



Brain-Inspired Reliability-Guided Multi-Level Interaction Network for Infrared–Visible Image Fusion

Yue Tian^{ID}, Linjun Liu, Changhui Zhu*, Zhao Zhang

Digital and Intelligent Operation Center, Guizhou Power Grid Co., Ltd., Guiyang 553099, China

Corresponding Author Email: juyaozhu@126.com

Copyright: ©2026 The authors. This article is published by IETA and is licensed under the CC BY 4.0 license (<http://creativecommons.org/licenses/by/4.0/>).

<https://doi.org/10.18280/ts.430419>

ABSTRACT

Received: 12 March 2026

Revised: 15 August 2026

Accepted: 20 August 2026

Available online: 31 August 2026

Keywords:

infrared–visible image fusion, brain-inspired computing, spiking neural network, modality reliability, cross-modal interaction, adaptive fusion

Infrared and visible images provide complementary information for scene perception, yet their different imaging mechanisms make it difficult to preserve thermal saliency and fine structural details simultaneously. This study proposes a Brain-Inspired Reliability-Guided Multi-Level Interaction Network (BMIF-Net) for infrared–visible image fusion (IVIF). The framework separates continuous image representation from modality reliability estimation. A dual-branch multi-scale encoder first extracts modality-specific features at three representation levels. A Spiking Modality Reliability Estimator (SMRE) then models local responses through leaky integrate-and-fire dynamics and derives reliability cues from temporally averaged membrane potentials and firing activity. These cues are introduced into a Reliability-Modulated Multi-Level Cross-Modal Interaction (RMCI) module to regulate bidirectional complementary feature transfer, while a Reliability-Guided Adaptive Fusion (RGAF) module generates spatially varying modality weights for final feature integration. Experiments on Multi-Spectral Road Scenarios (MSRS) show that BMIF-Net achieves an EN of 6.4131, an SD of 41.7063, an SF of 10.9309, an MI of 2.9500, a VIF of 2.1077, a QAB/F of 0.6369, and an SCD of 1.6084. In particular, the proposed method obtains the highest VIF among the evaluated methods and demonstrates competitive performance in spatial detail, edge preservation, and complementary information retention. Ablation experiments further show that the complete reliability-guided framework improves SF, VIF, QAB/F, and SCD over the baseline, while the joint membrane–spike representation provides stronger structural and perceptual preservation than membrane-state information alone. The complete model contains approximately 3.256 M trainable parameters and requires 7.843 GFLOPs for an input size of 128×128 . These results indicate that BMIF-Net provides a balanced hybrid continuous–spiking framework for reliability-aware IVIF with moderate computational complexity.

1. INTRODUCTION

Infrared and visible images describe the same scene through different physical sensing mechanisms. Visible-light imaging records reflected radiation and generally provides rich textures, edges, and structural details under favorable illumination, whereas infrared imaging responds to thermal radiation and can preserve target contrast under darkness, uneven illumination, or partial visual degradation. Infrared–visible image fusion (IVIF) therefore seeks to integrate complementary information from the two modalities into a single representation that retains both thermally salient targets and visually informative structures [1-4]. Such fusion has become relevant to surveillance, autonomous perception, object detection, and other vision tasks in which a single sensing modality may not provide sufficiently complete scene information.

Early IVIF methods relied mainly on multi-scale transforms, sparse representation, saliency analysis, subspace modeling, and manually designed fusion rules [4, 5]. Although these approaches offer relatively transparent processing

mechanisms, their performance depends strongly on the selected decomposition strategy, activity measure, and fusion rule. Deep learning subsequently shifted IVIF toward data-driven feature representation. DenseFuse introduced a densely connected encoder–decoder architecture for latent feature extraction and reconstruction [6], FusionGAN formulated fusion through adversarial image generation [7], and U2Fusion developed a unified unsupervised framework for multiple image-fusion tasks [8]. RFN-Nest and DIDFuse further explored residual feature fusion and deep image decomposition, respectively [9, 10]. These developments established the basic learning-based paradigm of extracting modality-specific features and reconstructing a fused image from learned representations.

More recent studies have increasingly focused on adaptive information preservation rather than uniform feature aggregation. PIAFusion introduced illumination-aware modeling to adapt fusion behavior to changing lighting conditions [11], MetaFusion incorporated object-detection-derived meta-features into the fusion process [12], and SGFusion used saliency guidance to control information

preservation [13]. MUFusion introduced memory-based unsupervised representation [14], while adaptive pixel-weighting strategies directly estimated spatially varying fusion weights for efficient infrared–visible integration [15]. These methods demonstrate that the relative importance of infrared and visible information is strongly dependent on local scene content and cannot be adequately described by a fixed global fusion rule.

A related challenge concerns how complementary information should be exchanged across representation levels. Shallow features mainly retain gradients, edges, and local textures, whereas deeper features progressively encode larger structures and salient content. Interactive Feature Embedding explicitly modeled information exchange between infrared and visible representations [16]. CoCoNet combined coupled contrastive learning with multi-level feature ensembles [17], MFIFusion introduced multi-level feature injection [18], and MDAN employed a multilevel dual-branch attention architecture [19]. BCMFIFuse further modeled bilateral cross-modal interaction in spatial and channel dimensions [20]. Transformer-based approaches extended this direction through progressive token exchange and interactive attention [21–23]. More recently, semantic interaction and selective state-space modeling have been incorporated into fusion networks, as illustrated by SFINet, S4Fusion, and CrossMamba [24–26]. Collectively, these studies show that hierarchical cross-modal interaction is important for preserving complementary information at different spatial and semantic scales.

However, effective interaction alone does not resolve a more fundamental issue: not every strong feature response is equally reliable. Infrared responses may dominate around thermally salient targets, whereas visible features may be more informative around road markings, building boundaries, vegetation, and fine textures. At the same time, thermal saturation, local sensor noise, low illumination, or overexposure can also generate strong activations that should not necessarily be transferred to the opposite modality. Conventional attention mechanisms generally estimate feature importance from continuous activations, but feature magnitude and modality reliability are not equivalent. Unrestricted information exchange may therefore propagate locally unreliable responses even when cross-modal interaction itself is well designed.

Spiking neural networks (SNNs) provide an alternative representation mechanism based on temporal membrane-state accumulation and threshold-triggered firing [27–29]. In a leaky integrate-and-fire neuron, incoming stimulation is accumulated in the membrane potential, while spike events are produced only when the neuronal state exceeds a firing threshold. The membrane potential therefore retains sub-threshold responses that may remain informative even when no spike is generated. This property is particularly relevant to image fusion, where weak edges and low-contrast textures should not be discarded simply because their responses are less pronounced than those of thermally salient targets. Recent work has also demonstrated that SNN-based architectures can be applied directly to IVIF, confirming the relevance of membrane dynamics and spike activity to multimodal fusion [30].

Nevertheless, using a fully spiking network as the principal reconstruction backbone is not necessarily desirable for detail-sensitive fusion. Repeated spike quantization may suppress weak but perceptually important image structures, whereas

continuous-valued features remain advantageous for fine-grained intensity and texture reconstruction. This motivates a different division of responsibility: continuous neural representations can be retained for image-content encoding and reconstruction, while spiking dynamics can be used as an auxiliary mechanism for estimating local modality reliability.

Based on this consideration, a Brain-Inspired Reliability-Guided Multi-Level Interaction Network (BMIF-Net) is proposed for IVIF. The framework separates continuous image representation from spiking-based reliability estimation. A dual-branch multi-scale encoder first extracts modality-specific features at three representation levels. A Spiking Modality Reliability Estimator (SMRE) then characterizes local modality responses through leaky integrate-and-fire dynamics and derives reliability cues from both membrane-state accumulation and firing activity. These cues are subsequently introduced into a Reliability-Modulated Multi-Level Cross-Modal Interaction (RMCI) module to regulate bidirectional feature exchange. A Reliability-Guided Adaptive Fusion (RGAF) module further combines interacted features and reliability cues to determine spatially varying modality contributions before multi-level reconstruction.

The main contributions of this study are summarized as follows:

- (1) A hybrid continuous–spiking fusion architecture is developed in which continuous neural features remain responsible for detailed image representation and reconstruction, while spiking dynamics are used specifically for modality reliability estimation. This design avoids forcing the complete fusion pathway into a discrete spiking representation.

- (2) A SMRE is introduced to characterize local modality reliability using both membrane-state accumulation and firing activity. The joint representation preserves continuous sub-threshold responses while retaining threshold-driven neuronal saliency.

- (3) A RMCI module is developed to regulate complementary information transfer according to the reliability of the contributing modality. This reduces the propagation of locally unreliable responses during feature interaction.

- (4) A RGAF module combines interacted continuous features with spiking-derived reliability cues to generate spatially varying modality weights, allowing the preferred information source to change with local scene content.

- (5) The proposed framework is evaluated using quantitative comparison, component ablation, reliability-representation analysis, temporal simulation analysis, and computational-complexity assessment. The experiments examine overall fusion performance together with the contributions of cross-modal interaction, the complete reliability-guided framework, and the incorporation of spike activity into the reliability representation.

2. RELATED WORK

2.1 Infrared–visible image fusion

IVIF aims to combine complementary information from heterogeneous sensors into a single image with enhanced scene representation. Traditional methods generally rely on multi-scale decomposition, sparse representation, subspace projection, saliency modeling, and manually designed fusion

rules [4, 5]. These methods provide relatively transparent processing mechanisms, but their performance is often sensitive to the selected decomposition strategy, activity measure, and fusion rule. Their adaptability is therefore limited when illumination conditions, thermal contrast, or background complexity vary substantially.

Deep learning has reduced the dependence on manually designed fusion rules by learning image representations directly from data. DenseFuse introduced a densely connected encoder–decoder framework for extracting and reconstructing infrared and visible features [6]. FusionGAN formulated infrared–visible fusion as an adversarial image-generation problem [7], while U2Fusion proposed a unified unsupervised framework capable of handling several image-fusion tasks within the same network [8]. RFN-Nest introduced residual fusion modules within a nested reconstruction architecture [9], and DIDFuse decomposed source images into common and modality-specific representations before fusion [10]. These methods established the encoder–fusion–decoder paradigm that remains influential in subsequent IVIF research.

Recent approaches increasingly emphasize adaptive information preservation. PIAFusion incorporates illumination awareness so that infrared and visible information can be treated differently under changing lighting conditions [11]. MetaFusion introduces object-detection-derived meta-features to improve the compatibility between low-level fusion and high-level semantic information [12]. SGFusion incorporates saliency information into an end-to-end fusion framework [13], while MUFusion employs memory units to preserve useful information during unsupervised fusion [14]. Adaptive pixel-weighting strategies further estimate spatially varying fusion weights to improve the balance between fusion quality and computational efficiency [15]. These studies demonstrate that a competitive fusion model should adapt to local scene characteristics rather than assign a fixed contribution to each modality.

More recent work has extended this adaptive paradigm toward semantic interaction and selective long-range modeling. SFINet jointly models image fusion and semantic segmentation and introduces semantic feature interaction to improve task-relevant representation [24]. S4Fusion employs a saliency-aware selective state-space mechanism to model global spatial information and cross-modal complementarity [25]. CrossMamba further introduces Mamba-based intra-modal and inter-modal interaction for long-range dependency modeling with approximately linear computational complexity [26].

These developments show a clear transition from static fusion rules toward content-dependent feature selection. Nevertheless, most existing approaches estimate modality importance directly from conventional continuous features. Feature importance and modality reliability are not necessarily equivalent: a large activation may correspond to useful scene content, but it may also originate from noise, saturation, poor illumination, or other modality-specific degradation. BMIF-Net therefore introduces a separate reliability-estimation mechanism and uses the resulting reliability information to regulate both cross-modal interaction and final modality weighting.

2.2 Multi-level and cross-modal feature interaction

Infrared and visible images contain complementary information at different representation levels. Shallow features

usually preserve local gradients, edges, and textures, whereas deeper representations capture broader spatial structures, salient regions, and higher-level contextual information. Fusion restricted to a single network depth may therefore fail to exploit useful information distributed across multiple scales.

Interactive Feature Embedding introduces explicit interaction between infrared and visible representations within a self-supervised framework, enabling hierarchical feature information to contribute to fusion [16]. CoCoNet employs coupled contrastive learning together with multi-level feature modeling to preserve modality-specific foreground and background characteristics [17]. MFIFusion injects information from multiple feature levels to distinguish and integrate superficial and deeper representations [18], whereas MDAN uses a multilevel dual-branch attention architecture to extract and fuse hierarchical features [19]. BCMFIFuse further performs bilateral cross-modal interaction to strengthen complementary feature exchange between the two modalities [20].

Transformer-based architectures provide another mechanism for cross-modal information communication. PTET introduces progressive token exchange so that useful information can be transferred gradually while less informative tokens are suppressed during the fusion process [21]. ITFuse alternates feature interaction and reconstruction to model both common and complementary information and employs interactive attention for cross-modal integration [23]. These methods demonstrate that effective information exchange should occur repeatedly and at multiple representation stages rather than through a single fusion operation.

However, stronger interaction does not automatically guarantee more reliable fusion. Existing interaction modules primarily determine whether features are relevant or complementary to each other, but they do not always explicitly distinguish useful responses from locally unreliable modality information. For example, a visible feature generated in an underexposed region may still participate strongly in interaction, while an infrared feature affected by local saturation may be propagated despite reduced reliability.

The proposed RMCI module addresses this issue by separating interaction strength from modality reliability. Instead of allowing complementary features to be transferred with unrestricted strength, independently estimated reliability cues modulate the information transmitted from one modality to the other. In this way, BMIF-Net retains the advantages of hierarchical cross-modal communication while reducing the propagation of locally unreliable responses.

2.3 Brain-inspired spiking computation

SNNs represent neuronal activity through temporal state accumulation and discrete firing events. Unlike conventional artificial neurons that propagate continuous activations directly, spiking neurons maintain an internal membrane state and generate a spike only when the accumulated response reaches a firing threshold. This temporal and event-driven representation has motivated substantial research in neuromorphic and brain-inspired computing [27].

Modern surrogate-gradient techniques have made deep SNNs increasingly compatible with gradient-based optimization [28]. In addition, learnable neuronal parameters, such as membrane time constants, have been introduced to improve the adaptability and representation capability of

spiking models [29]. These developments make it possible to integrate neuronal dynamics into conventional deep-learning pipelines rather than restricting SNNs to hardware-oriented or classification-specific applications.

For image fusion, the distinction between membrane states and firing events is particularly relevant. A spike records whether the neuronal response crosses a predefined threshold, whereas the membrane potential also retains accumulated sub-threshold information. Weak edges, low-contrast textures, or partially degraded image structures may therefore influence the membrane state even when they do not generate frequent firing events. Such information is potentially useful when estimating whether a local modality response is reliable enough to influence fusion.

The relevance of spiking computation to IVIF has recently become more explicit. SpikeVFuse employs an SNN-based architecture in which synchronized infrared and visible images are represented over multiple simulation steps, and membrane-potential dynamics are used to support cross-modal information processing [30]. Its emergence demonstrates that spiking computation is no longer limited to classification or neuromorphic recognition tasks and can be applied directly to multimodal image fusion.

Consequently, the novelty of BMIF-Net does not rest on the simple introduction of SNNs into IVIF. Instead, it assigns a different role to spiking dynamics. The spiking branch is not used as the principal carrier of fused image content or as a

replacement for the continuous reconstruction pathway. Continuous convolutional features remain responsible for preserving detailed intensity, texture, and structural information. Spiking dynamics are employed specifically to estimate modality reliability from membrane-state accumulation and firing activity.

The resulting reliability information is then used at two stages of the fusion process. First, it regulates bidirectional feature exchange in the RMCI module. Second, it is combined with the interacted continuous features in RGAF to support spatially adaptive modality weighting. Thus, spiking computation acts as a control signal for heterogeneous information flow rather than as the primary representation space for image reconstruction.

3. PROPOSED METHOD

3.1 Overall architecture

The proposed BMIF-Net integrates continuous image representation with spiking-based modality reliability estimation. The network consists of five main components: a dual-branch multi-scale encoder, a SMRE, a RMCI module, a RGAF module, and a lightweight multi-level reconstruction decoder. The overall architecture is illustrated in Figure 1.

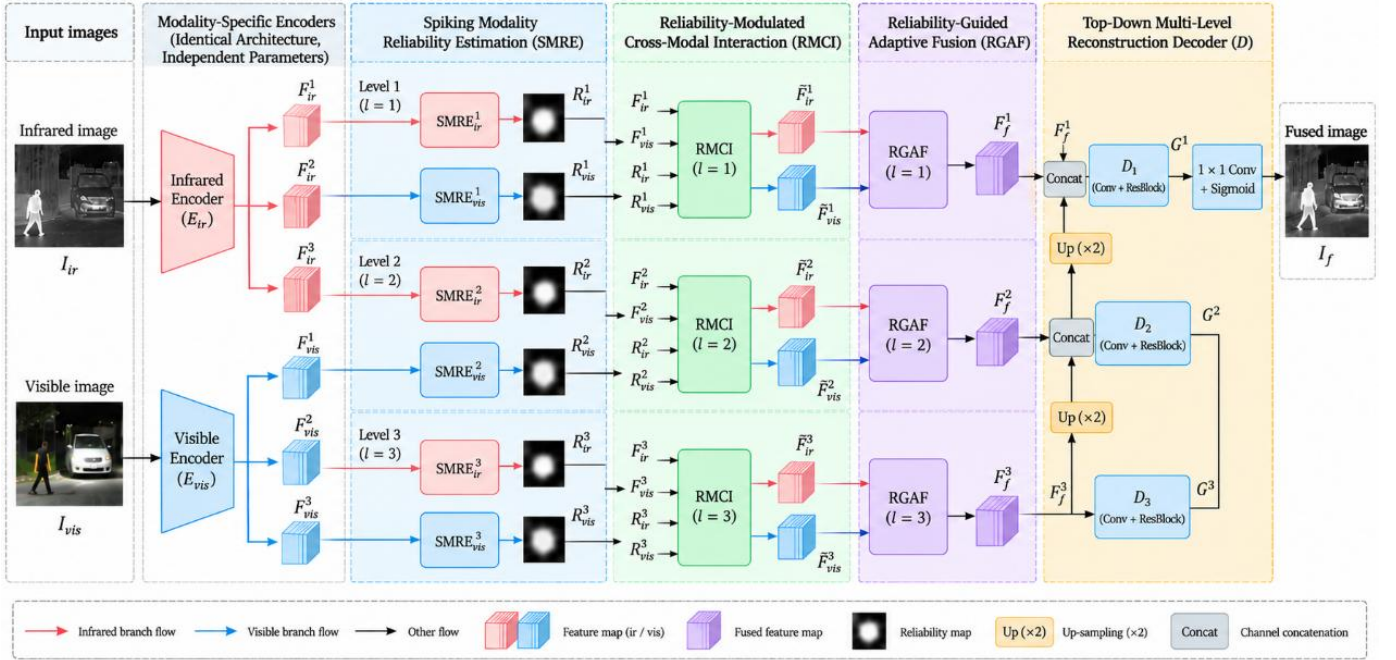


Figure 1. Overall architecture of BMIF-Net for reliability-guided IVIF

Note: BMIF-Net = Brain-Inspired Reliability-Guided Multi-Level Interaction Network; IVIF = infrared-visible image fusion.

Let I_{ir} and I_{vis} denote a registered infrared image and visible image, respectively. For each representation level l , the complete fusion process can be expressed as

$$I_f = \mathcal{D} \left(\left\{ \mathcal{G}^l \left(\mathcal{M}^l (F_{ir}^l, F_{vis}^l; R_{ir}^l, R_{vis}^l); R_{ir}^l, R_{vis}^l \right) \right\}_{l=1}^3 \right) \quad (1)$$

where, $\mathcal{D}$, $\mathcal{G}^l$, and $\mathcal{M}^l$ denote the decoder, RGAF, and RMCI at level l , respectively, while R_{ir}^l and R_{vis}^l are the modality reliability maps estimated by SMRE.

The dual encoders extract three levels of modality-specific features:

$$\{F_m^l\}_{l=1}^3 = \mathcal{E}_m(I_m), m \in \{ir, vis\} \quad (2)$$

The overall information flow is therefore summarized as

$$(I_{ir}, I_{vis}) \rightarrow \{F_{ir}^l, F_{vis}^l\} \rightarrow \{R_{ir}^l, R_{vis}^l\} \rightarrow \{\tilde{F}_{ir}^l, \tilde{F}_{vis}^l\} \rightarrow \{F_f^l\} \rightarrow I_f \quad (3)$$

Reliability estimation is performed before extensive cross-modal mixing. This ordering is deliberate: locally unreliable responses are first identified and can subsequently be attenuated before they propagate through cross-modal interaction and final fusion.

3.2 Dual-branch multi-scale feature encoder

Infrared and visible images exhibit different intensity distributions, structural characteristics, and modality-specific response patterns. A fully shared encoder may restrict the ability of the network to specialize for these heterogeneous sensing mechanisms. BMIF-Net therefore employs two encoder branches with identical architectures but independent learnable parameters.

For modality $m \in \{ir, vis\}$, the encoder produces three hierarchical feature representations:

$$F_m^1 \in \mathbb{R}^{H \times W \times 32}, F_m^2 \in \mathbb{R}^{\frac{H}{2} \times \frac{W}{2} \times 64}, F_m^3 \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times 128} \quad (4)$$

The first level retains fine spatial information, including gradients, local boundaries, and textures. The second level represents larger local structures, whereas the third level provides broader contextual and salient information.

Each level contains a residual feature-extraction block consisting of two 3×3 convolutions. The second and third levels are preceded by stride-2 convolutions that simultaneously reduce the spatial resolution and increase the channel dimension. Only two downsampling operations are employed because excessive spatial reduction may suppress

weak boundaries and low-contrast structures that remain important for fusion.

Although the infrared and visible branches have the same topology, their parameters are independent. This allows the infrared encoder to specialize in thermal contrast and target-related responses, while the visible encoder can retain greater sensitivity to edges, textures, and structural details.

Before the extracted features enter SMRE, each feature level is projected through a learnable 1×1 convolution:

$$Z_m^l = P_m^l(F_m^l), l \in \{1, 2, 3\} \quad (5)$$

where, P_m^l denotes a learnable 1×1 convolutional projection and Z_m^l is the representation supplied to the spiking reliability estimator. The projection preserves the spatial resolution and channel dimension of the corresponding encoder feature.

This design separates two responsibilities. Continuous features retain detailed image content for interaction and reconstruction, whereas the spiking branch operates on their projected representations to estimate local modality reliability.

3.3 Spiking Modality Reliability Estimator

SMRE estimates how reliably each modality responds to local scene content through leaky integrate-and-fire neuronal dynamics. Rather than replacing the continuous feature representation with discrete spikes, the module uses neuronal state evolution as an auxiliary source of reliability information. Its structure is illustrated in Figure 2.

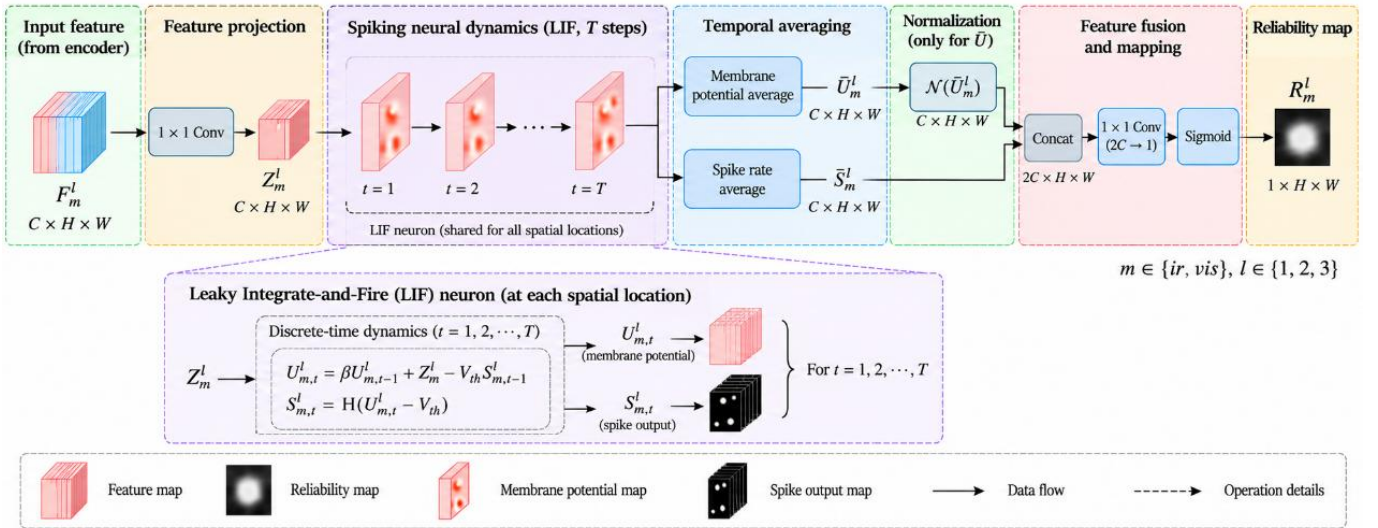


Figure 2. Detailed structure of the Spiking Modality Reliability Estimator (SMRE)

Because the source images are static, the projected representation is repeatedly supplied to the LIF neurons during a short simulation window:

$$Z_{m,t}^l = Z_m^l, t = 1, 2, \dots, T \quad (6)$$

The membrane potential is updated according to

$$U_{m,t}^l = \beta U_{m,t-1}^l + Z_{m,t}^l - V_{th} S_{m,t-1}^l \quad (7)$$

where, $U_{m,t}^l$ denotes the membrane potential at simulation step t , β is the membrane decay factor, V_{th} is the firing threshold,

and $S_{m,t-1}^l$ represents the firing state at the preceding step. The final term implements reset by subtraction following neuronal firing.

The firing state is determined by

$$S_{m,t}^l = H(U_{m,t}^l - V_{th}) \quad (8)$$

where, $H(\cdot)$ is the threshold function. Its non-differentiable derivative is replaced by a smooth surrogate gradient during backpropagation, enabling joint optimization of the spiking and continuous components.

After T simulation steps, the membrane state and firing

activity are temporally averaged:

$$\bar{U}_m^l = \frac{1}{T} \sum_{t=1}^T U_{m,t}^l, \bar{S}_m^l = \frac{1}{T} \sum_{t=1}^T S_{m,t}^l \quad (9)$$

The averaged membrane state preserves continuous accumulated evidence, including sub-threshold responses, while the averaged spike activity represents the frequency with which local responses exceed the neuronal firing threshold.

The two representations are jointly used to construct the modality reliability map:

$$R_m^l = \sigma[C_r^l([\mathcal{N}(\bar{U}_m^l), \bar{S}_m^l])] \quad (10)$$

where, $\mathcal{N}(\cdot)$ denotes channel-wise normalization of the mean membrane state, C_r^l denotes a learnable 1×1 convolution, $[\cdot, \cdot]$ denotes channel concatenation, and $\sigma(\cdot)$ denotes the sigmoid function.

Accordingly,

$$0 < R_m^l(x, y) < 1 \quad (11)$$

for every spatial position (x, y)

A larger value represents a stronger learned reliability response derived jointly from accumulated membrane evidence and firing behavior. The resulting map is used as a relative control signal for subsequent feature interaction and fusion rather than as a physically calibrated sensor-confidence probability. Importantly, the membrane-state component prevents the reliability estimator from depending exclusively on binary firing events. Weak but persistent structures can therefore contribute to reliability estimation even when they do not repeatedly cross the firing threshold.

3.4 Reliability-Modulated Multi-Level Cross-Modal Interaction

RMCI exchanges complementary infrared and visible information while preventing features from being transferred indiscriminately. Its detailed structure is illustrated in Figure 3.

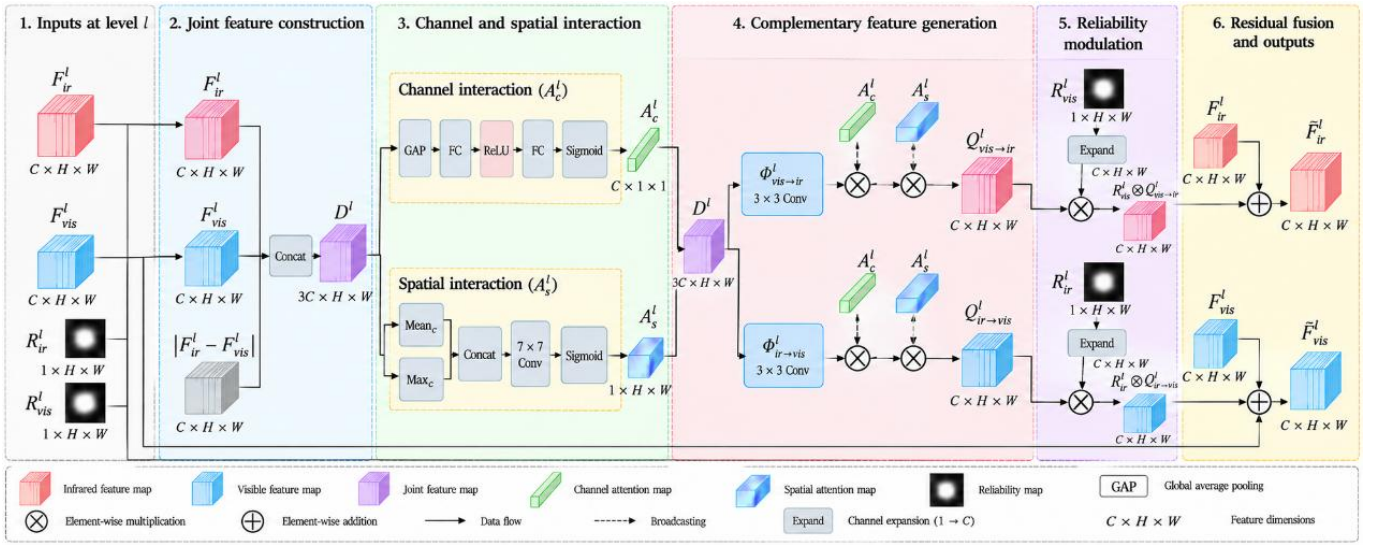


Figure 3. Detailed structure of the Reliability-Modulated Multi-Level Cross-Modal Interaction (RMCI) module

At each feature level, the infrared feature, visible feature, and their absolute difference are first concatenated:

$$D^l = [F_{ir}^l, F_{vis}^l, |F_{ir}^l - F_{vis}^l|] \quad (12)$$

The difference term explicitly represents disagreement between the two modalities and allows the interaction module to detect regions where their information differs substantially.

Channel interaction is estimated from globally pooled features:

$$A_c^l = \sigma[W_{c2}^l(\delta[W_{c1}^l(\text{GAP}(D^l))])] \quad (13)$$

where, GAP denotes global average pooling, W_{c1}^l and W_{c2}^l denote learnable 1×1 convolutional mappings, and ReLU is the rectified linear activation function.

A spatial interaction map is obtained from channel-wise average and maximum responses:

$$A_s^l = \sigma[C_s^l([\text{Mean}_c(D^l), \text{Max}_c(D^l)])] \quad (14)$$

where, channel-wise average and maximum pooling are applied to D^l , and the resulting two-channel spatial descriptor is processed by a 7×7 convolution followed by a sigmoid activation.

The candidate information transferred from the visible branch to the infrared branch is

$$Q_{vis \rightarrow ir}^l = \Phi_{vis \rightarrow ir}^l(D^l) \odot A_c^l \odot A_s^l \quad (15)$$

while the reverse-direction complementary information is

$$Q_{ir \rightarrow vis}^l = \Phi_{ir \rightarrow vis}^l(D^l) \odot A_c^l \odot A_s^l \quad (16)$$

where, $\Phi_{vis \rightarrow ir}^l$ and $\Phi_{ir \rightarrow vis}^l$ are independent learnable 3×3 convolutional mappings and $\odot$ represents element-wise multiplication.

The reliability of the contributing modality subsequently regulates the amount of information transferred:

$$\tilde{F}_{ir}^l = F_{ir}^l + R_{vis}^l \odot Q_{vis \rightarrow ir}^l \quad (17)$$

$$\tilde{F}_{vis}^l = F_{vis}^l + R_{ir}^l \odot Q_{ir \rightarrow vis}^l \quad (18)$$

The residual formulation retains the modality-specific representation and adds only a reliability-weighted complementary component. Consequently, a locally unreliable visible response contributes less to the infrared representation, and vice versa.

This distinction between feature interaction and modality reliability is central to RMCI. Cross-modal relevance

determines what information is potentially useful, whereas the independently estimated reliability map determines how strongly that information should influence the opposite modality.

3.5 Reliability-Guided Adaptive Fusion

After cross-modal interaction, RGAF determines the relative contributions of the two interacted modalities to the fused representation. Its structure is illustrated in Figure 4.

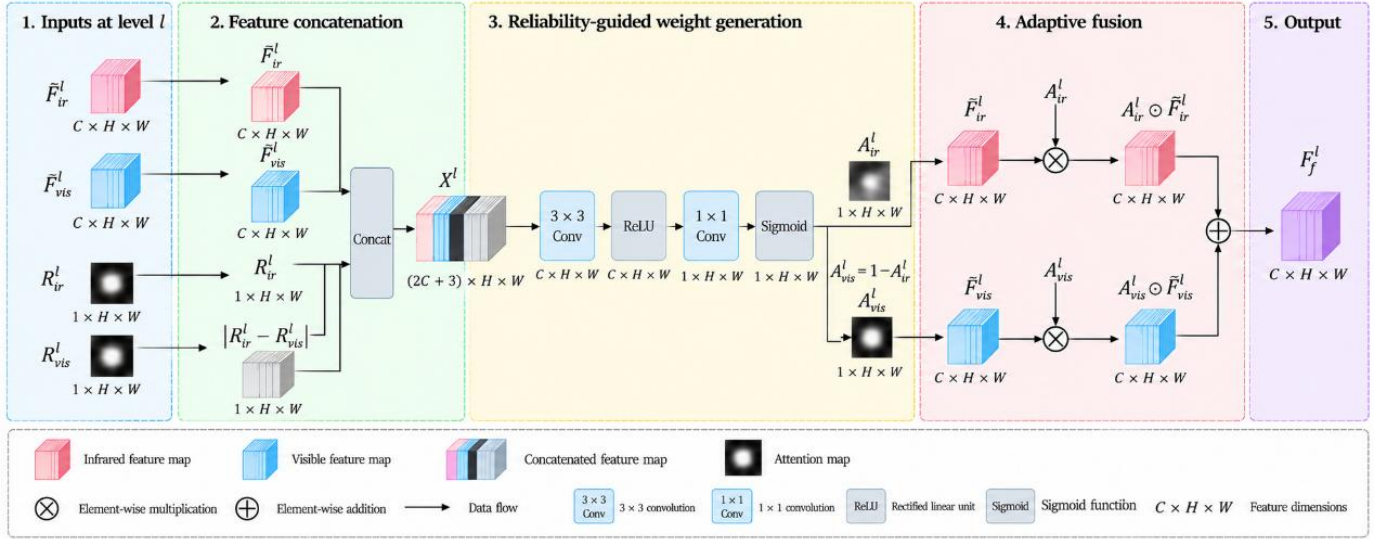


Figure 4. Detailed structure of the Reliability-Guided Adaptive Fusion (RGAF) module

At representation level l , the interacted features, reliability maps, and absolute reliability difference are concatenated:

$$X^l = [\tilde{F}_{ir}^l, \tilde{F}_{vis}^l, R_{ir}^l, R_{vis}^l, |R_{ir}^l - R_{vis}^l|] \quad (19)$$

The difference between R_{ir}^l and R_{vis}^l explicitly represents local disagreement in modality confidence.

A lightweight convolutional predictor generates the infrared contribution map:

$$A_{ir}^l = \sigma[C_2^l(\delta[C_1^l(X^l)])], A_{vis}^l = 1 - A_{ir}^l \quad (20)$$

where, C_1^l and C_2^l are learnable convolutional mappings.

The final fused feature at level l is

$$F_f^l = A_{ir}^l \odot \tilde{F}_{ir}^l + A_{vis}^l \odot \tilde{F}_{vis}^l \quad (21)$$

Because the weighting map varies spatially, the preferred modality can change from one image region to another. Thermally salient targets may receive a larger infrared contribution, whereas texture-rich regions can retain a stronger visible contribution.

The fusion weights are not determined by reliability maps alone. The interacted continuous features are also incorporated into Eq. (19), ensuring that the final fusion decision depends jointly on local modality confidence and actual feature content.

3.6 Multi-level reconstruction

The decoder reconstructs the fused image from the three fused feature levels in a top-down manner. The deepest fused representation is first transformed as

$$G^3 = \mathcal{D}_3(F_f^3) \quad (22)$$

where, $\mathcal{D}_3$ consists of a 3×3 convolution, a ReLU activation, and a residual feature block.

The second-level reconstruction combines the corresponding fused representation with the upsampled deeper feature:

$$G^2 = \mathcal{D}_2([F_f^2, \text{Up}(G^3)]) \quad (23)$$

The full-resolution representation is reconstructed as

$$G^1 = \mathcal{D}_1([F_f^1, \text{Up}(G^2)]) \quad (24)$$

where, $\text{Up}(\cdot)$ denotes bilinear interpolation to the spatial resolution of the corresponding shallower feature map.

Finally, a 1×1 convolution followed by a sigmoid activation produces the fused image:

$$I_f = \sigma[\mathcal{C}_{\text{out}}(G^1)] \quad (25)$$

This hierarchical reconstruction process retains high-level contextual information while progressively restoring local boundaries, textures, and fine spatial structures from shallower fused representations.

3.7 Training objective

A unique ground-truth fused image is generally unavailable for IVIF. BMIF-Net is therefore trained using complementary unsupervised constraints designed to preserve source intensity,

gradients, and structural information.

The intensity loss encourages the fused image to preserve locally prominent source responses:

$$\mathcal{L}_{\text{int}} = \|I_f - \max(I_{ir}, I_{vis})\|_1 \quad (26)$$

where, the maximum operation is performed element-wise.

The gradient loss encourages preservation of strong boundaries and local structural details:

$$\mathcal{L}_{\text{grad}} = \|\nabla I_f - \max(\nabla I_{ir}, \nabla I_{vis})\|_1 \quad (27)$$

where, ∇ denotes the gradient-magnitude operator calculated from horizontal and vertical finite differences.

Structural consistency with both source modalities is further encouraged using

$$\mathcal{L}_{\text{ssim}} = 1 - \frac{1}{2} [\text{SSIM}(I_f, I_{ir}) + \text{SSIM}(I_f, I_{vis})] \quad (28)$$

The complete optimization objective is

$$\mathcal{L} = \lambda_{\text{int}} \mathcal{L}_{\text{int}} + \lambda_{\text{grad}} \mathcal{L}_{\text{grad}} + \lambda_{\text{ssim}} \mathcal{L}_{\text{ssim}} \quad (29)$$

where, λ_{int} , λ_{grad} , and λ_{ssim} control the relative contributions of intensity preservation, gradient preservation, and structural consistency. Their values are specified in Section 4.2 and are kept unchanged throughout the reported experiments.

4. EXPERIMENTAL SETUP

4.1 Datasets and preprocessing

Experiments are conducted on the publicly available MSRS IVIF dataset. MSRS is used for model training, validation, and quantitative evaluation. The overall experimental protocol is illustrated in Figure 5.

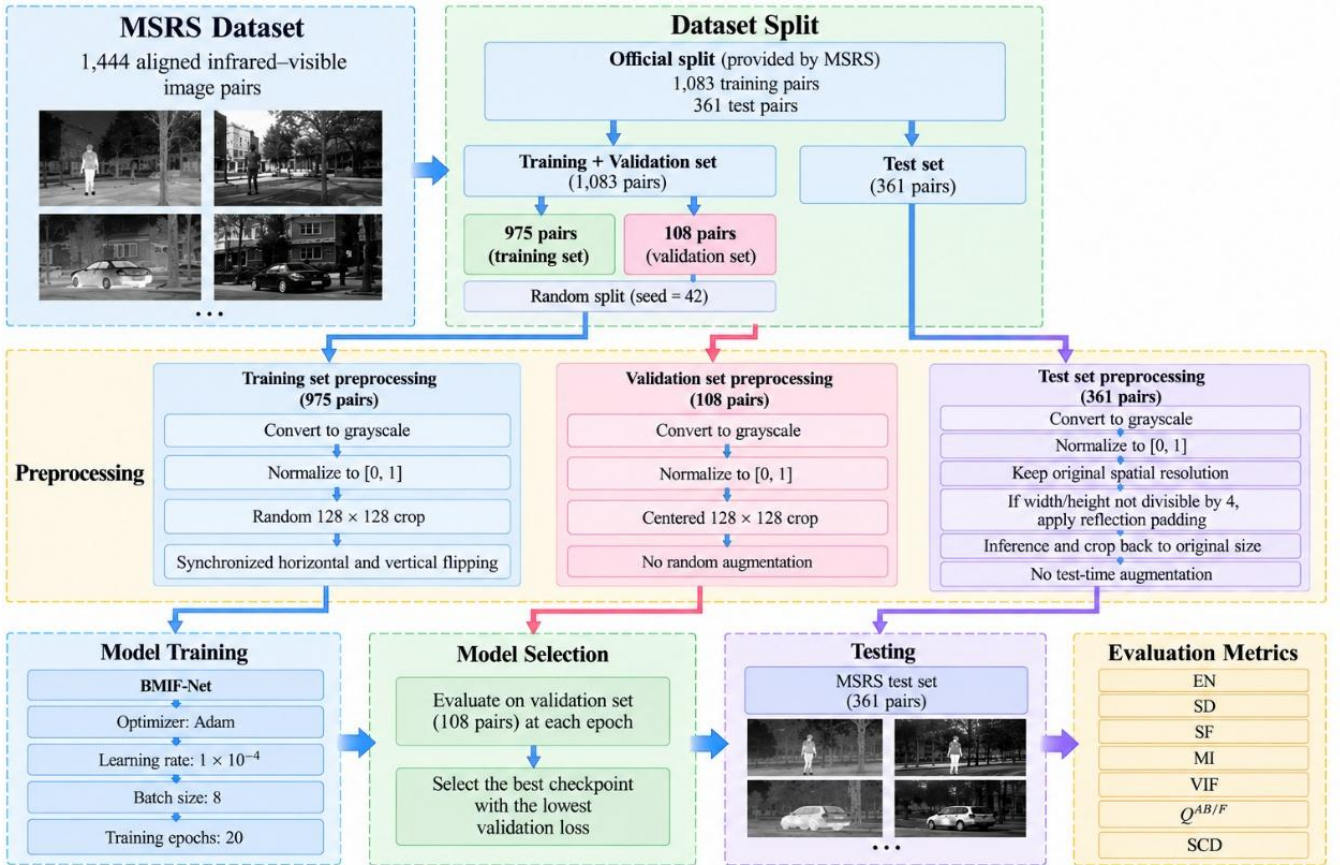


Figure 5. Experimental protocol for training, validation, and testing of BMIF-Net on the MSRS dataset

Note: BMIF-Net = Brain-Inspired Reliability-Guided Multi-Level Interaction Network; MSRS = Multi-Spectral Road Scenarios.

The MSRS dataset contains 1,444 registered infrared-visible image pairs. Following the official split, 1,083 pairs are assigned to the training set and 361 pairs are retained as the test set. To perform model selection without accessing the test data, 10% of the official training split is separated as a validation set using a fixed random seed of 42. This produces 975 training pairs and 108 validation pairs. The original 361-pair test split is used only for final evaluation.

Infrared and visible inputs are converted to single-channel grayscale images and normalized to the interval $[0, 1]$. During training, spatially corresponding infrared and visible images are randomly cropped into 128×128 patches. Synchronized

horizontal and vertical flipping is independently applied with a probability of 0.5 so that geometric correspondence between the two modalities is preserved.

For validation, a centered 128×128 crop is used without random augmentation. During testing, the original spatial resolution is retained. If either image dimension is not divisible by four, reflection padding is applied to the right or bottom boundary before inference, and the output is subsequently cropped back to the original spatial size. No test-time augmentation is employed.

The same source-image pairs and evaluation implementation are used for all reproducible comparison

methods. This ensures that differences in the reported metrics originate from the fusion results rather than from inconsistent dataset subsets or evaluation scripts.

4.2 Implementation details

BMIF-Net is implemented in PyTorch. All experiments reported in this study are conducted using a fixed software and computational environment, and the same training and evaluation pipeline is maintained across the ablation configurations. Because the final complexity analysis focuses on model parameters and floating-point operations rather than hardware-dependent runtime, no GPU-specific inference-time claim is made.

The network is optimized using Adam with an initial learning rate of 1×10^{-4} . The first- and second-order momentum coefficients are set to $\beta_1 = 0.9$ and $\beta_2 = 0.999$, respectively. The batch size is 8, and all reported BMIF-Net models are trained for 20 epochs. A cosine annealing learning-rate schedule is used throughout training. The random seed is fixed to 42 for dataset splitting, parameter initialization, and stochastic training operations whenever deterministic execution is supported.

For the SMRE, the default simulation length is set to $T = 4$, the membrane decay factor is $\beta = 0.75$, and the firing threshold is $V_{th} = 1.0$. These values provide a short temporal accumulation window while limiting the additional computation introduced by repeated neuronal-state updates.

The complete training objective defined in Eq. (29) uses the following coefficients:

$$\lambda_{\text{int}} = 1, \lambda_{\text{grad}} = 10, \lambda_{\text{ssim}} = 5 \quad (30)$$

Accordingly, the loss used in all reported experiments is:

$$\mathcal{L} = \mathcal{L}_{\text{int}} + 10\mathcal{L}_{\text{grad}} + 5\mathcal{L}_{\text{ssim}} \quad (31)$$

The same hyperparameters are used for the baseline and all ablation configurations so that observed differences can be attributed to architectural changes rather than altered optimization settings.

Model selection is based exclusively on validation loss. Whenever the validation loss improves, the corresponding model parameters are retained, and the best validation checkpoint is used for final testing.

The complete BMIF-Net contains approximately 3.256 million trainable parameters. Its computational complexity is reported in Section 5.5 using an input resolution of 128×128 .

4.3 Comparison methods

The quantitative comparison uses four fully reproducible classical fusion baselines: Average fusion, Max fusion, Principal Component Analysis (PCA), and Discrete Wavelet Transform (DWT). These methods are included as transparent reference baselines representing uniform averaging, local intensity selection, statistical projection, and multi-resolution transform-based fusion, respectively.

Average fusion computes the pixel-wise arithmetic mean of the infrared and visible inputs and therefore represents a uniform fusion strategy without modality selection.

Max fusion retains the larger source intensity at each spatial location. It provides a simple local-selection baseline that often preserves strong thermal or high-contrast responses.

PCA fusion projects the two source images into a statistically derived subspace and determines their fusion contributions according to principal-component information. It therefore represents a global statistical fusion strategy.

DWT fusion decomposes the source images into frequency subbands and performs fusion in the transform domain before image reconstruction. It provides a representative multi-resolution baseline.

All four baselines are evaluated using exactly the same infrared-visible test pairs and the same metric implementation used for BMIF-Net. No values are copied from previously published tables. This avoids differences caused by alternative preprocessing procedures, dataset subsets, or metric implementations.

The quantitative comparison is performed on the MSRS test set using the same source-image pairs and metric implementation for all evaluated methods. Recent deep-learning-based fusion methods discussed in Section 2 are included to establish the methodological context of the field, whereas the present quantitative comparison focuses on fully reproducible reference baselines under a unified evaluation pipeline.

In addition to inter-method comparison, controlled ablation experiments are conducted on MSRS. These experiments examine the contribution of cross-modal interaction, the effect of spiking-derived reliability representation, and the influence of the temporal simulation length. All ablation models use the same dataset split, optimizer, loss function, batch size, training duration, and evaluation procedure as the complete model.

4.4 Evaluation metrics

IVIF generally lacks a unique ground-truth fused image. Performance is therefore assessed using seven complementary source-referenced or reference-free metrics: Entropy (EN), Standard Deviation (SD), Spatial Frequency (SF), Mutual Information (MI), Visual Information Fidelity (VIF), QAB/F, and Sum of Correlations of Differences (SCD).

EN measures the information content and intensity diversity of the fused image. A larger EN generally indicates that the output contains a wider range of image information.

SD measures the dispersion of image intensities and provides an indication of global contrast. Higher SD values usually correspond to stronger intensity variation.

SF evaluates spatial activity according to horizontal and vertical intensity changes. It is used to characterize the amount of local detail and structural variation retained in the fused image.

MI measures the statistical information transferred from the two source images to the fused result. The reported value is obtained by summing the MI between the fused image and each source modality.

VIF measures perceptually meaningful information preserved from the source images. The final fusion score combines the fidelity between the fused image and the infrared source with that between the fused image and the visible source.

QAB/F evaluates edge-information preservation from the two source modalities. It considers both gradient magnitude and edge-orientation consistency and is particularly useful for assessing whether source-image boundaries are retained after fusion.

SCD evaluates the amount of complementary information transferred from the two source images by measuring the

correlations between source images and fusion-induced difference components.

For all seven metrics, a larger value indicates better performance. This is indicated by the upward arrow (↑) in the quantitative tables.

The metrics are interpreted jointly rather than individually. EN and MI emphasize statistical information content, SD and SF reflect contrast and spatial activity, while VIF, QAB/F, and SCD provide complementary evidence concerning perceptual fidelity, edge preservation, and source-information transfer. Consequently, the evaluation does not assume that the method obtaining the highest value for a single metric necessarily provides the best overall fusion result.

Quantitative analysis is further supported by component ablation, reliability-representation analysis, temporal simulation analysis, and computational-complexity evaluation.

5. EXPERIMENTAL RESULTS AND ANALYSIS

5.1 Quantitative comparison

The quantitative performance of BMIF-Net is first evaluated on the MSRS test set using seven complementary fusion metrics: EN, SD, SF, MI, VIF, QAB/F, and SCD. All methods are evaluated using the same test pairs and the same metric implementation. The results are reported in Table 1.

Table 1. Quantitative comparison on the MSRS

Method	EN ↑	SD ↑	SF ↑	MI ↑	VIF ↑	QAB/F ↑	SCD ↑
Average	5.7482	23.5892	5.9905	1.9227	0.7000	0.3728	1.2519
Max	6.6242	42.2841	10.9924	4.5242	1.0406	0.7126	1.6297
PCA	6.4675	40.0897	9.9440	8.1954	0.9766	0.6324	1.2519
DWT	5.9784	24.1072	9.0055	2.4599	0.6126	0.4419	1.2552
BMIF-Net	6.4131	41.7063	10.9309	2.9500	2.1077	0.6369	1.6084

Note: EN = entropy; SD = standard deviation; SF = spatial frequency; MI = mutual information; VIF = visual information fidelity; QAB/F = Edge-Based Fusion Quality Index; SCD = sum of the correlation of differences; MSRS = Multi-Spectral Road Scenarios; PCA = Principal Component Analysis; DWT = Discrete Wavelet Transform.

BMIF-Net achieves an EN of 6.4131, an SD of 41.7063, an SF of 10.9309, an MI of 2.9500, a VIF of 2.1077, a QAB/F of 0.6369, and an SCD of 1.6084. Its most pronounced advantage is observed in VIF, where the proposed method substantially exceeds all four classical baselines. Compared with the strongest classical VIF result of 1.0406 obtained by Max fusion, BMIF-Net increases VIF to 2.1077. This indicates that the proposed framework preserves substantially more perceptually relevant information from the two source modalities.

The proposed method also achieves strong SF, QAB/F, and SCD values. Its SF of 10.9309 is very close to the highest value of 10.9924 obtained by Max fusion, indicating that BMIF-Net preserves a high level of spatial variation and fine detail. The QAB/F value of 0.6369 and SCD value of 1.6084 further indicate effective preservation of edge information and complementary source content.

BMIF-Net does not achieve the highest EN or MI. Max fusion obtains a higher EN, while PCA produces the highest MI. These differences are expected because individual metrics emphasize different statistical properties. For instance, PCA directly maximizes variance-related information and can therefore produce a very high MI value without necessarily yielding stronger perceptual fidelity. Similarly, Max fusion

favors locally stronger intensities and may increase entropy or edge-related measures.

The performance of BMIF-Net is therefore better interpreted from a multi-metric perspective. Its dominant VIF and competitive SF, QAB/F, and SCD values indicate that the proposed reliability-guided mechanism places greater emphasis on perceptual fidelity, structural preservation, and complementary information transfer rather than on maximizing a single statistical criterion.

5.2 Core component ablation study

Ablation experiments are conducted on the MSRS test set to examine the contribution of the proposed interaction and reliability-guided fusion mechanisms. All configurations use the same dual-branch encoder, reconstruction decoder, training set, optimizer, objective function, training schedule, and evaluation protocol. Only the fusion mechanism is changed.

The Baseline configuration removes SMRE, RMCI, and RGAF. Infrared and visible features are directly averaged at each representation level:

$$F_f^l = \frac{1}{2}(F_{ir}^l + F_{vis}^l) \quad (32)$$

This configuration provides a reference for evaluating whether the proposed interaction and reliability-guided modules contribute beyond the underlying encoder-decoder architecture.

The RMCI-only configuration removes SMRE and RGAF. To isolate the contribution of cross-modal interaction, the two reliability maps are fixed to one:

$$R_{ir}^l = R_{vis}^l = 1 \quad (33)$$

RMCI therefore performs bidirectional information exchange without reliability modulation. The interacted features are subsequently averaged:

$$F_f^l = \frac{1}{2}(\tilde{F}_{ir}^l + \tilde{F}_{vis}^l) \quad (34)$$

The Full BMIF-Net retains SMRE, RMCI, and RGAF. SMRE estimates modality reliability, RMCI uses the reliability maps to regulate bidirectional feature exchange, and RGAF generates spatially varying fusion weights according to Eq. (21).

The quantitative results are shown in Table 2.

Table 2. Core component ablation on MSRS

Configuration	EN ↑	SD ↑	SF ↑	MI ↑	VIF ↑	QAB/F ↑	SCD ↑
Baseline	6.3961	41.5264	9.6709	2.9028	1.9584	0.5638	1.5754
RMCI only	6.3899	41.7174	10.0537	2.9561	1.9866	0.5957	1.5776
Full BMIF-Net	6.4131	41.7063	10.9309	2.9500	2.1077	0.6369	1.6084

Note: EN = entropy; SD = standard deviation; SF = spatial frequency; MI = mutual information; VIF = visual information fidelity; QAB/F = Edge-Based Fusion Quality Index; SCD = sum of the correlation of differences; MSRS = Multi-Spectral Road Scenarios; RMCI = Reliability-Modulated Multi-Level Cross-Modal Interaction; BMIF-Net = Brain-Inspired Reliability-Guided Multi-Level Interaction Network.

Introducing RMCI improves SF from 9.6709 to 10.0537, MI from 2.9028 to 2.9561, VIF from 1.9584 to 1.9866, QAB/F from 0.5638 to 0.5957, and SCD from 1.5754 to 1.5776. These

changes demonstrate that explicit cross-modal interaction facilitates the exchange of complementary infrared and visible information.

The full model produces a larger improvement in the metrics most closely associated with structural and perceptual preservation. Relative to the Baseline, SF increases from 9.6709 to 10.9309, VIF from 1.9584 to 2.1077, QAB/F from 0.5638 to 0.6369, and SCD from 1.5754 to 1.6084.

The comparison between the RMCI-only configuration and the complete model shows that introducing the full reliability-guided fusion framework provides additional gains beyond unrestricted cross-modal interaction. In the complete model, spiking-derived reliability information is used to modulate bidirectional feature transfer and to support spatially adaptive modality weighting. The improvements in VIF, QAB/F, and SCD therefore support the combined contribution of reliability estimation and reliability-guided fusion, although the individual effects of SMRE and RGAF are not isolated separately in the present ablation design.

5.3 Analysis of spiking reliability representation

The reliability representation used in SMRE is further examined to determine whether spike activity provides useful information beyond continuous membrane responses.

In the membrane-only configuration, the reliability map is estimated exclusively from the normalized average membrane potential:

$$R_{m,U}^l = \sigma[C_{r,U}^l(\mathcal{N}(\bar{U}_m^l))] \quad (35)$$

The complete SMRE jointly uses the average membrane potential and average firing activity:

$$R_m^l = \sigma[C_r^l([\mathcal{N}(\bar{U}_m^l), \bar{S}_m^l])] \quad (36)$$

The corresponding results are presented in Table 3.

Table 3. Effect of incorporating spike activity into SMRE reliability representation on MSRS

SMRE Representation	EN ↑	MI ↑	VIF ↑	QAB/F ↑	SCD ↑
Membrane only	6.4596	2.9982	2.0991	0.6329	1.5959
Membrane + Spike	6.4131	2.9500	2.1077	0.6369	1.6084

Note: EN = entropy; MI = mutual information; VIF = visual information fidelity; QAB/F = Edge-Based Fusion Quality Index; SCD = sum of the correlation of differences; MSRS = Multi-Spectral Road Scenarios; SMRE = Spiking Modality Reliability Estimator.

The membrane-only representation achieves higher EN and MI, indicating that continuous membrane states retain strong statistical information from the encoded features. However, adding spike activity improves VIF from 2.0991 to 2.1077, QAB/F from 0.6329 to 0.6369, and SCD from 1.5959 to 1.6084.

Membrane potential and firing activity therefore capture complementary aspects of modality responsiveness. The membrane state accumulates continuous input evidence and therefore retains sub-threshold responses, whereas spike activity emphasizes responses that repeatedly exceed the neuronal firing threshold. Combining the two representations slightly shifts the fusion behavior away from maximizing statistical information content and toward stronger perceptual fidelity, edge preservation, and complementary structural transfer.

The result also clarifies the role of spiking computation in BMIF-Net. Spike activity does not uniformly improve every metric, nor is it used to replace continuous image representation. Instead, the present comparison indicates that incorporating spike activity in addition to membrane-state information provides an additional reliability cue and improves several metrics associated with perceptual and structural fusion quality. Because a spike-only configuration is not evaluated, the experiment is intended to assess the incremental contribution of spike activity rather than to provide a complete comparison among alternative reliability representations.

5.4 Effect of the spiking simulation length

The simulation length T determines the number of neuronal-state updates performed by SMRE. A longer simulation window allows additional membrane accumulation and firing events but also increases computational cost. To examine the effect of a short temporal simulation window on fusion performance, two representative settings are compared:

$$T \in \{2, 4\} \quad (37)$$

All remaining model and training parameters are kept unchanged.

The quantitative results are shown in Table 4.

Table 4. Effect of the spiking simulation length on MSRS

T	EN ↑	VIF ↑	QAB/F ↑	SCD ↑
2	6.4190	2.1317	0.6325	1.6139
4	6.4131	2.1077	0.6369	1.6084

Note: EN = entropy; VIF = visual information fidelity; QAB/F = Edge-Based Fusion Quality Index; SCD = sum of the correlation of differences; MSRS = Multi-Spectral Road Scenarios; T denotes the simulation length (number of neuronal-state updates in SMRE).

The differences between the two settings are relatively small. With $T = 2$, the model achieves slightly higher EN, VIF, and SCD, whereas $T = 4$ produces a marginally higher QAB/F.

The limited variation across the two settings shows that useful reliability information can be obtained within a short simulation window. Even two state updates are sufficient to generate useful membrane and firing responses. Increasing the simulation length to four steps does not produce a uniform improvement across all metrics, but it provides a moderately longer response-accumulation window and slightly stronger edge preservation.

The default configuration therefore uses

$$T = 4 \quad (38)$$

as a moderate temporal setting for the reliability-estimation process. This choice retains a short simulation window while allowing both membrane accumulation and spike activity to contribute to reliability estimation.

The results should not be interpreted as evidence that larger values of T necessarily improve fusion quality. Instead, they indicate that satisfactory reliability estimation can be achieved with a small number of neuronal-state updates, which is favorable for controlling the additional computational cost introduced by the spiking branch.

5.5 Computational complexity

The computational cost of BMIF-Net is evaluated in terms of trainable parameters and floating-point operations. The results are reported in Table 5.

Table 5. Computational complexity of BMIF-Net

Model	Parameters (M)	FLOPs (G)	Input Size
BMIF-Net	3.256	7.843	128×128

Note: BMIF-Net = Brain-Inspired Reliability-Guided Multi-Level Interaction Network.

For an input resolution of 128×128 , BMIF-Net contains approximately 3.256 million trainable parameters and requires approximately 7.843 GFLOPs.

The parameter count remains moderate despite the use of dual modality-specific encoders, three-level reliability estimation, cross-modal interaction, and adaptive fusion. This is partly because SMRE operates on projected intermediate representations and generates single-channel spatial reliability maps rather than introducing a second large reconstruction backbone.

The additional computation associated with the brain-inspired component mainly arises from repeated LIF-state updates over T simulation steps. RMCi further introduces channel and spatial interaction operations at three representation levels. Nevertheless, these modules operate on hierarchical intermediate features and therefore do not result in excessive model growth.

The reported FLOPs should be interpreted together with the temporal nature of the spiking branch. Conventional FLOP-counting tools primarily account for convolutional and matrix operations and may not fully represent the cost of element-wise membrane-state updates. For this reason, the present analysis uses parameter count and standard FLOPs primarily as architecture-level complexity indicators rather than as hardware-specific runtime measures.

Overall, the complexity results show that BMIF-Net maintains a relatively compact model size while supporting explicit reliability estimation and hierarchical cross-modal interaction. This provides a practical balance between representational capability and computational cost for IVIF.

5.6 Summary of experimental findings

The experiments provide three main observations.

First, BMIF-Net demonstrates its clearest quantitative advantage in perceptual information fidelity. On MSRS, the proposed method achieves a VIF of 2.1077, substantially exceeding the classical baselines, while maintaining competitive SF, QAB/F, and SCD values. This indicates that the network preserves a strong balance between perceptually meaningful information, spatial detail, and complementary source content.

Second, the ablation results show that multi-level cross-modal interaction improves the baseline model, while the complete reliability-guided framework provides further gains in SF, VIF, QAB/F, and SCD. The results therefore support the use of reliability information not only during final modality weighting but also during cross-modal information transfer.

Third, the reliability-representation analysis indicates that membrane states and spike activity play complementary roles. Continuous membrane information provides strong statistical representation, whereas adding firing activity improves

several structural and perceptual metrics. The temporal analysis further shows that useful reliability estimates can be obtained using a relatively short neuronal simulation.

Taken together, these results support the central design principle of BMIF-Net: spiking dynamics are most useful in this framework as a reliability-estimation mechanism that controls continuous multimodal information flow, rather than as a replacement for continuous image representation and reconstruction.

6. DISCUSSION

BMIF-Net is designed around a clear separation between image-content representation and modality reliability estimation. Continuous convolutional features are retained throughout the main fusion pathway so that thermal intensity, edges, textures, and structural details can be reconstructed without being repeatedly quantized into discrete spike events. Spiking dynamics are instead used to estimate how reliably each modality responds to local scene content. This division of responsibility allows neuronal dynamics to guide fusion decisions while preserving the continuous representations required for detail-sensitive image reconstruction.

The experimental results support this design choice. BMIF-Net shows its clearest advantage in perceptual information fidelity while maintaining competitive performance in spatial detail, edge preservation, and complementary information transfer. This performance profile suggests that the reliability-guided architecture primarily improves the preservation of perceptually meaningful source information without sacrificing structural content.

The component ablation results further clarify the role of cross-modal interaction. RMCi improves the baseline by enabling explicit bidirectional exchange of complementary modality information, while the complete BMIF-Net provides additional gains after reliability-guided modulation and adaptive fusion are introduced. This difference indicates that cross-modal interaction alone is not sufficient; controlling the strength of transferred information according to modality reliability is also important.

Feature relevance and modality reliability are not necessarily equivalent. A feature may exhibit a strong response while still being locally unreliable because of poor illumination, thermal saturation, sensor noise, or other imaging degradation. RMCi therefore does not treat all complementary information as equally trustworthy. Instead, the estimated reliability of the contributing modality regulates the amount of information transferred to the opposite branch. RGAF subsequently combines the interacted continuous features with the reliability maps to determine spatially adaptive modality contributions. Reliability thus functions as a control signal at both the cross-modal interaction stage and the final fusion stage.

The reliability-representation experiment provides additional insight into the role of spiking computation. The results indicate that membrane-state information and spike activity contribute differently to reliability estimation. Membrane potentials preserve accumulated sub-threshold evidence, whereas spike activity emphasizes repeated threshold-crossing responses. Their combination therefore provides complementary information for controlling multimodal fusion.

The membrane potential is particularly useful for retaining

weak edges, low-contrast textures, and other structures that may remain informative even when neuronal firing is infrequent. Spike activity, by contrast, reflects responses that repeatedly exceed the firing threshold and therefore emphasizes stronger and more persistent neuronal activity. The joint representation combines accumulated response strength with threshold-driven saliency, providing a richer reliability cue than membrane-state information alone.

The results also show that spiking computation should not be interpreted as a universal performance enhancer. Incorporating spike activity does not improve every evaluation metric. Instead, it changes the balance of the fused representation and provides additional support for perceptual fidelity, edge preservation, and complementary information transfer. This observation is consistent with the intended role of SMRE, in which spiking dynamics are used to characterize modality reliability rather than to replace continuous image representation or to maximize every statistical measure.

The temporal simulation experiment indicates that useful reliability estimates can be obtained with a short neuronal sequence, as the results for $T = 2$ and $T = 4$ remain close across the evaluated metrics. $T = 2$ produces slightly higher EN, VIF, and SCD, whereas $T = 4$ provides a slightly higher QAB/F. Useful membrane accumulation and firing information can therefore be obtained within a short simulation window.

The default setting of $T = 4$ should be regarded as a moderate operating point rather than as an empirically dominant optimum. It provides additional neuronal-state updates while keeping the temporal simulation short. The relatively small difference between the two evaluated settings also suggests that the proposed reliability mechanism is not highly sensitive to moderate changes in simulation length, which reduces the need for extensive temporal hyperparameter tuning.

The quantitative results also highlight the importance of interpreting fusion metrics jointly. Different methods emphasize different statistical or perceptual properties, and a high score on a single metric does not necessarily correspond to the most balanced fusion result. BMIF-Net is primarily characterized by strong perceptual fidelity together with competitive structural and complementary-information preservation, rather than by maximizing every individual metric.

The multi-level design also contributes to this behavior. Infrared and visible information have different significance at different representation depths. Shallow features retain fine gradients, local boundaries, and textures, whereas deeper representations encode broader structures and salient regions. Applying reliability estimation, cross-modal interaction, and adaptive fusion at all three representation levels allows BMIF-Net to regulate information flow according to both spatial location and representation scale. Restricting reliability-guided fusion to the deepest level could reduce sensitivity to fine spatial information that may already have been attenuated by downsampling.

From a computational perspective, BMIF-Net maintains a relatively compact architecture despite incorporating multi-level reliability estimation and cross-modal interaction. The reported parameter count and FLOPs provide architecture-level complexity indicators, although standard FLOP profiling does not fully capture the cost of repeated membrane-state updates. Hardware-specific runtime therefore depends on the implementation and execution platform.

Several limitations remain. First, BMIF-Net assumes that

the infrared and visible inputs are already spatially registered. Registration errors may introduce inconsistent structures into cross-modal interaction and reduce the reliability of spatially aligned fusion weights. Second, modality reliability is inferred from learned feature responses rather than from an explicit physical uncertainty model. Severe sensor noise, thermal saturation, motion artifacts, or extreme visible-light degradation may therefore still lead to inaccurate reliability estimates. Incorporating explicit uncertainty estimation or sensor-quality modeling could improve the interpretability and robustness of the reliability mechanism.

Third, the present temporal analysis is limited to two representative simulation lengths. Although the results show that useful reliability estimates can be obtained with short simulations, a broader investigation of simulation length, membrane decay, firing threshold, and spike sparsity would provide a more complete understanding of how neuronal dynamics affect fusion behavior.

Fourth, the current quantitative comparison focuses on fully reproducible classical baselines and controlled ablation experiments. Broader comparison with additional recent deep-learning-based fusion models under a unified implementation and evaluation pipeline would provide a more comprehensive assessment of the proposed method. Such comparisons are complicated by differences in official checkpoints, preprocessing procedures, dataset subsets, and metric implementations, but they remain an important direction for further evaluation.

Future work may therefore focus on four directions: integrating image registration and fusion within a unified framework, incorporating explicit uncertainty modeling into modality reliability estimation, developing more efficient event-driven or adaptive temporal computation, and extending the proposed reliability-guided interaction strategy to other heterogeneous imaging tasks such as infrared–event fusion, multimodal medical imaging, and remote-sensing image fusion.

Overall, the main contribution of BMIF-Net lies in using spiking dynamics as a selective reliability-estimation mechanism within an otherwise continuous fusion framework. This design preserves the advantages of continuous image reconstruction while introducing an additional mechanism for regulating heterogeneous information flow. The resulting framework provides a practical basis for reliability-aware IVIF with strong perceptual fidelity, competitive structural preservation, and moderate computational complexity.

7. CONCLUSIONS

This study presents BMIF-Net, a BMIF-Net for IVIF. The framework separates continuous image-content representation from spiking-based modality reliability estimation. Continuous convolutional features are retained for detailed image encoding and reconstruction, while membrane-state dynamics and spike activity are used to estimate local modality reliability.

The proposed SMRE module derives reliability information from temporally averaged membrane potentials and firing activity. These reliability cues are subsequently introduced into RMCI to regulate bidirectional cross-modal feature transfer and into RGAF to support spatially adaptive modality weighting. In this way, spiking dynamics function as a control mechanism for heterogeneous information flow rather than as

a replacement for continuous image representation.

Experiments on the MSRS dataset show that BMIF-Net achieves an EN of 6.4131, an SD of 41.7063, an SF of 10.9309, an MI of 2.9500, a VIF of 2.1077, a QAB/F of 0.6369, and an SCD of 1.6084. The clearest quantitative advantage is observed in VIF, while the proposed method also maintains competitive performance in spatial-detail preservation, edge-information retention, and complementary-information transfer.

The ablation experiments further show that explicit cross-modal interaction improves the baseline model and that the complete reliability-guided framework provides additional gains beyond unrestricted interaction. The reliability-representation analysis indicates that membrane-state information and spike activity provide complementary cues: membrane potentials preserve accumulated sub-threshold evidence, whereas spike activity contributes additional threshold-driven information for reliability estimation.

The temporal analysis shows that useful reliability estimates can be obtained with a short neuronal simulation, with relatively small differences between $T = 2$ and $T = 4$. The default setting of $T = 4$ therefore serves as a moderate operating point rather than an empirically dominant optimum. In terms of model complexity, BMIF-Net contains approximately 3.256 million trainable parameters and requires approximately 7.843 GFLOPs for an input resolution of 128×128 .

Several limitations remain. The current framework assumes spatially registered infrared and visible inputs, estimates modality reliability from learned feature responses rather than explicit physical uncertainty, and evaluates temporal behavior using only two representative simulation lengths. In addition, the present quantitative comparison focuses on fully reproducible classical baselines and controlled ablation experiments. Broader evaluation against recent deep-learning-based fusion methods under a unified implementation would further strengthen the empirical assessment.

Future work will investigate joint registration and fusion, uncertainty-aware reliability modeling, more efficient event-driven or adaptive temporal computation, and broader evaluation across heterogeneous multimodal fusion scenarios.

Overall, BMIF-Net demonstrates that spiking dynamics can be effectively incorporated into a continuous image-fusion architecture as a selective reliability-estimation mechanism. By combining spiking-derived reliability cues with reliability-modulated interaction and adaptive fusion, the framework provides a practical approach for preserving perceptually meaningful and structurally complementary information in infrared-visible image fusion.

ACKNOWLEDGMENT

This work was supported by the Science and Technology Project of China Southern Power Grid Co., Ltd. (Grant No.: GZKJXM20232513) and the Science and Technology Program of Guizhou Province (Grant No.: [2024] 055).

REFERENCES

- [1] Liu, J., Wu, G., Liu, Z., et al. (2024). Infrared and visible image fusion: From data compatibility to task adaption. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(4): 2349-2369. <https://doi.org/10.1109/tpami.2024.3521416>
- [2] Zhang, X., Demiris, Y. (2023). Visible and infrared image fusion using deep learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8): 10535-10554. <https://doi.org/10.1109/TPAMI.2023.3261282>
- [3] Yang, K., Xiang, W., Chen, Z., Zhang, J., Liu, Y. (2024). A review on infrared and visible image fusion algorithms based on neural networks. *Journal of Visual Communication and Image Representation*, 101: 104179. <https://doi.org/10.1016/j.jvcir.2024.104179>
- [4] Ma, J., Ma, Y., Li, C. (2019). Infrared and visible image fusion methods and applications: A survey. *Information Fusion*, 45: 153-178. <https://doi.org/10.1016/j.inffus.2018.02.004>
- [5] Zhang, X., Ye, P., Xiao, G. (2020). VIFB: A visible and infrared image fusion benchmark. In *2020 IEEE/CVF conference on computer vision and pattern recognition workshops (CVPRW)*, Seattle, WA, USA, pp. 468-478. <https://doi.org/10.1109/CVPRW50498.2020.00060>
- [6] Li, H., Wu, X. (2018). DenseFuse: A fusion approach to infrared and visible images. *IEEE Transactions on Image Processing*, 28(5): 2614-2623. <https://doi.org/10.1109/tip.2018.2887342>
- [7] Ma, J., Yu, W., Liang, P., Li, C., Jiang, J. (2019). FusionGAN: A generative adversarial network for infrared and visible image fusion. *Information Fusion*, 48: 11-26. <https://doi.org/10.1016/j.inffus.2018.09.004>
- [8] Xu, H., Ma, J., Jiang, J., Guo, X., Ling, H. (2020). U2Fusion: A unified unsupervised image fusion network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1): 502-518. <https://doi.org/10.1109/tpami.2020.3012548>
- [9] Li, H., Wu, X., Kittler, J. (2021). RFN-Nest: An end-to-end residual fusion network for infrared and visible images. *Information Fusion*, 73: 72-86. <https://doi.org/10.1016/j.inffus.2021.02.023>
- [10] Zhao, Z., Xu, S., Zhang, C., Liu, J., Li, P., Zhang, J. (2020). DIDFuse: Deep image decomposition for infrared and visible image fusion. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence Main track*, Yokohama, Japan, pp. 970-976. <https://doi.org/10.24963/ijcai.2020/135>
- [11] Tang, L., Yuan, J., Zhang, H., Jiang, X., Ma, J. (2022). PIAFusion: A progressive infrared and visible image fusion network based on illumination aware. *Information Fusion*, 83-84: 79-92. <https://doi.org/10.1016/j.inffus.2022.03.007>
- [12] Zhao, W., Xie, S., Zhao, F., He, Y., Lu, H. (2023). Metafusion: Infrared and visible image fusion via meta-feature embedding from object detection. In *2023 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, Vancouver, BC, Canada, pp. 13955-13965. <https://doi.org/10.1109/CVPR52729.2023.01341>
- [13] Liu, J., Dian, R., Li, S., Liu, H. (2022). SGFusion: A saliency guided deep-learning framework for pixel-level image fusion. *Information Fusion*, 91: 205-214. <https://doi.org/10.1016/j.inffus.2022.09.030>
- [14] Cheng, C., Xu, T., Wu, X. (2022). MUFusion: A general unsupervised image fusion network based on memory unit. *Information Fusion*, 92: 80-92. <https://doi.org/10.1016/j.inffus.2022.11.010>

- [15] Zhang, X., Zhai, H., Liu, J., Wang, Z., Sun, H. (2023). Real-time infrared and visible image fusion network using adaptive pixel weighting strategy. *Information Fusion*, 99: 101863. <https://doi.org/10.1016/j.inffus.2023.101863>
- [16] Zhao, F., Zhao, W., Lu, H. (2023). Interactive feature embedding for infrared and visible image fusion. *IEEE Transactions on Neural Networks and Learning Systems*, 35(9): 12810-12822. <https://doi.org/10.1109/tnnls.2023.3264911>
- [17] Liu, J., Lin, R., Wu, G., Liu, R., Luo, Z., Fan, X. (2023). CoCoNet: Coupled contrastive learning network with multi-level feature ensemble for multi-modality image fusion. *International Journal of Computer Vision*, 132(5): 1748-1775. <https://doi.org/10.1007/s11263-023-01952-1>
- [18] Dong, A., Wang, L., Liu, J., Lv, G., Zhao, G., Cheng, J. (2024). MFIFusion: An infrared and visible image enhanced fusion network based on multi-level feature injection. *Pattern Recognition*, 152: 110445. <https://doi.org/10.1016/j.patcog.2024.110445>
- [19] Wang, J., Jiang, M., Kong, J. (2024). MDAN: Multilevel dual-branch attention network for infrared and visible image fusion. *Optics and Lasers in Engineering*, 176: 108042. <https://doi.org/10.1016/j.optlaseng.2024.108042>
- [20] Gao, X., Liu, S. (2024). BCMFIFuse: A bilateral cross-modal feature interaction-based network for infrared and visible image fusion. *Remote Sensing*, 16(17): 3136. <https://doi.org/10.3390/rs16173136>
- [21] Huang, J., Chen, Z., Ma, Y., Fan, F., Tang, L., Xiang, X. (2024). PTET: A progressive token exchanging transformer for infrared and visible image fusion. *Image and Vision Computing*, 144: 104957. <https://doi.org/10.1016/j.imavis.2024.104957>
- [22] Chen, B., Luo, S., Chen, M., Zhang, F., He, C., Wu, H. (2024). Infrared and visible image fusion based on a two-stage fusion strategy and feature interaction block. *Optics and Lasers in Engineering*, 182: 108461. <https://doi.org/10.1016/j.optlaseng.2024.108461>
- [23] Tang, W., He, F., Liu, Y. (2024). ITFuse: An interactive transformer for infrared and visible image fusion. *Pattern Recognition*, 156: 110822. <https://doi.org/10.1016/j.patcog.2024.110822>
- [24] Song, W., Li, Q., Gao, M., Chehri, A., Jeon, G. (2024). SFINet: A semantic feature interactive learning network for full-time infrared and visible image fusion. *Expert Systems with Applications*, 261: 125472. <https://doi.org/10.1016/j.eswa.2024.125472>
- [25] Ma, H., Li, H., Cheng, C., Wang, G., Song, X., Wu, X. (2025). S4Fusion: Saliency-aware selective state space model for infrared and visible image fusion. *IEEE Transactions on Image Processing*, 34: 4161-4175. <https://doi.org/10.1109/tip.2025.3583132>
- [26] Zhao, Z., Ma, Y., Huang, J., Wu, K., Wang, G., Fan, F. (2025). CrossMamba: Cross-modal features mixing via Mamba for infrared and visible image fusion. *Infrared Physics & Technology*, 152: 106234. <https://doi.org/10.1016/j.infrared.2025.106234>
- [27] Roy, K., Jaiswal, A., Panda, P. (2019). Towards spike-based machine intelligence with neuromorphic computing. *Nature*, 575: 607-617. <https://doi.org/10.1038/s41586-019-1677-2>
- [28] Eshraghian, J.K., Ward, M., Neftci, E.O., et al. (2023). Training spiking neural networks using lessons from deep learning. *Proceedings of the IEEE*, 111(9): 1016-1054. <https://doi.org/10.1109/jproc.2023.3308088>
- [29] Fang, W., Yu, Z., Chen, Y., Masquelier, T., Huang, T., Tian, Y. (2021). Incorporating learnable membrane time constant to enhance learning of spiking neural networks. In 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, pp. 2641-2651. <https://doi.org/10.1109/ICCV48922.2021.00266>
- [30] Cheng, M., Mo, H. (2026). SpikeVFuse: Enhancing infrared and visible image fusion with spiking neural networks. *IEEE Transactions on Multimedia*, 1-13. <https://doi.org/10.1109/tmm.2026.3729390>