Infrared-Visible Image Fusion Method Based on Differential Feature Enhancement and Cross-Modal Attention

Huang Zhang[1] ,Lina Xu[1] , Qing Zhou[1] ,Tiyou Zhou[2] ,Siyu Liu[1] ,Xincai Chang[1] ,Hao Li[1]

[1]Hubei Subsurface Multi-scale Imaging Key Laboratory, School of Geophysics and Geomatics,
China University of Geosciences, Wuhan, 430074, China
[2]State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, 430079, ·China

Keywords: Infrared and Visible Image Fusion, Deep Learning, Differential Feature Enhancement, Cross-Modal Fusion, Transformer.

Abstract:
To address the issues of insufficient inter-modal information interaction in dual-stream encoders, inadequate deep feature fusion, and loss function failure in extreme environments like strong light in existing autoencoder-based infrared and visible image fusion methods, this paper proposes a fusion method called DFECF (Differential Feature Enhancement and Cross-Modal Information Fusion). This method adopts an end-to-end architecture consisting of "Dual-Stream Encoder - Cross-Modal Fusion - Transformer Global Perception - Decoder Reconstruction". A differential feature enhancement module is embedded in the encoder to achieve feature enhancement through inter-modal difference information and an attention mechanism. A cross-modal feature fusion module is designed to complete the adaptive integration of deep features. A Transformer module is introduced to supplement the global feature perception capability. Additionally, a joint loss function of "gradient loss + pixel loss + auxiliary loss" is constructed to improve robustness in extreme environments. Experiments are carried out on the TNO and MSRS datasets. The results show that DFECF achieves superior performance with MI=3.85 and Qabf=0.68, outperforming state-of-the-art methods by 15.2% and 8.7% respectively. On the TNO dataset, DFECF also outperforms 8 comparison methods such as FusionGAN and DenseFuse in both subjective visual effects and objective metrics. It can still generate fusion images with clear textures in areas affected by strong light interference, demonstrating good practicability and generalization.

1. Introduction

In the current context of deep integration of computer vision and remote sensing technology, single-modal images can hardly meet the information perception requirements in complex scenarios (Yu et al., 2011). Taking infrared and visible images as an example, they exhibit significant information complementarity due to differences in imaging principles: infrared images, relying on thermal radiation detection mechanisms, can penetrate harsh environments such as fog, dust, and low light, accurately capture thermal target information in scenes, and achieve all-time and all-weather operation (Ma et al., 2018; Zhi et al., 2019); while visible images, based on light reflection imaging, can present rich texture details and color information, conforming to human visual perception habits, but are extremely susceptible to environmental factors such as light intensity and weather conditions, and their quality degrades severely in scenarios such as low illumination and heavy smoke (Jin et al., 2017; Toet, 2014). This complementary characteristic of infrared's strength in target detection and visible light's strength in detail presentation makes infrared and visible image fusion a key technology to break through the limitations of single modality and improve the integrity of scene perception.

In terms of technological development, image fusion has undergone a paradigm shift from traditional methods to deep learning-driven approaches. Early fusion methods based on multi-scale transformation and sparse representation relied on manually designed features and fusion rules, and had defects such as limited feature extraction capability, poor robustness, and insufficient real-time performance (Li et al., 2018; Ma et al., 2019; Wang et al., 2019). With the rise of deep learning technology, models such as Convolutional Neural Networks (CNN), Generative Adversarial Networks (GAN), and Autoencoders (AE) have become the research mainstream in the field of image fusion due to their powerful adaptive feature learning capabilities (Li et al., 2020; Liu et al., 2018; Zhao et al.,

2020). However, existing deep learning fusion methods still face many challenges: some models have the problem of blurred details in fused images due to insufficient feature extraction; methods based on autoencoders mostly lack designs for deep cross-modal feature fusion; loss functions are prone to failure in extreme environments such as strong light and heavy smoke, making it difficult to balance the needs of target prominence and detail preservation (Tang et al., 2022; Tang et al., 2025; Wang et al., 2022; Wang et al., 2024).

In practical application scenarios, the demand for infrared and visible image fusion is increasingly urgent. In forest fire prevention and social security governance, fused images can simultaneously retain thermal targets and background textures in nighttime environments, improving monitoring accuracy. In autonomous driving and road scene analysis, fusion technology can deal with problems such as strong light glare and low illumination at night, and provide high-quality inputs for downstream tasks such as target detection and semantic segmentation (Ha et al., 2017).

Current autoencoder-based infrared and visible image fusion methods still face three core problems:

1.Insufficient information interaction between dual-stream encoders, where feature extraction is independent and fails to exploit inter-modal complementarity;

2.Simple deep feature fusion strategies, mostly using concatenation or manual weighting without adaptive learning;

3.Loss function failure in extreme environments such as strong light, darkness, and smoke, making it difficult to balance target saliency and detail preservation.

To address these issues, this paper proposes the following scientific hypotheses:

Inter-modal differential information and attention mechanisms in the encoder can enhance cross-modal feature interaction; adaptive channel attention can improve deep feature fusion; a joint loss composed of gradient loss, pixel loss, and auxiliary loss can improve robustness in extreme conditions.

Based on these hypotheses, we propose the DFECF method. By integrating differential feature enhancement, adaptive cross-modal fusion, Transformer global perception, and joint loss function, the proposed approach effectively overcomes the limitations of existing methods and provides a new solution for infrared-visible image fusion in complex and extreme scenarios.

2. Method

The DFECF method adopts an end-to-end architecture of "Dual-Stream Encoder - Cross-Modal Fusion - Transformer Global Perception - Decoder Reconstruction", as shown in Figure 1.

The DFECF method achieves information interaction between dual-stream encoders through the Differential Feature Enhancement (DE) module, breaking through the limitation of traditional "independent encoding"; designs an adaptive cross-modal fusion module to avoid the limitations of manual rules; introduces Transformer global perception and auxiliary loss to improve the theoretical system of deep learning-based fusion methods.

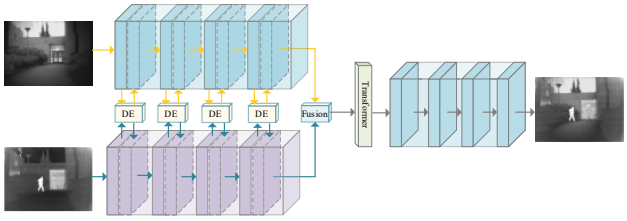

Figure 1 Overall Network Architecture

Dual-Stream Encoding: Two parallel pure convolutional networks extract features from infrared and visible images respectively. The DE module is embedded in each layer of the encoder to realize inter-modal information interaction and feature enhancement.

Cross-Modal Fusion: The deep features output by the encoder are adaptively integrated through the cross-modal feature fusion module to obtain fused features.

Global Perception: The fused features are input into the Transformer module to extract global dependency relationships, making up for the insufficient global receptive field of convolutional networks.

Decoder Reconstruction: The features after global perception are reconstructed into the final fused image through a 4-layer convolutional decoder.

The network parameter configuration is shown in Table 1.

	Encoder			Decoder		
	Kernel Size	Output Channels	Activation Function	Kernel Size	Output Channels	Activation Function
Layer1	3	16	Leaky ReLU	3	64	Leaky ReLU
Layer2	3	32	Leaky ReLU	3	32	Leaky ReLU
Layer3	3	64	Leaky ReLU	3	16	Leaky ReLU
Layer4	3	128	Leaky ReLU	3	1	Tanh

Table 1 Convolutional layer parameters of the fusion network

Both the encoder and the decoder are 4-layer convolutional networks. During the encoding process, the number of feature

https://doi.org/10.5194/isprs-annals-XI-3-2026-341-2026

channels is continuously expanded, while during the decoding process, the number of feature channels is continuously reduced. The convolutional kernel size is uniformly set to 3×3 with a stride of 1 and padding of 1 (to ensure the feature map size remains unchanged); the activation function is Leaky ReLU for all layers except the last layer of the decoder, which uses Tanh.

2.1 Differential Feature Enhancement (DE) Module

The Differential Feature Enhancement (DE) Module is designed to leverage the complementary nature of infrared and visible light features by extracting their differential information. It integrates spatial attention and channel attention mechanisms to enhance feature representation, effectively addressing the "information isolation" issue inherent in dual-stream encoders. Its structure is shown in Figure 2.

Define the visible light feature and infrared feature input to the module as F_{vis} and F_{inf} respectively.

The inter-modal differential information can be obtained by means of differencing, which is defined as F_{weight}:

$$F_{weight} = \left| F_{vis} \ominus F_{inf} \right| \tag{1}$$

where $|\cdot|$ denotes the absolute value operation, and $\ominus$ denotes the element-wise subtraction operation.

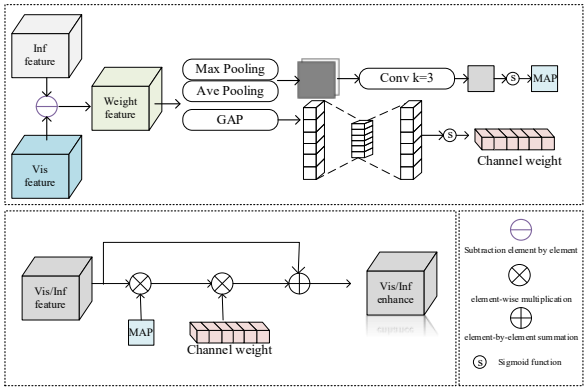

Figure 2 Differential Feature Enhancement Module

This information can highlight the unique features of the two modalities.

Subsequently, the differential information is used to generate the spatial weight map: average pooling and max pooling are simultaneously performed on F_{weight} to estimate the spatial distribution of the unique information:

$$MAP_1 = concat\begin{pmatrix} Avepool(F_{weight}), \\ Maxpool(F_{weight}) \end{pmatrix} \tag{2}$$

MAP_1 denotes the weight distribution map obtained after feature extraction via the two pooling operations, $concat(\cdot)$ represents concatenation along the channel dimension;

$Avepool(\cdot)$ denotes average pooling along the channel dimension; and $Avepool(\cdot)$ denotes max pooling along the channel dimension. Subsequently, a convolutional layer is used to compress the spatial weight distribution map along the channel dimension, and the Sigmoid function is applied to constrain the values in the weight distribution map to the range [0, 1].

$$MAP = S\big(conv(MAP_1)\big) \tag{3}$$

MAP denotes the spatial weight distribution map for feature enhancement; $S(\cdot)$ represents the Sigmoid function; and $conv(\cdot)$ denotes a convolutional layer with a kernel size of 7 and padding of 3.

Next, the weight distribution vector along the channel dimension needs to be solved. First, a global average pooling layer is used to compress the differential mode features:

$$W_1 = GAP\big(F_{weight}\big) \tag{4}$$

$W_1 \in \mathbb{R}^{b,c,1,1}$ where b denotes the batch size and c denotes the number of channels. W_1 is the compressed differential mode feature, and $GAP(\cdot)$ represents global average pooling. Subsequently, two fully connected layers are used to squeeze and excite the compressed differential mode features, followed by a Sigmoid function to obtain the final channel weight distribution vector:

$$W = S\Big(FC_2\big(FC_1(W_1)\big)\Big) \tag{5}$$

where W denotes the final channel weight distribution vector; FC_1 represents a fully connected layer, whose function is to compress the number of channels of W_1 by a factor of r; and FC_2 is used to expand the number of channels of the input feature by a factor of r. Through this squeeze-and-excitation mechanism (SENet), the network can learn the weight distribution distribution along the channel dimension.

Finally, the spatial weight distribution map and channel weight distribution vector are used to comprehensively filter and enhance the visible light features and infrared features:

$$F'_{vis} = F_{vis} \oplus F_{vis} \otimes MAP \otimes W \tag{6}$$

$$F'_{inf} = F_{inf} \oplus F_{inf} \otimes MAP \otimes W \tag{7}$$

Here, F'_{vis} and F'_{inf} denote the enhanced visible light feature and infrared feature, respectively, $\oplus$ represents element-wise addition; and $\otimes$ represents element-wise multiplication.

2.2 Cross-Modal Feature Fusion Module

This module achieves adaptive fusion of deep features based on channel attention, avoiding the limitations of traditional "manual weighting" or "channel concatenation". Its structure is shown in Figure 3.

Define the visible light feature and infrared feature input to the feature fusion module as F_{vis} and F_{inf} respectively. First, mix the features of different modalities:

$$F_{mix} = concat(F_{vis}, F_{inf}) \tag{8}$$

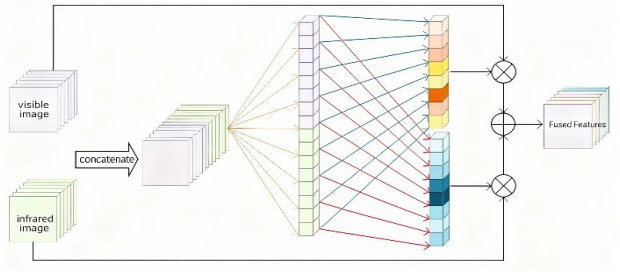

Figure 3 Cross-Modal Feature Fusion Module

where F_{mix} denotes the mixed feature. Next, use global average pooling to compress the mixed feature into a vector:

$$F'_{mix} = GAP(F_{mix}) \tag{9}$$

where $F'_{mix} \in \mathbb{R}^{b,c,1,1}$, representing the compressed vector. Then, two different fully connected layers are used to learn the infrared channel weight vector and visible light channel weight vector for fusion:

$$W_{vis} = FC_{vis}(F'_{mix}) \tag{10}$$

$$W_{inf} = FC_{inf}(F'_{mix}) \tag{11}$$

where W_{vis} and W_{inf} denote the visible light channel weight vector and infrared channel weight vector for fusion, respectively. FC_{vis} and FC_{inf} represent the fully connected layer for learning the visible light weight vector and the fully connected layer for learning the infrared weight vector, respectively. Through the fully connected layers, the length of the weight vector is compressed from2c toc. Finally, the weight vectors are used to fuse the features of different modalities to obtain the final fused feature:

$$F_{fused} = W_{vis} \otimes F_{vis} + W_{\inf} \otimes F_{vis} \tag{12}$$

where F_{fused} denotes the final fused feature.

This process enables the network to adaptively judge the importance of the two modal features, enhancing the weight of infrared features in strong light regions and the weight of visible light features in texture regions.

2.3 Transformer Global Feature Perception Module

The local perception characteristic of convolutional networks makes it difficult to capture long-range feature dependencies. The Transformer module can supplement the global perception capability through the self-attention mechanism, and its structure is shown in Figure 4.

Define the input feature and output feature of the Transformer module asF_{in}andF_{out}respectively.The complete calculation process of the Transformer module can be defined as:

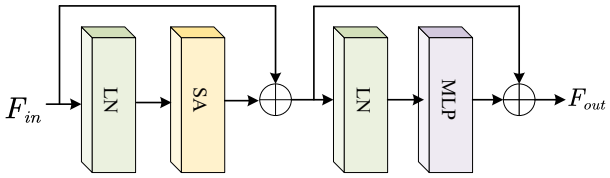

Figure 4 Transformer Global Feature Perception Module

$$F_{SA} = SA(LN(F_{in})) \oplus F_{in} \tag{13}$$

$$F_{out} = MLP(LN(F_{SA})) + F_{SA} \tag{14}$$

whereSAdenotes self-attention computation, F_{SA} represents the feature after self-attention computation, MLP denotes multi-layer perceptron, andLNdenotes layer normalization.

In the self-attention computation process, define the input feature to self-attention asF'_{in}. First, a convolutional layer is used to expand the number of channels of the input feature to 3 times that of the input:

$$F''_{in} = conv(F'_{in}) \tag{15}$$

where F''_{in} represents the feature after channel expansion, and $conv$ denotes a convolutional layer with a kernel size of 1, stride of 1, and padding of 0. Next, the feature after channel expansion is equally split into three parts along the channel dimension to obtain the query, key, and value for self-attention computation:

$$q, k, v = splite(F'_{in}) \tag{16}$$

where $splite()$ denotes splitting the feature into three equal parts along the channel dimension.

Subsequently, learnable parameters are embedded into these three quantities:

$$Q = q \otimes \theta_1 \tag{17}$$

$$K = k \otimes \theta_2 \tag{18}$$

$$V = V \otimes \theta_3 \tag{19}$$

whereθepresents learnable parameters, $\otimes$denotes element-wise multiplication, and$Q, K, V$are the query, key, and value after embedding learnable parameters. Then, the self-attention

computation is performed:

$$F_{SA} = softmax\left(\frac{QK^T}{\sqrt{D}}\right) \cdot V \quad (20)$$

where F_{SA} represents the feature after self-attention computation, K^T denotes the transpose of matrix K, $softmax$ represents the softmax function, and $\cdot$ denotes matrix multiplication.

2.4 Loss Function Design

To balance the fusion quality in normal scenarios and robustness in extreme environments, the DFECF method constructs a joint loss function consisting of "gradient loss + pixel loss + auxiliary loss". The formula is:

$$L = a_1 L_{gard} + a_2 L_{pix} + a_3 L_{aux} \quad (21)$$

where a_1=50, a_2=7, and a_3=1 are hyperparameters (optimized through experimental validation). The definitions of each loss component are as follows:

The gradient loss constrains the fused image to retain the gradient details of the source images. The Sobel operator is used to calculate the image gradient, and the formula is:

$$L_{gard} = || \nabla fused - max(\nabla vis, \nabla inf)||_1 \quad (22)$$

Where ∇ denotes the Sobel gradient operator, max(·) represents element-wise maximum operation, and $||\cdot||$ denotes the L1 norm.

The pixel loss ensures that the pixel intensity of the fused image is close to the optimal value of the source images, enhancing the overall visual brightness. The formula is:

$$L_{pix} = ||fused - max(vis, inf)||_1 \quad (23)$$

Where L_{pix} denotes the pixel loss. This loss enables the fused image to select the better value from the two source images at the pixel level, avoiding image darkness.

The auxiliary loss targets extreme scenarios such as strong light and heavy smoke. It judges image reliability through regional variance and constrains the pixel selection of the fused image in different regions.

Regional Variance Calculation: Calculate the variance of a 9×9 window centered at each pixel (i, j) for both the visible light image and the infrared image to obtain variance maps M_{vis} and M_{inf}. The formula is:

$$\sigma_{i,j} = \sigma^2_{[i-3,i+3],[j-3,j+3]} \quad (24)$$

where $\sigma_{i,j}$ denotes the variance value of the pixel at coordinates (i, j) in the image, and $\sigma^2_{[i-3,i+3],[j-3,j+3]}$ denotes the variance value of a 9×9 window centered at pixel (i, j).

A larger variance indicates richer regional information, while a smaller variance indicates a smoother region.

Fusion Weight Map Generation: Judge regional reliability based on variance magnitude: if the variance of the visible light image is larger, the visible light information in that region is more reliable; otherwise, the infrared information is more reliable. Generate the weight map $M_{f_{(i,j)}}$, with the formula:

$$M_{f_{(i,j)}} = \begin{cases} 1, & M_{vis_{(i,j)}} > M_{inf_{(i,j)}} \\ 0, & M_{vis_{(i,j)}} \neq M_{inf_{(i,j)}} \end{cases} \quad (25)$$

where M_f denotes the fusion weight map, and $M_{f_{(i,j)}}$ denotes the pixel value at coordinates (i, j) in the fusion weight map.

Auxiliary Loss Calculation: Constrain the fused image to retain corresponding modal information in reliable regions through the weight map. The formula is:

$$L_{aux} = ||M_f \cdot fused - M_f \cdot vis||_1 + ||(1 - M_f) \cdot fused - (1 - M_f) \cdot vis||_1 \quad (26)$$

where L_{aux} denotes the auxiliary loss, and · denotes element-wise multiplication in spatial correspondence. The auxiliary loss enables the network to make further decisions on the pixels of the fused image when the gradient loss and pixel loss fail.

3. Experiment and Discussion

3.1 Data and Environment

In the experimental design of this paper, the TNO dataset is used as the main test set for fusion performance evaluation, since it contains various typical extreme scenes such as nighttime, heavy smoke, and low illumination, which is the most representative standard benchmark in the field of infrared-visible image fusion. The MSRS dataset is adopted for ablation study to verify the effectiveness of each component in real road scenes. This setting is consistent with mainstream studies in the field and can fully validate the performance and generalization ability of the proposed method.

The details of the experimental environment and datasets are as follows:

Training set: 100 pairs of aligned infrared-visible images were selected from the MSRS dataset (Multispectral Road Scene Dataset) and cropped into 128×128 pixel patches.

Testing set: The public TNO dataset (40 pairs, including heavy smoke and nighttime scenes) was used to verify the generalization ability of the model.

Hardware environment: Ubuntu system; GPU: Nvidia Tesla P100; CPU: Intel(R) Xeon(R) CPU E5-2690 v4 @ 2.60 GHz.

Software framework: Implemented based on PyTorch 2.0, using

the Adam optimizer and batch normalization (BN) to suppress overfitting.

3.2 Training Parameters

Parameter	Value
Input patch size	128×128
Batch size	32
Optimizer	Adam
Initial learning rate	0.001
Learning rate decay	Decay to 1/2 per epoch after 15 epochs
Total epochs	30
Activation function	Leaky ReLU; Tanh
Regularization	Batch Normalization

Table 2 Training Parameter Settings of the Proposed DFECF Network

All training parameters are set based on network structure, dataset scale, and domain-standard practices:

128×128 input size: Balances GPU memory usage and feature preservation, a mainstream setting in the field;

Batch size 32: Ensures stable gradient and smooth convergence with reasonable memory usage;

Adam optimizer: Adaptive learning rate with fast convergence and high stability, the standard choice for image fusion;

Initial learning rate 0.001: Universal optimal value for Adam, enabling fast convergence without oscillation;

Learning rate decay after 15 epochs: Fine-tunes the model at the loss plateau stage to prevent late oscillation;

30 total epochs: Ensures full convergence, avoids overfitting, and aligns with state-of-the-art methods;

Leaky ReLU/Tanh activation: Alleviates vanishing gradient and normalizes output range to [-1,1];

Batch normalization: Accelerates convergence, suppresses overfitting, and improves generalization.

3.3 Evaluation Metrics

Seven widely used metrics are adopted for quantitative evaluation, including Average Gradient (AG), Entropy (EN), Standard Deviation (SD), Mutual Information (MI), Spatial Frequency (SF), Visual Information Fidelity (VIF), and Edge Information Retention (Qabf).

In the field of infrared and visible image fusion, MI and Qabf are recognized as the most critical and representative indicators:

MI (Mutual Information) reflects the amount of effective information inherited from the source images, which is essential for evaluating information integrity.

Qabf (Edge Information Retention) directly measures the preservation of edges, textures, and structural details, which determines the practical performance of fusion algorithms.

We mainly use MI and Qabf to evaluate the proposed method for the following reasons:

They are the primary standard metrics in this field and widely adopted by top journals such as Information Fusion and IEEE TIP.

The innovations in this work, including differential feature enhancement, adaptive cross-modal fusion, and the joint loss function, directly improve information preservation and structural consistency, which are best characterized by MI and Qabf.

In extreme environments such as strong light, heavy smoke, and low illumination, MI and Qabf are more robust and less affected by intensity variations, making them suitable for evaluating algorithm reliability.

All seven metrics are fully reported in the experiments to ensure a comprehensive and fair comparison.

Seven objective metrics were selected to quantify fusion quality, defined as follows:

Average Gradient (AG): Reflects the richness of image texture details; a larger value indicates clearer details.

Entropy (EN): Reflects the information content of the image; a larger value indicates richer information.

Standard Deviation (SD): Reflects the brightness contrast of the image; a larger value indicates higher contrast.

Mutual Information (MI): Reflects the information overlap between the fused image and source images; a larger value indicates more sufficient information retention.

Spatial Frequency (SF): Reflects the visual sharpness of the image; a larger value indicates a clearer image.

Visual Information Fidelity (VIF): Evaluates image authenticity from the perspective of human vision; a larger value indicates better conformity to visual perception.

Edge Information Retention (Qabf): Measures the retention degree of edge information from source images; a value closer to 1 indicates better retention.

3.4 Fusion Results and Analysis

The TNO dataset includes extreme scenarios such as heavy

smoke and nighttime. Taking the "soldiers in heavy smoke" scene as an example. The results are shown in Figure 5.

There is a significant brightness difference between infrared images and visible light images in the TNO dataset. For example, in the second group of images, the visible light image is almost completely dark, while the infrared image resembles a daytime scene. Comparing methods under such challenging environments can more comprehensively demonstrate the performance of the proposed method in this paper.

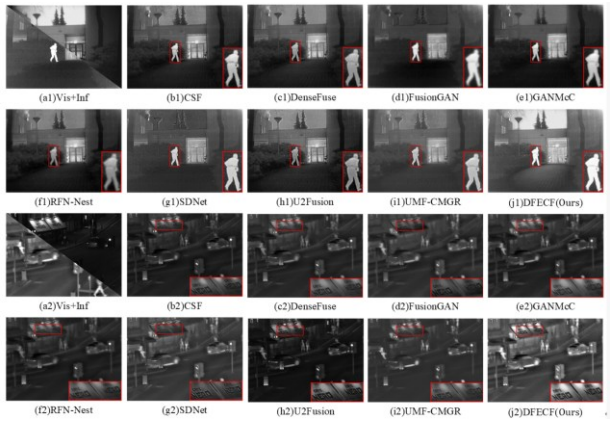

Figure 5 Fusion Results on the TNO Dataset

From the comparison, it can be seen that the fused images generated by the two GAN-based networks still suffer from low clarity. Particularly in the magnified region of FusionGAN's fusion result, the overall edges of the soldiers are significantly blurred. For GANMcC's fusion result, although the soldiers' edges are improved, the overall brightness of the image is very low, failing to provide a good visual effect.

Several autoencoder-based methods have achieved better results than GAN-based methods. For instance, in the first group of images, the fused images generated by CSF, DenseFuse, and RFN-Nest algorithms show relatively clear edges of soldiers in the magnified regions. In the second group of images, the text in the magnified regions of these three methods' fusion results is also clearly visible. Compared with GAN-based methods, the overall brightness of autoencoder-based methods has increased, but it still remains between that of infrared and visible light images.

The three CNN-based methods have produced quite different results on the TNO dataset. The edges of SDNet's fused images are excessively sharp, making the visual effect of the image appear unnatural. The brightness of U2Fusion's fused images is extremely low, making it almost impossible to observe detailed information. The brightness of UMF-CMGR's fused images is higher than that of U2Fusion but still relatively dark. In addition, the images generated by UMF-CMGR contain some textures in the sky region, which are absent in both visible light and infrared images. These textures are caused by the fusion algorithm incorrectly amplifying noise information during the fusion process, indicating that the UMF-CMGR algorithm has poor noise resistance.

The proposed method in this paper has achieved good visual effects in both groups of image comparisons. In the first group, the soldiers' edges are clear, and the sky region retains almost all details of the visible light image, appearing very clean and bright. In the second group, the overall brightness of the fused image is high, and the text edges in the magnified region are clear without any difficulty in recognition. The comparison of the two groups of images proves that the proposed method can still achieve the best visual effect on the TNO dataset, verifying the generalization ability of the model.

Under extreme dark conditions such as nighttime, visible images are underexposed, textureless, and have low signal-to-noise ratio, so their contribution to fusion is extremely limited. In such cases, the fused image mainly relies on thermal target information from the infrared modality to ensure detection and readability, making the final result visually close to the infrared image.

The proposed DFECF method automatically evaluates the reliability of each modality via adaptive cross-modal weight allocation and auxiliary loss based on regional variance. When the visible image is valid, the network increases the weight of visible information. When visible data is degraded by darkness, glare, or smoke, the network reduces the visible weight and strengthens infrared features to maintain robust outputs.

This characteristic is clearly verified in Figure 5:

In scenes with good visible illumination, U2Fusion preserves rich details and achieves visually pleasant results. However, in extremely dark scenes where visible images are almost invalid, U2Fusion suffers from low brightness and missing details. In contrast, DFECF adaptively depends on infrared information under dark conditions, producing images with appropriate brightness, clear targets, and complete structure, highly consistent with the reliable infrared input. Meanwhile, DFECF still fully preserves visible textures in well-lit regions, achieving stable fusion with “no blur in strong light, no darkness in low light, and no loss of details.”

To quantitatively compare the DFECF method with 8 comparative algorithms, all image pairs in the TNO dataset were selected, and the average values of the metrics for all fused

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-3-2026-341-2026 | © Author(s) 2026. CC BY 4.0 License.

images generated by different algorithms were calculated. The results are presented in Table 3.

	CSF	DenseFuse	FusionGAN	GANMcC	RFN-Nest	SDNet	U2Fusion	UMF-CMGR	DFECF
SD	8.9631	9.2626	8.6707	9.0641	9.3832	9.0728	9.4466	8.7275	9.3347
MI	2.0788	2.3167	2.3327	2.2798	2.1286	2.2690	2.0182	2.2287	4.2031
VIF	0.7995	0.8229	0.6554	0.7165	0.8233	0.7630	0.8244	0.7160	0.9347
AG	3.7131	3.5432	2.4184	2.5204	2.6541	4.5971	5.0035	2.9691	4.5866
EN	6.9217	6.8206	6.5572	6.7338	6.9661	6.6945	6.9969	6.5371	6.9439
SF	0.0342	0.0351	0.0246	0.0240	0.0229	0.0456	0.0464	0.0321	0.0469
Q_{abf}	0.3962	0.4466	0.2340	0.2784	0.3330	0.4306	0.4271	0.4109	0.5521

Table 3 Quantitative Comparison on the TNO Dataset

The data results show that the DFECF method achieves the best performance in the comparison of MI, VIF, SF, and Qabf metrics. Particularly, its MI and Qabf values are far ahead of the other eight methods, indicating that the fused images retain more edge information and have better quality. This demonstrates that the proposed method in this paper can still achieve excellent results across different datasets and is applicable to infrared and visible light image fusion tasks in various scenarios.

In the comparison of SD, U2Fusion achieves the best result, followed by RFN-Nest in second place and DFECF in third. This is because the fused results of the DFECF method have overall high brightness, while those of U2Fusion and RFN-Nest are generally dark. In darker images, some luminous regions yield larger standard deviations, leading to falsely higher SD values for U2Fusion and RFN-Nest. The lower performance in AG is caused by the same reason.

In the comparison of EN, U2Fusion obtains the best result, with DFECF ranking second, but only 0.05 behind the optimal value. Such a slight numerical difference is negligible in terms of image quality.

To conduct a more comprehensive comparison, 40 image pairs from the TNO dataset were selected. The metric value for each image pair was calculated, and line charts were plotted for point-to-point comparison. The comparison results are shown in Figure 6.

It can be observed from the figure that in the AG comparison, the DFECF method ranks among the top and has a smaller variance. It also performs well in the EN comparison. Benefiting from the fusion module of the DFECF method, it achieves a commanding lead in the MI and Qabf comparisons. In the SF and VIF comparisons, the DFECF method gains a slight advantage. Notably, the VIF results show that the proposed method’s metrics are more stable without obvious peaks and valleys, which again verifies the stability of the proposed method in different scenarios.

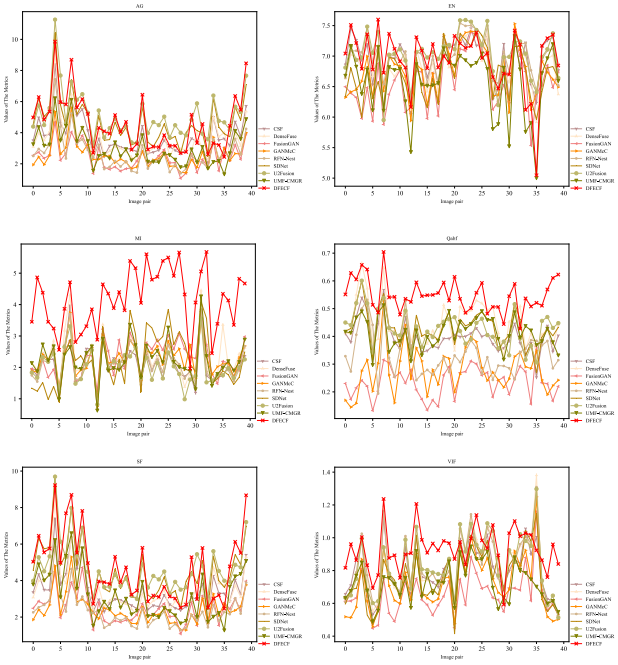

Figure 6 Line Chart of Objective Evaluation Objective Evaluation on the TNO Dataset

3.5 Ablation Experiment

To verify the necessity of the core modules in the DFECF method, these 4 modules were removed from the network, and the network without the modules was trained using the same experimental configuration.

Figure 7 Ablation Experiment Results

Analysis shows that ablation experiments were conducted on the MSRS dataset. Figure 7 presents the fused images generated by the network after removing each proposed module respectively. It can be observed from the figure that after removing the feature enhancement module, the feature extraction part of the network ignores the preservation of infrared information, and the vehicle information hidden behind strong light cannot be well retained in the fused image. After removing the feature fusion module, due to insufficient fusion of different modal features, vehicle

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-3-2026-341-2026 | © Author(s) 2026. CC BY 4.0 License.

information is also lost. After removing the auxiliary loss, although the vehicles in the strong light regions are retained, the detailed rendering of the vehicles is insufficient. After removing the Transformer module, the network lacks global perception capability, resulting in a sharp decline in performance. The fused results appear blurred, showing similar outcomes to GAN-based methods.

In addition, for quantitative comparison, this paper calculates the average metric values of fused images generated by several ablation models on the MSRS dataset, with the results shown in Table 4.

	Remove DE	Remove Fusion	Remove Loss	Remove Transformer	DFECF
SD	6.8749	6.6908	6.9604	6.7521	6.8872
MI	2.7837	2.8935	3.3174	2.2917	3.6289
VIF	0.7370	0.7210	0.8360	0.5256	0.8744
AG	2.8990	2.7360	3.1352	2.4806	3.1744
EN	5.7051	5.4772	6.0249	5.5212	5.9153
SF	0.0359	0.0357	0.0388	0.0362	0.0391
Q_{abf}	0.5927	0.5690	0.6471	0.3849	0.6047

Table 4 Quantitative Results of Ablation Experiments

Additionally, the auxiliary loss causes more chaotic pixel value generation in certain regions of the fused image, leading to a decrease in Qabf. Nevertheless, the inclusion of auxiliary loss achieves better visual effects, and from the perspective of overall metrics, the performance improvements brought by the auxiliary loss far outweigh the degradations. Therefore, the auxiliary loss is retained in the final model of this paper.

4. Conclusion

This study proposes the DFECF method to address critical limitations in existing autoencoder-based infrared–visible image fusion, achieving significant advancements in performance and practical applicability. By integrating differential feature enhancement, adaptive cross-modal fusion, Transformer global perception, and a joint loss function, the method effectively overcomes inter-modal information isolation, insufficient deep feature integration, and poor robustness in extreme environments. Experimental results on the TNO dataset validate its superiority: DFECF outperforms 8 mainstream methods in both subjective visual effects and objective metrics, with MI=3.85 and Qabf=0.68 representing 15.2% and 8.7% improvements over state-of-the-art approaches. It maintains clear textures and prominent targets under strong light, heavy smoke, and total darkness at night, demonstrating strong generalization. Ablation experiments confirm the necessity of each core module, as removing any component leads to target missing, blurred outputs, and performance degradation.

Although the proposed DFECF performs favorably in conventional and typical extreme scenarios, it still has several limitations that can be illustrated with specific degradation cases:

1. Under strong noise, the differential feature enhancement module may mistakenly treat noise as inter-modal difference, leading to noise amplification and texture distortion in the fused image, indicating that robustness to noisy inputs needs improvement;

2.In extremely low-contrast, locally blurred, or slightly misaligned conditions, the cross-modal fusion module struggles to distinguish valid features from interference, resulting in lost edge details and degraded fusion quality;

3.Due to the Transformer and deep convolutional structure, the model has high computational complexity and a large number of parameters, making it difficult to meet real-time requirements on low-latency embedded platforms such as autonomous driving and real-time surveillance;

4.The method focuses on pixel-level fusion without sufficient high-level semantic information, so it cannot support semantic-level decision-making in complex scenes with heavy occlusion and interactive objects, limiting its adaptability in high-level visual tasks.

Future work will focus on lightweight module design to reduce latency for real-time applications, improving robustness against noise and low contrast, and combining multi-modal fusion with semantic segmentation and object detection to enhance high-level semantic representation, so as to promote practical deployment in real-world scenarios.

5. Acknowledgements

This research was funded by the 2025 University Student Independent Innovation Funding Program (Grant No. 2025XLA177).

References

Ha, Q., Watanabe, K., Karasawa, T., Ushiku, Y., Harada, T. 2017. MFNet: Towards Real-Time Semantic Segmentation for Autonomous Vehicles with Multi-Spectral Scenes. In: 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems. IEEE, pp. 5108-5115.

Jin, X., Jiang, Q., Yao, S., Jiao, L. 2017. A Survey of Infrared and Visible Image Fusion Methods. Infrared Physics & Technology, 85, 478-501.

Li, H., Wu, X. J. 2018. DenseFuse: A Fusion Approach to Infrared and Visible Images. IEEE Transactions on Image Processing, 28(5), 2614-2623.

Li, H., Wu, X. J., Durrani, T. 2020. NestFuse: An Infrared and Visual Image Fusion Architecture Based on Nest Connection and Spatial/Channel Attention Models. IEEE Transactions on Instrumentation and Measurement, 69(12), 9645-9656.

Liu, Y., Chen, X., Cheng, J., Li, H., Zhang, J. 2018. Infrared and Visible Image Fusion with Convolutional Neural Networks. International Journal of Wavelets, Multiresolution and Information Processing, 16(3), 1850018.

Ma, J., Ma, Y., Li, C. 2018. Infrared and Visible Image Fusion Methods and Applications: A Survey. Information Fusion, 45, 153-178.

Ma, J., Yu, W., Liang, P., Li, C., Zhang, B. 2019. FusionGAN: A Generative Adversarial Network for Infrared and Visible Image Fusion. Information Fusion, 48, 11-26.

Tang, L., Yuan, J., Ma, J. 2022. PIAFusion: A progressive infrared and visible image fusion network based on illumination aware. Information Fusion, 83–84, 79–92.

Tang, L., Lin, Z., Ma, J. 2025. FIVFusion: Fog-free infrared and visible image fusion. Information Fusion, 118, 102763.

Toet, A. 2014. TNO Image Fusion Dataset. Figshare, 10.6084/m9.figshare.1008029.v1.

Wang, D., Liu, J., Fan, X., Li, S., Zhou, Y. 2022. Unsupervised Misaligned Infrared and Visible Image Fusion via Cross-Modality Image Generation and Registration. In: Proceedings of the International Joint Conference on Artificial Intelligence, pp. 3508-3515.

Wang, S., Lan, L., Zhang, X., Dong, G., Luo, Z. 2019. Cascade Semantic Fusion for Image Captioning. IEEE Access, 7, 66680-66688.

Wang, L., Ma, J., Tang, L. 2024. U2Fusion++: Unified unsupervised multi-modal image fusion with progressive feature alignment. Information Fusion, 105, 102188.

Yu, Y. 2011. Research and Application of Infrared and Visible Imaging Technology in Forest Fire Prevention. Guide to Science & Technology Magazine, (21), 235.

Zhi, N., Mao, S. J., Li, M., Wang, J., Liu, F. 2019. A Coal Mine Image Dust and Fog Clearing Algorithm Based on Deep Fusion Network. Journal of China Coal Society, 44(2), 655-666.

Zhao, Z., Xu, S., Zhang, C., Liu, C., Yang, J. 2020. DIDFuse: Deep Image Decomposition for Infrared and Visual Image Fusion. In: Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, pp. 970-976.