DIFFERENCE IS THE SIGNAL: MODALITY RESIDUAL CURVE FOR FREQUENCY-ADAPTIVE INFRARED AND VISIBLE FUSION

Anonymous authors
Paper under double-blind review

ABSTRACT

The fusion of infrared and visible images is typically viewed as the integration of complementary information from the two modalities. However, when the two modalities provide inconsistent responses to the same scene, fusion models face another challenge: how to explicitly incorporate this cross-modal discrepancy into the fusion decision without ground-truth. Cross-modal residual provides direct but incomplete evidence. Residual energy can describe how strong the discrepancy is and which frequency regions it is concentrated, but it cannot provide directional information. To address this limitation, we propose the Modality Residual Curve (MRC), which combines two frequency variables with distinct roles: the discrepancy-magnitude curve, which characterizes the strength of cross-modal divergence across each frequency band; and the modality-direction curve, which compares the structural responses of the two modalities within the same frequency band. Their product forms a signed frequency prior. Its magnitude represents the strength of the intervention routing in that frequency band, while its sign indicates the direction of relative structural support rather than prescribing which modality the final fusion result must choose. The projected signed prior is added as an explicit bias to the content-driven expert routing scores. A dual-branch architecture aggregates the resulting modality-guided responses. Experiments on four public datasets, downstream applications, component ablations, and controlled degradation analyses consistently support the effectiveness and interpretability of the proposed method.

1 INTRODUCTION

Infrared–visible image fusion (IVIF) aims to synthesize a single image that preserves thermal saliency from the infrared modality while retaining structural detail from the visible modality (Shao et al., 2026). Complementarity does not imply consistent evidence: the same structure may be clear in one modality but weakened or degraded in the other, with differences varying across frequency bands. Without a unique fusion ground truth, discrepancy requires additional evidence for fusion decisions. Fusion is comparatively unambiguous when the modalities agree but underdetermined when they disagree. Therefore, we focuses on a specific question: Can frequency-resolved evidence of cross-modal discrepancy be extracted from the current input pair and further explicitly influence the decision-making process of the fusion model?

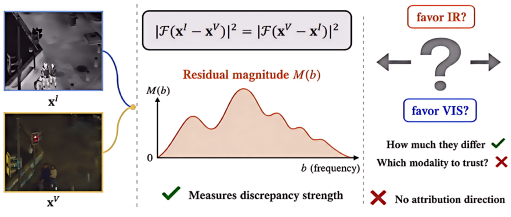

Figure 1: The modality residual shows the magnitude of discrepancy, but does not indicate the attribution direction.

Existing methods derive fusion cues from low-level measurements such as intensity, local contrast, edges, gradients, and spectral responses (Du et al., 2017; Li et al., 2020; Ren et al., 2021; Tang et al., 2023; Zhao et al., 2024; Wu et al., 2025). Task-driven and text-

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.

guided methods further introduce semantic knowledge to prioritize downstream-relevant content (Liu et al., 2024; Yi et al., 2024; Bai et al., 2025; Cheng et al., 2025). More closely related studies separate discrepant or frequency-specific information (Zheng et al., 2024; Guan et al., 2025; Wang et al., 2025), while mixture-of-experts (MoE) models route tokens, regions, or frequency components according to learned embeddings (Cao et al., 2023; Zhu et al., 2024). These methods improve content representation and cross-modal interaction, yet modality attribution largely remains implicit in latent features, attention maps, or learned fusion weights. Most existing formulations provide no explicit constraint that relates the magnitude and direction of cross-modal discrepancy to a corresponding change in fusion weights. This limitation is frequency-dependent because the modality with stronger structural support may vary across spectral bands and change as one source degrades.

This limitation calls for a signed, frequency-resolved attribution cue derived from the input pair. Let $\mathbf{\Delta} = \mathbf{x}^I - \mathbf{x}^V$ denote the residual between infrared and visible observations. As shown in Fig. 1, exchanging the modalities changes $\mathbf{\Delta}$ to $-\mathbf{\Delta}$, while leaving its power spectrum unchanged:

$$\left|\mathcal{F}\left(\mathbf{x}^I - \mathbf{x}^V\right)\right|^2 = \left|\mathcal{F}\left(\mathbf{x}^V - \mathbf{x}^I\right)\right|^2, \tag{1}$$

where $\mathcal{F}(\cdot)$ denotes the Fourier transform. Thus, residual energy measures the strength and frequency distribution of discrepancy but cannot alone identify the direction of relative structural support. We do not attempt to recover direction from residual energy. Instead, we pair an exchange-invariant measure of discrepancy strength with a separate, exchange-antisymmetric measure of relative structural response. Their product provides a signed prior that enters an observable decision variable.

We introduce Modality Residual Curve (MRC) to represent these requirements through two coupled frequency-domain variables. The discrepancy-magnitude curve $M(b)$ measures normalized residual energy in frequency band b by the joint energy of both inputs and remains unchanged under source–reference exchange. While the modality-direction curve $D(b)$ compares the band-wise structural responses of the source and reference modalities, indicating which role has the stronger structural response and reverses when the roles are exchanged. Their product, $r_b = M(b)D(b)$, forms a signed prior whose magnitude determines the strength of the routing bias and whose sign encodes the source–reference structural preference. By construction, $M(b)$ remains unchanged under source–reference exchange, whereas $D(b)$ and r_b reverse sign. When the modalities agree, the small magnitude suppresses unnecessary bias; when their roles are exchanged, the directional bias reverses consistently. Using $M(b)$ alone yields a directionless residual description, while coupling it with $D(b)$ supplies the missing attribution direction.

We use the distribution of $M(b)$ to allocate finer adaptive frequency bands where discrepancy is concentrated. The network extracts frequency features using these bands and projects r_b into band-aligned priors r_k. Each r_k is added to a content-dependent expert routing logit, explicitly perturbing expert weights without replacing content-based routing. Two parameter-unshared branches exchange the source and reference roles, so the prior supplies oppositely signed biases to corresponding bands. Cross-attention combines their responses and learns the final modality contributions. This connects discrepancy measurement, frequency allocation, and routing in a single decision path. Experiments on four datasets, downstream segmentation and detection, component ablations, and controlled-degradation settings evaluate both fusion performance and the intended band-wise behavior of MRC.

The main contributions are summarized as follows:

- We show that frequency-resolved residual energy measures discrepancy but cannot provide directional information because it is invariant to source–reference exchange. This motivates a separate measure of magnitude and direction support.
- We propose MRC, combining exchange-invariant $M(b)$ and exchange-antisymmetric $D(b)$ into a signed prior. Discrepancy strength determines adaptive bands, and the projected prior directly biases frequency-expert routing.
- We evaluate performance on four public fusion datasets and two downstream tasks. Component ablations and controlled-degradation routing analyses examine the prior's contribution and behavior.

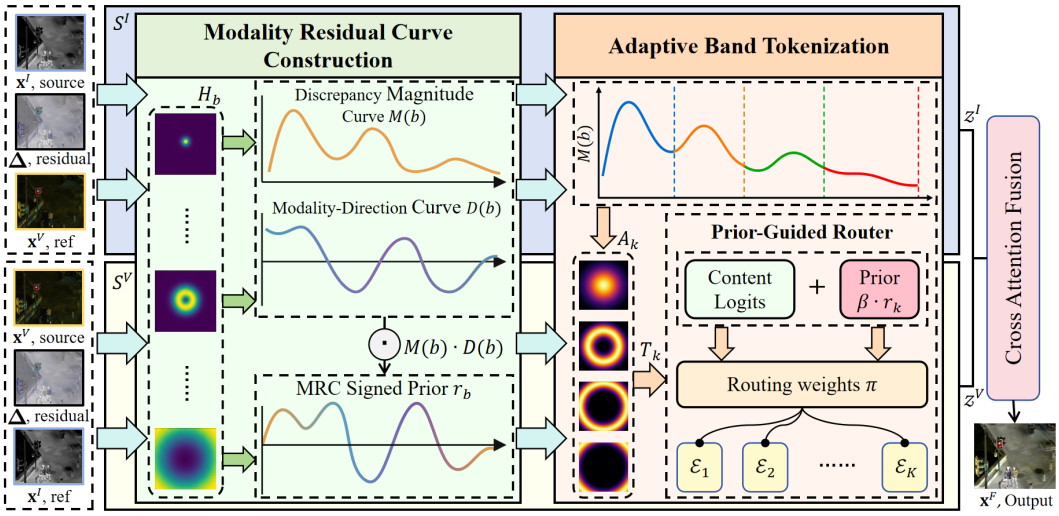

Figure 2: Overview of the proposed MRC framework. Each role-swapped branch constructs modality residual curve, forms adaptive frequency tokens, and applies prior-guided expert weighting before cross-attention fusion.

2 Related Work

2.1 Decision Cues in IVIF

Infrared-visible fusion methods differ primarily in the cues used to determine modality contributions. Early fusion methods (Du et al., 2017; Li et al., 2020; Ren et al., 2021) produce fused images from intensity, local contrast or gradients. Although such criteria are interpretable, they lack generalization and heavily rely on prior design. Instead, deep learning-based methods mainly employ convolutional neural networks (CNNs) (LeCun et al., 1998), Transformer (Vaswani et al., 2017), Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) and diffusion models (Ho et al., 2020) to learn hierarchical local representations, capture long-range cross-modal dependencies, impose adversarial perceptual constraints, and exploit generative priors for image fusion. More recently, an increasing number of studies have sought to equip models with high-level semantic information, enabling models to understand the semantic content of images. TDFusion (Bai et al., 2025) employs meta-learning to integrate downstream task-related training into a single training framework. SDSFusion (Chen et al., 2025) introduces semantic priors, multi-scale feature interaction, and a coarse-to-fine enhancement-fusion strategy to improve fusion quality and downstream perception. Text-IF (Yi et al., 2024) introduces textual descriptions and the CLIP model (Radford et al., 2021) into image fusion framework. TextFusion (Cheng et al., 2025) further utilizes text to obtain a controllable fusion network. BSPFusion (Shao et al., 2026) proposes a bottom-up and semantically progressive fusion model, establishing fine-grained entity-region priors through entity anchoring to enhance high-level semantic perception while retaining low-level texture details.

2.2 Residual and Frequency Modeling

The modality residual characterizes information that differs between aligned observations (Guan et al., 2025). Recent fusion methods increasingly recognize that common and discrepant information should be treated differently. Discrepancy-aware attention modules enhance modality-specific responses that may be suppressed by correlation-centered interaction (Yan et al., 2026). Frequency-domain methods (Zheng et al., 2024; Wang et al., 2025) decompose source images or latent features into low- and high-frequency components so that coarse structures and fine details can be processed separately. These studies demonstrate that discrepancy extraction and frequency decomposition are valuable for IVIF. However, modality residual is currently treated as an additional feature to be incorporated into latent representations. Such a representation allows the model to learn complementary information between two modalities, but does not explicitly characterize how modality residual af-

fects modality contributions during fusion. In contrast, MRC converts discrepancy into a signed, frequency-resolved prior whose magnitude and direction directly modulate the fusion decision.

2.3 Mixture-of-Experts Fusion Methods

Mixture-of-experts models dynamically activate suitable experts for different inputs, enabling adaptive modeling of different data distributions and thereby achieving robust and superior performance across diverse data (He et al., 2024). Their effectiveness depends critically on the routing function, which determines which experts are activated (Zhu et al., 2024). In image fusion field, MoE variants (Cao et al., 2023; Zhu et al., 2024) commonly route image tokens, spatial regions, or frequency components according to learned latent features.

MRC uses MoE routing for a different purpose. The router receives not only content-dependent token descriptors but also an explicit signed prior derived from cross-modal discrepancy. Adding this prior to the routing logits allows its magnitude and sign to modify the relative expert weights.

3 Methodology

3.1 Overview

Fig. 2 presents the framework of MRC. Given a registered infrared image $\mathbf{x}^I$ and a visible image $\mathbf{x}^V$, our goal is to generate a fused image $\mathbf{x}^F$ that preserves thermal saliency from the infrared modality while retaining the structural details of the visible modality. The network explicitly estimates how strong the discrepancy is and which modality provides stronger structural support.

The proposed framework contains two structurally identical but parameter-unshared branches. The infrared-guided branch $\mathcal{S}^I$ regards $\mathbf{x}^I$ as the source and $\mathbf{x}^V$ as the reference, whereas the visible-guided branch $\mathcal{S}^V$ reverses these roles:

$$\mathbf{z}^I = \mathcal{S}^I(\mathbf{x}^I, \mathbf{x}^V), \mathbf{z}^V = \mathcal{S}^V(\mathbf{x}^V, \mathbf{x}^I), \tag{2}$$

where $\mathcal{S}^I$ and $\mathcal{S}^V$ have identical architectures but do not share parameters, and $\mathbf{z}^I$ and $\mathbf{z}^V$ denote the two branch outputs, respectively.

For each branch, MRC measures band-wise discrepancy magnitude and relative structural preference, uses the magnitude distribution to construct adaptive frequency bands, and projects the signed curve into band-aligned routing priors. Two branch outputs are integrated by cross-attention:

$$\mathbf{x}^F = \mathrm{CA}(\mathbf{z}^I, \mathbf{z}^V), \tag{3}$$

where $\mathrm{CA}(\cdot)$ denotes the cross-attention fusion. Finally, the entire model is optimized using a composite objective $\mathcal{L}$ consisting of a spatial-domain $\mathcal{L}_1$ loss and a frequency-domain $\mathcal{L}_1$ loss as follows,

$$\mathcal{L} = \|\mathbf{x}^F - \mathbf{x}^I\|_1 + \|\mathbf{x}^F - \mathbf{x}^V\|_1 + \lambda_{fre}\big(\|\mathcal{F}(\mathbf{x}^F) - \mathcal{F}(\mathbf{x}^I)\|_1 + \|\mathcal{F}(\mathbf{x}^F) - \mathcal{F}(\mathbf{x}^V)\|_1\big), \tag{4}$$

where $||\cdot||_1$ denotes the $\mathcal{L}_1$ norm, and λ_{fre} denotes the weight of frequency-domain $\mathcal{L}_1$ loss.

3.2 Modality Residual Curve

For a branch, let $\mathbf{x}^s$ and $\mathbf{x}^r$ denote the source and reference images, respectively. MRC computes the modality residual:

$$\mathbf{\Delta} = \mathbf{x}^s - \mathbf{x}^r. \tag{5}$$

The residual captures cross-modal differences; however, this energy spectrum is sign invariant, as shown in Eq. (1). Therefore, it can quantify discrepancy but cannot determine which modality should be preferred.

To measure discrepancy at different frequencies, we construct $B_0 = 25$ normalized radial Gaussian masks. Let $\rho(u, v) \in [0, 1]$ be the normalized distance from frequency coordinate (u, v) to the

spectrum center. The b-th mask H_b can be written as

$$H_b(u,v) = \frac{\exp\left[-\frac{\left(\rho(u,v)-\mu_b\right)^2}{2\sigma_b^2}\right]}{\sum_{j=1}^{B_0}\exp\left[-\frac{\left(\rho(u,v)-\mu_j\right)^2}{2\sigma_j^2}\right]}, b = 1,\ldots,B_0, \tag{6}$$

where center $\mu_b = (\frac{b-1}{B_0-1})^2$ increases from low to high frequency and width σ_b controls the overlap. For any image $\mathbf{x}$, its band energy $E_x(b)$ in band b can be written as:

$$E_{\mathbf{x}}(b) = \sum_{u,v} H_b(u,v)|\mathcal{F}(\mathbf{x})(u,v)|^2. \tag{7}$$

The unsigned MRC magnitude is defined by normalizing residual energy with the joint source/reference energy:

$$M(b) = \frac{E_\Delta(b)}{E_{\mathbf{x}^s}(b) + E_{\mathbf{x}^r}(b) + \varepsilon}, \tag{8}$$

where $M(b) \geq 0$ and ε is a small positive constant to avoid division by zero. The denominator is symmetric in the two modalities, preventing $M(b)$ from being biased toward the lower-energy modality. Thus, $M(b)$ measures only the strength of modality residual.

To realize directionality, MRC computes Sobel gradient magnitudes $\mathbf{g}^s$ and $\mathbf{g}^r$ and evaluates their frequency band energies $G_s(b) = E_{\mathbf{g}^s}(b)$ and $G_r(b) = E_{\mathbf{g}^r}(b)$. The preferred direction is

$$D(b) = \log(\frac{G_s(b)+\varepsilon}{G_r(b)+\varepsilon}). \tag{9}$$

A positive $D(b)$ indicates stronger structural responses from source image, whereas a negative value favors the reference image. Gradient energy is used instead of raw intensity energy because it more directly reflects edges, contours, and textures relevant to perceptual fusion.

The signed routing prior r_b is obtained by coupling $M(b)$ and $D(b)$:

$$r_b = M(b) \cdot D(b). \tag{10}$$

Accordingly, $|r_b|$ expresses how strongly routing should be influenced, while its sign specifies the preferred modality. We additionally normalize $M(b)$ into a probability function $p(b)$,

$$p(b) = \frac{M(b)}{\sum_j M(j) + \varepsilon}. \tag{11}$$

3.3 Adaptive Band Tokenization and Prior-Guided Routing

The distribution $\{p(b)\}_{b=1}^{B_0}$ is converted into K input-adaptive frequency bands. Let $C(b) = \sum_{j=1}^{b} p(j)$ be the cumulative distribution and let a_k satisfy $C(a_k) = k/K$, with $a_0 = 0$ and $a_K = 1$. Band centers c_k and widths w_k are computed as

$$c_k = \frac{a_{k-1}+a_k}{2}, w_k = \frac{a_k - a_{k-1}}{2}, k = 1,\ldots,K. \tag{12}$$

Therefore, frequency regions with concentrated modality residual receive finer partitions, rather than being forced into fixed uniform bands. The parameters define soft radial masks $\mathbf{A}_k$, which can be formulated as

$$\mathbf{A}_k(u,v) = \exp\left[-\frac{(\rho(u,v)-c_k)^2}{2w_k^2}\right]. \tag{13}$$

For any feature G encoded from image $\mathbf{x}$, corresponding frequency token T_k is then extracted by Fourier filtering and band-specific convolution:

$$T_k = \psi_k\left(\mathcal{F}^{-1}[\mathcal{F}(\psi_0(\mathbf{G})) \odot \mathbf{A}_k]\right), \tag{14}$$

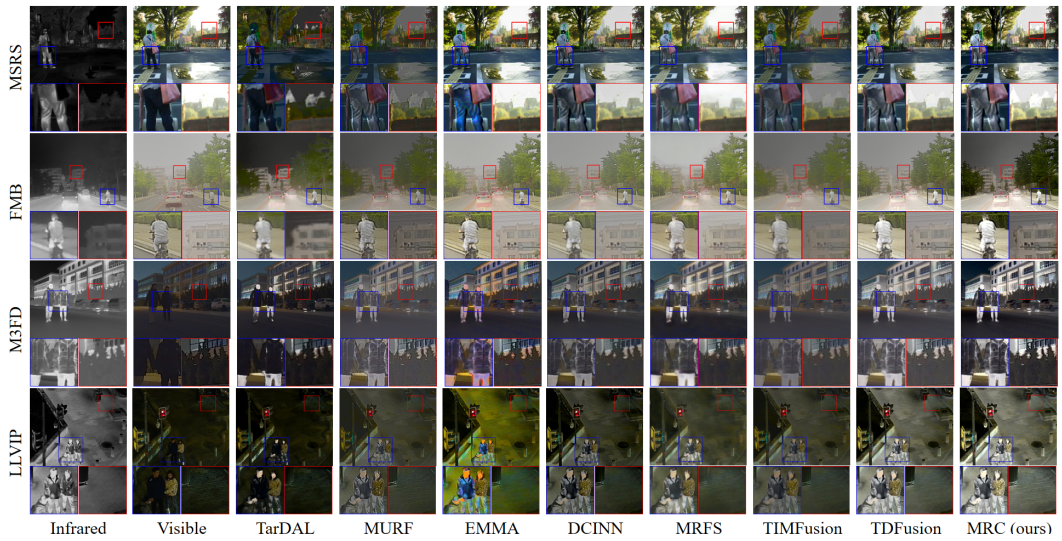

Figure 3: Visual comparison of fusion results on MSRS, FMB, M3FD, and LLVIP.

where $\odot$ denotes Hadamard product, ψ_0 denotes point-wise convolution, and ψ_k denotes depth-wise convolution. Consequently, T_k is a spatial feature map that represents the frequency token response in the k-th adaptive frequency band.

Since the original MRC contains B_0 bins and is adaptively divided into K bands, the signed curve is projected with the same Gaussian geometry. Let t_b be the normalized location of bin b, the projection coefficient of the k-th expert on the b-th position $\omega_{k,b}$ and the prior r_k can be formulated as

$$\omega_{k,b} = \frac{\exp\left[-\dfrac{(t_b - c_k)^2}{2w_k^2}\right]}{\displaystyle\sum_{j=1}^{B_0} \exp\left[-\dfrac{(t_j - c_k)^2}{2w_k^2}\right]}, \qquad r_k = \sum_{b=1}^{B_0} \omega_{k,b}\, r_b, \tag{15}$$

where $\omega_{k,b}$ aligns the MRC prior with the adaptive bands, while r_k summarizes the signed prior associated with the k-th adaptive frequency band. Its magnitude reflects the amplitude of the corresponding logit bias within the band, while its sign indicates the direction of structural preference.

Each token T_k is then globally pooled and mapped to a feature-dependent routing logit. The MRC prior is then added directly to the logits and the routing weight π_k is

$$\pi_k = \frac{\exp[(\phi_g(\mathrm{GAP}(T_k)) + \beta \cdot r_k)/\tau]}{\displaystyle\sum_{q=1}^{K} \exp[(\phi_g(\mathrm{GAP}(T_q)) + \beta \cdot r_q)/\tau]}, \tag{16}$$

where $\mathrm{GAP}(\cdot)$ denotes global average pooling, $\phi_g(\cdot)$ denotes linear projection, τ is the softmax temperature, and β controls the strength of MRC guidance. This additive formulation guarantees that the explicit modality residual participates in expert weighting. The MoE routed response is an aggregation of frequency-conditioned experts:

$$\mathbf{z} = \mathbf{W}\Big(\sum_{k=1}^{K} \pi_k \cdot \mathcal{E}_k(\mathbf{G}, T_k)\Big), \tag{17}$$

where $\mathcal{E}_k$ denotes the k-th expert and $\mathbf{W}(\cdot)$ is a 1×1 output projection.

Table 1: Quantitative comparison of infrared-visible image fusion on four datasets. The **bold** and underlined values indicate the best and second-best results, respectively.

Method	MSRS						FMB					
	EN ↑	SF ↑	SCD ↑	VIF ↑	$Q^{AB/F}$ ↑	SSIM ↑	EN ↑	SF ↑	SCD ↑	VIF ↑	$Q^{AB/F}$ ↑	SSIM ↑
TarDAL	5.28	5.98	0.71	0.21	0.18	0.47	6.63	6.94	1.03	0.28	0.29	0.74
MURF	5.04	10.40	1.62	0.22	0.37	0.60	6.37	13.88	1.34	0.22	0.37	0.68
EMMA	6.73	11.56	1.62	0.49	0.64	0.70	6.77	15.00	1.50	0.42	0.65	0.72
DCINN	6.00	10.51	1.49	0.41	0.57	0.52	6.47	11.47	1.39	0.38	0.59	0.74
MRFS	7.00	8.86	1.42	0.37	0.49	0.55	6.78	12.42	1.24	0.38	0.62	0.73
TIMFusion	6.27	9.67	1.34	0.32	0.48	0.68	6.51	12.23	1.24	0.35	0.59	0.73
TDFusion	6.74	11.30	**1.86**	0.50	**0.67**	0.70	6.86	14.16	1.76	**0.43**	**0.68**	0.75
MRC	**7.07**	**11.99**	1.83	**0.55**	0.60	**0.74**	**7.02**	**15.89**	**1.84**	**0.43**	0.66	**0.77**

Method	M3FD						LLVIP					
	EN ↑	SF ↑	SCD ↑	VIF ↑	$Q^{AB/F}$ ↑	SSIM ↑	EN ↑	SF ↑	SCD ↑	VIF ↑	$Q^{AB/F}$ ↑	SSIM ↑
TarDAL	6.87	7.63	1.29	0.27	0.30	0.71	6.32	7.42	1.04	0.27	0.22	0.58
MURF	6.50	12.55	1.46	0.21	0.32	0.64	6.13	15.08	0.96	0.21	0.31	0.57
EMMA	6.99	15.21	1.49	0.38	0.59	0.69	7.35	15.37	1.57	0.41	0.64	0.66
DCINN	6.59	11.21	1.46	0.34	0.51	0.72	6.98	13.34	1.43	0.38	0.52	0.64
MRFS	6.94	12.07	1.26	0.34	0.55	0.70	6.83	11.04	1.23	0.31	0.42	0.64
TIMFusion	6.75	12.31	1.37	0.35	0.53	0.70	6.58	13.52	1.14	0.33	0.46	0.64
TDFusion	6.99	14.49	1.83	0.41	**0.65**	0.72	7.36	**16.38**	1.75	0.46	**0.70**	**0.67**
MRC	**7.03**	**15.50**	**1.92**	**0.44**	0.63	**0.73**	**7.46**	15.93	**1.85**	**0.54**	0.67	0.66

Table 2: Performance comparison of downstream tasks. The **bold** and underlined values represent the best and second-best results, respectively.

Methods	Semantic Segmentation				Object Detection			
	MSRS		FMB		M3FD		LLVIP	
	mAcc	mIoU	mAcc	mIoU	mAP_{50}	mAP_{75}	mAP_{50}	mAP_{75}
TarDAL	81.93	71.35	62.86	55.33	83.16	56.39	93.79	62.71
MURF	85.03	74.08	64.10	56.96	80.58	54.22	94.24	68.04
EMMA	85.19	74.48	62.45	56.28	81.71	56.91	94.00	66.21
DCINN	84.11	74.35	61.09	54.81	82.69	57.37	94.92	68.34
MRFS	84.76	74.50	61.93	55.71	83.28	57.74	93.03	67.21
TIMFusion	83.67	73.58	63.70	57.24	83.22	56.08	93.76	61.33
TDFusion	86.04	75.09	67.17	**60.50**	**86.27**	**59.71**	95.00	69.18
MRC	**86.49**	**77.74**	**68.22**	60.00	86.06	58.46	**95.79**	**69.72**

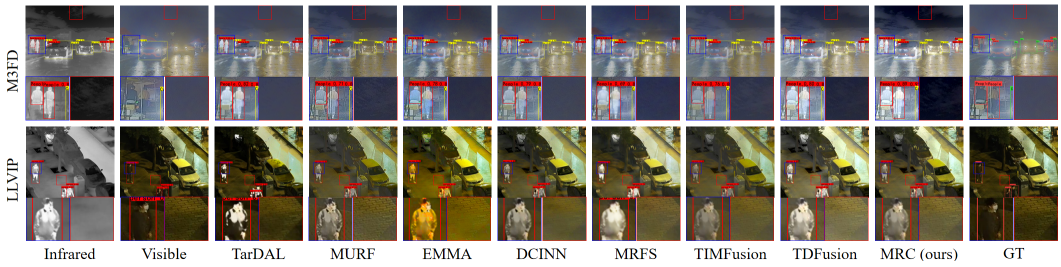

Figure 4: Visual results for object detection on M3FD and LLVIP.

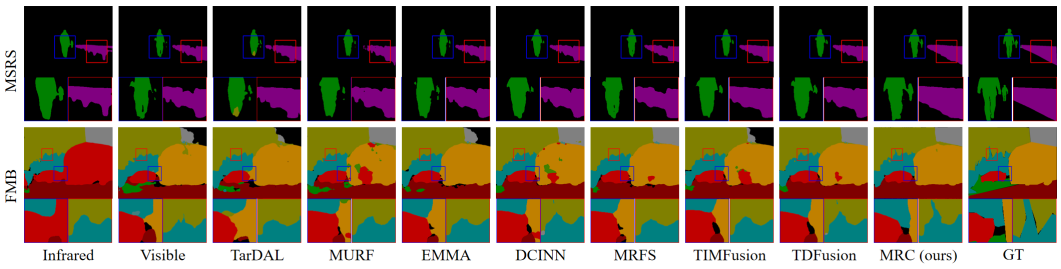

Figure 5: Visual results for semantic segmentation on MSRS and FMB.

4 Experiment

4.1 Experimental Setup

To evaluate the performance of our method, we conduct experiments on four public datasets, including M3FD (Liu et al., 2022), LLVIP (Jia et al., 2021), FMB (Liu et al., 2023) and MSRS (Tang et al., 2022). The dataset splits follow TDFusion (Bai et al., 2025). Additionally, MSRS and FMB are used for semantic segmentation. M^3FD and LLVIP are also used for object detection. We use the Adam (Kingma & Ba, 2014) optimizer with a learning rate of 1×10^{-4}. The training epoch is set to 120. The batch-size is set to 16. All experiments are conducted with a single NVIDIA A800 GPU. B_0 in Eq. (6) is set to 25. K in Eq. (12) is set to 4. β and τ in Eq. (16) are set to 1.0 and 1.0. We evaluate the MRC performance on IVIF with other representative state-of-the-art methods, including TarDAL (Liu et al., 2022), MURF (Xu et al., 2023), EMMA (Zhao et al., 2024), DCINN (Wang et al., 2024), MRFS (Zhang et al., 2024), TIMFusion (Liu et al., 2024), and TDFusion (Bai et al., 2025).

4.2 Fusion Results

The proposed method is evaluated from both quantitative and qualitative perspectives to provide a comprehensive assessment of its fusion performance. Tab. 1 presents the quantitative comparison results on four public datasets. MRC achieves the best or tied-best result on 17 of the 24 metrics and ranks within the top two on 23 metrics. The gains are particularly consistent in EN, SCD, VIF, and SSIM. Fig. 3 shows the qualitative comparison results on four different datasets. The visual comparison result also demonstrates that MRC effectively preserves salient target characteristics from the infrared modality while retaining rich background details from the visible modality without introducing severe luminance or contrast artifacts. The fused images exhibit superior performance in preserving distinct details, maintaining balanced luminance, and preserving clear structural outline across diverse environmental conditions. Overall, both the quantitative and qualitative results demonstrate that MRC provides stable fusion performance and effectively integrates the complementary characteristics of infrared and visible images.

4.3 Ablation Studies

To investigate the contribution of each component, this section conducts a series of ablation experiments. The results are shown in Tab. 3. The scalar variant averages $M(b)$ and $D(b)$ over frequency, thereby removing band-wise variation. The random-prior variant preserves the prior dimensionality but replaces the data-dependent values with random samples. The fixed-band variant replaces adaptive partitioning with uniformly spaced bands. The shared-branch variant uses the same parameter for both role-swapped branches. $D(b)$ only and $M(b)$ only variants remove discrepancy-strength and directional information, respectively. The full model performs best on all metrics, supporting the complementarity of the two curves, adaptive partitioning, and branch-specific parameters.

Table 3: Ablation results on MSRS dataset. The **bold** markers represent the best values.

Configs	EN	SF	SCD	VIF	$Q^{AB/F}$	SSIM
Scalar $M(b), D(b)$	6.82	11.25	1.65	0.52	0.55	0.63
Random $M(b), D(b)$	7.00	11.08	1.69	0.49	0.51	0.68
Fixed bands	6.93	11.38	1.60	0.44	0.56	0.57
Shared branches	6.78	10.62	1.58	0.43	0.47	0.61
$D(b)$ only	6.74	10.85	1.59	0.43	0.48	0.67
$M(b)$ only	6.87	10.79	1.66	0.51	0.51	0.61
MRC	**7.07**	**11.99**	**1.83**	**0.55**	**0.60**	**0.74**

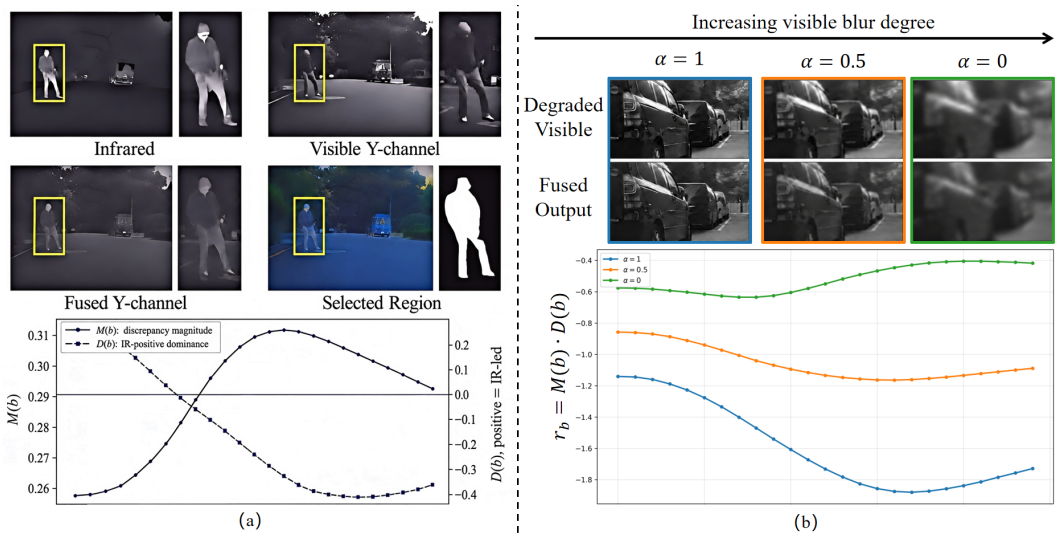

Figure 6: Figures of controlled behavior analysis. (a) Band-wise modality preference within a selected semantic region. (b) Response of the signed MRC prior to progressively blurred visible inputs.

4.4 Downstream Applications

The performance on downstream tasks is evaluated through object detection and semantic segmentation. Specifically, SegFormer (Xie et al., 2021) and YOLOv8 (Varghese & Sambath, 2024) are retrained using the same data split and optimization protocol. Tab. 2 shows that MRC ranks first on five of the eight reported metrics and second on the remaining three. Figs. 4 and 5 present the visual results of object detection and semantic segmentation, respectively. For semantic segmentation, the predictions obtained from MRC-fused images show a higher consistency with the ground truth. For object detection, the fused image provides more accurate localization and clearer boundaries for salient targets. These results indicate that MRC achieves a favorable balance between detail preservation and target localization.

4.5 Controlled Behavioral Analysis

In this section, we conduct two controlled analyses to examine the behavior of the proposed prior. First, we verify whether MRC can adaptively switch the preferred modality across different frequency bands within the same target region. Particularly, we select a masked semantic region and observe $M(b)$ and $D(b)$ curve over different frequency bands. Fig. 6 (a) shows that the infrared modality dominates at lower frequencies, whereas the visible modality provides stronger structural responses at higher frequencies. This masked-region computation is used only for analysis and does not imply spatially localized MRC routing during inference. Further, we investigate whether MRC can automatically reduce the reliance on visible modality when degradation occurs. Specifically, we progressively increase the blur degree (decreasing the clarity factor α) and observe r_b curve. As shown in Fig. 6 (b), the reliance on the visible modality decreases as blurring increases. Moreover, as the frequency increases, the differences in visible modality reliance among different blur degrees also become increasingly pronounced.

5 Conclusion

This work formulates infrared-visible image fusion under cross-modal discrepancy as a frequency-resolved modality-attribution problem. MRC combines normalized residual magnitude with a modality-direction curve to produce a signed prior whose magnitude is invariant and whose direction is antisymmetric under source-reference exchange. The magnitude distribution defines input-adaptive frequency bands, while the projected prior modulates band-specific expert weights in two role-swapped branches. Results across four fusion datasets, two downstream tasks, component ablations, and controlled degradation analyses support both the performance gains and the intended behavior of MRC.

AI USE STATEMENT

Generative AI tools were employed for language polishing and proofreading during the paper writing process. All suggestions generated with AI assistance were carefully reviewed and revised by the authors before being incorporated into the paper. We did not utilize generative AI tools for method design, experimental implementation or experiment analysis.

REPRODUCIBILITY STATEMENT

We will release the source code and instructions for reproducing the reported results when this work is accepted for ensuring reproducibility and transparency.

REFERENCES

Haowen Bai, Jiangshe Zhang, Zixiang Zhao, Yichen Wu, Lilun Deng, Yukun Cui, Tao Feng, and Shuang Xu. Task-driven image fusion with learnable fusion loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7457–7468, 2025.

Bing Cao, Yiming Sun, Pengfei Zhu, and Qinghua Hu. Multi-modal gated mixture of local-to-global experts for dynamic image fusion. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 23498–23507, 2023. doi: 10.1109/ICCV51070.2023.02153.

Jun Chen, Liling Yang, Wei Yu, Wenping Gong, Zhanchuan Cai, and Jiayi Ma. Sdsfusion: A semantic-aware infrared and visible image fusion network for degraded scenes. *IEEE Transactions on Image Processing*, 34:3139–3153, 2025. doi: 10.1109/TIP.2025.3571339.

Chunyang Cheng, Tianyang Xu, Xiao-Jun Wu, Hui Li, Xi Li, Zhangyong Tang, and Josef Kittler. Textfusion: Unveiling the power of textual semantics for controllable image fusion. *Information Fusion*, 117:102790, 2025.

Jiao Du, Weisheng Li, and Bin Xiao. Anatomical-functional image fusion by information of interest in local laplacian filtering domain. *IEEE Transactions on Image Processing*, 26(12):5855–5866, 2017.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, NIPS'14, pp. 2672–2680, Cambridge, MA, USA, 2014. MIT Press.

Zheng Guan, Xue Wang, Wenhua Qian, Peng Liu, and Runzhuo Ma. Residual prior-driven frequency-aware network for image fusion. *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025. URL https://api.semanticscholar.org/CorpusID:280017187.

Xuanhua He, Keyu Yan, Rui Li, Chengjun Xie, Jie Zhang, and Man Zhou. Frequency-adaptive pan-sharpening with mixture of experts. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(3):2121–2129, Mar. 2024. doi: 10.1609/aaai.v38i3.27984. URL https://ojs.aaai.org/index.php/AAAI/article/view/27984.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf.

Xinyu Jia, Chuang Zhu, Minzhen Li, Wenqi Tang, and Wenli Zhou. Llvip: A visible-infrared paired dataset for low-light vision. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3496–3504, 2021.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 2014.

Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

Hui Li, Xiao-Jun Wu, and Josef Kittler. Mdlatlrr: A novel decomposition method for infrared and visible image fusion. *IEEE Transactions on Image Processing*, 29:4733–4746, 2020.

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, 2022.

Jinyuan Liu, Zhu Liu, Guanyao Wu, Long Ma, Risheng Liu, Wei Zhong, Zhongxuan Luo, and Xin Fan. Multi-interactive feature learning and a full-time multi-modality benchmark for image fusion and segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8115–8124, 2023.

Risheng Liu, Zhu Liu, Jinyuan Liu, Xin Fan, and Zhongxuan Luo. A task-guided, implicitly-searched and meta-initialized deep model for image fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(10):6594–6609, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, A. Ramesh, and Gabriel Goh. Learning transferable visual models from natural language supervision. *International Conference on Machine Learning*, 2021.

Long Ren, Zhibin Pan, Jianzhong Cao, Hui Zhang, and Hao Wang. Infrared and visible image fusion based on edge-preserving guided filter and infrared feature decomposition. *Signal Processing*, 186:108108, 2021.

W Shao, H Liu, W Ding, and R Wang. Bspfusion: A bottom-up semantically progressive network for infrared and visible image fusion. *Information Fusion*, 2026.

Linfeng Tang, Jiteng Yuan, Hao Zhang, Xingyu Jiang, and Jiayi Ma. Piafusion: A progressive infrared and visible image fusion network based on illumination aware. *Information Fusion*, 83: 79–92, 2022.

Wei Tang, Fazhi He, Yu Liu, Yansong Duan, and Tongzhen Si. Datfuse: Infrared and visible image fusion via dual attention transformer. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(7):3159–3172, 2023.

Rejin Varghese and M Sambath. Yolov8: A novel object detection algorithm with enhanced performance and robustness. In *2024 International conference on advances in data engineering and intelligent computing systems (ADICS)*, pp. 1–6. IEEE, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

Qinghua Wang, Ziwei Li, Shuqi Zhang, Nan Chi, and Qionghai Dai. Wavefusion: A novel wavelet vision transformer with saliency-guided enhancement for multimodal image fusion. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(8):7526–7542, 2025. doi: 10.1109/TCSVT.2025.3549459.

Wu Wang, Liang-Jian Deng, Ran Ran, and Gemine Vivone. A general paradigm with detail-preserving conditional invertible network for image fusion. *International Journal of Computer Vision*, 132(4):1029–1054, 2024.

Xiao Wu, Zi-Han Cao, Ting-Zhu Huang, Liang-Jian Deng, Jocelyn Chanussot, and Gemine Vivone. Fully-connected transformer for multi-source image fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(3):2071–2088, 2025.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 12077–12090. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/64f1f27bf1b4ec22924fd0acb550c235-Paper.pdf.

Han Xu, Jiteng Yuan, and Jiayi Ma. Murf: Mutually reinforcing multi-modal image registration and fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):12148–12166, 2023.

Han Yan, Songlei Xiong, Long Wang, Lihua Jian, and Gemine Vivone. Atfusion: An alternate cross-attention transformer network for infrared and visible image fusion. *Infrared Physics & Technology*, 152:106252, 2026. ISSN 1350-4495. doi: https://doi.org/10.1016/j.infrared.2025.106252. URL https://www.sciencedirect.com/science/article/pii/S1350449525005456.

Xunpeng Yi, Han Xu, Hao Zhang, Linfeng Tang, and Jiayi Ma. Text-if: Leveraging semantic text guidance for degradation-aware and interactive image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27026–27035, 2024.

Hao Zhang, Xuhui Zuo, Jie Jiang, Chunchao Guo, and Jiayi Ma. Mrfs: Mutually reinforcing image fusion and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26974–26983, 2024.

Z Zhao, H Bai, J Zhang, and Y Zhang. Equivariant multi-modality image fusion. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Naishan Zheng, Man Zhou, Jie Huang, and Feng Zhao. Frequency integration and spatial compensation network for infrared and visible image fusion. *Information Fusion*, 109:102359, 2024. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2024.102359. URL https://www.sciencedirect.com/science/article/pii/S1566253524001374.

Pengfei Zhu, Yang Sun, Bing Cao, and Qinghua Hu. Task-customized mixture of adapters for general image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7099–7108, 2024. doi: 10.1109/CVPR52733.2024.00678.

A Appendix

A.1 Dataset Split

We follow the dataset partitions adopted by TDFusion to ensure a consistent comparison across fusion quality and downstream tasks. For MSRS, 1,083 registered infrared–visible image pairs are used for training and the remaining 361 pairs are used for testing. The same split is used for both fusion evaluation and semantic segmentation. For FMB, 1,220 image pairs are used for training and 280 pairs are used for testing; the associated semantic labels are retained for the downstream segmentation experiments. M3FD contains 4,200 annotated infrared visible image pairs. Following the adopted protocol, 3,150 pairs are assigned to the training set and 1,050 pairs to the test set for object detection. A fixed subset of 300 pairs from the detection test set is used for image-fusion evaluation, ensuring that no fusion-evaluation image is included in the training partition. The original LLVIP split contains 12,025 training pairs and 3,463 test pairs. Because consecutive frames are highly redundant and the complete dataset is substantially larger than the other benchmarks, every tenth pair is selected from the original training and test partitions, resulting in 1,203 training pairs and 347 test pairs. These reduced LLVIP partitions are used consistently for fusion and object-detection experiments.

A.2 Computational Efficiency

Table 4 reports the number of trainable parameters, floating-point operations, and GPU memory. MRC contains 14.4 million parameters and requires 85.4 GFLOPs for a 256×256 input. Its computational cost is higher than that of the most lightweight baselines because the model contains two

role-swapped branches, adaptive frequency processing, and cross-attention fusion. Nevertheless, its FLOPs remain substantially lower than those of EMMA and DCINN, while its parameter count and memory consumption are far below those of EMMA and MRFS. These results indicate that the performance gains of MRC are achieved with a moderate, rather than extreme, computational overhead.

Method	Params. (M)	FLOPs (G)	Peak Memory (MB)
TarDAL	7.8	41.9	134.8
MURF	3.5	59.9	165.0
EMMA	80.3	218.1	409.5
DCINN	4.3	702.5	182.2
MRFS	135.0	61.9	602.3
TIMFusion	1.2	58.7	179.7
TDFusion	3.8	17.4	179.1
MRC	14.4	85.4	220.5

Table 4: Computational-efficiency comparison at an input resolution of 256×256. All measurements are obtained on one NVIDIA A800 GPU under the same evaluation environment.

A.3 Training Pseudocode

Algorithm 1 summarizes the end-to-end optimization procedure without repeating the MRC equations given in the main paper. The two branches process the same registered pair with reversed source–reference roles and do not share parameters. Within each branch, the modality residual curve, adaptive frequency bands, band tokens, and signed routing priors are computed from the current input pair. Cross-attention then integrates the two branch outputs. The spatial- and frequency-domain losses are evaluated on the fused image, and gradients are propagated through the complete network in a single optimization step.

Algorithm 1 Simplified Training Procedure of MRC

Require: Registered training pairs $\mathcal{D}$; model parameters Θ
Ensure: Trained parameters $\Theta^\star$
Initialize Θ
for each training epoch **do**
  for each mini-batch (x^I, x^V) from $\mathcal{D}$ **do**
    $z^I \leftarrow \mathcal{S}^I(x^I, x^V)$
    $z^V \leftarrow \mathcal{S}^V(x^V, x^I)$
    $x^F \leftarrow \mathrm{CA}(z^I, z^V)$
    Update Θ by backpropagation and Adam
  end for
end for
$\Theta^\star \leftarrow \Theta$
return $\Theta^\star$

A.4 Evaluation Metrics

The fusion performance is evaluated using metrics including entropy (EN), spatial frequency (SF), sum of correlation differences (SCD), visual information fidelity (VIF), $Q^{AB/F}$ and SSIM. For all reported metrics, a larger value indicates better performance.

Entropy (EN). EN measures the information richness of the fused image from its intensity distribution. A larger value generally indicates that the result contains a wider and more informative range of visual content.

Spatial Frequency (SF). SF measures spatial activity using intensity variations along the horizontal and vertical directions. A larger SF value usually corresponds to sharper transitions, stronger local contrast, and richer texture details.

Sum of Correlation Differences (SCD). SCD evaluates how effectively complementary information from the two source images is transferred to the fused result. It measures correlation-based differences between the fused image and each source modality. A larger value indicates that the fused image preserves more complementary source information rather than merely reproducing one modality.

Visual Information Fidelity (VIF). VIF estimates the amount of perceptually meaningful information retained in the fused image under a human-visual-system-inspired model. A larger value indicates better preservation of source information and higher perceptual fidelity.

$Q^{AB/F}$. $Q^{AB/F}$ is an edge-based fusion-quality metric. It evaluates whether edge strength and edge orientation from the infrared and visible images are correctly preserved in the fused result. A larger value indicates more faithful transfer of salient boundaries and structural details.

Structural Similarity Index (SSIM). SSIM assesses structural consistency between the fused image and the source images by considering luminance, contrast, and structural patterns. We compute the score with respect to both source modalities and use their average. A larger value indicates that the fused result is structurally more consistent with the input pair.

Mean Accuracy (mAcc). mAcc is used for semantic segmentation and averages the per-class pixel classification accuracy. It reduces the dominance of frequent classes and reflects how reliably the model recognizes all semantic categories.

Mean Intersection over Union (mIoU). mIoU is the average class-wise overlap between predicted segmentation regions and ground-truth regions. It jointly penalizes false-positive and false-negative pixels and is the principal metric for evaluating semantic segmentation quality.

Mean Average Precision at IoU 0.50 (mAP50). mAP50 is used for object detection and averages class-wise detection precision when a predicted bounding box is considered correct at an intersection-over-union threshold of 0.50. It primarily reflects the model's ability to recognize and approximately localize objects.

Mean Average Precision at IoU 0.75 (mAP75). mAP75 uses the stricter intersection-over-union threshold of 0.75. It places greater emphasis on accurate bounding-box localization and is therefore more sensitive to boundary precision than mAP50.