathbf{1}_N^\top,\ \rho]^\top$, which assigns each real thermal source a uniform supply of $(1-\rho)/N$ and reserves a fixed mass ρ for the null source. The uniform source marginal constrains excessive reuse of individual thermal sources rather than estimating their quality, while REF evaluates the reliability of the reconstructed evidence before injection. Since ρ fixes the total null mass within each window, the null source provides relative target-wise rejection allocation under a fixed window-level budget rather than adaptive estimation of overall window-level correspondence quality. We solve Eq. equation 2 using log-domain Sinkhorn iterations (Cuturi, 2013), with the per-head temperature τ_k learned within a bounded range. The head-wise plans are averaged as $\Pi^{\text{avg}} = \frac{1}{H}\sum_{k=1}^{H}\Pi^{\star,k}$ and row-normalized as $\bar{\Pi}_{ij} = \Pi^{\text{avg}}_{ij}/\sum_{j'}\Pi^{\text{avg}}_{ij'}$, where j' ranges over the thermal tokens and the null source.

Why structured transport? Let $K^k_{ij} = \exp(-C^k_{ij}/\tau_k)$. The entropic OT solution can be written in Sinkhorn form as $\Pi^{\star,k} = \text{diag}(u^k)K^k\text{diag}(v^k)$, where u^k and v^k are the target- and source-side scaling factors determined by the marginal constraints. After conditioning on a target row, the target-side factor cancels, yielding $\pi^{\text{row},k}_{ij} \propto \exp(-C^k_{ij}/\tau_k + \log v^k_j)$. Unlike independent row-wise matching, the source-side factor v^k_j is jointly determined by all RGB targets through the column constraint. Consequently, the assignment of one target depends on the aggregate demand placed on the same thermal sources by other targets, preventing unrestricted reuse of a locally dominant response. Setting $v^k_j = 1$ removes this source-side coupling and recovers ordinary row-wise softmax matching, which we evaluate as a controlled counterpart in Sec. 4.3.

Barycentric reconstruction. Because part of the mass may be absorbed by the null source, we separate how much usable correspondence is found from how it is distributed. We define the retained mass $m_i = \sum_{j\leq N}\bar{\Pi}_{ij} = 1 - \bar{\Pi}_{i\varnothing}$ and the retained plan $\widetilde{\pi}_{ij} = \bar{\Pi}_{ij}/(m_i+\epsilon)$, so that m_i reflects the relative non-null mass retained for target i within the window, while $\widetilde{\pi}_i$ describes how that retained evidence is distributed over real source locations. Since the total null mass is fixed within each window, m_i should be interpreted as a relative target-wise retention signal rather than an estimate of absolute window-level correspondence quality. Rather than using the plan as final fusion weights, we use it for barycentric reconstruction of thermal evidence in the RGB coordinate system:

$$\widetilde{y}_i = \sum_{j=1}^{N}\widetilde{\pi}_{ij}W_v y_j. \quad (3)$$

where W_v is a learnable value projection into the common feature space. Thus, $\widetilde{y}_i$ reconstructs RGB-aligned thermal evidence from plausible nearby sources beyond the co-located response y_i.

3.3 Reliability-Controlled Evidence Fusion (REF)

Cross-modal transport provides an alternative source of thermal evidence, but it should not replace a reliable co-located response at every position. Even after transport, the retrieved evidence may remain ambiguous or unreliable. UCMIT therefore retains both evidence paths and regulates their contribution before the RGB representation is updated.

Direct and transported evidence. REF arbitrates between the two paths according to the local transport state. Let $y^{\text{dir}}_i = W_v y_i$ denote the co-located thermal response. We characterize the transport state using the normalized transport entropy $h_i = -\frac{1}{\log(N-1)}\sum_{j\neq i}\widetilde{\pi}_{ij}\log(\widetilde{\pi}_{ij}+\epsilon)$, the expected semantic cost $c_i = \sum_{j\neq i}\widetilde{\pi}_{ij}\overline{C}^{\text{sem}}_{ij}$, the retained mass m_i, and the direct cross-modal agreement $a_i = 1 - \overline{C}^{\text{sem}}_{ii}$. Here h_i measures transport ambiguity, c_i measures the semantic mismatch

of transported correspondence, m_i measures relative retained non-null mass, and a_i measures the cross-modal agreement of the co-located response. A lightweight predictor uses these statistics to balance the two evidence paths:

$$\beta_i = \sigma(f_\beta(h_i, c_i, a_i, m_i)), \qquad \widehat{y}_i = (1-\beta_i)y_i^{\text{dir}} + \beta_i \widetilde{y}_i. \tag{4}$$

A small β_i preserves the co-located thermal response, while a large β_i favors the transported evidence when a coherent alternative correspondence is identified.

Reliability-controlled residual injection. The arbitrated evidence $\widehat{y}_i$ determines which thermal information is retained, but not how strongly it should modify the RGB representation. Two lightweight predictors are applied in sequence. We first assess evidence reliability using $g_i^r = [x_i, \widehat{y}_i, |\widehat{y}_i - x_i|, h_i, c_i, m_i]$. Conditioned on this estimate, a second predictor produces a base injection score from $g_i^s = [x_i, \widehat{y}_i, \widehat{y}_i - x_i, h_i, c_i, m_i, r_i]$:

$$r_i = \sigma(f_r(g_i^r)), \qquad s_i = \sigma(f_s(g_i^s)). \tag{5}$$

To align the final update with evidence arbitration, we define transported certainty as $t_i = m_i(1-h_i)$ and direct certainty as $d_i = (1+a_i)/2$. Their arbitration-aware combination is $\kappa_i = (1-\beta_i)d_i + \beta_i t_i$. The final injection strength is then defined as:

$$\alpha_i = s_i\,[r_0 + (1-r_0)r_i]\,[\kappa_0 + (1-\kappa_0)\kappa_i], \tag{6}$$

where r_0 and κ_0 are fixed floor values that prevent excessive suppression of potentially useful thermal evidence. The arbitrated thermal evidence is converted into a residual proposal using $e_i = [\widehat{y}_i, \widehat{y}_i - x_i, \widehat{y}_i \odot \tanh(x_i)]$, yielding:

$$\Delta_i = f_\Delta(e_i), \qquad z_i = x_i + \alpha_i \Delta_i. \tag{7}$$

The asymmetric update keeps RGB as the anchor, with LCIT determining where complementary thermal evidence is retrieved and REF controlling how it updates the RGB representation.

3.4 Multi-Scale Prediction

At each encoder stage, LCIT and REF are sequentially applied to X^l and Y^l, yielding a fused representation Z^l for $l = 1, \ldots, L$. The resulting multi-scale representations $\{Z^l\}_{l=1}^L$ are progressively aggregated by a lightweight decoder. For SOD, the decoder produces a binary saliency prediction, while for semantic segmentation, the fused representation is mapped to a K-class prediction through the corresponding output head. Task supervision follows the respective dense prediction setting, using multi-scale saliency for SOD (Wang et al., 2026), and multiclass prediction with region-overlap supervision for semantic segmentation (Zhou et al., 2025).

4 Experiments

4.1 Experimental Setup

Datasets and metrics. For RGB-T SOD, following prior unaligned SOD protocols (Wang et al., 2025), we jointly train a single model on the official training splits of UVT20K and un-VT5000, and evaluate it on UVT20K, UVT2000 (Wang et al., 2024a), un-VT5000, un-VT1000, and un-VT821. For semantic segmentation, we follow the standard train/test split of each benchmark and report KUST4K (Ouyang et al., 2025) in the main paper. We report S-measure S_α, E-measure E_ξ, and weighted F-measure F_β^w for SOD, and class-wise IoU and mIoU for semantic segmentation.

Implementation details. UCMIT employs two independent Swin-B encoders (Wang et al., 2025) for the RGB and thermal modalities. LCIT operates with 8×8 local windows at the first three encoder stages and a 4×4 window at the deepest stage. We use four transport heads and five Sinkhorn iterations, with $\lambda_g = 0.05$, $\rho = 0.5$, $r_0 = 0.05$, and $\kappa_0 = 0.25$. For SOD, the input resolution is 384×384, and the model is trained for 100 epochs using AdamW with cosine learning-rate decay, an initial learning rate of 4×10^{-5}, and a batch size of 4. The training objective combines weighted binary cross-entropy and IoU losses with deep supervision. For semantic segmentation, the input resolution is 512×640, and the training objective combines cross-entropy, Lovasz-Softmax, and Dice losses. The model is trained for 25 epochs under the same optimizer and learning-rate schedule, with an initial learning rate of 2×10^{-5} and a batch size of 2.

Table 1: Quantitative comparison (%) on RGB-T SOD benchmarks. The best is shown in bold.

Method	Source	UVT20K			UVT2000			un-VT5000			un-VT1000			un-VT821		
		S_α	E_ξ	F_β^w	S_α	E_ξ	F_β^w	S_α	E_ξ	F_β^w	S_α	E_ξ	F_β^w	S_α	E_ξ	F_β^w
MIDD (Tu et al., 2021)	TIP21	83.6	83.8	71.5	77.8	76.8	55.3	83.0	84.5	72.0	89.3	90.9	83.4	84.3	84.8	74.0
CGFNet (Wang et al., 2021)	TCSVT21	83.0	82.7	68.9	76.4	74.9	52.0	83.3	84.3	70.3	89.4	90.8	82.6	83.7	84.1	72.4
SwinNet (Liu et al., 2021)	TCSVT21	84.1	88.7	72.8	79.0	82.1	59.1	83.7	90.1	74.4	89.2	93.8	83.9	85.4	90.6	76.8
TNet (Cong et al., 2022)	TMM22	85.6	88.4	76.9	79.2	80.6	61.0	85.8	90.2	78.4	90.4	93.6	86.3	87.6	90.9	80.9
DCNet (Tu et al., 2022)	TIP22	80.9	84.2	71.5	76.7	80.0	58.8	81.2	86.1	73.7	87.6	91.1	83.9	81.7	85.8	73.8
HRTransNet (Tang et al., 2022)	TCSVT22	85.2	87.0	72.4	75.8	75.3	52.9	87.2	90.3	76.7	91.6	93.9	85.3	87.3	89.1	77.5
MCFNet (Ma et al., 2023)	AAAI23	84.2	86.6	75.7	77.4	78.6	59.3	83.6	87.3	76.1	89.5	92.8	85.7	85.0	88.2	78.0
LSNet (Zhou et al., 2023)	TIP23	83.5	85.7	70.0	76.3	76.2	53.0	84.7	88.9	74.5	89.7	92.8	83.5	84.2	87.0	74.1
LAFB (Wang et al., 2024b)	TCSVT24	84.7	88.8	75.6	77.8	80.0	59.1	85.1	91.1	78.3	90.4	94.6	86.7	83.4	88.3	77.7
SACNet (Wang et al., 2024a)	TMM24	85.2	90.2	75.0	79.5	83.3	61.6	86.2	93.3	79.9	91.0	95.7	87.4	87.6	93.0	81.0
PCNet (Wang et al., 2025)	AAAI25	87.1	91.1	80.8	81.9	86.2	67.9	87.9	93.5	83.9	92.2	96.4	90.4	89.3	93.8	85.4
TPS-SCL (Hu et al., 2026)	AAAI26	89.0	90.2	–	83.1	83.5	–	91.0	95.0	–	94.1	95.4	–	90.7	93.4	–
UMFNet (Wang et al., 2026)	CVPR26	89.0	91.8	83.6	83.7	85.5	**70.8**	91.4	95.5	**88.7**	94.1	97.2	**92.7**	**90.8**	94.2	**87.0**
UCMIT	–	**89.7**	**92.2**	**83.8**	**84.3**	**85.7**	**70.8**	**92.1**	**96.0**	88.1	**94.4**	**97.7**	92.3	90.6	**94.4**	84.6

Table 2: Quantitative comparison (%) on the KUST4K test set. The best is shown in bold.

Method	Source	Road	Building	Motorcycle	Car	Truck	Tree	Person	Traffic Facilities	mIoU
RTFNet (Sun et al., 2019)	RAL19	94.60	81.80	21.30	78.20	80.20	81.60	0.00	20.70	60.00
FEANet (Deng et al., 2021)	IROS21	94.70	82.30	53.90	83.00	83.00	82.50	0.00	47.80	67.60
EAEFNet (Liang et al., 2023)	RAL23	95.00	85.00	60.60	84.50	85.50	84.00	0.00	53.10	70.10
CMX (Zhang et al., 2023a)	TITS23	95.40	86.90	62.10	84.20	85.10	85.30	23.00	49.90	72.90
CMNeXt (Zhang et al., 2023b)	CVPR23	95.20	85.20	57.80	81.30	83.50	82.70	22.70	52.20	71.40
MRFS (Zhang et al., 2024)	CVPR24	95.51	85.82	61.27	84.20	84.62	84.70	20.73	49.31	72.26
DFormer (Yin et al., 2024)	ICLR24	94.52	78.06	52.55	77.68	81.42	79.23	21.09	45.63	67.80
DFormerV2 (Yin et al., 2025)	CVPR25	95.54	86.64	60.42	84.81	86.64	84.86	30.25	51.71	73.90
AMDANet (Zhong et al., 2025)	ICCV25	95.10	86.09	64.24	84.92	86.16	84.63	34.71	**58.43**	75.30
Mul-VMamba (Ni et al., 2025)	KBS25	95.41	84.70	63.33	85.17	86.79	84.98	36.06	56.11	75.15
U3M (Li et al., 2025)	PR25	94.36	80.35	54.98	78.57	81.79	80.24	22.47	47.47	68.89
MiLNet (Liu et al., 2025)	TIP25	95.39	87.45	63.15	85.42	86.75	84.94	39.25	55.24	75.76
Sigma (Wan et al., 2025)	WACV25	94.80	86.30	64.60	**86.50**	86.50	84.10	35.40	58.10	75.30
MambaSeg (Gu et al., 2026)	AAAI26	95.48	86.84	61.51	84.82	86.50	85.13	33.33	55.57	74.84
UCMIT	–	**95.71**	**87.53**	**66.07**	85.81	**87.74**	**85.77**	**42.08**	57.15	**77.01**

4.2 Comparison with State-of-the-Art Methods

RGB-T salient object detection. Table 1 compares UCMIT with thirteen representative RGB-T SOD methods on five unaligned benchmarks. UCMIT achieves the best S_α on UVT20K, UVT2000, un-VT5000, and un-VT1000, and the best E_ξ on UVT20K, un-VT5000, un-VT1000, and un-VT821. These results demonstrate the effectiveness of UCMIT in exploiting complementary thermal evidence under spatial inconsistency. As shown in the upper part of Fig. 3, UCMIT also produces more complete objects and cleaner boundaries under low illumination, cross-modal displacement, and small-target scenarios.

RGB-T semantic segmentation. Table 2 compares UCMIT with representative RGB-T segmentation methods on the unaligned KUST4K test set. UCMIT achieves the best overall mIoU of 77.01% and the highest IoU in six of the eight semantic categories. These results further demonstrate that the proposed cross-modal evidence transport remains effective for dense multi-class prediction under spatial inconsistency. As shown in the lower part of Fig. 3, UCMIT yields more accurate object regions and cleaner category boundaries in cluttered scenes with displaced thermal responses.

4.3 Ablation and Analysis

Component analysis. The ablation study evaluates the contributions of individual components to cross-modal evidence construction and fusion. "Local OT" retrieves nearby thermal evidence according to semantic similarity, while "+ Geometric Cost" further favors spatially plausible corre-

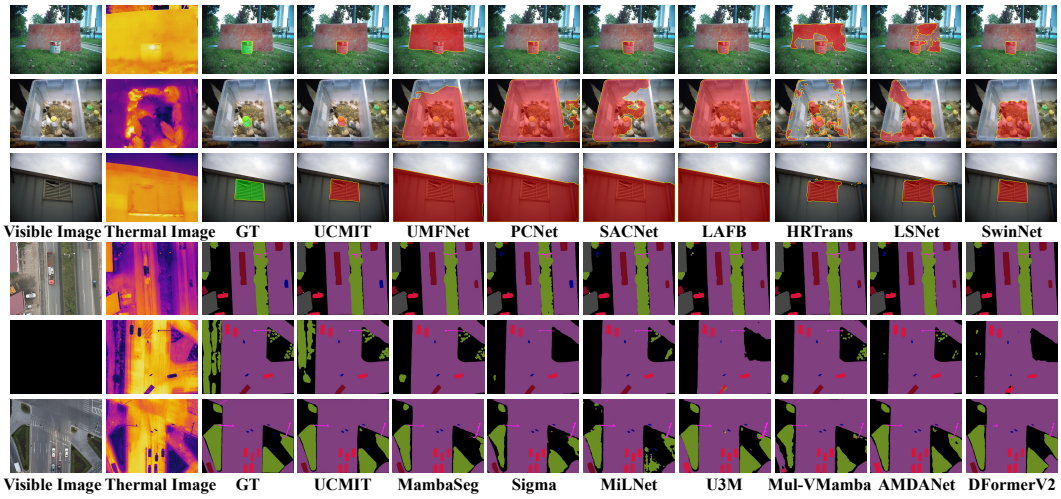

Figure 3: Qualitative comparison with representative methods. The upper part shows RGB-T salient object detection results, and the lower part shows semantic segmentation results.

Table 3: Controlled fusion comparison and component analysis. All variants use the same backbone, decoder, losses, and training protocol. The "+" variants are cumulative, whereas Row-wise Softmax replaces the structured transport in the otherwise complete configuration.

Dataset	Metric	Fusion baselines						Transport variants					
		RGB	Add	Concat	Gate	Cross-Attn.	Deformable	Local OT	+ Geometric Cost	+ Direct Path	+ Reliability Gate	Row-wise Softmax	UCMIT
UVT20K	$S_\alpha \uparrow$	85.6	87.0	87.3	87.7	88.8	88.6	88.4	88.7	88.5	89.1	88.8	**89.7**
	$E_\xi \uparrow$	88.9	90.0	90.3	91.7	92.1	91.5	91.7	91.7	91.4	92.0	91.5	**92.2**
	$F_\beta^w \uparrow$	76.6	79.6	80.0	80.8	80.2	79.9	79.6	80.3	79.6	81.9	82.4	**83.8**
KUST4K	mIoU↑	66.7	68.8	67.8	69.5	71.1	73.3	72.8	73.4	75.5	75.7	76.9	**77.01**

spondences. "+ Direct Path" preserves valid co-located responses when relocation is unnecessary, whereas "+ Reliability Gate" suppresses unreliable thermal contributions before RGB update. The degradation of "Row-wise Softmax" further demonstrates the benefit of structured source-side allocation over independent target-wise matching. Together, these components consistently improve upon the basic "Local OT" variant. The final UCMIT additionally assigns the fixed null budget toward targets with relatively weak real-source matches.

Feature representation analysis. Figure 4 visualizes the intermediate evidence evolution of UCMIT, with the first two rows corresponding to RGB-T salient object detection and the last two rows to semantic segmentation. Starting from the modality-specific features X^l and Y^l, LCIT reconstructs the transported representation $\widetilde{Y}^l$ so that thermal responses become more consistent with the spatial structure of the RGB branch. REF further selects between co-located and transported evidence to obtain $\widehat{Y}^l$, while the direct-agreement and injection maps indicate where complementary thermal information is considered reliable and useful for representation update. The resulting gated residual $\alpha^l \Delta^l$ is concentrated on task-relevant structures rather than being uniformly injected across the scene. Consequently, the fused representation Z^l preserves the dominant RGB structure while enhancing object regions and boundaries with complementary thermal evidence. This behavior is consistently observed in both binary salient-object prediction and multi-class semantic segmentation, supporting the generality of the proposed transport-and-update mechanism.

Robustness analysis. We evaluate robustness while keeping the RGB input and ground truth unchanged. For spatial robustness, we randomly translate the thermal image within ± 50 pixels. For thermal robustness, we add Gaussian noise with $\sigma = 0.1$ to the normalized thermal input. Δ_{spa} and Δ_{cor} denote the absolute performance drops from the clean setting under spatial perturbation and

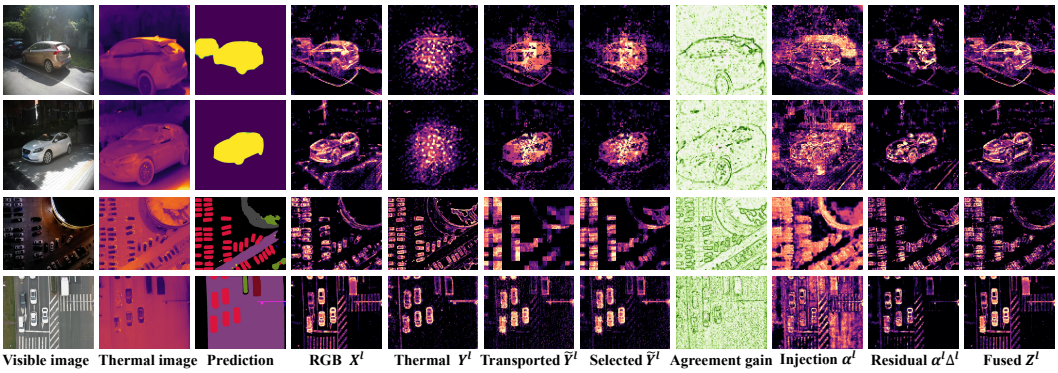

Figure 4: Feature analysis of UCMIT on RGB-T salient object detection (top two rows) and semantic segmentation (bottom two rows). From left to right are the input pair and prediction, followed by the RGB feature X^l, thermal feature Y^l, transported evidence $\widetilde{Y}^l$, selected evidence $\widehat{Y}^l$, direct agreement a^l, injection strength α^l, gated residual $\alpha^l\Delta^l$, and fused representation Z^l.

Table 4: Robustness under controlled spatial perturbation and thermal degradation.

Fusion	SOD (S_α)			Seg. (mIoU)		
	Clean	$\Delta_{\text{spa}}\downarrow$	$\Delta_{\text{cor}}\downarrow$	Clean	$\Delta_{\text{spa}}\downarrow$	$\Delta_{\text{cor}}\downarrow$
Addition	87.0	2.8	2.6	68.8	3.2	1.7
Gated fusion	87.7	2.0	2.2	69.5	2.3	1.5
Cross-Attn.	88.8	2.6	2.0	71.1	1.7	1.1
Deformable	88.6	1.8	1.7	73.3	1.4	1.8
UCMIT	**89.7**	**0.5**	**0.8**	**77.0**	**0.8**	**1.0**

Table 5: Efficiency comparison under the same implementation setting.

Setting	Fusion	Params (M) ↓	FLOPs (G) ↓	FPS↑
SOD	Cross-Attn.	221.31	141.42	28.4
	Deformable	224.90	143.05	26.8
	UCMIT	**215.12**	**139.21**	**30.1**
Seg.	Cross-Attn.	221.39	**338.39**	12.1
	Deformable	224.98	342.01	11.4
	UCMIT	**215.69**	339.56	**12.8**

thermal corruption, respectively. As shown in Table 4, UCMIT exhibits the smallest performance degradation on both SOD and semantic segmentation. The reduced Δ_{spa} supports the effectiveness of LCIT in improving robustness to spatial inconsistency, while the reduced Δ_{cor} is consistent with REF suppressing unreliable thermal contributions.

Efficiency analysis. Table 5 compares UCMIT with local cross-attention and deformable sampling under the same backbone, decoder, hardware, and inference protocol. LCIT confines correspondence construction to local windows and uses only a small number of Sinkhorn iterations. Consequently, UCMIT uses fewer parameters and achieves higher inference speed than both alternatives for both tasks. It also requires fewer FLOPs in the SOD setting, while its segmentation FLOPs remain comparable to local cross-attention and lower than deformable sampling.

5 Conclusion

This work studies unaligned RGB-T dense prediction as cross-modal information transport. UCMIT separates correspondence construction from representation update through LCIT and REF. LCIT reconstructs RGB-aligned thermal evidence from plausible nearby locations using semantic-geometric transport, while REF arbitrates between direct and transported evidence and controls residual injection according to evidence reliability and path certainty. This design reduces the influence of spatially mismatched or unreliable thermal responses before they update the RGB representation. Experiments on RGB-T salient object detection and semantic segmentation demonstrate strong performance across unaligned benchmarks, supporting the generality of the transport-based formulation. Current limitations arise from local transport neighborhoods and the fixed RGB anchor. Future work will explore longer-range correspondence modeling and adaptive cross-modal anchoring.

AI USE STATEMENT

In this work, generative AI tools were used solely for language editing, including grammar correction and improvements to the readability and linguistic clarity of the manuscript. Generative AI tools were not used for research tasks requiring disclosure under the ICLR policy, such as developing theoretical models or conceptual frameworks, formulating mathematical claims, designing research methodology or experiments, implementing the proposed method, or interpreting experimental results. All AI-assisted text was reviewed and verified by the authors. The authors take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

We have made efforts to ensure the reproducibility of this work. Detailed descriptions of the proposed method, experimental settings, datasets, evaluation protocols, and implementation details are provided in the main paper and appendix. Anonymous source code, configuration files, and training and evaluation scripts are available at `https://anonymous.4open.science/r/UCMIT-076F`. These materials are intended to facilitate reproduction of the reported results.

REFERENCES

Runmin Cong, Kepu Zhang, Chen Zhang, Feng Zheng, Yao Zhao, Qingming Huang, and Sam Kwong. Does thermal really always matter for rgb-t salient object detection? *IEEE Transactions on Multimedia*, 2022.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in Neural Information Processing Systems*, 26:2292–2300, 2013.

Fuqin Deng, Hua Feng, Mingjian Liang, Hongmin Wang, Yong Yang, Yuan Gao, Junfeng Chen, Junjie Hu, Xiyue Guo, and Tin Lun Lam. Feanet: Feature-enhanced attention network for rgb-thermal real-time semantic segmentation. In *2021 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pp. 4467–4473. IEEE, 2021.

Siyi Du, Xinzhe Luo, Declan O'Regan, and Chen Qin. Inference-time dynamic modality selection for incomplete multimodal classification. In *The Fourteenth International Conference on Learning Representations*, 2026.

Haixiao Gao, Chong Li, Yimin Zheng, Zhi-Ri Tang, Guanghua Yang, and Rammohan Mallipeddi. Unaligned red-green-blue and thermal salient object detection via structure-aware prior guidance. *Engineering Applications of Artificial Intelligence*, 176:114668, 2026.

Fuqiang Gu, Yuanke Li, Xianlei Long, Kangping Ji, Chao Chen, Qingyi Gu, and Zhenliang Ni. Mambaseg: harnessing mamba for accurate and efficient image-event semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 4302–4310, 2026.

Zirun Guo and Tao Jin. Smoothing the shift: Towards stable test-time adaptation under complex multimodal noises. In *The Thirteenth International Conference on Learning Representations*, 2025.

Qishen Ha, Kohei Watanabe, Takumi Karasawa, Yoshitaka Ushiku, and Tatsuya Harada. Mfnet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5108–5115. IEEE, 2017.

Jiahao He, Keren Fu, Xiaohong Liu, and Qijun Zhao. Samba: A unified mamba-based framework for general salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25314–25324, 2025.

Lupiao Hu, Fasheng Wang, Fangmei Chen, Fuming Sun, and Haojie Li. Breaking alignment barriers: Tps-driven semantic correlation learning for alignment-free rgb-t salient object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 4887–4895, 2026.

Wenjie Lai, Fanyu Zeng, Xiao Hu, Shaowei He, Ziji Liu, and Yadong Jiang. Regseg: An end-to-end network for multimodal rgb-thermal registration and semantic segmentation. *IEEE Transactions on Image Processing*, 33:6676–6690, 2024.

Bingyu Li, Da Zhang, Zhiyuan Zhao, Junyu Gao, and Xuelong Li. U3m: Unbiased multiscale modal fusion model for multimodal semantic segmentation. *Pattern Recognition*, 168:111801, 2025.

Gongyang Li, Yike Wang, Zhi Liu, Xinpeng Zhang, and Dan Zeng. Rgb-t semantic segmentation with location, activation, and sharpening. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(3):1223–1235, 2023.

Mingjian Liang, Junjie Hu, Chenyu Bao, Hua Feng, Fuqin Deng, and Tin Lun Lam. Explicit attention-enhanced fusion for rgb-thermal perception tasks. *IEEE Robotics and Automation Letters*, 8(7):4060–4067, 2023.

Chang Liu, Haizhuang Liu, Junbao Zhuo, Bochao Zou, Jiansheng Chen, Qianchuan Zhao, and Huimin Ma. Implicit alignment and query refinement for rgb-t semantic segmentation. *Pattern Recognition*, 169:111951, 2026.

Jinfu Liu, Hong Liu, Xia Li, Jiale Ren, and Xinhua Xu. Milnet: Multiplex interactive learning network for rgb-t semantic segmentation. *IEEE Transactions on Image Processing*, 34:1686–1699, 2025.

Yanbin Liu, Linchao Zhu, Makoto Yamada, and Yi Yang. Semantic correspondence as an optimal transport problem. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4463–4472, 2020.

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.

Shuai Ma, Kechen Song, Hongwen Dong, Hongkun Tian, and Yunhui Yan. Modal complementary fusion network for rgb-t salient object detection. *Applied Intelligence*, 53(8):9038–9055, 2023.

Rongrong Ni, Yuanhui Guo, Biao Yang, Yi Liu, Hai Wang, and Chuan Hu. Mul-vmamba: multimodal semantic segmentation using selection-fusion-based vision-mamba. *Knowledge-Based Systems*, pp. 115119, 2025.

Junlin Ouyang, Qingwang Wang, Ying Shang, Pengcheng Jin, Hangwei Zhong, Liman Zhou, and Tao Shen. An rgb-tir dataset from uav platform for robust urban traffic scenes semantic segmentation. *Scientific Data*, 12:1701, 2025.

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.

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.

Yuxiang Sun, Weixun Zuo, and Ming Liu. Rtfnet: Rgb-thermal fusion network for semantic segmentation of urban scenes. *IEEE Robotics and Automation Letters*, 4(3):2576–2583, 2019.

Bin Tang, Zhengyi Liu, Yacheng Tan, and Qian He. Hrtransnet: Hrformer-driven two-modality salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(2): 728–742, 2022.

Hao Tang, Zechao Li, Dong Zhang, Shengfeng He, and Jinhui Tang. Divide-and-conquer: Confluent triple-flow network for rgb-t salient object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(3):1958–1974, 2025.

Zhengzheng Tu, Zhun Li, Chenglong Li, Yang Lang, and Jin Tang. Multi-interactive dual-decoder for rgb-thermal salient object detection. *IEEE Transactions on Image Processing*, 30:5678–5691, 2021.

Zhengzheng Tu, Zhun Li, Chenglong Li, and Jin Tang. Weakly alignment-free rgbt salient object detection with deep correlation network. *IEEE Transactions on Image Processing*, 31:3752–3764, 2022.

Zifu Wan, Pingping Zhang, Yuhao Wang, Silong Yong, Simon Stepputtis, Katia Sycara, and Yaqi Xie. Sigma: Siamese mamba network for multi-modal semantic segmentation. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1734–1744. IEEE, 2025.

Jie Wang, Kechen Song, Yanqi Bao, Liming Huang, and Yunhui Yan. Cgfnet: Cross-guided fusion network for rgb-t salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(5):2949–2961, 2021.

Kunpeng Wang, Danying Lin, Chenglong Li, Zhengzheng Tu, and Bin Luo. Alignment-free rgbt salient object detection: Semantics-guided asymmetric correlation network and a unified benchmark. *IEEE Transactions on Multimedia*, 26:10692–10707, 2024a.

Kunpeng Wang, Zhengzheng Tu, Chenglong Li, Cheng Zhang, and Bin Luo. Learning adaptive fusion bank for multi-modal salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(8):7344–7358, 2024b.

Kunpeng Wang, Keke Chen, Chenglong Li, Zhengzheng Tu, and Bin Luo. Alignment-free rgb-t salient object detection: A large-scale dataset and progressive correlation network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 7780–7788, 2025.

Mianzhao Wang, Fan Shi, Xu Cheng, Chen Jia, and Shengyong Chen. Uncertainty-aware modality fusion for unaligned rgb-t salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 41489–41498, 2026.

Bo-Wen Yin, Jiao-Long Cao, Ming-Ming Cheng, and Qibin Hou. Dformerv2: Geometry self-attention for rgbd semantic segmentation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19345–19355. IEEE, 2025.

Bowen Yin, Xuying Zhang, Zhong-Yu Li, Li Liu, Ming-Ming Cheng, and Qibin Hou. Dformer: Rethinking rgbd representation learning for semantic segmentation. In *International Conference on Learning Representations*, volume 2024, pp. 51803–51825, 2024.

Hao Zhang, Xuhui Zuo, Jie Jiang, Chunchao Guo, and Jiayi Ma. Mrfs: Mutually reinforcing image fusion and segmentation. In *2024 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 26964–26973. IEEE, 2024.

Jiaming Zhang, Huayao Liu, Kailun Yang, Xinxin Hu, Ruiping Liu, and Rainer Stiefelhagen. Cmx: Cross-modal fusion for rgb-x semantic segmentation with transformers. *IEEE Transactions on intelligent transportation systems*, 24(12):14679–14694, 2023a.

Jiaming Zhang, Ruiping Liu, Hao Shi, Kailun Yang, Simon Reiß, Kunyu Peng, Haodong Fu, Kaiwei Wang, and Rainer Stiefelhagen. Delivering arbitrary-modal semantic segmentation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1136–1147. IEEE, 2023b.

Shenlu Zhao, Yichen Liu, Qiang Jiao, Qiang Zhang, and Jungong Han. Mitigating modality discrepancies for rgb-t semantic segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 35(7):9380–9394, 2024.

Shenlu Zhao, Ziniu Jin, Qiang Jiao, Qiang Zhang, and Jungong Han. Resolving semantic conflicts in rgb-t semantic segmentation. *Pattern Recognition*, 162:111398, 2025.

Haifeng Zhong, Fan Tang, Zhuo Chen, Hyung Jin Chang, and Yixing Gao. Amdanet: Attention-driven multi-perspective discrepancy alignment for rgb-infrared image fusion and segmentation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10645–10655. IEEE, 2025.

Heng Zhou, Zhenxi Zhang, Chengyang Li, Chunna Tian, Yongqiang Xie, Zhongbo Li, and Xiao-Jun Wu. Deformation-resilient multigranularity learning for unaligned rgb-t semantic segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 36(10):18530–18544, 2025.

Wujie Zhou, Yun Zhu, Jingsheng Lei, Rongwang Yang, and Lu Yu. Lsnet: Lightweight spatial boosting network for detecting salient objects in rgb-thermal images. *IEEE Transactions on Image Processing*, 32:1329–1340, 2023.

A Appendix

A.1 Source-Coupled Softmax Interpretation of the Transport Plan

This section further clarifies why the transport plan used by UCMIT differs from independent local similarity matching. Although entropy-regularized optimal transport also involves exponential normalization of matching costs, its correspondence weights are not obtained by applying a softmax independently to each RGB target. The source marginal introduces coupling across target locations and regulates how thermal evidence is allocated within each local window. For transport head k, the entropy-regularized OT problem in Eq. equation 2 admits the Sinkhorn factorization:

$$\Pi^{\star,k} = \mathrm{diag}(u^k) K^k \mathrm{diag}(v^k), \qquad K^k_{ij} = \exp\left(-C^k_{ij}/\tau_k\right). \tag{8}$$

where $u^k \in \mathbb{R}^N$ and $v^k \in \mathbb{R}^{N+1}$ are the target- and source-side Sinkhorn scaling vectors, respectively, and the last entry of v^k corresponds to the null source. For convenience, we define the row-normalized transport plan of head k as:

$$\bar{\Pi}^k_{ij} = \frac{\Pi^{\star,k}_{ij}}{\sum_{j'} \Pi^{\star,k}_{ij'}}. \tag{9}$$

where $\bar{\Pi}^k_{ij}$ denotes the normalized transport weight from source j to RGB target i, and j' ranges over all real thermal sources and the null source. Substituting Eq. equation 8 into Eq. equation 9 yields:

$$\bar{\Pi}^k_{ij} = \frac{u^k_i K^k_{ij} v^k_j}{u^k_i \sum_{j'} K^k_{ij'} v^k_{j'}} = \mathrm{softmax}_j\left(-\frac{C^k_{ij}}{\tau_k} + \log v^k_j\right). \tag{10}$$

where u^k_i is the target-side scaling factor, v^k_j is the source-side scaling factor, and τ_k controls the sharpness of transport assignment. The target-side factor cancels after row normalization, whereas the source-side factor remains. Hence, the row-normalized transport plan can be interpreted as a local softmax whose logits are modulated by the source-dependent term $\log v^k_j$. Importantly, $\log v^k_j$ is not an independently learned attention bias. The source-side scaling vector is jointly determined by all RGB targets through the source marginal constraint:

$$\sum_i \Pi^{\star,k}_{ij} = b_j. \tag{11}$$

where b_j denotes the prescribed transport supply of source j within the local window. Consequently, the correspondence assigned to one RGB target depends not only on its own semantic-geometric matching cost, but also on the aggregate demand placed on the same thermal sources by other targets. Target assignments are therefore coupled through aggregate source demand. The uniform source marginal serves as an allocation regularizer that limits unrestricted reuse of a dominant source without explicitly estimating source reliability. It should therefore be interpreted as an anti-collapse allocation prior rather than a source-quality model. Source reliability is instead assessed by REF before representation update. This distinction can also be observed from the relative preference between two source locations j and ℓ:

$$\frac{\bar{\Pi}^k_{ij}}{\bar{\Pi}^k_{i\ell}} = \exp\left(-\frac{C^k_{ij} - C^k_{i\ell}}{\tau_k}\right)\frac{v^k_j}{v^k_\ell}. \tag{12}$$

where the exponential term represents the pairwise semantic-geometric preference between sources j and ℓ, while v^k_j/v^k_ℓ reflects the source-side coupling induced by the transport constraints. If $v^k = \mathbf{1}$, this coupling disappears and Eq. equation 10 reduces to independent row-wise softmax over the same matching costs. We use this variant as a controlled counterpart in Table 3 to evaluate the benefit of structured source-side marginal coupling over independent target-wise matching. The

Table 6: Ablation of the transport design.

Variant	$S_\alpha \uparrow$	mIoU↑
w/o co-located mask	88.9	76.9
w/o null source	89.1	75.7
Fixed null cost	89.3	75.4
Row-wise softmax	88.8	76.9
UCMIT	**89.7**	**77.0**

Table 7: Ablation of REF statistics.

h_i	c_i	m_i	a_i	$S_\alpha \uparrow$	mIoU↑
				88.5	75.5
✓				89.2	76.0
✓	✓			89.3	76.5
✓	✓	✓		89.3	76.8
✓	✓	✓	✓	**89.7**	**77.0**

null source follows the same transport formulation but serves a different purpose. Its marginal fixes the window-level null mass:

$$b_\varnothing = \rho. \tag{13}$$

where ρ fixes the total null mass within each local window. The corresponding target-dependent null cost is defined as:

$$C^k_{i\varnothing} = f^k_\varnothing(q^k_i). \tag{14}$$

where $f^k_\varnothing$ is a lightweight predictor that estimates the rejection cost for RGB target i. This target-dependent cost determines how the fixed null mass is distributed across RGB targets. Locations with weak cross-modal correspondence can therefore allocate a larger fraction of their demand to the null source, while the total null mass remains fixed at ρ. In the row-wise softmax counterpart, the same null-source cost is retained, but its total mass is no longer constrained by a source marginal. The head-wise transport plans are averaged before row normalization:

$$\Pi^{\text{avg}} = \frac{1}{H}\sum_{k=1}^{H} \Pi^{\star,k}. \tag{15}$$

where H denotes the number of transport heads and Π^{avg} is their averaged transport plan. Under the prescribed target marginal, each head satisfies:

$$\Pi^{\star,k}\mathbf{1} = \frac{1}{N}\mathbf{1}. \tag{16}$$

where every RGB target receives the same total transport mass $1/N$. Therefore, the relation between averaging and row normalization can be written as:

$$\bar{\Pi} = N\Pi^{\text{avg}} = \frac{1}{H}\sum_{k=1}^{H} N\Pi^{\star,k} = \frac{1}{H}\sum_{k=1}^{H} \bar{\Pi}^k. \tag{17}$$

where $\bar{\Pi}$ denotes the row-normalized averaged transport plan. Therefore, averaging the transport plans before row normalization is equivalent to averaging their row-normalized counterparts. In practice, the truncated Sinkhorn recursion is terminated with a final target-side scaling update, which enforces the prescribed row marginal up to numerical precision. Finally, the transport problem remains feasible because the total target demand equals the total source supply:

$$N \cdot \frac{1}{N} = N \cdot \frac{1-\rho}{N} + \rho = 1. \tag{18}$$

Equation equation 18 also implies that the window-averaged retained mass satisfies $\frac{1}{N}\sum_i m_i = 1 - \rho$ up to numerical precision. Therefore, m_i is a relative target-wise retention signal within a window rather than an estimate of absolute window-level match quality. The structured plan regularizes local source allocation under this fixed null mass, but it is not the final correspondence output in UCMIT: it drives barycentric evidence reconstruction, followed by direct-path arbitration and reliability-controlled representation update.

A.2 Additional Implementation Details

UCMIT uses two modality-specific Swin-B encoders with identical architectures and independent parameters, both initialized with ImageNet-22K pretrained weights. LCIT and REF are applied at multiple encoder stages, and the resulting representations are progressively aggregated by a lightweight top-down decoder. LCIT uses 8×8 local windows at the first three encoder stages

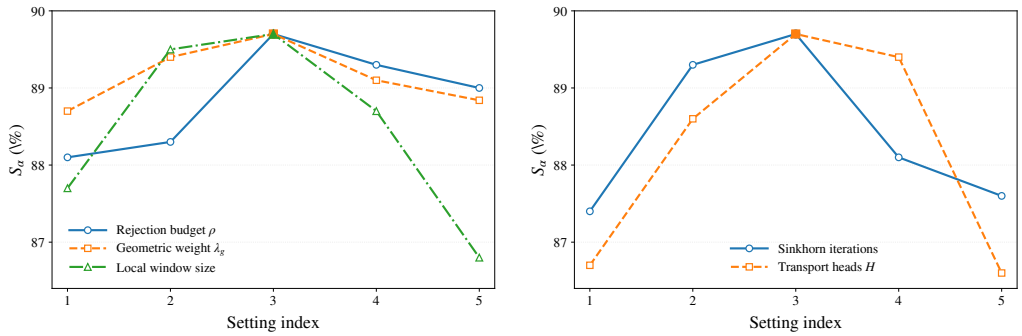

Figure 5: Parameter sensitivity of UCMIT on UVT20K. The horizontal axis denotes the setting index rather than the numerical hyperparameter value. For indices 1-5, the tested settings are $\rho = \{0.1, 0.3, 0.5, 0.7, 0.9\}$, $\lambda_g = \{0, 0.02, 0.05, 0.10, 0.20\}$, window size at the first three stages $= \{4 \times 4, 6 \times 6, 8 \times 8, 10 \times 10, 12 \times 12\}$, with the deepest-stage window fixed to 4×4, Sinkhorn iterations $= \{1, 3, 5, 7, 10\}$, and $H = \{1, 2, 4, 6, 8\}$. The left panel shows the sensitivity to ρ, λ_g, and the local window size, while the right panel shows the sensitivity to the number of Sinkhorn iterations and transport heads.

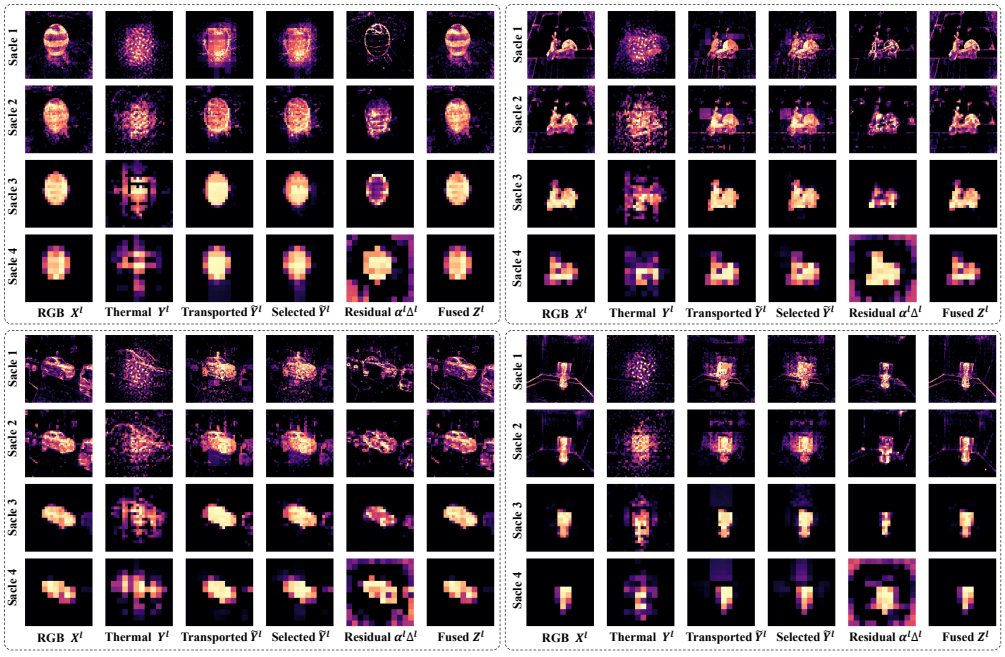

Figure 6: Multi-scale feature representation visualization of UCMIT for RGB-T salient object detection. Each row corresponds to one encoder stage, from shallow to deep. From left to right, we show the RGB anchor X^l, direct thermal response Y^l_{dir}, transported evidence $\widetilde{Y}^l$, selected evidence $\widehat{Y}^l$, gated residual $\alpha^l\Delta^l$, and fused feature Z^l.

and a 4×4 window at the deepest stage. We employ four transport heads and five log-domain Sinkhorn iterations. The geometric weight and fixed null mass are set to $\lambda_g = 0.05$ and $\rho = 0.5$, respectively, while the floor values in REF are set to $r_0 = 0.05$ and $\kappa_0 = 0.25$. The learnable transport temperature at each stage is bounded within $[0.05, 0.25]$. For RGB-T salient object detection, the input resolution is 384×384. The model is trained for 100 epochs using AdamW with an initial learning rate of 4×10^{-5}, a weight decay of 10^{-4}, and a batch size of 4. Weighted binary cross-entropy and IoU losses are used with deep supervision. For semantic segmentation, the input resolution is 512×640 following the KUST4K protocol. The training objective combines cross-

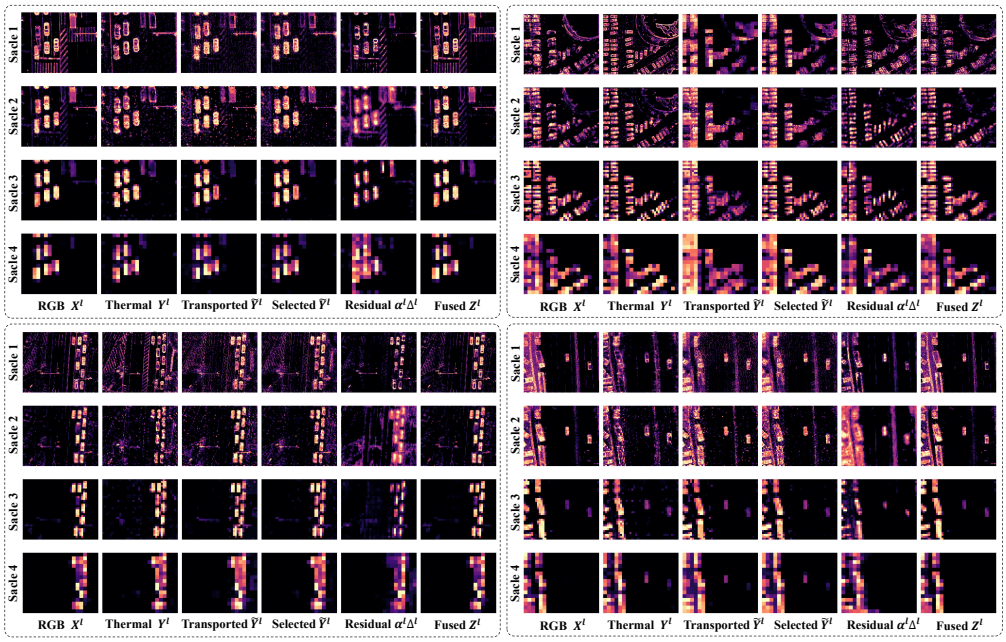

Figure 7: Multi-scale feature visualization of UCMIT for RGB-T semantic segmentation. Rows show encoder stages (shallow to deep); columns show X^l, Y^l_{dir}, $\widetilde{Y}^l$, $\widehat{Y}^l$, $\alpha^l\Delta^l$, and Z^l. LCIT progressively reconstructs thermal evidence that is more consistent with the RGB structure across scales, while REF gates the residual to yield compact, semantically coherent fused features.

Table 8: Quantitative comparison (%) on the aligned MFNet benchmark. The best result in each column is shown in bold.

Method	Source	Car	Person	Bike	Curve	Car Stop	Guardrail	Color Cone	Bump	mIoU
MFNet	IROS17	65.9	58.9	42.9	29.9	9.9	0.0	25.2	27.7	39.7
RTFNet	RAL19	87.4	70.3	62.7	45.3	29.8	0.0	29.1	55.7	53.2
FEANet	IROS21	87.8	71.1	61.1	46.5	22.1	6.6	55.3	48.9	55.3
GMNet	TIP21	86.5	**73.1**	61.7	44.0	**42.3**	14.5	48.7	47.4	57.3
EGFNet	AAAI22	87.6	69.8	58.8	42.8	33.8	7.0	48.3	47.1	54.8
MFTNet	GRSL22	87.9	66.8	**64.4**	47.1	36.4	8.4	**55.5**	**62.2**	57.3
LASNet	TCSVT22	84.2	67.1	56.9	41.1	39.6	**18.9**	48.8	40.1	54.9
MDRNet	TNNLS23	87.1	69.8	60.9	**47.8**	34.2	8.2	50.2	55.0	56.8
DML	TNNLS25	88.8	72.8	63.1	44.7	39.4	8.4	51.8	48.5	57.4
UCMIT	–	**89.3**	71.9	62.6	45.2	37.3	9.2	52.5	54.2	**57.8**

entropy, Lovasz-Softmax, and Dice losses. The model is optimized for 25 epochs using AdamW with an initial learning rate of 2×10^{-5}, a weight decay of 10^{-5}, and a batch size of 2. Cosine learning-rate decay is adopted for both tasks.

A.3 Additional Ablation Studies

A.3.1 Transport Design

We first examine several structural choices in LCIT that are not separately reported in the main component analysis. As shown in Table 6, removing the co-located mask allows the direct thermal response to enter the same transport competition as transported evidence, weakening the separation between evidence transport and subsequent arbitration. Removing the null source forces each RGB target to receive real thermal evidence even when no plausible local counterpart exists. Replacing

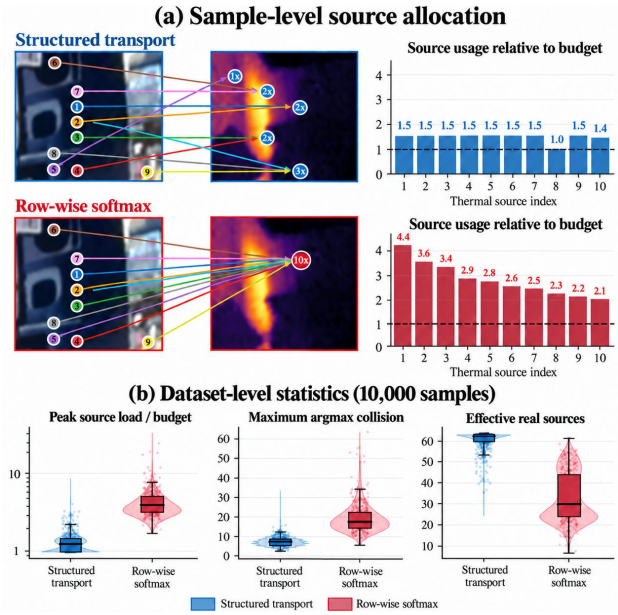

Figure 8: Comparison between structured transport and independent row-wise matching. The upper part visualizes source allocation for a representative sample. Structured transport distributes RGB targets across multiple thermal sources under the shared source marginal, whereas row-wise softmax may repeatedly reuse locally dominant sources. The lower part reports statistics over 10,000 samples using peak source load relative to the prescribed budget, maximum argmax collision, and the number of effective real sources. Structured transport exhibits more balanced source utilization and fewer correspondence collisions.

Table 9: Quantitative comparison (%) on the unaligned U-MFNet benchmark. The best result is shown in bold.

Method	Source	Car	Person	Bike	Curve	Car Stop	Guardrail	Color Cone	Bump	mIoU
MFNet	IROS17	59.6	39.2	42.8	16.1	8.9	0.0	18.2	14.0	32.7
RTFNet	RAL19	83.4	51.0	58.9	32.5	25.2	0.0	23.8	34.5	45.2
FEANet	IROS21	84.4	51.9	58.2	32.9	19.7	7.2	43.7	33.3	47.6
GMNet	TIP21	82.6	53.5	59.0	31.8	36.8	12.1	43.2	37.6	50.4
EGFNet	AAAI22	85.2	57.9	57.7	34.4	33.4	6.1	46.2	40.7	51.0
MFTNet	GRSL22	84.6	51.7	54.0	39.6	35.4	7.9	50.4	**56.0**	53.0
LASNet	TCSVT22	81.9	56.2	55.0	35.6	**38.6**	**19.9**	46.4	35.6	51.8
MDRNet	TNNLS23	84.0	55.8	58.9	36.0	28.5	6.5	46.7	45.9	51.1
DML	TNNLS25	87.6	64.9	63.1	39.3	35.1	5.4	51.0	45.1	54.4
UCMIT	–	**89.7**	**67.4**	**64.8**	**42.2**	32.9	3.6	**52.8**	47.2	**55.4**

the target-dependent null cost with a fixed value reduces the flexibility of rejection allocation, while row-wise softmax removes the structured source-side marginal coupling and performs independent target-wise matching. These results support the main structural choices of LCIT for reliable cross-modal evidence transport.

A.3.2 Reliability-Controlled Evidence Fusion

Table 7 evaluates the progressive combination of the transport-state statistics used by REF. These statistics characterize complementary aspects of transport ambiguity, semantic mismatch, retained

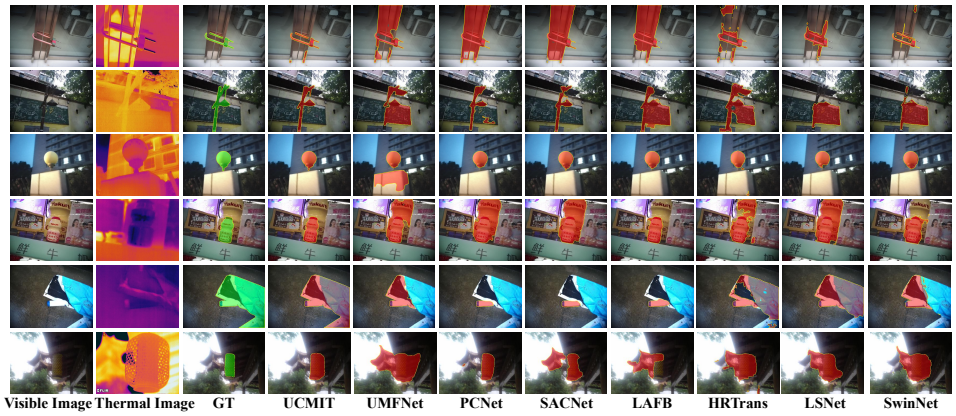

Figure 9: Additional qualitative results for RGB-T salient object detection on representative unaligned benchmarks.

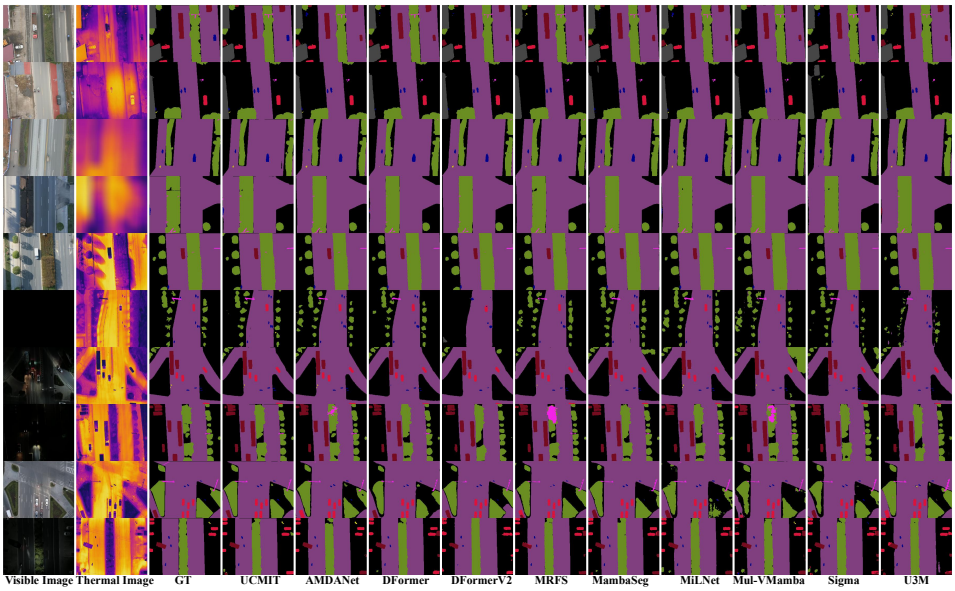

Figure 10: Additional qualitative results for RGB-T semantic segmentation on KUST4K.

mass, and direct-path agreement. Their joint use achieves the best overall performance on both SOD and semantic segmentation, supporting the full REF design.

A.3.3 Parameter Sensitivity

We further analyze the sensitivity of UCMIT to several key hyperparameters, including the fixed null mass ρ, geometric weight λ_g, local transport window size, number of Sinkhorn iterations, and number of transport heads H. As shown in Fig. 5, UCMIT maintains stable performance over a reasonable range of these settings. Moderate values of ρ and λ_g provide a favorable balance between null allocation, semantic compatibility, and spatial locality. A moderate transport window also performs best, while an excessively large range introduces less relevant thermal evidence. Similarly, increasing the number of Sinkhorn iterations or transport heads improves performance up to a moderate setting, after which further increases bring no additional benefit. Overall, these results indicate that UCMIT does not rely on delicate hyperparameter tuning.

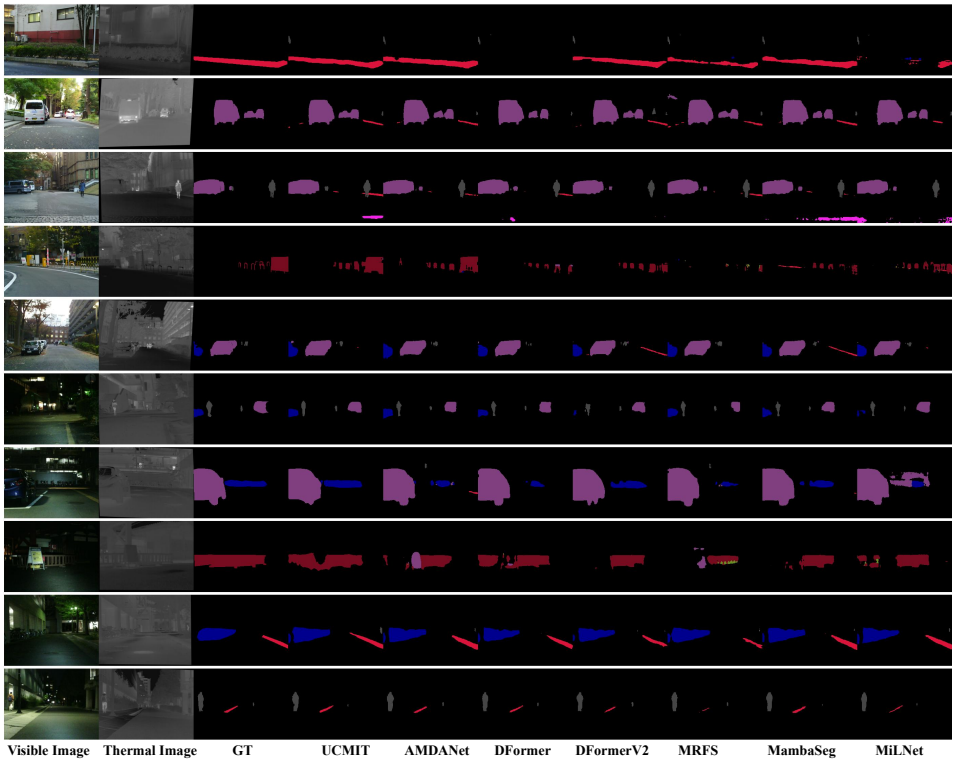

Figure 11: Additional qualitative results for RGB-T semantic segmentation on U-MFNet.

A.4 Multi-scale Feature Representation Visualizations

To further examine the multi-scale behavior of UCMIT, Figs. 6 and 7 present intermediate representations across encoder stages for RGB-T salient object detection and semantic segmentation, respectively. From shallow to deep stages, the transported and selected thermal evidence becomes increasingly compact and semantically focused, while the gated residual suppresses irrelevant responses and preserves task-relevant structures. As a result, the fused representations maintain the dominant RGB structure while progressively incorporating complementary thermal cues across scales.

A.5 Structured Transport versus Row-wise Matching

We further compare structured transport with independent row-wise matching. Both use the same semantic-geometric costs, but only structured transport imposes source-side marginal coupling. As shown in Fig. 8, row-wise softmax may repeatedly assign multiple RGB targets to the same dominant thermal source, whereas structured transport produces more balanced source allocation. Statistics over 10,000 samples further show fewer source-reuse collisions and broader utilization of effective thermal sources, confirming the benefit of coordinated cross-target evidence allocation.

A.6 Additional Semantic Segmentation Results

A.6.1 Results on Aligned MFNet

MFNet is a widely used RGB-T semantic segmentation benchmark containing 1,569 aligned RGB-thermal image pairs with a resolution of 640×480, including 820 daytime and 749 nighttime samples. The dataset contains eight foreground semantic classes and is commonly divided into 784 training, 392 validation, and 393 test samples. Table 8 reports the class-wise IoU and mIoU results of UCMIT and representative RGB-T semantic segmentation methods.

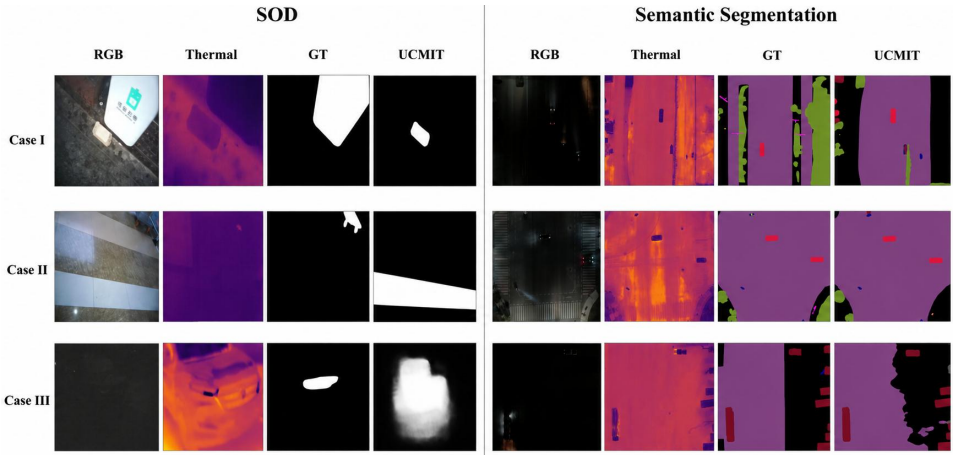

Figure 12: Representative failure cases of UCMIT on SOD and semantic segmentation under severe cross-modal spatial inconsistency. For each example, we show the RGB image, thermal image, ground truth (GT), and UCMIT prediction. When the displacement between the two modalities substantially exceeds the predefined local transport range, the correct complementary evidence may fall outside the candidate neighborhood, resulting in inaccurate localization, incomplete object structures, or erroneous semantic regions.

A.6.2 Results on U-MFNet

U-MFNet is an unaligned RGB-T semantic segmentation benchmark constructed from MFNet to evaluate cross-modal fusion under spatial inconsistency. Following the official U-MFNet protocol, we keep the RGB images unchanged and apply the prescribed affine deformation to the corresponding thermal images with a deformation strength of 20 pixels. Consequently, U-MFNet retains the 1,569 RGB-thermal image pairs, 640×480 image resolution, semantic categories, and data splits of MFNet, while introducing spatial displacement between the two modalities. This setting provides a controlled benchmark for evaluating the robustness of multimodal segmentation methods to cross-modal misalignment. Table 9 compares UCMIT with representative RGB-T semantic segmentation methods on U-MFNet.

A.7 Additional Qualitative Results

We provide additional qualitative results for RGB-T salient object detection and semantic segmentation under challenging conditions, including spatial misalignment, low illumination, cluttered backgrounds, small objects, and ambiguous boundaries. As shown in Fig. 9, UCMIT yields more complete salient regions and cleaner boundaries under diverse cross-modal inconsistencies. Figures 10 and 11 further show that UCMIT preserves object boundaries and small semantic regions more effectively on KUST4K and U-MFNet. These results consistently demonstrate the benefit of retrieving and reconstructing complementary thermal evidence before cross-modal representation update.

A.8 Failure Cases and Limitations

Although UCMIT improves robustness to cross-modal spatial inconsistency, it may still fail under extreme misalignment, as shown in Fig. 12. Since UCMIT performs evidence transport within a bounded local neighborhood, the correct thermal correspondence may fall outside the candidate region when the cross-modal displacement is too large, leading to inaccurate localization or incomplete semantic structures. This limitation reflects a trade-off between transport range and matching ambiguity. Enlarging the search region may cover larger displacement but also increases computation and introduces more irrelevant candidates. Future work could explore adaptive or coarse-to-fine transport strategies to handle more severe cross-modal misalignment.