Engineering Applications of Artificial Intelligence 132 (2024) 107898

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

SCGRFuse: An infrared and visible image fusion network based on spatial/channel attention mechanism and gradient aggregation residual dense blocks

Yong Wang [a], Jianfei Pu [a,*], Duoqian Miao [b], L. Zhang [c], Lulu Zhang [a], Xin Du [a]

[a] School of Artificial Intelligence, Chongqing University of Technology, Chongqing, 401135, China
[b] Department of Computer Science and Technology, Tongji University, Shanghai, 201804, China
[c] School of Mechanical & Aerospace Engineering, Rehabilitation Research Institute of Singapore, Nanyang Technological University, Singapore, 308232, Singapore

ARTICLE INFO

Keywords:
Image fusion
Infrared image
Visible image
Attention mechanisms
Gradient-aggregation residual dense block

ABSTRACT

The goal of image fusion is to retain the strengths of different images in the fused result. However, existing fusion algorithms are often complex in design and overlook the influence of attention mechanisms on deep features. To address these issues, we propose an image fusion network based on spatial/channel attention mechanisms and gradient-aggregated residual dense blocks(SCGRFuse). Firstly, we design a novel gradient-aggregated residual dense block (GRXDB) that combines the advantages of ResNeXt and DenseNet, which integrating the Sobel and Laplacian operators to preserve both strong and weak texture features. Then, we introduce spatial and channel attention mechanisms to refine the channel and spatial information of feature maps, enhancing their information capturing capability. Additionally, we leverage a pooling fusion block to merge the refined spatial and channel feature maps, yielding high-quality fusion features. Compared to the existing state-of-the-art methods, experimental results on the MSRS, RoadScene and TNO datasets demonstrate the outstanding fusion performance of our proposed approach. In addition, in the task-driven experiments, SCGRFuse achieved an mIoU accuracy of 71.37%.

1. Introduction

With the advancement of information collection technologies, samples are often observed from different perspectives or in different ways, resulting in multi-view data. The task of integrating these views to aid in identifying underlying grouping structures is commonly referred to as the multi-view clustering problem (Chao et al., 2017). Due to limitations in optical imaging, hardware devices, and theoretical techniques, images captured by single sensors or single-modal sensors often only capture partial details of a scene and cannot effectively and comprehensively depict the entire scene. For example, visible light images are easily affected by external factors such as illumination and weather, resulting in lower quality. Infrared sensors can capture the thermal radiation emitted by objects, which make it easy to distinguish between background and targets, and they can work in all weather conditions. However, infrared images often overlook texture, fail to effectively describe details, and are also susceptible to noise. In contrast, visible light images have a high spatial distribution rate, strong adaptability to visual perception, and typically contain abundant texture and structural information, which are beneficial for enhancing object recognition capability. but, visible images are sensitive to lighting and occlusion. In multi-view learning, the complementarity principle assumes that each view of the data contains information that is not present in the other views. Therefore, effectively and accurately utilizing information from multiple views is expected to result in better models (Chao and Sun, 2016). To obtain complementary information from both types of images and achieve higher-quality images, image fusion technology has emerged (Ali et al., 2020). Image fusion is an important image enhancement technique that extracts meaningful information from different source images and combines them into a new image. The fused image typically exhibits strong robustness, information richness, and represent a more complex and detailed scene representation. It not only reduces data redundancy but also promotes the development of provided subsequent applications and decision-making. Therefore, the fusion of infrared and visible light images has been widely used in preprocessing modules for high-level vision tasks, such as re-identification (Lu et al., 2020), object detection (Li et al., 2017), target tracking (Li et al., 2018), human motion prediction (Liu et al., 2022b, 2023; Zhang et al., 2024;

* Corresponding author.
E-mail addresses: ywang@cqut.edu.cn (Y. Wang), pjf674786205@163.com (J. Pu), dqmiao@tongji.edu.cn (D. Miao), longbin.zhang@ntu.edu.sg (L. Zhang), 1286852139@qq.com (L. Zhang), Deante@stu.cqut.edu.cn (X. Du).

https://doi.org/10.1016/j.engappai.2024.107898
Received 27 August 2023; Received in revised form 9 January 2024; Accepted 11 January 2024
Available online 23 January 2024

Wang et al., 2024), semantic segmentation (Ha et al., 2017). In the application of the digital economy, image fusion holds extensive potential value. In smart city planning and monitoring, the fusion of image data from different sensors allows city decision-makers to gain a better understanding of urban environments, such as traffic conditions, facilitating more effective urban development planning. In the medical field, the fusion of images from various imaging technologies such as X-rays, MRI, and CT scans enables doctors to comprehensively assess patients' conditions, enhancing diagnostic accuracy. In the transportation sector, the fusion of infrared and visible light images contributes to achieving safer and more efficient traffic systems. Integrating information from cameras, radar, and lidar sensors provides a comprehensive view of road conditions and vehicle behaviors, supporting the development of autonomous driving technologies (Li et al., 2020b).

In recent years, numerous image fusion techniques have been developed due to the increasing practicality of both infrared and visible light images. These techniques can be broadly classified into two categories: traditional methods and deep learning-based methods. Traditional methods can be further divided based on different mathematical transformations, including multi-scale transform-based methods such as Discrete Wavelet Transform (DWT) (Li et al., 1995, 2011; Ben Hamza et al., 2005), representation learning-based methods such as sparse representation (SR) (Wang et al., 2014) and joint sparse representation (JSR) (Zhang et al., 2013), subspace-based methods (Bavirisetti et al., 2017; Kong et al., 2014), saliency-based methods (Zhang et al., 2017; Zhao et al., 2014), and hybrid models (Liu et al., 2015; Ma et al., 2017, 2016). On the other hand, deep learning-based methods can be classified into three categories based on network architectures: models based on autoencoders (AE) (Li et al., 2021), models based on convolutional neural networks (CNN) (Zhang et al., 2020; Xu et al., 2020b), and models based on generative adversarial networks (GAN) (Ma et al., 2019b, 2020a).

Although existing methods have been able to produce satisfactory fused images for their respective tasks, there are still many challenges in the field of image fusion. Firstly, the selection of fusion strategies and the complexity of manual design in traditional methods limit the improvement of performance. These manually designed fusion strategies not only lack the ability to learn but also introduce artifacts into the fused results. Secondly, existing methods have introduced overly complex structures into the network and overlooked the impact of attention mechanisms on deep features. This may result in the loss of relevant information and interference from some irrelevant details, ultimately leading to subpar quality of the fused images. Whether traditional methods or deep learning-based methods, most of these algorithms primarily measure the fusion performance and visual effects of the fused images, but rarely consider whether the fused images can promote high-level vision tasks in a systematic manner.

To address the aforementioned issues, inspired by SeAFusion (Tang et al., 2022), We propose a novel end-to-end fusion network called SCGRFuse for infrared and visible light image fusion. Specifically, our network eliminates the need for manually designed fusion rules and instead incorporates spatial and channel attention mechanisms to design a learnable module suitable for image fusion tasks. Firstly, a multi-scale feature extraction block is employed to capture multi-scale deep features. These features are then passed through spatial and channel attention branches to enhance the ability to capture semantic information and spatial details. Finally, an efficient fusion is performed using a pooling fusion block to effectively merge the two feature maps. Additionally, we have developed a new gradient-aggregated residual dense block (GRXDB) consisting of three branches. Two branches comprise a residual block and a dense residual block, integrating Sobel and Laplacian operators to preserve both strong and weak texture features, thereby enhancing feature diversity in feature extraction.

In summary, our contributions are as follows:

- We propose a novel fusion network called SCGRFuse, which effectively avoids the need for manually designing complex fusion rules.
- We introduce spatial and channel attention mechanisms, enabling the images to contain rich semantic and spatial information, thereby improving the performance of image fusion.
- The designed gradient-aggregated residual dense block effectively integrates deep features and strong-weak texture details, further enhancing the quality of the fused images.
- We conducted experiments on publicly available datasets and compared our algorithm with state-of-the-art methods both qualitatively and quantitatively. The results demonstrate that our algorithm achieves outstanding performance. In addition, we evaluated the quality of our images using high-level vision tasks, and the results further validate the beneficial impact of our algorithm on high-level visual tasks.

The remaining structure of this paper is as follows. In Section 2, we provide a brief overview of relevant deep learning methods and attention mechanisms for image fusion tasks. In Section 3, we provide a detailed description of the SCGRFuse network and its application in infrared and visible light image fusion. In Section 4, we present the experimental details and results of our algorithm. Finally, we conclude the paper with a discussion and summary of the findings.

2. Related work

In this section, we provide a concise review of existing methods in the field of image fusion, encompassing both traditional approaches and deep learning-based methods. Furthermore, we delve into a comprehensive exploration of attention mechanisms within the realm of deep learning.

2.1. Traditional image fusion methods

Traditional image fusion mainly addresses two problems: feature extraction and feature fusion. The most commonly used fusion method is based on multi-scale transformations, such as non-subsampled shearlet transform (NSST) (Kong et al., 2014) and discrete wavelet transform (DWT) (Li et al., 1995, 2011; Ben Hamza et al., 2005), which decompose the original image into different scale subbands. Each subband corresponds to information at different frequencies or scales. Then, based on fusion rules or strategies, these subbands are weighted and combined layer by layer to obtain the final fused image. Additionally, methods based on sparse representation are also commonly used for feature extraction. In these methods, an image is represented by a sparse coefficient matrix that describes the image's representation on a set of basis functions or dictionaries. The commonly used basis functions include wavelet bases, sparse dictionaries, and so on. By element-wise fusion or combination of the sparse coefficient matrices of two or more input images, the final sparse coefficient matrix is obtained. Then, through an inverse transform or reconstruction process, the sparse coefficient matrix is transformed into the fused image. Subspace-based methods have also received considerable attention. They assume that each input image can be represented in a specific subspace, which may have different structures and features. Subspace-based methods utilize the representation characteristics of multiple input images in different subspaces to fuse them into a comprehensive subspace and obtain the final fused image. Representative methods include principal component analysis (PCA) (Fu et al., 2016) and non-negative matrix factorization (NMF) (Mou et al., 2013). Although traditional image fusion methods often rely on manually designed rules and strategies for feature extraction, weight allocation, and fusion operations, and these rules may only be effective for specific scenarios and may not adapt well to different data and tasks, they provide new ideas and prospects for image fusion and have laid a solid foundation for deep learning methods.

2.2. Deep learning-based image fusion algorithms

Due to the excellent feature learning ability of neural networks, deep learning has gained popularity in various tasks. Compared to traditional methods, deep learning models can not only extract detailed features from source images but also preserve richer post-processing information. In 2018, Li et al. proposed a simple autoencoder (AE) fusion framework (Li and Wu, 2018) consisting of an encoder, fusion layer, and decoder. The framework effectively extracts more useful features through convolutional layers and dense blocks in the encoder. Subsequently, the fusion layer combines the high-level features using element-wise addition and l1 norm fusion strategy. The decoder, composed of four convolutional layers, is used to reconstruct the fused image.

However, the aforementioned methods suffer from limitations imposed by manually designed fusion rules, which severely restrict their fusion performance. CNN-based methods inherit the core concepts of traditional optimization methods and have been widely adopted in various image fusion domains due to their powerful feature extraction capabilities. LP-CNN (Liu et al., 2017) pioneered the application of Convolutional Neural Networks (CNNs) in the field of image fusion. It combines Low-Pass filtering (LP) with a classification-based CNN, where LP is applied as a preprocessing step to extract the low-frequency information and reduce noise in the input images. Subsequently, the classification-based CNN leverages these LP-processed images for feature learning and classification tasks, thus achieving image fusion. Since then, an increasing number of CNN-based algorithms have been developed, and researchers have explored end-to-end CNN fusion frameworks to overcome the limitations of handcrafted rules.

In the field of image fusion, obtaining ground truth is often challenging, and GAN networks have demonstrated unique advantages in addressing unsupervised deep learning problems. Collecting and annotating ground truth data for infrared images is prohibitively expensive. To overcome this challenge, Ali et al. introduced an Attention-based Generative Adversarial Network (AGAN) (Ali and Cha, 2022) for enhancing infrared images and training networks. This not only significantly reduces costs but also applies the approach to real-world bridge systems. In the field of image fusion, Ma et al. first proposed an end-to-end image fusion framework based on GANs (FusionGAN) (Ma et al., 2019b), which utilizes a discriminator to generate fusion images with rich textures. Additionally, they introduced detail loss and edge-enhancement loss to improve the quality of detail information and sharpen the edges of thermal targets. However, since there is only one discriminator in the framework, the generated fusion images tend to be biased towards either the visible or infrared image. To address this issue, Ma et al. further proposed a dual-discriminator conditional generative adversarial network (Ma et al., 2020b) based on FusionGAN to maintain a balance between infrared and visible light images.

Although these methods have achieved impressive results in the field of image fusion, they primarily focus on fusion quality and overlook the requirements of high-level vision tasks. Furthermore, the goal of Fine-Grained Image Recognition (FGIR) is to recognize differences among images classified within subordinate categories (Kang et al., 2022). However, existing image fusion network architectures cannot effectively extract fine-grained detailed features. In 2022, Tang et al. introduced a novel image fusion framework called SeAFusion (Tang et al., 2022), which integrates high-level vision tasks into the fusion process. Instead of making significant innovations in network architecture or learning paradigms, they approached the image fusion task from a new perspective, leveraging high-level vision tasks to drive the fusion process. The emergence of SeAFusion (Tang et al., 2022) has opened up new possibilities for image fusion by incorporating the considerations of high-level vision tasks. SeAFusion integrates a semantic segmentation network after the image fusion network, providing semantic feedback to the fusion network through gradient backpropagation. Moreover, the newly designed GRDB feature extraction module enhances the fusion network's ability to describe fine-grained spatial details. This approach enhances the quality of image fusion and promotes the application of fused images in semantic segmentation tasks. Similarly, Liu et al. proposed TarDAL (Liu et al., 2022a), wherein object detection is employed instead of a segmentation network. TarDAL constrains the fusion network from the perspective of object detection to retain rich semantic information. Additionally, Sun et al. trained two object detection models based on infrared and visible light images, using both models to jointly constrain the fusion network (Sun et al., 2022). The attention maps generated in the detection network are transferred to the fusion network, facilitating comprehensive information aggregation and improving the recognition capabilities of fused images in object detection tasks.

2.3. Deep attention mechanism

The attention mechanism is a technique used to enhance the focus and processing capability of neural networks on important parts of input data. In neuroscience, researchers have discovered the mechanism of selective attention in human perception and cognition, where individuals selectively focus on interesting parts of complex information for further processing and analysis. This has sparked interest in applying attention mechanisms in the fields of machine learning and artificial intelligence. To address the performance limitations of traditional segmentation decoders, Kong and Cha designed a unique decoder (Kang and Cha, 2022). It maximizes the use of attention operations by configuring attention decoders, upsampling, and coarse upsampling, thereby reducing heavy computational costs while maintaining real-time processing performance. For control over environmental noise, Mostafavi et al. introduced an attention module after each encoder (Mostafavi and Cha, 2023). This module maintains lower-level features through skip connections and intelligently selects and utilizes effective feature mappings. Lewis et al. combined CNN and transformer to develop a novel dual-encoder–decoder medical segmentation network (Lewis et al., 2023). The two structures interact, with one component of the network primarily focusing on global relationships between pixels, while the other concentrates on extracting local and smaller-scale feature information. Due to its excellent performance, the attention mechanism has been widely applied and developed, particularly in the field of computer vision, including tasks such as image fusion. In 2020, Li et al. proposed a spatial/channel attention fusion strategy model called NestFuse (Li et al., 2020a), which integrates multi-scale deep features and applies them to infrared and visible light image fusion. Additionally, Chen et al. extended the Transformer model by introducing spatial transformers and channel transformers (Chen et al., 2023) to capture global information of each feature and the inter-channel relationships. Li et al. also incorporated the attention mechanism into GAN models (Li et al., 2022) by establishing spatial and channel attention modules to suppress unimportant information such as image backgrounds, thereby further enhancing the extraction of the original feature information from the image data.

3. Method

In this section, we will provide a comprehensive introduction to the infrared and visible light image fusion network based on spatial/channel attention mechanisms and a gradient-aggregated residual dense block. Firstly, we will briefly introduce the research motivation and the network architecture proposed in this paper. Then, we will provide a detailed analysis of the gradient-aggregated residual dense block used in the network. Finally, we will present a detailed description of our spatial/channel attention mechanisms and loss function.

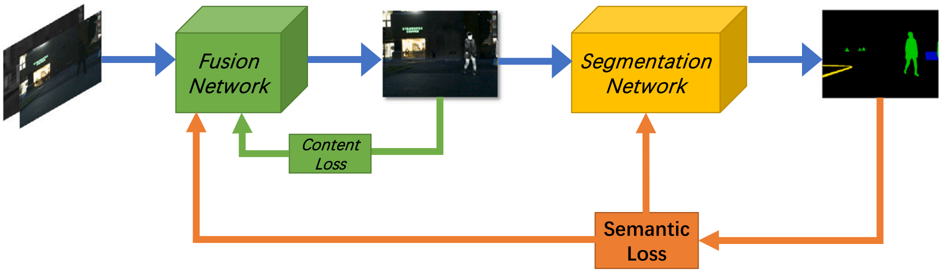

Fig. 1. The overall framework of the proposed infrared and visible image fusion algorithm.

3.1. Motivation

In the field of image fusion, dual-scale decomposition separates the input image into a background image containing low-frequency information with large-scale pixel intensity variations and a detail image containing high-frequency information with small-scale pixel intensity variations. Currently, most algorithms incorporate certain prior knowledge and use filters or optimization-based methods for image decomposition, making them manually designed decomposition algorithms. We emphasize that image decomposition algorithms are essentially feature extractors. In formal terms, they transform the source image from the spatial domain to the feature domain. As is well known, deep neural networks are a promising data-driven feature extraction method that offers significant advantages over traditional manually designed methods. Utilizing CNN to generate fused images overcomes the difficulties associated with handcrafted activity level measurements and fusion rules. Therefore, we propose a new design called the Gradient-aggregated Residual Dense Block (GRXDB) as a feature extraction block. GRXDB achieves feature reuse through the main dense stream and enhances the descriptive capability for fine-grained details and coarse-texture details through the residual gradient stream. Additionally, existing image fusion methods predominantly focus on the quality of the fused image while neglecting the connection with downstream tasks. Some studies (Haris et al., 2021) indicate that only considering visual quality and quantitative metrics may not assist high-level visual tasks. To address this issue, we introduce spatial/channel attention mechanisms in the network to provide more accurate semantic information and spatial information for the images. We evaluate the impact of our fused images on segmentation performance by incorporating a semantic segmentation task.

3.2. Problem formulation

Given a pair of precisely registered infrared image $I_{ir} \in \mathbb{R}^{H\times W\times 1}$ and visible light image $I_{vi} \in \mathbb{R}^{H\times W\times 3}$, guided by a customized loss function, image fusion is achieved through feature extraction, aggregation, and reconstruction. To enhance fusion performance and facilitate the application of advanced visual tasks, we propose a fusion network comprised of spatial/channel attention and gradient-aggregation residual dense blocks. The structure of the framework is shown in Fig. 1. Firstly, a feature extraction block based on the gradient-aggregation residual dense block is designed. Specifically, we utilize this feature extraction block to extract deep features with rich detailed information from infrared and visible light images. Additionally, we employ two convolutional blocks to extract shallow features, which can be represented as follows:

$$\left\{F_{ir}, F_{vi}, F_{conv}\right\} = \left\{E_F(F_{ir}), E_F(F_{vi}), E_{conv}(F_{ir}, F_{vi})\right\}, \tag{1}$$

where F_{ir}, F_{vi} and F_{conv} represent the infrared features, visible light features, and shallow features, respectively. Furthermore, the GRXDB module is deployed to enhance the capability of extracting coarse and fine texture features (its network structure will be discussed in the Section 3.3). Given F^i the input of GRXDB, its output F^{i+1} can be denoted as:

$$\begin{aligned} F^{i+1} &= GRXDB(F^i) \\ &= Conv(\nabla^2 F^i \oplus F^i) \oplus Conv(\nabla F^i) \oplus F^i, \end{aligned} \tag{2}$$

where $Conv(\cdot)$ represents multiple convolution operations, ∇ and ∇^2 correspond to Sobel gradient and Laplace gradient operators, respectively, *i.e.* , specially designed convolution operations with manually crafted kernels. Additionally, $\oplus$ denotes element-wise summation.

Subsequently, employing a concatenation fusion strategy, the deep infrared and visible light features are concatenated together. Before computing attention weights, we expand the network's receptive field using four different-sized convolution kernels (1 × 1, 3 × 3, 5 × 5, 7 × 7). The process can be represented as follows:

$$F_{cout} = Conv(C(Conv_{n\times n}(C(F_{ir}, F_{vi})))), n \in \{1, 3, 5, 7\}, \tag{3}$$

where $C(\cdot)$ denotes the concatenation operation. Then, the output features are fed into spatial and channel attention masks to calculate weights. Finally, a pooling fusion method is employed to generate our ultimate feature map. The process is represented as follows:

$$F_{SC} = PFB(Conv(F_{cout} \otimes B_c(M_s)), Up(F_{cout} \otimes B_s(M_c))), \tag{4}$$

where M_c and M_s represent the channel attention mask and spatial attention mask, respectively, and $\otimes$ denotes element-wise multiplication. Before the multiplication, attention values are appropriately broadcasted: spatial broadcasting is applied along the spatial dimension using the operation $B_s(\cdot)$, broadcasting channel attention values, and channel broadcasting is applied along the channel dimension using the operation $B_c(\cdot)$, broadcasting spatial attention values. $Up(\cdot)$ represents the upsampling operation.

The final step involves concatenating shallow features with refined features, and through the image reconstructor R_I, reconstructing the fused image I_f from the merged feature values:

$$I_f = R_I(C(F_{conv}, F_{SC})), \tag{5}$$

Finally, by introducing a segmentation model N_s the fused image I_f is segmented, further guiding the training of the fusion network. Given the fused image I_f the semantic perception process is represented as follows:

$$I_s = N_s(I_f), \tag{6}$$

3.3. Network architecture

The overall framework proposed in this paper is similar to SeA-Fusion (Tang et al., 2022). First, the source images are input into the fusion network. Then, the fused image generated by the fusion network is further fed into the segmentation network. By introducing semantic loss, more semantic information is integrated into the fused image to optimize the overall fusion result. To generate high-quality fused images and achieve outstanding performance in high-level vision tasks, we propose a fusion network based on spatial/channel attention mechanisms and a gradient-aggregated residual dense block, as

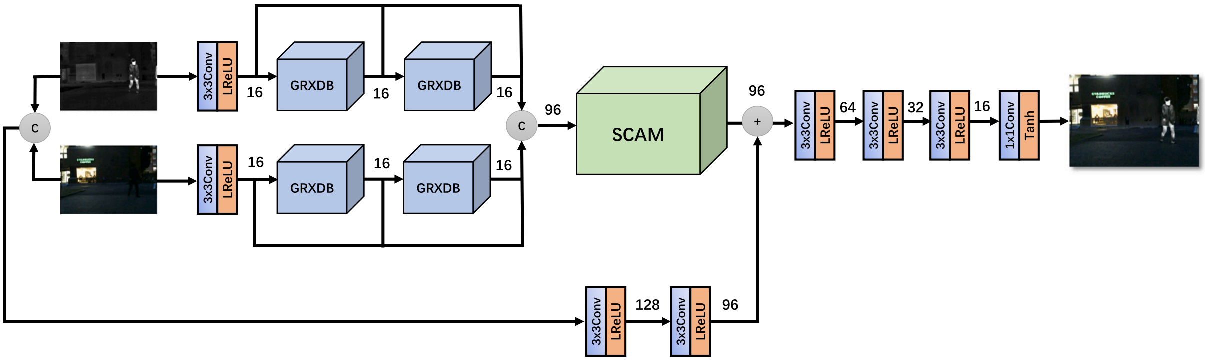

Fig. 2. The architecture of the infrared and visible light image fusion is based on spatial/channel attention mechanisms and gradient-aggregated residual dense blocks.

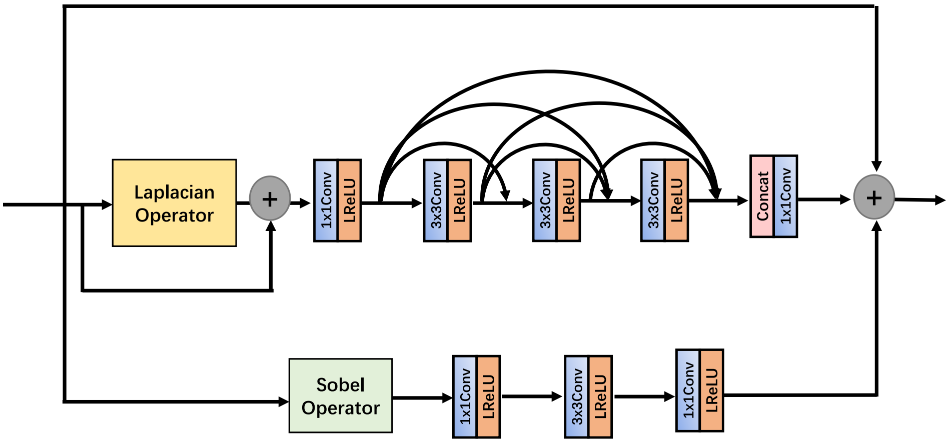

Fig. 3. The specific design of the gradient-aggregated residual dense block (GRXDB).

shown in Fig. 2. Our fusion network consists of a feature extractor, spatial/channel attention modules, and a feature reconstructor.

During the training stage, the infrared and visible light images are first input into the feature extractor separately, then merged along the channel dimension. The merged features are then passed through two convolutional blocks to extract shallow-level features. The feature extractor consists of two parallel branches for infrared and visible light feature extraction, each containing an identical convolutional block and two GRXDBs. To better extract feature information, we fuse the features from the two GRXDBs to obtain dense features. After refinement through the spatial/channel attention modules, the generated features are added to the shallow-level features and input into the feature reconstructor to obtain the final fused image. The loss function is then calculated to optimize the entire framework until the training process is completed and the desired fusion framework is obtained. During the testing stage, our network does not require any manually designed strategies and can directly generate fused images using the trained fusion network.

3.4. Gradient-aggregated residual dense block

The Aggregated Residual Dense Block (RXDB) (Long et al., 2021) is a multi-branch architecture that utilizes a split-transform-merge strategy. It provides an effective way of feature extraction and representation for neural networks, leveraging feature reuse, strong feature expression, parameter efficiency, and gradient propagation advantages. Inspired by RXDNFuse (Long et al., 2021), we design a Gradient-aggregated Residual Dense Block (GRXDB) as shown in Fig. 3. Our module consists of a residual block and a residual dense block. In particular, two branches are integrated with gradient operators to extract strong and weak textures. More specifically, our GRXDB consists of three branches. The main branch is composed of three 3 × 3 convolutional blocks and two 1 × 1 convolutional blocks. To fully utilize various convolutional layers for feature extraction, we introduce dense connections in the main branch and handle the differences between channels using 1 × 1 convolutional blocks. Additionally, the main branch incorporates the Laplacian operator to further extract weak texture features. The residual branch consists of two 3 × 3 convolutional blocks, one 1 × 1 convolutional block, and the Sobel operator to preserve strong texture features. The third branch remains unchanged, serving as the residual branch to retain the input feature information. Finally, the outputs of the main branch, residual branch, and residual branch are combined using element-wise addition to integrate deep features. It is worth noting that the ReLU activation function discards negative activations, which may be effective for classification tasks but not suitable for image fusion tasks. To better meet the requirements of image fusion, we set the activation function of GRXDB to Leaky ReLU, which retains negative activation information.

3.5. Attention mechanisms

The attention mechanism allows for the allocation of different weights to amplify useful information while suppressing harmful features. Inspired by NestFuse, we propose a lightweight and trainable multi-scale spatial/channel attention refinement module, which we refer to as SCAM. Its main structure is shown in Fig. 4. As same-scale convolutions may hinder the extraction of multi-scale information from different features, we employ four convolutional blocks with different kernel sizes (1 × 1, 3 × 3, 5 × 5, and 7 × 7) to capture deep features with multiple receptive fields and scales. The activation function for these convolutional blocks remains Leaky ReLU. Afterward, the obtained deep features are concatenated along the channel axis

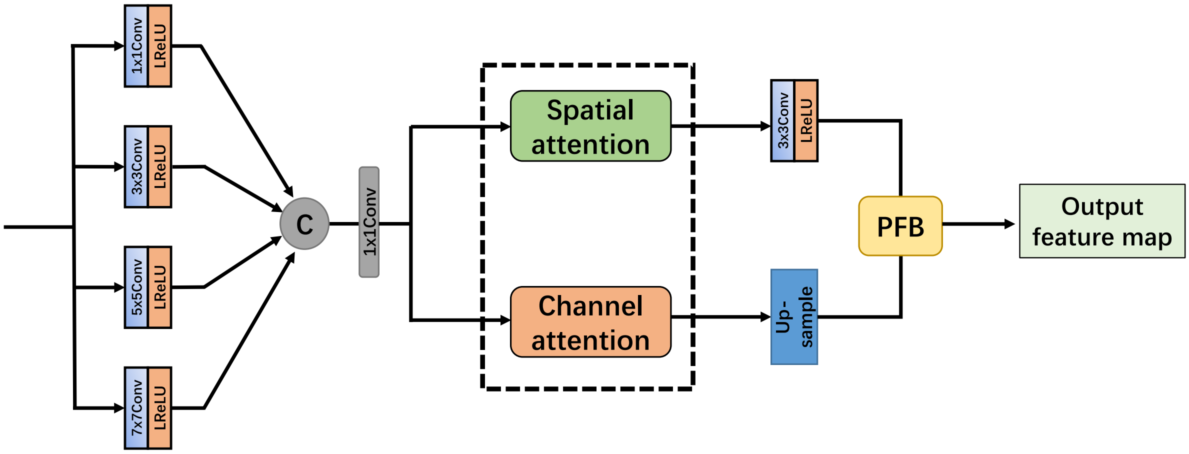

Fig. 4. The network architecture of the spatial/channel attention module (SCAM). PFB:pooling fusion block.

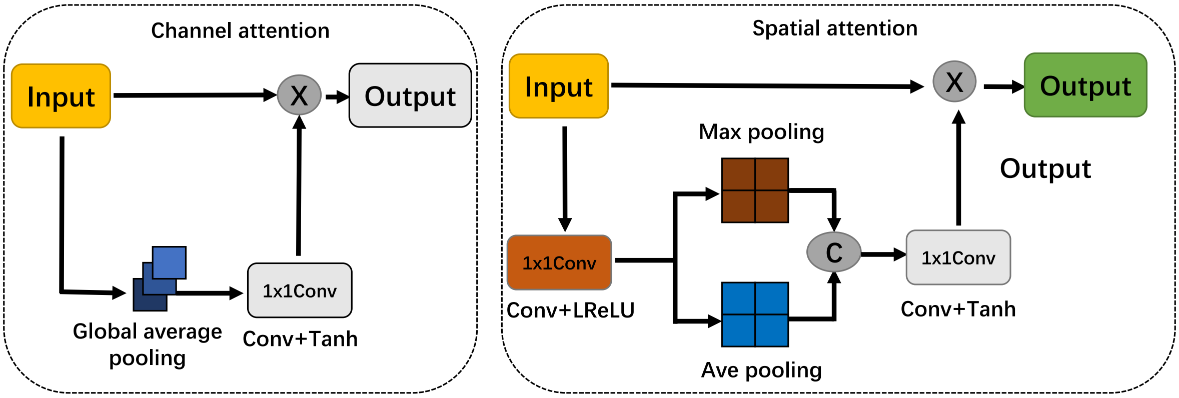

Fig. 5. The process to generate the channel attention and spatial attention masks.

and then passed through a 1 × 1 convolutional block to adjust the channel number before being sent to the spatial and channel attention layers. The generation of attention masks Ms and Mc is a crucial process in the attention mechanism, aiming to enhance the network's information capturing capability and obtain more accurate spatial and semantic information. Inspired by Woo et al. (2018), we use self-attention functions to generate spatial and channel attention masks, as illustrated in Fig. 5

To generate the channel attention mask, we first perform global average pooling on each channel of the feature maps to obtain the average value of all pixels within that channel. This process yields a new 1×1 channel map. Then, we use the Tanh activation function and a 1×1 convolutional block to adjust the values of the feature map. It is worth noting that we divide the computed values of the Tanh activation function by 2 and add 0.5, which maps the values to the range of $[0, 1]$. Finally, we multiply the computed channel map with the input feature map to generate the output of the channel attention branch. Similarly, to obtain the spatial attention mask, we first reduce the channel dimension of the input feature map using a 1×1 convolutional block. Then, we use both max pooling and average pooling to obtain two feature maps, each having a single channel, and their sizes remain the same as the input feature map. Next, we concatenate the two feature maps and use a 1 × 1 convolutional block to reduce their channel dimension to 1. The Tanh activation function is applied to this convolutional block, consistent with the channel attention mechanism. Finally, we multiply the generated spatial attention mask with the input feature map to obtain the output of the spatial attention branch.

After the refinement of the attention branches, we use upsampling to restore the channel feature map to its original size, and we use convolutional operations to control the channel dimension of the spatial feature map. In order to generate feature maps that simultaneously meet the requirements of intra-class similarity and inter-class diversity, we employ the Pooling Fusion Block (PFB) (Peng et al., 2021) to fuse the generated spatial and channel attention feature maps, as shown in Fig. 6. We perform 3 × 3 average pooling on the upsampled channel feature map to achieve smoothing, and use reflection padding to construct the true boundaries to reduce boundary artifacts. Finally, we concatenate it with the spatial feature map and use a 1 × 1 convolutional block with the Leaky ReLU activation function to correct

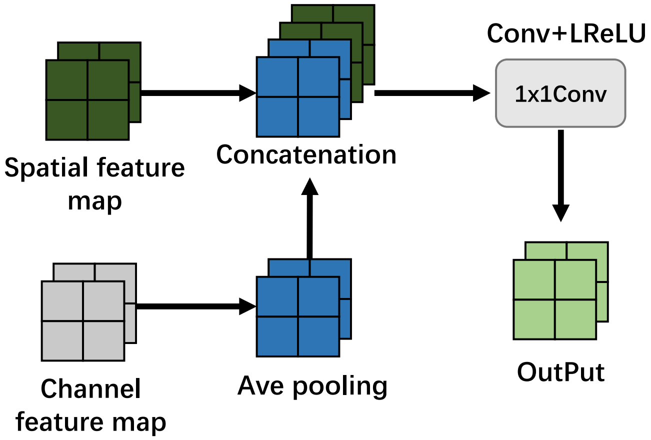

Fig. 6. The whole structure of the pooling fusion block.

the boundaries, thus generating the fused feature map that satisfies the desired criteria.

3.6. Loss function

Considering that the fusion network needs to fully integrate complementary information from the source images, ensure the visual fidelity of the fusion image, and provide effective support for high-level vision tasks, we use content loss and semantic loss (Tang et al., 2022) to guide the training of the model. These loss functions are specifically designed for image fusion tasks; therefore, they cannot be directly applied to image classification. The definition of the total loss is as follows:

$$L_{total} = L_{content} + \beta L_{semantic}, \tag{7}$$

where β represents the hyperparameter that controls the importance of the semantic loss. It is essential to emphasize that β is dynamically adjusted during the training process, following the approach in the original paper (Tang et al., 2022), where we set it to $\gamma(m-1)$, with m denoting the mth iteration. γ is a constant used to evaluate the content loss and semantic loss. For detailed training strategies, please refer to SeAFusion (Tang et al., 2022).

To enhance the visual quality and quantitative metrics of the fusion model, we employ content loss to optimize the entire model. The content loss consists of two components: intensity loss L_{int} and texture loss $L_{texture}$, defined as follows:

$$L_{content} = L_{int} + \alpha L_{texture}, \tag{8}$$

where L_{int} is used to constrain the overall intensity representation of the fusion image, and $L_{texture}$ is used to enforce the fusion image to contain more detailed texture features. α is used to balance the intensity loss and texture loss.

The intensity loss is used to measure the differences between the source image and the fusion image at the pixel level. Therefore, the intensity loss is defined as follows:

$$L_{int} = \frac{1}{HW}\left\| I_f - max(I_{ir}, I_{vi})\right\|_1, \tag{9}$$

where $I_f \in \mathbb{R}^{H\times W\times 1}$ belongs to the registered infrared image set, and $I_{vi} \in \mathbb{R}^{H\times W\times 3}$ represents the corresponding visible light image set. H and W denote the height and width of the images, respectively. $\|\cdot\|_1$ denotes the l_1-norm used to calculate the absolute differences between pixels. $max(\cdot)$ represents the maximum selection strategy, used to choose the maximum value during the computation process.

We aim to ensure that the fused image not only maintains the optimal intensity distribution but also preserves the rich texture details from the source images. Therefore, we introduce the texture loss, which imposes constraints on the fused image to contain more abundant texture details. The texture loss is defined as follows:

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

The ∇ represents the Sobel gradient operator, which is used to measure the fine-grained texture of the image. The $|\cdot|$ denotes the absolute value operation.

In addition to the content loss, we also use semantic loss (Tang et al., 2022) to enhance the semantic information in the fused image. Specifically, we introduce a segmentation model (Peng et al., 2021) to perform segmentation on the fused image. The output of the segmentation network includes the segmentation result $I_s \in \mathbb{R}^{H\times W\times C}$ and the auxiliary segmentation result $I_{sa} \in \mathbb{R}^{H\times W\times C}$. Therefore, the semantic loss consists of the main semantic loss and the auxiliary semantic loss, defined as follows:

$$L_{semantic} = L_{main} + \lambda L_{aux}, \tag{11}$$

$$L_{main} = \frac{-1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W}\sum_{c=1}^{C} L_{so}^{(h,w,c)} log(I_s^{h,w,c}), \tag{12}$$

$$L_{aux} = \frac{-1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W}\sum_{c=1}^{C} L_{so}^{(h,w,c)} log(I_{sa}^{h,w,c}), \tag{13}$$

The parameter λ, as inspired by SeAFusion (Tang et al., 2022), is set to 0.1 to balance the main semantic loss and the auxiliary semantic loss. $L_{so} \in \mathbb{R}^{H\times W\times C}$ represents the one-hot vector transformed from the segmentation label $L_s \in \mathbb{R}^{H\times W\times C}$.

4. Experimental validation

In this section, we first describe our experimental setup and experimental details. Based on this, we validate the effectiveness of our method through comparative experiments and generalization experiments. Additionally, we perform semantic segmentation on the generated fusion images to demonstrate the superiority of our algorithm in high-level vision tasks. Finally, we conduct ablation experiments to verify the rationality of the network architecture.

4.1. Experimental configurations

To comprehensively evaluate the proposed algorithm, we conducted extensive quantitative and qualitative assessments on three datasets: MSRS (Tang et al., 2022), TNO (Toet and Hogervorst, 2012) and RoadScene (Xu et al., 2020a). In addition, we compared our algorithm with nine SOTA algorithms, including one traditional approach: GTF (Ma et al., 2016), one autoencoder-based approach: DenseFuse (Li and Wu, 2018), two GAN-based approaches: FusionGAN (Ma et al., 2019b), GANMcC (Ma et al., 2020c), four CNN-based approaches: IFCNN (Zhang et al., 2020), SDNet (Zhang and Ma, 2021), U2Fusion (Xu et al., 2020a) and SeAFusion (Tang et al., 2022), one image decomposition-based approach: DeFusion (Liang et al., 2022). The implementations of these nine algorithms are publicly available, and we kept the parameters consistent with the original papers. Specifically, DenseFuse and IFCNN adopted element-wise addition and element-wise maximum fusion strategies to fuse the deep features, respectively.

For quaive evaluation, we selected seven metrics to objectively assess the fusion performance. including entropy(EN) (Roberts et al., 2008), mutual information(MI) (Qu et al., 2002), visual information fidelity(VIF) (Han et al., 2013), spatial frequency(SF) (Eskicioglu and Fisher, 1995), standard deviation(SD) (Rao, 1997), sum of the correlations of differences(SCD) (Aslantas and Bendes, 2015) and Q_{abf} (Ma et al., 2019a). Among them, EN is computed based on information theory to quantify the information content in the fused image. A higher EN indicates a more abundant information presence in the fused image. MI refers to an information-theoretic measure assessing the amount of information transferred from the source images to the fused image. Increased MI in the fused image suggests a greater transfer of information from the source images. VIF is an index based on natural scene statistics and the human visual system, quantifying the shared information between the fused image and the source images. A higher VIF implies that the fusion result aligns more closely with human visual perception. SF unveils details and texture information in the fused image by measuring its gradient distribution. A higher SF signifies richer edge and texture details. SD is an indicator reflecting the contrast and distribution of the fused image. Regions with higher contrast are often more appealing to the human visual system, so a higher SD in the fusion result suggests better contrast. SCD is a metric indicating the quality of fusion algorithms by measuring the difference between the fused image and the original image. A higher SCD implies that the fused image contains more information from the source images. Qabf is utilized to gauge the edge information transferred from the source images to the fused image. Additionally, we utilized the Intersection over Union (IoU) to quantify the segmentation performance. Larger values of these metrics indicate better fusion performance.

Algorithm 1 Training procedure

```
Input: Infrared images $I_{ir}$ and visible images $I_{vi}$;
Output: Fused images $I_f$;
for $m \le Maxiterations\ M$ do
    for $p\ step$ do
        Select $c$ infrared images $\{I_{ir}^1, I_{ir}^2, I_{ir}^3, \cdots I_{ir}^c\}$;
        Select $c$ visible images $\{I_{vi}^1, I_{vi}^2, I_{vi}^3, \cdots I_{vi}^c\}$;
        Update the weight of semantic loss;
        Update the parameters of the fusion network $N_F$ by Adam Optimizer: $\bigtriangledown_{N_F}(L_{total}(N_F))$;
    end for
    Generate fused images from infrared and visible images in the training set;
    for $q\ step$ do
        Select $c$ fused images $\left\{I_f^1, I_f^2, I_f^3, \cdots I_f^c\right\}$;
        Update the parameters of the segmentation network $N_S$ by SGD Optimizer: $\bigtriangledown_{N_S}(L_{semantic}(N_S))$;
    end for
end for
```

4.2. Implementation details

We trained our fusion network on the MSRS dataset, which consists of infrared and visible light image pairs captured in both daytime and nighttime scenes with a spatial resolution of 480 × 640. The training set contains 1083 pairs of infrared and visible light images, while the test set includes 361 pairs of images. Additionally, the MSRS dataset provides semantic labels for 9 objects, including Background, Car, Person, Bike, Curve, Car Stop, Guardrail, Color cone, and Bump. Moreover, the images are normalized to the range of $[0, 1]$ before being fed into the network.

According to the low-level and high-level joint adaptive training strategy (Tang et al., 2022), we iteratively train the fusion network and segmentation network. Our training parameters are set as follows: the maximum number of iterations $M = 4$, fusion network iterations $p = 3610$, segmentation network iterations $q = 20\,000$, and the hyperparameter $\gamma = 1$. The hyperparameter for content loss is set as $\alpha = 10$. We use the Adam optimizer to optimize our fusion model, with batch size of 3, β_1 of 0.9, β_2 of 0.99, epsilon of $1e^{-8}$, weight decay of 0.0002, and an initial learning rate of 0.0001. Additionally, for the segmentation model, we use a mini-batch SGD optimizer with batch-size of 16, momentum of 0.9, and weight decay of 0.0005. The learning rate is updated using the initial learning rate multiplied by $(1 - \frac{iter}{max_{iter}}^{(power)})$, with an initial learning rate set to 0.01 and a power factor of 0.9. Our method is implemented on the PyTorch platform (Paszke et al., 2019). We further summarized the training process of SCGRFuse in Algorithm 1. All experiments are conducted on NVIDIA RTX A6000 and Intel(R) Core(TM) i9-10850K CPU @ 3.60 GHz.

Furthermore, since the MSRS and RoadScene datasets contain color visible light images, we adopt a special strategy (Ram Prabhakar et al., 2017) to process color information. Specifically, except for DeFusion, we first convert RGB images to YCbCr images, then use different fusion algorithms to fuse the Y channel of the visible light image and the infrared image. This is because the YCbCr color space preserves the structural details mainly in the Y channel, which emphasizes the luminance variation and chrominance channels. Finally, we convert the fused image back to the RGB color space with visible light Cr and Cb channels.

4.3. Comparative experiment

In order to thoroughly evaluate our algorithm, we first compare the proposed SCGRFuse with nine other approaches on the MSRS dataset.

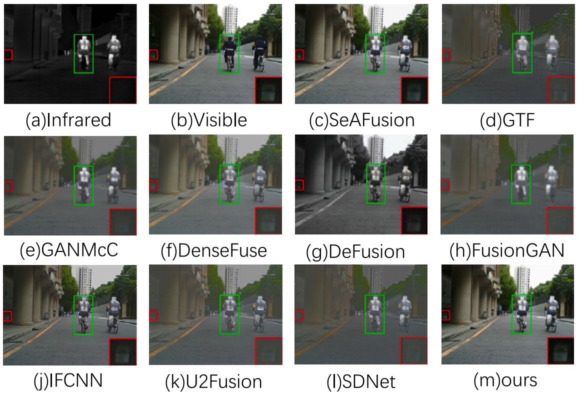

Fig. 7. Qualitative comparison of SCGRFuse with 9 state-of-the-art methods on 00537D image from the MSRS dataset.

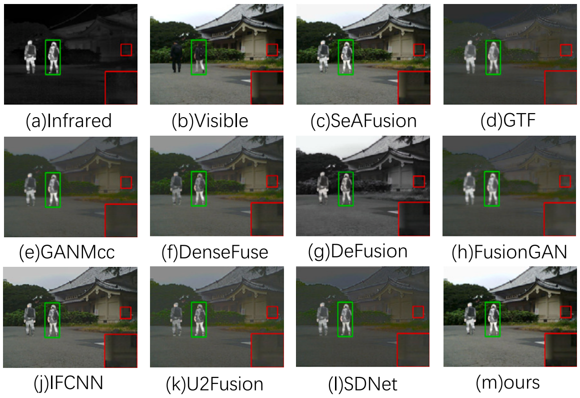

Fig. 8. Qualitative comparison of SCGRFuse with 9 state-of-the-art methods on 00633D image from the MSRS dataset.

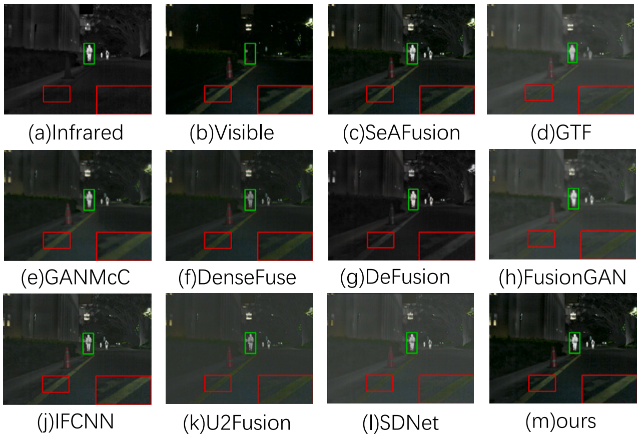

Fig. 9. Qualitative comparison of SCGRFuse with 9 state-of-the-art methods on 00858N image from the MSRS dataset.

4.3.1. Qualitative results

We visualize the generated fusion images, as shown in Figs. 7 to 10. From these four images, it can be observed that our proposed method, along with the other nine algorithms, achieves good fusion performance. In Figs. 7 and 8, during daytime scenes, GTF and FusionGAN fail to preserve the texture details of the visible light images effectively, and other methods are also affected by spectral contamination to some extent. We use red boxes to zoom in on a region to illustrate the varying degrees of spectral contamination on texture details. Additionally, we

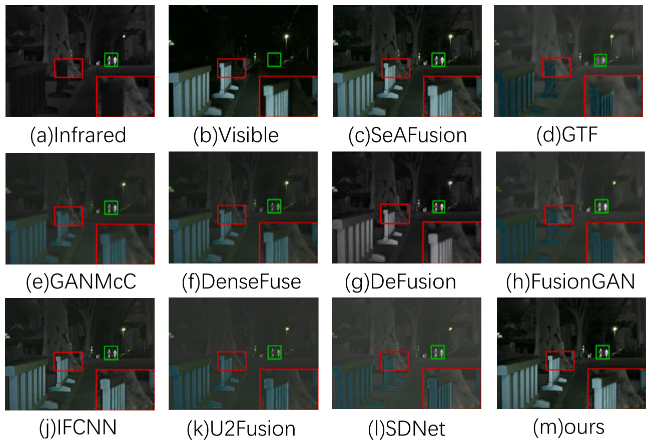

Fig. 10. Qualitative comparison of SCGRFuse with 9 state-of-the-art methods on 01024N image from the MSRS dataset.

Table 1
Quantitative comparisons of the seven metrics, i.e., EN, SD, SF, MI, SCD, VIF and Q_abf, on 361 image pairs from the MSRS dataset. The best result is indicated by RED and the second best result is represented by BLUE.

	EN	SD	SF	MI	SCD	VIF	Qabf
GTF (Ma et al., 2016)	5.47	19.56	8.48	1.67	0.76	0.51	0.40
DenseFuse (Li and Wu, 2018)	5.94	23.57	6.03	2.65	1.25	0.69	0.37
FusionGAN (Ma et al., 2019b)	5.44	17.07	4.42	1.87	0.98	0.44	0.14
IFCNN (Zhang et al., 2020)	6.28	31.50	10.76	2.82	1.38	0.77	0.60
GANMcC (Ma et al., 2020c)	5.91	22.84	4.92	2.53	1.24	0.59	0.25
SDNet (Zhang and Ma, 2021)	5.25	17.35	8.67	1.65	0.99	0.45	0.38
U2Fusion (Xu et al., 2020a)	5.56	27.71	9.24	1.96	1.26	0.55 N	0.42
DeFusion (Liang et al., 2022)	6.46	37.63	8.60	2.16	1.35	0.77	0.54
SeAFusion (Tang et al., 2022)	6.65	41.84	11.11	4.04	1.69	0.97	0.67
Ours	6.68	42.76	11.28	5.08	1.67	1.04	0.69

Table 2
Quantitative comparisons of the seven metrics, i.e., EN, SD, SF, MI, SCD, VIF and Q_abf, on 50 image pairs from the RoadScene dataset. The best result is indicated by RED and the second best result is represented by BLUE.

	EN	SD	SF	MI	SCD	VIF	Qabf
GTF (Ma et al., 2016)	7.23	44.33	8.70	3.24	1.02	0.45	0.32
DenseFuse (Li and Wu, 2018)	6.66	28.61	8.35	2.65	1.41	0.51	0.35
FusionGAN (Ma et al., 2019b)	6.73	31.92	7.13	2.70	1.06	0.35	0.24
IFCNN (Zhang et al., 2020)	6.96	35.91	13.14	2.99	1.45	0.61	0.53
GANMcC (Ma et al., 2020c)	6.93	36.01	7.40	2.74	1.53	0.47	0.30
SDNet (Zhang and Ma, 2021)	7.14	40.20	13.70	2.21	1.49	0.60	0.51
U2Fusion (Xu et al., 2020a)	7.09	38.12	13.25	1.87	1.70	0.60	0.51
DeFusion (Liang et al., 2022)	7.23	44.44	10.22	2.25	1.69	0.63	0.48
SeAFusion (Liang et al., 2022)	7.36	51.15	15.33	3.24	1.71	0.67	0.52
Ours	7.33	50.45	15.60	3.39	1.72	0.70	0.52

use green boxes to highlight the issue of weakening targets due to the introduction of irrelevant information. Only our method and SeAFusion manage to preserve rich texture details and emphasize targets, but our images have higher contrast compared to SeAFusion. For nighttime scenes, as shown in Figs. 9 and 10, all algorithms fuse complementary information from infrared and visible light images to some extent. However, most algorithms introduce irrelevant information in the fusion images, leading to the weakening of significant targets and contamination of texture background details. For example, GTF exhibits severe spectral contamination in the texture regions.

4.3.2. Quantitative results

The quantitative results for 361 pairs of images using 7 statistical metrics are shown in Table 1. Additionally, we presented cumulative distribution curve plots to illustrate the credibility of our results, as shown in the Fig. 11. Our method exhibits significant advantages in terms of SD, SF, and MI. A higher SD value indicates that the fused image has the highest contrast. A higher SF value indicates that the fused image is clearer and of better quality. A higher MI value indicates that the fused image conveys more information. Furthermore, our SCGRFuse achieves the best VIF, indicating that our fused images are more consistent with the human visual system. Our method also obtains the best Qabf, suggesting that the fused images preserve more edge information. Moreover, our method shows the highest EN, indicating that our images contain the most information. Only a marginal difference separates our method from SeAFusion in terms of the SCD metric.

Table 3
Quantitative comparisons of the seven metrics, i.e., EN, SD, SF, MI, SCD, VIF and Q_abf, on 25 image pairs from the TNO dataset. The best result is indicated by RED and the second best result is represented by BLUE.

	EN	SD	SF	MI	SCD	VIF	Qabf
GTF (Ma et al., 2016)	6.69	40.05	9.54	2.91	0.95	0.52	0.40
DenseFuse (Li and Wu, 2018)	6.42	26.00	6.78	2.30	1.54	0.57	0.35
FusionGAN (Ma et al., 2019b)	6.48	28.84	6.27	2.36	1.27	0.42	0.22
IFCNN (Zhang et al., 2020)	6.80	35.15	12.81	2.50	1.65	0.64	0.52
GANMcC (Ma et al., 2020c)	6.67	32.19	6.22	2.33	1.63	0.51	0.27
SDNet (Zhang and Ma, 2021)	6.64	32.66	12.05	1.52	1.49	0.56	0.44
U2Fusion (Xu et al., 2020a)	6.83	34.55	11.52	1.37	1.71	0.58	0.44
DeFusion (Liang et al., 2022)	6.95	38.41	8.21	1.78	1.64	0.60	0.41
SeAFusion (Tang et al., 2022)	7.10	44.20	12.41	2.89	1.72	0.69	0.51
Ours	7.09	43.55	12.67	3.38	1.71	0.78	0.53

4.4. Generalization experiment

To demonstrate the generalization performance of our proposed SCGRFuse, we conducted generalization experiments on the RoadScene and TNO datasets. It is worth noting that our model was trained on the MSRS dataset and directly tested on the RoadScene and TNO datasets.

4.4.1. Qualitative results

The qualitative comparison of different algorithms on the RoadScene dataset is shown in Figs. 12 and 13. Almost all methods are affected by thermal radiation, resulting in the weakening of salient objects. GTF, DenseFuse, FusionGAN, GANMcC, and SDNet exhibit particularly noticeable effects. Similarly, we use red boxes to magnify regions with rich texture details and green boxes to highlight salient objects. It is worth mentioning that IFCNN and SeAFusion are only slightly affected by irrelevant information. Furthermore, our fusion results show similarities with visible light images in the background region, and the pixel intensities of salient objects are mostly consistent with the infrared images.

The qualitative results of different algorithms on the TNO dataset are shown in Figs. 15 and 16. From the figures, it can be observed that DenseFuse and U2Fusion significantly weaken the salient objects, while GTF, FusionGAN, and GANMcC blur the edges of salient objects. Additionally, other methods exhibit some degree of spectral pollution in the background region. Only our method and SeAFusion successfully preserve the texture details of visible light images and the intensity of salient objects.

4.4.2. Quantitative results

We randomly selected 50 pairs and 25 pairs of images from the RoadScene and TNO datasets, respectively, for quantitative evaluation. The comparison results of different algorithms on the 7 metrics are shown in Tables 2 and 3. The cumulative distribution curve plots are shown in Figs. 14 and 17. From Table 2 and Fig. 14, it can be observed that SCGRFuse is generally in the leading position, showing significant advantages in SF, MI, and VIF. This indicates that our images are not only clear and contain rich texture details but also more in line with the human visual system. As shown in Table 3 and Fig. 17, on

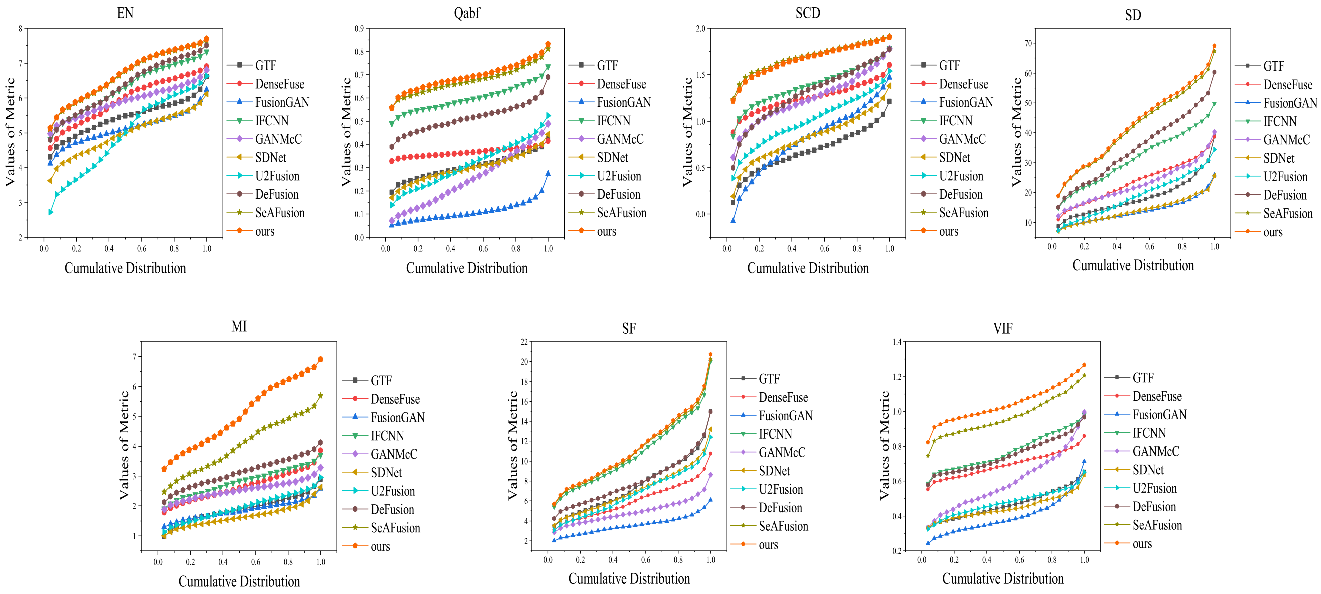

Fig. 11. Quantitative comparisons of the seven metrics, i.e., EN, SD, SF, MI, SCD, VIF and Q_abf, on 361 image pairs from the MSRS dataset. A point (x, y) on the curve denotes that there are $100 * x$ percent of image pairs which have metric values no more than y.

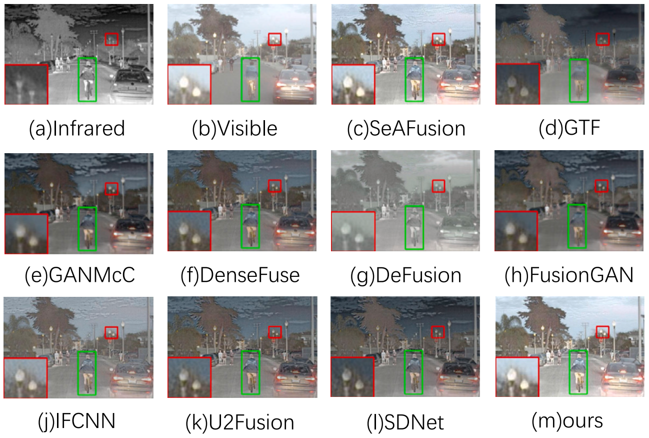

Fig. 12. Qualitative comparison of SCGRFuse with 9 state-of-the-art methods on $FLIR$_06832 image from the RoadScene dataset.

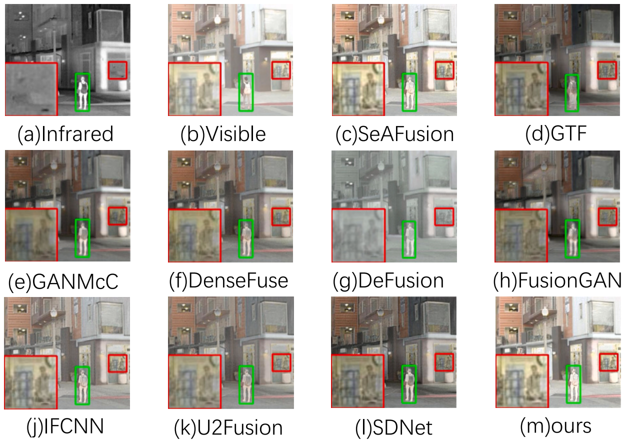

Fig. 13. Qualitative comparison of SCGRFuse with 9 state-of-the-art methods on $FLIR$_08835 image from the RoadScene dataset.

the TNO dataset, SCGRFuse ranks first in MI, VIF, and Qabf, and our performance is also close to the first place in other metrics. Objectively speaking, our algorithm outperforms others in the 7 objective evaluation metrics, demonstrating better stability and accuracy of our model.

4.5. Task-driven evaluation

The fused images can be used not only for visual observation but also for advanced visual tasks. Therefore, we perform semantic segmentation on the fused images and evaluate the segmentation performance of different fusion methods. For a fair comparison, we retrain the Deeplabv3+ (Chen et al., 2018) network with different fusion methods on the MSRS dataset. Specifically, we first generate fused images using each fusion method. Then, we separately train Deeplabv3+ on the infrared images, visible light images, and the nine fused image training sets. The segmentation performance is measured using the Intersection over Union (IoU) metric. Our training configuration is as follows: We use MobileNetv2 (Sandler et al., 2018) as the backbone network and apply both cross-entropy and Dice loss as supervision for the model. The training is performed using Stochastic Gradient Descent (SGD) with a batch size of 4 and 100 epochs. The initial learning rate is set to 7e-3, and we reduce it with cosine annealing. The segmentation results are shown in Table 4, where The abbreviations in the table represent the following categories: Background, Cars, Person, Bike, Curves, Car Stop, Guardrail, Color cone, Bump, etc. It can be observed that our algorithm generally takes a leading position in various IoU categories, with a first-place ranking in terms of MIoU. We attribute our advantage to two main factors. Firstly, our network effectively integrates the complementary information from both infrared and visible light images, which helps the segmentation model comprehensively understand the imaging scene. Secondly, our SCGRFuse, with the utilization of spatial and channel attention mechanisms, enhances information capturing ability, and guided by the semantic loss, strengthens spatial and semantic information, enabling the segmentation network to describe the imaging scene more accurately.

In addition, we provide some visual examples to demonstrate the segmentation results of both infrared and visible light images and different fused images. We only present the segmentation results of four representative fusion algorithms, namely GTF, DenseFuse, U2Fusion,

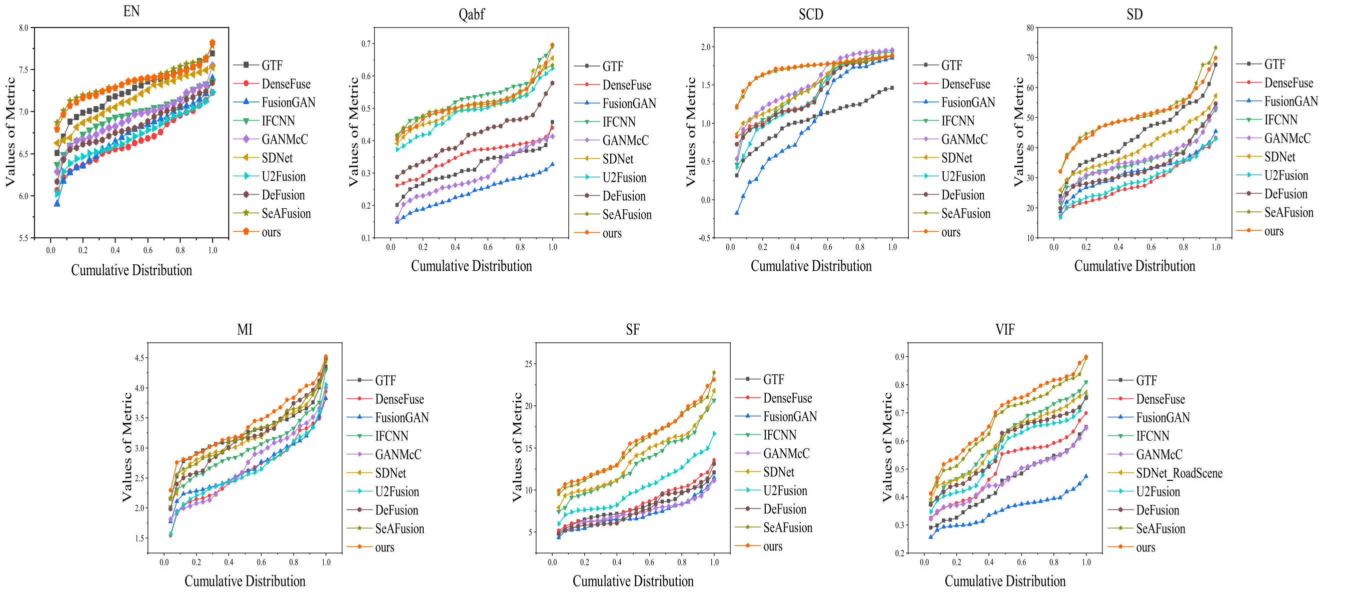

Fig. 14. Quantitative comparisons of the seven metrics, i.e., EN, SD, SF, MI, SCD, VIF and $Q_a bf$, on 50 image pairs from the RoadScene dataset. A point (x, y) on the curve denotes that there are $100 * x$ percent of image pairs which have metric values no more than y.

Table 4
Segmentation performance (mIoU) of visible, infrared and fused images on the MSRS dataset. RED indicates the best result and BLUE represents the second best result.

	BG	car	Per	Bik	Cur	CS	Gr	CC	Bu	mIoU
Infrared	97.58	85.43	70.18	64.61	50.14	53.02	43.39	45.13	58.02	63.05
Visible	97.85	87.18	59.63	68.54	51.72	66.14	73.27	56.48	65.23	69.56
GTF (Ma et al., 2016)	97.81	86.88	69.72	65.66	50.61	61.81	47.80	48.33	63.45	65.79
DenseFuse (Li and Wu, 2018)	97.97	87.16	68.82	67.71	52.80	67.60	63.73	54.15	66.25	69.58
FusionGAN (Ma et al., 2019b)	97.95	87.20	69.59	67.74	51.99	64.69	57.74	54.49	60.93	68.04
IFCNN (Zhang et al., 2020)	97.92	87.51	69.13	68.01	52.41	64.13	71.11	52.21	65.28	69.75
GANMcC (Ma et al., 2020c)	97.87	87.09	68.81	67.98	49.44	65.72	64.67	55.40	62.22	68.80
SDNet (Zhang and Ma, 2021)	97.96	87.42	70.45	67.68	53.71	60.50	49.44	54.36	64.43	67.33
U2Fusion (Xu et al., 2020a)	97.88	87.46	68.79	68.26	51.41	63.67	63.74	53.00	57.88	68.01
DeFusion (Liang et al., 2022)	97.87	87.09	69.06	66.29	51.66	62.83	72.55	50.30	62.50	68.91
SeAFusion (Tang et al., 2022)	98.00	87.66	69.30	67.71	53.09	67.07	74.90	53.46	69.12	71.15
Ours	98.07	87.96	69.23	67.73	54.17	67.95	73.66	54.11	69.43	71.37

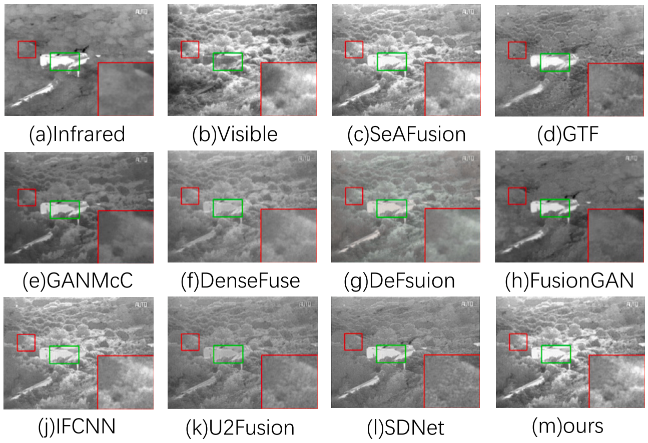
(a)Infrared (b)Visible (c)SeAFusion (d)GTF
(e)GANMcC (f)DenseFuse (g)DeFsuion (h)FusionGAN
(j)IFCNN (k)U2Fusion (l)SDNet (m)ours

Fig. 15. Qualitative comparison of SCGRFuse with 9 state-of-the-art methods on $Kaptein\_1123$ image from the TNO dataset.

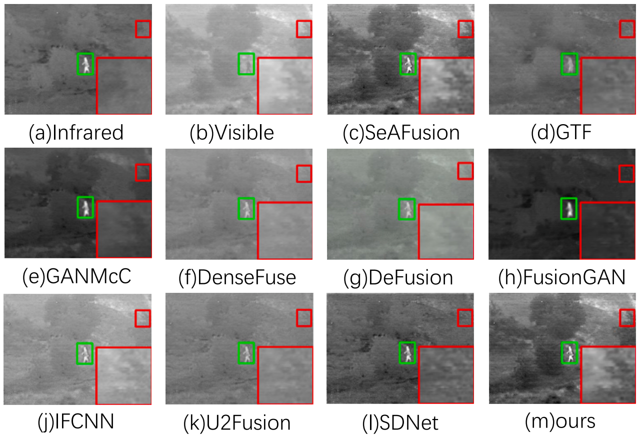
(a)Infrared (b)Visible (c)SeAFusion (d)GTF
(e)GANMcC (f)DenseFuse (g)DeFusion (h)FusionGAN
(j)IFCNN (k)U2Fusion (l)SDNet (m)ours

Fig. 16. Qualitative comparison of SCGRFuse with 9 state-of-the-art methods on $Tree\_4915$ image from the TNO dataset.

and SeAFusion, as shown in Fig. 18. From the results, it can be observed that the infrared image focuses more on salient objects such as pedestrians, while the visible light image better describes the background information. It is worth noting that our fusion method fully integrates the semantic information of the images during the fusion process. As a result, the segmentation model can produce better segmentation results on our fused images, such as accurately segmenting pedestrians in image 00127D and identifying curves in scene 00504D.

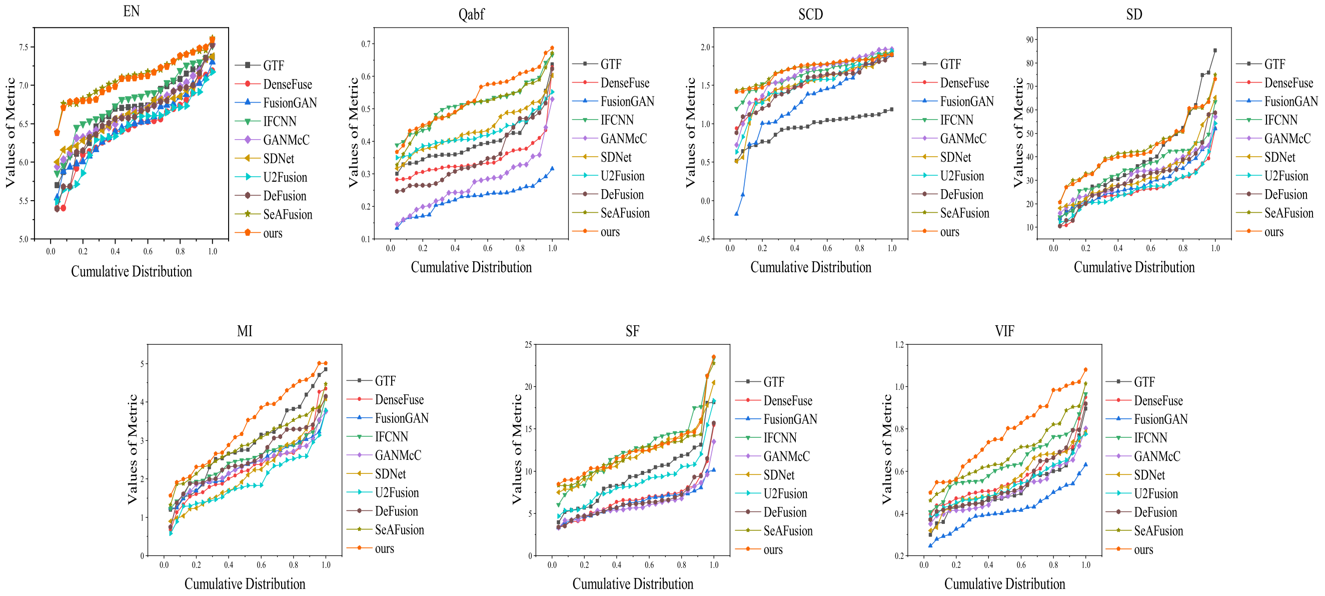

Fig. 17. Quantitative comparisons of the seven metrics, i.e., EN, SD, SF, MI, SCD, VIF and Q_abf, on 25 image pairs from the TNO dataset. A point (x, y) on the curve denotes that there are $100 * x$ percent of image pairs which have metric values no more than y.

Table 5
The fusion evaluation metrics of ablation studies. RED indicates the best result and BLUE represents the second best result.

	EN	SD	MI	VIF	Qabf
Without GRXDB	6.51	31.40	3.14	0.49	0.61
Without SCAM	6.54	41.33	4,78	0.79	0.66
Ours	6.68	42.76	5.08	1.04	0.69

Table 6
The segmentation performance of ablation studies. RED indicates the best result and BLUE represents the second best result.

	CS	GR	Car	Bu	mIoU
Without GRXDB	65.30	63.75	87.49	67.22	69.64
Without SCAM	64.91	63.95	87.83	67.01	69.98
Ours	67.72	73.54	87.96	68.16	71.37

4.6. *Ablation studies*

To validate the effectiveness of our GRXDB and spatial/channel attention mechanism module (SCAM), we conducted ablation experiments. For the GRXDB, we removed the gradient operator and reverted the Leaky ReLU activation function to ReLU. Additionally, we no longer fused the features from the two GRXDB. From the visual results shown in Fig. 19, it can be observed that the fused images highlight salient objects, but there are issues of information loss and spectral contamination in terms of texture details. This demonstrates that our GRXDB module effectively integrates feature information and texture details, further improving the visual quality of the fused image. Regarding the spatial/channel attention mechanism module, we conducted an ablation experiment by removing the entire SCAM. Specifically, we trained a fusion model consisting of only the encoder and decoder parts, retaining the feature extraction and reconstruction parts. The results are shown in Fig. 19. It can be noticed that without SCAM, the fused image lacks amplification of useful information and suppression of harmful information. As a consequence, the fused image fails to highlight salient objects and effectively combine the complementary information from the infrared and visible light images.

Furthermore, we also provide fusion evaluation metrics and partial segmentation results for ablation experiments to illustrate the importance of GRXDB and SCAM. As shown in Table 5, the fusion performance significantly declines when the feature extraction module is removed. For the attention mechanism, coarse feature extraction can lead to inaccurate focusing of the attention mechanism on key areas, introducing more noise and interference. This makes it challenging for the attention mechanism to identify and focus on genuinely useful features. Without the spatial-channel attention module, the visual fidelity of the image and the contained information are reduced, further emphasizing the role of the attention mechanism in preserving critical information from the source image and highlighting the importance of focusing on the structure and edge information in the source image. As shown in the Table 6 We only present IoU for the Car stop, Guardrail, Car and Bump and mIoU for all categories. It can be observed that without GRXDB and SCAM, the segmentation performance of the fused image is significantly reduced. In contrast, our SCGRFuse not only effectively improves the segmentation performance but also preserves salient objects and maintains texture details.

5. Conclusion

This paper proposes an image fusion framework called SCGRFuse, which effectively fuses infrared and visible light images. The Gradient Aggregation Residual Dense Blocks designed in the Encoder part can efficiently extract deep features and preserve strong and weak texture details effectively. Additionally, we introduce a Scale Channel Attention Module to emphasize the importance of source images, magnifying useful features, and suppressing harmful features, such as differences between infrared and visible light images. Through extensive qualitative and quantitative experiments and ablation analysis on three publicly available infrared and visible light datasets, the effectiveness of SCGRFuse is demonstrated. Furthermore, we evaluate the quality of our fusion images through high-level visual tasks, and task-driven evaluation experiments reveal that our framework achieves better performance on high-level visual tasks.

However, there are still limitations in our proposed method: (1) The hyperparameters of the model, such as the loss function, are determined based on the experience and experiments from other related literature, which may face the problem of fine-tuning. (2) SCGRFuse proposed in this paper is only verified for the fusion of infrared and visible light images, but there are certain relationships between various image fusion tasks, and this method may lack generalization.

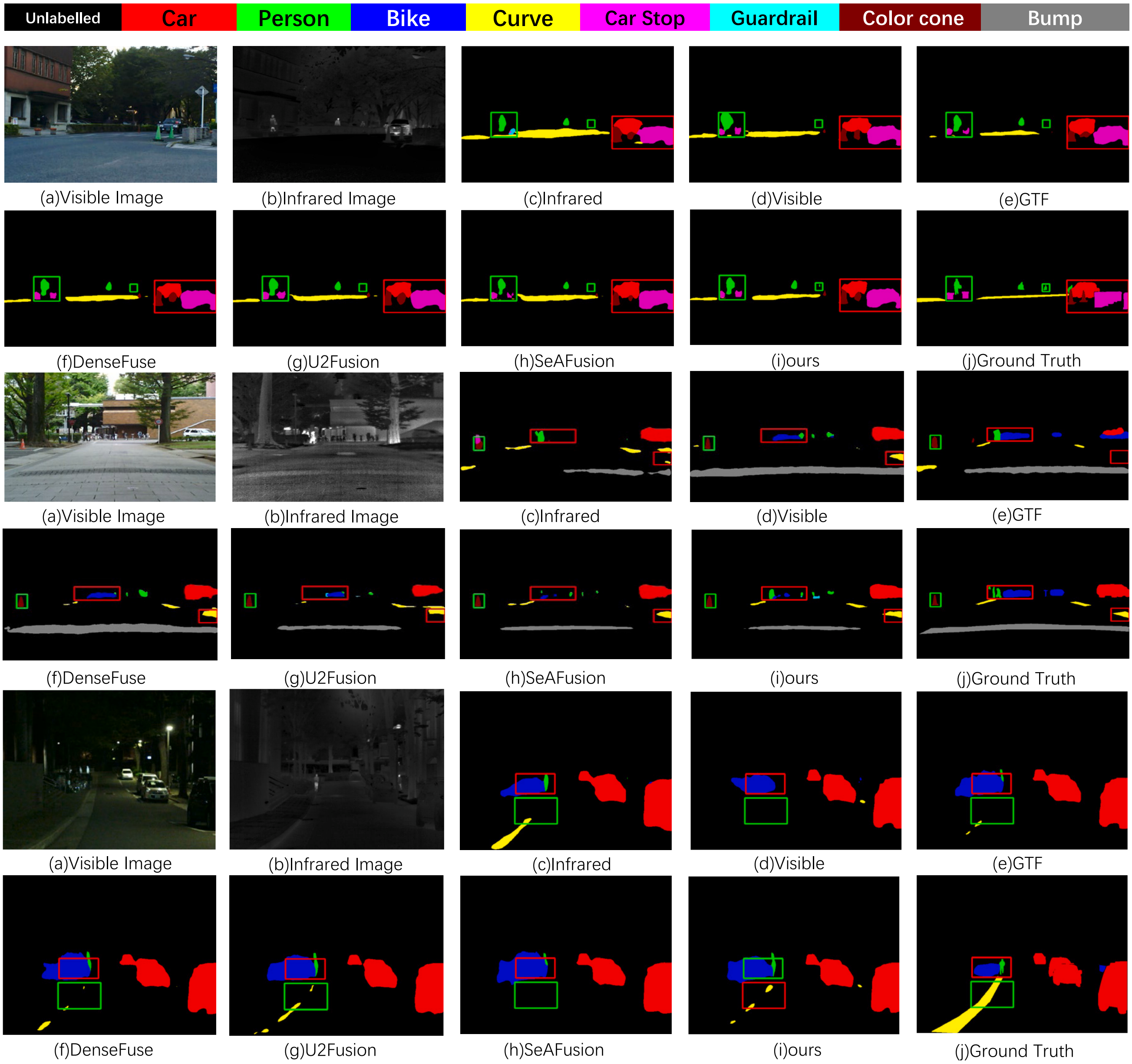

Fig. 18. Segmentation results for infrared, visible and fused images from the MSRS dataset. Each two rows represent a scene, and from top to bottom are: 00127D, 00504D and 01066N.

Therefore, future research should focus on finding reasonable parameter settings and determining the proportional relationship between various hyperparameters adaptively. In the future, we will continue to explore image fusion tasks and expand SCGRFuse to other fusion tasks, such as medical image fusion, and multi-focus fusion.

CRediT authorship contribution statement

Yong Wang: Project administration, Writing – review & editing. **Jianfei Pu:** Investigation, Validation, Visualization, Writing – original draft. **Duoqian Miao:** Project administration, Supervision. **L. Zhang:** Supervision. **Lulu Zhang:** Methodology, Supervision. **Xin Du:** Investigation, Supervision.

Declaration of competing 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.

Data availability

Data will be made available on request.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grant 61976158 and Grant 61673301

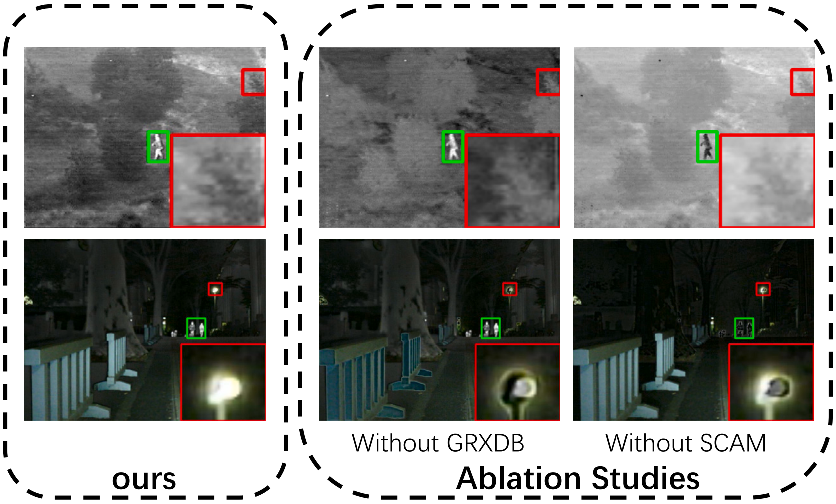

Fig. 19. Visualized results of ablation.

References

Ali, R., Cha, Y.-J., 2022. Attention-based generative adversarial network with internal damage segmentation using thermography. Autom. Constr. 141, 104412.

Ali, R., Zeng, J., Cha, Y.-J., 2020. Deep learning-based crack detection in a concrete tunnel structure using multispectral dynamic imaging. In: Smart Structures and NDE for Industry 4.0, Smart Cities, and Energy Systems. Vol. 11382, SPIE, pp. 12–19.

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.

Bavirisetti, D.P., Xiao, G., Liu, G., 2017. Multi-sensor image fusion based on fourth order partial differential equations. In: 2017 20th International Conference on Information Fusion. Fusion, IEEE, pp. 1–9.

Ben Hamza, A., He, Y., Krim, H., Willsky, A., 2005. A multiscale approach to pixel-level image fusion. Integr. Comput.-Aided Eng. 12 (2), 135–146.

Chao, G., Sun, S., 2016. Consensus and complementarity based maximum entropy discrimination for multi-view classification. Inform. Sci. 367, 296–310.

Chao, G., Sun, S., Bi, J., et al., 2017. A survey on multi-view clustering. arXiv preprint arXiv:1712.06246.

Chen, J., Ding, J., Yu, Y., Gong, W., 2023. THFuse: An infrared and visible image fusion network using transformer and hybrid feature extractor. Neurocomputing 527, 71–82.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European Conference on Computer Vision. ECCV, pp. 801–818.

Eskicioglu, A.M., Fisher, P.S., 1995. Image quality measures and their performance. IEEE Trans. Commun. 43 (12), 2959–2965.

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.

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, IEEE, pp. 5108–5115.

Han, Y., Cai, Y., Cao, Y., Xu, X., 2013. A new image fusion performance metric based on visual information fidelity. Inf. Fusion 14 (2), 127–135.

Haris, M., Shakhnarovich, G., Ukita, N., 2021. Task-driven super resolution: Object detection in low-resolution images. In: Neural Information Processing: 28th International Conference, ICONIP 2021, Sanur, Bali, Indonesia, December 8–12, 2021, Proceedings, Part V 28. Springer, pp. 387–395.

Kang, D.H., Cha, Y.-J., 2022. Efficient attention-based deep encoder and decoder for automatic crack segmentation. Struct. Health Monit. 21 (5), 2190–2205.

Kang, Y., Chao, G., Hu, X., Tu, Z., Chu, D., 2022. Deep learning for fine-grained image recognition: a comprehensive study. In: Proceedings of the 2022 4th Asia Pacific Information Technology Conference. pp. 31–39.

Kong, W., Lei, Y., Zhao, H., 2014. Adaptive fusion method of visible light and infrared images based on non-subsampled shearlet transform and fast non-negative matrix factorization. Infrared Phys. Technol. 67, 161–172.

Lewis, J., Cha, Y.-J., Kim, J., 2023. Dual encoder–decoder-based deep polyp segmentation network for colonoscopy images. Sci. Rep. 13 (1), 1183.

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.

Li, H., Manjunath, B., Mitra, S.K., 1995. Multisensor image fusion using the wavelet transform. Graph. Models Image Process. 57 (3), 235–245.

Li, J., Song, M., Peng, Y., 2017. Infrared and visible image fusion based on saliency detection and infrared target segment. In: DEStech Transactions on Computer Science and Engineering. CII.

Li, H., Wu, X.-J., 2018. DenseFuse: A fusion approach to infrared and visible images. IEEE Trans. Image Process. 28 (5), 2614–2623.

Li, H., Wu, X.-J., Durrani, T., 2020a. NestFuse: An infrared and visible image fusion architecture based on nest connection and spatial/channel attention models. IEEE Trans. Instrum. Meas. 69 (12), 9645–9656.

Li, H., Wu, X.-J., Kittler, J., 2021. RFN-nest: An end-to-end residual fusion network for infrared and visible images. Inf. Fusion 73, 72–86.

Li, S., Yang, B., Hu, J., 2011. Performance comparison of different multi-resolution transforms for image fusion. Inf. Fusion 12 (2), 74–84.

Li, Y., Zhao, H., Hu, Z., Wang, Q., Chen, Y., 2020b. IVFuseNet: Fusion of infrared and visible light images for depth prediction. Inf. Fusion 58, 1–12.

Li, C., Zhu, C., Huang, Y., Tang, J., Wang, L., 2018. Cross-modal ranking with soft consistency and noisy labels for robust RGB-T tracking. In: Proceedings of the European Conference on Computer Vision. ECCV, pp. 808–823.

Liang, P., Jiang, J., Liu, X., Ma, J., 2022. Fusion from decomposition: A self-supervised decomposition approach for image fusion. In: European Conference on Computer Vision. Springer, pp. 719–735.

Liu, Y., Chen, X., Cheng, J., Peng, H., 2017. A medical image fusion method based on convolutional neural networks. In: 2017 20th International Conference on Information Fusion. Fusion, IEEE, pp. 1–7.

Liu, J., Fan, X., Huang, Z., Wu, G., Liu, R., Zhong, W., Luo, Z., 2022a. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5802–5811.

Liu, Y., Liu, S., Wang, Z., 2015. A general framework for image fusion based on multi-scale transform and sparse representation. Inf. Fusion 24, 147–164.

Liu, M., Meng, F., Liang, Y., 2022b. Generalized pose decoupled network for unsupervised 3d skeleton sequence-based action representation learning. Cyborg Bionic Syst..

Liu, J., Wang, X., Wang, C., Gao, Y., Liu, M., 2023. Temporal decoupling graph convolutional network for skeleton-based gesture recognition. IEEE Trans. Multimed..

Long, Y., Jia, H., Zhong, Y., Jiang, Y., Jia, Y., 2021. RXDNFuse: A aggregated residual dense network for infrared and visible image fusion. Inf. Fusion 69, 128–141.

Lu, Y., Wu, Y., Liu, B., Zhang, T., Li, B., Chu, Q., Yu, N., 2020. Cross-modality person re-identification with shared-specific feature transfer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13379–13389.

Ma, J., Chen, C., Li, C., Huang, J., 2016. Infrared and visible image fusion via gradient transfer and total variation minimization. Inf. Fusion 31, 100–109.

Ma, J., Liang, P., Yu, W., Chen, C., Guo, X., Wu, J., Jiang, J., 2020a. Infrared and visible image fusion via detail preserving adversarial learning. Inf. Fusion 54, 85–98.

Ma, J., Ma, Y., Li, C., 2019a. Infrared and visible image fusion methods and applications: A survey. Inf. Fusion 45, 153–178.

Ma, J., Xu, H., Jiang, J., Mei, X., Zhang, X.-P., 2020b. DDcGAN: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion. IEEE Trans. Image Process. 29, 4980–4995.

Ma, J., Yu, W., Liang, P., Li, C., Jiang, J., 2019b. FusionGAN: A generative adversarial network for infrared and visible image fusion. Inf. Fusion 48, 11–26.

Ma, J., Zhang, H., Shao, Z., Liang, P., Xu, H., 2020c. GANMcC: A generative adversarial network with multiclassification constraints for infrared and visible image fusion. IEEE Trans. Instrum. Meas. 70, 1–14.

Ma, J., Zhou, Z., Wang, B., Zong, H., 2017. Infrared and visible image fusion based on visual saliency map and weighted least square optimization. Infrared Phys. Technol. 82, 8–17.

Mostafavi, A., Cha, Y.-J., 2023. Deep learning-based active noise control on construction sites. Autom. Constr. 151, 104885.

Mou, J., Gao, W., Song, Z., 2013. Image fusion based on non-negative matrix factorization and infrared feature extraction. In: 2013 6th International Congress on Image and Signal Processing. CISP, Vol. 2, IEEE, pp. 1046–1050.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al., 2019. Pytorch: An imperative style, high-performance deep learning library. Adv. Neural Inf. Process. Syst. 32.

Peng, C., Tian, T., Chen, C., Guo, X., Ma, J., 2021. Bilateral attention decoder: A lightweight decoder for real-time semantic segmentation. Neural Netw. 137, 188–199.

Qu, G., Zhang, D., Yan, P., 2002. Information measure for performance of image fusion. Electron. Lett. 38 (7), 1.

Ram Prabhakar, K., Sai Srikar, V., Venkatesh Babu, R., 2017. Deepfuse: A deep unsupervised approach for exposure fusion with extreme exposure image pairs. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4714–4722.

Rao, Y.-J., 1997. In-fibre Bragg grating sensors. Meas. Sci. Technol. 8 (4), 355.

Roberts, J.W., Van Aardt, J.A., Ahmed, F.B., 2008. Assessment of image fusion procedures using entropy, image quality, and multispectral classification. J. Appl. Remote Sens. 2 (1), 023522.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.-C., 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4510–4520.

Sun, Y., Cao, B., Zhu, P., Hu, Q., 2022. Detfusion: A detection-driven infrared and visible image fusion network. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 4003–4011.

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. Inf. Fusion 82, 28–42.

Toet, A., Hogervorst, M.A., 2012. Progress in color night vision. Opt. Eng. 51 (1), 010901.

Wang, J., Peng, J., Feng, X., He, G., Fan, J., 2014. Fusion method for infrared and visible images by using non-negative sparse representation. Infrared Phys. Technol. 67, 477–489.

Wang, X., Zhang, W., Wang, C., Gao, Y., Liu, M., 2024. Dynamic dense graph convolutional network for skeleton-based human motion prediction. IEEE Trans. Image Process..

Woo, S., Park, J., Lee, J.-Y., Kweon, I.S., 2018. Cbam: Convolutional block attention module. In: Proceedings of the European Conference on Computer Vision. ECCV, pp. 3–19.

Xu, H., Ma, J., Jiang, J., Guo, X., Ling, H., 2020a. U2Fusion: A unified unsupervised image fusion network. IEEE Trans. Pattern Anal. Mach. Intell. 44 (1), 502–518.

Xu, H., Ma, J., Le, Z., Jiang, J., Guo, X., 2020b. Fusiondn: A unified densely connected network for image fusion. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 34, pp. 12484–12491.

Zhang, Q., Fu, Y., Li, H., Zou, J., 2013. Dictionary learning method for joint sparse representation-based image fusion. Opt. Eng. 52 (5), 057006.

Zhang, Y., Liu, Y., Sun, P., Yan, H., Zhao, X., Zhang, L., 2020. IFCNN: A general image fusion framework based on convolutional neural network. Inf. Fusion 54, 99–118.

Zhang, H., Ma, J., 2021. SDNet: A versatile squeeze-and-decomposition network for real-time image fusion. Int. J. Comput. Vis. 129, 2761–2785.

Zhang, X., Ma, Y., Fan, F., Zhang, Y., Huang, J., 2017. Infrared and visible image fusion via saliency analysis and local edge-preserving multi-scale decomposition. J. Opt. Soc. Amer. A 34 (8), 1400–1410.

Zhang, Y., Xu, X., Zhao, Y., Wen, Y., Tang, Z., Liu, M., 2024. Facial prior guided micro-expression generation. IEEE Trans. Image Process..

Zhao, J., Chen, Y., Feng, H., Xu, Z., Li, Q., 2014. Infrared image enhancement through saliency feature analysis based on multi-scale decomposition. Infrared Phys. Technol. 62, 86–93.