Received: 20 December 2022 | Revised: 17 February 2023 | Accepted: 20 March 2023

DOI: 10.1002/cpe.7712

SPECIAL ISSUE PAPER

WILEY

Multiscale aggregation and illumination-aware attention network for infrared and visible image fusion

Wenhao Song[1] | Wenzhe Zhai[1] | Mingliang Gao[1] | Qilei Li[2] | Abdellah Chehri[3] | Gwanggil Jeon[1,4]

[1]School of Electrical and Electronic Engineering, Shandong University of Technology, Zibo, China

[2]School of Electronic Engineering and Computer Science, Queen Mary University of London, London, UK

[3]Department of Mathematics and Computer Science, Royal Military College of Canada, Kingston, Canada

[4]Department of Embedded Systems Engineering, Incheon National University, Incheon, South Korea

Correspondence
Mingliang Gao and Gwanggil Jeon, School of Electrical and Electronic Engineering, Shandong University of Technology, Zibo, China.
Email: mlgao@sdut.edu.cn and gjeon@inu.ac.kr

Funding information
Nature Science Foundation of China, Grant/Award Numbers: 61601266, 61801272; National Natural Science Foundation of Shandong Province, Grant/Award Numbers: ZR2021QD041, ZR2020MF127

Abstract

Image fusion plays a significant role in computer vision since numerous applications benefit from the fusion results. The existing image fusion methods are incapable of perceiving the most discriminative regions under varying illumination circumstances and thus fail to emphasize the salient targets and ignore the abundant texture details of the infrared and visible images. To address this problem, a multiscale aggregation and illumination-aware attention network (MAIANet) is proposed for infrared and visible image fusion. Specifically, the MAIANet consists of four modules, namely multiscale feature extraction module, lightweight channel attention module, image reconstruction module, and illumination-aware module. The multiscale feature extraction module attempts to extract multiscale features in the images. The role of the lightweight channel attention module is to assign different weights to each channel so as to focus on the essential regions in the infrared and visible images. An illumination-aware module is employed to assess the probability distribution regarding the illumination factor. Meanwhile, an illumination perception loss is formulated by the illumination probabilities to enable the proposed MAIANet to better adjust to the changes in illumination. Experimental results on three datasets, that is, MSRS, TNO, and RoadSence, verify the effectiveness of the MAIANet in both qualitative and quantitative evaluations.

KEYWORDS

attention mechanism, autoencoder, illumination awareness, image fusion, multiscale feature

1 | INTRODUCTION

Numerous cameras feature both visible and infrared imaging sensors to capture visible and infrared pictures. Visible images include abundant textural information, but are vulnerable to environmental and climatic concerns. Since infrared photographs are recorded on thermal radiation, they are resistant to influence from unfavorable environments.[1] Infrared photos, however, have poor spatial resolution and texture information. In order to take full advantage of these images, more and more infrared and visible image fusion (IVIF) techniques have been developed in the past decade.[2,3]

The existing IVIF methods are broadly categorized into traditional-based methods and deep learning-based methods.[3] The traditional-based methods typically consist of three essential parts, that is, feature extraction, fusion strategy, and reconstruction.[4-6] However, these methods have major drawbacks. The same transformation strategies should be adopted to extract features in the infrared and visible images to ensure the viability of later feature fusion. Therefore, these methods need to consider the feature differences between infrared and visible images, which would lead to

a lack of information on the extracted features. Meanwhile, the traditional feature fusion strategies are relatively rudimentary, resulting in degraded fusion performance.[2]

The explosion of deep learning greatly promotes the development of image fusion technology.[7] The deep learning-based fusion methods are generally divided into three categories according to their adopted architectures, that is, autoencoder (AE)-based methods, convolutional neural Network (CNN)-based methods, and generative adversarial network (GAN)-based methods.[3,7] The AE-based methods usually utilize a pre-trained autoencoder for feature extraction and image reconstruction, and feature fusion is done by traditional fusion rules.[8]

CNNs are typically introduced into image fusion frameworks in two ways. One is to construct an end-to-end fusion network and design a loss function to train the fusion network end-to-end.[9] The other is to train the CNNs to guide feature fusion, while feature extraction and image reconstruction are implemented by traditional methods.[10] The GAN-based approach utilizes a generator to generate the fused image, and the fusion network is optimized by adversarial between the generator and the discriminator.[11]

However, the existing fusion networks are incapable of extracting intensity information in infrared images as well as detail information in visible images under different illumination environments. To this end, a multiscale aggregation and illumination-aware attention network (MAIANet) is proposed in this paper for the IVIF task. A multiscale feature extraction (MFE)module is built to extract multiscale features from infrared and visible images. Meanwhile, a cross-modal difference-aware fusion (CMDAF) unit[4] is adopted in the MFE model to merge complementary features of infrared and visible images. Moreover, a lightweight channel attention (LCA) module is built to determine the weights of each feature channel. Finally, an image reconstructor (IRC) module is applied to generate the final fused images. To make the network better handle the illumination intensity variations in visible images, the illumination-aware (IA) module is adopted to calculate the distribution regarding the illumination factor of the visible image. In sum, the main contributions of this work are:

1. A Multiscale Aggregation and Illumination-aware Attention Network (MAIANet) constructed for IVIF tasks. It is capable of extracting discriminative and representative multiscale features from pairs of infrared and visible images, and pays particular attention to fluctuations in light.
2. An MFE module is constructed to extract multiscale information of the infrared and visible images.
3. An LCA model is built to focus on the intensity information of infrared and the detailed information of visible images.
4. An IA module is adopted to assess the probability distribution regarding the illumination factor. Meanwhile, an illumination-aware loss function is formulated to guide the network to adjust the illumination variations.

The remaining parts of this paper are structured as follows. Related works of image fusion are introduced in Section 2. Details of the proposed MAIANet are presented in Section 3. Comparative experiments are conducted in Section 4. Conclusions are derived in Section 5.

2 | RELATED WORK

2.1 | Traditional image fusion methods

The traditional image fusion methods mainly consist of three parts, namely feature extraction, feature fusion, and feature reconstruction.[2,12] Thereinto, feature extraction, and feature fusion play key roles in image fusion methods because feature reconstruction is the inverse feature extraction process.

Traditional image fusion methods typically depend on spatial transformation techniques to extract features. To efficiently fuse the source images, Bhat et al.[13] proposed a multi-focus image fusion method by combining the Neutrosophic set and stationary wavelet transform. Mei et al.[14] proposed a medical image fusion method based on the non-subsampling contourlet transform and adaptive pulse-coupled neural network. The technique effectively preserves the spectral features of the source images. Also, the saliency map is widely used in infrared images to improve the visualization of the visible image. Han et al.[15] presented a saliency-aware fusion method. The method first generates a saliency map by performing saliency target detection on the infrared image, and a subsequent fusion step biases the fusion result towards the visible image. Meng et al.[16] proposed an image fusion algorithm that combines saliency maps and interest points to extract saliency maps from infrared images. The saliency results are more accurate and better able to locate the object accurately.

Besides, the optimization-based approaches present the image fusion community with fresh perspectives.[17] The loss function for image fusion is defined as a weighted combination of intensity loss and texture loss. In addition, the hybrid models, combined with the advantages of different frameworks, are employed to pursue better image fusion performance.[7,18] Specifically, Huang et al.[19] proposed a synthetic aperture radar and multispectral image fusion via combining non-subsampled shearlet transform and activity measure, which could improve the interpretation ability of SAR images. Zhang et al.[20] developed an image fusion framework by combining the Laplacian pyramid and sparse representation to integrate the information from multispectral and synthetic aperture radar images.

2.2 | Autoencoder-based image fusion methods

Deep learning has emerged as a powerful tool for resolving image fusion problems, leveraging the substantial capabilities of feature learning.[17] Thereinto, the autoencoder (AE)-based method is an important branch. It achieves feature extraction and reconstruction by a trained autoencoder network, while feature fusion is performed by traditional strategies.[21]

For example, Prabhakar et al.[22] proposed a CNN-structured autoencoder network termed Deepfuse for image fusion. They adopted an encoder to extract features from five layers and utilized a decoder to reconstruct the fused image. Li et al.[8] presented an encoder that incorporates a dense block and convolutional layer to obtain more beneficial features. The fusion layer employs two different strategies, that is, addition strategy and the l_1-norm strategy, and it utilizes four 3×3 convolutional layers to reconstruct the fused image. In the aforementioned methods, manually designed fusion rules are utilized which limits the performance. To address this issue, the end-to-end fusion framework is introduced for the better fusion of depth features. Li et al.[23] proposed an end-to-end fusion network architecture in which a learnable fusion strategy termed residual fusion network (RFN) is designed to improve the performance of the image fusion framework. Ma et al.[24] proposed a cascade network to generate the decision map and fused images, which aims to improve the fused image in both quality and structure.

2.3 | CNN-based image fusion methods

The CNN-based image fusion methods are developed to avoid manually designed fusion rules. Generally, CNN is introduced into the image fusion framework in two forms. One is end-to-end image fusion by constructing a network structure and a well-designed loss function, where feature extraction, feature fusion, and image reconstruction are done in an implicit manner.[2]

For example, Zhang et al.[9] introduced an end-to-end CNN-based method in which a proportionate maintenance loss of gradient and intensity was adopted to generate the fused image directly. Another is to use a pre-trained CNN model to formulate fusion rules, while the modules of feature extraction and image reconstruction are performed by traditional methods.[10] Liu et al.[25] utilized a two-branch Siamese network to generate fusion weights, and employed the Laplace pyramids to implement image decomposition and reconstruction. Moreover, Xu et al.[26] proposed a unified model for multi-fusion missions. The cross-fusion between different image fusion tasks is considered and consolidated with elastic weights. However, because there is a lack of ground truth, the CNN-based network is unable to demonstrate its full potential in the picture fusion domain.

2.4 | GAN-based image fusion methods

In the GAN-based methods, the probability distribution of the target is estimated depending on the adversarial game between the generator and the discriminator, which can perform feature extraction, feature fusion, and image reconstruction synchronously in an implicit manner.[3,27,28]

For example, Ma et al.[11] proposed a GAN-based method to fuse infrared and visible images, referred to as FusionGAN. The method treated the IVIF as a game between the generator and discriminator. The generator aims to generate a fused image with major infrared intensities as well as additional visible gradients and maintain the allocation balance. The function of the discriminator is to compel the generator to generate a fused image containing more textures. However, the fused images generated by a single discriminator may be biased toward the infrared or visible images. To ensure the fused image simultaneously maintain the structural and detailed information from these multi-modal sources, Ma et al.[29] recently designed DDcGAN to learn unbiased knowledge with a dual-discriminator GAN model. In addition, Yang et al. designed two additional loss functions in ResNetFusion,[30] in which a detailed loss is to improve the detail quality, and a target edge enhancement loss is to sharpen the edges of targets.

3 | PROPOSED METHOD

3.1 | Network architecture

The framework of the proposed MAIANet is depicted in Figure 1. As shown in Figure 1, a multiscale feature extraction (MFE) module is introduced to extract multiscale features of infrared and visible images. By this means, the network can learn features at different scales adequately and improve the effectiveness of the features. Then, a lightweight channel attention (LCA) module is adopted to further suppress irrelevant channel features and enhance key channel features, the lightweight channel attention module is adopted. Subsequently, an image reconstruction module is established to generate the fused image. In addition, an illumination-aware module[4] is employed to estimate the illumination of the visible image.

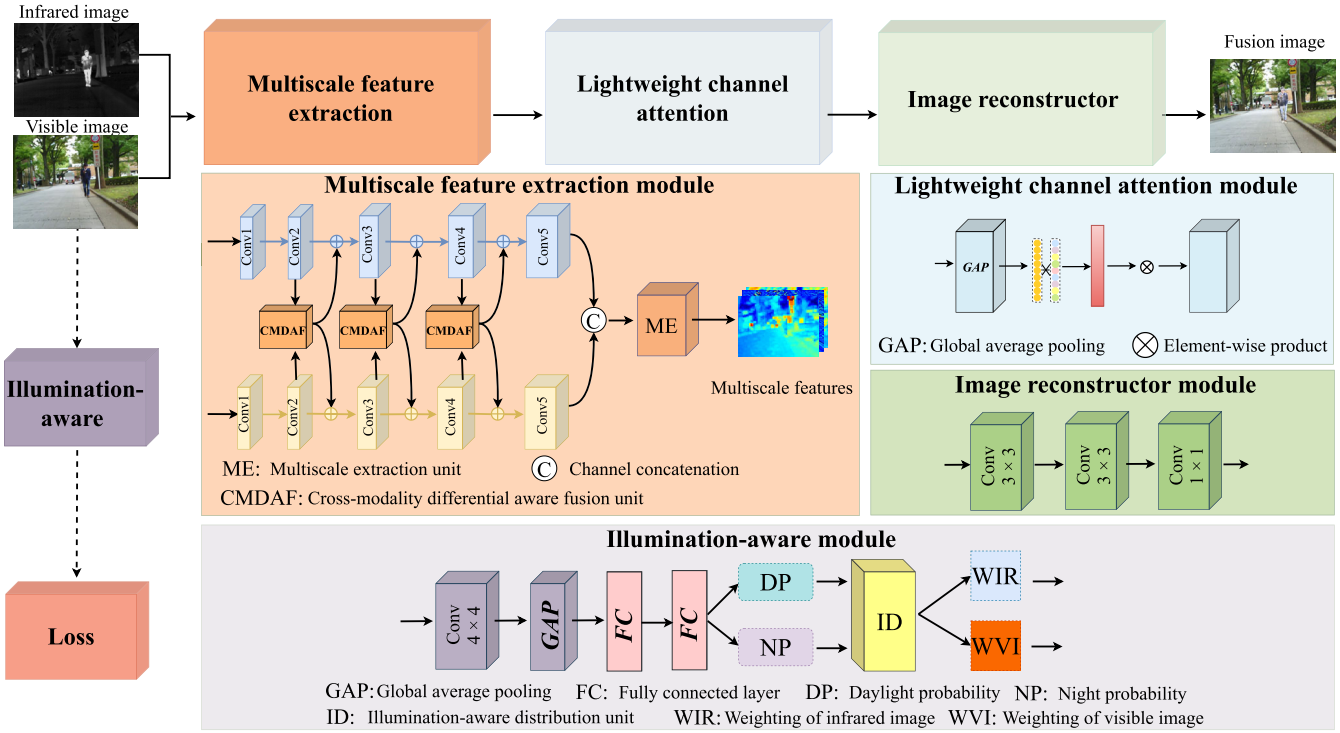

FIGURE 1 Framework of the MAIANet for infrared and visible image fusion.

TABLE 1 Architecture of the multiscale feature extraction module.

Layers	Kernel size	Input channels	Output channels	Activation function
Layer1	1 × 1	1	16	LReLU
Layer2	3 × 3	16	16	LReLU
Layer3	3 × 3	16	32	LReLU
Layer4	3 × 3	32	64	LReLU
Layer5	3 × 3	64	128	LReLU

3.2 | Multiscale feature extraction module

The MFE module is built to extract multiscale features from both visible and infrared images. As illustrated in Figure 1, five identical convolutional layers are deployed in a parallel way to extract low-level features. The configuration lists the specifics of each convolutional layer in the feature extraction in Table 1. Specifically, the multi-modal images are fed into a 1 × 1 convolutional layer to initialize the low-level feature representation. We aim to fully extract the complementary information across the low-level feature by adopting a CMDAF unit[4] which is illustrated in Figure 2. It's embedded in the first four layers and formulated as

$$\begin{aligned}\hat{F}_{ir}^{i} &= F_{ir}^{i} \oplus \delta\left(GAP\left(F_{vi}^{i} - F_{ir}^{i}\right)\right) \odot \left(F_{vi}^{i} - F_{ir}^{i}\right),\\ \hat{F}_{vi}^{i} &= F_{vi}^{i} \oplus \delta\left(GAP\left(F_{ir}^{i} - F_{vi}^{i}\right)\right) \odot \left(F_{ir}^{i} - F_{vi}^{i}\right),\end{aligned} \tag{1}$$

where $\oplus$ stands for the element-wise summation, and $\odot$ denotes the channel-wise multiplication. The $\delta(\cdot)$ and GAP$(\cdot)$ refer to the Sigmoid activation function and the global average pooling operation, respectively. The value of ith convolutional layer features for the infrared and visible images are F_{ir}^{i} and F_{vi}^{i}, respectively.

Thus, the common and complementary features could be fully extracted from the infrared and visible images by the MFE module, which can be formulated as,

$$\{F_{ir}, F_{vi}\} = \{E_F(I_{ir}), E_F(I_{vi})\}, \tag{2}$$

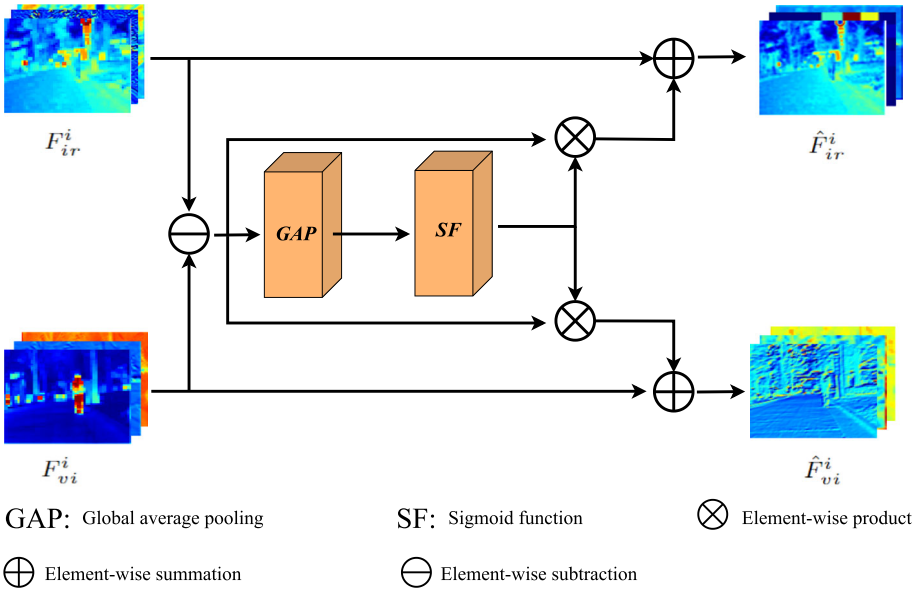

FIGURE 2 Framework of cross-modality differential aware fusion (CMDAF) unit.

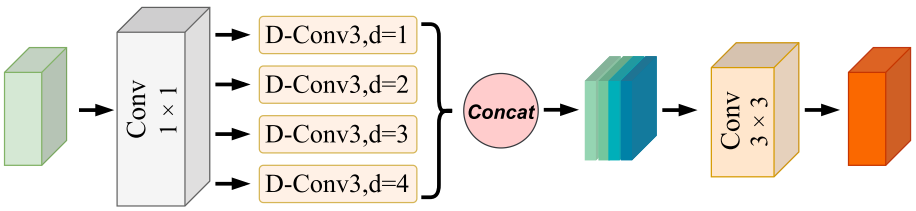

FIGURE 3 Framework of multiscale extraction (ME) unit. "D-Conv3, d = t" indicates the dilated convolution with a 3 × 3 convolution kernel with dilation ratios of t.

where F_{ir} / F_{vi} respectively denote the infrared/visible features. I_{ir} and I_{vi} indicates the infrared and visible images

Subsequently, these extracted features from the multi-modal sources are concatenated and fed into the multiscale extraction (ME) unit to extract the multiscale features. The framework of the ME unit is shown in Figure 3. Specifically, the ME unit squeezes the channels of the feature maps by 1 × 1 convolution layer. After that, the squeezed feature map is handled by dilated convolution with different dilation ratios of 1, 2, 3, and 4 to conserve the multiscale features of the source images. The feature maps are concatenated and fed into a 3 × 3 convolution layer. The final feature map is the same size as the input.

3.3 | Lightweight channel attention module

To further emphasize the intensity information of the infrared images and the detailed information of the visible images, an LCA module is built, as shown in Figure 1.

The LCA module employs a local cross-channel interaction strategy without dimensionality reduction. Meanwhile, it utilizes a dynamic convolution kernel to adapt to the size of the input channel feature maps and extract different ranges of dimensional features. In addition, this module can better focus the significant target and detailed information of infrared and visible images on the promise of small model complexity.

Specifically, the multiscale feature $H \times W \times C$ is compressed by global averaging pooling of spatial information to generate a 1-dimensional feature map. Then, the compressed feature maps are convolved with a 1 × 1 convolution kernel to learn the weight of different channels.

The dimension of the channel attention feature map is formulated as follows:

$$C = 2^{\phi(k)}, \tag{3}$$

where $\phi(k) = \gamma * k - b$ is a simple linear function. The kernel size k is denoted as,

$$k = \psi(C) = \left| \frac{\log_2(C)}{\gamma} + \frac{b}{\gamma} \right|_{odd}, \tag{4}$$

where ψ is the mapping function. The C stands for the number of channels. $|t|_{odd}$ denotes the nearest odd number of t. In this paper, γ and b are set to 1 and 2, respectively. Finally, the obtained channel attention feature map is multiplied by the original feature map channel by channel, and the feature map with channel attention is generated.

3.4 | Image reconstruction module

The IRC module consists of five convolutional layers to produce the fused image. Table 2 displays the Architecture of the IRC module. The first four layers use kernel size of 3×3, while the last layer adopts kernel size of 1×1. All the convolution layers employ the LReLU activation function except for the final convolution which applies a Tanh activation function to generate the output.

The padding is set to the "SAME" in all convolutional layers to prevent information loss during image fusion. In addition, the stride is set to 1 for all convolution layers except for the first and the last layers. By these means, the size of the fused image remains the same as the size of the source images.

3.5 | Illumination-aware module

We further estimate the distribution of the illumination factor by adopting the IA module,[4] as shown in Figure 1. The configuration of the convolutional layer in the IA module is depicted in Table 3.

Given a visible image I_{vi}, the illumination-aware module process is defined as,

$$\{P_d, P_n\} = N_{IA}(I_{vi}), \tag{5}$$

where N_{IA} stands for the illumination-aware module. P_d and P_n indicate the probability that the visible image belongs to the day and night, respectively. Since the infrared image captures thermal radiation emitted by objects, the infrared image lacks of illumination information. In contrast, the visible images capture illumination reflection information and therefore contain rich illumination information. The visible image is fed into the convolutional layer, and the illumination probability is calculated by two fully connected layers.

The spatial information is compressed, and illumination data is extracted using a 4×4 convolutional layer with a 2 stride. A Leaky ReLU (LReLU) activation function is adopted in all convolutional layers. The padding is set to the "SAME" to ensure that the size of fused images is consistent with

TABLE 2 Architecture of the image reconstructor module.

Layers	Kernel size	Input channels	Output channels	Activation function
Layer1	3×3	256	256	LReLU
Layer2	3×3	256	128	LReLU
Layer3	3×3	128	64	LReLU
Layer4	3×3	64	32	LReLU
Layer5	1×1	32	1	Tanh

TABLE 3 Architecture of the multiscale feature extraction module.

Layers	Kernel size	Input channels	Output channels	Activation function
Layer1	4×4	1	16	LReLU
Layer2	4×4	16	32	LReLU
Layer3	4×4	32	64	LReLU
Layer4	4×4	64	128	LReLU

source images. After that, the illumination information is integrated using a global average pooling manipulation. Finally, two fully connected layers are employed to calculate the illumination probabilities.

3.6 | Loss function

The illumination-aware loss is presented in MAIANet to fuse significant information in illumination conditions which are formulated as,

$$\mathcal{L}_{\text{illum}} = W_{ir} \cdot \mathcal{L}_{int}^{ir} + W_{vi} \cdot \mathcal{L}_{int}^{vi}, \tag{6}$$

where $\mathcal{L}_{\text{int}}^{ir}$ and $\mathcal{L}_{\text{int}}^{vi}$ stands for the intensity losses that are corresponding to the infrared and visible images, respectively. W_{ir} and W_{vi} denote the illumination-aware weights calculated over infrared and visible sources. They are formulated as:

$$\begin{aligned} W_{ir} &= \frac{P_n}{P_d + P_n}, \\ W_{vi} &= \frac{P_d}{P_d + P_n}, \end{aligned} \tag{7}$$

where P_d and P_n represent the probability regarding the day time or night time of the given image. To adaptively tune the intensity factor of the fused image, With the formulation of Equation (7), we thereby calculate the illumination-aware weights, that is, W_{ir} and W_{vi}.

To minimize the difference between the fused image and the source images, we design the intensity loss as

$$\begin{aligned} \mathcal{L}_{int}^{ir} &= \frac{1}{HW} \left\| I_f - I_{ir} \right\|_1, \\ \mathcal{L}_{int}^{vi} &= \frac{1}{HW} \left\| I_f - I_{vi} \right\|_1, \end{aligned} \tag{8}$$

where H and W are the height and width of the images. $\| \cdot \|_1$ is the l_1-norm. I_{ir} and I_{vi} are the infrared and visible images. I_f represents the fused image.

The proposed method is driven by the illumination loss to dynamically preserve intensity information from the source image based on illumination circumstances. However, it cannot guarantee the overall distribution of the intensity factor, which may result in suboptimal fusion results caused by imbalanced intensity distribution. To solve this problem, an auxiliary intensity loss is further employed, which is formulated as follows:

$$\mathcal{L}_{\text{aux}} = \frac{1}{HW} \left\| I_f - \max(I_{ir}, I_{vi}) \right\|_1, \tag{9}$$

A texture loss is defined as the following, and it is used to inject more detailed texture information into the fused image:

$$\mathcal{L}_{\text{texture}} = \frac{1}{HW} \left\| \left| \nabla I_f \right| - \max(\left| \nabla I_{ir} \right|, \left| \nabla I_{vi} \right|) \right\|_1, \tag{10}$$

where ∇ indicates the gradient operator (Note: in this work the Sobel operator is adopted). $| \cdot |$ denotes the absolute operation.

The final loss function of the proposed model is the weighted combination of several components. It is represented as:

$$\mathcal{L}_{\text{fusion}} = \lambda_1 \cdot \mathcal{L}_{\text{illum}} + \lambda_2 \cdot \mathcal{L}_{\text{aux}} + \lambda_3 \cdot \mathcal{L}_{\text{texture}}, \tag{11}$$

where the hyperparameters of $\lambda_1, \lambda_2, \lambda_3$ are set to 3, 7, and 50, respectively.

In addition, the cross-entropy loss is used to constrain the training process of the illumination-aware module, which is formulated as:

$$\mathcal{L}_{IA} = -z \log \sigma(y) - (1 - z) \log(1 - \sigma(y)), \tag{12}$$

where z represent the $y = [P_n, P_d]$ illumination of the input image, $y = [P_n, P_d]$ represent the probability regarding the day time or night time of the input image, and σ refers to the softmax function.

4 | EXPERIMENTAL RESULTS AND ANALYSIS

4.1 | Dataset

To verify the effectiveness of the proposed MAIANet, qualitative and quantitative experiments are conducted on the MSRS,[4] TNO,[31] and RoadSence[26] datasets. The MSRS dataset contains 1,444 pairs of aligned infrared and visible images with high quality. The TNO dataset[31] consists of multi-spectral nighttime images in various military-related scenes.

The TNO dataset is divided into three sequences, containing 19, 23, and 32 pairs of images, respectively. To compensate for the lack of quantity in the existing datasets, Xu et al.[26] built the RoadSence dataset based on the FLIR video.

The RoadSence dataset contains 221 matched visible and infrared image pairings in a variety of different roadside contexts, including roads, cars, and people.

The representative infrared and visible images from these three datasets are depicted in Figure 4. Inspired by Tang et al.,[4] 752 pairs of infrared and visible images of the MSRS dataset are randomly selected as the training set, and 360, 21, and 40 image pairs from the MSRS, TNO, and RoadSence datasets are used as the test set.

FIGURE 4 Representative infrared and visible images from the MSRS dataset, TNO dataset, and RoadScene dataset. The first and the second rows in each dataset denote the infrared and visible images, respectively.

TABLE 4 Parameters of the training phase.

	Batch size	Training step	Epoch
Illumination-aware module	128	438	100
Multiscale feature extraction	64	819	30
Image reconstruction module	64	819	30

Algorithm 1. Pseudocode of the training phase

Require: Infrared images I_{ir} and visible images I_{vi}
Ensure: Fused images I_f
for M_1 **do**
  for p_1 **do**
    Select visible images;
    Generate the illumination probability P_n and P_d with the illumination-aware module;
    Calculate the cross-entropy loss $\mathcal{L}_{IA}$ according to Eq∼12;
    Update the parameters of the illumination-aware module by Adam Optimizer;
  end for
end for
for M_2 **do**
  for p_2 **do**
    Select infrared images;
    Select visible images;
    Generate the illumination probability P_n and P_d with the illumination-aware module;
    synthesize the fused images with the fusion network;
    Calculate the total loss L_{fusion} according to Eq∼11;
    Update the parameters of the fusion network by Adam Optimizer;
  end for
end for

4.2 | Training details

The MSRS dataset is employed to train the MAIANet. The training of the proposed method is divided into two stages. First, the illumination-aware module is trained using 427 daytime images and 376 nighttime images. In addition, the multiscale feature extraction and image reconstruction module are trained by 376 pairs of daytime and 376 pairs of nighttime infrared and visible image pairs. To augment the training samples, the images are cropped into patches with a size of 64×64 and the cropping stride is set to 64. To minimize the side effects caused by internal covariate shift, these patches are normalized to $[-1, 1]$. The daytime vector is set to $[1, 0]$ and vise-verse for the nighttime scene as $[0, 1]$. The parameter of the training phase is listed in Table 4. The training step in one epoch is set as p, the batch size is set to b and it takes M epochs to train a model. The illumination-aware module is trained for $M_1 = 100$ epochs, train steps as $p_1 = 438$ and the batch size is defined as $b_1 = 128$. The multiscale feature extraction and image reconstruction are trained for $M_2 = 30$ epochs, training steps as $p_2 = 819$ and the batch size is defined as $b_2 = 64$. The Adam is introduced as the optimizer. The learning rates of the multiscale feature extraction and image reconstruction are initialized as 0.001.

Considering that some of the visible images in MSRS and RoadSence datasets are color images, a special fusion strategy[22] is utilized to retain their color information in the fused images. Specifically, the visible image is first converted to the YCbCr color space. The Y channel of the infrared and visible images are then merged using the proposed method. In the end, the fused image is converted to RGB color space. The pseudocode of the training phase is shown in Algorithm 1.

4.3 | Qualitative evaluation

The qualitative evaluation is the degree of human satisfaction with the image from the human observation criteria, including brightness, contrast, color, and naturalness metrics. Seven state-of-the-art competitors, that is, Deepfuse,[22] Densefuse,[8] Dualbranch,[32] FusionDN,[26] GAN-FM,[33]

GTF,[34] and SwinFusion[35] are compared with the proposed MAIANet. Comparative results on the MSRS, TNO, and RoadSence datasets are shown in Figures 5,6, and 7, respectively.

In Figure 5, the sample depicts a representative pair of infrared and visible images on the MSRS dataset. Specifically, a bright region (in the red box) and a dark region (in the green box) are selected for analysis. In the bright region, the thermally radiated information of the infrared image can be utilized as complementary information to the visible image. Therefore, a superior fusion method should emphasize the important targets while maintaining the textural features of the visible images. Comparative results prove that the GTF,[34] Deepfuse,[22] Densefuse,[8] Dual-branch,[32] and FusionDN[26] fail to maintain the detailed information of the visible image. Although the GAN-FM[33] and SwinFusion[35] combine the texture information of the visible image with the salient target information in the infrared image, the detailed information of the background region is missing. The MAIANet not only highlights salient target information but also retains background detail information well in visible images.

In Figure 6, the sample illustrates a representative pair of infrared and visible images in the TNO dataset, a person walking down a forest path with houses nearby. The illumination is too dark to discover the person in the visible image, and it is only in the infrared image that the person can be detected. Meanwhile, the infrared image lacks detailed information about the trees and houses. It shows that the Deepfuse,[22] Densefuse,[8] Dualbranch,[32] FusionDN,[26] and GTF[34] fail to capture the intensity information of the person. As shown in Figure 6, the MAIANet retains the edge and detail information. Although the GAN-FM[33] performs well in persevering intensity information from images, it does not perform well in maintaining the edge and detail information of images. The green contour in the visible image shows the eaves and windows, which contain a lot of edge and detail information. The fusion of GAN-FM[33] results in the loss of edge and detail information. Compared with other competitors, the proposed MAIANet

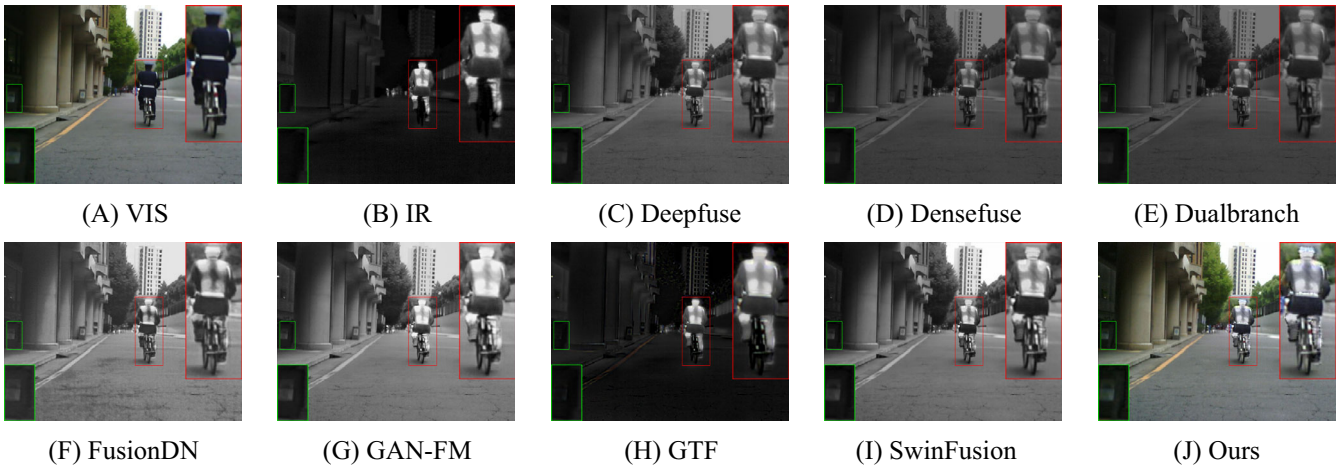

FIGURE 5 Qualitative comparison results on the MSRS dataset.

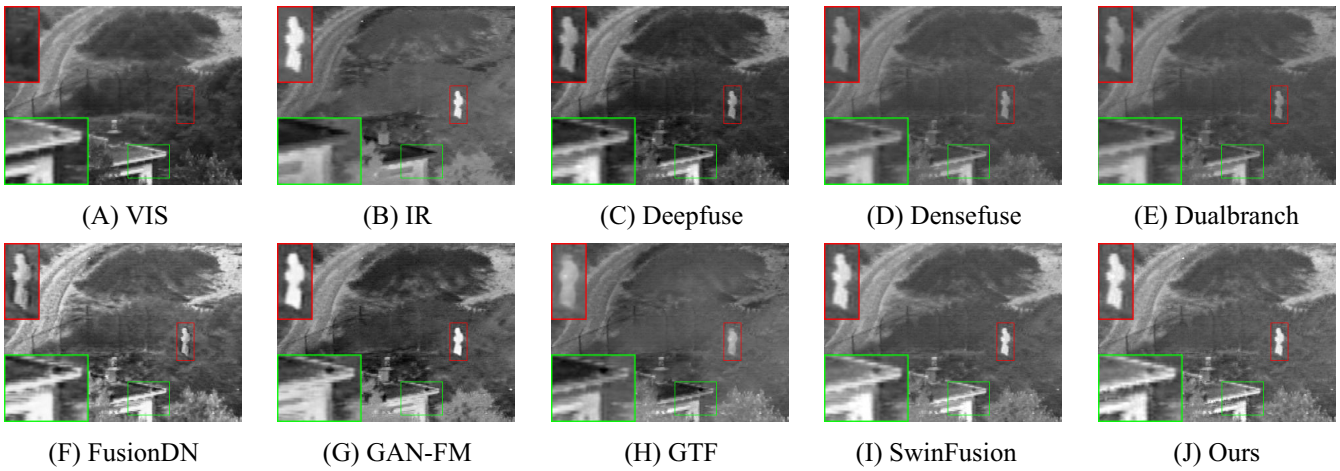

FIGURE 6 Qualitative comparison results on the TNO dataset.

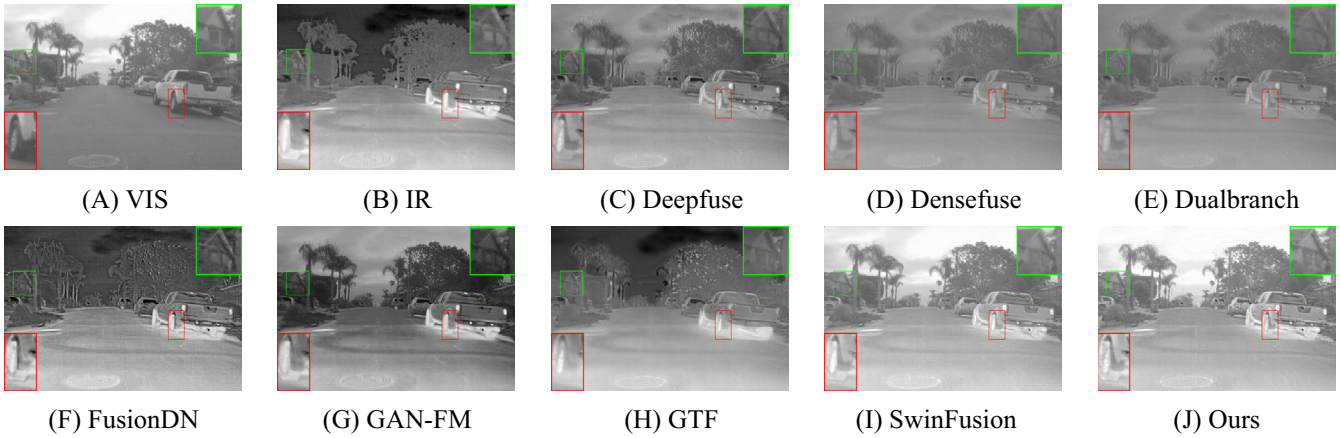

FIGURE 7 Qualitative comparison results on the RoadSence dataset.

can retain more intense information with the person in the image being more clear and more natural. Meanwhile, the edge information and texture detail are well preserved.

Figure 7 illuminates a scenario from the RoadScene dataset in which cars and puddles are thermal targets with strong intensity in the infrared image, while manhole covers, trees, and house windows are textured details in the visible image. The subjective comparison result proves that only the GAN-FM,[33] SwinFusion,[35] and MAIANet preserve full intensity information, while Deepfuse,[22] Densefuse,[8] Dualbranch,[32] FusionDN,[26] and GTF[34] have diminished performance. In terms of visual perception, the proposed method is more in alignment with human visual perception and more natural. Besides, the MAIANet performs well in the reconstruction of textured areas (the textured details of the window in the red contour and the tree in the green box). However, the Densefuse,[8] Dualbranch,[32] and FusionDN[26] miss a lot of edge and detail information.

4.4 | Quantitative evaluation

For quantitative evaluation, six metrics, that is, Edge Intensity (EI),[36] spatial frequency (SF),[37] Qabf,[38] FMI_pixel,[39] Q_E[40] and average gradient (AG),[41] are employed to evaluate the fusion performance of the MAIANet and the competitors. Among them, EI represents the contrast intensity of adjacent pixels. SF denotes the degree of mutation in the image. Qabf measures the amount of edge information transferred from source images to the fusion result. FMI can describe the amount of feature information transferred and preserved in the fused image. Q_E considers the edge information of the human visual system. AG quantifies the gradient information of the fused image to measure the detail and texture information of the fused image. Quantitative comparison experiments are conducted on 360, 21, and 40 image pairs of the MSRS, TNO, and RoadScene datasets. The comparative results are averaged as the final results which are presented in Tables 5, 6, and 7, respectively.

TABLE 5 Quantitative comparison results on the MSRS dataset.

Methods	EI ↑	SF ↑	Qabf ↑	FMI_pixel ↑	Q_E ↑	AG ↑
Deepfuse[22]	26.5760	7.3778	0.4967	0.923123	0.5624	2.5058
Densefuse[8]	21.7100	6.0216	0.3678	0.923126	0.3824	2.0552
Dualbranch[32]	21.5734	5.9915	0.3491	0.3622	0.3622	2.0329
FusionDN[26]	52.1984	13.7831	0.4292	0.9003	0.5773	4.9097
GAN-FM[33]	42.5906	11.6002	0.5376	0.9071	0.6119	4.0157
GTF[34]	24.2052	7.6823	0.3458	0.9048	0.3027	2.3052
SwinFusion[35]	38.0957	11.0911	0.6545	0.9306	0.8479	3.5709
Ours	42.7669	12.0459	0.6585	0.9288	0.8709	4.0418

Note: The best, second-best, and third-best results are marked in red, blue, and green, respectively.

TABLE 6 Quantitative comparison results on the TNO dataset.

Methods	EI ↑	SF ↑	Qabf ↑	FMI_pixel ↑	Q_E ↑	AG ↑
Deepfuse[22]	34.7373	8.9177	0.4967	0.9047	0.5042	3.5129
Densefuse[8]	23.3069	6.0417	0.3678	0.9086	0.3098	2.3534
Dualbranch[32]	23.3851	5.7463	0.3491	0.9079	0.3002	2.3162
FusionDN[26]	53.6711	12.8069	0.4292	0.8803	0.4171	5.2768
GAN-FM[33]	44.4676	12.0983	0.5376	0.8860	0.4353	4.6080
GTF[34]	32.5279	9.2044	0.3458	0.9047	0.3521	3.3587
SwinFusion[35]	38.8169	11.9617	0.6545	0.9117	0.5502	3.9276
Ours	44.8592	11.9617	0.6585	0.9169	0.7039	4.5583

Note: The best, second-best, and third-best results are marked in red, blue, and green, respectively.

TABLE 7 Quantitative comparison results on the RoadSence dataset.

Methods	EI ↑	SF ↑	Qabf ↑	FMI_pixel ↑	Q_E ↑	AG ↑
Deepfuse[22]	46.5114	11.2044	0.4827	0.86312	0.4376	4.4297
Densefuse[8]	35.2111	8.4654	0.3802	0.86318	0.2879	3.3600
Dualbranch[32]	35.3197	8.3459	0.3760	0.8626	0.2843	3.3623
FusionDN[26]	66.4588	15.5675	0.4808	0.8519	0.4696	6.2816
GAN-FM[33]	59.1052	15.0467	0.4600	0.8590	0.4186	5.6422
GTF[34]	35.4243	9.4159	0.3480	0.8710	0.1849	3.3632
SwinFusion[35]	47.7495	12.0794	0.4571	0.8573	0.3734	4.4535
Ours	65.6589	17.7281	0.5252	0.8696	0.6361	6.4276

Note: The best, second-best, and third-best results are marked in red, blue, and green, respectively.

Table 5 illustrates that the proposed MAIANet performs best in the Qabf and Q_E indicators on the MSRS dataset. The MAIANet ranked second in EI, SF, FMI_pixel, and AG. Especially, compared with the second-best method of SwinFusion,[35] the proposed method improves Qabf and Q_E by 0.6%, and 2.7%, respectively. Meanwhile, the proposed method improves the EI, SF, and AG by 0.4%, 3.8%, and 0.6% compared with the GAN-FM[33] respectively. Specifically, compared with the SwinFusion[35] which is a CNN-based method, the MAIANet improves the EI, SF, Qabf, and AG by 12.3%, 8.6%, 0.6%, 2.7%, and 13.1%.

Table 6 shows that the proposed MAIANet ranks first in Qabf, FMI_pixel, and Q_E on the TNO datasets. Meanwhile, the MAIANet ranks second place in EI and third place in both SF and AG. Particularly, compared with Deepfuse[22] which is an autoencoder-based method, the proposed method improves the EI, SF, Qabf, FMI_pixel, Q_E and AG improved by 29.1%, 34.1%, 32.6%, 1.3%, 39.6% and 29.8%, respectively. Compared with SwinFusion,[35] which is a CNN-based method, the proposed method improves the EI, Qabf, FMI_pixel, Q_E, and AG by 15.6%, 0.6%, 0.5%, 27.9%, and 16.7%, respectively.

Table 7 demonstrates that the MAIANet performs best in SF, Qabf, Q_E, and AG on the RoadSence dataset. Also, it ranks second place in EI and FMI_pixel. Particularly, compared to the second-best method FusionDN,[26] the MAIANet improves SF, Q_E, and AG by 13.9%, 35.5% and 2.3%, respectively. Meanwhile, the proposed MAIANet improves the QABF by 8.8% compared to the second-ranked Deepfuse.[22] In addition, compared with SwinFusion,[35] the proposed method improves the EI, SF, Qabf, FMI_pixel, Q_E and AG improved by 37.5%, 46.8%, 14.9%, 1.4%, 0.7%, and 44.3%, respectively.

4.5 | Ablation study

To further investigate the effectiveness of different components of MAIANet, we conduct ablation studies on the TNO dataset. The ablation studies are divided into two aspects, that is, the effectiveness of the multiscale extraction unit and the effectiveness of the lightweight channel attention module. The detailed network configurations are depicted as follows.

TABLE 8 Ablation analysis on the key components in MAIANet.

Methods	EI ↑	SF ↑	Qabf ↑	FMI_pixel ↑	Q_E ↑	AG ↑
Baseline	42.3462	11.0184	0.5756	0.9166	0.6630	4.2651
Baseline + ME	43.8706	11.5155	0.5808	**0.9185**	**0.7254**	4.4513
Baseline + LCA	43.1541	11.2222	0.5657	0.9173	0.6954	4.3707
Baseline + ME-LCA	**44.8592**	**11.9617**	**0.6585**	0.9169	0.7039	**4.5583**

Note: The best results are marked in **bold**.

1. "baseline" refers to the vanilla model without any component.
2. "baseline + ME" denotes the baseline model with single ME unit.
3. "baseline + LCA" represents the baseline model with single LCA module.
4. "baseline + ME-LCA" refers to the baseline model with the ME unit and LCA module sequentially connected.

Comparison results are shown in Table 8. It proves that the ME unit and LCA module contribute to substantial improvements in the baseline approach. As shown in Table 8, the "baseline" method achieves the worst performance. Compared with the "baseline" method, the "baseline + ME", "baseline + LCA", and "baseline + ME-LCA" methods synergy multiscale information and channel attention. Therefore, the latter is better than the former. Specifically, compared to the "baseline" method, the "baseline + ME" and "baseline + LCA" methods improve EI by 3.6%, 1.9%, and 5.9% respectively. In addition, it can be seen that the addition of the ME unit and the LCA module are helpful to enhance the SF, Qabf, FMI_pixel, Q_E, and AG. Specifically, the "baseline + ME" method achieves 4.5%, 0.9%, 0.2%, 9.4%, and 4.4% improvement in SF, Qabf, FMI_pixel, Q_E, and AG, respectively. The "baseline + LCA" method obtains 1.8%, 0.07%, 4.9%, and 2.5% enhancement in SF, FMI_pixel, Q_E, and AG, respectively. The "baseline + ME-LCA" method performs best in EI, SF, Qabf, and AG. Particularly, compared with the "baseline" method, the "baseline + ME-LCA" method improves EI, SF, Qabf, FMI_pixel, Q_E, and AG by 5.9%, 8.6%, 14.4%, 0.003%, 6.2%, and 6.9%, respectively.

5 | CONCLUSION

A multiscale aggregation and illumination-aware attention network (MAIANet) is proposed for the IVIF task. The proposed MAIANet utilizes the multiscale feature extraction module to extract multiscale information from infrared and visible images. In addition, the suggested method constructs an illumination-aware module to evaluate the illumination probability and distribution. An illumination-aware loss function is developed to assist the network in adjusting to differences in illumination. Meanwhile, a lightweight attention module is employed to focus on the salient targets and texture details of the infrared and visible images. The proposed fusion network is verified with state-of-the-art techniques on three challenging infrared and visible datasets in both qualitative and quantitative evaluation. Comparison results prove that the proposed strategies can preserve the intensity information of the infrared image and the detailed information of the visible image under various illumination circumstances. In the future, more efforts are expected to detail preserving, as the visible images are susceptible to loss of detail information and are highly susceptible to interference.

ACKNOWLEDGMENTS

This work is supported in part by the Nature Science Foundation of China (Nos. 61601266 and 61801272) and National Natural Science Foundation of Shandong Province (Nos. ZR2021QD041 and ZR2020MF127).

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no conflict of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available Baidu Netdisk at https://pan.baidu.com/s/1ErwoKxhlEKQKHeyrPKBA7g (PASSWORD: 7412). Data sharing is not applicable to this article as no new data were created or analyzed in this study.

ORCID

Mingliang Gao https://orcid.org/0000-0001-7273-7499
Qilei Li https://orcid.org/0000-0002-9675-9016
Gwanggil Jeon https://orcid.org/0000-0001-8848-3243

REFERENCES

1. Liu S, Gao M, John V, Liu Z, Blasch E. Deep learning thermal image translation for night vision perception. *ACM Trans Intell Syst Technol*. 2020;12(1):1-18.
2. Zhang H, Xu H, Tian X, Jiang J, Ma J. Image fusion meets deep learning: a survey and perspective. *Inf Fusion*. 2021;76:323-336.
3. Ma J, Ma Y, Li C. Infrared and visible image fusion methods and applications: a survey. *Inf Fusion*. 2019;45:153-178.
4. 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;83:79-92.
5. Huang Y, Li W, Gao M, Liu Z. Algebraic multi-grid based multi-focus image fusion using watershed algorithm. *IEEE Access*. 2018;6:47082-47091.
6. Li Q, Wu W, Lu L, Li Z, Ahmad A, Jeon G. Infrared and visible images fusion by using sparse representation and guided filter. *J Intell Transp Syst*. 2020;24(3):254-263.
7. Azam MA, Khan KB, Salahuddin S, et al. A review on multimodal medical image fusion: compendious analysis of medical modalities, multimodal databases, fusion techniques and quality metrics. *Comput Biol Med*. 2022;144:105253.
8. Li H, Wu XJ. DenseFuse: a fusion approach to infrared and visible images. *IEEE Trans Image Process*. 2018;28(5):2614-2623.
9. 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. 34 of Proceedings of the AAAI Conference on Artificial Intelligence. AAAI. 2020:12797-12804.
10. Wang K, Zheng M, Wei H, Qi G, Li Y. Multi-modality medical image fusion using convolutional neural network and contrast pyramid. *Sensors*. 2020;20(8):2169.
11. Ma J, Yu W, Liang P, Li C, Jiang J. FusionGAN: a generative adversarial network for infrared and visible image fusion. *Inf Fusion*. 2019;48:11-26.
12. Chen J, Zhang L, Lu L, Li Q, Hu M, Yang X. A novel medical image fusion method based on rolling guidance filtering. *Internet Things*. 2021;14:100172.
13. Bhat S, Koundal D. Multi-focus image fusion using neutrosophic based wavelet transform. *Appl Soft Comput*. 2021;106:107307.
14. Mei Q, Li M. Nonsubsampled contourlet transform and adaptive PCNN for medical image fusion. *J Appl Sci Eng*. 2022;26(2):213-220.
15. Han J, Pauwels EJ, De Zeeuw P. Fast saliency-aware multi-modality image fusion. *Neurocomput*. 2013;111:70-80.
16. Meng F, Guo B, Song M, Zhang X. Image fusion with saliency map and interest points. *Neurocomput*. 2016;177:1-8.
17. 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;82:28-42.
18. Chen J, Yang X, Lu L, Li Q, Li Z, Wu W. A novel infrared image enhancement based on correlation measurement of visible image for urban traffic surveillance systems. *J Intell Transp Syst*. 2020;24(3):290-303.
19. Huang D, Tang Y, Wang Q. An image fusion method of SAR and multispectral images based on non-subsampled shearlet transform and activity measure. *Sensors*. 2022;22(18):7055.
20. Zhang H, Shen H, Yuan Q, Guan X. Multispectral and SAR image fusion based on laplacian pyramid and sparse representation. *Remote Sens*. 2022;14(4):870.
21. Jian L, Yang X, Liu Z, Jeon G, Gao M, Chisholm D. SEDRFuse: a symmetric encoder–decoder with residual block network for infrared and visible image fusion. *IEEE Trans Instrum Meas*. 2021;70:1-15.
22. Ram Prabhakar K, Sai Srikar V, Venkatesh Babu R. Deepfuse: a deep unsupervised approach for exposure fusion with extreme exposure image pairs. Proceedings of the IEEE International Conference on Computer Vision, IEEE. 2017:4714-4722.
23. Li H, Wu XJ, Kittler J. RFN-Nest: an end-to-end residual fusion network for infrared and visible images. *Inf Fusion*. 2021;73:72-86.
24. Ma B, Yin X, Wu D, Shen H, Ban X, Wang Y. End-to-end learning for simultaneously generating decision map and multi-focus image fusion result. *Neurocomput*. 2022;470:204-216.
25. Liu Y, Chen X, Cheng J, Peng H. A medical image fusion method based on convolutional neural networks. Paper presented at: 2017 20th International Conference on Information Fusion (Fusion), IEEE. 2017:1-7.
26. Xu H, Ma J, Le Z, Jiang J, Guo X. Fusiondn: a unified densely connected network for image fusion. 34 of Proceedings of the AAAI Conference on Artificial Intelligence, AAAI. 2020:12484-12491.
27. Wang X, Jiang J, Gao M, Liu Z, Zhao C. Activation ensemble generative adversarial network transfer learning for image classification. *J Electron Imag*. 2021;30(1):013016.
28. Li Q, Lu L, Li Z, et al. Coupled GAN with relativistic discriminators for infrared and visible images fusion. *IEEE Sens J*. 2019;21(6):7458-7467.
29. Ma J, Xu H, Jiang J, Mei X, Zhang XP. DDcGAN: a dual-discriminator conditional generative adversarial network for multi-resolution image fusion. *IEEE Trans Image Process*. 2020;29:4980-4995.
30. 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*. 2020;70:1-15.
31. Toet A. The TNO multiband image data collection. *Data Brief*. 2017;15:249-251.
32. Fu Y, Wu XJ. A dual-branch network for infrared and visible image fusion. Paper presented at: 2020 25th International Conference on Pattern Recognition (ICPR), IEEE. 2021:10675-10680.
33. Zhang H, Yuan J, Tian X, Ma J. GAN-FM: infrared and visible image fusion using GAN with full-scale skip connection and dual Markovian discriminators. *IEEE Trans Comput Imag*. 2021;7:1134-1147.
34. Ma J, Chen C, Li C, Huang J. Infrared and visible image fusion via gradient transfer and total variation minimization. *Inf Fusion*. 2016;31:100-109.
35. 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/CAA J Automat Sin*. 2022;9(7):1200-1217.
36. Xydeas CS, Pv V. Objective image fusion performance measure. *Mil Tech Cour*. 2000;56(4):181-193.
37. Eskicioglu AM, Fisher PS. Image quality measures and their performance. *IEEE Trans Commun*. 1995;43(12):2959-2965.
38. Qu G, Zhang D, Yan P. Information measure for performance of image fusion. *Electron Lett*. 2002;38(7):313-315.
39. Haghighat M, Razian MA. Fast-FMI: non-reference image fusion metric. Paper presented at: 2014 IEEE 8th International Conference on Application of Information and Communication Technologies (AICT), IEEE. 2014:1-3.

40. Piella G, Heijmans H. A new quality metric for image fusion. 3 of Proceedings 2003 international conference on image processing (Cat. No. 03CH37429), IEEE. 2003:3-173.
41. Cui G, Feng H, Xu Z, Li Q, Chen Y. Detail preserved fusion of visible and infrared images using regional saliency extraction and multi-scale image decomposition. *Optics Commun*. 2015;341:199-209.

How to cite this article: Song W, Zhai W, Gao M, Li Q, Chehri A, Jeon G. Multiscale aggregation and illumination-aware attention network for infrared and visible image fusion. *Concurrency Computat Pract Exper*. 2023;e7712. doi: 10.1002/cpe.7712