Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.083216

ARTICLE

DSCAttFuseNet: A Structure–Detail–Luminance Decoupled Network for Low-Light Infrared–Visible Image Fusion

Kezhen Xie, Syed Mohd Zahid Syed Zainal Ariffin* and Muhammad Izzad Ramli

Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Shah Alam, Malaysia

*Corresponding Author: Syed Mohd Zahid Syed Zainal Ariffin. Email: zahidzainal@uitm.edu.my

Received: 31 March 2026; Accepted: 13 July 2026; Published: 13 August 2026

ABSTRACT: Low-light infrared–visible image fusion remains challenging due to severe modality imbalance caused by visible-image degradation under insufficient illumination. In low-light conditions, visible images often suffer from luminance attenuation, blurred details, and amplified noise, whereas infrared images preserve stable structural information but lack texture representation. Existing fusion frameworks commonly perform feature interaction within a shared representation space, which may cause degraded visible responses to be progressively suppressed by dominant infrared structures during fusion. To address this issue, this paper proposes DSCAttFuseNet for low-light infrared–visible image fusion. The proposed framework adopts a structure–detail–luminance decoupled modeling strategy to model infrared structural perception, visible-detail enhancement, and luminance-aware reconstruction through complementary branches, thereby alleviating visible-detail degradation and modality imbalance under low-light conditions. In addition, depthwise separable convolution and efficient channel attention are introduced as compact feature extraction components to reduce redundant feature extraction and strengthen discriminative feature responses. A luminance-guided reconstruction strategy together with multi-objective optimization is further employed to preserve structural consistency, improve visible-detail representation, and maintain balanced luminance representation in fused images. Experiments on the LLVIP dataset and cross-dataset evaluation on the unseen TNO dataset demonstrate that the proposed method achieves competitive performance against representative and recent fusion methods in both qualitative and quantitative evaluations. Ablation studies further verify the effectiveness of the proposed decoupled modeling strategy and compact feature extraction design for low-light infrared–visible image fusion.

KEYWORDS: Infrared–visible image fusion; low-light image fusion; modality imbalance; structure–detail–luminance modeling; compact feature representation; luminance-guided fusion

1 Introduction

In practical applications such as nighttime surveillance, intelligent transportation, and security monitoring, visual perception systems often need to operate in low-light environments [1]. In such scenarios, the quality of visible images is usually severely affected by insufficient brightness, blurred details, and increased noise, thereby significantly reducing the ability to acquire reliable visual cues. Compared with visible-light imaging, infrared (IR) sensors are less sensitive to illumination variations and can preserve object contours and structural information even in dark environments [2]. However, infrared images usually lack fine texture details and natural color representation, resulting in limited detail representation. Therefore, relying solely on either visible or infrared images cannot simultaneously satisfy the requirements of structural clarity and natural visual appearance in low-light scenes. In low-light infrared–visible image fusion, severe visible-image degradation reduces the effective contribution of visible information during fusion, causing many

methods to rely more heavily on infrared structural responses. Consequently, the fused images may preserve infrared-dominant contours but still suffer from insufficient visible details and unbalanced luminance distribution [3,4]. As shown in Fig. 1, MGFF [5] shows limited ability to restore visible details and balance infrared-dominant structural responses under low-light degradation.

Figure 1: Visible (VI), infrared (IR), and MGFF [5] fusion results under low-light conditions. VIS suffers from low brightness and noise, IR preserves structural contours, and MGFF shows limited visible-detail restoration and luminance balance.

In low-light scenarios, the key challenge is not only to improve visual quality, but also to prevent degraded visible features from being weakened during multimodal interaction. However, many existing fusion models pay limited attention to how degraded visible information should be represented and balanced during fusion.

Early infrared–visible image fusion mainly relied on traditional strategies based on transform domains or model decomposition, such as wavelet transform [6], NSCT [7], DTCWT [8], as well as sparse representation [9] and low-rank decomposition [10]. These methods are able to enhance edge and structural information to a certain extent and exhibit good interpretability [11]. However, their fusion rules are often manually designed, making it difficult to adaptively characterize complex nonlinear complementary relationships across modalities [12]. When visible images suffer from severe degradation, this limitation is further amplified, often causing the fusion results to be dominated by a single modality and making modality imbalance difficult to alleviate in low-light scenarios. With the rapid development of deep learning, infrared–visible image fusion has benefited from data-driven feature learning [13]. By learning cross-modal representations, these methods are able to capture more complex nonlinear relationships between modalities and generally achieve improved fusion performance. Nevertheless, most deep learning-based fusion approaches are developed mainly under normal illumination and lack explicit modeling of visible degradation in low-light scenarios. When visible inputs are severely degraded, these methods may still rely heavily on infrared responses, resulting in insufficient detail preservation and persistent modality imbalance [14].

Based on the above observations, this paper proposes DSCAttFuseNet, a branch-wise structure–detail–luminance modeling framework for low-light infrared–visible image fusion. Existing fusion frameworks generally perform feature interaction within a shared representation space, which may cause degraded visible features to be progressively suppressed by dominant infrared structural responses under severe illumination degradation. To mitigate this issue, DSCAttFuseNet adopts a structure–detail–luminance decoupled modeling strategy, where infrared structural perception, visible detail enhancement, and luminance estimation are modeled through separate yet complementary branches. Such a design enables the fusion process to preserve infrared structural information, enhance degraded visible details, and maintain a more balanced luminance representation. In addition, depthwise separable convolution and efficient channel attention are

introduced as compact feature extraction components to reduce redundant feature extraction and strengthen discriminative multimodal responses.

The main contributions of this work are summarized as follows:

1. We propose a structure–detail–luminance decoupled fusion framework for low-light infrared–visible image fusion. Instead of directly mixing degraded visible features with infrared features in a shared representation space, the proposed framework separately models infrared structural perception, visible detail enhancement, and luminance estimation, thereby alleviating visible-detail suppression under low-light degradation.
2. We develop a compact feature representation module based on depthwise separable convolution and efficient channel attention to reduce redundant feature extraction and strengthen discriminative multimodal responses. This design supports the proposed decoupled modeling strategy with limited model complexity increase.
3. We introduce a luminance-guided reconstruction strategy together with edge-aware constraints to regulate the reconstruction process under low-light degradation. By jointly constraining structural consistency, luminance representation, and visible-detail preservation, this strategy helps compensate for weakened visible responses during fusion.

2 Related Work

2.1 Traditional Fusion Methods

Early studies on infrared–visible image fusion mainly relied on traditional image processing methods [2]. These methods typically decompose and recombine features of source images in the spatial or frequency domains, and integrate complementary information between modalities into a single image through manually designed fusion rules [15]. According to different processing strategies, traditional approaches can be broadly categorized into multiscale transforms [16], sparse representation [17], low-rank decomposition [18], as well as filter-based methods [19].

Multiscale transform methods, such as wavelet transform [20], NSCT [21], and DWT [22], decompose source images into low-frequency structures and high-frequency textures. These components are fused separately and reconstructed, thereby preserving edges and details to some extent. However, their fixed fusion rules make it difficult to adapt to complex scenarios [23]. Sparse representation-based methods use dictionary coding to represent and reconstruct image features. Such methods are able to highlight local detail information in source images [24]. However, the iterative optimization process often limits their adaptability in complex imaging scenarios [25]. Low-rank decomposition methods, such as LatLRR [18], achieve relatively balanced fusion results by separating global structural components from local details. However, in practical applications, the high-order matrix operations involved introduce significant computational overhead, thereby limiting their applicability in complex real-world environments. Filter-based fusion strategies (such as guided filtering and bilateral filtering [26]) are commonly used for edge enhancement and noise suppression. However, these methods are highly sensitive to filter size and smoothing parameters, and often exhibit insufficient robustness when applied under different imaging conditions.

Traditional fusion methods offer advantages such as interpretability, clear methodological structure, and straightforward implementation [27]. However, they rely heavily on handcrafted features and fixed fusion rules, which limits their adaptability under complex and degraded imaging conditions [2]. Under low-light conditions, severe visible-image degradation further weakens the effectiveness of these handcrafted strategies. As a result, traditional methods often struggle to preserve complementary visible details and may produce infrared-dominant fusion results [15].

2.2 Deep Learning Methods

In recent years, deep learning–based infrared–visible image fusion methods have developed rapidly [2]. Unlike traditional methods that rely on handcrafted features, deep fusion networks automatically extract multi-level representative and structural features across modalities through end-to-end feature learning [15]. Most deep learning-based fusion methods follow an overall "encode–fuse–decode" framework, in which multimodal information is integrated through feature encoding and fusion. Existing methods mainly improve fusion performance through variations in network architecture and feature interaction mechanisms. For example, DenseFuse [28] enhances information transmission among multi-scale features by introducing dense connections, thereby improving reconstruction quality. FusionGAN [29] employs adversarial learning to enhance the detail fidelity and visual realism of fused results, although its training process often suffers from stability issues. In addition,

PMGI [30] formulates image fusion as the proportional maintenance of gradient and intensity information and employs separate gradient and intensity paths with cross-path information exchange to preserve complementary source information, while RFN-Nest [31] achieves multi-scale feature aggregation through a nested residual architecture and demonstrates competitive performance. CDDFuse [32] further explores correlation-driven dual-branch feature decomposition to model global and local multimodal representations in infrared–visible image fusion. Contrast-aware detail preservation has also been explored. For example, weighted contrast map–based fusion frameworks preserve complementary texture and contrast information through adaptive weight-map construction and detail-enhancement mechanisms [33].

In addition, edge-preserving fusion strategies have been investigated to enhance structural detail retention under complex environments. For example, quantum-inspired edge-preserving fusion frameworks generate adaptive weight maps to preserve complementary structural information while reducing redundant noise [34].

Attention mechanisms have also been incorporated into fusion frameworks to enhance salient feature responses and suppress irrelevant interference during feature interaction. For example, SeAFusion [35] introduces channel–spatial joint attention to improve feature selection and noise suppression during multimodal fusion. More recently, Transformer-based and self-supervised fusion frameworks have further improved multimodal feature interaction through global representation learning and progressive cross-attention mechanisms. For example, SMAE-Fusion [36] integrates saliency-aware masked autoencoders with hybrid attention transformers to enhance semantic representation and cross-modal interaction. Although these methods improve global contextual representation, most existing frameworks still rely on implicit shared feature interaction during multimodal representation, which may weaken degraded visible-detail responses under severe low-light conditions.

Meanwhile, lightweight fusion frameworks have also attempted to balance fusion quality and computational efficiency. For example, SIFusion [37] introduces semantic injection and structural re-parameterization mechanisms to improve semantic representation while reducing computational redundancy. Although lightweight architectures reduce redundant feature interaction, explicit modeling of visible-detail degradation under low-light conditions remains limited.

Under low-light degradation, visible images often suffer from luminance attenuation, texture degradation, and noise amplification. When feature interaction is mainly performed within a shared representation space, these degraded visible responses may be weakened by dominant infrared structural information, resulting in infrared-biased fused representations with insufficient visible-detail preservation and unbalanced luminance representation. Although existing fusion frameworks improve feature interaction and visual quality through decomposition strategies, attention mechanisms, Transformer architectures, and

lightweight representation designs, explicit modeling of visible-detail degradation and luminance-aware representation under low-light conditions remains limited. Therefore, developing a fusion framework that can jointly address modality imbalance, visible-detail degradation, and luminance-aware reconstruction remains a challenging issue in low-light infrared–visible image fusion.

3 Proposed Method

3.1 Overall Framework

The proposed DSCAttFuseNet is designed for low-light infrared–visible image fusion, where the visible image usually suffers from luminance attenuation, texture degradation, and noise disturbance, while the infrared image provides relatively stable structural information. In this study, the fusion task is formulated as a structure–detail–luminance coordinated reconstruction problem. The goal is to generate a fused image with preserved infrared structural cues, effective visible details, and balanced luminance representation. To achieve this, the proposed framework consists of luminance–chrominance preprocessing, compact feature encoding, branch-wise feature recalibration, multi-branch decoding, and luminance-guided reconstruction, as illustrated in Fig. 2.

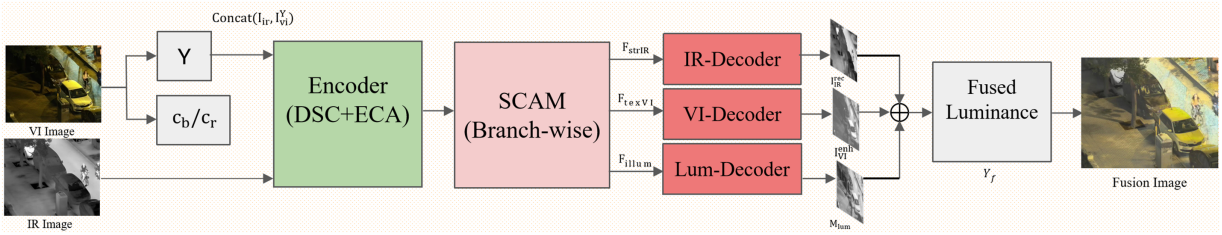


Figure 2: Overall architecture of DSCAttFuseNet, including compact feature extraction, branch-wise structure–detail–luminance modeling, and luminance-guided reconstruction.

In the preprocessing stage, the input visible image I_{vi} is converted into the YCbCr color space, where the luminance component I_{vi}^{Y} and the chrominance components ($I_{vi}^{C_b}$, $I_{vi}^{C_r}$) are separated. Since low-light degradation mainly affects the luminance and texture visibility of the visible image, the network performs fusion at the luminance level. Specifically, I_{vi}^{Y} is concatenated with the infrared image I_{ir} along the channel dimension to construct a two-channel input tensor:

$$X_{in} = \mathrm{Concat}\left(I_{ir}, I_{vi}^{Y}\right) \tag{1}$$

The chrominance components $I_{vi}^{C_b}$ and $I_{vi}^{C_r}$ are not involved in feature encoding, but are retained for the final RGB reconstruction. This setting allows the network to focus on structural preservation, visible-detail enhancement, and luminance reconstruction while maintaining the color information of the visible modality.

Under this formulation, the low-light visible luminance I_{vi}^{Y} is regarded as a degraded observation that still contains useful but weakened detail information. In contrast, the infrared image I_{ir} provides more stable contour and structural responses. Therefore, the fusion process should not directly treat all visible responses as reliable details. Instead, DSCAttFuseNet adopts branch-wise modeling to separately emphasize infrared structure, visible detail, and luminance-related information before reconstruction.

The two-channel input tensor X_{in} is first encoded by a compact feature extraction module based on DSC and ECA. DSC is used to reduce redundant spatial feature extraction, while ECA recalibrates channel responses to enhance structure- and detail-sensitive information. The encoded shared feature is expressed as:

$$F = E(X_{in}), F \in \mathbb{R}^{H \times W \times C} \tag{2}$$

where $E(\cdot)$ denotes the compact encoder, and F is the shared multimodal feature with spatial size $H \times W$ and channel number C. The shared feature F is then fed into the SCAM module for branch-wise functional recalibration. Instead of enforcing strict semantic separation, SCAM generates three complementary feature representations from F:

$$F_m = \alpha_m \odot F, m \in \{\text{strIR}, \text{texVI}, \text{illum}\} \tag{3}$$

where α_m denotes the branch-specific channel attention vector, and $\odot$ denotes element-wise multiplication with broadcasting along the spatial dimensions. The three branches correspond to infrared structure representation, visible texture representation, and luminance representation, respectively.

The branch-specific decoders then generate three outputs: an infrared reconstruction image I_{IR}^{rec}, an enhanced visible luminance image I_{VI}^{enh} and a luminance estimation map M_{lum}. These outputs are combined through a luminance-guided reconstruction strategy to obtain the final fused luminance:

$$Y_f = \max(I_{IR}^{rec}, I_{VI}^{enh} \odot M_{lum}) \tag{4}$$

where $\max(\cdot,\cdot)$ denotes a pixel-wise maximum operation. In this formulation, I_{IR}^{rec} provides stable structural responses, while $I_{VI}^{enh} \odot M_{lum}$ represents luminance-modulated visible-detail information. The pixel-wise maximum operation selects the stronger response between infrared structure and luminance-enhanced visible details, thereby reducing the suppression of weakened visible information during reconstruction.

Finally, the fused luminance Y_f is combined with the retained visible chrominance components (Ivicb, Ivicr) and converted back to the RGB space to generate the final color fusion result. The entire network is trained in an end-to-end manner using a joint loss function that constrains infrared reconstruction, visible-detail reconstruction, luminance consistency, perceptual quality, and edge preservation.

3.2 Compact Multimodal Feature Representation

Based on the problem formulation in Section 3.1, the encoder aims to extract a compact shared feature from the two-channel input while preserving structure- and detail-sensitive responses. Under low-light degradation, visible features are easily affected by luminance attenuation, texture degradation, and noise disturbance, whereas infrared features remain more stable in structural perception. Therefore, the encoder adopts depthwise separable convolution (DSC) and efficient channel attention (ECA). DSC is used to reduce redundant spatial feature extraction, while ECA recalibrates channel responses to emphasize informative structure and detail features.

As described in Section 3.1, the encoder input is the two-channel tensor X_{in}. A 3×3 standard convolution is first used to expand the channel dimension to 64. Then, three stacked DSC modules are employed to progressively extract compact multimodal features. Finally, ECA is introduced at the end of the encoder to strengthen discriminative channel responses. The resulting shared feature F is used as the unified input for SCAM and the multi-branch decoders. The architecture of the compact encoder is shown in Fig. 3.

(1) Depthwise Separable Convolution (DSC)

Each DSC block decomposes a standard convolution into a depthwise convolution and a pointwise 1×1 convolution:

$$DSC(X) = Conv_{pw}(Conv_{dw}(X)) \tag{5}$$

where $Conv_{dw}$ denotes channel-wise spatial convolution, and $Conv_{pw}$ denotes pointwise convolution for cross-channel aggregation. The depthwise convolution captures spatial responses within each channel, while the pointwise convolution integrates information across channels. Each DSC block is combined with Batch Normalization and LeakyReLU with $\alpha = 0.2$ to improve training stability.

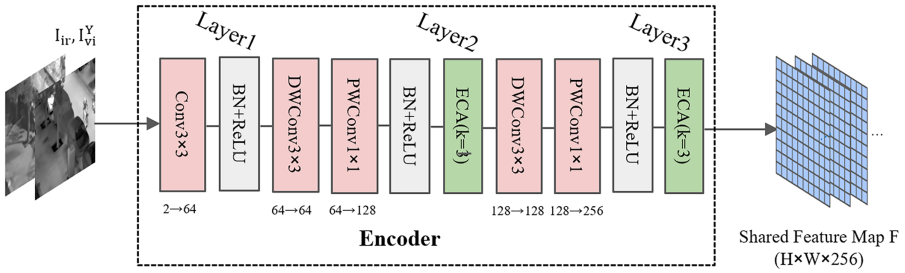


Figure 3: Architecture of the proposed compact feature extraction module based on DSC and ECA.

To further illustrate its computational efficiency, the computational costs of standard convolution and DSC are expressed as:

$$\mathrm{Cost}_{std} = k^2 C_{in} C_{out},\ \mathrm{Cost}_{DSC} = k^2 C_{in} + C_{in} C_{out},\ \frac{\mathrm{Cost}_{DSC}}{\mathrm{Cost}_{std}} \approx 0.115, k = 3, C_{in} = C_{out} = 256 \tag{6}$$

where k is the kernel size, and C_{in} and C_{out} denote the numbers of input and output channels, respectively. This indicates that DSC requires only about 11.5% of the computation of a standard convolution under this setting. Thus, DSC reduces redundant feature extraction while maintaining sufficient representation capability for subsequent branch-wise modeling.

(2) Efficient Channel Attention (ECA)

Although DSC reduces computational redundancy, different feature channels contribute unequally to structure and detail representation under low-light degradation. Therefore, ECA is introduced to perform lightweight channel recalibration. Let U denote the feature map output from the convolutional layer. The feature map and its channel descriptor are defined as:

$$U \in \mathbb{R}^{H \times W \times C}, z \in \mathbb{R}^{C}, z_c = \frac{1}{HW} \sum_{i=1}^{H} \sum_{j=1}^{W} U_{i,j,c}, c = 1, \ldots, C \tag{7}$$

where H and W denote the spatial resolution, C is the number of channels, and z_c denotes the global average response of the c-th channel. Next, a one-dimensional convolution with kernel size $k = 3$ is applied along the channel dimension to model local dependencies among adjacent channels. The channel attention vector and the recalibrated feature map are obtained as:

$$s = \sigma(\mathrm{Conv1D}_{k=3}(z)), U_{ECA}(i, j, c) = U(i, j, c) \cdot s_c \tag{8}$$

where $s \in (0,1)^C$ denotes the channel attention weight vector, s_c is the attention weight of the c-th channel, $\sigma(\cdot)$ is the Sigmoid activation function, and U_{ECA} denotes the channel-recalibrated output feature map. Compared with traditional Squeeze-and-Excitation (SE) or non-local attention mechanisms, ECA discards fully connected layers and instead captures channel correlations through a lightweight one-dimensional convolution. This introduces only a small number of additional parameters. Placing ECA at the end of the encoder helps highlight structure- and detail-relevant channels while suppressing noise-affected responses, thereby improving discriminative shared features for subsequent decoupled modeling.

3.3 *Structure–Detail–Luminance Decoupled Representation*

After compact multimodal feature extraction, the shared feature F contains mixed infrared structural responses, visible texture cues, and luminance-related information. Under low-light degradation, directly using such shared features may weaken visible-detail responses and bias the fusion process toward dominant infrared structures. To alleviate this issue, a Structure-aware Channel Attention Mechanism (SCAM) is introduced to perform branch-wise functional recalibration of the shared feature.

Rather than enforcing strict semantic separation at the channel level, SCAM guides the shared representation toward three complementary functional branches: infrared structure representation, visible texture representation, and luminance representation. As illustrated in Fig. 4, SCAM first summarizes the shared feature F into a channel descriptor and then generates three branch-specific attention vectors to recalibrate the shared feature toward different functional responses. The shared feature F and its channel descriptor z are defined as:

$$F \in \mathbb{R}^{H\times W\times C}, z = \mathrm{GAP}(F), z \in \mathbb{R}^{C} \tag{9}$$

where H and W denote the spatial resolution, C denotes the number of feature channels, and GAP(·) denotes global average pooling. The descriptor z summarizes the global response of each channel in F.

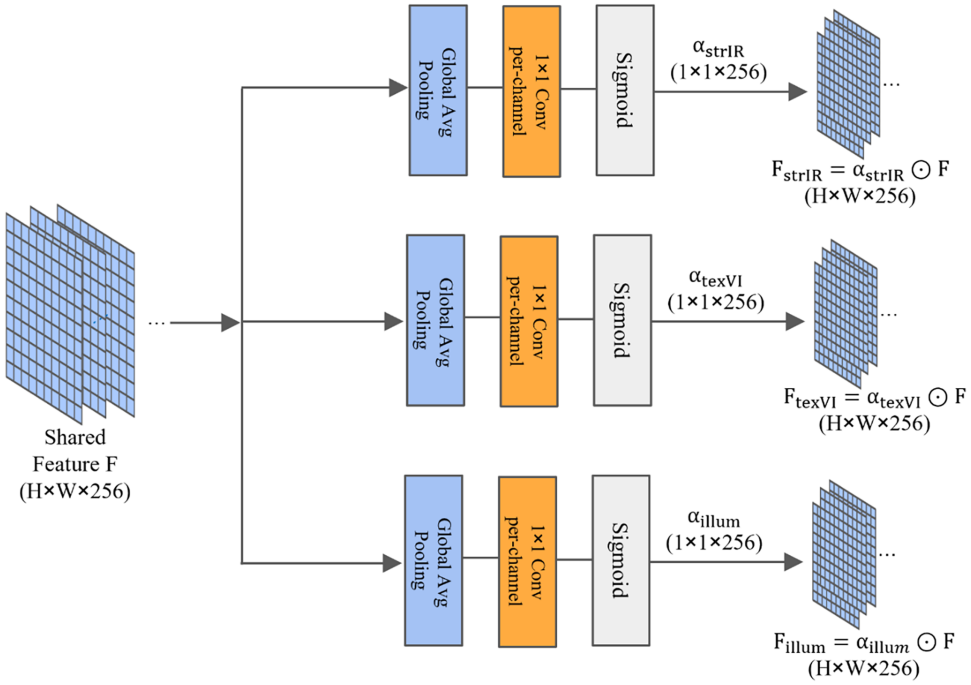


Figure 4: Branch-wise structure–detail–luminance representation through channel recalibration in SCAM.

Then, three independent branch-specific transformations followed by Sigmoid activation are applied to generate channel attention weights for the structure, texture, and luminance branches:

$$\alpha_{str}^{IR} = \sigma(W_{IR}z), \alpha_{tex}^{VI} = \sigma(W_{VI}z), \alpha_{illum} = \sigma(W_{L}z) \tag{10}$$

where α_{str}^{IR}, α_{tex}^{VI}, and α_{illum} denote the channel attention weights of the infrared structure branch, visible texture branch, and luminance branch, respectively. Here, $\sigma(\cdot)$ denotes the Sigmoid activation function, and W_{IR}, W_{VI}, and W_{L} are branch-specific transformation parameters.

Finally, the three branch features are obtained by applying the corresponding attention weights to the shared feature F:

$$F_{str}^{IR} = \alpha_{str}^{IR} \odot F, F_{tex}^{VI} = \alpha_{tex}^{VI} \odot F, F_{illum} = \alpha_{illum} \odot F \tag{11}$$

where $\odot$ denotes element-wise multiplication with broadcasting along the spatial dimensions. Through this branch-wise recalibration, each branch selectively emphasizes task-relevant channel responses from the shared feature space.

Specifically, the infrared structure branch F_{str}^{IR} emphasizes contour and edge-related responses for stable structural perception. The visible texture branch F_{tex}^{VI} enhances weakened texture and detail responses from the visible modality under low-light degradation. The luminance branch F_{illum} models illumination-related responses and provides luminance-aware priors for subsequent reconstruction. Through the following branch-specific decoders and reconstruction constraints, these three branches are encouraged to play complementary roles in the final fusion process.

3.4 Luminance-Guided Reconstruction

After branch-wise recalibration by SCAM, the three branch features F_{str}^{IR}, F_{tex}^{VI}, and F_{illum} are decoded through structure, detail, and luminance reconstruction branches. As illustrated in Fig. 5, the reconstruction stage consists of three corresponding decoders, namely the IR-Decoder, VI-Decoder, and Lum-Decoder. The decoding process is formulated as:

$$I_{IR}^{rec} = D_{IR}(F_{str}^{IR}), I_{VI}^{enh} = D_{VI}(F_{tex}^{VI}), M_{lum} = D_{lum}(F_{illum}) \in \mathbb{R}^{H\times W\times 1} \tag{12}$$

where D_{IR}, D_{VI}, D_{lum} denote the IR-Decoder, VI-Decoder, and Lum-Decoder, respectively. The infrared reconstruction I_{IR}^{rec} preserves stable contour and edge information from the infrared modality. The enhanced visible luminance I_{VI}^{enh} strengthens weakened texture and detail responses from the visible modality. The luminance estimation map M_{lum} models illumination-related responses and provides luminance-aware priors for the final reconstruction.

In the fusion stage, I_{IR}^{rec}, I_{VI}^{enh}, and M_{lum} are combined to generate the final fused luminance:

$$Y_f = \max(I_{IR}^{rec}, I_{VI}^{enh} \odot M_{lum}) \tag{13}$$

where $\odot$ denotes element-wise multiplication, and $\max(\cdot,\cdot)$ denotes a pixel-wise maximum operation. In this formulation, I_{IR}^{rec} provides stable structural responses, while I_{VI}^{enh}, M_{lum} represents luminance-modulated visible-detail information. The pixel-wise maximum operation selects the stronger response between infrared structure and luminance-enhanced visible details, thereby reducing the suppression of weakened visible information during reconstruction.

3.5 Loss Function Design

To jointly constrain structure preservation, visible-detail enhancement, luminance consistency, perceptual quality, and edge preservation under low-light conditions, multiple complementary loss terms are introduced during training. As illustrated in Fig. 6, the overall objective consists of infrared reconstruction loss, visible reconstruction loss, luminance consistency loss, perceptual loss, and edge-aware loss. The total loss is defined as:

$$\mathcal{L}_{total} = \lambda_{ir}\mathcal{L}_{ir} + \lambda_{vi}\mathcal{L}_{vi} + \lambda_{lum}\mathcal{L}_{lum} + \lambda_{per}\mathcal{L}_{per} + \lambda_{edge}\mathcal{L}_{edge} \tag{14}$$

where λ_{ir}, λ_{vi}, λ_{lum}, λ_{per}, and λ_{edge} denote the weights of the infrared reconstruction loss, visible reconstruction loss, luminance consistency loss, perceptual loss, and edge-aware loss, respectively. The numerical values and implementation details of these loss weights are summarized in Table 1. These weights were empirically set according to the relative magnitudes of different loss terms and the observed training stability, and were kept fixed in all experiments. Since the pixel-level reconstruction losses have relatively small raw numerical values, larger weights are assigned to $\mathcal{L}_{ir}$ and $\mathcal{L}_{vi}$ to balance their contributions with the luminance, perceptual, and edge-aware constraints.

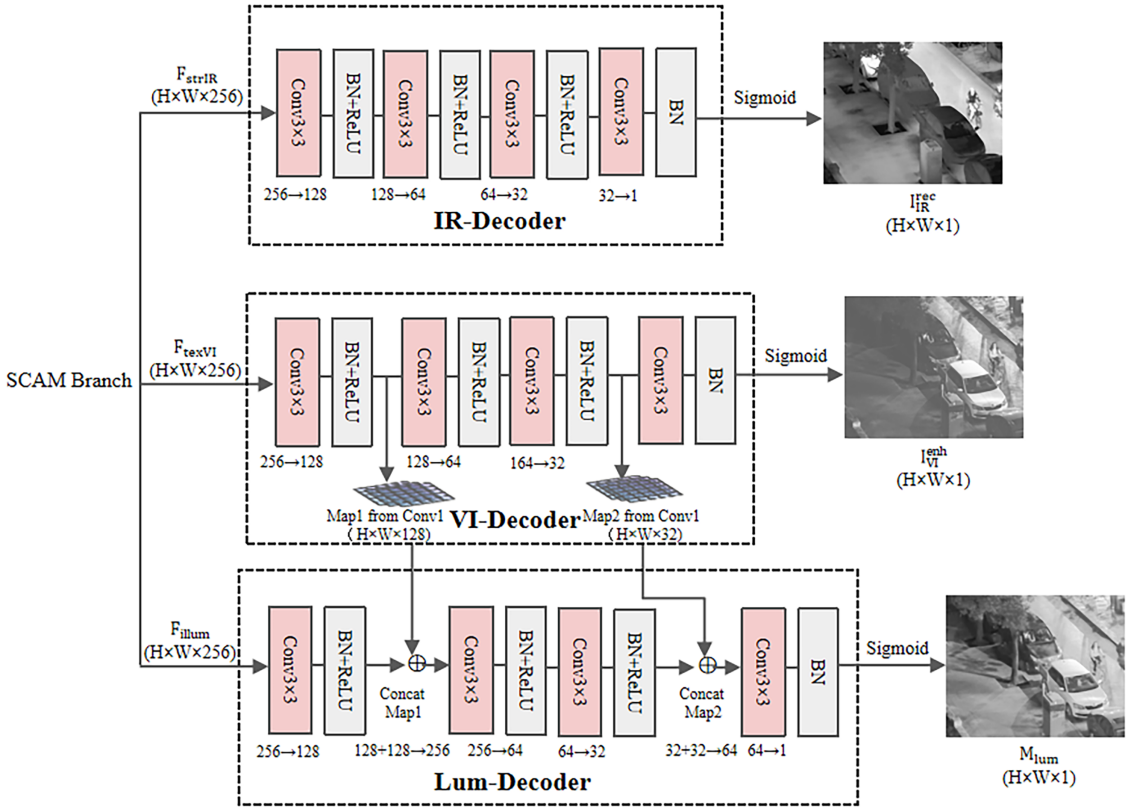


Figure 5: Luminance-guided reconstruction through branch-wise decoding of structure, visible detail, and illumination representation.

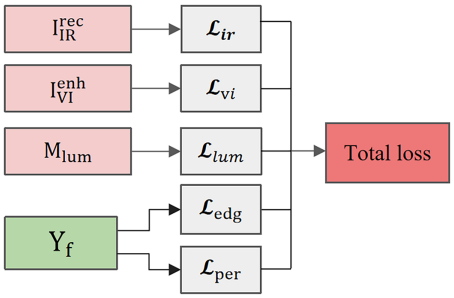


Figure 6: Collaborative multi-objective loss design for structure, detail, luminance, perceptual, and edge constraints.

Table 1: Weight values of different loss terms in Eq. (14).

Loss Term	Symbol	Weight Value	Implementation Detail
Infrared reconstruction loss	λ_{ir}	2000.0	Infrared reconstruction constraint
Visible reconstruction loss	λ_{vi}	1000.0	Visible luminance reconstruction constraint
Luminance consistency loss	λ_{lum}	8.0	Luminance consistency constraint
Perceptual loss	λ_{per}	40.0	VGG-based perceptual constraint
Edge-aware loss	λ_{edge}	0.3	Sobel-gradient-based edge constraint

The infrared reconstruction loss is used to preserve stable structural responses from the infrared modality:

$$\mathcal{L}_{ir} = \|I_{IR}^{rec} - I_{ir}\|_1 \tag{15}$$

where I_{IR}^{rec} denotes the reconstructed output of the infrared branch, and I_{ir} is the input infrared image. The visible reconstruction loss constrains the visible branch to retain texture and detail information from the visible luminance component:

$$\mathcal{L}_{vi} = \|I_{VI}^{enh} - I_{vi}^{Y}\|_1 \tag{16}$$

where I_{VI}^{enh} denotes the enhanced visible luminance image, and I_{vi}^{Y} is the luminance channel of the input visible image. The luminance consistency loss is imposed on the luminance estimation branch to guide the estimation of illumination-related responses:

$$\mathcal{L}_{lum} = \|M_{lum} - I_{vi}^{Y}\|_1 \tag{17}$$

where M_{lum} denotes the luminance estimation map. This loss encourages the luminance branch to model the global illumination distribution of the visible modality. Pixel-level constraints alone are insufficient to ensure perceptual consistency. Therefore, a VGG16-based [38] perceptual loss is introduced to constrain the fused luminance in the feature space:

$$\mathcal{L}_{per} = \sum_{l \in S} \|\Phi_l(Y_f) - \Phi_l(I_{ref})\|_2^2 \tag{18}$$

where Y_f is the final fused luminance, $\Phi_l(\cdot)$ denotes the feature map extracted from the l-th selected layer of the fixed VGG16 network, and S denotes the set of selected feature layers. In this implementation, features from $conv4_1$ and $conv5_1$ are used to capture both texture-level and high-level semantic information. Since no ground-truth fused image is available in the unsupervised fusion setting, I_{ref} is not used as a ground-truth supervision target. Instead, I_{ref} is defined as the visible luminance component I_{vi}^{Y} and is used only as a perceptual reference. Although I_{vi}^{Y} is degraded under low-light conditions, it still contains residual texture, local appearance, and semantic cues from the visible modality. Thus, the perceptual loss does not force the fused luminance to reproduce the degraded visible image, but encourages Y_f to retain effective visible information in the VGG feature space. Before being fed into VGG16, the single-channel luminance images are replicated into three channels and normalized according to the ImageNet setting. An edge-aware loss based on Sobel gradient alignment is further introduced to preserve salient contour and edge responses:

$$\mathcal{L}_{edge} = \|G(Y_f) - \max(G(I_{ir}), G(I_{vi}^{Y}))\|_1 \tag{19}$$

where $G(\cdot)$ denotes the Sobel gradient magnitude operator, and $\max(\cdot,\cdot)$ denotes a pixel-wise maximum operation. This loss encourages the fused luminance to preserve strong edge responses from both infrared

and visible modalities. Overall, the proposed multi-objective loss design jointly constrains structure, detail, luminance, perceptual, and edge information during reconstruction, as shown in Fig. 6.

4 Experiments and Results

This section presents experimental evaluations of the proposed DSCAttFuseNet on low-light infrared–visible image fusion tasks. We first introduce the datasets and experimental settings, followed by comparisons with representative and recent fusion methods in terms of both qualitative and quantitative results. Furthermore, ablation studies are carried out to analyze the contributions of key components and the effectiveness of the proposed decoupled reconstruction strategy.

4.1 Datasets and Experimental Settings

This study adopts the LLVIP dataset [39] as the primary platform for training and evaluation. LLVIP contains diverse low-light infrared–visible scenes with severe visible degradation. To further evaluate the cross-dataset generalization of the proposed model, independent tests are conducted on the public TNO dataset [40], which includes military surveillance scenarios and challenging infrared–visible image pairs.

For model training, the Adam optimizer [41] is employed with an initial learning rate of 1×10^{-4}, a batch size of 5, and a total of 2000 epochs. The model is trained using paired infrared–visible image samples from the LLVIP dataset, which cover diverse low-light scenarios with varying illumination conditions and scene complexities.

A relatively large number of epochs is adopted considering the small batch size and training stability requirements. The overall loss function jointly constrains infrared reconstruction, visible-detail enhancement, luminance consistency, perceptual consistency, and edge preservation during training. The loss-weight settings in Eq. (14) are reported in Table 1 and were kept fixed throughout all experiments. The perceptual loss is implemented based on the publicly available VGG16 [38] network pretrained on ImageNet, where fixed feature maps are used for perceptual constraint during training.

To evaluate the proposed framework under low-light fusion conditions, multiple objective metrics are used, including Entropy [42] for information content, AG [43] for detail response, VIF [44] for visual fidelity, Variance and SF [45] for contrast and texture representation, CC [46] for correlation consistency, and QCV [47] for comprehensive visual quality. In addition, $Q^{AB/F}$ [48], $L^{AB/F}$ [49], and $N^{AB/F}$ [50] are introduced to assess edge information transfer, information loss, and artificial noise introduced during fusion, respectively. Higher values are preferred for Entropy, AG, VIF, Variance, SF, CC, and $Q^{AB/F}$, while lower values indicate better performance for QCV, $L^{AB/F}$, and $N^{AB/F}$. Model parameters and inference statistics are also reported to compare model size and inference cost. For ablation studies, all variants are evaluated under the same training and testing settings for fair comparison.

4.2 Comparison with Existing Methods

To evaluate the effectiveness of the proposed DSCAttFuseNet, we conduct comparative experiments against several representative infrared–visible image fusion methods, including DenseFuse [28], MGFF [5], PIAFusion [51], PSFusion [52], RFN-Nest [31], SDNet [53], CDDFuse [32], SIFusion [37], and SwinFusion [54]. These comparison methods cover traditional fusion approaches, CNN-based models, and recent decomposition-, attention-, semantic-injection-, and Transformer-based frameworks.

4.2.1 Qualitative Results

As shown in Fig. 7, the visible images suffer from severe luminance attenuation and texture degradation under low-light conditions, whereas the infrared images preserve relatively stable structural information

but lack fine-grained texture details. MGFF and early CNN-based approaches such as DenseFuse, RFN-Nest, and SDNet can preserve partial structural information, but their fused results still show blurred textures, over-smoothed local details, and luminance imbalance in low-light regions. For example, weakened texture responses can be observed in the red-box regions of RFN-Nest and SDNet, while incomplete target contours appear in several green-box regions. Recent methods such as PIAFusion, PSFusion, CDDFuse, SIFusion, and SwinFusion improve illumination-aware fusion, feature decomposition, semantic injection, or global representation; however, texture smoothing and weakened edge details can still be observed in some highlighted regions. In comparison, the proposed DSCAttFuseNet preserves relatively clearer pedestrian contours and more stable local textures, while maintaining a more balanced luminance distribution in the highlighted target and texture regions.

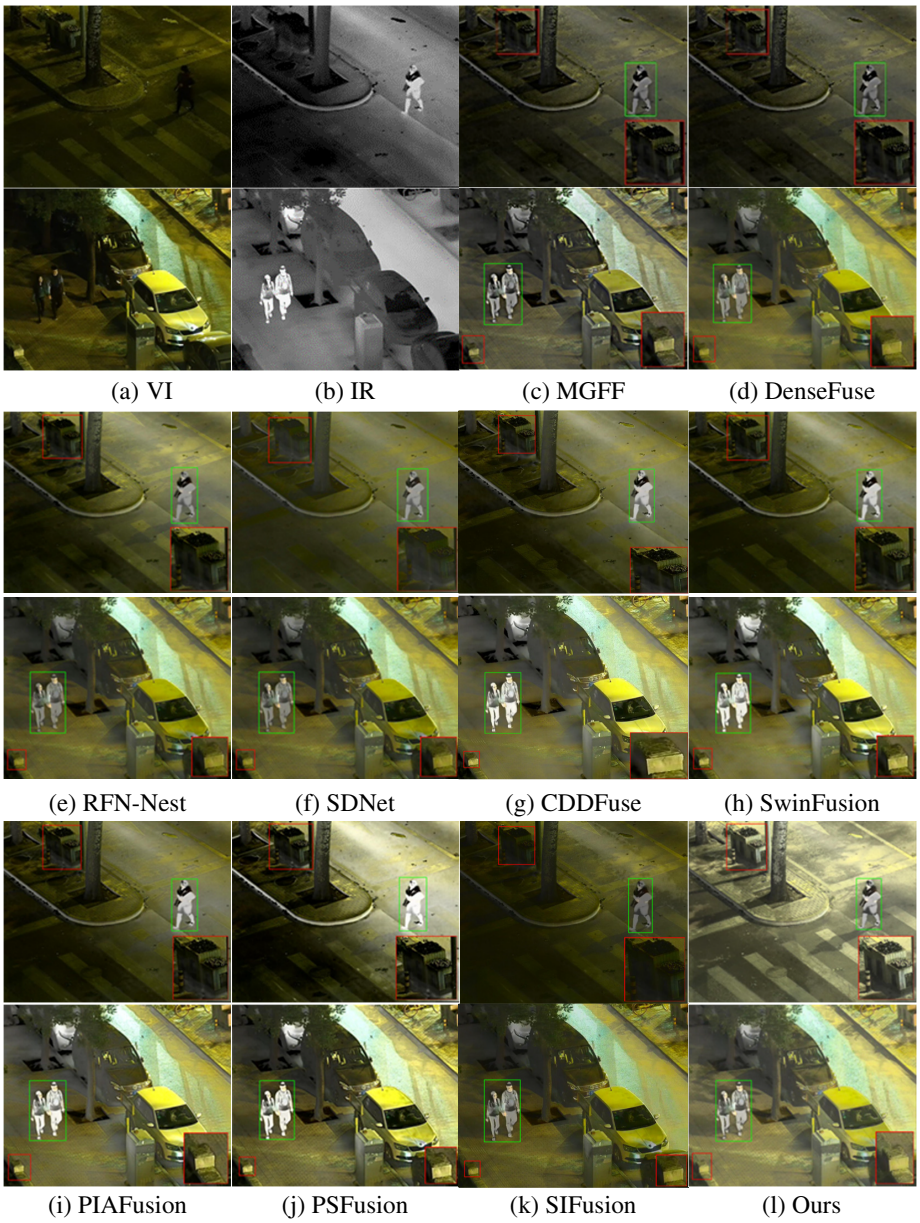


Figure 7: Qualitative comparison on the LLVIP dataset (#19308 and #210074). Red boxes highlight local texture regions, and green boxes indicate target regions.

4.2.2 Quantitative Results

In the quantitative experiments, 60 low-light infrared–visible image pairs are selected from the LLVIP test set. The evaluation includes information-based, gradient-based, fidelity-based, correlation-based, comprehensive quality, and edge-based metrics, namely Entropy, AG, VIF, Variance, SF, CC, QCV, $Q^{AB/F}$, $L^{AB/F}$, and $N^{AB/F}$. The average value of each metric over the 60 test image pairs is reported in Table 2.

Table 2: Average objective evaluation results of different method on 60 low-light infrared–visible image pairs from the LLVIP test set.

Methods	Entropy	AG	VIF	Variance	SF	CC	QCV	$Q^{AB/F}$	$L^{AB/F}$	$N^{AB/F}$
DenseFuse	6.923	2.429	0.791	36.944	8.277	0.755	0.540	0.844	0.154	0.002
MGFF	6.837	3.091	0.947	36.258	11.134	0.745	0.229	0.856	0.112	0.032
SIFusion	7.504	4.861	1.046	51.591	15.232	0.732	1.248	0.912	0.060	0.028
PIAFusion	7.073	3.560	0.968	44.779	12.545	0.728	1.167	0.923	0.056	0.021
PSFusion	7.486	4.589	0.989	66.733	16.703	0.719	1.658	0.872	0.063	0.065
RFN-Nest	6.689	1.931	0.705	37.408	5.704	0.750	1.911	0.639	0.357	0.004
SDNet	6.393	2.613	0.590	29.156	9.303	0.701	0.634	0.778	0.214	0.008
CDDFuse	7.488	4.913	1.031	50.936	15.384	0.736	1.274	0.914	0.061	0.025
SwinFusion	7.009	3.165	0.962	44.501	11.352	0.727	0.644	0.861	0.105	0.034
Ours	7.517	4.827	1.048	51.877	15.196	0.743	1.191	0.927	0.058	0.015

As shown in Table 2, DSCAttFuseNet achieves the highest Entropy, VIF, and $Q^{AB/F}$ values, indicating strong information preservation, visual information fidelity, and edge information transfer. It also obtains competitive AG, Variance, SF, and $L^{AB/F}$ values, showing its ability to maintain detail and contrast-related responses under low-light conditions. Meanwhile, the relatively low $N^{AB/F}$ suggests that the proposed method suppresses artificial edge responses to a certain extent during fusion. Although DSCAttFuseNet does not achieve the best performance on all metrics, especially on CC and QCV, it provides an overall balanced performance in information preservation, visible-detail enhancement, edge transfer, and artifact suppression.

4.3 Ablation Studies

To evaluate the contribution of the proposed modules and the effectiveness of the decoupled reconstruction strategy, ablation studies are conducted on 60 low-light infrared–visible image pairs from the LLVIP test set. All variants are trained and evaluated under the same dataset split, optimizer, and training protocol to ensure a fair comparison. In addition to removing individual modules, a shared fusion baseline is constructed by replacing the branch-wise structure–detail–luminance reconstruction with a single shared decoder. This baseline is used to examine whether explicit branch-wise modeling benefits low-light infrared–visible image fusion.

4.3.1 Ablation Analysis of ECA and SCAM

To investigate the effectiveness of ECA and SCAM, each module is individually removed, and the quantitative results are reported in Table 3. Since the proposed three-branch structure is designed for branch-wise functional modeling rather than three isolated modules, this subsection focuses on the effects of ECA and SCAM within the decoupled reconstruction framework. The contribution of the branch-wise design itself is further examined through a shared fusion baseline in the following subsection.

Table 3: Average ablation results of ECA and SCAM within the proposed decoupled reconstruction framework on 60 low-light image pairs from the LLVIP dataset.

Method	Entropy	AG	VIF	Variance	SF	CC	QCV
Ours	7.517	4.827	1.048	51.877	15.196	0.743	1.191
–ECA	6.804	1.874	0.776	44.675	9.284	0.702	3.442
–SCAM	6.863	2.015	0.812	45.962	9.623	0.706	3.521

When ECA is removed, Entropy, AG, and SF decrease noticeably, indicating weakened information preservation and detail representation during reconstruction. When SCAM is removed, most metrics also decrease, especially VIF and CC, suggesting that the absence of branch-wise recalibration weakens structural consistency and multimodal feature representation during fusion. For QCV, the ablated models show substantially larger values, indicating degraded comprehensive visual quality. In contrast, the complete model achieves a more balanced performance across the evaluated metrics.

Qualitative ablation comparisons are further presented in Fig. 8. Without ECA, the fused results exhibit weakened local textures and insufficient detail representation in dark regions. Without SCAM, the fused images show blurred target boundaries and less complete contour structures. In contrast, the complete model preserves relatively clearer contours, more stable local textures, and more balanced luminance in low-light regions.

(a) VI (b) IR (c) w/o ECA (d) w/o SCAM (e) Ours

Figure 8: Qualitative comparison results of the ablation study. (**a**) Visible image (VI), (**b**) infrared image (IR), (**c**) w/o ECA, (**d**) w/o SCAM, and (**e**) the complete model. Red boxes indicate local texture regions, while green boxes highlight target contours.

4.3.2 Comparison with a Shared Fusion Baseline

To further evaluate the role of the branch-wise decoupled reconstruction strategy, a shared fusion baseline is constructed for comparison. In this baseline, the input, compact encoder, optimizer, and training settings are kept the same as those of the complete model, while the structure–detail–luminance branches are replaced with a single shared decoder. Therefore, this baseline does not explicitly model structure, detail, and luminance-related responses through separate branches. Since this modification changes both the decoder structure and the number of parameters, the shared fusion baseline is not strictly parameter-matched with the complete model. Thus, this comparison is mainly used to examine the effect of branch-wise modeling, and the parameter numbers of different variants are reported in Table 4.

As shown in Table 4, the shared fusion baseline has fewer parameters because it uses a single shared decoder instead of branch-wise decoders. However, its EN, AG, VIF, and CC values are lower than those of

the complete model, and its QCV value is higher. Introducing the branch-wise structure–detail–luminance design improves the results over the shared baseline, indicating that explicit modeling of structure, visible detail, and luminance-related responses is beneficial for low-light fusion. Adding SCAM further improves the performance with only a small increase in parameters from 1.783 to 1.796 M. Since the shared fusion baseline is not strictly parameter-matched with the complete model, this comparison is mainly used to analyze the contribution of branch-wise decoupled modeling rather than as an equal-parameter comparison.

Table 4: Average quantitative comparison with the shared fusion baseline on 60 low-light image pairs from the LLVIP dataset.

Method	Branches	SCAM	Params (M)	EN	AG	VIF	CC	QCV
Shared fusion baseline	×	×	1.124	6.704	1.712	0.766	0.698	3.607
Ours w/o SCAM	√	×	1.783	6.863	2.015	0.812	0.706	3.521
Ours	√	√	1.796	7.517	4.827	1.048	0.743	1.191

4.3.3 Analysis of DSC-Based Encoder

To evaluate the effect of the DSC-based encoder, depthwise separable convolution (DSC) is replaced with standard convolution (StdConv), while the remaining network structure and training settings are kept unchanged. The comparison is conducted using the models trained on the LLVIP dataset and evaluated on the same 60 LLVIP test image pairs used in the main quantitative comparison. The evaluated metrics include Entropy, VIF, CC, the number of parameters, processing time, and throughput. The quantitative results are summarized in Table 5.

Table 5: Quantitative comparison between StdConv and DSC-based encoders under the same evaluation settings.

Method	Entropy	VIF	CC	Params (M)	Processing Time (ms)	Throughput (FPS)
StdConv	7.426	1.124	0.736	2.644	2479.6	0.403
Ours (DSC)	7.517	1.048	0.743	1.796	2311.5	0.433

As shown in Table 5, StdConv obtains a slightly higher VIF value, whereas the DSC-based encoder achieves higher Entropy and comparable CC values with fewer parameters. Specifically, DSC reduces the number of parameters from 2.644 to 1.796 M and decreases the processing time from 2479.6 to 2311.5 ms under the same evaluation settings. These results indicate that DSC can reduce convolutional redundancy while maintaining comparable overall fusion quality.

Overall, the DSC-based encoder reduces the parameter scale and slightly improves inference efficiency while maintaining comparable fusion performance. Further optimization is still needed for real-time scenarios.

4.3.4 Sensitivity Analysis of Loss Weights

In addition to the architectural ablation studies, a sensitivity analysis is conducted to examine the influence of the loss-weight settings in Eq. (14). Since the infrared and visible reconstruction losses provide the main pixel-level reconstruction supervision, λ_{ir} and λ_{vi} are kept unchanged to maintain the basic reconstruction objective. The auxiliary weights λ_{lum}, λ_{per}, and λ_{edge} are scaled by 0.5× and 2× based on the default setting. All variants are trained and evaluated under the same experimental protocol, and the results are reported in Table 6.

Table 6: Sensitivity analysis of auxiliary loss weights on the LLVIP dataset.

Setting	λ_{lum}	λ_{per}	λ_{edge}	EN	AG	VIF	CC	QCV
0.5× auxiliary weights	4.0	20.0	0.15	7.496	4.681	1.022	0.746	1.164
Default	8.0	40.0	0.3	7.517	4.827	1.048	0.743	1.191
2× auxiliary weights	16.0	80.0	0.6	7.509	4.914	1.041	0.737	1.326

As shown in Table 6, the average metric values vary only slightly under different auxiliary loss-weight settings. The default setting achieves the highest EN and VIF and maintains balanced CC and QCV values, whereas the 2× setting slightly improves AG but increases QCV. This suggests that stronger auxiliary constraints may enhance local details but may not always improve comprehensive visual quality. Since the luminance, perceptual, and edge-aware losses respectively guide illumination estimation, feature-level consistency, and contour preservation, the limited variation among the evaluated settings indicates that the proposed method is not highly sensitive to moderate changes in these auxiliary constraints.

4.4 Generalization Evaluation

To evaluate the cross-dataset generalization capability of the proposed DSCAttFuseNet, experiments are further conducted on the TNO dataset, which is not involved in training and is used only for testing. Since its imaging scenarios, including military surveillance and long-range targets, differ significantly from the urban night scenes in LLVIP, it provides a challenging evaluation setting for cross-dataset analysis. For evaluation, 42 representative infrared–visible image pairs are selected from the TNO dataset, covering long-range targets, complex backgrounds, and dark or low-contrast scenes. All experiments are performed by directly applying the model trained on LLVIP without additional fine-tuning. The evaluation metrics include Entropy, AG, VIF, Variance, SF, CC, QCV, $Q^{AB/F}$, $L^{AB/F}$ and $N^{AB/F}$. The average value of each metric over the 42 test image pairs is reported.

4.4.1 Cross-Dataset Quantitative Evaluation on the TNO Dataset

Cross-dataset quantitative results are reported in Table 7. On the unseen TNO dataset, DSCAttFuseNet achieves the highest Entropy, AG, Variance, and SF values, as well as the lowest $L^{AB/F}$ value, indicating strong information preservation, detail response, contrast-related representation, and reduced edge information loss without additional fine-tuning. It also obtains competitive $Q^{AB/F}$, suggesting effective source edge information transfer across datasets. However, the proposed method is not superior on all metrics. CDDFuse achieves a slightly higher $Q^{AB/F}$, while RFN-Nest obtains the lowest NAB/F. Nevertheless, the low NAB/F of RFN-Nest is accompanied by a much lower QAB/F and higher LAB/F, suggesting substantial source-edge information loss. In addition, several methods perform better in VIF, CC, and QCV, showing that DSCAttFuseNet still has limitations in visual fidelity, correlation consistency, and QCV-based comprehensive quality evaluation under cross-dataset testing. Overall, the TNO results demonstrate competitive generalization in information and detail preservation, while cross-dataset visual quality consistency and artificial edge-response control remain areas for further improvement.

4.4.2 Qualitative Results

Representative fusion results on the TNO dataset are shown in Fig. 9. In complex long-range and low-light scenes, MGFF, DenseFuse, RFN-Nest, and SDNet preserve partial structural information, but their results still show blurred contours, weak local textures, and insufficient brightness in several dark regions.

Recent methods such as PIAFusion, PSFusion, CDDFuse, SIFusion, and SwinFusion improve illumination-aware fusion, feature decomposition, semantic injection, or global representation to different degrees; however, texture smoothing and weakened edge details can still be observed in some highlighted regions.

Table 7: Average cross-dataset quantitative results on 42 image pairs from the TNO dataset.

Method	Entropy	AG	VIF	Variance	SF	CC	QCV	$Q^{AB/F}$	$L^{AB/F}$	$N^{AB/F}$
DenseFuse	6.819	3.560	0.667	34.825	8.985	0.531	15.681	0.786	0.198	0.016
MGFF	6.726	4.704	0.658	31.881	11.896	0.518	13.852	0.812	0.141	0.047
PIAFusion	6.906	4.397	0.777	41.354	11.254	0.462	5.854	0.884	0.083	0.033
PSFusion	7.264	5.882	0.737	48.015	15.121	0.500	11.861	0.842	0.071	0.087
RFN-Nest	6.963	2.669	0.563	36.897	5.874	0.524	20.756	0.646	0.341	0.013
SDNet	6.352	3.699	0.528	26.908	9.420	0.460	15.733	0.721	0.255	0.024
CDDFuse	7.321	5.984	0.792	47.562	14.876	0.503	12.486	0.895	0.054	0.051
SwinFusion	6.907	4.219	0.760	39.647	10.753	0.474	7.105	0.862	0.108	0.030
SIFusion	7.012	5.126	0.724	42.168	12.874	0.478	8.247	0.867	0.090	0.043
Ours	7.554	6.393	0.639	52.851	16.191	0.468	19.861	0.889	0.046	0.065

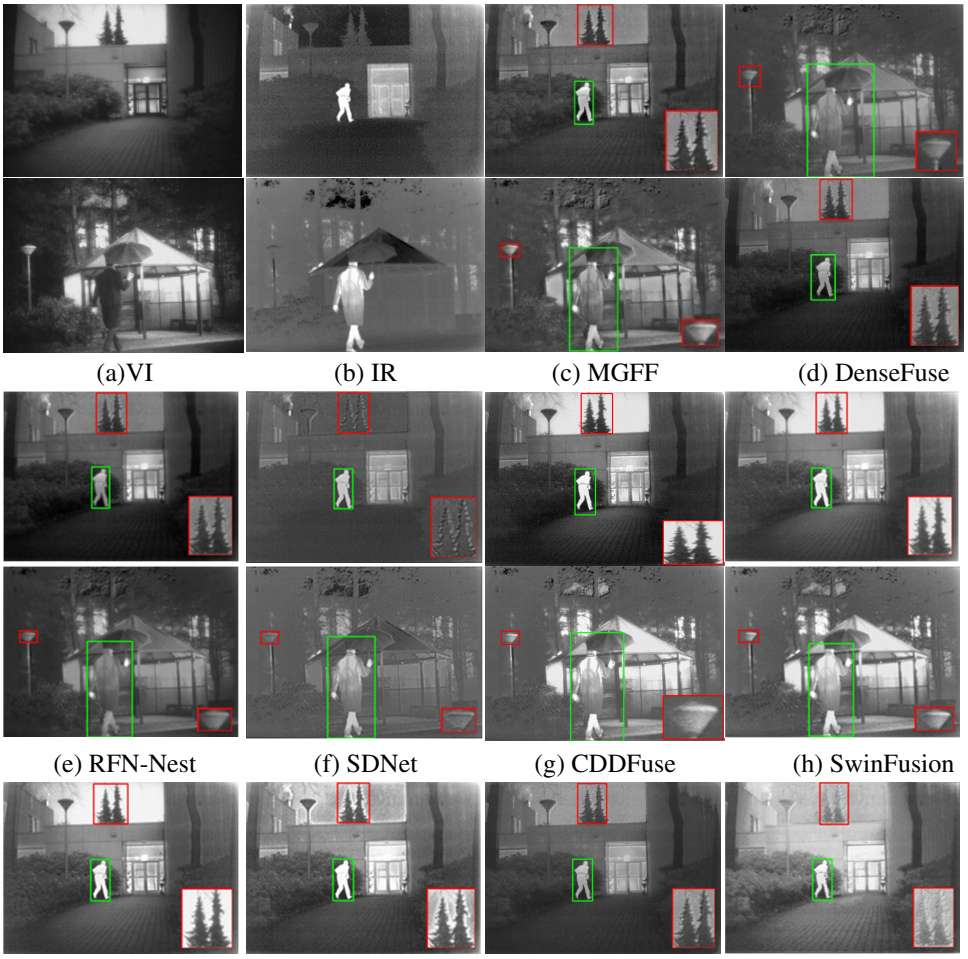
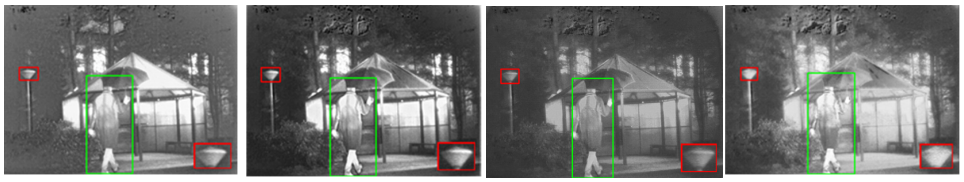


Figure 9: (Continued)



(i) PIAFusion (j) PSFusion (k) SIFusion (l) Ours

Figure 9: Qualitative comparison results on the TNO dataset: (**a**) VI, (**b**) IR, (**c**) MGFF, (**d**) DenseFuse, (**e**) RFN-Nest, (**f**) SDNet, (**g**) CDDFuse, (**h**) SwinFusion, (**i**) PIAFusion, (**j**) PSFusion, (**k**) SIFusion, and (**l**) ours. Red boxes highlight local texture regions, while green boxes indicate long-range target areas.

In comparison, DSCAttFuseNet shows relatively clearer local details, more balanced brightness, and better target visibility in the selected highlighted regions. These observations are consistent with the quantitative results in Table 7, where the proposed method performs strongly in Entropy, AG, Variance, and SF, while its visual fidelity and QCV-based quality still require further improvement under cross-dataset testing.

5 Conclusion

This study addresses low-light infrared–visible image fusion by focusing on visible-information degradation and modality imbalance. A branch-wise structure–detail–luminance modeling strategy is proposed to preserve infrared structural cues, enhance weakened visible details, and maintain balanced luminance representation during reconstruction. Experimental results on the LLVIP dataset show that DSCAttFuseNet achieves competitive performance compared with representative and recent fusion methods, particularly in information preservation, visual information fidelity, and edge information transfer. The shared fusion baseline further indicates that the performance improvement is closely related to the decoupled reconstruction design, while the sensitivity analysis shows that the method is not highly sensitive to moderate changes in auxiliary loss weights.

Cross-dataset evaluation on TNO demonstrates that the proposed method maintains effective information and detail preservation on unseen scenes, suggesting its generalization potential beyond the low-light LLVIP setting. However, the method still has limitations. Its QCV and NAB/F results on the TNO dataset are less favorable than several competing methods, indicating that cross-dataset visual quality consistency and artificial edge-response control require further improvement. In addition, a more fine-grained analysis of individual loss terms and the luminance-guided reconstruction operation would help further clarify their independent contributions. Future work will focus on improving visual quality consistency across datasets, introducing chrominance-aware refinement, optimizing implementation efficiency, and extending the evaluation to more challenging low-light scenarios and downstream vision tasks such as object detection and tracking.

Acknowledgement: The authors would like to acknowledge Universiti Teknologi MARA (UiTM) for the support provided in relation to this research. The authors also sincerely thank their supervisors for their guidance and encouragement throughout this work.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: The authors confirm their contributions to the paper as follows: study conception and overall framework design: Kezhen Xie; methodological development, algorithm implementation, experimental design, data analysis, and writing—original draft preparation: Kezhen Xie. Syed Mohd Zahid Syed Zainal Ariffin provided supervision, guidance on the research direction, and constructive suggestions on the methodological design at the early stage of this work. He also contributed to the interpretation of experimental results and the critical review and revision

of the manuscript. Muhammad Izzad Ramli provided supervisory support, academic advice, and manuscript review. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The datasets used in this study are publicly available from their original sources. The data supporting the findings of this study are available from the corresponding author upon reasonable request.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Dong X, Cappuccio ML. Applications of computer vision in autonomous vehicles: methods, challenges and future directions. arXiv:2311.09093. 2023.
2. Ma J, Ma Y, Li C. Infrared and visible image fusion methods and applications: a survey. Inf Fusion. 2019;45(4):153–78. doi:10.1016/j.inffus.2018.02.004.
3. Ma J, Zhang H, Shao Z, Liang P, Xu H. GANMcC: a generative adversarial network with multiclassification constraints for infrared and visible image fusion. IEEE Trans Instrum Meas. 2021;70:5005014. doi:10.1109/TIM.2020.3038013.
4. Zhang X, Demiris Y. Visible and infrared image fusion using deep learning. IEEE Trans Pattern Anal Mach Intell. 2023;45(8):10535–54. doi:10.1109/tpami.2023.3261282.
5. Bavirisetti DP, Xiao G, Zhao J, Dhuli R, Liu G. Multi-scale guided image and video fusion: a fast and efficient approach. Circuits Syst Signal Process. 2019;38(12):5576–605. doi:10.1007/s00034-019-01131-z.
6. Zhan L, Zhuang Y, Huang L. Infrared and visible images fusion method based on discrete wavelet transform. J Comput. 2017;28(2):57–71. doi:10.1117/12.2216054.
7. Fu Z, Wang X, Xu J, Zhou N, Zhao Y. Infrared and visible images fusion based on RPCA and NSCT. Infrared Phys Technol. 2016;77(1):114–23. doi:10.1016/j.infrared.2016.05.012.
8. Paramanandham N, Rajendiran K. Infrared and visible image fusion using discrete cosine transform and swarm intelligence for surveillance applications. Infrared Phys Technol. 2018;88(3):13–22. doi:10.1016/j.infrared.2017.11.006.
9. Yang Y, Zhang Y, Huang S, Zuo Y, Sun J. Infrared and visible image fusion using visual saliency sparse representation and detail injection model. IEEE Trans Instrum Meas. 2021;70:5001715. doi:10.1109/TIM.2020.3011766.
10. Xie K, Ariffin SMZSZ, Ramli MI. A mask-guided latent low-rank representation method for infrared and visible image fusion. Comput Mater Contin. 2025;84(1):997–1011. doi:10.32604/cmc.2025.063469.
11. Singh S, Singh H, Bueno G, Deniz O, Singh S, Monga H, et al. A review of image fusion: methods, applications and performance metrics. Digit Signal Process. 2023;137(6):104020. doi:10.1016/j.dsp.2023.104020.
12. Kalamkar S, Geetha Mary A. Multimodal image fusion: a systematic review. Decis Anal J. 2023;9(3):100327. doi:10.1016/j.dajour.2023.100327.
13. Tan MC, Nie RC, Zhang GC, Zhang Y. Deep learning-based infrared and visible image fusion: a survey. J Yunnan Univ Nat Sci Ed. 2023;45(2):326–43. (In Chinese). doi:10.7540/j.ynu.20220646.
14. Shopovska I, Jovanov L, Philips W. Deep visible and thermal image fusion for enhanced pedestrian visibility. Sensors. 2019;19(17):3727. doi:10.3390/s19173727.
15. Li H, Wu XJ, Kittler J. Infrared and visible image fusion using a deep learning framework. In: 2018 24th International Conference on Pattern Recognition (ICPR); 2018 Aug 20–24; Beijing, China. p. 2705–10. doi:10.1109/ICPR.2018.8546006.
16. Chen J, Li X, Luo L, Mei X, Ma J. Infrared and visible image fusion based on target-enhanced multiscale transform decomposition. Inf Sci. 2020;508(4):64–78. doi:10.1016/j.ins.2019.08.066.
17. Li X, Tan H, Zhou F, Wang G, Li X. Infrared and visible image fusion based on domain transform filtering and sparse representation. Infrared Phys Technol. 2023;131(2):104701. doi:10.1016/j.infrared.2023.104701.
18. Li H, Wu XJ. Infrared and visible image fusion using latent low-rank representation. arXiv:1804.08992. 2018.

19. Xiang W, Shen J, Zhang L, Zhang Y. Infrared and visual image fusion based on a local-extrema-driven image filter. Sensors. 2024;24(7):2271. doi:10.3390/s24072271.
20. Sappa A, Carvajal J, Aguilera C, Oliveira M, Romero D, Vintimilla B. Wavelet-based visible and infrared image fusion: a comparative study. Sensors. 2016;16(6):861. doi:10.3390/s16060861.
21. Li H, Qiu H, Yu Z, Zhang Y. Infrared and visible image fusion scheme based on NSCT and low-level visual features. Infrared Phys Technol. 2016;76(8):174–84. doi:10.1016/j.infrared.2016.02.005.
22. Singh S, Singh H, Gehlot A, Kaur J, Gagandeep. IR and visible image fusion using DWT and bilateral filter. Microsyst Technol. 2023;29(4):457–67. doi:10.1007/s00542-022-05315-7.
23. Fan Z, Guan N, Wang Z, Su L, Wu J, Sun Q. Unified framework based on multiscale transform and feature learning for infrared and visible image fusion. Opt Eng. 2021;60(12):123102. doi:10.1117/1.oe.60.12.123102.
24. Zhang Q, Liu Y, Blum RS, Han J, Tao D. Sparse representation based multi-sensor image fusion for multi-focus and multi-modality images: a review. Inf Fusion. 2018;40(3):57–75. doi:10.1016/j.inffus.2017.05.006.
25. Liu Y, Chen X, Ward RK, Wang ZJ. Image fusion with convolutional sparse representation. IEEE Signal Process Lett. 2016;23(12):1882–6. doi:10.1109/lsp.2016.2618776.
26. Li S, Kang X, Hu J. Image fusion with guided filtering. IEEE Trans Image Process. 2013;22(7):2864–75. doi:10.1109/tip.2013.2244222.
27. Li S, Kang X, Fang L, Hu J, Yin H. Pixel-level image fusion: a survey of the state of the art. Inf Fusion. 2017;33(6583):100–12. doi:10.1016/j.inffus.2016.05.004.
28. Li H, Wu XJ. DenseFuse: a fusion approach to infrared and visible images. IEEE Trans Image Process. 2019;28(5):2614–23. doi:10.1109/tip.2018.2887342.
29. Ma J, Yu W, Liang P, Li C, Jiang J. FusionGAN: a generative adversarial network for infrared and visible image fusion. Inf Fusion. 2019;48(4):11–26. doi:10.1016/j.inffus.2018.09.004.
30. Zhang H, Xu H, Xiao Y, Guo X, Ma J. Rethinking the image fusion: a fast unified image fusion network based on proportional maintenance of gradient and intensity. In: Proceedings of the AAAI Conference on Artificial Intelligence. Washington, DC, USA: AAAI Press; 2020. p. 12797–804. doi:10.1609/aaai.v34i07.6975.
31. Li H, Wu XJ, Kittler J. RFN-Nest: an end-to-end residual fusion network for infrared and visible images. Inf Fusion. 2021;73(9):72–86. doi:10.1016/j.inffus.2021.02.023.
32. Zhao Z, Bai H, Zhang J, Zhang Y, Xu S, Lin Z, et al. CDDFuse: correlation-driven dual-branch feature decomposition for multi-modality image fusion. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023 Jun 17–24; Vancouver, BC, Canada. p. 5906–16. doi:10.1109/CVPR52729.2023.00572.
33. Panda MK, Parida P, Rout DK. A weight induced contrast map for infrared and visible image fusion. Comput Electr Eng. 2024;117(18):109256. doi:10.1016/j.compeleceng.2024.109256.
34. Parida P, Panda MK, Rout DK, Panda SK. Infrared and visible image fusion using quantum computing induced edge preserving filter. Image Vis Comput. 2025;153(2):105344. doi:10.1016/j.imavis.2024.105344.
35. Tang L, Yuan J, Ma J. Image fusion in the loop of high-level vision tasks: a semantic-aware real-time infrared and visible image fusion network. Inf Fusion. 2022;82(10):28–42. doi:10.1016/j.inffus.2021.12.004.
36. Wang Q, Li Z, Zhang S, Luo Y, Chen W, Wang T, et al. SMAE-Fusion: integrating saliency-aware masked autoencoder with hybrid attention transformer for infrared-visible image fusion. Inf Fusion. 2025;117(1):102841. doi:10.1016/j.inffus.2024.102841.
37. Qian S, Yang L, Xue Y, Li P. SIFusion: lightweight infrared and visible image fusion based on semantic injection. PLoS One. 2024;19(11):e0307236. doi:10.1371/journal.pone.0307236.
38. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556. 2014.
39. Jia X, Zhu C, Li M, Tang W, Zhou W. LLVIP: a visible-infrared paired dataset for low-light vision. In: 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW); 2021 Oct 11–17; Montreal, QC, Canada. p. 3489–97. doi:10.1109/iccvw54120.2021.00389.
40. Toet A. The TNO multiband image data collection. Data Brief. 2017;15(2):249–51. doi:10.1016/j.dib.2017.09.038.
41. Kingma DP, Ba J. Adam: a method for stochastic optimization. arXiv:1412.6980. 2014.

42. Zhang Y. Methods for image fusion quality assessment—a review, comparison and analysis. Int Arch Photogramm Remote Sens Spat Inf Sci. 2008;37(PART B7):1101–10.
43. Hossny M, Nahavandi S, Creighton D, Bhatti A. Image fusion performance metric based on mutual information and entropy driven quadtree decomposition. Electron Lett. 2010;46(18):1266–8. doi:10.1049/el.2010.1778.
44. Tang L, Xiang X, Zhang H, Gong M, Ma J. DIVFusion: darkness-free infrared and visible image fusion. Inf Fusion. 2023;91(1):477–93. doi:10.1016/j.inffus.2022.10.034.
45. Luo Y, Luo Z. Infrared and visible image fusion: methods, datasets, applications, and prospects. Appl Sci. 2023;13(19):10891. doi:10.3390/app131910891.
46. Ma W, Wang K, Li J, Zhai Y. Infrared and visible image fusion technology and application: a review. Sensors. 2023;23(2):599. doi:10.3390/s23020599.
47. Zhang X, Ye P, Xiao G. VIFB: a visible and infrared image fusion benchmark. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); 2020 Jun 14–19; Seattle, WA, USA. p. 468–78. doi:10.1109/cvprw50498.2020.00060.
48. Shreyamsha Kumar BK. Image fusion based on pixel significance using cross bilateral filter. Signal Image Video Process. 2015;9(5):1193–204. doi:10.1007/s11760-013-0556-9.
49. Fan X, Kong F, Shi H, Guo Y. Infrared and visible image fusion algorithm based on NSCT and improved FT saliency detection. Sci Rep. 2026;16(1):7144. doi:10.1038/s41598-026-37670-0.
50. Wang R, Zhou Z, Li S, Zhang Z. Advances and challenges in infrared-visible image fusion: a comprehensive review of techniques and applications. Artif Intell Rev. 2025;59(1):18. doi:10.1007/s10462-025-11426-0.
51. Tang L, Yuan J, Zhang H, Jiang X, Ma J. PIAFusion: a progressive infrared and visible image fusion network based on illumination aware. Inf Fusion. 2022;83–84(5):79–92. doi:10.1016/j.inffus.2022.03.007.
52. Tang L, Zhang H, Xu H, Ma J. Rethinking the necessity of image fusion in high-level vision tasks: a practical infrared and visible image fusion network based on progressive semantic injection and scene fidelity. Inf Fusion. 2023;99(1):101870. doi:10.1016/j.inffus.2023.101870.
53. Zhang H, Ma J. SDNet: a versatile squeeze-and-decomposition network for real-time image fusion. Int J Comput Vis. 2021;129(10):2761–85. doi:10.1007/s11263-021-01501-8.
54. Ma J, Tang L, Fan F, Huang J, Mei X, Ma Y. SwinFusion: cross-domain long-range learning for general image fusion via swin transformer. IEEE. 2022;9(7):1200–17. doi:10.1109/jas.2022.105686.