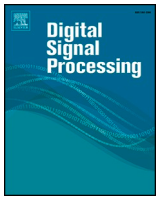
Digit. Signal Process. 168 (2026) 105493

Contents lists available at ScienceDirect

Digital Signal Processing

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



Multi-focus image fusion via multi-scale attention and Siamese networks

Hao Zhai [a], Nannan Luo [a,*], You Yang [b], Zhendong Xu [a], Bo Lin [a]

[a] College of Computer and Information Science, Chongqing Normal University, Chongqing 401331, China
[b] National Center for Applied Mathematics in Chongqing, Chongqing 401331, China

ARTICLE INFO

Keywords:
Multi-focus image fusion
Multi-scale dilated convolution
Siamese networks
Local feature attention
Deep learning

ABSTRACT

Multi-focus Image Fusion (MFIF) technology aims to generate a full-focus image with an extended focus range by combining multiple images with different focal depths. This has significant implications in fields such as image restoration and medical imaging. This paper proposes a new MFIF method based on deep learning, which utilizes multi-scale attention and a Siamese network structure to efficiently extract local depth features from images and enhance the fusion effect. The design of the Siamese network structure allows the model to process paired multi-focus images and share the feature extraction process in the deeper layers of the network. This not only enhances the expressive capability but also improves the model's ability to recognize images with different focal depths. Consequently, the network can effectively capture local depth features, which provides rich information for subsequent fusion. By incorporating a multi-scale dilated convolution attention module, which dynamically adapts the receptive field size to encompass a larger number of pixels, the process of information aggregation is facilitated across a wider area, thereby enhancing the optimization of the feature reconstruction process. Furthermore, binary segmentation and small-area filtering methods are employed to enhance the consistency of the fused image. Experimental results show that the proposed method surpasses existing multi-focus image fusion methods in terms of both subjective visual effects and objective evaluation metrics.

1. Introduction

Due to the camera's depth of field (DOF) constraints, achieving a fully focused image is often challenging. Objects within the depth of field remain sharp, while those outside range become blurred, making it difficult to obtain a completely clear image during practical photography. Multi-focus image fusion (MFIF) technology creates an image that is overall in focus by merging multiple images with different focal points. MFIF technology has been widely applied in various fields, including image restoration, medical imaging fusion [1], visual sensor networks, visual power detection [2], and optical microscopy observation [3]. MFIF technology is primarily categorized into three categories: spatial domain-based methods, transform domain-based methods and deep learning methods.

The MFIF technology based on the spatial domain [4–7] operates directly within the spatial domain of an image. It can be subdivided into three fusion methods: pixel-based [8], block-based [9], and region-based [10]. The pixel-based method merges corresponding pixels from multiple images directly. While this method is simple, it may not fully preserve the details and textures of the images. The block-based method divides the image into small blocks and fuses the pixels within each block. However, this approach is vulnerable to noise and highly dependent on the fusion strategy. The region-based method divides the image into multiple non-overlapping regions and fuses the pixels within each region. This approach more effectively preserves the details and textures of the image but requires careful selection of the region division method and fusion strategy. While spatial domain methods are effective in addressing intricate image features, they can be susceptible to noise and inconsistencies as they depend on pixel-level weighted averages. As a result, they frequently necessitate preprocessing and postprocessing procedures.

In addition to spatial domain-based methods, transform domain-based MFIF methods, such as wavelet transform and Fast Fourier Transform (FFT), achieve image fusion by analyzing and representing images with varying focal points in the transform domain. This technique first analyzes the features of images with different focal points in the transform domain, and then generates a fused image by weighted summation of the transform domain coefficients. The determination of the weighting coefficients is usually based on the feature analysis of the image's energy, contrast, and other characteristics to ensure a smooth

* Corresponding author.
E-mail address: 2023110516027@stu.cqnu.edu.cn (N. Luo).

https://doi.org/10.1016/j.dsp.2025.105493

Available online 19 July 2025

transition between the foreground and background while preserving details. Compared to traditional transform domain-based methods [11–14], the transform domain-based MFIF methods are more effective in extracting image features and details, and they exhibit better robustness against noise and inconsistencies. However, the drawback of this method is its high computational cost, as it requires transform domain analysis and inverse transform processing, and it has certain requirements for the choice of transform domain and parameter settings.

Recent studies have highlighted deep learning's success in image processing. These methods are divided into regression-based [15–17], directly mapping source to synthetic images, and classification-based [18–20], which classify pixel focus using neural networks. The latter can introduce edge artifacts and exaggerate differences due to neural networks' limited expressiveness. Liu et al. [15] pioneered neural network applications in multi focus image fusion, advancing the field. Multi focus image fusion techniques include discriminative, which detect focus for null-space fusion, generative, which create synthetic images, and multi-task learning methods. The latter, despite requiring complex strategies and more parameters, offer better fusion outcomes by overcoming the limitations of the first two.

To address the shortcomings of aforementioned multi-focus image fusion (MFIF) algorithms in terms of robustness, multi-scale feature utilization, and useful information retention, this research proposes an innovative MFIF algorithm. The algorithm incorporates a local feature attention module and a Siamese network architecture to enhance its performance and overcome existing limitations. Initially, a pair of complementary raw images and first use the GCB (GenClean Block) to denoise, improving image quality and enhancing feature representation. A LAM (Local Attention Module) block is combined with residual blocks to construct the Siamese network, effectively extracting local features while preventing overfitting to enhance the robustness of the network. In the feature extraction phase, we introduce the DFEB (Deep Features Extraction Block) module, which can more efficiently extract deep features from the images. Additionally, in the feature reconstruction process, we introduce the DCA (Dilated Convolutional Attention) module, which improves boundary recognition capabilities and more comprehensively and accurately extracts multi-scale global feature information. To reduce information loss from the original images, we directly generate the fused image. The main contributions of the proposed method are as follows:

(1) Through the local feature attention module and the DFEB module, the ability to extract image features is effectively enhanced, to capture local deep features. A Siamese network architecture has been developed, incorporating residual modules to enable deep extraction of intricate image features. This approach aims to mitigate the issue of gradient vanishing and enhance the efficiency of feature extraction.
(2) Introduced a multi-scale dilated convolution attention module, which enhances boundary recognition capabilities through multi-scale feature extraction and attention mechanisms, allowing for a more comprehensive and accurate extraction of multi-scale global feature information, thereby effectively enhancing the quality of fused images.
(3) To validate the performance of the proposed method, we conducted comparative experiments on three public test sets. These experiments involved a comprehensive evaluation against 11 classic methods. The experimental results demonstrate the superiority and effectiveness of the proposed method.

The subsequent sections of this paper are organized as follows. In Section 2, we provide a brief overview of related work relevant to the methods proposed in this paper. Section 3 details the proposed methods and the construction of the dataset. Section 4 presents the experimental details and comparative tests. Finally, Section 5 provides a summary of our work, highlights key findings, and discusses potential future directions.

2. Related work

2.1. Traditional methods

Based on spatial domain methods, pixel processing is directly performed through a specific algorithm for image fusion. Li, Kang, and Hu (2013) [21] proposed a new method in 2013 based on guided filtering weighted averaging. This method utilizes average filtering to decompose the source image into a base layer and a detail layer at dual scales, fully utilizing spatial consistency to fuse the base and detail layers. However, guided filtering methods are sensitive to registration errors, leading to blurred edges in the fused image. Liu et al. (2015) [22] proposed a multi-focus image fusion method based on dense scale-invariant feature transform (DSIFT). The initial decision map was obtained by measuring the activity level of source image blocks using DSIFT, and the best decision map is obtained through feature matching and comparing local fusion methods. However, the fixed size of image blocks resulted in a blocking effect. To address the blocking effect, region-based image fusion methods have been subsequently proposed. Li, Kang, Hu, and Yang (2013) [23] proposed a method for obtaining fused images through morphological filtering, which initially segments and obtains an accurate decision map through image matting. Recently, Chen et al. (2021) [24] proposed a fusion method based on multidimensional gradients and image matting (MGIMF), which differs from Li, Kang, Hu, and Yang (2013) [23]. Despite the excellent fusion results of these methods, the application of matting consumes a significant amount of time. Zhang et al. (2017) [25] proposed an edge detection method for multi-focus image fusion (BF). Ma et al. (2019) [26] introduced a random walk model based on dual scale focus images, which fuses complementary information from small scale focus images and large scale focus images. Bai et al. (2015) [27] proposed a multi-focus image fusion method based on a quadtree decomposition strategy. Additionally, Ma et al. (2017) [28] proposed a multi-focus image fusion method (GRW) based on multi-scale fusion measurement and generating random walks. Recently, Wang et al. (2022) [29] proposed a multi-focus image fusion method based on optimal block decomposition. Their innovative approach involves processing source images based on optimal decomposition and quadtree. However, these methods perform poorly in small-area fusion and fail to achieve satisfactory fusion results.

The transform domain methods hold an important position within the realm of image processing. They operate on the source image based on transform coefficients, utilizing methods such as Wavelet Transform (WT), Discrete Cosine Transform (DCT), gradient domain, and sparse representation. Subsequently, the image is restored to the spatial domain through inverse transformation. This method typically follows three main steps: image transformation, coefficient fusion, and inverse transformation (Liu et al., 2020) [30]. For instance, Wei and Ke introduced a multi-focus image fusion method based on Dual-Tree Complex Wavelet Transform (DT-CWT) in 2007 [31]. With a deeper understanding of the human visual system, multi-scale geometric analysis methods, such as those based on Non-Subsampled Contourlet Transform (NSCT), have been developed to better align with the way humans perceive images (Zhang and Guo, 2009) [32]. Furthermore, He et al. proposed the Guided Filter Fusion (GFF) algorithm in 2013 [33], which decomposes the image into a base layer and a detail layer. This method utilizes spatial consistency for weighted guided filter fusion, specifically addressing the fusion of low-frequency and high-frequency coefficients. Liu et al. introduced a fusion method based on Convolutional Sparse Representation (CSR) in 2016 [34]. This method involves learning a concise dictionary from the gradient information of a large number of high-quality image patches and selecting suitable sub-dictionaries for the source image patches to complete the fusion task. Although transform domain methods achieve smooth outcomes, they exhibit different sensitivities to various frequency components, which may lead to

brightness and color distortion. Consequently, enhancing image fidelity remains a research challenge. In 2018, Zhang et al. proposed a multi-focus image fusion method based on sparse representation [35]. This method concurrently considers the local information and spatial context information of the source images, effectively addressing the issue of artificial edges in the fused image. Amin-Naji and Aghagolzadeh proposed a DCT-based multi-focus image fusion method in the same year [36], introducing two novel fusion criteria: Laplacian energy and variance. Additionally, Zhou et al. proposed a multi-scale weighted gradient fusion method in 2014 [37], which addressed anisotropic blur and misalignment issues, similar to the related methods proposed by Kou et al. [38] and Wang [39] in 2018.

2.2. Deep learning based methods

With the rapid development of internet technology, deep learning methods have been widely applied in the field of image fusion (Tang et al., 2023) [40]. In recent years, many deep learning-based multi-focus image fusion methods have been proposed (Liu et al., 2018) [41]. Zhang et al. (2020) [42] introduced an innovative convolutional neural network framework that enables end-to-end training without any preprocessing. U2Fusion [43] is a groundbreaking algorithm in the MFIF field, which is a general method based on the transform domain capable of handling multi-modal, multi-exposure, and multi-focus image fusion tasks. U2Fusion cleverly integrates information from different images in the transform domain to generate fused images with greater depth of field and rich details. Additionally, unsupervised learning-based multi-focus image fusion methods have begun to emerge, reducing reliance on dataset labels. Jung et al. (2020) [44] proposed an unsupervised network for MFIF (DIF-Net), whose loss function is trained based on the vector representation of images. Generative Adversarial Network (GAN) methods, as powerful generative models, have been widely applied in image generation and MFIF. The GAN network consists of a generator and a discriminator, where the generator focuses on producing high-quality fused images, while the discriminator continuously improves its ability to judge image quality. For example, methods like MFFGAN [45] and mif-GAN are based on GAN networks and perform pixel-level image fusion. Recently, Bouzos et al. (2023) [46] proposed a convolutional neural network (mf-CNNCRF) that is robust to noise based on conditional random fields. Furthermore, Ma et al. (2022) [47] proposed a unified fusion structure based on the Swin Transformer. Cheng et al. (2023) [48] introduced the MUFusion structure, which is based on a memory unit that uses intermediate fusion results during the training process. However, these methods have significant limitations as they heavily depend on the quality of the training data. Therefore, Hu et al. (2023) [49] proposed a zero-shot multi-focus image fusion method (ZMFF) based on deep prior networks to address this issue. To solve the problem of multi-focus image fusion methods being unable to accurately identify small out-of-focus (or in-focus) areas covered by large focus regions, as well as issues like edge blurriness, Qi et al. (2024) [50] proposed a multi-focus image fusion framework based on a multi-channel Rybak neural network (MCRYNN), which innovatively utilizes the information interaction effect of the multi-channel network structure to accurately generate decision maps. Recently, Chen et al. (2024) [51] addressed the shortcomings of existing multi-focus image fusion methods that do not consider the continuous blur changes in close-up photography and camera shake issues in sequential image capture, as well as the inability to process multiple images simultaneously, by proposing an end-to-end deep learning network (DSAF-Net) for generating clear panoramic images from multi-focus and non-aligned images. Ouyang et al. [52] proposed a method for multi focus image fusion using superpixel feature generation GCN and pixel-level feature reconstruction CNN.

2.3. Attention mechanism and Siamese network

2.3.1. Attention mechanism

In the field of computer vision, attention mechanisms have gained significant attention for their ability to focus on key areas of an image while ignoring irrelevant parts. The human visual system efficiently understands complex scenes through similar mechanisms, prompting researchers to introduce them into computer vision to enhance performance. The development of attention mechanisms over the past decade can be divided into four stages: The first stage began with the RAM (Recurrent Attention Model) [53], which was the first to combine deep neural networks with attention mechanisms, using policy gradients to repeatedly predict key areas and update the network. The STN (Spatial Transformer Network) [54] was also an important achievement of this stage, as it used a sub-network to predict affine transformations to select important regions. The second stage is characterized by the explicit prediction of discriminative input features. DCNs (Deformable Convolutional Networks) [55] are a typical representative of this stage, employing deformable convolutions to better capture the shapes and positions of targets. The third stage is marked by the introduction of SENet (Squeeze-and-Excitation Network) [56], which introduced a novel channel attention mechanism that can implicitly and adaptively predicting key features. CBAM (Convolutional Block Attention Module) [57] is representative works of this stage, further strengthening the channel attention mechanism. The final stage is the era of self-attention. Vaswani et al. [58] first to propose the self-attention mechanism, achieving significant progress in the field of natural language processing. Recently, pure deep self-attention networks, such as Vision Transformers [59] have emerged one after another, demonstrating the immense potential of attention-based models. In this research, we further refines the feature extraction process by introducing attention mechanisms, including both channel attention and spatial attention dimensions. This mechanism allows our network to more precisely identify and highlight crucial information in images, achieving higher precision and resilience in image processing tasks. Consequently, our model can effectively capture the visual structure of images, mirroring the selective focusing characteristics of the human visual system.

2.3.2. Siamese network

The "Siamese" network architecture has been further developed by Chopra S and Hadse R, among others [60], who utilized Siamese networks for face similarity determination in their research. The Siamese network comprises two primary components: the initial segment is tasked with feature extraction, taking two images as input and outputting corresponding feature vectors; the second half measures the distance based on these feature vectors to assess the similarity between the input images, thereby achieving face similarity discrimination. Subsequently, Zagoruyko S and Komodakis N [61] made innovative improvements to the Siamese network in the literature, combining Spatial Pyramid Pooling (SPP) technology with the Siamese network for the first time [62]. The Siamese network is essentially a similarity measurement algorithm, particularly suitable for multi-class classification problems with limited sample sizes. In recent years, Siamese networks have been widely used in face recognition and similarity detection domains. In this research, the proposed method is based on the Siamese network for feature extraction to generate decision maps. The Siamese network consists of two identical branches that share weights and structure. This symmetrical design enables the network to effectively learn to extract discriminative features from paired samples.

3. Proposed methods

3.1. Network architecture

The autoencoder architecture has been increasingly utilized in multi-focus image fusion (MFIF) in recent years. However, in situations where

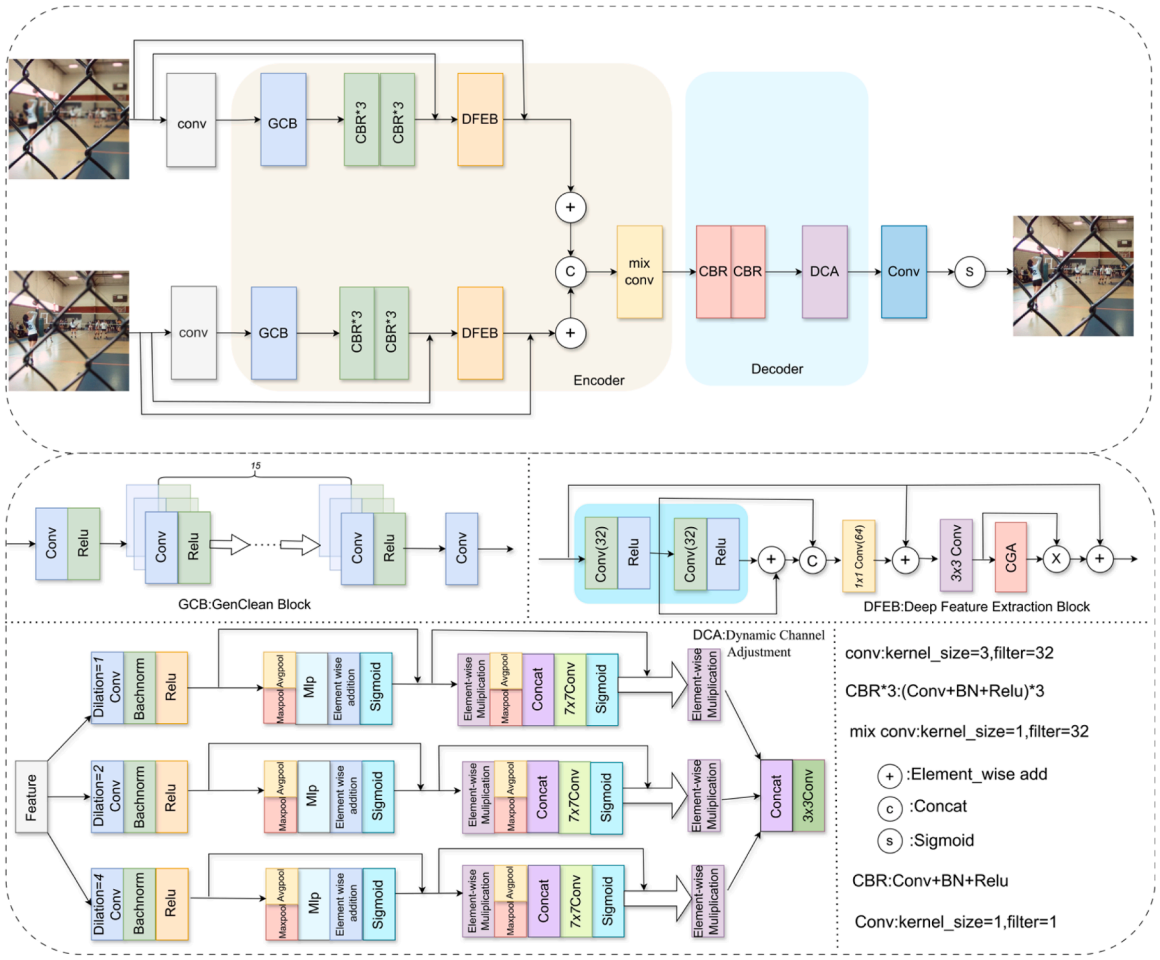


Fig. 1. Network structure diagram.The overall structure of the network is presented, including the Genclean Block (GCB), Deep Feature Extraction Block (DFEB), and Dynamic Channel Adjustment (DCA) module.

training data is scarce or the training cycle is limited, the issue of model overfitting becomes particularly prominent. To tackle this issue, this research proposes an improved strategy that integrates a Siamese network into the encoder and combines it with Convolution-BatchNorm-ReLU (CBR) [63] units to enhance the model's generalization ability. The Batch Normalization (BN) layer in the CBR unit helps stabilize data distribution and accelerates the model's convergence process. In the feature extraction phase, we first introduce a Generating Clear Image Block (GenClean Block, GCB) for noise suppression to strengthen the expressiveness and fusion accuracy. Subsequently, through continuous convolution and pooling operations, the input data is gradually abstracted to a higher level. The application of the BN layer further normalizes the mean and variance of each batch of data, limiting the fluctuation range of input features, thereby effectively accelerating network training and suppressing overfitting. To fully utilize the details and deep features, this paper introduces a Deep Feature Extraction Block (DFEB) to enhance feature learning, thereby improving the dehazing effect. Moreover, the ReLU activation function enhances the network's nonlinear expression capability by setting negative value features to zero, improving the effectiveness of feature extraction.

The decoder plays a crucial role in feature reconstruction by converting the extracted features back into their original input form. For example, in the MFF-GAN model used in the MFIF field, the decoder achieves feature reconstruction through deconvolution and skip connections to obtain clearer images. However, existing decoders may face issues of feature loss, failing to fully utilize multi-scale feature information. To overcome these limitations, this paper proposes a novel Dynamic Channel Adjustment (DCA) module. The DCA module dynamically adjusts channel and spatial weights during the feature fusion process, allowing the model to focus on key features, enhance feature expressiveness, mitigate overfitting, and expedite model convergence. The multi-scale residual structure within the DCA module further enhances feature expressiveness and image reconstruction quality, resulting in clearer and more detailed reconstructed images. Moreover, by incorporating convolution kernels of different sizes in the multi-scale residuals, feature extraction at different scales is achieved, and different scale feature maps are integrated through concatenation or addition to obtain a richer feature representation. The skip connection mechanism in the residual blocks effectively prevents information loss during the cross-scale feature extraction process. Fig. 1 illustrates the network model architecture proposed in this paper.

3.2. DCA module

In the realm of Convolutional Neural Networks (CNNs), the Convolutional Block Attention Module (CBAM) [57] proposed by Sanghyun W et al. has played a key role. In this paper, during the feature reconstruction phase, an innovative Dynamic Channel Adjustment (DCA) module is introduced based on the design concept of CBAM. The DCA module integrates Dilated Convolution and Attention Mechanism, demonstrating significant utility in image processing and computer vision tasks. Dilated Convolution expands the receptive field of convolution by introducing the concept of dilation into the convolution kernel, thereby increasing the range of input pixels that influence the

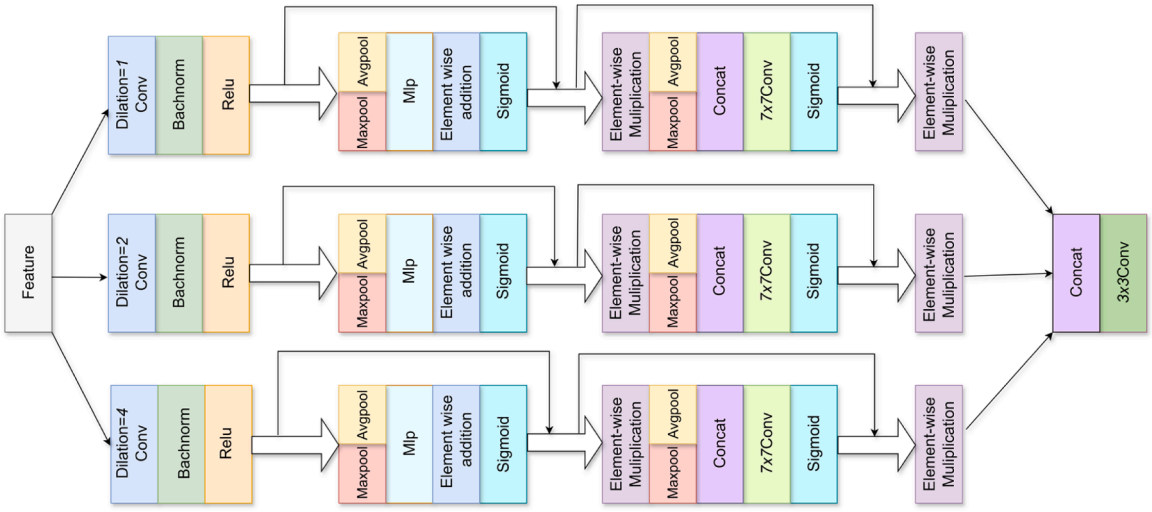


Fig. 2. DCA structure diagram.

convolution output. Compared to traditional convolution, dilated convolution can span more pixels, achieving information aggregation over a broader area. The attention mechanism further refines the feature extraction process, encompassing both channel attention and spatial attention. Channel attention enhances feature representation by assigning differentiated importance weights to different channels, while spatial attention focuses on key spatial regions of the image, improving the localization capability of the features.

In the DCA module, the feature maps obtained from dilated convolution and attention mechanisms can be effectively fused through weighted summation or feature concatenation. The integration facilitates the incorporation of multi-scale and multi-focus area information. Following the feature fusion process, a nonlinear activation function, such as ReLU or Leaky ReLU, is typically used to enhance the model's nonlinear expressive capability. The design of the DCA module not only works in synergy with other convolutional modules, batch normalization, and pooling layers to construct a more complex neural network architecture, but also significantly improves the model's deep understanding of image content by combining the advantages of multi-scale feature extraction from dilated convolution and the focusing characteristics of attention mechanisms, thereby optimizing performance across various visual tasks.

In this research, the DCA module receives feature inputs extracted by the encoder and processes them through a series of steps to enhance feature representation. First, the features are passed through three convolutional layers with dilation rates of 1, 2, and 4, respectively, to achieve multi-scale feature extraction. Subsequently, a ReLU activation function is applied, and the feature maps are downsampled from $C \times H \times W$ to $C \times 1 \times 1$ using max-pooling and average-pooling. This downsampling process is implemented through a shared multi-layer perceptron (MLP) which first compresses the number of channels and then restores it to the original number of channels, applying the ReLU activation function again to produce two activation outputs.

These outputs are combined through element-wise addition and use the Mc (F)sigmoid activation function to generate the channel attention output Mc (F), as shown in formula (1).

$$M_c(F) = \sigma(MLP(Avg(F)) + MLP(Max(F))) \tag{1}$$

The output is multiplied by the original feature F to obtain the adjusted feature F', as shown in formula (2).

$$F' = M_c(F) \otimes F \tag{2}$$

Furthermore, the spatial feature maps obtained through max pooling and average pooling are concatenated and subjected to convolution operations to generate a 1-channel spatial attention map Ms (F), as shown in formula (3).

$$M_s(F) = \sigma\left(f^{7x7}([Avg(F); Max(F)])\right) \tag{3}$$

Finally, the output of the spatial attention is multiplied by the adjusted feature F', restoring it to the original feature $C \times H \times W$, as shown in formula (4).

$$F'' = M_s(F') \otimes F' \tag{4}$$

To minimize the loss of original feature information, the DCA module uses residual connections to extract key information from the input features and concatenates it with the reduced-dimensional features for efficient feature fusion. Fig. 2 shows the architecture of the DCA module proposed in this paper, which effectively retains important feature information and optimizes the feature fusion process through the aforementioned operations. With this design, the DCA module not only enhances the expressive power of the features but also ensures the integrity and richness of the original information during the feature reconstruction process through a residual learning mechanism, thereby achieving better performance in visual tasks.

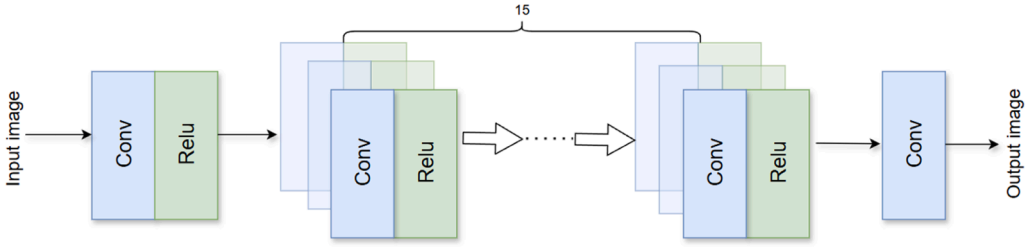


Fig. 3. GCB structure diagram.

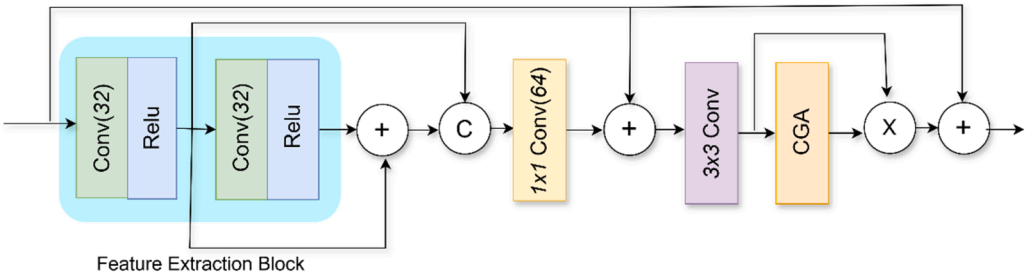


Fig. 4. DFEB structure diagram.

3.3. GCB module

The DnCNN [64] model is used as the clean generator in this research, with the objective of reconstructing noise-free output images. This module is a convolutional neural network module called GenClean Block, designed for the image denoising task in multi-focus image fusion. The model is intended to learn and generate high-resolution, low-noise fused images from a sequence of multi-focus images. The GCB module adopts a classic convolutional neural network structure, consisting of multiple convolutional layers and the ReLU activation functions. The module takes a sequence of multi-focus images as input. It contains 17 convolutional layers, each using a 3×3 convolution kernel and 1-pixel padding to maintain the output image size. Each convolutional layer is followed by a ReLU activation function, introducing non-linearity and enhancing the model's fitting ability. The last convolutional layer converts the feature maps back to the number of channels of the input image, resulting in a denoised clean image. The GCB module is trained using the multi-focus image sequence, learning the image reconstruction task by minimizing the loss function between the predicted image and the ground truth image. Commonly used loss functions include Mean Squared Error (MSE) loss and Structural Similarity (SSIM) loss, among others.

Fig. 3, 4, 6

The trained GCB module is used for multi-focus image fusion tasks, merging a sequence of multi-focus images into a high-quality fused image. The resulting fused image has a high resolution and low noise level, effectively enhancing the visual effect and application value of the image. This model has advantages such as a simple structure, fewer parameters, and stable training.

3.4. DFEB module

Inspired by DEAB [65], we propose the DFEB module. It effectively extracts image features and provides attention guidance by extracting rich features and directing the model's focus to important areas in each channel, which can significantly enhance the effectiveness of image fusion. This module mainly consists of two neural network components: CGA (content-guided attention) and Feature Extraction Block.

The Feature Extraction Block uses two convolutional layers for feature extraction and performs residual connections to enhance the expressive power of the features. The extracted features are concatenated along the channel dimension to obtain a richer feature map. A convolutional layer is then used to reduce the dimensionality of the concatenated feature map, resulting in the final output feature map.

The function of the CGA [57] module is to generate a unique Spatial Importance Map (SIM) for each channel of every input feature map. This process aids in directing the model's attention towards significant regions within each channel, consequently enhancing the image fusion performance. The module offers several advantages, including channel specificity by creating a unique SIM for each channel, which effectively addresses feature-level non-uniformity. Moreover, it merges both coarse and fine versions of the SIMs to more accurately capture spatial details. By facilitating information exchange through the combination of channel attention and spatial attention weights, the module ensures improved model performance.

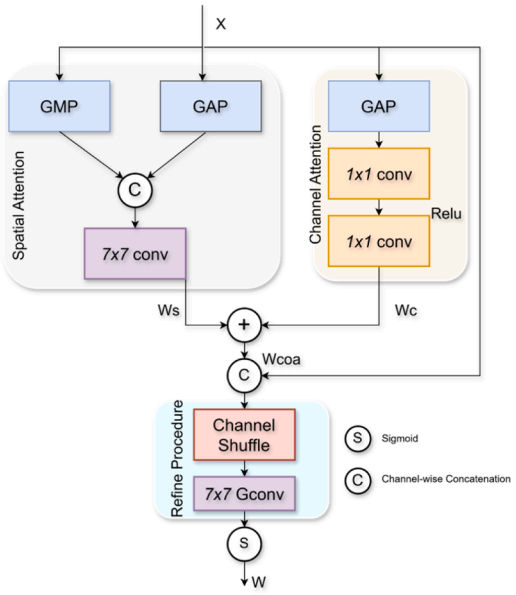


Fig. 5. The diagram of content-guided attention(CGA).

The detailed process of CGA is shown in Fig. 5. CGA is a coarse-to-fine process:the coarse version of SIMs(i.e.,Wcoa $\in R^{C\times H\times W}$)is generated firstly and then every channel is refined by the guided of input features. Let X $\in$ R^($C \times H \times W$) represent the subsequent input features. The goal of CGA is to generate SIMs for specific channels that have the same dimensions as X (i.e., W $\in$ R^($C \times H \times W$)). We first calculate the corresponding Wc and Ws using the following formulas [57,66].

$$\begin{aligned} W_C &= C_{1x1}\left(\max\left(0, C_{1x1}\left(X^c_{GAP}\right)\right)\right) \\ W_s &= C_{7x7}\left(\left[X^s_{GAP}, X^s_{GMP}\right]\right) \end{aligned} \tag{5}$$

In this research, we used the ReLU activation function, represented as max(0, x), and performed convolution operations with k x k convolution kernels, denoted as $C_{k\times k}(\cdot)$. Additionally, we employed channel-level concatenation operations, represented as $[\cdot]$. The features underwent global average pooling (GAP) across the spatial dimensions, denoted as X^c_{GAP}, global average pooling across the channel dimensions, denoted as X^s_{GAP}, and global maximum pooling (GMP) across the channel dimensions, denoted as X^s_{GMP}.

To reduce the number of parameters in the model and control the complexity of the model, a 1×1 convolution is applied for the first time to reduce the number of feature channels from C to C/r, where r is the reduction ratio. Subsequently, a second 1×1 convolution restores the number of channels to the original C. For this particular experiment, we choose to reduce the number of channels to a constant value of 16, so the reduction ratio r is set to C/16. This design strategy aims to optimize the network structure while maintaining computational efficiency and

Fig. 6. Visualisations of the training set for the VOC2012 dataset, including images from various sources along with their corresponding ground truths and masks.

performance.

Then, according to the broadcasting rules, we combine Wc and Ws through simple addition to obtain the coarse SIMs Wcoa $\in R^{C\times H\times W}$.

$$W_{coa} = W_c + W_s \tag{6}$$

In this research, we adopted a channel-sensitive alignment strategy to ensure the consistency of the weight matrix Wcoa with the input features X in the channel dimension. The design of the weight matrix Wcoa is based on channels, allowing for fine-tuning for each channel. Our method utilizes the intrinsic content of the input features to guide the generation of specific channel SIMs W, enabling customized adjustments for each channel. Specifically, Wcoa is reordered with each channel of X through a channel shuffling operation, which alternately adjusts the channel order [67]. This strategy not only enhances information interaction between channels but also, when combined with subsequent group convolution layers, effectively reduces the number of parameters in the model. In this way, we optimized the network structure, improved parameter utilization efficiency, while maintaining the model's performance and generalization ability.

$$W = \sigma(GC_{7x7}(CS([X, W_{coa}]))) \tag{7}$$

In this, σ represents the sigmoid operation, CS(·) denotes the channel shuffle operation, and $GC_{k\times k}(\cdot)$ indicates a group convolution layer with a kernel size of $k \times k$, where the group number is set to c in our implementation. CGA assigns a unique SIM to each channel, guiding the model to focus on the important areas of each channel. Therefore, more useful information encoded in the features can be emphasized.

DFEB first uses a Feature Extraction Block for feature extraction. The extracted features are connected to the input features via a residual connection to enhance the stability of the features. Convolutional layers are used for feature extraction, followed by a residual connection. Attention weights are then generated. The feature map is element-wise multiplied with the final pixel attention weights, and the result is connected to the input features through a residual connection to obtain the final output feature map. The advantages of this module include the extraction of rich features; through the Feature Extraction Block and convolutional layers, it can effectively extract image features. Guided by attention, through spatial attention, channel attention, and pixel attention, it can direct the model's focus to important areas within the image, thereby improving the performance of image fusion. Additionally, it is computationally efficient; by using the ReLU activation function and residual connections, it can enhance the stability and learning ability of the model, and by employing attention mechanisms, it can reduce the computational load.

3.5. Detailed integration plan

3.5.1. Binarization

Binarization technology simplifies images by limiting pixel values to two states: black and white. The core of this process is thresholding, which converts grayscale levels into binary form to enhance features or segment images. This process not only simplifies the image structure for easier analysis and processing but also aids in feature extraction and background separation while reducing storage and transmission requirements. In this research, binarization is used for the post-processing of decision maps, assisting in the identification of source image regions and image fusion. Specifically, we employ a 3×3 convolution layer and a sigmoid function to reduce the dimensionality of the feature map and map it to the range [0, 1], forming the decision map d. Through binarization, pixels in the decision map M with values greater than or equal to 0.5 are set to 1, while the rest are set to 0, resulting in the binary image B. This strategy effectively extracts key information while ensuring the retention of image features, providing a foundation for subsequent processing.

$$B(x,y) = \begin{cases} 1, & M(x,y) > 0.5 \\ 0, & M(x,y) \le 0.5 \end{cases} \tag{8}$$

3.5.2. Small-area filtering

Small-area filtering is a local image processing technique that focuses on specific local regions of an image to enhance quality or extract features while preserving detail and edge information. Compared to global filtering, it reduces noise while avoiding excessive blurring of edges. This technique plays an important role in image analysis and recognition

Fig. 7. Test set. The datasets used in the image from top to bottom are as follows: Grayscale, Lytro, MFFW and MFI-WHU.

by enhancing local features such as texture and shape, and it can smooth transition areas and adjusting contrast to improve visual presentation. In this research, we propose an area threshold-based algorithm to address the issue of small regions or "holes" in binary images. If the area of a region is <0.01 times the size of the image, it is classified as a small region, and a bit-flipping operation is performed to improve the consistency of the decision map. The mathematical model for small area filtering is described by formula (9), where the area(B_region) represents the area of the small region, B is the binary image, and D is the final decision map. This strategy effectively enhances the accuracy of binary image processing and provides a solid foundation for subsequent image analysis.

$$D(x,y) = \begin{cases} 1 - B(x,y), if\ area(B\_region) < 0.01 * H * W \\ B(x,y),\ \ if\ area(B\_region) \geq 0.01 * H * W \end{cases} \tag{9}$$

3.5.3. Image fusion

In this research, we adopted a weight-based graph fusion strategy to synthesize high-quality multi-focus images. This strategy merges two source images A and C, along with a decision map D, into a final image F according to specific rules. For color images, we employed channel-independent fusion methods, processing the red, green, and blue channels separately to ensure effective integration of features and details. The fused channels are then recombined to form a high-quality image F that is rich in details and distinct in features. This method not only enhances fusion accuracy and quality but also provides a new solution for multi-focus image processing, laying the foundation for image analysis and further processing.

$$F(x,y) = A(x,y) * D(x,y) + C(x,y) * (1 - D(x,y)) \tag{10}$$

4. Experiment

4.1. Experimental setup

4.1.1. Training dataset

In the field of Multi-Focus Image Fusion (MFIF), obtaining a training set of real images is a challenge. To address this issue, this research utilized the VOC2012 (Pascal Visual Object Classes 2012) dataset [68] to train the network, which is widely used in computer vision tasks such as image segmentation. Prior to the training phase, necessary preprocessing was performed on the dataset. We randomly selected 2900 images and their corresponding masks from the VOC2012 dataset. Considering the varying sizes of the original images, we standardized their dimensions to 256×256 pixels to fit the training process. Additionally, to generate focused images with different levels of blur, we applied Gaussian filtering three times to the adjusted images, each time using a different kernel size. This step aims to simulate the lack of a real image training set in the MFIF task. Through this method, we were able to prepare a set of images with varying focus levels for network training, allowing us to train an effective MFIF model even without a real MFIF training set. This data preprocessing and augmentation strategy provided the network with rich visual information, helping to improve its performance in image fusion tasks.

4.1.2. Test dataset

To evaluate the performance of the method proposed in this research, we selected four widely used public image datasets for assessment during the experimental phase, which include a grayscale image dataset [69], the Lytro dataset [70], the MFFW dataset [71], and the MFI-WHU dataset [72]. Among them, the grayscale image dataset

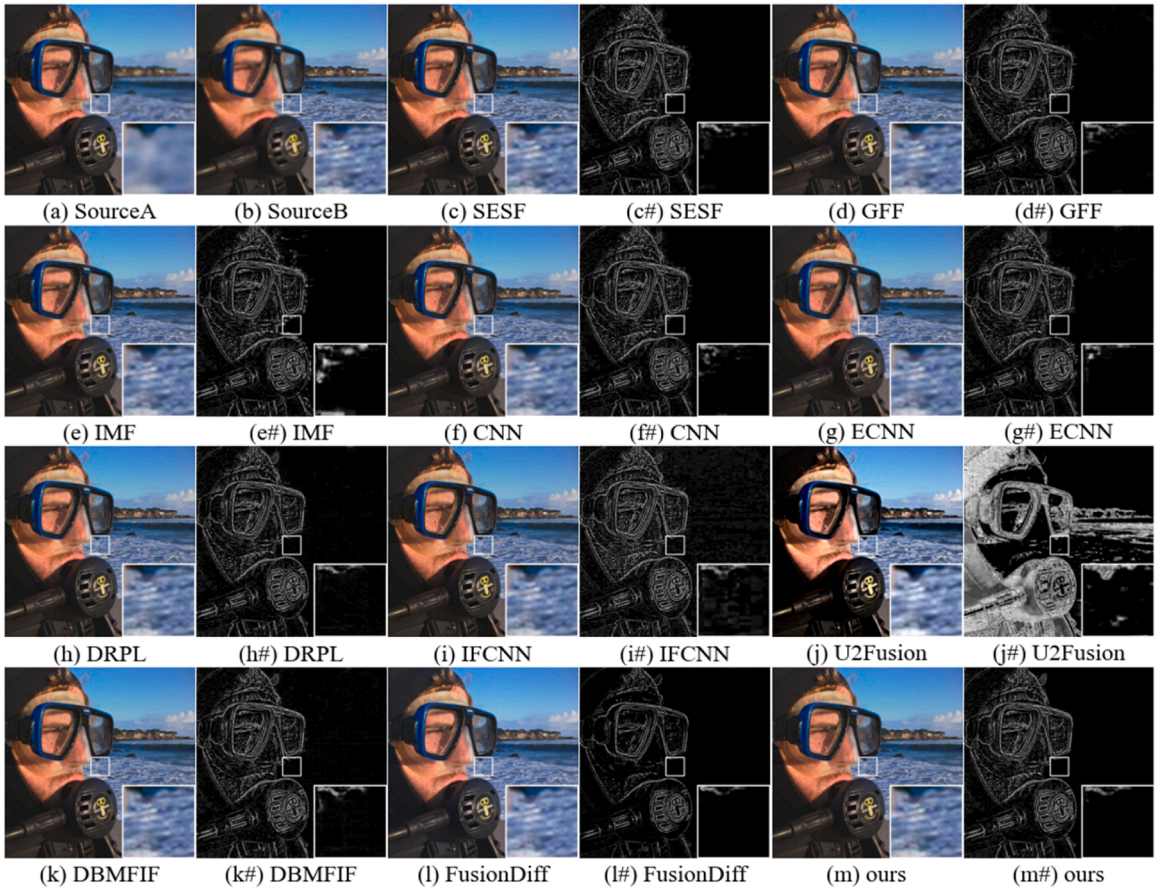


Fig. 8. "Man and Sea" from the Lytro dataset Fusion Results and Difference Map.

contains 20 pairs of commonly used grayscale multi-focus images, the Lytro dataset contains 20 pairs of color images, the MFFW dataset contains 13 pairs of color images, and the MFI-WHU dataset contains 30 pairs of commonly used color images. Fig. 7 shows representative samples from these datasets, with the first row displaying the grayscale dataset, the second row showing the Lytro dataset, the third row presenting the MFFW dataset, and the last row showcasing samples from the MFI-WHU dataset.

4.1.3. Implementation details

First, the images in the training and validation sets are read to train the network. The learning rate of the network is set to 0.0005, decreasing by a factor of 0.88 every two epochs. In this paper, we optimize the network parameters using the Adam optimizer. The batch size and the number of model training epochs are set to 8 and 200, respectively. Each epoch consists of two phases: training and validation. At the end of each epoch, the learning rate scheduler is updated, and it is checked whether the early stopping condition has been triggered. The EarlyStopping class is used to monitor the validation loss; if the validation loss does not improve for a consecutive number of epochs defined by the patience parameter, training is stopped. In this paper, the patience parameter is set to 6. The experiments in this paper were conducted on a computing platform equipped with a GTX 3060 Ti GPU (8GB VRAM) and an Intel i5–12490F processor. This experimental setup aims to ensure the accuracy and reproducibility of the results while providing a standardized evaluation benchmark for research in the field of image fusion.

4.1.4. Metrics

To objectively evaluate the performance of various fusion algorithms, we conducted a comprehensive assessment using six objective evaluation metrics across four datasets. These metrics are Q_{NMI} [73], Q_G [74], Q_{CB} [75], Q_{TE} [76], Q_{SF} [77], and Q_{Piella} [78]. Q_{NMI} primarily measures the information correlation between the source image and the fused image, assessing the degree of information transfer from the source image to the fused image. Q_G is based on image gradients and reflects the details and clarity of the image by quantifying the average gradient. Q_{CB} measures the similarity between the source and fused images in terms of major features that align with those perceived by the human visual system. Q_{TE} evaluates the similarity between the two images from the perspective of entropy by comparing the Tsallis entropy of the source and fused images. Q_{SF} serves as a measure of image clarity, with values closer to zero indicating a clearer image. Q_{Piella} measures the structural similarity between the source and fused images.

4.1.5. Fusion methods for comparison

In the comparative experiment, we selected 10 industry-recognized image fusion algorithms, including GFF [33], CNN [15], IMF [23], DBMFIF [79], ECNN [16], FusionDiff[80], DRPL [20], IFCNN [42], SESF [17], and U2Fusion [19]. All these algorithms were configured and executed according to the parameters recommended in their original papers. Among them, GFF is based on traditional methods, while the rest are based on deep learning methods.

4.2. Comparison with existing methods

4.2.1. Qualitative comparison

We conducted experiments using three representative color images: "Man and Sea" from the Lytro dataset, "Rocks and Sea" from the MFFW dataset, "Playing Tennis" from the MFI dataset, and "Stone Lion" from the Grayscale dataset. The experimental results of our proposed method were compared with those of other image fusion methods. For visual

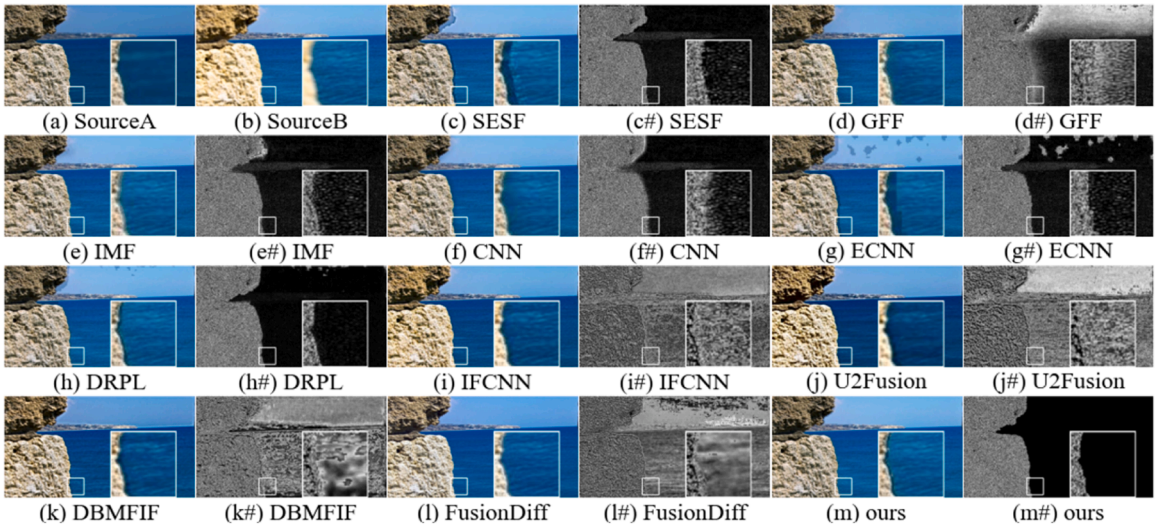


Fig. 9. "Stone and Sea" from the MFFW dataset Fusion Results and Difference Map.

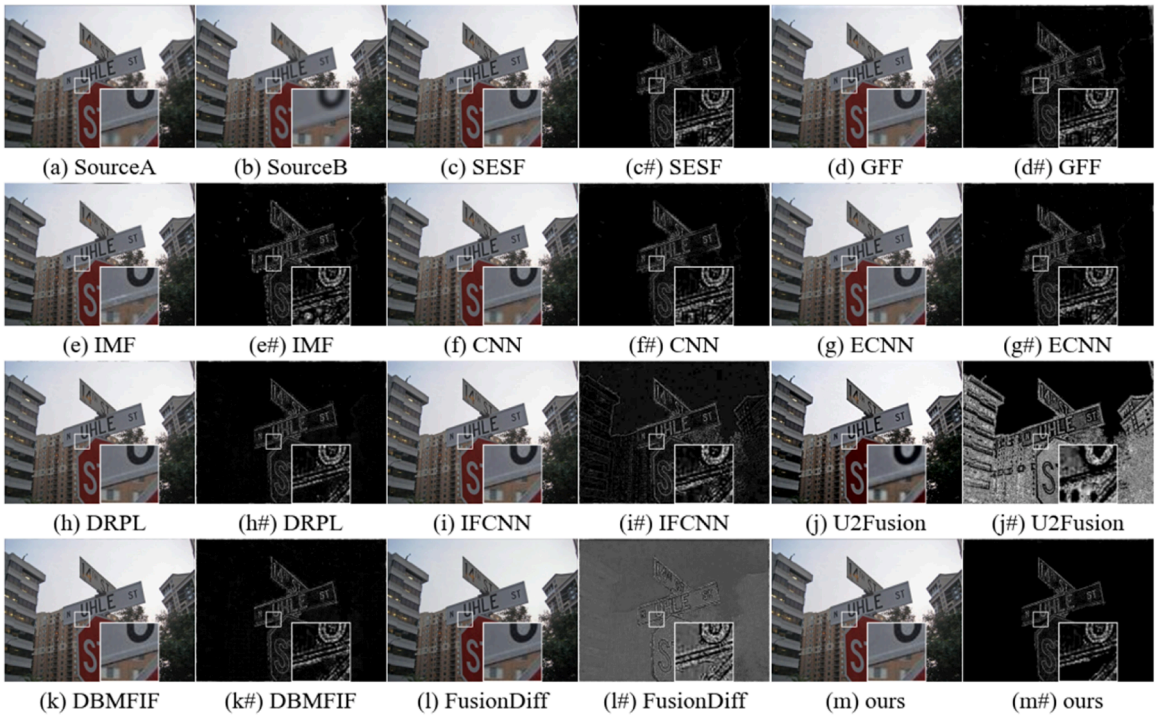


Fig. 10. "Road Marking" from the MFI dataset Fusion results and difference map.

comparison, we generated difference images by subtracting the distant focus source image from the fused image. The difference images for "Man and Sea" from the Lytro dataset, "Rocks and Sea" from the MFFW dataset, "Playing Tennis" from the MFI dataset, and "Stone Lion" from the Grayscale dataset are shown in the figures. From Fig. 8, it can be seen that IFCNN and U2Fusion were unable to fully detect the distant focus areas in this experiment, resulting in their difference images having the distant focus image as the background. GFF, CNN, IMF, and ECNN exhibited artifacts at the boundary between near and distant focus. As we can see from the figure, DBMFIF does not work as well on the border, and FusionDiff is slightly less effective at extracting the details of the near-focus image directly below the picture. In this experiment, only the SESF, DRPL, and our model achieved better results than the other models. In Fig. 9, we compared 11 models, among which SESF, IMF, GFF, CNN, ECNN, IFCNN, U2Fusion, DBMFIF and FusionDiff were not able to fully identify the near focus areas. The models GFF, and ECNN exhibited noise anomalies in the distant focus area, while GFF and CNN showed varying degrees of artifacts and light leakage at the boundaries. In Fig. 10, GFF, IMF, CNN, IFCNN, U2Fusion and FusionDiff did not effectively recognize the distant focus areas, leading to the presence of distant focus backgrounds in the difference images. As can be seen from the zoomed-in region, SESF, GFF, CNN, and ECNN cannot distinguish the boundary between near-focus and far-focus well. As shown in Fig. 11, compared with other SOTA methods, the method proposed in this paper not only avoids the region misclassification problem, but also maximizes the preservation of the edge details of the stone lions and achieves accurate segmentation. Fig. 11 demonstrates the excellent performance of the proposed network in processing both color and grayscale data. From Figs. 8, 9, 10 and 11, we can conclude that the proposed model not only outperforms other models in recognizing

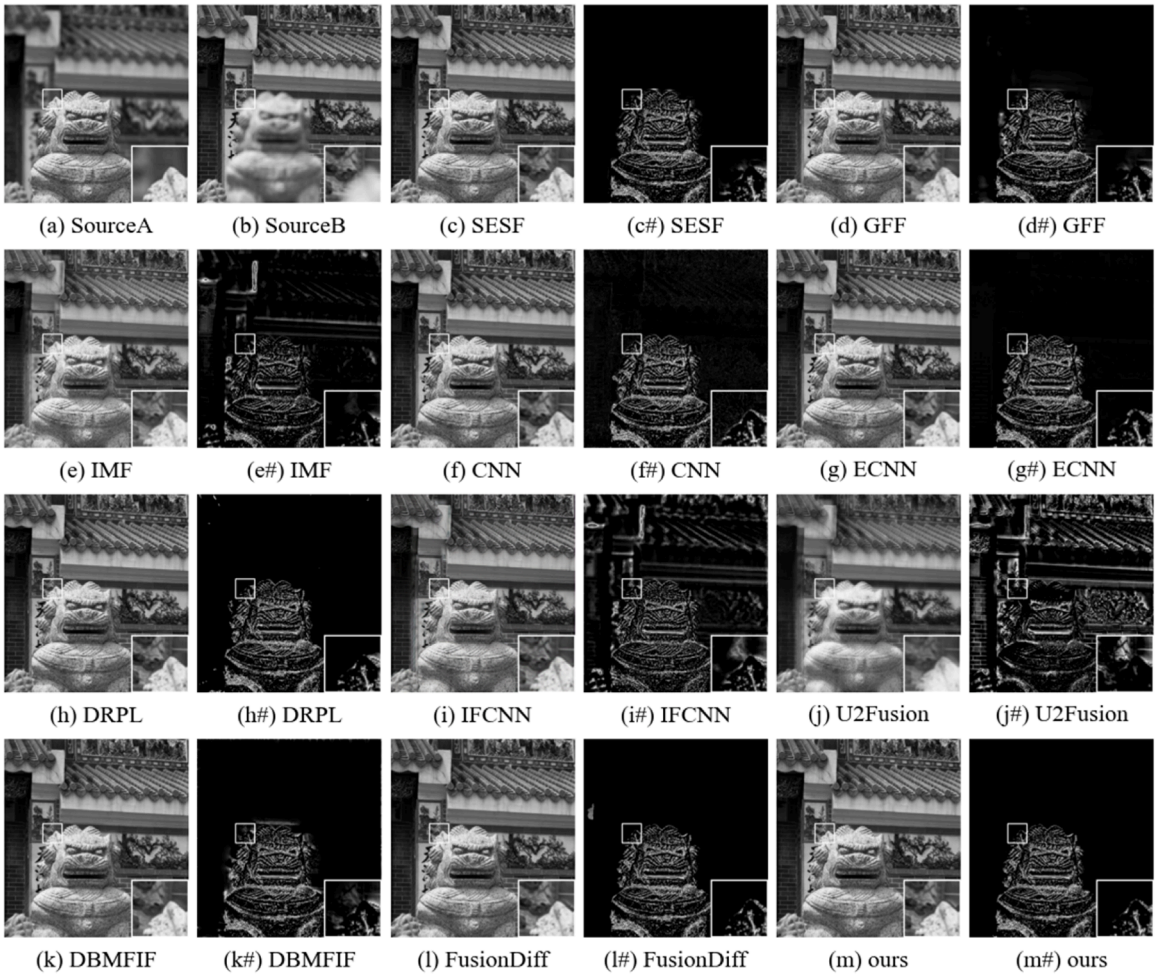


Fig. 11. "Stone Lion" from the Grayscale dataset Fusion results and difference map.

Table 1
The average value of objective indicators to evaluate the performance of the comparison methods and our proposed method on the GrayScale dataset.

Methods	Metrics					
	Q_{NMI}	Q_G	Q_{CB}	Q_{TE}	Q_{SF}	Q_{Piella}
SESF [17]	1.17045	0.70197*	**0.78356**	0.41698*	−0.03735	0.91503*
GFF [33]	1.05967	0.69251	0.76010	0.40977	−0.04603	0.91171
IMF [23]	1.07795	0.68418	0.76674	0.41059	−0.04735	0.91410
CNN [15]	0.83163	0.61144	0.66257	0.39573	−0.19937	0.91486
ECNN [16]	0.76374	0.55559	0.69529	0.39358	−0.03605	0.88316
DRPL [20]	1.19386*	0.69036	0.76617	0.41664	−0.02938	0.91454
IFCNN [42]	0.86369	0.61572	0.69850	0.39121	−0.04075	0.91436
U2Fusion [19]	0.80461	0.48614	0.60503	0.41654	−0.40932	0.86590
DBMFIF [79]	0.96347	0.63591	0.72581	0.40016	**−0.00607**	0.91041
FusionDiff [80]	0.86015	0.68582	0.68201	0.40263	−0.02981	0.91245
Ours	**1.20957**	**0.70574**	0.77653*	**0.42140**	−0.02934*	**0.91582**

distant and near focus areas but also excels in recognizing the boundary areas.

4.2.2. *Quantitative comparison*

This research used a test set consisting of 83 images, which includes 20 pairs of grayscale images, 20 pairs of Lytro images with a resolution of 520×520, 13 pairs of MFFW images, and 30 MFI test images. The bold numbers represent the optimal results, and the blue font indicates the suboptimal results. As shown in the results of Table 1, the proposed method demonstrates its effectiveness in grayscale image fusion. Except for QCB and QSF, all other metrics achieve optimal results. This indicates that the proposed method generally outperforms other algorithms in preserving key details and structural information in grayscale images. The superior performance of the proposed method can be attributed to its more precise feature extraction and more effective fusion of key information. Compared to the SESF method, the proposed method performs worse than SESF in the QCB metric, as SESF can better adjust the brightness and contrast of the image, making the fused image visually more layered. In the QSF metric, the DBMFIF method performs better, attributed to its advantage in capturing high-frequency details of the image, enabling it to more accurately preserve edge and texture information. As shown in Table 2, although the proposed algorithm ranks second in the QSF metric, it achieves the best performance in the remaining five metrics. This indicates that the proposed method has a

Table 2
The average value of objective indicators to evaluate the performance of the comparison methods and our proposed method on the Lytro dataset.

Methods	Metrics					
	Q_{NMI}	Q_G	Q_{CB}	Q_{TE}	Q_{SF}	Q_{Piella}
SESF [17]	1.10225	0.70268	0.79627	0.39477	−0.02222	0.94354
GFF [33]	1.06617	0.70489	0.79041	0.39404	−0.02780	0.94519
IMF [23]	1.09437	0.70088	0.78841	0.39249	−0.02164	0.94092
CNN [15]	1.10968	0.70791	0.79988*	0.39628	−0.03072	0.94513
ECNN [16]	1.12732*	0.70457	0.79580	0.39696	−0.02152	0.94134
DRPL [20]	1.09154	0.71305*	0.79420	0.39662	−0.01855	0.94454
IFCNN [42]	0.92684	0.66225	0.72342	0.38243	−0.02356	0.94521
U2Fusion [19]	0.77246	0.51516	0.56821	0.39941*	0.01067	0.83335
DBMFIF [79]	1.05885	0.69078	0.77747	0.39365	**−0.01377**	0.94511*
FusionDiff [80]	0.89209	0.63197	0.67784	0.39928	−0.03983	0.93513
ours	**1.22945**	**0.76305**	**0.80683**	**0.40402**	−0.01640*	**0.94522**

Table 3
The average value of objective indicators to evaluate the performance of the comparison methods and our proposed method on the MFFW dataset.

Methods	Metrics					
	Q_{NMI}	Q_G	Q_{CB}	Q_{TE}	Q_{SF}	Q_{Piella}
SESF [17]	0.74523	0.49274	0.67678	0.35343	0.01213	0.89375
GFF [33]	0.74564	0.50024	0.68031	0.36315	−0.03938	0.90621
IMF [23]	0.76353	0.50045	0.68000	0.36186	−0.02963	0.90045
CNN [15]	0.77379	0.49858	0.67828	0.36381	−0.04921	0.90457
ECNN [16]	0.75197	0.48554	0.67284	0.35871	−0.02756	0.89074
DRPL [20]	0.92944*	0.63383*	0.70754*	0.38653	**−0.00336**	0.91699
IFCNN [42]	0.75479	0.48228	0.62713	0.36257	−0.04626	0.90773
U2Fusion [19]	0.69922	0.42110	0.54987	0.38981	−0.03251	0.81420
DBMFIF [79]	0.87020	0.58581	0.66415	0.37836	−0.00695	**0.96125**
FusionDiff [80]	0.80625	0.57498	0.56875	0.40787*	−0.05754	0.88412
ours	**1.17817**	**0.70635**	**0.74994**	**0.41256**	−0.00675*	0.92268*

Table 4
The average value of objective indicators to evaluate the comparative methods and the performance of our proposed method on the MFI-WHU dataset.

Methods	Metrics					
	Q_{NMI}	Q_G	Q_{CB}	Q_{TE}	Q_{SF}	Q_{Piella}
SESF [17]	1.14466	0.73163	0.82131	0.39125	−0.02656	0.94765
GFF [33]	1.11751	0.72470	0.81641	0.38185	−0.02857	0.94807
IMF [23]	1.14651	0.72206	0.80296	0.38520	−0.02203	0.94186
CNN [15]	1.17412	0.73598	0.82720*	0.39077	−0.03106	0.94876*
ECNN [16]	1.19094*	0.73842*	0.82579	0.39158*	−0.02340	0.94803
DRPL [20]	1.10337	0.73140	0.81515	0.38660	**−0.01993**	0.94848
IFCNN [42]	0.91046	0.68618	0.73640	0.36608	−0.02687	0.94653
U2Fusion [19]	0.69906	0.50220	0.51557	0.37202	−0.06306	0.83827
DBMFIF [79]	1.07013	0.67432	0.78242	0.38335	−0.02199	0.94853
FusionDiff [80]	0.82402	0.65327	0.71576	0.38348	−0.03051	0.93273
ours	**1.22404**	**0.74172**	**0.83061**	**0.39751**	−0.02196*	**0.94899**

good performance balance across different aspects of image fusion. This is attributed to its advanced feature extraction and fusion capabilities. Its multi-scale attention mechanism efficiently identifies and highlights key information, leading to better results in most metrics. The twin network structure further enhances feature learning capabilities by effectively capturing the similarities and differences between input images, thereby improving the overall robustness and effectiveness of the method. In Table 3, the proposed method once again demonstrates its advantages in image fusion tasks. Except for QSF, all other results are optimal. This further confirms the effectiveness of the proposed method in capturing image visual structure and details. The success of this method on this dataset can be attributed to its ability to adaptively adjust attention weights based on input image features. The integration of channel and spatial attention mechanisms enables the model to effectively balance global and local features, thereby generating high-quality fusion images. The DRPL method performs better on the QSF metric because it better captures the mapping relationships between image pairs through its deep regression-based learning framework, thereby achieving outstanding performance on the QSF metric. Table 4 further demonstrates the strong performance of the proposed method. Except for the second-best QSF results, all other results are the best. This indicates that the proposed method can effectively handle different types of images and fusion challenges. The advanced attention mechanism and twin network structure in the method work synergistically to ensure high-quality fusion images, with key details and structural information preserved. These results indicate that the method proposed in this research outperforms the other 11 algorithms across multiple evaluation dimensions, thereby validating its effectiveness and superiority.

4.3. More analysis

4.3.1. Fusion results of more source images

To verify the generalization ability of the proposed method, this research further demonstrates the fusion effects of multiple sets of typical multi-focus source images. As shown in Fig. 12, the fusion results

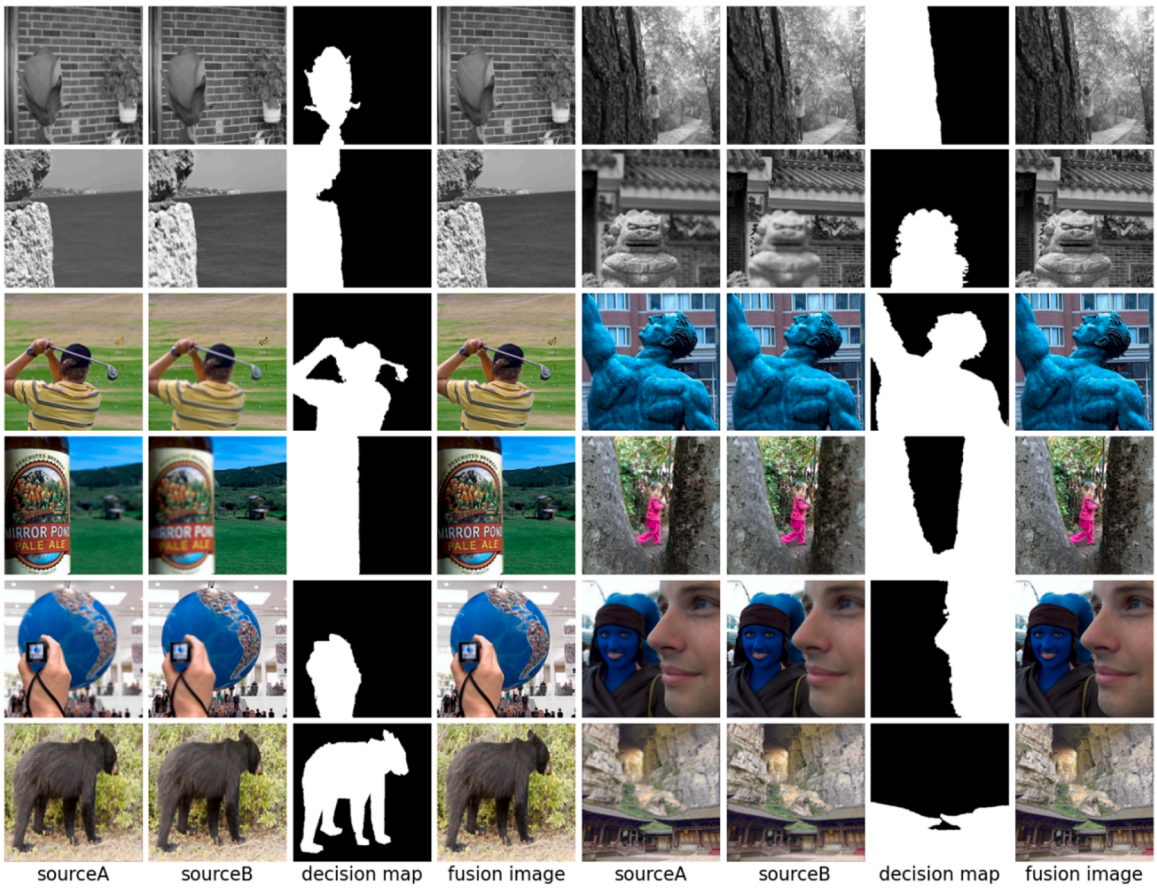


Fig. 12. More fusion results of multi-focus source images.

indicate that this method not only accurately identifies and extracts the focused areas in the source images but also exhibits excellent segmentation performance when processing transitional areas. This results in the final fused image achieving a high level of visual quality, showcasing an ideal fusion effect.

4.3.2. *Fusion results of multiple source images*

This research selected three highly complementary source images from the Lytro dataset to showcase the exceptional performance and innovation of the algorithm. Detailed image fusion experiments were conducted on these images, each containing distinct visual information at varying focal depths, making them ideal for algorithm testing. The fusion results are visually presented in Fig. 13, demonstrating the algorithm's capability to identify and merge focal areas from the source images effectively. The fused image exhibits an impressive all-focus effect, significantly enhancing clarity, image contrast, and detail representation. The algorithm's advanced fusion methods accurately preserve and present detailed foreground textures and smooth background transitions, showcasing its ability to identify and utilize complementary information from source images while maintaining naturalness and realism in the fused image. A noticeable improvement in overall image quality is observed when comparing the images before and after fusion. Previously blurred areas in the source images become clear, and objects with unclear edges show sharp contours in the fused image. These enhancements not only create a striking visual impact but also objectively validate the effectiveness and superiority of the algorithm.

4.3.3. *Efficiency*

In this research section, we conducted efficiency comparison experiments on the ten aforementioned state-of-the-art (SOTA) techniques. The experiments were executed on a computing system equipped with an NVIDIA GTX 3060 Ti GPU (with 8 GB of video random access memory) and an Intel i5–12490F central processor. In order to exclude the influence of performance differences between computing platforms on the results, we used the open source code provided by the authors of each method for independent testing. In particular, we note that for traditional algorithms that do not integrate CUDA optimizations, we will execute them on the CPU, while other algorithms utilize GPU resources. We recorded the average elapsed time for each method when merging 20 pairs of multi focus images of the same resolution (520 $\times$ 520 pixels) provided by Lytro. As shown in Table 5, most of the algorithms were able to complete the image fusion task in a relatively short time. Among all the evaluated algorithms, U2Fusion and IFCNN exhibit very short processing times, while CNN and ECNN take longer. Given that the processing time is $<$0.1 s, the proposed algorithms are considered feasible for practical applications. The additional computation time of the method is reasonable in order to obtain better fusion results. Compared with other SOTA techniques, its efficiency shows significant competitiveness.

4.3.4. *Fusion results under low light conditions*

In various photographic settings, low-light conditions are frequently encountered, leading to image noise as imaging devices compensate for insufficient illumination to enhance brightness. This issue poses a challenge for industries demanding high-precision imagery. Particularly in medicine, procedures like CT scans and MRIs require capturing clear images in dim settings, which are vital for diagnostic purposes and health assessment. Similarly, in security surveillance, cameras must function effectively in low-light scenarios to capture clear footage essential for identifying and documenting events. To assess algorithm performance objectively and offer insights for future research, we selected a low-light multifocal image fusion (MFIF) dataset for

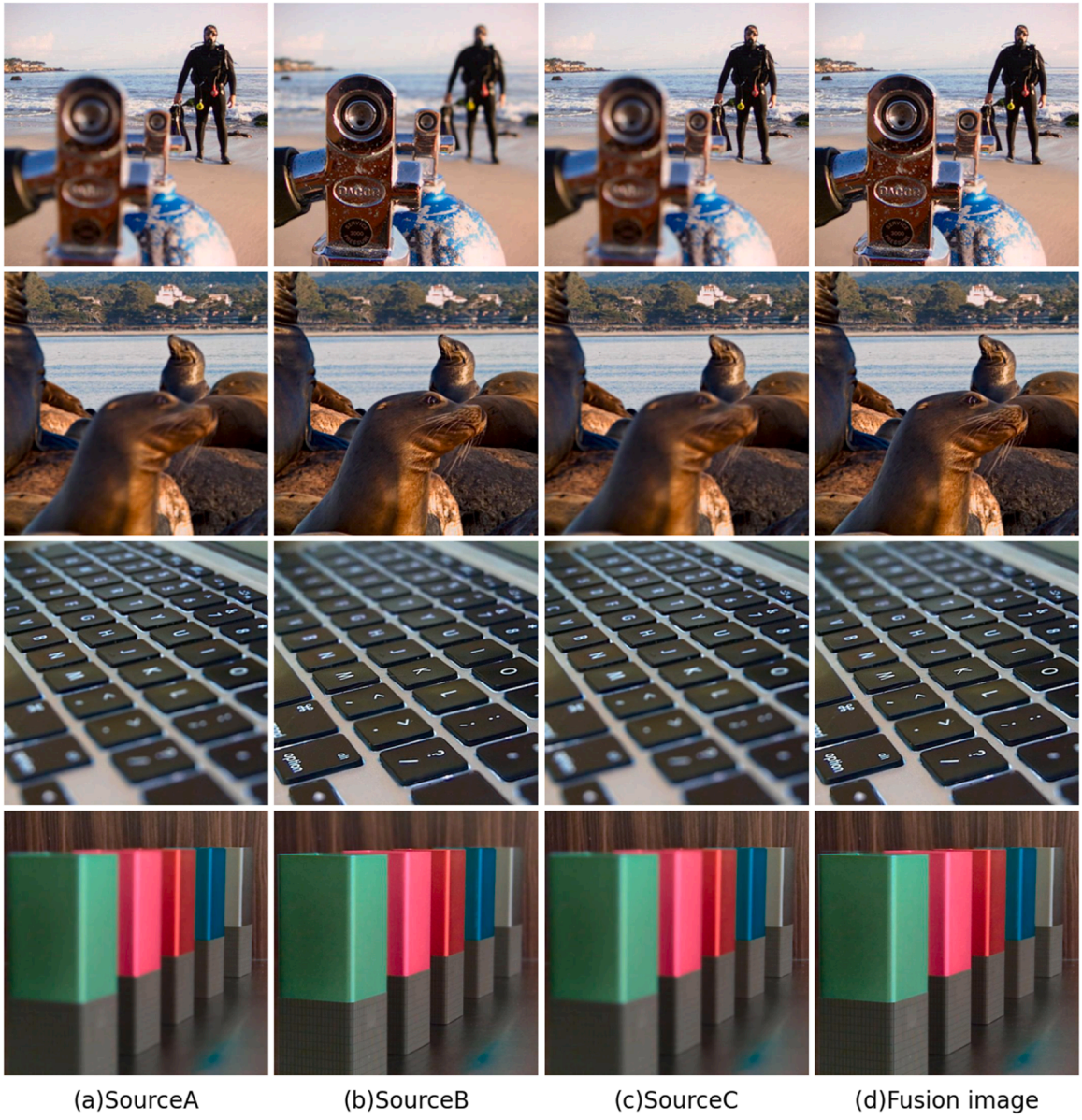


Fig. 13. Fusion of three source images.

Table 5
The average running time of different methods on Lytro.

Method	Time/s	Method	Time/s	Method	Time/s
SESF [17]	0.462	ECNN [16]	124.055	DBMFIF [79]	0.190
GFF [33]	0.277	DRPL [20]	0.091	FusionDiff [80]	70.823
IMF [23]	1.498	IFCNN [42]	0.041	Ours	0.089
CNN [15]	407.484	U2Fusion [19]	0.053		

evaluating fusion effectiveness in dark conditions, as illustrated in Fig. 14. The results demonstrate that our algorithm maintains high accuracy and adaptability in low-light scenarios, suitable for practical fusion tasks.

4.4. Ablation experiment

4.4.1. Discussion of different parameter values

To ascertain the optimal convolution window size, this research conducted a detailed experimental analysis of four main sliding window sizes: 4 × 4, 8 × 8, 16×16, and 32×32. The experimental results were used to evaluate and compare the performance of different window sizes. For the entire dataset, the objective evaluation metrics results for different window sizes are summarized in Table 6, with the optimal results highlighted in bold. Through this data, this research aims to identify the window size that most enhances the fusion effect in multi-focus image fusion tasks.

4.4.2. Discussion on the use of different modules

This research conducts ablation experiments to ascertain the significance of each component within the network. The model's baseline comprises the encoder-decoder architecture designed for multi-focus image fusion. From the observation in Fig. 15(a), it can be inferred that the removal of the DCA module has the most significant impact on the model performance, which suggests that the DCA module greatly enhances the model's ability to deeply parse the image content by integrating the multi-scale feature extraction advantage of dilated convolution and the focusing property of the attention mechanism, and brings about a significant improvement, which has become a key component in the structure of the proposed network. The DCA module

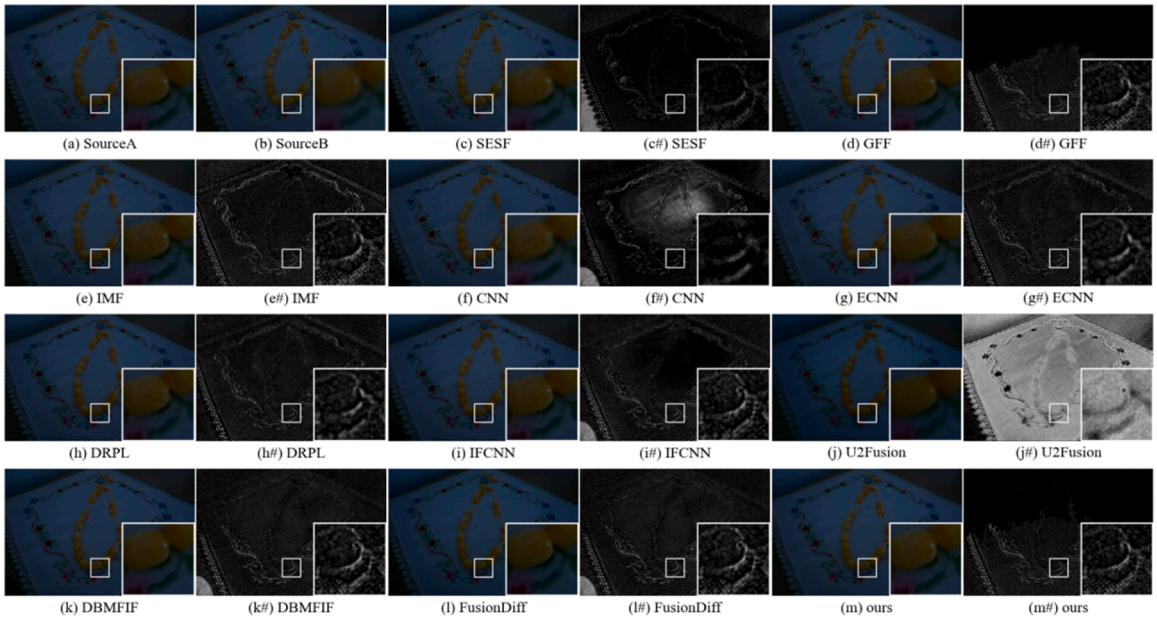


Fig. 14. Low light image fusion results and difference maps.

Table 6
Measurement results for different window sizes.

Window size	Metrics					
	Q_{NMI}	Q_G	Q_{CB}	Q_{TE}	Q_{SF}	Q_{Piella}
4 × 4	1.0029	0.7157	0.7345	0.3563	−0.0421	0.8145
8 × 8	0.1665	0.7312	0.7275	0.3954	−0.0285	0.9067
32×32	0.1734	0.7275	0.7909	0.4048	−0.0231	0.9203
16×16	**1.2294**	**0.7675**	**0.8068**	**0.4089**	**−0.0164**	**0.9342**

not only enhances the expressiveness of the features, but also ensures the completeness and richness of the information during the feature reconstruction process through the residual learning mechanism, which in turn achieves superior performance in visual tasks. From Fig. 15(b), it can be seen that the impact of Siamese structure on the performance is also not negligible. The Siamese network, through its unique structural design, is able to efficiently recognize the differences between clear and blurred regions of an image, and utilize the learned similarity metrics to guide the fusion process to achieve higher quality image fusion, while reducing human intervention in the traditional methods and enhancing the level of automation and efficiency. Fig. 15(c) shows that the introduction of the DFEB module can significantly improve the network performance, and the DFEB module effectively extracts image features and directs attention by extracting rich features and guiding the model to focus on key regions in each channel. Finally, Fig. 15(d) reveals the role of GCB in noise suppression, which enhances the overall effectiveness of the network by increasing the expressiveness and fusion accuracy of the features.

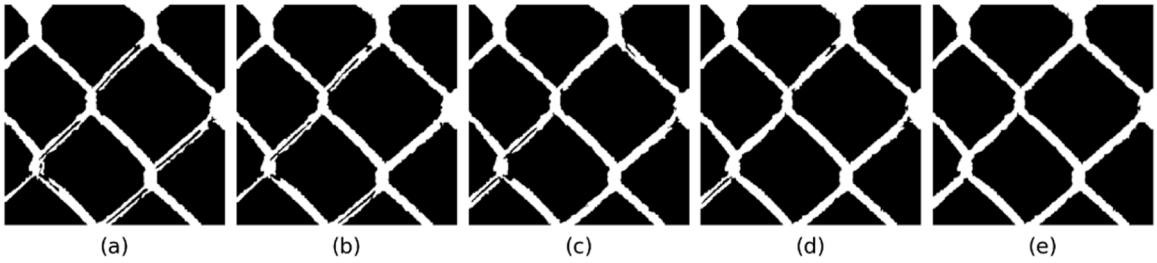


Fig. 15. Comparison of decision map under changes in network structure, (a) represents the decision map obtained by removing the DCA module from the proposed network, (b) represents the decision map obtained by removing the Siamese structure from the proposed network, (c) represents the decision map obtained by removing the DFEB module from the proposed network. (d) represents the decision map obtained by removing the GCB module for the proposed network, and (e) is the decision map obtained for the proposed network.

Table 7
Objective results of the ablation research in the Lytro dataset.

Methods	Metrics					
	Q_{NMI}	Q_G	Q_{CB}	Q_{TE}	Q_{SF}	Q_{Piella}
Baseline	1.18548	0.76227	0.80533	0.40269	−0.01729	0.94496
Baseline + DCA	1.18662	0.76259	0.80533	0.40269	−0.01729	0.94501
Baseline + DCA + Siamese	1.19316	0.72843	0.80716	0.40393	−0.01672	0.94509
Baseline + DCA + residual + DFEB	1.19357	0.72853	0.80777	0.40413	−0.01671	0.94510
Baseline + DCA + residual+ DFEB + GCB	1.22945	0.77510	0.81328	0.40834	−0.01311	0.94502

Table 8
Measurement results of different segmentation thresholds.

Segmentation threshold	Metrics					
	Q_{NMI}	Q_G	Q_{CB}	Q_{TE}	Q_{SF}	Q_{Piella}
0	1.1682	0.7402	0.7681	0.4012	−0.0310	0.8931
0.25	0.1757	0.7461	0.7882	0.4063	−0.0236	0.9164
0.75	0.1748	0.7475	0.7893	0.4069	−0.0231	0.9170
1	0.1692	0.7418	0.7692	0.4021	−0.0313	0.8938
0.5	**1.2294**	**0.7630**	**0.8068**	**0.4089**	**−0.0164**	**0.9342**

The model's performance is assessed by averaging various objective evaluation metrics across all images in the Lytro dataset. The experimental findings are presented in Table 7, demonstrating enhancements in the proposed model concerning the QNMI, QG, QCB, QTE, QSF, and QPiella metrics. These results indicate that the network structure introduced in this research effectively improves the performance of multi-focus image fusion. The experiment employs the same six evaluation metrics for comparative analysis and utilizes the identical Lytro dataset. Table 7 vividly showcases the crucial roles of the Siamese structure, DCA, and other components in our methodology. Upon the removal of the GCB module, QNMI decreased by approximately 3.01 %. Similarly, the removal of the DFEB module led to a decrease of about 4.75 % in QG. Furthermore, eliminating the DCA attention mechanism resulted in a decrease of approximately 4.92 % in QNMI. Despite only a 2.03 % decrease in QPiella upon removing the residual, we opted to integrate the proposed components into our model to achieve superior overall performance.

4.4.3. *Discussion on post-processing methods*

The selection of an appropriate segmentation threshold is a crucial aspect in binary segmentation to achieve accurate segmentation. This research extensively investigates a range of critical thresholds 0, 0.25, 0.5, 0.75, and 1 and assesses their varying impacts on segmentation performance through empirical studies. The primary objective of these experiments is to determine the threshold that yields the most effective segmentation outcomes at the dataset level. The findings reveal that different thresholds significantly influence segmentation precision, with a detailed breakdown of their individual performances presented in Table 8, offering empirical evidence for further analysis and discourse. This research endeavors to propose an optimal global threshold selection for binary image segmentation tasks using this approach.

5. Conclusion

This paper presents a novel deep learning-based approach for multi-focus image fusion. The method incorporates local feature attention modules, a Siamese network architecture, multi-scale residual convolution, and attention mechanisms to improve the quality and robustness of image fusion. The local feature attention module enhances the extraction of image features by capturing local deep features, thereby providing a more precise foundation for the fusion process. The Siamese network architecture, in conjunction with residual modules, facilitates the extraction of complex image features, mitigating the issue of gradient vanishing and improving feature extraction efficiency. The multi-scale dilated convolution attention module boosts boundary recognition by extracting features at multiple scales and employing attention mechanisms, leading to a more comprehensive and accurate extraction of multi-scale global feature information, thereby enhancing the quality of the fused image. While the proposed method exhibits promising performance in multi-focus image fusion, it still has certain limitations. Currently, for the problem of poor edge transitions in the model, we plan to improve the edge discrimination accuracy of the model by adopting the training method of multi-task learning or exploring improved network structure designs. Our goal is to provide a more efficient and high-quality solution for related applications. In future, we will investigate the combination of Granular-ball computing with other network architectures to achieve better generalization and adaptability.

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.

Acknowledgement

The authors are grateful to the anonymous reviewers for their valuable comments and suggestions. This work was supported by the National Natural Science Foundation of China (Grant No 62003065); the Natural Science Foundation of Chongqing (General Program) (Grant No CSTB2024NSCQ-MSX0527); the Science and Technology Research Program of Chongqing Municipal Education Commission (Grant No KJQN202200564); Innovation and Development Joint Fund of Chongqing Natural Science Foundation (Grant No 2023NSCQ-LZX0029); the Fund Project of Chongqing Normal University (Grant No 21XLB032).

Data availability

Data will be made available on request.

References

[1] R. Pei, K. Yao, X. Xu, et al., TransFusion-net for multifocus microscopic biomedical image fusion, Comput. Methods Programs Biomed. 240 (2023) 107688.
[2] X. Zhang, Deep learning-based multi-focus image fusion: a survey and a comparative study, IEEe Trans. Pattern. Anal. Mach. Intell. 44 (9) (2021) 4819–4838.
[3] R. Pei, K. Yao, X. Xu, et al., TransFusion-net for multi focus microscopic biomedical image fusion, Comput. Methods Programs Biomed. 240 (2023) 107688.
[4] Y. Li, Y. Sun, X. Huang, et al., An image fusion method based on sparse representation and sum modified-laplacian in NSCT domain, Entropy 20 (7) (2018) 522.
[5] J. Hu, S. Li, X. Kang, Filter-based image fusion, IEEE J. Image Process. 22 (6) (2013) 2864–2875.
[6] Y. Zhang, X. Bai, Y. Wang, Multi-focus image fusion based on multi-scale morphological focus measurement, Inf. Fusion. 35 (2017) 81–101.
[7] X. Bai, Y. Zhang, F. Zhou, et al., Quadtree-based multi-focus image fusion using a weighted focus-measure, Inf. Fusion. 22 (2015) 105–118.
[8] Y. Liu, Y. Liu, Z. Wang, Multi-focus image fusion based on multi-focus image fusion, Inf. Fusion. 23 (2015) 139–155.
[9] D. Guo, j. Yan, X. Qu, High quality multi-focus image fusion using self-similarity and depth information, Opt. Commun. 338 (2015) 138–144.
[10] C. Cheng, T. Xu, X.-J. Wu, MUFusion: a general unsupervised image fusion network based on memory unit, Inf. Fusion. 92 (2023) 80–92.
[11] Z. Zhou, S. Li, B. Wang, Multi-scale weighted gradient-based fusion for multi-focus images, Inf. Fusion. 20 (2014) 60–72.
[12] G. Liu, W. Yang, Multisensor image fusion based on wavelet transform, //, in: Process Control and Inspection for Industry 4222, SPIE, 2000, pp. 219–223.
[13] O. Kováč, I. Gladišová, Multifocal images fusion, Acta Electrotech. Inform. 17 (3) (2017) 22–26.
[14] Bo Yang, Sheng Li, Feng Sun, Image fusion based on non-downsampled contourlet transform, in: Fourth International Conference on Image Graphics (ICIG 2007), IEEE, 2007, pp. 719–724.
[15] Y. Liu, X. Chen, H. Peng, Z. Wang, Multi-focus image fusion with a deep convolutional neural network, Inf. Fusion. 36 (2017) 191–207.

[16] M. Amin-Naji, A. Aghagolzadeh, M. Ezoji, Ensemble of CNN for multi-focus image fusion, Inf. Fusion. 51 (2019) 201–214.
[17] B. Ma, Y. Zhu, X. Yin, X. Ban, H. Huang, M. Mukeshimana, SESF-fuse: an unsupervised deep model for multi-focus image fusion, Neural Comput. Appl. 33 (11) (2021) 5793–5804.
[18] Y. Zhang, Y. Liu, P. Sun, et al., IFCNN: a general image fusion framework based on convolutional neural network, Inf. Fusion. 54 (2020) 99–118.
[19] Ma Xu, Jun Jiang, Xiang Guo, Hui Ling, A unified unsupervised image fusion network- U2fusion, IEEE J. Pattern Anal. Ma-China Intell. 44 (16) (2020) 502–518.
[20] J. Li, et al., DRPL: deep regression pair learning for multi-focus image fusion, IEEe Trans. Image Process. 29 (2020) 4816–4831.
[21] S. Li, X. Kang, J. Hu, Image fusion with guided filtering, IEEe Trans. Image Process. 22 (7) (2013) 2864–2875.
[22] Y. Liu, S. Liu, Z. Wang, Multi-focus image fusion with dense SIFT, Inf. Fusion. 23 (2015) 139–155.
[23] S. Li, X. Kang, J. Hu, B. Yang, Image matting for fusion of multi-focus images in dynamic scenes, Inf. Fusion. 14 (2) (2013) 147–162.
[24] J. Chen, X. Li, L. Luo, J. Ma, Multi-focus image fusion based on multi-scale gradients and image matting, IEEE Trans Multimed. 24 (2021) 655–667.
[25] Y. Zhang, X. Bai, T. Wang, Boundary finding based multi-focus image fusion through multi-scale morphological focus-measure, Inf. Fusion. 35 (2017) 81–101.
[26] J. Ma, Z. Zhou, B. Wang, et al., Multi-focus image fusion using boosted random walks-based algorithm with two-scale focus maps, Neurocomputing 335 (2019) 9–20.
[27] X. Bai, Y. Zhang, F. Zhou, B. Xue, Quadtree-based multi-focus image fusion using a weighted focus-measure, Inf. Fusion. 22 (2015) 105–118.
[28] J. Ma, Z. Zhou, B. Wang, M. Dong, Multi-focus image fusion based on multi-scale focus measures and generalized random walk, in: In 2017 36th Chinese control conference, 2017, pp. 5464–5468.
[29] J. Wang, H. Qu, Y. Wei, et al., Multi-focus image fusion based on quad-tree decomposition and edge-weighted focus measure, Signal Process. 198 (2022) 108590.
[30] Y. Liu, L. Wang, J. Cheng, et al., Multi-focus image fusion: a survey of the state of the art, Inf. Fusion. 64 (2020) 71–91.
[31] S. Wei, W. Ke, A multi-focus image fusion algorithm with DT-CWT, //, in: 2007 International Conference on Computational Intelligence and Security (CIS 2007), IEEE, 2007, pp. 147–151.
[32] Q. Zhang, B.-l. Guo, Multifocus image fusion using the nonsubsampled contourlet transform, Signal Process 89 (7) (2009) 1334–1346.
[33] K. He, J. Sun, X. Tang, Guided image filtering, IEEe Trans. Pattern. Anal. Mach. Intell. 35 (6) (2012) 1397–1409.
[34] Y. Liu, X. Chen, R.K. Ward, Z. Jane Wang, Image fusion with convolutional sparse representation, IEEE Signal Process 23 (12) (2016) 1882–1886.
[35] Q. Zhang, T. Shi, F. Wang, et al., Robust sparse representation based multi-focus image fusion with dictionary construction and local spatial consistency, Pattern. Recognit. 83 (2018) 299–313.
[36] M. Amin-Naji, A. Aghagolzadeh, Multi-focus image fusion in DCT domain using variance and energy of Laplacian and correlation coefficient for visual sensor networks, J. AI Data Min. 6 (2) (2018) 233–250.
[37] Z. Zhou, S. Li, B. Wang, Multi-scale weighted gradient-based fusion for multi-focus images, Inf. Fusion. 20 (2014) 60–72.
[38] L. Kou, L. Zhang, K. Zhang, et al., A multi-focus image fusion method via region mosaicking on laplacian pyramids, PLoS. One 13 (5) (2018) e0191085.
[39] H. Wang, Multi-focus image fusion algorithm based on focus detection in spatial and NSCT domain, PLoS. One 13 (9) (2018) e0204225.
[40] L. Tang, X. Xiang, H. Zhang, et al., DIVFusion: darkness-free infrared and visible image fusion, Inf. Fusion. 91 (2023) 477–493.
[41] Y. Liu, X. Chen, Z. Wang, et al., Deep learning for pixel-level image fusion: recent advances and future prospects, Inf. Fusion. 42 (2018) 158–173.
[42] Y. Zhang, Y. Liu, P. Sun, et al., IFCNN: a general image fusion framework based on convolutional neural network, Inf. Fusion. 54 (2020) 99–118.
[43] Ma Xu, Jun Jiang, Xiang Guo, Hui Ling, A unified unsupervised image fusion network- U2fusion, IEEE J. Pattern Anal. Ma-China Intell. 44 (16) (2020) 502–518.
[44] H. Jung, Y. Kim, H. Jang, N. Ha, K. Sohn, Unsupervised deep image fusion with structure tensor representations, IEEe Trans. Image Process. 29 (2020) 3845–3858.
[45] C.Li Zhang, C. Shao, C. Xu, C. Ma, Multi-focus image fusion with unsupervised generative adversarial networks based on joint constraints of adaptive and gradient, Inf. Fusion. 26 (2021) 40–53.
[46] O. Bouzos, I. Andreadis, N. Mitianoudis, A convolutional neural network-based conditional random field model for structured multi-focus image fusion robust to noise, IEEe Trans. Image Process. 32 (2023) 2915–2930.
[47] J. Ma, L. Tang, F. Fan, J. Huang, X. Mei, Y. Ma, SwinFusion: cross-domain long-range learning for general image fusion via swin transformer, IEEE/CAA J. Autom. Sin. 9 (7) (2022) 1200–1217.
[48] X. Hu, J. Jiang, X. Liu, J.ZMFF Ma, Zero-shot multi-focus image fusion, Inf. Fusion. 92 (2023) 127–138.
[49] X. Zhang, Deep learning-based multi-focus image fusion: a survey and a comparative study, IEEe Trans. Pattern. Anal. Mach. Intell. 44 (9) (2021) 4819–4838.
[50] Y. Qi, Z. Yang, X. Lu, et al., A multi-channel neural network model for multi-focus image fusion, Expert. Syst. Appl. 247 (2024) 123244.
[51] P. Chen, J. Jiang, L. Li, et al., A defocus and similarity attention-based cascaded network for multi-focus and misaligned image fusion, Inf. Fusion. 103 (2024) 102125.
[52] Yuncan Ouyang, Hao Zhai, Hanyue Hu, et al., FusionGCN: multi-focus image fusion using superpixel features generation GCN and pixel-level feature reconstruction CNN, Expert. Syst. Appl. 262 (2025) 125665.
[53] V. Mnih, N. Heess, A. Graves, et al., Recurrent models of visual attention, Adv. Neural Inf. Process. Syst. (2014) 27.
[54] M. Jaderberg, K. Simonyan, A. Zisserman, Spatial transformer networks, Adv. Neural Inf. Process. Syst. (2015) 28.
[55] X. Zhu, H. Hu, S. Lin, et al., Deformable convnets v2: more deformable, better results[C], //, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 9308–9316.
[56] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks[C], //, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7132–7141.
[57] S. Woo, J. Park, J.Y. Lee, et al., Cbam: convolutional block attention module, //, in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 3–19.
[58] A. Vaswani, N. Shazeer, N. Parmar, et al., Attention is all you need, Adv. Neural Inf. Process. Syst. (2017) 30.
[59] L. Yuan, Y. Chen, T. Wang, et al., Tokens-to-token vit: training vision transformers from scratch on imagenet, //, in: Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 558–567.
[60] S. Chopra, R. Hadsell, Y. LeCun, Learning a similarity metric discriminatively, with application to face verification, //, in: 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05) 1, IEEE, 2005, pp. 539–546.
[61] S. Zagoruyko, N. Komodakis, Learning to compare image patches via convolutional neural networks, //, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 4353–4361.
[62] K. He, X. Zhang, S. Ren, et al., Spatial pyramid pooling in deep convolutional networks for visual recognition, IEEe Trans. Pattern. Anal. Mach. Intell. 37 (9) (2015) 1904–1916.
[63] K. He, X. Zhang, S. Ren, et al., Deep residual learning for image recognition, //, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
[64] K. Zhang, W. Zuo, Y. Chen, et al., Beyond a gaussian denoiser: residual learning of deep cnn for image denoising, IEEe Trans. Image Process. 26 (7) (2017) 3142–3155.
[65] Z. Chen, Z. He, Lu Z-M. DEA-Net, Single image dehazing based on detail-enhanced convolution and content-guided attention, IEEe Trans. Image Process. 33 (2024) 1002–1015.
[66] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, //, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7132–7141.
[67] X. Zhang, X. Zhou, M. Lin, et al., Shufflenet: an extremely efficient convolutional neural network for mobile devices, //, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 6848–6856.
[68] M. Everingham, S.M.A. Eslami, L. Van Gool, et al., The pascal visual object classes challenge: a retrospective, Int. J. Comput. Vis. 111 (2015) 98–136.
[69] X. Guo, R. Nie, J. Cao, et al., FuseGAN: learning to fuse multi-focus image via conditional generative adversarial network, IEEE Trans Multimed. 21 (8) (2019) 1982–1996.
[70] M. Nejati, S. Samavi, S. Shirani, Multi-focus image fusion using dictionary-based sparse representation, Inf. Fusion. 25 (2015) 72–84.
[71] S. Xu, X. Wei, C. Zhang, et al., MFFW: a new dataset for multi-focus image fusion, arxiv prepr. arxiv:2002.04780 (2020).
[72] H. Zhang, Z. Le, Z. Shao, et al., MFF-GAN: an unsupervised generative adversarial network with adaptive and gradient joint constraints for multi-focus image fusion, Inf. Fusion. 66 (2021) 40–53.
[73] G. Qu, D. Zhang, P. Yan, Information measure for performance of image fusion, Electron. Lett. 38 (7) (2002) 313–315.
[74] Z. Liu, E. Blasch, Z. Xue, et al., Objective assessment of multiresolution image fusion algorithms for context enhancement in night vision: a comparative study, IEEe Trans. Pattern. Anal. Mach. Intell. 34 (1) (2011) 94–109.
[75] Y. Chen, R.S Blum, A new automated quality assessment algorithm for image fusion, Image Vis. Comput. 27 (10) (2009) 1421–1432.
[76] N. Cvejic, C.N. Canagarajah, D.R Bull, Image fusion metric based on mutual information and Tsallis entropy, Electron. Lett. 42 (11) (2006) 1.
[77] Y. Zheng, E.A. Eeeock, B.C. Hansen, et al., A new metric based on extended spatial frequency and its application to DWT based fusion algorithms, Inf. Fusion. 8 (2) (2007) 177–192.
[78] G. Piella, H. Heijmans, A new quality metric for image fusion, //, in: Proceedings 2003 international conference on image processing (Cat. No. 03CH37429) 3, IEEE, 2003, pp. 111–173.
[79] J. Zhang, Q. Liao, H. Ma, et al., Exploit the best of both end-to-end and map-based methods for multi-focus image fusion, IEEE Trans Multimed. 26 (2024) 6411–6423.
[80] M. Li, R. Pei, T. Zheng, et al., FusionDiff: multi-focus image fusion using denoising diffusion probabilistic models, Expert. Syst. Appl. 238 (2024) 121664.