Complex & Intelligent Systems (2024) 10:705–719
https://doi.org/10.1007/s40747-023-01185-2

ORIGINAL ARTICLE

Multi-scale attention-based lightweight network with dilated convolutions for infrared and visible image fusion

Fuquan Li[1] · Yonghui Zhou[1] · YanLi Chen[1] · Jie Li[2] · ZhiCheng Dong[3] · Mian Tan[4]

Received: 26 April 2023 / Accepted: 15 July 2023 / Published online: 3 August 2023

Abstract

Infrared and visible image fusion aims to generate synthetic images including salient targets and abundant texture details. However, traditional techniques and recent deep learning-based approaches have faced challenges in preserving prominent structures and fine-grained features. In this study, we propose a lightweight infrared and visible image fusion network utilizing multi-scale attention modules and hybrid dilated convolutional blocks to preserve significant structural features and fine-grained textural details. First, we design a hybrid dilated convolutional block with different dilation rates that enable the extraction of prominent structure features by enlarging the receptive field in the fusion network. Compared with other deep learning methods, our method can obtain more high-level semantic information without piling up a large number of convolutional blocks, effectively improving the ability of feature representation. Second, distinct attention modules are designed to integrate into different layers of the network to fully exploit contextual information of the source images, and we leverage the total loss to guide the fusion process to focus on vital regions and compensate for missing information. Extensive qualitative and quantitative experiments demonstrate the superiority of our proposed method over state-of-the-art methods in both visual effects and evaluation metrics. The experimental results on public datasets show that our method can improve the entropy (EN) by 4.80%, standard deviation (SD) by 3.97%, correlation coefficient (CC) by 1.86%, correlations of differences (SCD) by 9.98%, and multi-scale structural similarity (MS_SSIM) by 5.64%, respectively. In addition, experiments with the VIFB dataset further indicate that our approach outperforms other comparable models.

Keywords Image fusion · Attention model · Dilated convolution · Infrared image · Visible image

Introduction

Images collected by a single mode sensor fail to effectively and comprehensively describe imaging scenes due to theoretical and technical limitations [1]. Infrared sensors capture thermal radiation emitted by objects and can generate infrared images with significant targets, even in adverse conditions such as low brightness, occlusions, or harsh weather. However, infrared images are susceptible to noise and lack textural details. In contrast, visible images offer abundant texture and structural information but are susceptible to imaging conditions. As such, infrared and visible image fusion tasks involve reconstructing a single image with comprehensive information from multimodal data, providing both significant targets and valuable texture information. Motivated by variations in imaging scenes, several excellent fusion algorithms have been proposed for broad applications in various advanced vision tasks, including object detection [2], semantic segmentation [3], pedestrian re-recognition [4], and visual tracking [5].

In recent years, the fusion of infrared and visible images has attracted the attention of many scholars and has developed rapidly as a result. Existing technologies can be categorized into two groups: traditional methods [6, 7] and deep learning-based methods [8–12]. Traditional image fusion algorithms are typically implemented using multi-

✉ YanLi Chen
yanli_027@163.com

[1] School of Big Data and Computer Science, Guizhou Normal University, Guizhou, China

[2] School of Intelligent Technology and Engineering, Chongqing University of Science and Technology, Chongqing, China

[3] School of Information Science and Technology, Tibet University, Lhasa, China

[4] Guizhou Key Laboratory of Pattern Recognition and Intelligent System, Guizhou Minzu University, Guizhou, China

scale transform (MST)-based methods [13], sparse representation (SR)-based methods [14], low-rank representation (LRR) [15], saliency-based methods [16], subspace-based methods [17], and other methods [18]. Although traditional methods have shown superior fusion performance in some aspects, they are also known to encounter specific challenges. (1) They generally require manually selected feature representations and accurate fusion rules when generating high-quality fused images, which require manual intervention and can degrade fusion performance. (2) In the case of SR and LRR techniques, it can be difficult to construct a suitable overcomplete dictionary. The runtime for corresponding fusion algorithms is, thus, not conducive to real-time image fusion. (3) Complex feature extraction and fusion strategies often introduce halos and blurred edges, due to the overlapping of asymmetric feature information.

To address these issues, deep learning-based methods have been introduced for infrared and visible image fusion. These frameworks can typically be divided into three categories: auto-encoder (AE) [8, 9], convolutional neural network (CNN) [19], and generative adversarial network (GAN) based architectures [20]. Deep learning offers several advantages for improved representation capabilities, but with certain limitations. First, to reduce the complexity of the network, introducing a down-sampling operation to reduce the image resolution inevitably results in the loss of important information in the fusion image. On the other hand, modern convolutional networks are not shift-invariant[21], as small shifts or translations in the input cause substantial changes in the output. Second, some methods use only simple feature fusion rules, such as addition and connection, which can cause artifacts or blurred edges in fused images. Third, the existing infrared and visible images used for training and testing are mainly derived from the TNO [22] and RoadScene [23] datasets, which restricts the comprehensive evaluation of the model's generalization performance. FusionGAN [24] crops the source images into image patches by setting the stride to 14, while lacking global information to learn over long distances and failing to handle complex scenes.

A novel deep learning architecture for the fusion of infrared and visible images is proposed in this paper to address the issues discussed above.

Inspired by previous traditional multi-scale frameworks, we design an encoder network, consisting of hybrid dilated convolutional blocks used to obtain multi-scale depth salient features by implementing different dilation rates. It is worth noting that no down-sampling is used during feature extraction, so the resulting feature map is the same size as the source images. In addition, to make full use of multi-scale layer characteristics, we introduce different attention modules for each scale, to ensure the network pays attention to specific features and compensates for information loss. To demonstrate the effectiveness of our approach, a representative fusion sample is shown in Fig. 1 and compared with three other excellent deep learning-based algorithms. Our method not only produces higher image contrast (e.g., the person in the infrared image is brighter using our technique), but also improves visual effects (e.g., smoke is preserved in the visible images, and trees in the background exhibit clearer edges). The primary contributions of our work can be summarized as follows:

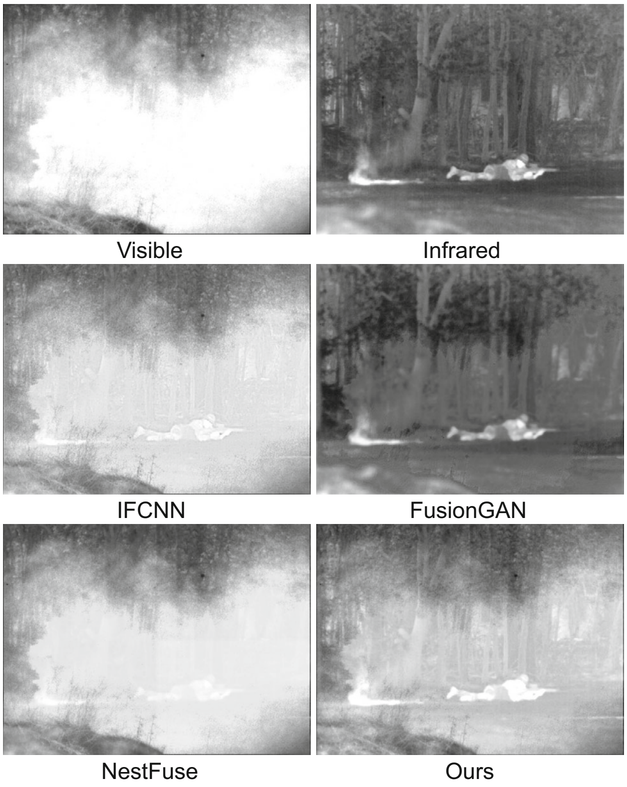


Fig. 1 A schematic illustration of the proposed method. The first row displays the source images, the second row presents fused images with IFCNN [19] and FusionGAN [24], and the last row provides results produced by NestFuse [8] and our proposed method

- We propose a lightweight network architecture for infrared and visible image fusion, which can capture fine-grained detailed features with a high semantic level and does not require a down-sampling operation.
- Both spatial attention and channel attention mechanisms are introduced in the encoder-decoder framework at different scales. The proposed method not only forces the network to focus on foreground targets of the infrared image and the background information in the visible image, it also enhances local and global contextual information and attenuates noise.
- A total loss function is designed to jointly focus on pixel distribution information and texture details in both

infrared and visible images, to preserve essential complementary information in each modality.
- Extensive experiments demonstrate our method's superiority over state-of-the-art methods. The experimental results on public datasets reveal that our method achieves significant enhancements in entropy (EN) by 4.80%, standard deviation (SD) by 3.97%, correlation coefficient (CC) by 1.86%, correlations of differences (SCD) by 9.98%, and multi-scale structural similarity (MS_SSIM) by 5.64%.

Related work

Traditional image fusion methods

Traditional image fusion algorithms can be divided into three steps: feature extraction, fusion, and reconstruction. The feature extraction and reconstruction steps are typically opposite operations. Several multi-scale techniques such as Gaussian pyramid [7], shearlet [25], and nonsubsampled contourlet [26] transforms have been proposed in the past few decades, some of which are utilized in deep learning-based fusion frameworks. In addition, feature extraction methods used for sparse representations include joint sparse representation [14] and latent low-rank representations [27]. Inspired by human visual perception, this process requires an overcomplete dictionary. As such, the computational complexity of sparse representations has always been an issue. In addition, by reducing the dimensionality of the original features into low-dimensionality of features that are independent of each other, representative techniques can be developed using subspace feature extraction, including independent component analysis [28], principal component analysis [29], and non-negative matrix factorization [30].

Deep learning-based fusion methods

Convolutional neural networks can learn prior knowledge from large image quantities and have been widely used for image fusion and other related tasks. Image fusion methods based on deep learning include AE-based algorithms, convolutional neural networks, and GAN-based image fusion models. Liu et al. [31] first proposed a CNN-based fusion framework. Since the purpose of the network is to generate a decision map, this approach is only suitable for multi-focus images. Li et al. [8] proposed a fusion method of nest connection-based architecture comprised of three parts: encoder network, fusion strategy, and decoder network, which extract deep features at different scales. This feature fusion is manually supervised by rules that affect fusion performance to a certain extent. Later, residual end-to-end auto-encoder fusion networks have been proposed to overcome the issue [9].

In addition, by forcing the network to focus on intensity distribution and texture structures in images, infrared and visible image fusion algorithms based on the end-to-end convolutional neural network provide a solution to this problem. For example, Ma et al. [32] used salient mask to force the network on texture details in visible images and salient information in infrared images. However, it can be difficult to provide ground truth data to the network for image fusion tasks. Considering extreme illumination conditions for source images, Tang et al. [33] introduced an illumination-aware sub-network that maintains intensity distributions in salient targets and preserves texture information in the background. Furthermore, to facilitate advanced visual tasks, this group introduced semantic segmentation into the image fusion module to improve the semantic information in the fused images. They also proposed a joint low-level and high-level adaptive training strategy to simultaneously achieve superior performance and close the gap in both image fusion and high-level vision tasks [34].

In 2019, Ma et al. [24] first introduced the generative adversarial networks into the field of infrared and visible image fusion. Specifically, content loss and adversarial loss are employed to preserve details of thermal radiation in the fused images generated from connected source images. However, a single discriminator cannot focus on both infrared and visible regions. As such, Li et al. [35] not only introduced a dual-discriminator conditional generative adversarial network, but also used a multi-scale attention mechanism to constrain the discriminator and focus more on regions of interest, to balance the data distribution and improve fused image fidelity.

Dilated convolutional and attention mechanism applications

Dilated convolution, inspired by wavelet decomposition, enhances the receptive field of a convolutional kernel by inserting zeros between its pixels. This expansion aids the network in capturing detailed information within the scene. Dilated convolution has been widely applied in image classification, object detection, and semantic segmentation. Yu et al. [36] addressed the issue of gridding artifacts introduced by dilation by designing dilated residual networks, which can be effectively employed in downstream tasks such as object localization and semantic segmentation.

The attention mechanism, motivated by the human visual system, has been successfully incorporated into computer vision systems such as image recognition, object detection, semantic segmentation, and action recognition [37]. Channel attention focuses on important objects by assigning new weights to the channels of the feature map. Hu et

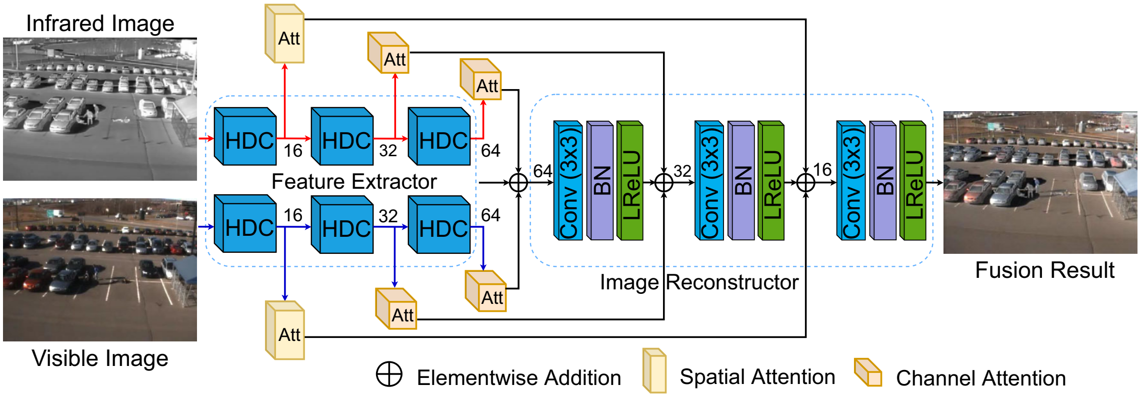


Fig. 2 The overall framework for the infrared and visible image fusion algorithm based on HDC blocks and different attention mechanisms

al. [38] first proposed the concept of channel attention, known as SENet. The core squeeze-and-excitation (SE) block of SENet effectively captures the channel-wise relationship, thereby enhancing the representation capability of the network model. Qin et al. [39] demonstrated that global average pooling can be viewed as a special case of the discrete cosine transform and designed a multi-spectral channel attention mechanism to further enhance the model's representation capabilities. Spatial attention, on the other hand, can be seen as the adaptive selection of important spatial regions. Hu et al. [40] designed GENet to capture long-distance spatial contextual information in feature maps, enabling the highlighting of important features while suppressing noise. Building upon the success of self-attention in natural language processing, Wang et al. [41] proposed Non-Local networks that expand the receptive fields of the network, enabling the capture of global information. In the context of image fusion, Ma et al. [42] introduced Swin Transformer and proposed intra-domain and inter-domain fusion units based on self-attention and cross-attention, respectively. This approach achieves the integration of complementary information and captures global long-range dependencies, facilitating the effective fusion of multi-domain images.

Methodology

This section describes the proposed lightweight infrared and visible image fusion network architecture in detail. First, we present the overall network pipeline. Hybrid dilated convolutional (HDC) blocks and multi-scale spatial/channel attention are then introduced. Finally, the proposed loss function is discussed.

Problem formulation

Given a pair of registered infrared $I_{ir} \in R^{H \times W \times 1}$ and visible images $I_{vis} \in R^{H \times W \times 3}$, under the guidance of a total loss function, the fused image $I_f \in R^{H \times W \times 3}$ can be generated by feature extraction, feature fusion, and reconstruction. The previous deep learning methods emphasized the importance of feature extraction on the quality of fusion results, which led to designing complex feature extractors. However, the real-time image fusion requirement was ignored. In order to improve the ability of feature representation, while ensuring real-time infrared and visible image fusion, key design components for the lightweight HDC blocks and multi-scale attention mechanisms are designed to produce high-quality fused images and prevent artifacts. (we will discuss its network architecture in Section "Network architecture"). The overall framework for our proposed infrared and visible image fusion algorithm is shown in Fig. 2.

First, a fusion network based on HDC blocks is devised to fully extract the high-level semantic information in source images. More specifically, we apply a feature extraction module F_E to extract fine-grained feature information from infrared and visible images. This process can be represented as:

$$\{F_{ir}, F_{vis}\} = \{F_E\left(I_{ir}\right), F_E\left(I_{vis}\right)\}, \tag{1}$$

where F_{ir} and F_{vis} represent feature maps for infrared and visible images, respectively. Moreover, HDC blocks are deployed in the feature extraction module to expand the receptive field while ensuring that important coarse-grained and fine-grained feature information is extracted, as shown in Fig. 3. Given the HDC input F_i, the corresponding output F_{i+1} can be represented as:

$$F_{i+1} = HDC\left(F_i\right) = \phi\left(DConv^n\left(F_i\right)\right), \tag{2}$$

where $DConv^n$ is an n-cascaded 3 × 3 dilated convolutional layer and ϕ represents the LReLU activation function. Information flow is processed in HDC blocks using respective hierarchical levels in the pipeline. In this paper, HDC blocks

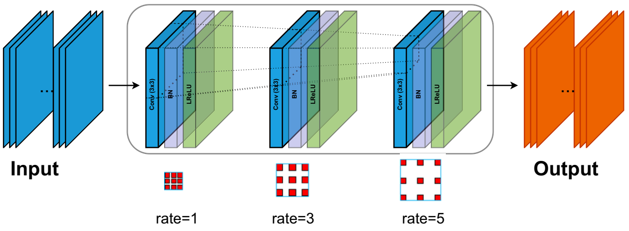


Fig. 3 The specific arrangement of the hybrid dilated convolutional blocks. Lift to right: convolutional layers with a kernel size of 3 × 3 and dilation rates of 1, 3, and 5, respectively. The HDC clocks naturally enlarge the network receptive field without adding extra modules

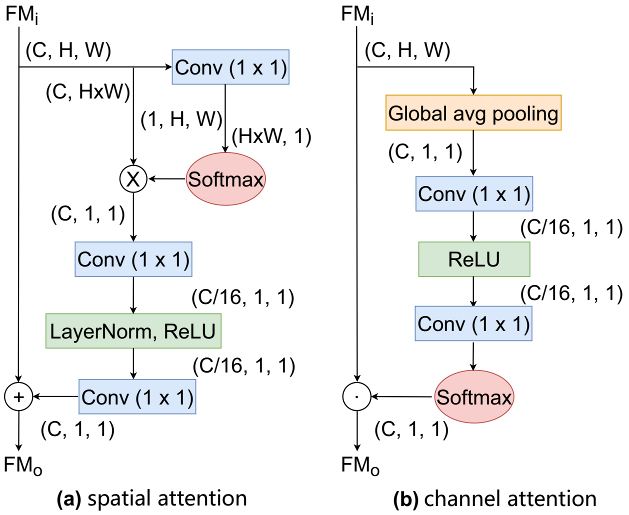


Fig. 4 A diagram of the spatial and channel attention-based modules is shown. $C \times H \times W$ denotes feature maps with channel number C, height H and width W. $\otimes$ denotes matrix multiplication, $\oplus$ represents element-wise addition, and $\odot$ indicates element-wise multiplication

capture local and global information of the source image to effectively facilitate feature representation capabilities.

The feature fusion and reconstruction module is responsible for converting the feature maps into the fused image. However, simply reconstructing the fused image using convolution operations may result in information loss. Therefore, we introduce different attention modules at different layers of the extractor to fully exploit contextual information from the source images and alleviate the information loss of the feature maps in reconstruction.

To integrate the abundant fine-grained detailed features in infrared and visible images and reconstruct the fused image, the element-wise addition strategy in [43] is used. The formula for this fusion process is as follows:

$$F_f = \text{Add}\left(\alpha_i\left(F_{ir}\right), \alpha_i\left(F_{vis}\right)\right), \tag{3}$$

where F_f is fused feature maps, $Add(\cdot, \cdot)$ represents an element-wise addition strategy, and α_i denotes an attention mechanism corresponding to multiple scales. Specifically, α_1 is employed to focus on coarse-grained information from infrared and visible images using a spatial attention mechanism. Both α_2 and α_3 are devoted to strengthening a large amount of fine-grained feature information using a channel attention mechanism. Finally, the fused image I_f is reconstructed from F_f via an image reconstructor R_i as follows:

$$I_f = R_i\left(F_f\right). \tag{4}$$

Network architecture

The framework for the proposed lightweight fusion network based on hybrid dilated convolutional blocks (HDCBs), shown in Fig. 2, consists of encoder and decoder networks for feature extraction and image reconstruction, respectively.

The feature extractor utilizes three HDCBs to increase the size of the receptive field in the network and capture more contextual information, while ensuring fine-grained features are extracted from infrared and visible images. In addition, a multi-scale spatial/channel attention module is also proposed to retain valuable information and reduce artifacts in multi-modality images. In the feature extractor, a multi-scale shallow layer in the encoder focuses on the elemental features using a spatial attention module, while a channel attention module is used to pay attention to fine-grained features in source images on multi-scale deep layers of the encoder. These multi-scale attention features are added as inputs to corresponding layer features of the decoder network to reconstruct the fused image. As shown in Fig. 2, two parallel encoder modules are used to extract features from infrared and visible images containing three HDCBs with dilation rates of 1, 3, and 5, respectively. The special design of the HDCB is shown in Fig. 3. The block mainly changes the dilation rates of ordinary convolutions, which is set to prevent the occurrence of gridding problems. The mainstream applies to three convolutional layers with a kernel size of 3 × 3 and stride of 1, the batch normalization (BN) layers, and the LReLU layers. To preserve more diverse and important contextual information, the different attention modules are introduced to each scaling layer of the encoder, as shown in Fig. 4. The FM_i serves as the input to the attention module, acquired from the feature maps of each HDCB output in the encoder, while the FM_o provides the output of the attention module. The spatial attention mechanism is used by shallow features of the first HDCB, while the channel attention mechanism is exploited in the deep scaling layers.

Attention maps for infrared and visible images at different scales are then integrated via an element-wise addition strategy, and the results are fed into the decoder network to achieve image reconstruction. The decoder network in the image reconstructor generates fused images using three 3 × 3 convolutional layers and three BN layers, all of which are followed by an LReLU activation function. The stride is set to

1 in the fused network with no down-sampling operation, to reduce information loss. As such, fused images are the same size as the source images.

Loss function

A total loss function is proposed in this study to facilitate more comprehensive detail in the resulting images, obtained from salient target information in infrared images and fine-grained features in visible images. This total loss function consists of intensity loss $L_{intensity}$ and detail loss L_{detail} terms, which is defined as follows:

$$L_{total} = L_{intensity} + \gamma L_{detail}, \tag{5}$$

where γ is a weight factor used to balance the intensity loss $L_{intensity}$ and detail loss L_{detail}.

The intensity loss is designed to constrain intensity similarity between the fused and input images at the pixel level. Therefore, the intensity loss is expressed as:

$$L_{intensity} = \frac{1}{HW} \left\| I_f - (pI_{ir} + (1-p)I_{vis}) \right\|_1, \tag{6}$$

where W and H represent the width and height of the image, respectively, $\|\cdot\|_1$ is the l_1-norm, and p denotes the weight of constraints used to integrate the distribution of pixel intensities in infrared and visible images.

However, fused images not only include the pixel intensity distribution of the source images, but also exhibit a fine-grained detail distribution. Hence, a detail loss is introduced to force the fused image to preserve more structure and fine-grained texture information. Detail loss can be expressed as:

$$L_{detail} = \frac{1}{HW} \left\| \left|\nabla I_f\right| - (q\left|\nabla I_{ir}\right| + (1-q)\left|\nabla I_{vis}\right|) \right\|_1, \tag{7}$$

where ∇ indicates the Sobel gradient operation used to measure the fine-grained information in the source images, q is a weight parameter that constrains the fine-grained features in infrared and visible images, and $|\cdot|$ indicates the absolute value operation.

Finally, guided by the total loss function, our proposed fused network based on HDCBs and multi-scale attention provides fused images with a better pixel intensity distribution and larger quantities of detail information, to efficiently generate high-quality images.

Experiments

In this section, we first describe the experimental settings and training details. Then, we conduct both quantitative and qualitative comparative experiments and generalization experiments to fully evaluate the performance of our proposed fusion algorithm. Finally, we introduce ablation experiments to demonstrate the effectiveness of the model design, including detail loss and multi-scale spatial/channel attention.

Experimental settings

We perform extensive quantitative and qualitative experiments using the TNO [22], RoadScene [23], and VIFB [44] datasets to comprehensively evaluate the proposed fusion method. In addition, seven state-of-the-art image fusion algorithms are selected for comparison with our approach, including three typical traditional methods, i.e., IFEVIP [45], GTF [18] and CBF [46], two AE-based models, i.e., MFEIF [47] and NestFuse [8], one CNN-based method IFCNN [19], and one GAN-based method FusionGAN [24]. Implementations of these algorithms are publicly available and corresponding parameters are set in agreement with those in their respective papers.

Nine statistical evaluation indicators are used to quantitatively evaluate our method and the seven other excellent fusion methods. They are entropy (EN) [48], modified fusion artifacts measure (Nabf) [49], correlations of differences (SCD) [50], spatial frequency (SF) [51], standard deviation (SD) [52], peak signal to noise ratio(PSNR) [53], multi-scale structural similarity (MS_SSIM) [54], feature mutual information (FMI) and correlation coefficient (CC). These values increase as the fusion performance improved (excluding Nabf).

The EN measures the amount of information contained in a fused image as follows:

$$EN = -\sum_{l=0}^{L} p_l \log_2 p_l, \tag{8}$$

where L and p_l represent the total number of gray levels and the normalized histogram of the corresponding gray level in the fused image, respectively. A large EN indicates that a large amount of information is available, representing better fusion performance. Larger EN values may also be caused by noises.

The Nabf, which quantifies the number of noises or artifacts added in the fused image due to the fusion process, can be expressed as:

$$N_m^{\frac{AB}{F}} = \frac{\sum_{\forall i}\sum_{\forall j} AM_{i,j}\left[\left(1-Q_{i,j}^{AF}\right)w_{i,j}^{A} + \left(1-Q_{i,j}^{BF}\right)w_{i,j}^{B}\right]}{\sum_{\forall i}\sum_{\forall j}\left(w_{i,i}^{A} + w_{i,i}^{B}\right)}, \tag{9}$$

$$AM_{i,j} = \begin{cases} 1, \ g_{i,j}^{F} > g_{i,j}^{A} \text{ and } g_{i,j}^{F} > g_{i,j}^{B} \\ 0, \ \text{otherwise} \end{cases}, \tag{10}$$

where $AM_{i,j}$ indicates locations of fusion artifacts when fused gradients are stronger than input, $Q_{i,j}^{AF}$ and $Q_{i,j}^{BF}$ denote the gradient information preservation estimates of source images A and B, respectively, $w_{i,i}^{A}$ and $w_{i,i}^{B}$ are the perceptual weights of source images, respectively, $g_{i,j}^{A}$, $g_{i,j}^{B}$ and $g_{i,j}^{F}$ are the edge strength of A, B, and fused image F, respectively. A low Nabf value is indicative of superior visual performance in the fused image.

The SCD, which measures the amount of information transmitted from source images to the fused image, can be represented as:

$$SCD = r(D_1, S_1) + r(D_2, S_2), \tag{11}$$

where $r(\cdot)$ denotes the correlation function.

The SF metric effectively measures the gradient distribution of images, which reveals the details and texture of images. It can be defined as follows:

$$SF = \sqrt{RF^2 + CF^2}, \tag{12}$$

$$RF = \sqrt{\sum_{i=1}^{M}\sum_{j=1}^{N}(F(i,j) - F(i,j-1))^2}, \tag{13}$$

$$CF = \sqrt{\sum_{i=1}^{M}\sum_{j=1}^{N}(F(i,j) - F(i-1,j))^2}, \tag{14}$$

where RF and CF are the spatial row frequency (RF) and column frequency (CF) based on horizontal and vertical gradients, respectively.

The CC metric measures the degree of linear correlation between the fused image and the source images, as defined below:

$$CC = \frac{r_{af} + r_{bf}}{2}, \tag{15}$$

$$r_{xf} = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N}(x_{i,j} - \mu_x)(f_{i,j} - \mu_f)}{\sqrt{\sum_{i=1}^{M}\sum_{j=1}^{N}(x_{i,j} - \mu_x)^2 \sum_{i=1}^{M}(f_{i,j} - \mu_f)^2}}, \tag{16}$$

where μ_x and μ_f indicate the mean values of the input image x and the fused image f, respectively. A higher value of CC indicates a better correlation and higher image quality for the fused image.

The SD reflects the distribution and contrast of the fused image from a statistical perspective and can be defined mathematically as:

$$SD = \sqrt{\sum_{i=1}^{M}\sum_{j=1}^{N}(f(i,j) - \mu)^2}, \tag{17}$$

where μ denotes the average of the fused image. A positive SD value indicates that the fused image exhibits favorable visual effects.

The MS_SSIM represents a calibration definition for the difference between two images across scales. The corresponding multi-scale SSIM index is given by:

$$SSIM(x, y) = [l_M(x, y)]^{\alpha_M} \cdot \prod_{j=1}^{M}[c_j(x, y)]^{\beta_j}[s_j(x, y)]^{\gamma_j}, \tag{18}$$

where M is the highest scale, α_M, β_j and γ_j are used to adjust the relative importance of different components, and $c_j(x, y)$ and $s_j(x, y)$ provide a comparison of contrast and structure at the j-th scale image, respectively, while $l_M(x, y)$ is only the luminance comparison at scale M.

The PSNR is used to evaluate the ratio of peak signal power to noise power and therefore reflects the amount of distortion during the fusion process. This metric is defined as follows:

$$PSNR = 10\log_{10}\frac{r^2}{MSE}, \tag{19}$$

where r indicates the peak value of the fused image. The higher PSNR value indicates that the fused image is closer to the source images and has less distortion in terms of image quality.

The FMI is used to measure the amount of feature information transmitted from the source images to the fused image. It is defined as follows:

$$FMI_F^{AB} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{I_i(A; F)}{H_i(A) + H_i(F)} + \frac{I_i(B; F)}{H_i(B) + H_i(F)}\right), \tag{20}$$

where $H_i(A)$ and $H_i(B)$ are the entropy of the corresponding windows from the input images, $I_i(A; F)$ and $I_i(B; F)$ indicate the regional mutual information between corresponding windows in the fused image and source images. A larger FMI value commonly implies that a considerable amount of feature information is transferred from the source images to the fused image.

Training details

We train the proposed fusion network on the Multi-Spectral Road Scenarios (MSRS) [33] dataset. This training set includes 1078 pairs of infrared and visible images, while the test set contains 361 image pairs. This dataset is constructed based on MFNet [55] and consists of a large number of nighttime and daytime scenes. Before feeding the training set to the fusion network, all images are normalized to [0, 1] and

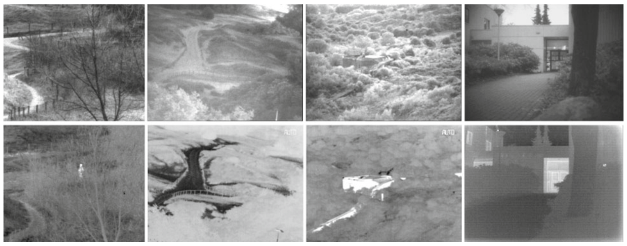


Fig. 5 Four pairs of source images. The top row contains visible images, and the second row displays infrared images

parameters are set as follows. The total loss hyper-parameter is set to $\gamma = 100$, $p = 0.68$, and $q = 0.08$. The batch size and epoch are set to 8 and 80, respectively. The model parameters are updated by the Adam optimizer with a learning rate of 0.001 and weight decay of 0.0001. All experiments are performed on an NVIDIA RTX A5000 GPU and a 2.40 GHz Intel(R) Xeon (R) Silver 4214R CPU. Since color visible images are included in MSRS, a specific fusion strategy[43] is used to process color image fusion. We first transfer the input visible images from the RGB color space to the YCbCr color space. The Y channel in the visible images is then employed to fuse the infrared images and obtain a new fused channel Y. Finally, the fused image is combined with the Cb and Cr channels of visible images and converted to the RGB color space.

Results analysis on TNO dataset

We compare the fusion performance for our method with the seven state-of-the-art algorithms applied to 24 image pairs acquired from the TNO dataset. All infrared and visible images display different scenes and are registered before being fed to the network. Samples of these images are shown in Fig. 5.

Qualitative results

For quantitative experiments, fused images produced by existing fusion methods and our proposed method are shown in Figs. 6 and 7. Some representative regions from the fused images are selected and enlarged near the bottom, to more intuitively display and analyze visual effects in the fused results. A significant target is evident in the green box and abundant textural details can be seen in the red box.

As shown, nearly all methods generate some meaningless information due to thermal radiation contamination in the background. However, our method not only highlights the target but also preserves detail information. The region in the green box indicates that although the CBF results include a bright target, the pixel distribution in this area suffers heavily from noise compared to the proposed method. Also, the IFEVIP, GTF, and FusionGAN models severely weaken significant targets in the fused images. In the case of NestFuse, IFCNN, and MFEIF, the fused images indicate that while some of the target edges are highlighted, other salient features and textural details in the fused images are blurred. In contrast, our fusion method produces more realistic contrast and successfully preserves the intensity of significant areas and the texture detail of visible images, compared with other methods. For example, the proposed scheme preserves internal contours and details for cars and clouds intact in Fig. 7. This improvement demonstrates one of the primary advantages of our method.

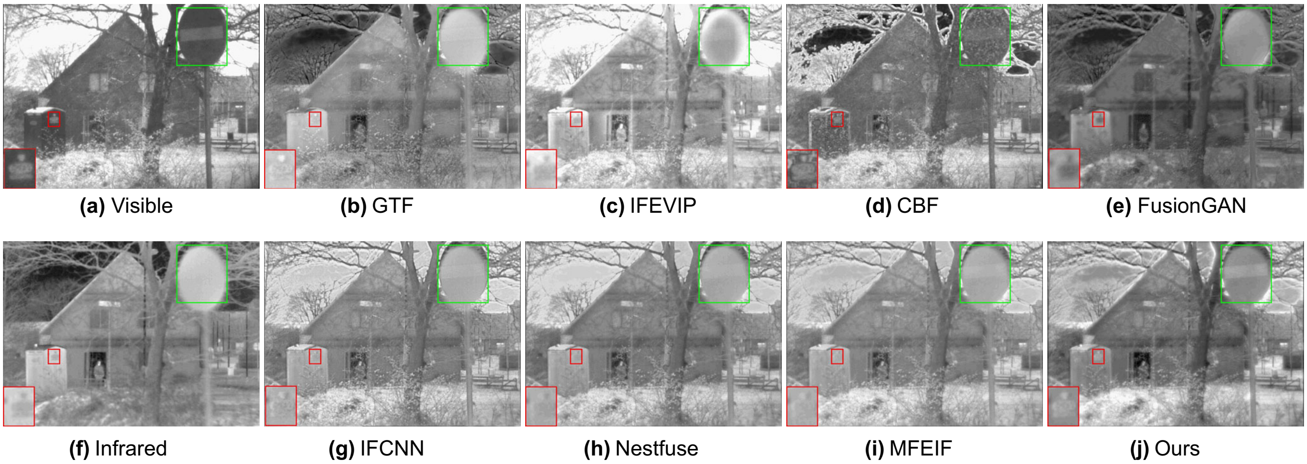
Quantitative results

Quantitative evaluation experiments are conducted using the TNO dataset, employing nine metrics to comprehensively compare our method with seven state-of-the-art methods. Average values for the compared fusion methods and the proposed algorithm are shown in Table 1 across nine metrics, where the two best values for each metric are bold and underlined, respectively. As demonstrated by the statistical results, the proposed fusion method achieves the largest average values in four of the metrics, including CC, SCD, MS_SSIM, and FMI. It also achieves reasonable performance in EN and SD, producing the second largest average values. Our method also achieves the best performance for SCD, indicating that the correlation between our fused images and the source images is the highest. In addition, the largest average values for CC and MS_SSIM indicate that our fused images transfer more considerable information while preserving structural information in the input images. The values for FMI also prove that our method well preserves feature information from the source images to the fused images. These results indicate that our method can transfer more meaningful information from the source images, especially the richest fine-grained details and significant structural information.

Results analysis on RoadScene dataset

Qualitative results

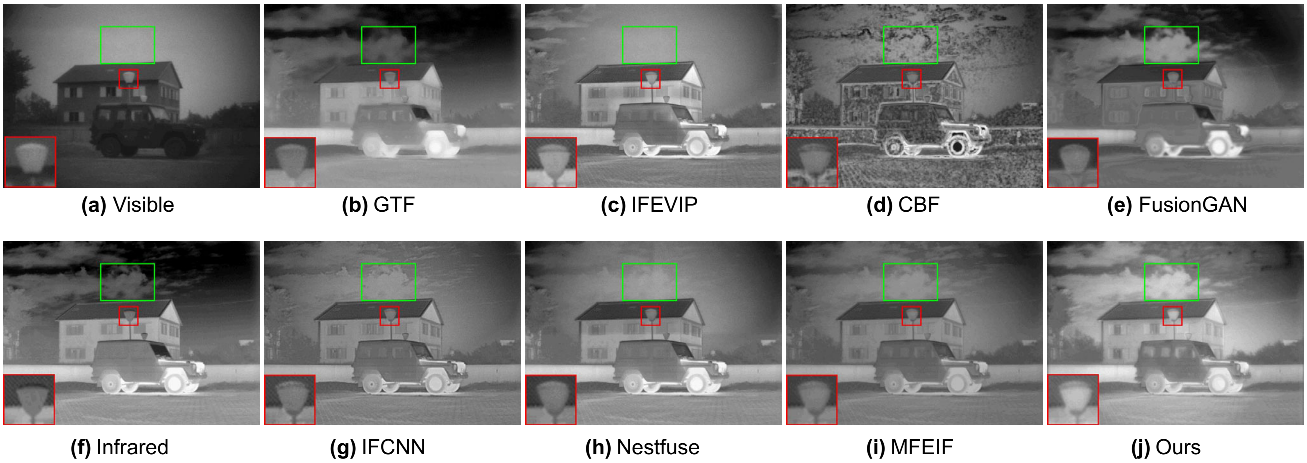
An additional 24 image pairs showing different day and night scenes are selected from the RoadScene dataset, including cars, streetlights, roads, pedestrians, bicycles, trees, and houses. The fused results produced by different fusion methods are shown in Figs. 8 and 9. It is evident that undesirable artifacts appear in the CBF results, while the GTF and IFEVIP fused images do not retain details from the infrared image. This results in significant information loss, particularly in the red box region. In addition, FusionGAN produces under-exposed results and could not retain the sharp target edges. On the contrary, the NestFuse, IFCNN, MFEIF,



(a) Visible **(b)** GTF **(c)** IFEVIP **(d)** CBF **(e)** FusionGAN

(f) Infrared **(g)** IFCNN **(h)** Nestfuse **(i)** MFEIF **(j)** Ours

Fig. 6 Visual result comparisons for different methods apply to the 'man-in-doorway' image from the TNO dataset. Our method excels at preserving abundant texture details, particularly in the zoomed-in region (i.e., the red box), and effectively highlights a salient region (i.e., the green box)



(a) Visible **(b)** GTF **(c)** IFEVIP **(d)** CBF **(e)** FusionGAN

(f) Infrared **(g)** IFCNN **(h)** Nestfuse **(i)** MFEIF **(j)** Ours

Fig. 7 Visual result comparisons for different methods apply to the 'Marne-04' image from the TNO dataset. Our method excels at preserving abundant texture details, particularly in the zoomed-in region (i.e., the red box), and effectively highlights a salient region (i.e., the green box)

Table 1 Average evaluation metric values for all methods apply to 24 image pairs from the TNO dataset

Metrics	GTF	IFEVIP	CBF	FusionGAN	IFCNN	NestFuse	MFEIF	Ours
EN	6.5999	6.6540	6.8784	6.4741	6.6637	**6.9888**	6.6295	6.9476
SF	0.0373	0.0425	**0.0553**	0.0259	0.0484	0.0429	0.0290	0.0395
SD	8.8174	8.9208	8.9346	8.2617	8.7769	**9.2871**	8.8902	9.2430
PSNR	62.7178	62.1595	63.8897	60.9702	64.3641	62.9549	**64.5550**	64.1134
CC	0.3877	0.4923	0.4377	0.4682	0.5322	0.5226	0.5527	**0.5630**
SCD	0.9237	1.5413	1.3249	1.2768	1.6169	1.7041	1.7044	**1.8574**
Nabf	0.0713	0.1189	0.2588	0.0780	0.1779	0.1308	**0.0047**	0.0993
MS_SSIM	0.8091	0.8443	0.7286	0.7362	0.9022	0.8544	0.8957	**0.9462**
FMI	0.8953	0.8917	0.8769	0.8788	0.8957	0.8965	0.8983	**0.8997**

The two best values for each metric are bold and underlined, respectively. The two types of numbers under each method name represent the number of best values and second best values, respectively

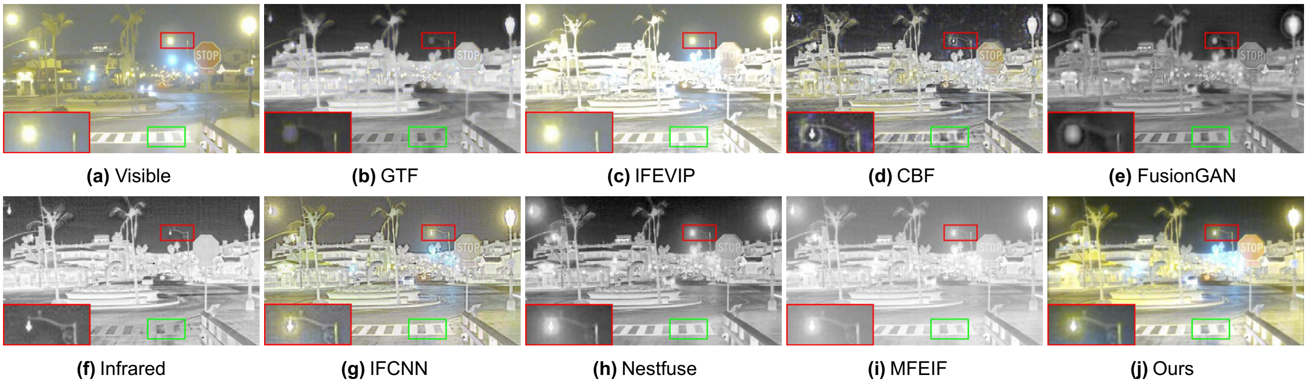

(a) Visible **(b)** GTF **(c)** IFEVIP **(d)** CBF **(e)** FusionGAN
(f) Infrared **(g)** IFCNN **(h)** Nestfuse **(i)** MFEIF **(j)** Ours

Fig. 8 Qualitative comparisons of the proposed method with seven state-of-the-art methods apply to ‘$FLIR$_07210’ from the RoadScene dataset. Our method excels at preserving abundant texture details, particularly in the zoomed-in region (i.e., the red box), and effectively highlights a salient region (i.e., the green box)

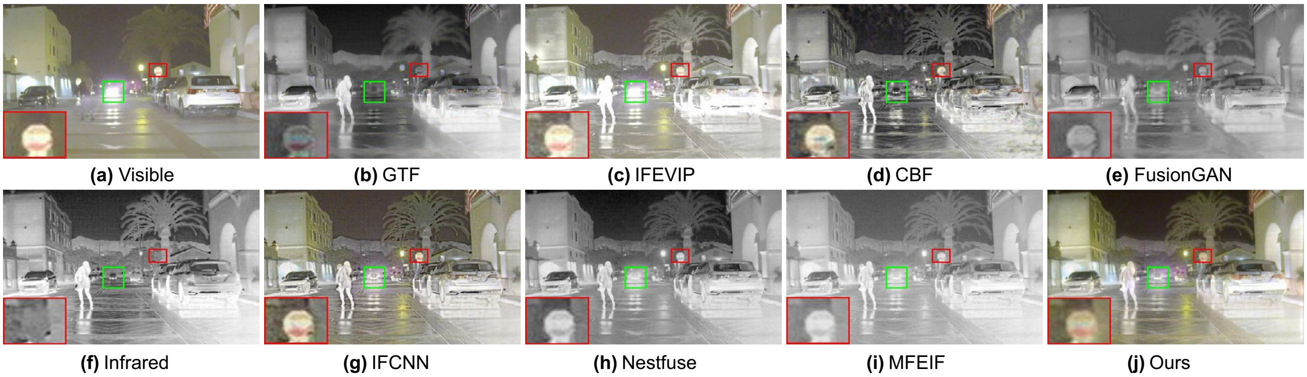

(a) Visible **(b)** GTF **(c)** IFEVIP **(d)** CBF **(e)** FusionGAN
(f) Infrared **(g)** IFCNN **(h)** Nestfuse **(i)** MFEIF **(j)** Ours

Fig. 9 Qualitative comparisons of the proposed method with seven state-of-the-art methods on ‘$FLIR$_08954’ from the RoadScene dataset. Our method excels at preserving abundant texture details, particularly in the zoomed-in region (i.e., the red box), and effectively highlights a salient region (i.e., the green box)

Table 2 Average evaluation metric values for all methods apply to 24 image pairs from the RoadScene dataset

Metrics	GTF	IFEVIP	CBF	FusionGAN	IFCNN	NestFuse	MFEIF	Ours
EN	**7.524**	7.0617	7.4704	7.1238	7.2134	7.5156	7.1476	7.3405
SF	0.0399	0.0555	**0.0658**	0.0358	0.0630	0.0558	0.0392	0.0545
SD	10.2173	9.8298	10.2358	9.9576	10.0158	10.3017	10.1937	**10.3180**
PSNR	62.6298	61.5192	63.493	60.5857	64.1579	62.6796	**64.1637**	63.9071
CC	0.5326	0.6244	0.5673	0.5974	0.6614	0.6628	0.6871	**0.6912**
SCD	0.9901	1.3149	1.1595	1.0931	1.3921	1.6465	1.5420	**1.6868**
Nabf	0.0634	0.1593	0.2576	0.1019	0.1796	0.1309	**0.0086**	0.0786
MS_SSIM	0.7861	0.8361	0.7985	0.7578	0.8991	0.8627	0.8813	**0.9088**
FMI	0.8628	0.8503	0.8516	0.8486	0.8592	**0.8631**	0.8627	0.8531

The two best values for each metric are bold and underlined, respectively. The two types of numbers under each method name represent the number of best values and second best values, respectively and the proposed method obtain better fusion performance in subjective evaluations compared with the other three fusion methods. However, the fused images obtained by the proposed method exhibit more reasonable luminance information.

Quantitative results

The results of quantitative comparisons between our method and other state-of-the-art algorithms are provided in Table 2. It shows that our method achieves the largest average across four metrics, including SD, CC, SCD, and MS_SSIM. Our proposed method presents the best SD value, indicating the fused images exhibit the highest contrast. In addition, our algorithm produces the highest CC and MS_SSIM values, suggesting the fused results share strong correlation and structural information with the source images. The highest SCD value further implies that our fused images have less pseudo-information and the strongest correlation with source images.

In summary, both qualitative and quantitative results demonstrate that our proposed method achieves excellent performance in transferring more considerable information and highlighting significant contrast, which has remarkable advantages over other methods.

Ablation studies

Multi-scale attention analysis

The multi-scale attention module plays a critical role in our fusion network as it enhances the contextual representation of the network on both local and global features. Therefore, we implement an ablation study using the multi-scale attention module, the results of which are shown in Fig. 10. The multi-scale attention module is excluded from the ablation experiment. It is evident that the fused images preserve texture details in the source images, but with low contrast. In addition, some of the visualized results exhibit a few artifacts.

Detail loss analysis

Ablation experiments are included to determine the role of detail loss in the results. More specifically, we train a network without additional detail loss, the results of which are shown in Fig. 10. Notice that when the detail loss is removed, the fusion network fails to preserve useful information of source images, specifically texture detail in background regions and pixel intensity and contours for salient targets. In addition, the results of quantitative comparisons are provided in Table 3, where all metrics are seen to decrease, excluding the SD metric. These experimental results demonstrate the importance of detail loss, which can preserve the texture details in the fused images.

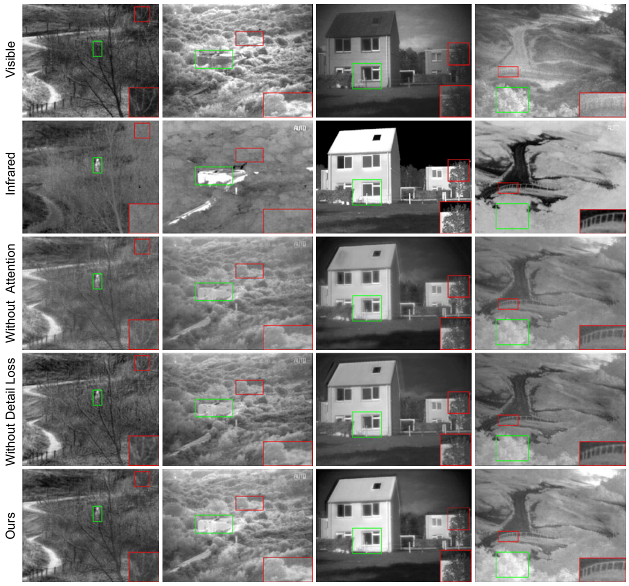


Fig. 10 Qualitative comparisons of ablation analysis results for four image pairs acquired from the TNO dataset. The source images are shown in the first two rows, followed by the fused images produced without a multi-scale attention network (Without Attention), fused images without detail loss (Without Detail Loss), and fused images produced by our method

Efficiency comparisons

To verify the computational efficiency of the fusion algorithm, the traditional methods are tested on the CPU, while the others are implemented on the GPU. As can be seen in Table 5, the average running time of the image fusion algorithms varies widely, and the running times of traditional methods are longer than that of deep learning-based methods that benefit from the GPU acceleration. Specifically, IFCNN with a simple network architecture is the fastest algorithm on all datasets. Our proposed fusion algorithm focuses on features at different scales and makes up for the missing comprehensive information via attention modules. As such, the running time for our method trails only IFCNN. Fortunately, the experiments show that our fusion algorithm has an efficiency advantage compared with other methods and will be thus feasible for real-time applications.

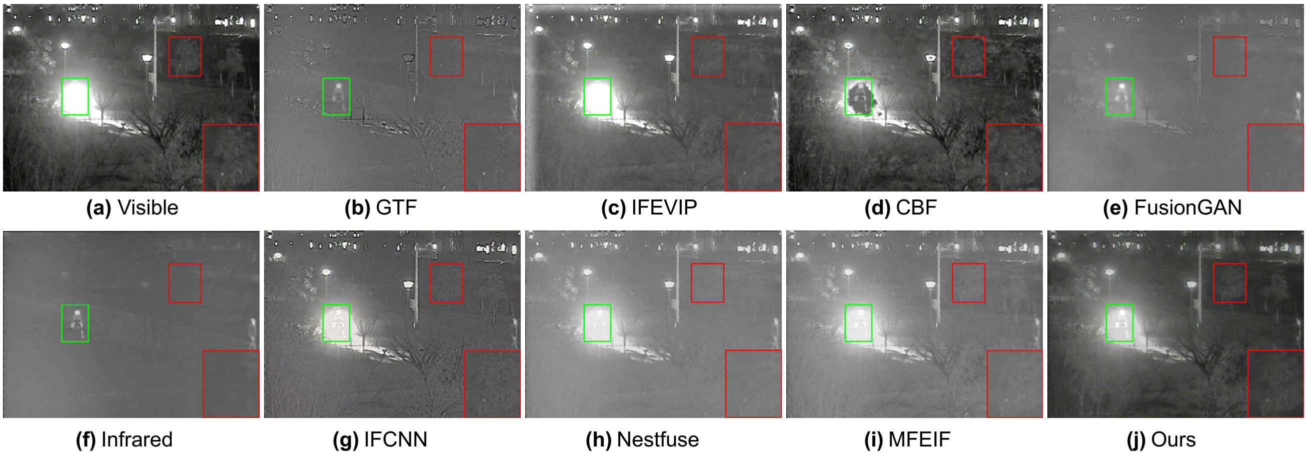
Extension to the VIFB dataset

To further verify our generalization of the proposed method, the experiment is also conducted using the VIFB dataset, which includes 21 pairs of registered visible and infrared images. These samples not only cover a wide range of environments and working conditions (e.g., indoor, outdoor, low

Table 3 Quantitative comparisons of ablation studies using the TNO dataset

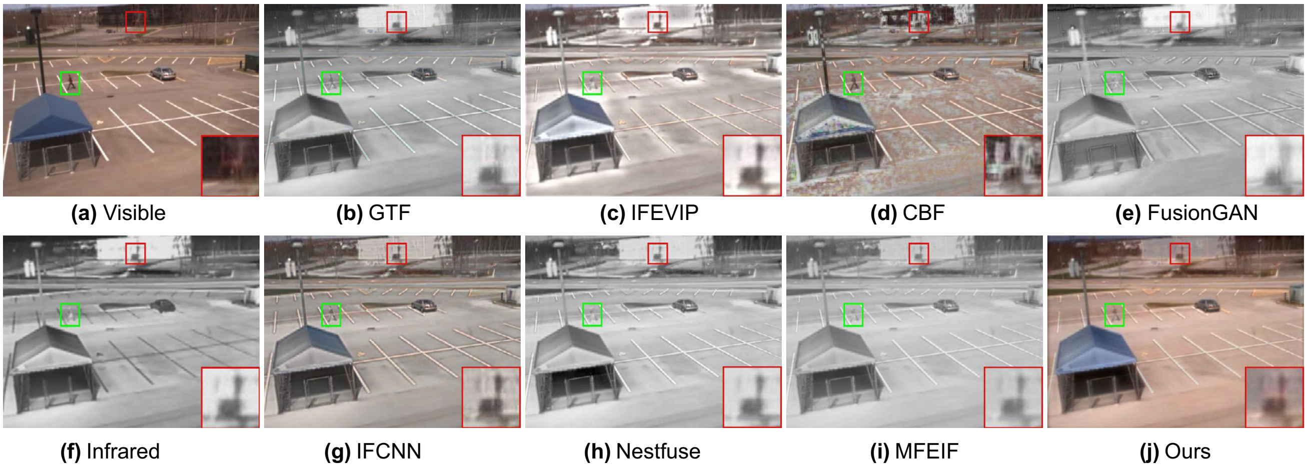
Method	EN	SF	SD	PSNR	CC	SCD	Nabf	MS_SSIM	FMI
without attention	6.7770	0.0300	9.1252	63.9204	0.5477	1.6720	**0.0543**	0.8728	0.8888
without detail loss	6.8592	0.0277	**9.2555**	63.9672	0.5472	1.7239	0.1110	0.8774	0.8870
Our	**6.9476**	**0.0395**	9.2430	**64.1134**	**0.5630**	**1.8574**	0.0993	**0.9462**	**0.8997**

Bold text indicates the best result


(a) Visible **(b)** GTF **(c)** IFEVIP **(d)** CBF **(e)** FusionGAN

(f) Infrared **(g)** IFCNN **(h)** Nestfuse **(i)** MFEIF **(j)** Ours

Fig. 11 Qualitative comparisons of eight methods apply to 'elecbike' image pairs from the extended VIFB dataset. Our method excels at preserving abundant texture details, particularly in the zoomed-in region (i.e., the red box), and effectively highlights a salient region (i.e., the green box)


(a) Visible **(b)** GTF **(c)** IFEVIP **(d)** CBF **(e)** FusionGAN

(f) Infrared **(g)** IFCNN **(h)** Nestfuse **(i)** MFEIF **(j)** Ours

Fig. 12 Qualitative comparisons of eight methods apply to 'manCar' image pairs from the extended VIFB dataset. Our method excels at preserving abundant texture details, particularly in the zoomed-in region (i.e., the red box), and effectively highlights a salient region (i.e., the green box)

illumination, and over-exposure), they also include various image resolutions, such as 320 × 240, 630 × 460, 512 × 184, and 452 × 332.

Fused results for the VIFB dataset are shown in Figs. 11 and 12, where it is evident that GTF, FusionGAN, and NestFuse lose vital information. CBF is also seen to suffer from noise interference and other undesirable artifacts. In addition, IEVIP fails to display significant targets due to overexposure to visible images. In contrast, MFEIf, IFCNN, and the proposed method preserve detail information and highlighted targets from the source images. Quantitative results for the VIFB dataset are provided in Table 4, where it is evident that our method achieves the largest average values across three metrics, including CC, SCD, and MS_SSIM. These metrics indicate the fused results exhibit a meaningful structure and texture information transferred from the source images. In contrast, the proposed method follows CBF in the EN metric because the fused images generated by CBF contain additional noise.

Table 4 Quantitative comparisons of 21 image pairs from the extended VIFB dataset

Metrics	GTF	IFEVIP	CBF	FusionGAN	IFCNN	NestFuse	MFEIF	Ours
EN	6.5061	6.9566	**7.3149**	6.3727	6.9083	6.9131	6.8695	7.0286
SF	0.0572	0.0612	**0.0789**	0.0361	0.0726	0.0579	0.0452	0.0529
SD	9.0553	9.3865	**9.7274**	8.3456	9.3688	9.5795	9.5951	9.7190
PSNR	61.7115	61.4836	62.4867	62.1179	63.3259	62.7005	**63.7930**	63.7914
CC	0.4845	0.5546	0.5144	0.5730	0.5942	0.5992	0.6216	**0.6323**
SCD	0.7584	1.2620	1.0509	0.8902	1.3786	1.4510	1.4925	**1.5522**
Nabf	0.0936	0.1528	0.3442	0.1092	0.1891	0.0925	**0.0209**	0.1075
MS_SSIM	0.7630	0.8481	0.7566	0.6757	0.9087	0.8585	0.8991	**0.9354**
FMI	0.8823	0.8908	0.8841	0.8807	0.8956	0.8949	**0.8974**	0.8923

The two best values for each metric are bold and underlined, respectively. The two types of numbers under each method name represent the number of best values and second best values, respectively

Table 5 Average running time for all methods across three datasets (unit: second)

Method	TNO	RoadScene	VIFB
GTF	6.152	8.448	8.504
IFEVIP	0.078	0.089	0.096
CBF	17.342	26.056	32.184
FusionGAN	0.697	0.440	0.535
IFCNN	**0.056**	**0.056**	**0.068**
NestFuse	3.764	2.175	2.744
MFEIF	0.084	0.068	0.079
Ours	0.061	0.063	0.073

Bold text indicates the best results and underlined text represents the second best results

Conclusion

In this paper, a novel lightweight deep learning fusion network based on multi-scale attention and hybrid dilated convolutional blocks is proposed to effectively improve the fusion of infrared and visible images. By designing hybrid dilated convolution blocks, the feature extraction module with a larger receptive field efficiently extracts more contextual information and fine-grained details without changing the size of the feature maps. The use of a unique total loss allows our proposed fusion network to simultaneously preserve texture features and salient target intensity from both infrared and visible images. In addition, the spatial/channel attention modules at different scales are designed to focus on shallow local and deep global detail features, which compensate for missing detail in the fusion process and improve the contrast of fused images. Experiments performed on two public infrared and visible image datasets demonstrate that our fused images not only include large amounts of detailed textural features but also reduce noise and artifacts. In addition, these experiments are extended to the VIFB dataset and further verify the generalizability of our proposed model.

Acknowledgements This work is supported by the National Key Research and Development Program of China (Grant No.2020YFC0833406), the National Natural Science Foundation of China (NSFC) under the grant No.62102112, the Basic Research Plan of Guizhou Province (Grant No.Qiankehejichu-ZK[2021]Yiban310), the Guizhou Provincial Science and Technology Projects (No. QKHJCZK2022YB195, QKHJCZK2023YB143) and the Youth Science and Technology Talents Cultivating Object of Guizhou Province (No. QJHKY2021104).

Data Availability The data underlying this article will be shared on reasonable request to the corresponding author.

Declarations

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Zhang H, Xu H, Tian X, Jiang J, Ma J (2021) Image fusion meets deep learning: a survey and perspective. Inform Fus 76:323–336. https://doi.org/10.1016/j.inffus.2021.06.008
2. Zhang Q, Xiao T, Huang N, Zhang D, Han J (2021) Revisiting feature fusion for RGB-T salient object detection. IEEE Trans Circ Syst Video Technol 31(5):1804–1818. https://doi.org/10.1109/TCSVT.2020.3014663
3. Kim Y-H, Shin U, Park J, Kweon IS (2021) Ms-uda: Multi-spectral unsupervised domain adaptation for thermal image semantic seg-

mentation. IEEE Robot Automation Lett 6(4):6497–6504. https://doi.org/10.1109/LRA.2021.3093652

4. Zeng X, Long J, Tian S, Xiao G (2023) Random area pixel variation and random area transform for visible-infrared cross-modal pedestrian re-identification. Expert Syst Appl 215:119307. https://doi.org/10.1016/j.eswa.2022.119307
5. Liu W, Liu W, Sun Y (2023) Visible-infrared dual-sensor fusion for single-object tracking. IEEE Sens J 23(4):4118–4128. https://doi.org/10.1109/JSEN.2023.3234091
6. Zhou, Z., Wang, B., Li, S., Dong, M.: Perceptual fusion of infrared and visible images through a hybrid multi-scale decomposition with gaussian and bilateral filters. Inform Fus 15–26 (2016) https://doi.org/10.1016/j.inffus.2015.11.003
7. Yan L, Hao Q, Cao J, Saad R, Li K, Yan Z, Wu Z: Infrared and visible image fusion via octave gaussian pyramid framework. Sci Rep 11(1) (2021) https://doi.org/10.1038/s41598-020-80189-1
8. Li H, Wu X-J, Durrani T (2020) NestFuse: an infrared and visible image fusion architecture based on nest connection and spatial/channel attention models. IEEE Trans Instrument Measure 69(12):9645–9656. https://doi.org/10.1109/TIM.2020.3005230. arXiv:2007.00328 [cs]
9. Li H, Wu X-J, Kittler J (2021) Rfn-nest: An end-to-end residual fusion network for infrared and visible images. Inform Fus 72–86https://doi.org/10.1016/j.inffus.2021.02.023
10. Ma J, Tang L, Xu M, Zhang H, Xiao G (2021) Stdfusionnet an infrared and visible image fusion network based on salient target detection.pdf. IEEE Trans Instrumentation Measurement 1–13 https://doi.org/10.1109/tim.2021.3075747
11. Wang Z, Shao W, Chen Y, Xu J, Zhang L (2023) A cross-scale iterative attentional adversarial fusion network for infrared and visible images. IEEE Trans Circ Syst Video Technol 1–1https://doi.org/10.1109/TCSVT.2023.3239627
12. Li J, Li B, Jiang Y, Cai W (2022) MSAt-GAN: a generative adversarial network based on multi-scale and deep attention mechanism for infrared and visible light image fusion. Complex Intell Syst 8(6):4753–4781. https://doi.org/10.1007/s40747-022-00722-9
13. Chen J, Li X, Luo L, Mei X, Ma J (2020) Infrared and visible image fusion based on target-enhanced multiscale transform decomposition. Inform Sci 64–78. https://doi.org/10.1016/j.ins.2019.08.066
14. Liu Y, Liu S, Wang Z (2015) A general framework for image fusion based on multi-scale transform and sparse representation. Inform Fus 147–164. https://doi.org/10.1016/j.inffus.2014.09.004
15. Li H, Wu X-J (2017) Multi-focus image fusion using dictionary learning and low-rank representation, pp 675–686. https://doi.org/10.1007/978-3-319-71607-7_59
16. Guo Z, Yu X, Du Q (2022) Infrared and visible image fusion based on saliency and fast guided filtering. Infrared Phys Technol 123:104178. https://doi.org/10.1016/j.infrared.2022.104178
17. Vargas H, Ramírez J, Pinilla S, Martínez-Torre JI (2022) Multi-sensor image feature fusion via subspace-based approach using ℓ_1-gradient regularization. IEEE J Selected Topics Signal Process 1–13. https://doi.org/10.1109/JSTSP.2022.3219357
18. Ma J, Chen C, Li C, Huang J (2016) Infrared and visible image fusion via gradient transfer and total variation minimization. Inform Fus 31:100–109. https://doi.org/10.1016/j.inffus.2016.02.001
19. Zhang Y, Liu Y, Sun P, Yan H, Zhao X, Zhang L (2020) IFCNN: a general image fusion framework based on convolutional neural network. Inform Fus 54:99–118. https://doi.org/10.1016/j.inffus.2019.07.011
20. Ma J, Xu H, Jiang J, Mei X, Zhang X-P (2020) Ddcgan: a dual-discriminator conditional generative adversarial network for multi-resolution image fusion. IEEE Trans Image Process 4980–4995. https://doi.org/10.1109/tip.2020.2977573
21. Zhang R (2019) Making convolutional networks shift-invariant again. In: ICML
22. Toet A (2022) TNO image fusion dataset. https://doi.org/10.6084/m9.figshare.1008029.v2
23. Xu H, Ma J, Jiang J, Guo X, Ling H (2022) U2Fusion: A Unified Unsupervised Image Fusion Network. IEEE Trans Pattern Anal Mach Intell 44(1):502–518. https://doi.org/10.1109/TPAMI.2020.3012548
24. Ma J, Yu W, Liang P, Li C, Jiang J (2019) FusionGAN: a generative adversarial network for infrared and visible image fusion. Inform Fus 48:11–26. https://doi.org/10.1016/j.inffus.2018.09.004
25. Liu X, Mei W, Du H (2017) Structure tensor and nonsubsampled shearlet transform based algorithm for ct and mri image fusion. Neurocomputing 131–139. https://doi.org/10.1016/j.neucom.2017.01.006
26. Zhang Q, Maldague X (2016) An adaptive fusion approach for infrared and visible images based on nsct and compressed sensing. Infrared Phys Technol 74:11–20. https://doi.org/10.1016/j.infrared.2015.11.003
27. Li H, Wu X-J, Kittler J (2020) Mdlatlrr: A novel decomposition method for infrared and visible image fusion. IEEE Trans Image Process 4733–4746. https://doi.org/10.1109/tip.2020.2975984
28. Huang Y, Yao K (2020) Multi-exposure image fusion method based on independent component analysis. In: Proceedings of the 2020 international conference on pattern recognition and intelligent systems. PRIS 2020. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3415048.3416099
29. Fu Z, Wang X, Xu J, Zhou N, Zhao Y (2016) Infrared and visible images fusion based on rpca and nsct. Infrared Phys Technol 77:114–123. https://doi.org/10.1016/j.infrared.2016.05.012
30. Ahmad T, Lyngdoh RB, Anand SS, Gupta PK, Misra A, Raha S (2021) Robust coupled non-negative matrix factorization for hyperspectral and multispectral data fusion. In: 2021 IEEE international geoscience and remote sensing symposium IGARSS, pp 2456–2459. https://doi.org/10.1109/IGARSS47720.2021.9553681
31. Liu Y, Chen X, Peng H, Wang Z (2017) Multi-focus image fusion with a deep convolutional neural network. Inform Fus 36:191–207. https://doi.org/10.1016/j.inffus.2016.12.001
32. Ma J, Tang L, Xu M, Zhang H, Xiao G (2021) STDFusionNet: an infrared and visible image fusion wetwork based on salient target detection. IEEE Trans Instrument Measure 70:1–13. https://doi.org/10.1109/TIM.2021.3075747
33. Tang L, Yuan J, Zhang H, Jiang X, Ma J (2022) PIAFusion: a progressive infrared and visible image fusion network based on illumination aware. Inform Fus 83–84:79–92. https://doi.org/10.1016/j.inffus.2022.03.007
34. Tang L, Yuan J, Ma J (2022) Image fusion in the loop of high-level vision tasks: A semantic-aware real-time infrared and visible image fusion network. Inform Fus 28–42. https://doi.org/10.1016/j.inffus.2021.12.004
35. Li J, Huo H, Li C, Wang R, Sui C, Liu Z (2021) Multigrained attention network for infrared and visible image fusion. IEEE Trans Instrument Measure 1–12. https://doi.org/10.1109/tim.2020.3029360
36. Yu F, Koltun V, Funkhouser T (2017) Dilated residual networks. In: 2017 IEEE conference on computer vision and pattern recognition (CVPR). https://doi.org/10.1109/cvpr.2017.75
37. Guo M-H, Xu T-X, Liu J-J, Liu Z-N, Jiang P-T, Mu T-J, Zhang S-H, Martin RR, Cheng M-M, Hu S-M (2022) Attention mechanisms in computer vision: a survey. Comput Vis Media 331–368. https://doi.org/10.1007/s41095-022-0271-y
38. Hu J, Shen L, Albanie S, Sun G, Wu E (2020) Squeeze-and-excitation networks. IEEE transactions on pattern analysis and machine intelligence 2011–2023. https://doi.org/10.1109/tpami.2019.2913372
39. Qin Z, Zhang P, Wu F, Li X (2021) Fcanet: Frequency channel attention networks. In: 2021 IEEE/CVF International conference on

computer vision (ICCV). https://doi.org/10.1109/iccv48922.2021.00082
40. Hu J, Shen L, Albanie S, Sun G, Vedaldi A (2018) Gather-excite: exploiting feature context in convolutional neural networks. In: Advances in Neural Information Processing Systems (NeurIPS)
41. Wang X, Girshick R, Gupta A, He K (2018) Non-local neural networks. In: 2018 IEEE/CVF conference on computer vision and pattern recognition. https://doi.org/10.1109/cvpr.2018.00813
42. Ma J, Tang L, Fan F, Huang J, Mei X, Ma Y (2022) Swinfusion: cross-domain long-range learning for general image fusion via swin transformer. IEEE/CAA J Automatica Sinica 9(7):1200–1217. https://doi.org/10.1109/JAS.2022.105686
43. Prabhakar KR, Srikar VS, Babu RV (2017) Deepfuse: a deep unsupervised approach for exposure fusion with extreme exposure image pairs. In: 2017 IEEE international conference on computer vision (ICCV). https://doi.org/10.1109/iccv.2017.505
44. Zhang X, Ye P, Xiao G (2020) Vifb: A visible and infrared image fusion benchmark. In: 2020 IEEE/CVF conference on computer vision and pattern recognition workshops (CVPRW). https://doi.org/10.1109/cvprw50498.2020.00060
45. Zhang Y, Zhang L, Bai X, Zhang L (2017) Infrared and visual image fusion through infrared feature extraction and visual information preservation. Infrared Phys Technol 83:227–237. https://doi.org/10.1016/j.infrared.2017.05.007
46. Shreyamsha Kumar BK (2015) Image fusion based on pixel significance using cross bilateral filter. Signal Image Video Process 9(5):1193–1204. https://doi.org/10.1007/s11760-013-0556-9
47. Liu J, Fan X, Jiang J, Liu R, Luo Z (2022) Learning a deep multi-scale feature ensemble and an edge-attention guidance for image fusion. IEEE Trans Circ Syst Video Technol 32(1):105–119. https://doi.org/10.1109/TCSVT.2021.3056725
48. Van Aardt J (2008) Assessment of image fusion procedures using entropy, image quality, and multispectral classification. J Appl Remote Sens 2(1):023522. https://doi.org/10.1117/1.2945910
49. Shreyamsha Kumar BK (2013) Multifocus and multispectral image fusion based on pixel significance using discrete cosine harmonic wavelet transform. Signal Image Video Process 7(6):1125–1143. https://doi.org/10.1007/s11760-012-0361-x
50. Aslantas V, Bendes E (2015) A new image quality metric for image fusion: the sum of the correlations of differences. AEU—Int J Electron Commun 69(12):1890–1896. https://doi.org/10.1016/j.aeue.2015.09.004
51. Eskicioglu AM, Fisher PS (1995) Image quality measures and their performance. IEEE Trans Commun 43(12):2959–2965. https://doi.org/10.1109/26.477498
52. Rao Y-J (1997) In-fibre bragg grating sensors. Measure Sci Technol 8(4):355–375. https://doi.org/10.1088/0957-0233/8/4/002
53. Jagalingam P, Hegde AV (2015) A review of quality metrics for fused image. Aquatic Proc 4:133–142. https://doi.org/10.1016/j.aqpro.2015.02.019
54. Wang Z, Simoncelli EP, Bovik AC (2003) Multiscale structural similarity for image quality assessment. In: The thrity-seventh asilomar conference on signals, systems & computers pp 1398–1402. IEEE, Pacific Grove, CA, USA. https://doi.org/10.1109/ACSSC.2003.1292216
55. 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 (IROS), pp 5108–5115. IEEE, Vancouver, BC. https://doi.org/10.1109/IROS.2017.8206396

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.