The Visual Computer
https://doi.org/10.1007/s00371-024-03637-3

RESEARCH

SwinMFF: toward high-fidelity end-to-end multi-focus image fusion via swin transformer-based network

Xinzhe Xie[1,2] · Buyu Guo[2,3] · Peiliang Li[1,2] · Shuangyan He[1,2] · Sangjun Zhou[1,2]

Accepted: 1 September 2024

Abstract
The end-to-end approach that directly learns the mapping from multi-focus images to fused images has been widely used recently, which achieves excellent performance in dealing with complex scenes. However, the fusion quality of this approach falls short of decision map-based methods, as this approach can preserve the original pixels of the focused regions in the fused image, while end-to-end methods use network inference results with pixel-wise regression errors, resulting in low fidelity of the fused images. To mitigate this limitation, we propose SwinMFF, which effectively captures long-range dependencies across the source images via the swin transformer to reduce pixel-wise regression errors, achieving high-fidelity end-to-end fusion while simultaneously alleviating edge artifacts in the fused image. Extensive experiments demonstrate that SwinMFF outperforms the other 28 state-of-the-art methods in both subjective visual quality and quantitative metrics. The codes are available at https://github.com/Xinzhe99/SwinMFF.

Keywords Deep learning · Multi-focus · Image fusion · End-to-end · Transformer

1 Introduction

The purpose of multi-focus image fusion (MFF) is to integrate multiple images captured with different focus settings into a single all-in-focus image, addressing focus issues arising from the physical constraints of optical lenses. It holds significant importance in fields such as biology, industry, and medicine, with a growing relevance to everyday life. In biology, it aids in acquiring comprehensive microscopic images of cells at various depths of field (DoF) [1]. In the industrial sector, it finds applications in measuring the structure of nonwoven textiles [2] and enhancing the quality of microscopic images of printed circuit boards (PCBs) [3]. In medicine, this technique can be utilized to fuse microscopic images of cervical cells for disease diagnosis [4]. Moreover, it can be used in the field of autonomous driving to improve the accuracy of target detection [5]. Existing MFF methods can be classified into two categories: conventional methods and deep learning-based methods.

Over the past decades, various conventional methods have been developed to fuse multi-focus images, including methods based on the transform domain and spatial domain. Transform domain-based MFF methods process the original image by converting it into another domain for processing, such as the frequency domain. Methods such as pyramid-based methods [6, 7], wavelet-based methods [8, 9], non-subsampled contourlet transform (NSCT) methods [10, 11], and sparse representation (SR) methods [12, 13] are adept at preserving detailed information in the original image and exhibiting superior performance in complex scenes. However, these methods are prone to errors in pixel intensity [14] and are susceptible to distortion because of their sensitivity to high-frequency components [15]. Spa-

✉ Buyu Guo
guobuyuwork@163.com

✉ Shuangyan He
hesy103@163.com

Xinzhe Xie
xiexinzhe@zju.edu.cn

Peiliang Li
lipeiliang@zju.edu.cn

Sangjun Zhou
sjune163@163.com

1 Ocean College, Zhejiang University, Zhoushan 316021, Zhejiang, People's Republic of China

2 Hainan Institute, Zhejiang University, Sanya 572025, Hainan, People's Republic of China

3 Donghai Laboratory, Zhoushan 316021, Zhejiang, People's Republic of China

Published online: 05 October 2024

tial domain-based MFF methods typically operate on image pixels to generate fused images. This category encompasses block-based methods [16, 17], region-based methods [18, 19], and pixel-based methods [20, 21]. These methods achieve satisfactory performance but rely on handcrafted features such as gradients, structural information, and texture intensity. However, these handcrafted features are not extensive and informative, particularly in homogeneous and flat regions, which constrains the fusion performance [22]. Moreover, these methods are prone to produce artifacts on the object boundaries [23].

To alleviate some of the shortcomings in conventional methods, researchers have recently proposed several new approaches to address these issues. MCDFD [24] introduced a novel MFF method based on multi-scale cross-difference and focus detection, which can effectively address edge blur and detail loss and improve the clarity and visual quality of fused images. Additionally, a method based on multi-dictionary linear sparse representation and region fusion mode was proposed, which can significantly reduce artifacts at image edges and preserve more image details [25]. Furthermore, SAMF [5] proposed an innovative small-region-aware MFF method to preserve information regarding details and texture. Moreover, [26] proposed to assess fusion quality with learning features based on convolutional sparse representation, rather than with those handcrafted low-level patterns.

Deep learning-based MFF methods, particularly those using convolutional neural networks (CNNs), have achieved outstanding performance [14, 27]. Nonetheless, CNNs have limited receptive fields, which makes learning global interactions challenging. To solve this problem, a hierarchical Swin Transformer was introduced [28], achieving state-of-the-art results on multiple vision tasks and demonstrating the capabilities of Transformers for computer vision. In image fusion, a unified fusion framework for merging images using a hybrid architecture of swin transformer and CNNs was designed [29], aiming to leverage the long-range dependency modeling capability of Transformers and the local feature extraction ability of CNNs for improved fusion performance. Recently, generative adversarial networks (GANs) have also entered this field [30], employing their generative capability to produce more realistic fused images, addressing common issues such as artifacts and blurriness. Furthermore, diffusion models have been explored for MFF task [31], harnessing their ability to learn the true data distribution and generate high-quality samples, aiming to enhance the visual quality and detail preservation of the fused images.

Deep learning-based MFF methods commonly encounter a choice: whether to let the network generate a decision map. Researchers initially used the network to generate binary decision maps to compute fused images [27]. The prevalence of this two-stage approach in research stems from its ability to reduce the learning difficulty of networks and preserve inherent image details by guaranteeing that fused pixels are directly sourced from the original images. However, some researchers indicate that methods based on decision maps may not be necessary, as they are prone to produce edge artifacts [32] and introduce additional problems such as decision map post-processing [33]. While many decision map-based MFF methods utilize morphological operations or conditional random fields (CRFs) to reduce obvious errors of the decision map output by the network, we believe that this approach still has inherent limitations. Firstly, it introduces subjectivity as it relies solely on human perception of decision maps. Secondly, it assumes that the in-focus areas of the input images are continuous, which is not always the case. As some researchers have found, the quality of the fused image can actually decrease after applying post-processing techniques [23]. Consequently, the necessity of post-processing emerges as an additional consideration in decision map-based MFF methods.

To avoid these problems, researchers have explored end-to-end methods for MFF [34, 35]. End-to-end fusion methods directly map the input images to the fused image, eliminating the process complexity and computational burden associated with generating and optimizing decision maps, leading to improved efficiency. Furthermore, end-to-end methods demonstrate superiority in handling complex images and reducing image edge artifacts compared to decision map-based approaches. Moreover, the end-to-end network architecture demonstrates greater generalizability, as it circumvents the need for decision map computations and can be seamlessly adapted to similar vision tasks, such as multi-modal image fusion and medical image fusion [29, 36–38].

However, end-to-end MFF networks face two significant challenges. First, they lack explicit prior knowledge, requiring the network to implicitly learn focus and defocus information from the training data, which places substantial demands on its expressive capacity. Second, since the pixel values of the fused image in end-to-end fusion networks are inferred by the network rather than directly taken from the source images, the fidelity of the fused image is usually lower than that of decision map-based methods. As a result, the fused images produced by end-to-end methods often perform worse than those from decision map-based methods on evaluation indicators, as shown in Table 3. To address these two challenges, we propose SwinMFF, leveraging the long-range dependency modeling capability of the swin transformer. The major contributions of this paper are summarized as follows:

(1) This paper presents a new end-to-end multi-focus image fusion network named SwinMFF, which utilizes the swin transformer solely to complete the entire fusion pipeline.

(2) By exploiting the swin transformer's powerful modeling capabilities, SwinMFF attains superior image fidelity compared to other end-to-end fusion networks and effectively mitigates the common issue of edge artifacts, a challenge in prior fusion methods.

(3) Extensive experimental results demonstrate that our method achieves state-of-the-art performance, both qualitatively and quantitatively, setting a new benchmark for end-to-end MFF techniques.

2 Related works

This section introduces two types of deep learning methods for MFF: decision map-based methods and end-to-end methods. We further discuss the application of swin transformer in computer vision and its promising potential for MFF task.

2.1 Decision map-based multi-focus image fusion network

Traditional MFF methods are primarily based on image transform domains or spatial domains, and their fusion performance is limited by manually designed feature extraction and fusion rules. To overcome this limitation, [27] began using CNNs to fuse multi-focus images, treating MFF as a classification task. They trained the CNN by using artificially synthesized binary masks as supervised images. Subsequent works by [14, 39–41] introduced more advanced network designs and post-processing technology to generate better decision maps and improve the quality of fused images. To avoid the need for large-scale annotated datasets for network training, SESF [42] combined traditional image spatial domain-based methods with CNNs, using focus operators to compute the spatial frequency of high-dimensional feature maps and obtain decision maps. Furthermore, to further understand the potential data distribution differences between focused and defocused regions in the input images, MFIF-GAN [43] utilized generative adversarial networks to generate decision maps. Additionally, ZMFF [15] proposed a novel MFF framework that leverages deep image priors to achieve zero-shot learning, presenting a new unsupervised approach for MFF. Recently, the potential of Transformer applications in the field of MFF has been further demonstrated by [22], who combined Transformers with CNNs and proposed a new MFF network. Furthermore, a novel framework for MFF, called incremental network prior adaptation (INPA), has been introduced by [44], in which a side network is incrementally fine-tuned to learn the deep prior in conjunction with the supervised pre-trained model.

Decision map-based methods have significantly succeeded in MFF because they directly utilize the network's output as the decision map to indicate the focal information in the source images, thus possessing clear interpretability. However, as mentioned in the introduction, this fusion process suffers from issues such as complexity in workflow, artifacts at the edges of the fused image, and lack of generality.

2.2 End-to-end multi-focus image fusion network

End-to-end networks have a significant advantage over traditional non-end-to-end networks as they can learn directly from raw input data to produce the desired output. This simplifies the system, reduces error propagation, and promises better performance. Although widely used in various visual tasks, such as medical image segmentation [45] and image depth estimation [46], end-to-end networks have been relatively underexplored in multi-focus image fusion (MFF) tasks compared to decision map-based methods.

Compared with decision map-based methods, end-to-end methods eliminate the need for decision map generation and optimization procedures. Therefore, they are mainly used in general image fusion frameworks [29, 37, 47] to simultaneously achieve the fusion of multi-focus images, multi-modal images, and medical images. Different from decision map-based methods, end-to-end methods cannot explicitly obtain targeted guidance by utilizing the prior knowledge in the binary labels provided by the training process. Instead, the network needs to capture the differences between defocus and focus regions on its own, which requires a higher expression ability of the network and limits the fusion performance of end-to-end methods. To better distinguish the distribution differences between focused and defocused regions in images, generative adversarial networks [35] and diffusion models [31] have also been introduced into end-to-end MFF, achieving good fusion quality. However, due to the pixel-wise regression errors of the fused image generated by the end-to-end networks, it does not have an advantage in image fidelity. The fusion of multi-focus images demands robust feature extraction and fusion capabilities to preserve fine details and maintain consistent visual quality across the entire image. Existing fusion networks often fail to strike an effective balance between preserving local context and integrating global contextual information, leading to artifacts or loss of crucial visual cues in the final fused result. Addressing these challenges requires the development of more powerful feature modeling and fusion strategies that can adaptively leverage the complementary information present in the inputs, thereby enabling high-fidelity multi-focus image fusion. This paper will attempt to build a more powerful MFF network, exploring the potential of achieving comparable results to decision map-based MFF in terms of image fidelity, human visual perception, and evaluation metrics using the end-to-end manner.

2.3 Application of swin transformer in the field of computer vision

CNNs have limited receptive fields, making it difficult to learn global interactions. In contrast, the self-attention mechanism in Transformers can effectively capture long-range dependencies in images, allowing for a more comprehensive understanding of global context and relationships. Motivated by the remarkable success of Transformers in natural language processing (NLP) [48], computer vision researchers have explored introducing Transformer models into the field. A significant advancement in computer vision is the Vision Transformer (ViT) introduced in [49], which achieved comparable results to CNNs across various vision benchmarks. Besides, [50] proposed an efficient attention pyramid transformer for image processing, this also highlights the promise of Transformers within computer vision. Moreover, [28] proposed a hierarchical swin transformer as a visual backbone network, achieving state-of-the-art results on multiple vision tasks. This further demonstrates the capabilities of Transformers for computer vision modeling. For instance, [51] utilized swin transformers for image restoration tasks, while in the related field of multi-exposure image fusion (MEF), [52, 53] also achieved notable advancements. Moreover, [29] designed a unified fusion framework for image fusion based on a hybrid architecture of swin transformer and CNNs. We believe that the primary factor hindering the fidelity of fused images generated by end-to-end approaches lies in their demanding network requirements. Specifically, these methods necessitate a network with a large receptive field, dynamic weights, strong expressive ability, and long-range modeling capability. Therefore, in this paper, we solely utilize the swin transformer as our backbone without employing any convolutional layers, aiming to fully exploit the advantages of self-attention. Meanwhile, we strive to ensure that the overall network architecture remains simple, intuitive, and efficient.

3 Methods

In this section, we present the architecture of proposed MFF network, the fusion process, and the loss function used to train the network.

3.1 Overview of the proposed network

In SwinFusion [29], a hybrid model combining CNN and Transformer is employed for MFF, with Transformer utilized for feature fusion and deep feature reconstruction, while conventional convolutional layers are retained for other steps. Unlike SwinFusion's hybrid design that combines CNNs and Transformers, our SwinMFF exclusively uses pure Transformer-based modules for the entire feature extraction, fusion, and reconstruction pipeline. By entirely omitting convolutional layers, SwinMFF aims to more effectively capture the long-range dependencies and global contextual relationships inherent in multi-focus image inputs. This design choice allows our model to maintain consistent feature representations throughout the end-to-end fusion process, potentially enhancing its ability to generate high-fidelity fused outputs. In contrast, the alternating use of convolutional layers and Transformers in SwinFusion [29] may be detrimental to feature consistency within the network, potentially leading to larger regression errors.

The design of the SwinMFF network follows three principles: modularity, intuitiveness, and simplicity. The overall architecture of the SwinMFF is displayed in Fig. 1, comprising three modules: feature extraction, feature fusion, and image reconstruction. In the feature extraction module, the input image is reshaped into non-overlapping patches of size 4×4, so that the shape of its features is $H/4 \times W/4 \times 16C$ (C is the number of image channels). Next, a linear embedding layer is applied to project the feature dimensions to a higher-dimensional (projected to D dimensions in SwinMFF, $D = 96$) feature space, which can increase the representational capacity of the features. Then, feed the patches into the swin transformer blocks for the shallow feature extraction. The patch merging layer is used to downsample the features, gradually reducing the resolution of the feature maps while retaining feature information. The feature dimension is then changed by patch merging layer from $H/4 \times W/4 \times D$ to $H/8 \times W/8 \times 2D$. Subsequently, it aims to extract the deeper semantic feature, the merged features are sent to the swin transformer blocks. After repeatedly executing the patch merging process and passing it through the swin transformer blocks again, the features of the two images are concatenated in the channel dimension, and the shape of concatenated features is $H/16 \times W/16 \times 8D$. Afterward, the concatenated features are fed into the feature fusion module to adaptively fuse the features. In the image reconstruction module, patch-expanding layers are utilized to gradually restore the shape of the features back to the original size, we first use three patch-expanding layers and two swin transformer blocks to change the feature dimension to $H \times W \times 2D$ and map the features with dimension $H \times \mathrm{W} \times 2D$ back to $\mathrm{H} \times \mathrm{W} \times C$ by a multi-layer perceptron (MLP) to complete the image reconstruction and the entire fusion process. Notably, the SwinMFF incorporates well-established deep learning techniques such as residual connections, upsampling, downsampling, and multi-scale feature interaction. By avoiding introducing overly intricate components, the network maintains a simple and effective structure, facilitating efficient training. Overall, the modular design, intuitive structure, and simplistic nature of the SwinMFF network are key distinguishing factors that set it apart from more intricate fusion approaches. This sim-

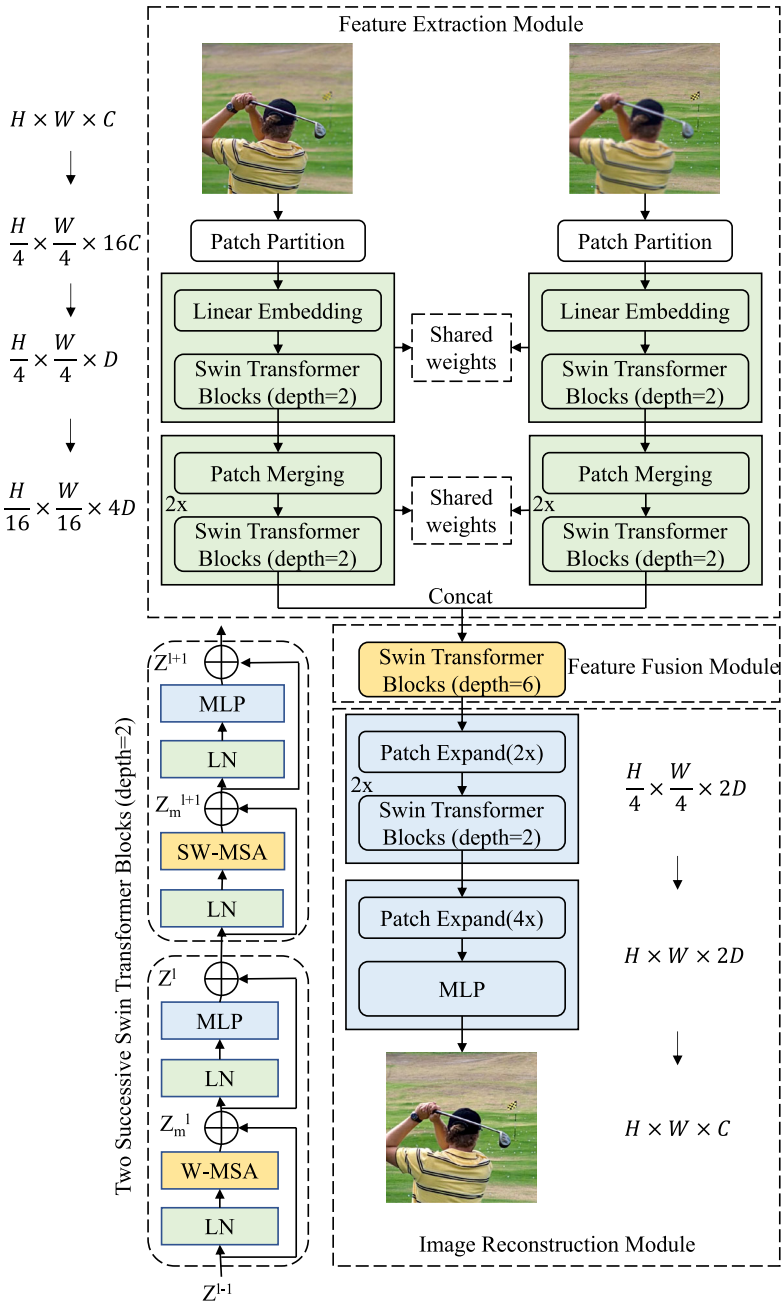

Fig. 1 The architecture of SwinMFF

ple yet powerful architecture enables the network to excel at high-fidelity multi-focus image fusion, while also ensuring robust performance and ease of implementation.

3.2 Swin transformer blocks

The swin transformer blocks [28] differ from the conventional multi-head self-attention (MSA) module by introducing window-based multi-head self-attention (W-MSA) and shifted window multi-head self-attention (SW-MSA). SW-MSA is a further improvement of MSA, which pans the window cyclically over the input sequence so that each window can interact with all other windows, further enhancing global information aggregation and having more substantial expressive power. Two layer-norm layers, a multi-head attention layer and an MLP layer compose each swin transformer block. Figure 1 displays two consecutive swin transformer blocks joined in series, with W-MSA and SW-MSA, respectively. Residual connections are constantly used to alleviate the problem of gradient disappearance, allowing deep transformer networks to be trained. Utilizing this window partitioning approach, a series of continuous swin

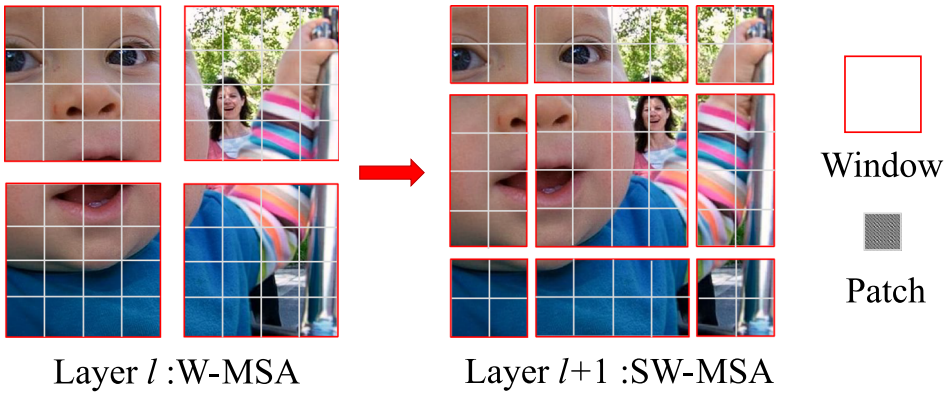

Fig. 2 Shifted window mechanism for computing attention in the swin transformer blocks

transformer blocks can be constructed as follows:

$$Z_m^l = W - \text{MSA}\left(\text{LN}\left(Z^l - 1\right)\right) + Z^{l-1} \tag{1}$$

$$Z^l = \text{MLP}\left(\text{LN}\left(Z_m^l\right)\right) + Z_m^l \tag{2}$$

$$Z_m^{l+1} = \text{SW} - \text{MSA}\left(LN\left(Z^l - 1\right)\right) + Z^l \tag{3}$$

$$Z^{l+1} = \text{MLP}\left(\text{LN}\left(Z_m{}^{l+1}\right)\right) + Z_m^{l+1} \tag{4}$$

Z_m^l denotes the outputs of the (S)W-MSA module, Z^l denotes the MLP module of the lth block, and the LN denotes layer norm layer. Self-attention is calculated as follows:

$$\text{Attention}(Q, K, V) = \text{SoftMax}\left(\frac{QK^T}{\sqrt{d_k}} + B\right)V \tag{5}$$

Q, K, and $V \in R^{M^2 \times d}$ are three matrices to measure similarity; M^2 represents the number of patches in a window. d is the query/key dimension; B is the bias matrix. Figure 2 illustrates a demonstration of the shifted window mechanism in computing attention for the swin transformer blocks.

3.3 Fusion scheme

3.3.1 Feature extraction

For multi-channel images, the input images are first converted from the standard RGB color space to the $YCbCr$ color space. The Y channel, being independent of color information, exclusively represents the brightness of each pixel. This brightness is determined by structural details like variations in light and darkness, texture, and edges within the image. As the Y channel contains the majority of visual information, it effectively reflects the focus information present in the multi-focus image inputs. Leveraging this property, our SwinMFF method performs end-to-end regression solely on the Y channel, rather than attempting to jointly optimize the reconstruction of all color channels [34]. This design choice allows the model to focus its capacity on accurately predicting the brightness details that are critical for high-fidelity fusion, potentially reducing the overall regression error and enhancing the visual quality of the final fused output. By concentrating the learning process on the most informative Y channel, SwinMFF can more effectively capture the nuanced focus variations across the input images, leading to a fusion result that preserves the structural integrity of the scene.

The procedure of the hierarchical image feature extraction is shown in Fig. 1, a pair of source images is fed into the feature extraction module. Take one of them as an example, the image with a shape of $H \times W \times C$ will be split into non-overlapping patches by a patch-splitting module. Each patch is treated as a "token" and its feature is set as a concatenation of the raw pixel values. We use the same 4×4 patch size as in the original swin transformer paper and thus the feature dimension of each patch is $4 \times 4 \times C = 16\text{C}$. Then, a linear embedding layer maps the low-dimensional input to a new feature space with D channels. Following, stacked swin transformer blocks are used to extract features. Next, the extracted features are fed into the patch merging layer, which aims to enhance fusion performance by collecting feature representations at different scales. Also, this pooling-like operation could aggregate richer contextual information. After the patch merging operation is completed, swin transformer blocks are used to further extract higher-level features. The same operation of patch merging and swin transformer blocks is performed again to further extract information at different scales, the output feature size would be mapped into $H/16 \times W/16 \times 4D$. In the end, concatenate every patch in the channel dimension as the output of the feature extraction module.

3.3.2 Feature fusion

For image fusion tasks, the performance of the fusion module plays a critical role in determining the quality of the fused image. Most earlier works relied on simple fusion strategies, such as element-wise addition, as seen in [37,

54, 55]. These simple, element-wise fusion strategies lack adaptability, limiting the network's fusion capabilities. To address this limitation, [34] proposed a novel fusion strategy based on a carefully designed attention mechanism, significantly improving performance. However, this approach increases the network architecture's complexity, compromising its overall conciseness. We believe that the local-global modeling capabilities of the Swin Transformer and its self-attention mechanism might enable a simpler fusion approach to be sufficient. This is because various complex designs, including those involving attention mechanisms, aim not only to make the fusion strategy learnable but also to allow the model to autonomously focus on crucial information and enhance long-range information interaction. Adhering to the idea of designing a simple and efficient network, the SwinMFF's fusion module leverages the inherent local-global modeling capabilities of the swin transformer architecture to achieve effective feature fusion through a much simpler approach. By simply concatenating the features from the extraction stage and passing them through a series of swin transformer blocks, SwinMFF is able to adaptively fuse the information from the two input images. Remarkably, our experiments demonstrate that this streamlined fusion strategy can achieve performance on par with, or even surpassing, the results obtained by more intricate fusion mechanisms. This design choice not only enhances the overall conciseness of the network but also highlights the versatility of swin transformers in solving challenging end-to-end image fusion tasks.

3.3.3 Image reconstruction

The hierarchical feature extraction process employed in the feature extraction module of SwinMFF inevitably leads to a significant reduction in spatial resolution. To recover the original image dimensions, the network utilizes a series of patch expansion layers with varying up-sampling factors. This approach is more effective than traditional linear interpolation or transposed convolution techniques, as patch expansion can better preserve the integrity of fine image details. Importantly, swin transformer blocks are strategically inserted between these up-sampling layers. This design choice serves two crucial purposes: first, it facilitates smooth feature transitions, ensuring a stable flow of information as the spatial resolution is gradually restored; second, it further enhances the reconstruction fidelity by enabling the Transformer's self-attention mechanism to adaptively model long-range dependencies within the features. Finally, rather than producing a decision map as the output, SwinMFF employs a multi-layer perceptron (MLP) to map the high-dimensional features back to the desired fused image. This direct regression to the target image, as opposed to an intermediate representation, allows the network to optimize the fusion process end-to-end and generate a high-quality fused output that effectively captures the original scene details. The careful integration of patch expansion, swin transformer blocks, and direct regression within the reconstruction module represents a thoughtful design choice that leverages the strengths of these individual components. The image reconstruction module in SwinMFF is also a key contributor to the impressive performance and visual fidelity achieved by the SwinMFF network.

3.4 Loss function

The loss function L applied in SwinMFF has the following definition:

$$L = \lambda L_{\text{MSE}} + (1 - \lambda) L_{SSIM} \tag{6}$$

The L_{MSE} is defined as follows:

$$L_{\text{MSE}} = (O - I)^2 \tag{7}$$

L_{MSE} is a common loss function for image reconstruction, which is used to measure the pixel-level difference between the output image O and the ground truth I and to output an accurate and clear image.

The L_{SSIM} is defined as follows:

$$\text{SSIM}(O, I) = \frac{(2\mu_O\mu_I + C_1)(2\sigma_{OI} + C_2)}{(\mu_O^2 + \mu_I^2 + C_1)(\mu_O^2 + \mu_I^2 + C_2)} \tag{8}$$

$$L_{\text{SSIM}} = 1 - \text{SSIM}(O, I) \tag{9}$$

L_{SSIM} measures the difference between two images based on structural similarity, accounting for brightness, contrast, and structure. In the formula of L_{SSIM}, μ is the mean value, σ is the standard deviation, C_1 and C_2 are constants. This loss function is consistent with human vision and provides more comprehensive and accurate information about image quality.

The parameter λ was set to 0.2 in SwinMFF. The combination of L_{MSE} and L_{SSIM} can drive the network to output a clearer fused image with preserved details.

4 Experiments

This section presents a comprehensive evaluation of the proposed SwinMFF, employing both qualitative and quantitative metrics across diverse datasets. The experiments were all performed on a hardware platform with two NVIDIA A6000 GPUs and an Intel(R) Xeon(R) Platinum 8375C CPU clocked at 2.90 GHz.

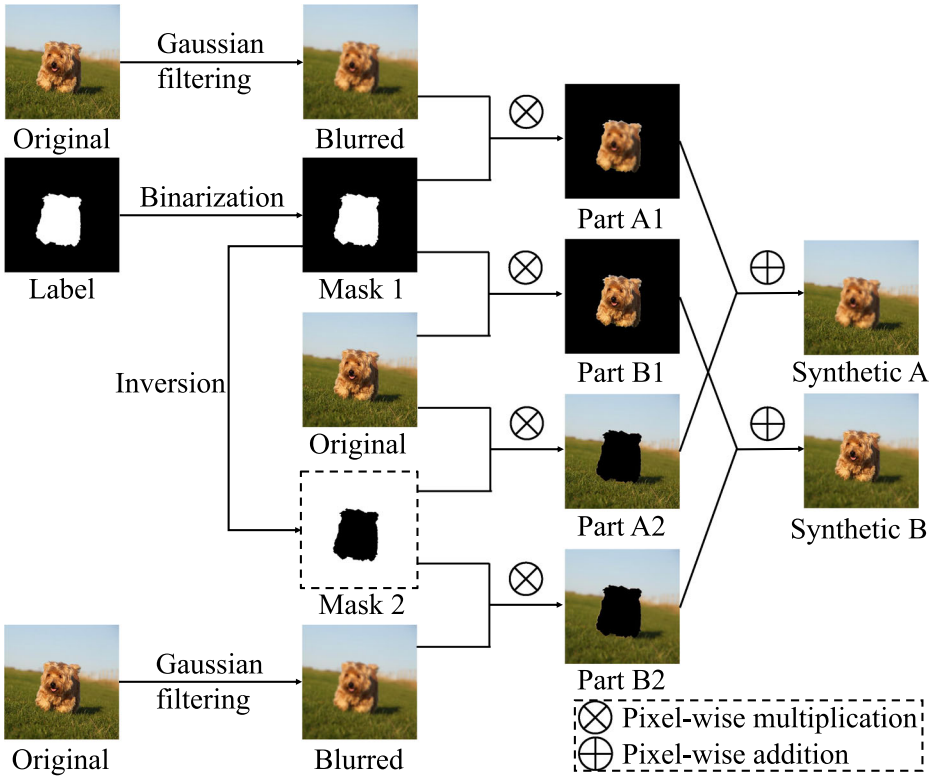

Fig. 3 Workflow diagram for synthetic dataset generation

4.1 Training details

It is widely acknowledged that the training of transformer networks necessitates an extensive number of datasets, and the dataset's quality significantly impacts the networks' inference ability. Our dataset originates from the DUTS dataset [56]. The flowchart to generate the dataset is shown in Fig. 3. We expand the dataset by applying different blur levels. A total of 52,765 training images and 25,095 validation images with the size of 256 × 256 are generated to train SwinMFF.

To optimize the network weights, the AdamW optimizer was employed with a learning rate of 5e−4. A CosineAnnealingLR scheduler was utilized to gradually decrease the learning rate to 0. For data augmentation, random flipping was implemented. Prior to training initiation, the network weights were initialized using a normal distribution. The proposed network was implemented in PyTorch, and training was conducted for 10 epochs on two Nvidia A6000 GPUs, taking approximately 4 h to complete.

4.2 Methods for comparison

To demonstrate the capability of the proposed SwinMFF, we compared it with 28 other SOTA MFF algorithms. Among these are 9 transform domain methods, including DWT [9], DTCWT [57], NSCT [10], GFF [58], SR [59], ASR [60], MWGF [61], ICA [62], NSCT-SR [63], 7 spatial domain methods including SSSDI [64], QUADTREE [65], DSIFT [63], SRCF [66], GFDF [67], BRW [68], MISF [69]. For deep learning-based methods, we compared 6 decision map-based methods, including CNN [27], ECNN [40], SESF [42], MSFIN [14], MFIF-GAN [43], [15], and 6 end-to-end methods including IFCNN-MAX [37], U2Fusion [38], SDNet [47], MFF-GAN [35], SwinFusion [29], FusionDiff [31]. Compared with [70], the end-to-end methods we compared in this paper combine some latest models proposed after 2020, such as Transformers [29] and diffusion model [31]. This represents an advance on the techniques discussed in previous work [70].

In the comparison, for non-deep learning methods, the fusion results of each method follow the default parameter settings, and most fusion results are from the results provided in [71]. For deep learning-based methods, we use the network parameters provided by the original authors. For methods such as FusionDiff [31], in which pre-trained network parameters are not provided, we follow the training code and strategies described in the original paper to retrain the network and obtain the network parameters. Detailed information and download links for these deep learning-based methods are given in Table 1.

4.3 Datasets for assessment

In order to qualitatively and quantitatively evaluate the SwinMFF proposed in this paper from multiple perspectives, we use 3 public datasets in the MFF field, they are Lytro

Table 1 The deep learning-based MFF algorithms used for comparison and the corresponding download links

Method	Journal/conference	Network	Download link
Decision map-based deep learning methods			
CNN (2017)	Information Fusion	CNN	https://github.com/yuliu316316/CNN-Fusion
ECNN (2019)	Information Fusion	CNN	https://github.com/mostafaaminnaji/ECNN
SESF (2020)	Neural. Comput. Appl	CNN	https://github.com/Keep-Passion/SESF-Fuse
MFIF-GAN (2021)	SPIC	GAN	https://github.com/ycwang-libra/MFIF-GAN
MSFIN (2021)	IEEE TIM	CNN	https://github.com/yuliu316316/MSFIN-Fusion
ZMFF (2023)	Information Fusion	DIP	https://github.com/junjun-jiang/ZMFF
End-to-end deep learning methods			
IFCNN-MAX (2020)	Information Fusion	CNN	https://github.com/uzeful/IFCNN
U2Fusion (2020)	IEEE TPAMI	CNN	https://github.com/hanna-xu/U2Fusion
SDNet (2021)	IJCV	CNN	https://github.com/HaoZhang1018/SDNet
MFF-GAN (2021)	Information Fusion	GAN	https://github.com/HaoZhang1018/MFF-GAN
SwinFusion (2022)	IEEE/CAA JAS	CNN & Transformer	https://github.com/Linfeng-Tang/SwinFusion
FusionDiff (2024)	ESWA	Diffusion Model	https://github.com/lmn-ning/ImageFusion
SwinMFF (2024)	The Visual Computer	Transformer	https://github.com/Xinzhe99/SwinMFF

[66], MFFW [72], MFI-WHU [35], respectively. The Lytro dataset contains 20 pairs of multi-focus images and 4 image sequences captured using a Lytro light-field camera. The MFFW dataset includes 13 pairs of multi-focus images. Compared to the Lytro dataset, images in MFFW exhibit a significant defocus spread effect, allowing us to evaluate the fusion performance of different methods under severe defocus conditions. Additionally, we conduct evaluations on the MFI-WHU dataset, which consists of 120 pairs of images synthesized by applying Gaussian blurring and contains more diverse image pairs, such as images of high-contrast scenes.

In the following sections, we will conduct a comprehensive evaluation of the fusion results for all the methods mentioned in the previous section using these datasets.

4.4 Evaluation metrics

Evaluating the quality of fused images is challenging due to the absence of ground truth data in multi-focus datasets. To evaluate the fusion results from multiple perspectives, we selected EI, $Q^{ab/f}$, STD, SF, and AVG from the image feature-based metrics, MI and EN from the information theory-based metrics, and VIF from the human perception inspired metrics. A total of eight metrics were chosen to comprehensively compare the fused images.

EI represents the edge intensity of an image, where a higher value of EI indicates clearer edges in the image. $Q^{ab/f}$ is a gradient-based metric, a larger $Q^{ab/f}$ indicates better clarity performance in the fused image. The STD represents the standard deviation of the image, indicating the extent of variation in the gray levels of the image. Therefore, it also can reflect the clarity of the image. SF reflects the rate of change of pixels in the horizontal and vertical directions, and it can measure the overall clarity level of the image. AVG is the average gradient of the image, which can reflect the details and edge information of the image. MI quantifies the information shared between the source images and the fused image. A high MI value indicates that the fused image effectively preserves the fidelity, or the original information content, from the source images. EN measures the information contained in the fused image, a higher EN value generally indicates a richer and more informative image. VIF evaluates the fidelity of the fused image aligning with human visual perception; For all indicators, the fusion result improves as the metrics increase in value.

4.5 Results on the Lytro dataset

In Fig. 4, we visualize the results of 24 SOTA MFF methods on the Lytro dataset, including our SwinMFF. The four rows of images in Fig. 4, from top to bottom, represent transform domain-based methods, spatial domain-based methods, decision map-based deep learning methods, and end-to-end deep learning methods, respectively. We can observe that methods such as QUADTREE, DSIFT, GFDF, CNN, ECNN, and SESF have evidently lost the information near the focused region boundaries of the source images. In contrast, all transform domain-based methods and end-to-end deep learning methods do not exhibit such obvious errors. Among the transform domain-based methods, the fusion results of DWT, DTCWT, GFF, SR, and ASR show noticeable noise in the iron mesh, while NSCT performs relatively better. All

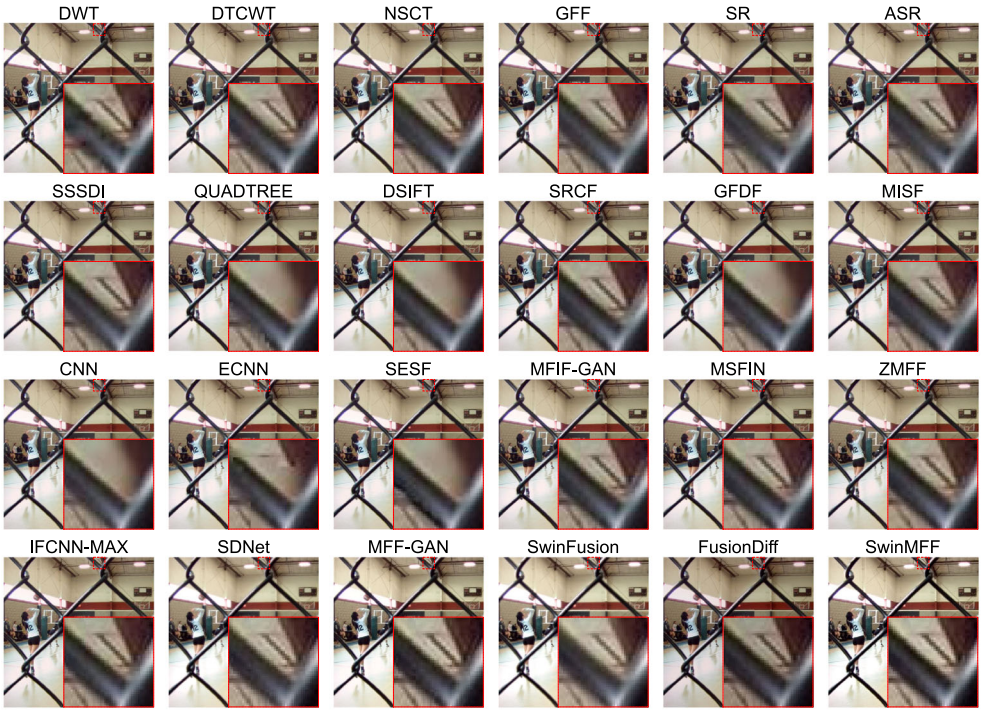

Fig. 4 The fusion results of various SOTA methods on the "gym" example from the Lytro dataset [66]

end-to-end deep learning methods demonstrate satisfactory performance, without noticeable fusion errors or artifacts, showcasing the advantages of end-to-end approaches in complex scenarios. It is important to note that the fused image of FusionDiff exhibits some color differences compared to other methods, and SwinFusion appears slightly defocused in the iron mesh. In comparison, our SwinMFF better preserves the information in the focused regions of the source images, and maintains excellent performance in complex areas.

Figure 5 specifically compares the fusion results of different methods at the transition zone between the defocus and focus regions. As can be observed from the Fig. 5, MISF and SwinFusion exhibit obvious artifacts, while only SwinMFF, ZMFF, MSFIN, MFIF-GAN, and ECNN achieve results with almost no visible artifacts. It is worth noting that SwinMFF is the only end-to-end deep learning method among them that does not generate any obvious artifacts.

To better illustrate the differences between different fusion algorithms, we use difference maps for comparison. Difference images can be obtained by subtracting the fused image from each of the two source images. If the fusion quality is near-perfect, the difference between the defocused regions of the source images and the fused image will be significantly higher than the difference between the focused regions and the fused image. Additionally, the focused regions should ideally approach zero difference. Therefore, from a visual perspective, if the two difference images exhibit a complementary trend and the difference is more pronounced, it indicates a higher fusion quality and greater fidelity to the source images.

Figure 6 compares the difference maps and fusion results of various SOTA decision map-based MFF methods implemented using deep learning with proposed SwinMFF. As the pixel values of the fused image output by the decision map-based MFF methods originate from the source images, the difference between their difference maps is more pronounced than in Fig. 7. Remarkably, despite SwinMFF being an end-to-end fusion network, where the pixel values of the fused image do not directly come from the source images, its performance on difference maps is comparable to that of decision map-based MFF methods. This further demonstrates the ability of SwinMFF to preserve image fidelity and retain image details.

To evaluate the performance differences among various end-to-end deep learning methods, we present their respective fusion results and two difference maps in Fig. 7. For SwinFusion, MFF-GAN, and SDNet, many regions in their respective difference maps exhibit similar content, indicating a significant difference between the pixel values of the fused image and the focused regions of the source images. These observations suggest the limitations of these methods in accurately recovering pixel values from the focused regions of the source images. Additionally, the significant deviations observed in the difference maps of U2Fusion and FusionDiff compared to other methods indicate potential difficulties in preserving brightness information during the fusion process

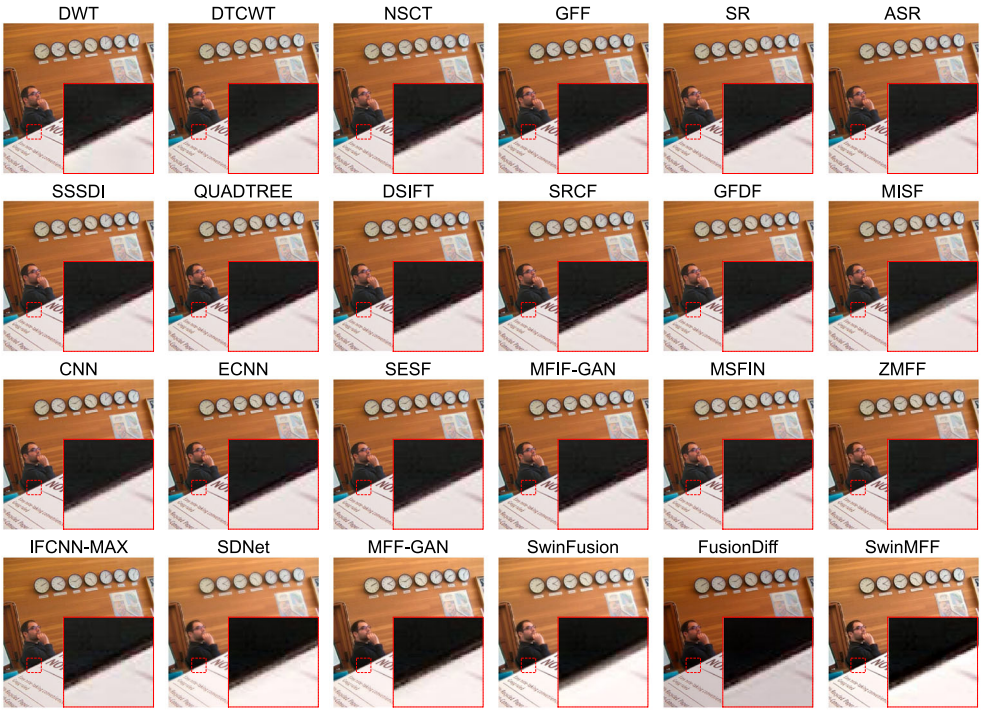

Fig. 5 The fusion results of various SOTA methods on the "notebook" example from the Lytro [66] dataset

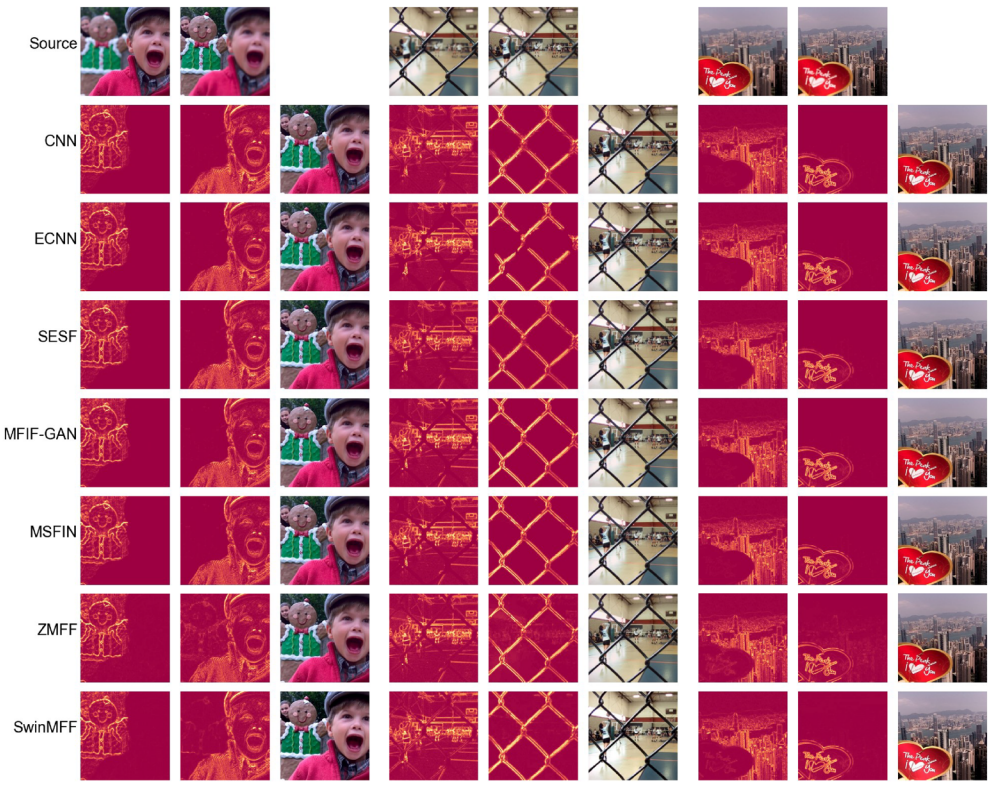

Fig. 6 The difference maps of various SOTA decision map-based MFF methods implemented using deep learning

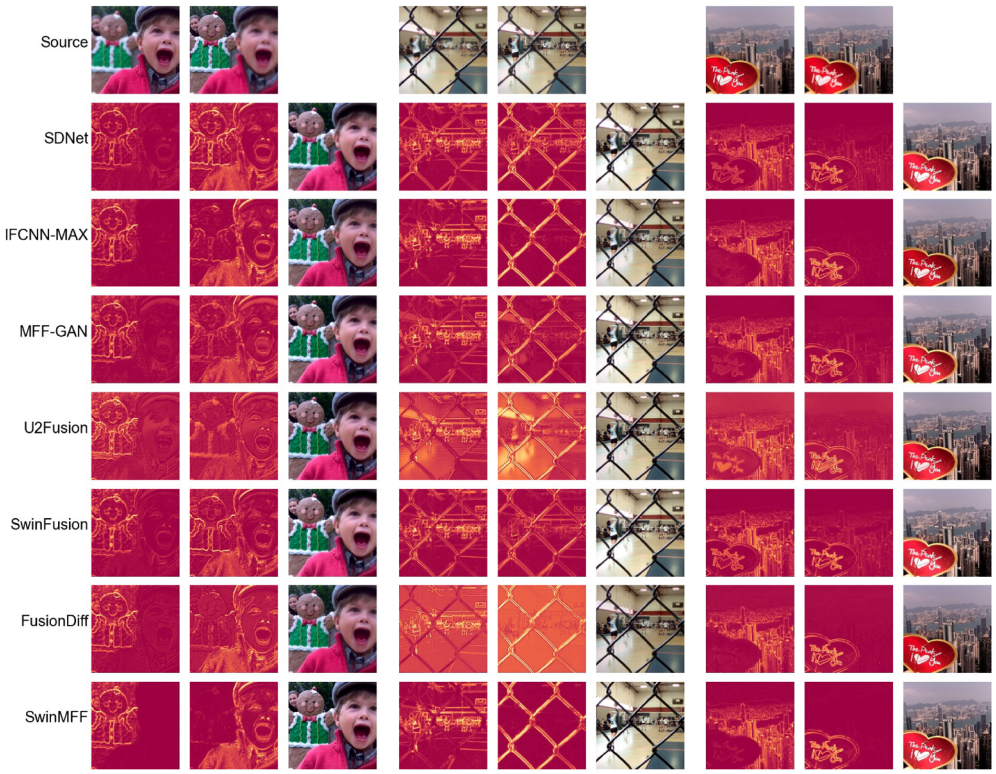

Fig. 7 The difference maps of different SOTA end-to-end MFF methods implemented using deep learning

for these networks. In comparison, IFCNN-MAX performs better, but it still contains content also present in the other difference map. For example, in the second difference picture, both the gingerbread man and the little boy can be clearly seen. However, in a single difference map of SwinMFF, only the foreground or background content is visible, indicating that SwinMFF, as an end-to-end MFF method, excels in preserving source image fidelity and retaining image details, despite the pixel values generated by the network rather than directly originating from the source images (Fig. 8).

Tables 2 and 3 present a comprehensive quantitative comparison of the proposed SwinMFF method against 28 state-of-the-art MFF algorithms on the Lytro dataset. The results are categorized into four groups: transform domain-based methods, spatial domain-based methods, decision map-based deep learning methods and end-to-end deep learning methods.

In Table 2, SwinMFF demonstrates superior performance across all metrics when compared to both transform domain-based and spatial domain-based traditional methods. Notably, SwinMFF outperforms all spatial domain-based methods listed in the table. Furthermore, when compared to transform domain-based methods, SwinMFF consistently achieves either the first or second-best scores across all evaluation metrics.

Table 3 compares SwinMFF with various deep learning-based methods. Among decision map-based deep learning methods, SwinMFF again outperforms all other approaches across all metrics. The performance gap is particularly notable in metrics such as EI (72.4041 vs. 71.0914 for the second-best MSFIN) and VIF (1.1810 vs. 1.1420 for MSFIN), indicating SwinMFF's superior ability in edge preservation and overall visual quality.

In comparison with end-to-end deep learning methods, SwinMFF's superiority is even more pronounced. It significantly outperforms other methods across all metrics, with substantial margins in most cases. For instance, the EI score of SwinMFF (72.4041) is notably higher than the second-best IFCNN-MAX (70.9193), and its VIF score (1.1810) far exceeds that of IFCNN-MAX (1.1322).

In summary, the quantitative results presented in Tables 2 and 3 provide strong evidence for the effectiveness of SwinMFF in the MFF task. Its consistent outperformance across various evaluation metrics and different categories of competing methods underscores its capability to produce high-quality and high-fidelity fused images.

4.6 Results on the MFFW dataset

Figure 9 presents the fusion results of various end-to-end methods on three image pairs ("sculpture", "coffee cup", and

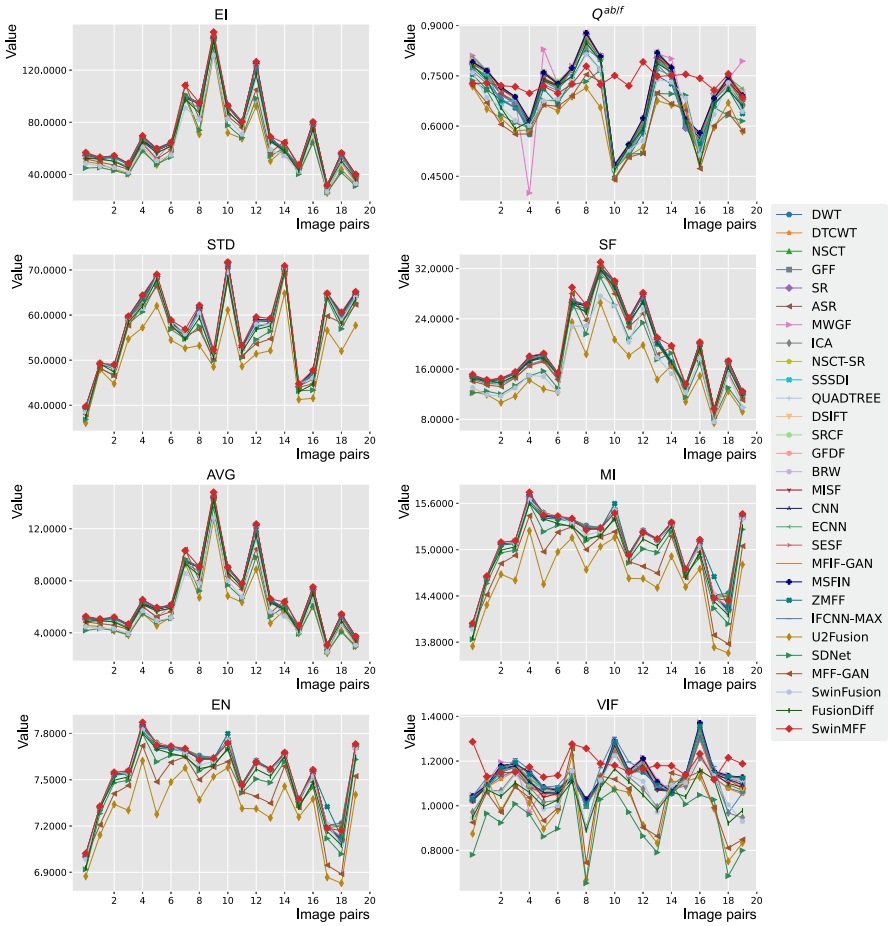

Fig. 8 Objective performance of different fusion methods on the Lytro [66] dataset

Table 2 Quantitative comparison with traditional methods on Lytro [66] dataset

Method	*EI*↑	$Q^{ab/f}$↑	*STD*↑	*SF*↑	*AVG*↑	*MI*↑	*EN*↑	*VIF*↑
Comparison with transform domain-based SOTA MFF methods								
DWT	70.7942	0.6850	57.2776	19.3342	6.8336	**15.0872**	**7.5436**	1.1114
DTCWT	70.5666	0.6929	57.2315	19.3204	6.8134	15.0791	7.5396	1.1079
NSCT	70.4289	0.6901	57.3601	19.2662	6.8027	15.0816	7.5408	1.1249
GFF	70.5179	0.6998	57.4451	19.2947	6.8058	15.0716	7.5358	1.1277
SR	70.2498	0.6944	57.3795	19.2819	6.7818	15.0650	7.5325	1.1208
ASR	70.3342	0.6951	57.3616	19.2818	6.7897	15.0654	7.5327	1.1201
MWGF	69.8052	0.7037	57.4136	19.1900	6.7273	15.0669	7.5334	1.1343
ICA	68.3180	0.6766	56.9383	18.5968	6.6125	15.0655	7.5327	1.0708
NSCT-SR	70.6705	0.6995	57.3924	19.3355	6.8213	15.0676	7.5338	1.1251
Proposed	**72.4041**	**0.7321**	**57.9737**	**19.7954**	**6.9734**	15.0826	7.5413	**1.1810**
Comparison with spatial domain-based SOTA MFF methods								
SSSDI	70.7102	0.6966	57.4770	19.3567	6.8234	15.0668	7.5334	1.1309
QUADTREE	70.8957	0.7027	57.5334	19.4163	6.8412	15.0684	7.5342	1.1368
DSIFT	70.9808	0.7046	57.5319	19.4194	6.8493	15.0688	7.5344	1.1381
SRCF	71.0810	0.7036	57.5394	19.4460	6.8607	15.0690	7.5345	1.1374
GFDF	70.6258	0.7049	57.4973	19.3312	6.8145	15.0674	7.5337	1.1336
BRW	70.6777	0.7040	57.5020	19.3433	6.8200	15.0675	7.5337	1.1336
MISF	70.4148	0.6984	57.4437	19.2203	6.7945	15.0671	7.5335	1.1222
Proposed	**72.4041**	**0.7321**	**57.9737**	**19.7954**	**6.9734**	**15.0826**	**7.5413**	**1.1810**

Bold indicates the best, underlined indicates the second best

Table 3 Quantitative comparison with deep learning-based methods on Lytro [66] dataset

Method	$EI\uparrow$	$Q^{ab/f}\uparrow$	$STD\uparrow$	$SF\uparrow$	$AVG\uparrow$	$MI\uparrow$	$EN\uparrow$	$VIF\uparrow$
Comparison with decision map-based deep learning methods								
CNN	70.3238	0.7019	57.4354	19.2295	6.7860	15.0663	7.5331	1.1255
ECNN	70.7432	0.7030	57.5089	19.3837	6.8261	15.0675	7.5338	1.1337
SESF	70.9403	0.7031	57.5495	19.4158	6.8448	15.0696	7.5348	1.1395
MFIF-GAN	71.0395	0.7029	57.5430	19.4370	6.8560	15.0690	7.5345	1.1393
MSFIN	71.0914	0.7045	57.5642	19.4438	6.8602	15.0695	7.5348	1.1420
ZMFF	70.8298	0.6635	57.0347	18.9707	6.8045	15.0735	7.5368	1.1331
Proposed	**72.4041**	**0.7321**	**57.9737**	**19.7954**	**6.9734**	**15.0826**	**7.5413**	**1.1810**
Comparison with end-to-end deep learning methods								
IFCNN-MAX	70.9193	0.6784	57.4896	19.3793	6.8463	15.0722	7.5361	1.1322
U2Fusion	59.8957	0.6190	51.9356	14.9334	5.6515	14.6153	7.3077	0.9882
SDNet	60.3437	0.6441	55.2655	16.9252	5.8725	14.9332	7.4666	0.9281
MFF-GAN	66.0601	0.6222	55.1920	18.4022	6.4089	14.8153	7.4076	1.0084
SwinFusion	62.8130	0.6597	56.8142	16.6430	5.9862	15.0476	7.5238	1.0685
FusionDiff	67.4911	0.6744	56.1372	18.8483	6.5325	14.9817	7.4909	1.0448
Proposed	**72.4041**	**0.7321**	**57.9737**	**19.7954**	**6.9734**	**15.0826**	**7.5413**	**1.1810**

Bold indicates the best, underlined indicates the second best

"flowerpot") from the MFFW [72] dataset, along with difference maps between the fused images and the source image. Notably, the MFFW [72] dataset exhibits a more pronounced defocus spread effect compared to the Lytro [66] dataset.

It is evident that SDNet and MFF-GAN yield relatively inferior results in the magnified regions highlighted by boxes. From a visual perspective, IFCNN-MAX, SwinFusion, U2Fusion, and the proposed method SwinMFF demonstrate comparatively superior performance with fewer artifacts. In the "sculpture" example, FusionDiff's result exhibits noticeable color discrepancies compared to the source images, clearly visible in the difference map. IFCNN-MAX's difference maps reveal that the method performs well only in the cane area, while the consistency in background regions indicates its ineffectiveness in distinguishing foreground from background. Conversely, SwinFusion, U2Fusion, and the proposed SwinMFF demonstrate good performance in addressing this issue. In the "coffee cup" example, the fused images from IFCNN-MAX, SwinFusion, and U2Fusion display prominent white artifacts in the boxed region containing digits and letters. However, our method and FusionDiff preserve these details with good clarity, avoiding such artifacts. In the "flowerpot" example, SwinMFF stands out as the only method that produces almost no artifacts at the edges of high-contrast regions. Interestingly, SwinFusion, a model with a hybrid CNN-Transformer architecture, shows the second-best results in this example. This may be attributed to the long-range modeling capabilities of both SwinMFF and SwinFusion models, which enable them to better handle such transition areas.

Table 4 presents a comprehensive quantitative comparison of various multi-focus image fusion methods on the MFFW dataset. The results are divided into two categories: decision map-based methods and end-to-end deep learning approaches.

Among decision map-based methods, the proposed SwinMFF demonstrates superior performance, ranking first in six out of eight metrics (EI, STD, AVG, MI, EN, and VIF). Notably, SwinMFF achieves the highest EI score, significantly outperforming the second-best ZMFF. This indicates SwinMFF's superior ability to preserve edge information and structural details. Similarly, the significantly higher VIF score compared to the second place further confirms SwinMFF's ability to generate images that align with human visual perception.

In comparison with end-to-end deep learning methods, SwinMFF exhibits competitive performance across most metrics. While MFF-GAN achieves the highest scores in several metrics (EI, SF, AVG, VIF), SwinMFF consistently ranks among the top performers. It is worth noting that SwinMFF outperforms SwinFusion in most metrics, despite both utilizing swin transformer architectures. This highlights the effectiveness of the proposed modifications and fusion strategy.

The consistently high performance of SwinMFF across various metrics underscores its robustness on this real-world dataset. This robustness mainly stems from the fact that the proposed SwinMFF avoids using explicit decision maps during the training process, but full-clear images consistent with human visual perception. Moreover, the long-range modeling

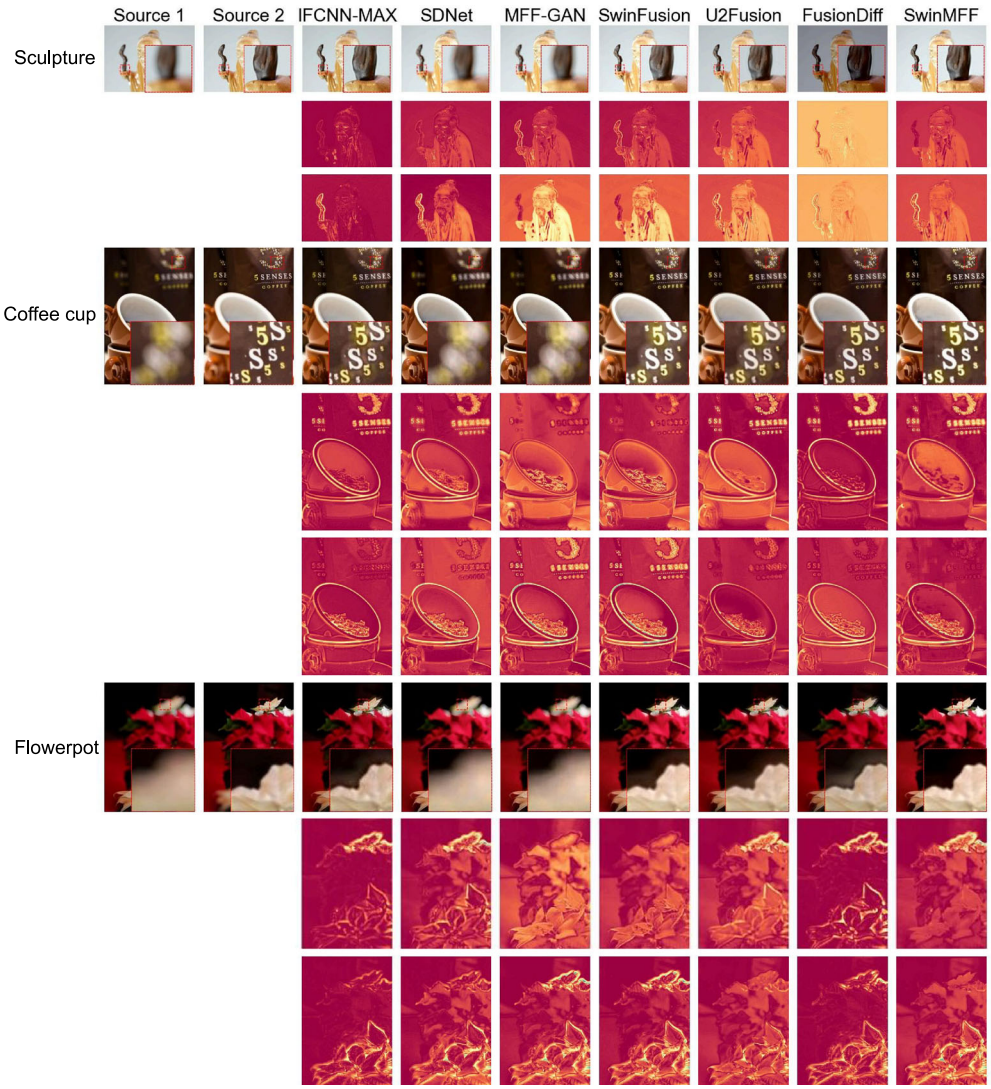

Fig. 9 The qualitative results of ours and other SOTAs in MFFW [72] dataset

capability of the swin transformer enables better adaptation to the continuous transitions between in-focus and out-of-focus regions in real-world scenes.

4.7 Results on the MFI-WHU dataset

Since MFI-WHU [35] is a dataset synthesized using Gaussian blur, the fusion difficulty is relatively lower compared to Lytro [66] and MFFW [72]. Therefore, we selected three challenging image pairs to compare the fusion performance differences of various methods. The results are shown in Fig. 10. Additionally, in Fig. 10, we used three common metrics PSNR, SSIM, and MSE to quantitatively describe the differences between the fused images and the ground truth images.

For the "Waffle" House example, we observe that SwinMFF performs exceptionally well in handling high-contrast scenes. It ranks first in PSNR, SSIM, and MSE, indicating that the fused image is very close to the ground truth image. The "Tennis" and "Baseball" examples allow us to evaluate the fusion capabilities of different methods in complex scenarios. In the "Tennis" example, the tennis racket in our method's fused image is almost identical to that in the first source image, whereas U2Fusion's result shows noticeable distortion, and FusionDiff exhibits significant color bias. For the "Baseball" example, although the visual differences among the methods are subtle, the quantitative results indicate that our method still performs the best.

Table 5 presents a comprehensive quantitative comparison of various multi-focus image fusion methods on the MFI-WHU [35] dataset. The results are also categorized into two groups: decision map-based methods and end-to-end deep learning methods.

Table 4 Quantitative results of different fusion methods on the MFFW [72] dataset

Method	$EI\uparrow$	$Q^{ab/f}\uparrow$	$STD\uparrow$	$SF\uparrow$	$AVG\uparrow$	$MI\uparrow$	$EN\uparrow$	$VIF\uparrow$
Comparison with decision map-based methods								
CNN	73.9624	0.6216	55.2288	22.1315	7.4426	14.3497	7.1749	1.0305
ECNN	75.7644	**0.7312**	54.3536	22.7149	7.6464	14.3807	7.1904	1.0541
SESF	76.9227	0.6247	54.6179	**23.1558**	7.7519	14.3841	7.1920	1.0542
MFIF-GAN	76.5417	0.7283	54.7778	22.9481	7.7225	14.3140	7.1570	1.0639
MSFIN	75.7969	0.6183	55.4647	22.8168	7.6375	14.3528	7.1764	1.0616
ZMFF	77.7055	0.6541	54.1892	21.4789	7.6592	14.3329	7.1665	1.0636
Proposed	**80.4903**	0.6636	**56.3653**	22.7120	**7.9646**	**14.3843**	**7.1921**	**1.1577**
Comparison with end-to-end deep learning methods								
IFCNN-MAX	76.3056	0.6022	55.8810	22.1333	7.6334	14.3420	7.1710	1.0344
U2Fusion	65.7906	0.5917	47.9726	18.1867	6.6806	14.2630	7.1315	1.1349
SDNet	78.6557	0.4551	60.1968	27.6084	8.0076	14.1937	7.0968	1.1543
MFF-GAN	**83.0560**	0.4372	60.5192	**28.2025**	**8.4157**	14.3462	7.1731	**1.2342**
SwinFusion	75.3649	0.6423	**60.5835**	20.5358	7.3528	14.2838	7.1419	1.1912
FusionDiff	69.6123	**0.6673**	53.0291	21.2969	7.0366	14.2275	7.1138	0.9052
Proposed	80.4903	0.6636	56.3653	22.7120	7.9646	**14.3843**	**7.1921**	1.1577

Bold indicates the best, underlined indicates the second best

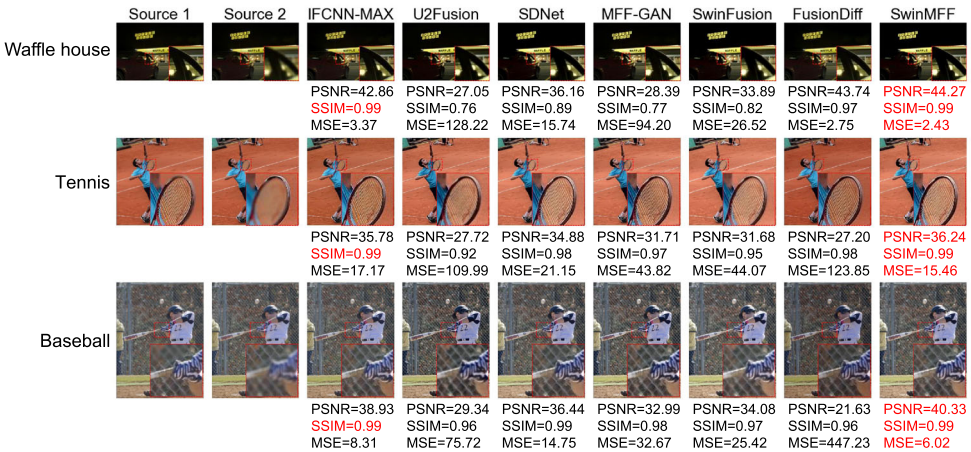

Fig. 10 Part of results of ours and other SOTAs in MFI-WHU [35] dataset

Compared with decision map-based methods, the proposed SwinMFF demonstrates competitive performance across all metrics. SwinMFF achieves the highest scores in MI and EN, with values of 14.6548 and 7.3274 respectively. This indicates superior performance in preserving information from source images. It ranks second in EI with a score of 78.9436, slightly behind ZMFF (82.0595), suggesting excellent edge preservation capabilities. SwinMFF also ranks second in STD and VIF, demonstrating its ability to maintain image contrast and align with human visual perception.

Among end-to-end deep learning methods, SwinMFF shows remarkable performance. It ranks first in $Q^{ab/f}$ (0.7008), and it consistently ranks second across all other metrics, closely following IFCNN-MAX. The performance gap is notably small in most metrics, such as EI (78.9436 vs. 79.3862) and STD (52.6293 vs. 52.6472). Our method significantly outperforms other end-to-end approaches like U2Fusion, SDNet, MFF-GAN, SwinFusion, and FusionDiff across all metrics. The superior performance in information-theoretic metrics (MI and EN) and visual quality metrics (VIF) underscores SwinMFF's ability to produce high-fidelity fused images.

It's worth noting that SwinMFF maintains consistently high performance across both categories, demonstrating its robustness. While ZMFF shows strong performance in some metrics among decision map-based methods, and IFCNN-MAX leads in the end-to-end category, SwinMFF is the only method that consistently performs at the top level across both categories.

In summary, the quantitative results in Table 5 also provide strong evidence for the effectiveness and robustness of SwinMFF in the MFF task. Its consistent top-tier performance

Table 5 Quantitative results of different fusion methods on the MFI-WHU [35] dataset

Method	$EI\uparrow$	$Q^{ab/f}\uparrow$	$STD\uparrow$	$SF\uparrow$	$AVG\uparrow$	$MI\uparrow$	$EN\uparrow$	$VIF\uparrow$
Comparison with decision map-based methods								
CNN	77.0123	0.7276	52.2883	26.4975	8.1720	14.6345	7.3173	1.0959
ECNN	77.9532	**0.7314**	52.3634	26.7520	8.2718	14.6411	7.3205	1.1038
SESF	77.6439	0.7267	52.4376	26.7527	8.2356	14.6404	7.3202	1.1078
MFIF-GAN	78.5272	0.7302	52.5995	**26.9048**	8.3274	14.6494	7.3247	1.1169
MSFIN	77.6764	0.7273	52.4399	26.8228	8.2380	14.6345	7.3173	1.1118
ZMFF	**82.0595**	0.6193	**58.0647**	24.9329	**8.3925**	14.5580	7.2790	**1.1742**
Proposed	78.9436	0.7008	52.6293	26.3398	8.3138	**14.6548**	**7.3274**	1.1379
Comparison with end-to-end deep learning methods								
IFCNN-MAX	**79.3862**	0.6936	**52.6472**	**26.6642**	**8.3756**	**14.6662**	**7.3331**	**1.1713**
U2Fusion	68.8453	0.5917	47.9726	18.1867	6.6806	14.2630	7.1315	1.1349
SDNet	70.8002	0.6889	51.0994	24.2105	7.5039	14.5238	7.2619	1.0236
MFF-GAN	76.7037	0.6496	51.4179	25.2436	8.0277	14.4673	7.2336	1.1368
SwinFusion	68.6117	0.6777	51.6143	20.6637	6.9656	14.5788	7.2894	1.1060
FusionDiff	72.3067	0.6762	50.9468	23.6592	7.5304	14.5516	7.2758	1.0399
Proposed	78.9436	**0.7008**	52.6293	26.3398	8.3138	14.6548	7.3274	1.1379

Bold indicates the best, underlined indicates the second best

across various evaluation metrics and different categories of competing methods highlights its capability to produce high-quality fused images.

4.8 Generalization comparison

Table 6 presents a comprehensive ranking analysis of various end-to-end deep learning methods across three different multi-focus image datasets. This analysis aims to evaluate the generalization capabilities of these networks in fusing diverse multi-focus images.

Key observations from the ranking analysis are as follows: for the Lytro [66] dataset, SwinMFF achieves a perfect score, ranking first across all eight metrics. For the MFFW [72] dataset, MFF-GAN performs exceptionally well, ranking first overall with top performances in EI, SF, AVG, and VIF. SwinMFF closely follows, ranking second overall and achieving top ranks in MI and EN. For the MFI-WHU [35] dataset, IFCNN-MAX demonstrates superior performance, ranking first in seven out of eight metrics. SwinMFF consistently ranks second across most metrics, showing robust performance.

SwinMFF shows remarkable consistency, ranking 1st, 2nd, and 2nd across the three datasets. This demonstrates excellent generalization capability and robustness across different image characteristics. IFCNN-MAX also shows strong generalization, ranking 2nd, 4th, and 1st. Its performance notably improves on the MFI-WHU dataset. MFF-GAN shows variable performance, ranking 5th, 1st, and 3rd, indicating dataset-specific strengths.

In conclusion, this ranking analysis provides strong evidence for the superior generalization capabilities of SwinMFF. Its consistent top-tier performance across diverse datasets demonstrates robustness to varying image characteristics and fusion scenarios. While other methods show strengths on specific datasets, SwinMFF maintains high performance across all evaluated datasets, making it a reliable choice for diverse multi-focus image fusion tasks. This analysis not only highlights the effectiveness of SwinMFF but also provides valuable insights into the relative strengths and weaknesses of various end-to-end deep learning methods in the context of multi-focus image fusion.

In Fig. 11, we conducted the Nemenyi post hoc test on these end-to-end MFF methods to identify significant differences between algorithms and variations in performance across different datasets. Each dot represents the average ranking of the methods, while the horizontal lines denote the range determined by the critical difference (CD) value. If the lines of the two algorithms do not overlap, it suggests a statistically significant difference between them. Our analysis of Fig. 11 reveals that SwinFusion exhibits relatively consistent performance across multiple datasets, whereas MFF-GAN demonstrates greater performance variability. MFF-GAN performs well on the more challenging MFFW dataset but performs poorly on the easier Lytro and MFI-WHU datasets. IFCNN-MAX and SwinMFF achieve significant advantages in quantitative performance compared to other methods, which aligns with our qualitative observations. Notably, SwinMFF exhibits smaller performance variations across different datasets compared to IFCNN-

Table 6 Ranking of different end-to-end deep learning methods on various datasets

Method	*EI*	$Q^{ab/f}$	*STD*	*SF*	*AVG*	*MI*	*EN*	*VIF*	*Average*	*Ranking*
Ranking in Lytro [66] *dataset*										
IFCNN-MAX	2	2	2	2	2	2	2	2	2.00	2
U2Fusion	7	7	7	7	7	7	7	6	6.88	7
SDNet	6	5	5	5	6	5	5	7	5.50	6
MFF-GAN	4	6	6	4	4	6	6	5	5.13	5
SwinFusion	5	4	3	6	5	3	3	3	4.00	4
FusionDiff	3	3	4	3	3	4	4	4	3.50	3
Proposed	1	1	1	1	1	1	1	1	1.00	1
Ranking in MFFW [72] *dataset*										
IFCNN-MAX	4	4	5	4	4	3	3	6	4.13	4
U2Fusion	7	5	7	7	7	5	5	5	6.00	7
SDNet	3	6	3	2	2	7	7	4	4.25	5
MFF-GAN	1	7	2	1	1	2	2	1	2.13	1
SwinFusion	5	3	1	6	5	4	4	2	3.75	3
FusionDiff	6	1	6	5	6	6	6	7	5.38	6
Proposed	2	2	4	3	3	1	1	3	2.38	2
Ranking in MFI-WHU [35] *dataset*										
IFCNN-MAX	1	2	1	1	1	1	1	1	1.13	1
U2Fusion	6	7	7	7	7	7	7	4	6.50	7
SDNet	5	3	5	4	5	5	5	7	4.88	6
MFF-GAN	3	6	4	3	3	6	6	3	4.25	3
SwinFusion	7	4	3	6	6	3	3	5	4.63	4
FusionDiff	4	5	6	5	4	4	4	6	4.75	5
Proposed	2	1	2	2	2	2	2	2	1.88	2

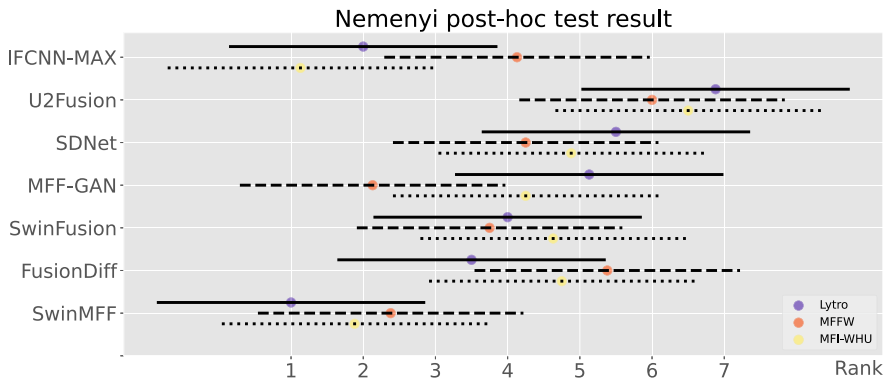

Fig. 11 Results of the Nemenyi post hoc test

MAX. We emphasize that our SwinMFF not only achieves superior quantitative and qualitative performance but also demonstrates strong generalization capabilities and consistent excellence across different datasets.

4.9 Model efficiency comparison

In Table 7, we compare the size, FLOPs (Floating Point Operations), and inference time of the proposed SwinMFF with other end-to-end methods. To evaluate the processing speed of our method, we calculated the average running time per image using the MFI-WHU dataset [35]. Although our SwinMFF, as a transformer-based model, requires the computation of self-attention and is relatively larger in size, it still maintains competitiveness in terms of FLOPs and processing time. Compared to the most similar method, SwinFusion [29], we achieve better fusion quality while utilizing only one-third of the computational load. As a result, the inference time is reduced to approximately one-fourth of that of SwinFusion.

Table 7 The comparison of total parameters, FLOPs and inference time for various end-to-end MFF methods

Method	IFCNN	U2Fusion	SDNet	MFF-GAN	SwinFusion	FusionDiff	SwinMFF
Size (M)	0.08	0.66	0.07	0.05	0.93	26.90	41.25
FLOPs (G)	8.54	86.4	8.81	3.08	63.73	58.13	22.38
Inference time (s)	0.09	0.16	0.10	0.06	1.79	81.47	0.46

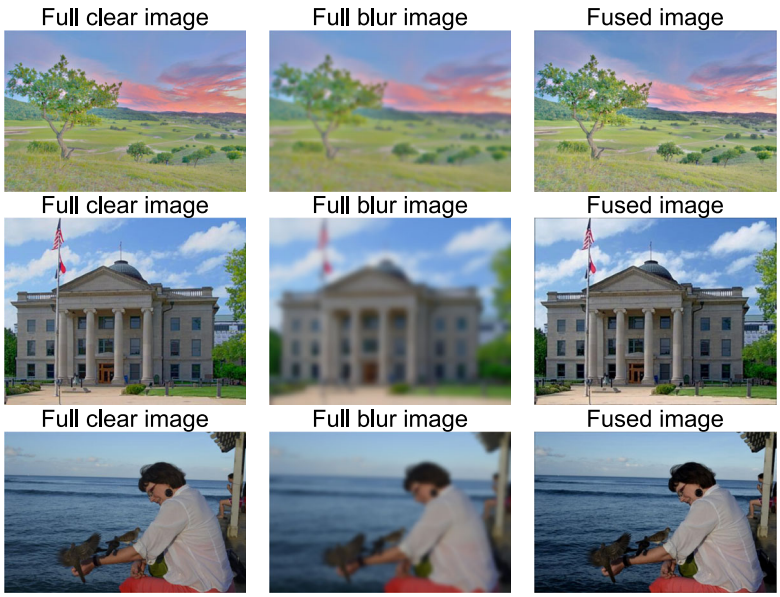

Fig. 12 Fusion performance in extreme situations

4.10 Performance in extreme conditions

To assess the robustness of the proposed SwinMFF under extreme conditions, we conducted a challenging experiment using images from the MFI-WHU dataset [35]. We carefully selected three diverse images representing different scenarios: a natural landscape, an urban building, and a lifestyle photo, as shown in Fig. 12. For each of these images, we synthesized a corresponding fully blurred version using Gaussian blur. This process simulates extreme cases of defocus that might occur in practical photography situations, such as severe camera shake or drastically misadjusted focus settings. We then fed pairs of images(consisting of the original clear image and its fully blurred counterpart) into the proposed model. Remarkably, even with one completely blurred source image as input, our model demonstrated an exceptional ability to output a fully clear image.

4.11 Consecutive multi-focus image fusion

To demonstrate the applicability of the proposed method to more than two images, we implemented it on four image sequences with multiple source images in Lytro [66] dataset. Specifically, we first fused the first two source images and then fused this intermediate result with the third source image to obtain the final output. As shown in Fig. 13, the fusion result by our method successfully retains all the in-focus regions from the multiple input images, producing a comprehensive full-clear image with visually pleasing quality. This experiment verifies that our method can be effectively extended to fuse multiple source images by applying it sequentially.

4.12 Ablation study

Swin transformer has demonstrated a positive correlation between network depth and performance in various visual tasks, such as image classification [28]. For the MFF task, it is crucial to investigate the impact of the number of swin transformer blocks in each module on fusion performance, as Transformer-based networks are generally more computationally intensive than convolutional neural networks. To explore this impact, we conducted an ablation study using the Lytro [66] dataset. Our default architecture consists of six swin transformer blocks with numbers of 2, 2, (2), (6), (2), 2, where the numbers in parentheses represent the layers in the feature extraction, feature fusion, and image reconstruction modules, respectively. This study aims to strike a balance between fusion performance and compu-

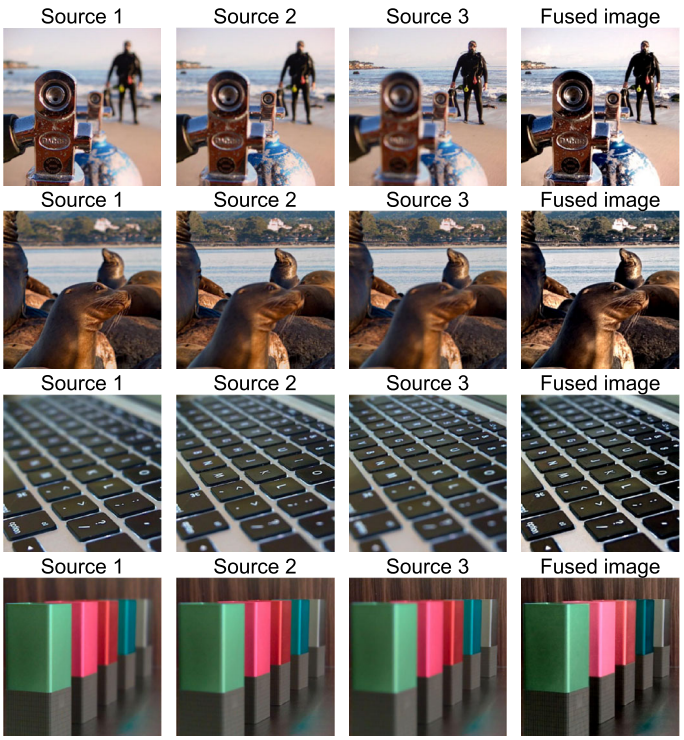

Fig. 13 Consecutive multi-focus image fusion results on Lytro [66] triplet dataset

tational efficiency, addressing the unique requirements of the MFF task.

Table 8 presents a comprehensive analysis of the impact of varying the number of swin transformer blocks in each module on fusion performance.

In the feature extraction module, we observe that increasing the number of blocks does not consistently improve performance across all metrics. Notably, the 2-block configuration achieves optimal results in three metrics (STD, MI, and EN). On the other hand, the 14-block configuration excels in four metrics (EI, SF, AVG, and VIF). This comparison highlights that a shallower network can be competitive with, and in some aspects outperform, a deeper network for specific MFF task requirements. This suggests that a shallower network in this module may be sufficient to capture the necessary features for MFF task. For the feature fusion module, the configuration with 14 blocks demonstrates superior performance in five out of eight metrics (EI, SF, AVG, MI, EN, and VIF). However, the 6-block configuration achieves the highest scores in $Q^{ab/f}$ and STD. In the image reconstruction module, the results are more varied. The 8-block configuration shows the best performance in four metrics (EI, STD, AVG, and VIF), while other configurations excel in individual metrics.

Based on these results and considering the trade-off between performance and computational efficiency, we opt for a balanced configuration of 2, 6, and 2 swin transformer blocks in three modules, respectively. This architecture strikes an optimal balance between fusion quality and model complexity, addressing the performance requirements of the MFF task while maintaining computational efficiency.

These findings highlight the complex relationship between network depth and fusion performance in MFF task. While deeper networks generally offer increased representational capacity, our results indicate that this does not always translate to improved fusion quality. This phenomenon may be attributed to the unique characteristics of the MFF task, where preserving and integrating low-level features is crucial, in contrast to tasks like image classification that benefit from high-level semantic information.

5 Discussion and conclusion

In this paper, we introduce a new simple yet powerful network architecture for end-to-end deep learning-based MFF. We construct the proposed network SwinMFF using a pure Transformer architecture, devoid of intricate design elements. We conducted a comprehensive series of qualitative and quantitative experiments across multiple public datasets. The results reveal that our end-to-end method exhibits superior image fidelity and edge processing capabilities when compared to prior methods, effectively addressing edge artifact issues. Additionally, our method stands out as not only the most potent end-to-end method presently but also as one

Table 8 The impact of the number of swin transformer blocks in different modules on fusion performance

Number	EI	$Q^{ab/f}$	STD	SF	AVG	MI	EN	VIF
Feature extraction module								
2	72.3557	0.7311	**57.9358**	19.7701	6.9656	**15.0813**	**7.5406**	1.1808
4	72.2525	0.7300	57.8962	19.7503	6.9593	15.0808	7.5404	1.1774
6	72.1691	**0.7321**	57.8876	19.7382	6.9511	15.0801	7.5400	1.1757
8	72.3211	0.7294	57.8968	19.7485	6.9638	15.0802	7.5401	1.1788
10	72.4920	0.7304	57.9074	19.7941	6.9793	15.0809	7.5405	1.1809
12	72.2922	0.7313	57.9125	19.7481	6.9616	15.0795	7.5398	1.1798
14	**72.5315**	0.7300	57.9100	**19.7989**	**6.9832**	15.0803	7.5402	**1.1814**
16	72.3543	0.7320	57.9202	19.7594	6.9675	15.0809	7.5404	1.1788
Feature fusion module								
2	72.4355	0.7290	57.9149	19.7786	6.9732	15.0794	7.5397	1.1801
4	72.4577	0.7295	57.9222	19.7709	6.9751	15.0812	7.5406	1.1809
6	72.4041	**0.7321**	**57.9737**	19.7954	6.9734	15.0826	7.5413	1.1810
8	72.3328	0.7308	57.8709	19.7607	6.9643	15.0792	7.5396	1.1781
10	72.4037	0.7312	57.9395	**19.8041**	6.9712	15.0810	7.5405	1.1820
12	72.3805	0.7293	57.8683	19.7636	6.9678	15.0787	7.5394	1.1784
14	**72.4897**	0.7277	57.9441	19.8003	**6.9801**	**15.0831**	**7.5416**	**1.1841**
16	72.1756	**0.7321**	57.9102	19.7555	6.9526	15.0800	7.5400	1.1779
Image reconstruction module								
2	72.3687	0.7315	57.9195	**19.7973**	6.9683	15.0803	7.5401	1.1796
4	72.3737	0.7286	57.9146	19.7457	6.9674	15.0813	7.5406	1.1793
6	72.3181	0.7298	57.9321	19.7599	6.9634	**15.0822**	**7.5411**	1.1793
8	**72.4778**	0.7278	**57.9420**	19.7611	**6.9763**	15.0815	7.5407	**1.1830**
10	72.3799	**0.7327**	57.9078	19.7941	6.9712	15.0802	7.5401	1.1796
12	72.2447	0.7306	57.8704	19.7482	6.9576	15.0776	7.5388	1.1742
14	72.3097	0.7320	57.9045	19.7851	6.9643	15.0792	7.5396	1.1788
16	72.3763	0.7254	57.9356	19.7483	6.9653	15.0814	7.5407	1.1803

of the top-performing methods for MFF, achieving excellent and stable performance in multiple datasets.

In future work, we plan to explore the application of this network architecture to unified image fusion tasks. Additionally, we will investigate methods to lighten the model further, as real-time multi-focus fusion is crucial in fields such as autonomous driving [5]. Moreover, considering the inherent differences between synthetic training datasets and real-world scenarios, we will continue to explore techniques for training networks with enhanced robustness and generalizability. By addressing these challenges, we hope to contribute to developing more robust, efficient, and versatile image fusion systems.

Acknowledgements Thanks for the help of the Hainan Provincial Observatory of Ecological Environment and Fishery Resource in Yazhou Bay. Also, we want to thank Yanjun Li and Chloe Alex Schaff for her contribution to polishing the article.

Author Contributions X. Xie implemented the proposed method, conducted all the experiments described in the manuscript, and wrote the main text of the manuscript. B. Guo performed the preliminary research, proposed the idea of this work, and helped write the Introduction and Related Work sections. P. Li contributed to the conceptualization of the study, organized the experimental data, provided computational resources, and acquired funding to support this work. S. He also conceived and designed the study, polished the manuscript, and supervised the project. S. Zhou prepared most of the figures in the manuscript and collected and preprocessed the datasets. All authors reviewed the manuscript.

Funding This work was supported by the Hainan Provincial Joint Project of Sanya Yazhou Bay Science and Technology City (No: 2021JJLH0079), Innovational Fund for Scientific and Technological Personnel of Hainan Province (NO. KJRC2023D19), and the Hainan Provincial Joint Project of Sanya Yazhou Bay Science and Technology City (No. 2021CXLH0020).

Data availability All the experiments are conducted utilizing publicly accessible datasets.

Code availability The code for the SwinMFF is available at https://github.com/Xinzhe99/SwinMFF. Researchers are welcome to access

the code, reproduce the results, and build upon this work for further research and applications.

Declarations

Conflict of interest The authors have no conflict of interest to declare that are relevant to the content of this article.

Ethical and informed consent for data used Not applicable.

References

1. Bacus, J.W., Grace, L.J.: Optical microscope system for standardized cell measurements and analyses. Appl. Opt. **26**(16), 3280–3293 (1987)
2. Chen, Y., Deng, N., Xin, B.-J., Xing, W.-Y., Zhang, Z.-Y.: Structural characterization and measurement of nonwoven fabrics based on multi-focus image fusion. Measurement **141**, 356–363 (2019)
3. Juočas, L., Raudonis, V., Maskeliūnas, R., Damaševičius, R., Woźniak, M.: Multi-focusing algorithm for microscopy imagery in assembly line using low-cost camera. Int. J. Adv. Manuf. Technol. **102**, 3217–3227 (2019)
4. Liang, Y., Mao, Y., Tang, Z., Yan, M., Zhao, Y., Liu, J.: Efficient misalignment-robust multi-focus microscopical images fusion. Signal Process. **161**, 111–123 (2019)
5. Li, X., Li, X., Tan, H., Li, J.: Samf: small-area-aware multi-focus image fusion for object detection. In: ICASSP 2024–2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3845–3849. IEEE (2024)
6. Burt, P.J., Adelson, E.H.: The Laplacian pyramid as a compact image code. In: Fischler MA, Firschein O (eds) Readings in Computer Vision, pp. 671–679. Morgan Kaufmann, Elsevier (1987). https://doi.org/10.1016/B978-0-08-051581-6
7. Burt, P.J., Kolczynski, R.J.: Enhanced image capture through fusion. In: 1993 (4th) International Conference on Computer Vision, pp. 173–182. IEEE (1993)
8. Lewis, J.J., O'Callaghan, R.J., Nikolov, S.G., Bull, D.R., Canagarajah, N.: Pixel-and region-based image fusion with complex wavelets. Inf Fusion **8**(2), 119–130 (2007)
9. Li, H., Manjunath, B., Mitra, S.K.: Multisensor image fusion using the wavelet transform. Graph. Models Image Process. **57**(3), 235–245 (1995)
10. Yang, B., Li, S., Sun, F.: Image fusion using nonsubsampled contourlet transform. In: Fourth International Conference on Image and Graphics (ICIG 2007), pp. 719–724. IEEE (2007)
11. Zhang, Q., Guo, B.: Multifocus image fusion using the nonsubsampled contourlet transform. Signal Process. **89**(7), 1334–1346 (2009)
12. Liu, Z., Chai, Y., Yin, H., Zhou, J., Zhu, Z.: A novel multi-focus image fusion approach based on image decomposition. Inf. Fusion **35**, 102–116 (2017)
13. Jiang, Y., Wang, M.: Image fusion with morphological component analysis. Inf. Fusion **18**, 107–118 (2014)
14. Liu, Y., Wang, L., Cheng, J., Chen, X.: Multiscale feature interactive network for multifocus image fusion. IEEE Trans. Instrum. Meas. **70**, 1–16 (2021)
15. Hu, X., Jiang, J., Liu, X., Ma, J.: Zmff: zero-shot multi-focus image fusion. Inf. Fusion **92**, 127–138 (2023)
16. Sujatha, K., Shalini Punithavathani, D.: Optimized ensemble decision-based multi-focus imagefusion using binary genetic grey-wolf optimizer in camera sensor networks. Multimed. Tools Appl. **77**, 1735–1759 (2018)
17. Kausar, N., Majid, A., Javed, S.G.: A novel ensemble approach using individual features for multi-focus image fusion. Comput. Electr. Eng. **54**, 393–405 (2016)
18. Huang, Y., Li, W., Gao, M., Liu, Z.: Algebraic multi-grid based multi-focus image fusion using watershed algorithm. IEEE Access **6**, 47082–47091 (2018)
19. Duan, J., Chen, L., Chen, C.P.: Multifocus image fusion with enhanced linear spectral clustering and fast depth map estimation. Neurocomputing **318**, 43–54 (2018)
20. Jagtap, N.S., Thepade, S.D.: High-quality image multi-focus fusion to address ringing and blurring artifacts without loss of information. Vis. Comput. **38**, 4353–4371 (2022). https://doi.org/10.1007/s00371-021-02300-5
21. Kong, W., Lei, Y.: Multi-focus image fusion using biochemical ion exchange model. Appl. Soft Comput. **51**, 314–327 (2017)
22. Duan, Z., Luo, X., Zhang, T.: Combining transformers with CNN for multi-focus image fusion. Expert Syst. Appl. **235**, 121156 (2024)
23. Li, J., Guo, X., Lu, G., Zhang, B., Xu, Y., Wu, F., Zhang, D.: Drpl: deep regression pair learning for multi-focus image fusion. IEEE Trans. Image Process. **29**, 4816–4831 (2020)
24. Li, X., Li, X., Cheng, X., Wang, M., Tan, H.: MCDFD: multifocus image fusion based on multiscale cross-difference and focus detection. IEEE Sens. J. **23**(24), 30913–30926 (2023). https://doi.org/10.1109/JSEN.2023.3330871
25. Wang, J., Qu, H., Zhang, Z., Xie, M.: New insights into multi-focus image fusion: a fusion method based on multi-dictionary linear sparse representation and region fusion model. Inf. Fusion **105**, 102230 (2024)
26. Hu, Y., Wu, P., Zhang, B., et al.: A new multi-focus image fusion quality assessment method with convolutional sparse representation. Vis. Comput. (2024). https://doi.org/10.1007/s00371-024-03351-0
27. Liu, Y., Chen, X., Peng, H., Wang, Z.: Multi-focus image fusion with a deep convolutional neural network. Inf. Fusion **36**, 191–207 (2017)
28. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022 (2021)
29. Ma, J., Tang, L., Fan, F., Huang, J., Mei, X., Ma, Y.: Swinfusion: cross-domain long-range learning for general image fusion via Swin Transformer. IEEE/CAA J. Autom. Sin. **9**(7), 1200–1217 (2022)
30. Guo, X., Nie, R., Cao, J., Zhou, D., Mei, L., He, K.: Fusegan: learning to fuse multi-focus image via conditional generative adversarial network. IEEE Trans. Multimed. **21**(8), 1982–1996 (2019)

31. Li, M., Pei, R., Zheng, T., Zhang, Y., Fu, W.: Fusiondiff: multi-focus image fusion using denoising diffusion probabilistic models. Expert Syst. Appl. **238**, 121664 (2024)
32. Lai, R., Li, Y., Guan, J., Xiong, A.: Multi-scale visual attention deep convolutional neural network for multi-focus image fusion. IEEE Access **7**, 114385–114399 (2019)
33. Ma, B., Yin, X., Wu, D., Shen, H., Ban, X., Wang, Y.: End-to-end learning for simultaneously generating decision map and multi-focus image fusion result. Neurocomputing **470**, 204–216 (2022)
34. Zang, Y., Zhou, D., Wang, C., Nie, R., Guo, Y.: UFA-FUSE: a novel deep supervised and hybrid model for multifocus image fusion. IEEE Trans. Instrum. Meas. **70**, 1–17 (2021)
35. Zhang, H., Le, Z., Shao, Z., Xu, H., Ma, J.: MFF-GAN: an unsupervised generative adversarial network with adaptive and gradient joint constraints for multi-focus image fusion. Inf. Fusion **66**, 40–53 (2021)
36. Liu, Y., Li, X., Liu Y., Zhong, W.: Simplifusion: a simplified infrared and visible image fusion network. Vis. Comput. 1–16 (2024). https://doi.org/10.1007/s00371-024-03423-1
37. Zhang, Y., Liu, Y., Sun, P., Yan, H., Zhao, X., Zhang, L.: IFCNN: a general image fusion framework based on convolutional neural network. Inf. Fusion **54**, 99–118 (2020)
38. Xu, H., Ma, J., Jiang, J., Guo, X., Ling, H.: U2Fusion: a unified unsupervised image fusion network. IEEE Trans. Pattern Anal. Mach. Intell. **44**(1), 502–518 (2020)
39. Xiao, B., Xu, B., Bi, X., Li, W.: Global-feature encoding U-Net (GEU-Net) for multi-focus image fusion. IEEE Trans. Image Process. **30**, 163–175 (2020)
40. Amin-Naji, M., Aghagolzadeh, A., Ezoji, M.: Ensemble of CNN for multi-focus image fusion. Inf. Fusion **51**, 201–214 (2019)
41. Guan, Z., Wang, X., Nie, R., Yu, S., Wang, C.: NCDCN: multi-focus image fusion via nest connection and dilated convolution network. Appl. Intell. **52**(14), 15883–15898 (2022)
42. Ma, B., Zhu, Y., Yin, X., Ban, X., Huang, H., Mukeshimana, M.: Sesf-fuse: an unsupervised deep model for multi-focus image fusion. Neural Comput. Appl. **33**, 5793–5804 (2021)
43. Wang, Y., Xu, S., Liu, J., Zhao, Z., Zhang, C., Zhang, J.: MFIF-GAN: a new generative adversarial network for multi-focus image fusion. Signal Process. Image Commun. **96**, 116295 (2021)
44. Hu, X., Jiang, J., Wang, C., Liu, X., Ma, J.: Incrementally adapting pretrained model using network prior for multi-focus image fusion. IEEE Trans. Image Process. **33**, 3950–3963 (2024). https://doi.org/10.1109/TIP.2024.3409940
45. Nazir, A., Cheema, M.N., Sheng, B., Li, H., Li, P., Yang, P., Jung, Y., Qin, J., Kim, J., Feng, D.D.: OFF-ENET: an optimally fused fully end-to-end network for automatic dense volumetric 3d intracranial blood vessels segmentation. IEEE Trans. Image Process. **29**, 7192–7202 (2020)
46. Kendall, A., Martirosyan, H., Dasgupta, S., Henry, P., Kennedy, R., Bachrach, A., Bry, A.: End-to-end learning of geometry and context for deep stereo regression. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 66–75 (2017)
47. Zhang, H., Ma, J.: SDNet: A versatile squeeze-and-decomposition network for real-time image fusion. Int. J. Comput. Vis. **129**(10), 2761–2785 (2021)
48. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł, Polosukhin, I.: Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, pp. 6000–6010 (2017)
49. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
50. Lin, X., Sun, S., Huang, W., Sheng, B., Li, P., Feng, D.D.: EAPT: efficient attention pyramid transformer for image processing. IEEE Trans. Multimed. **25**, 50–61 (2021)
51. Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R.: Swinir: Image restoration using swin transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1833–1844 (2021)
52. Vs, V., Valanarasu, J.M.J., Oza, P., Patel, V.M.: Image fusion transformer. In: 2022 IEEE International Conference on Image Processing (ICIP), pp. 3566–3570 (2022). IEEE
53. Qu, L., Liu, S., Wang, M., Song, Z.: Transmef: A transformer-based multi-exposure image fusion framework using self-supervised multi-task learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, pp. 2126–2134 (2022)
54. Ram Prabhakar, K., Sai Srikar, V., Venkatesh Babu, R.: 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 (2017)
55. Li, H., Wu, X.-J.: DenseFuse: A fusion approach to infrared and visible images. IEEE Trans. Image Process. **28**(5), 2614–2623 (2018)
56. Wang, L., Lu, H., Wang, Y., Feng, M., Wang, D., Yin, B., Ruan, X.: Learning to detect salient objects with image-level supervision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 136–145 (2017)
57. Rockinger, O.: Image sequence fusion using a shift-invariant wavelet transform. In: Proceedings of International Conference on Image Processing, vol. 3, pp. 288–291. IEEE (1997)
58. Li, S., Kang, X., Hu, J.: Image fusion with guided filtering. IEEE Trans. Image Process. **22**(7), 2864–2875 (2013)
59. Olshausen, B.A., Field, D.J.: Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature **381**(6583), 607–609 (1996)
60. Liu, Y., Wang, Z.: Simultaneous image fusion and denoising with adaptive sparse representation. IET Image Proc. **9**(5), 347–357 (2015)
61. Zhou, Z., Li, S., Wang, B.: Multi-scale weighted gradient-based fusion for multi-focus images. Inf. Fusion **20**, 60–72 (2014)
62. Paul, S., Sevcenco, I.S., Agathoklis, P.: Multi-exposure and multi-focus image fusion in gradient domain. J. Circuits Syst. Comput. **25**(10), 1650123 (2016)
63. Liu, Y., Liu, S., Wang, Z.: A general framework for image fusion based on multi-scale transform and sparse representation. Inf. Fusion **24**, 147–164 (2015)
64. Guo, D., Yan, J., Qu, X.: High quality multi-focus image fusion using self-similarity and depth information. Opt. Commun. **338**, 138–144 (2015)
65. De, I., Chanda, B.: Multi-focus image fusion using a morphology-based focus measure in a quad-tree structure. Inf. Fusion **14**(2), 136–146 (2013)
66. Nejati, M., Samavi, S., Shirani, S.: Multi-focus image fusion using dictionary-based sparse representation. Inf. Fusion **25**, 72–84 (2015)
67. Qiu, X., Li, M., Zhang, L., Yuan, X.: Guided filter-based multi-focus image fusion through focus region detection. Signal Process. Image Commun. **72**, 35–46 (2019)
68. Ma, J., Zhou, Z., Wang, B., Miao, L., Zong, H.: Multi-focus image fusion using boosted random walks-based algorithm with two-scale focus maps. Neurocomputing **335**, 9–20 (2019)
69. Zhan, K., Kong, L., Liu, B., He, Y.: Multimodal image seamless fusion. J. Electron. Imaging **28**(2), 023027–023027 (2019)
70. Zhang, X.: Deep learning-based multi-focus image fusion: a survey and a comparative study. IEEE Trans. Pattern Anal. Mach. Intell. **44**(9), 4819–4838 (2022). https://doi.org/10.1109/TPAMI.2021.3078906
71. Liu, Y., Wang, L., Cheng, J., Li, C., Chen, X.: Multi-focus image fusion: a survey of the state of the art. Inf. Fusion **64**, 71–91 (2020)

72. Xu, S., Wei, X., Zhang, C., Liu, J., Zhang, J.: Mffw: A new dataset for multi-focus image fusion. arXiv preprint arXiv:2002.04780 (2020)

Xinzhe Xie received the B.S. degree in Electronic Information Engineering from the College of Electrical and Electronic Engineering, Wenzhou University, China, in July 2022. He is currently working toward the Ph.D. degree in Ocean College, Zhejiang University. His current research interests include computer vision, image fusion, underwater sensing equipment.

Buyu Guo received the B.S. degree optical information science and technology from Weifang University, Weifang, China, in 2014 and the Ph.D degree in Marine detection technology from department of marine technology, Ocean University of China, Qingdao, China, in 2021. Since 2021, he has been a Postdoctoral Researcher Fellow with Ocean College, Zhejiang University, Hangzhou, China. His current research interests include learning-enabled smart sensors and advanced imaging techniques.

Peiliang Li received the B.S., M.S., and Ph.D. degrees in Ocean University of China, Qingdao, China, in 1994, 1998, and 2003, respectively. Between 2005 and 2018, he successively served as Associate Professor, Professor, Associate Dean and Dean in the Department of Marine Science, College of Marine and Environmental Sciences at Ocean University of China. Currently, he is the Director of the Institute of Physical Oceanography and Remote Sensing, Ocean College at Zhejiang University. His main research areas are applied oceanography, marine detection and intelligent oceanographic information sensing.

Shuangyan He received the B.S. degree in electronic information science and technology from Ocean University of China, Qingdao, China, in 2004, and the Ph.D. degree in ocean physics from Ocean University of China, Qingdao, China, in 2011. She was a postdoctoral fellow at Zhejiang University, Hangzhou, China, between 2011 and 2014, and then worked as a postdoctoral fellow at University of North Dakota, Grand Forks, USA, between 2015 and 2017. She worked as a lecturer at Zhejiang University, Zhoushan, China, from 2014 to 2018, and is an associate professor at Zhejiang University, Zhoushan, China, since 2019. Her research interest involves ocean optics, ocean color remote sensing, oceanic/atmospheric radiative transfer simulation, and remote sensing image processing.

Sangjun Zhou received the B.S. degree in Electronic Information Science and Technology from the College of Electrical and Electronic Engineering, Wenzhou University, China, in July 2022. She is currently working toward the M.S. degree in Ocean College, Zhejiang University. Her current research interests include artificial intelligence applications in the ocean.