Applied Intelligence (2025) 55:503
https://doi.org/10.1007/s10489-025-06383-8

StackMFF: end-to-end multi-focus image stack fusion network

Xinzhe Xie[1] · Jiang Qingyan[1] · Dong Chen[1] · Buyu Guo[2,3] · Peiliang Li[1,3] · Sangjun Zhou[1,3]

Accepted: 14 February 2025

Abstract
Existing end-to-end multi-focus image fusion (MFF) networks demonstrate excellent performance when fusing image pairs. However, when image stacks are processed, the necessity for iterative fusion leads to error accumulation, resulting in various types and degrees of image degradation, which ultimately limits the algorithms' practical applications. To address this challenge and expand the application scenarios of multi-focus fusion algorithms, we propose a relatively simple yet effective approach: utilizing 3D convolutional neural networks to directly model and fuse entire multi-focus image stacks in an end-to-end manner. To obtain large-scale training data, we developed a refocusing pipeline based on monocular depth estimation technology that can synthesize a multi-focus image stack from any all-in-focus image. Furthermore, we extended the attention mechanisms commonly used in image pair fusion networks from two dimensions to three dimensions and proposed a comprehensive loss function group, effectively enhancing the fusion quality. Extensive experimental results demonstrate that the proposed method achieves state-of-the-art performance in both fusion quality and processing speed while avoiding image degradation issues, establishing a simple yet powerful baseline for the multi-focus image stack fusion task. The codes are available at https://github.com/Xinzhe99/StackMFF.

Keywords Deep learning · Multi-focus image stack fusion · 3D CNNs · Synthetic dataset

1 Introduction

Owing to the inherent limitations of optical imaging systems, only objects within a specific depth of field (DoF) appear sharp during capture, whereas objects outside this range suffer from defocus blur. When the scene depth exceeds the DoF, a single image cannot maintain sharpness across all regions-a challenge particularly prevalent in microscopic imaging [1–3]. To address this limitation, multiple images are captured at varying focal settings to ensure that each scene object is sharp in at least one image. Multi-focus image fusion (MFF) techniques then merge these images into a single all-in-focus output where all objects appear sharp.

Traditional MFF approaches rely on handcrafted features and predefined fusion rules, resulting in limited adaptability and requiring manual parameter tuning. Deep learning has significantly advanced this field by enabling adaptive feature learning and data-driven fusion strategies for generating all-in-focus images. Current deep learning-based MFF methods can be categorized into two main approaches: decision map-based two-stage fusion and end-to-end fusion. Decision map-based methods first generate a binary decision map to guide pixel selection from source images for fusion. However, these methods often suffer from edge artifacts due to decision map inaccuracies [4], which compromise visual quality. In contrast, end-to-end methods directly synthesize the fused image from input pairs, eliminating intermediate steps. This direct approach demonstrates superior robust-

✉ Buyu Guo
guobuyuwork@163.com

Xinzhe Xie
xiexinzhe@zju.edu.cn

Jiang Qingyan
jqy0610@zju.edu.cn

Dong Chen
chendong@ouc.edu.cn

Peiliang Li
lipeiliang@zju.edu.cn

Sangjun Zhou
sjune163@163.com

1 Ocean College, Zhejiang University, Zhoushan 316021, Zhejiang, P. R. China

2 Donghai Laboratory, Zhoushan 316021, Zhejiang, P. R. China

3 Hainan Institute, Zhejiang University, Sanya 572025, Hainan, P. R. China

Published online: 04 March 2025

ness in complex scenes while offering improved efficiency and practicality. Given these advantages, end-to-end MFF networks are emerging as a promising paradigm in current research.

Existing end-to-end MFF networks [5–12] focus primarily on fusing two multi-focus images with complementary in-focus regions. However, practical applications often require fusing multiple images to achieve comprehensive scene clarity. While conventional wisdom suggests that multi-image fusion can be accomplished through pairwise iterative fusion, our investigation reveals significant limitations of this approach when processing larger image stacks (dozens or more images). First, iterative fusion leads to cumulative errors, manifesting as edge artifacts, noise amplification, blur, and color distortion in the final output, as illustrated in Fig. 1. Second, the sequential nature of iterative fusion creates an inherent bias where later images exert disproportionate influence on the final result, making the fusion quality heavily dependent on the input order. Third, the iterative approach introduces substantial computational overhead, as each fusion operation requires significant processing resources. Most current MFF networks lack native support for stack fusion, requiring users to implement custom iterative processing pipelines. This not only complicates deployment but also dramatically increases computational costs and processing time as the stack size increases. These limitations severely restrict the practical applicability of current MFF methods, particularly in scenarios requiring efficient processing of large image stacks.

Prior MFF approaches predominantly operate on image pairs [8, 13, 14]. Even widely used benchmarks such as the Lytro dataset [14] rarely include sequences with more than three images. This binary fusion paradigm not only constrains algorithmic development but also limits real-world applications where multiple focal planes are common. To overcome these limitations, we propose an end-to-end 3D CNN architecture that directly processes arbitrarily sized focal stacks. Our approach eliminates the error accumulation inherent in pairwise iterative fusion methods while significantly improving computational efficiency. By extending MFF to volumetric processing, this work opens new avenues for handling real-world focal stacks. The main contributions of this paper are as follows:

- This paper introduces StackMFF, an end-to-end network capable of fusing arbitrarily sized multi-focus image stacks. The proposed framework directly processes variable-sized focal stacks to generate all-in-focus images, establishing a new paradigm for multi-focus fusion. Extensive experiments demonstrate that StackMFF achieves superior performance in both fusion quality and computational efficiency compared with existing methods.
- By leveraging monocular depth estimation, we develop a physically motivated refocus pipeline that simulates the natural defocus process in optical systems. This approach faithfully reproduces the physical blur characteristics observed in real systems, enabling the synthesis of photorealistic multi-focus image stacks from arbitrary single images.
- To enable effective volumetric feature learning, we extend coordinate attention to 3D space and design a

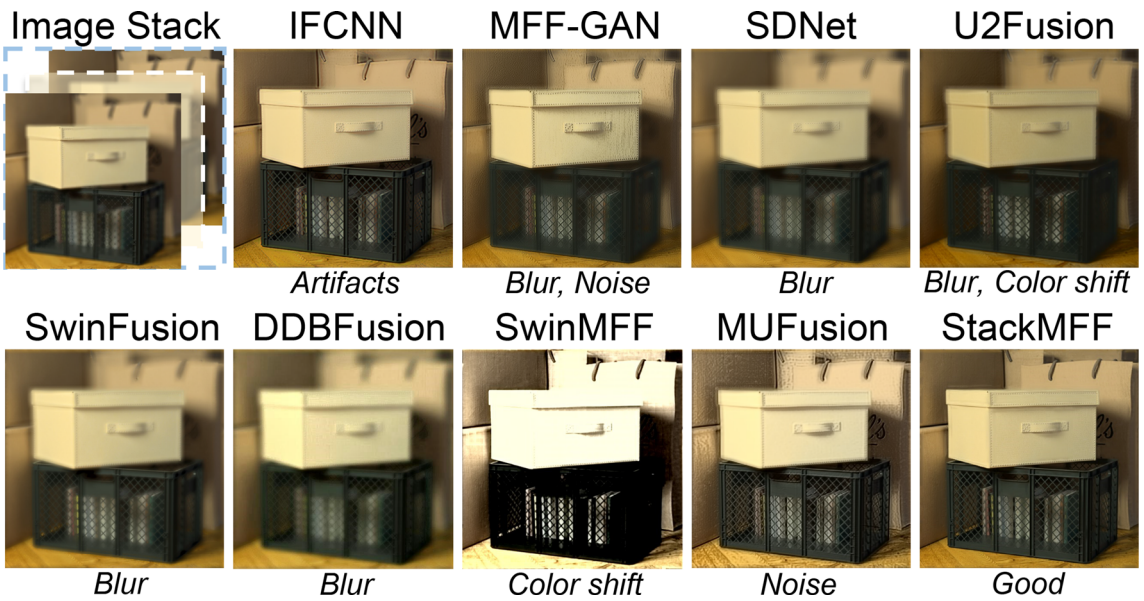

Fig. 1 Fusion results of different end-to-end fusion methods on a multi-focus image stack containing 10 images

comprehensive loss group, leading to improved fusion quality in end-to-end training.

2 Related works

2.1 Multi-focus image fusion

2.1.1 Methods based on traditional image processing

The existing MFF methods can be broadly categorized into transform domain methods and spatial domain methods. Transform domain approaches process images in alternative representations. Early works adopted the Laplacian pyramid (LP) [15] and discrete wavelet transform (DWT) [16] for multiscale decomposition. To address the limitations of traditional wavelets, such as shift sensitivity, researchers have explored more advanced transforms, including the dual-tree complex wavelet transform (DTCWT) [17] and nonsubsampled contourlet transform (NSCT) [18]. Sparse representation (SR) techniques [19] were later introduced to exploit signal sparsity. However, these methods often suffer from high computational complexity and parameter sensitivity [20].

Spatial domain methods operate directly at the pixel level through block-based, region-based, or pixelwise strategies. Block-based methods [21] partition source images into fixed-size blocks and measure activity levels via spatial frequency. Region-based approaches extend this concept by incorporating image segmentation for more flexible partitioning. Recent pixel-based methods, including guided filtering (GFF) [22], multiscale weighted gradient (MWG) [23], and SIFT features (DSIFT) [24], achieve precise fusion weight estimation. However, these approaches rely on handcrafted features such as gradients and structural information, which may not fully capture image characteristics for optimal fusion [25].

2.1.2 Methods based on deep learning

Recent years have witnessed the rapid development of deep learning-based MFF methods, which demonstrate superior performance by learning fusion strategies directly from data without manual parameter tuning. These approaches can be broadly categorized into two paradigms: decision map-based methods and end-to-end regression methods. Decision map-based methods focus on learning fusion weight maps. Early work [26] pioneered the use of CNNs for generating pixel-level decision maps. Subsequent studies enhanced the fusion quality through architectural innovations [20, 27] and the integration of traditional priors [28, 29]. Recent advances include adversarial learning [30], zero-shot fusion via deep image priors [31], and hybrid architectures that combine CNNs with transformers [25]. Notable progress has also been made in incremental learning through the INPA framework [32].

End-to-end regression methods directly synthesize fused images from input pairs, showing advantages in handling complex scenes and edge preservation. Representative works include IFCNN [5], U2Fusion [9], SDNet [6], MFF-GAN [8], SwinMFF [10], MUFusion [11], DDBFusion [12] and SwinFusion [7]. Recent exploration has extended to diffusion models [33], further pushing the boundaries of fusion quality. Despite these advances, existing methods are designed primarily for image pairs. When applied iteratively to image stacks, they suffer from error accumulation and image degradation. This limitation motivated our work to develop a more flexible architecture that directly processes arbitrarily sized focal stacks.

2.2 Image refocusing technology

Image refocusing technology has evolved through different approaches that are based on various input modalities. Light field cameras provide a specialized hardware solution for postcapture refocusing. While light field cameras enable postcapture refocusing through their unique 4D Light Field structure, this capability is contingent upon specialized hardware configurations, limiting their practical applications. To enable refocusing from conventional images, several single-image methods have been proposed. Bando et al. [34] pioneered single-image refocusing by estimating depth from defocus cues and modeling the physical defocus process. Zhang and Cham [35] further advanced this direction by developing a more robust depth estimation approach based on local blur analysis and introducing an efficient defocus rendering pipeline. However, these single-image methods often struggle with complex scenes because of the inherent ambiguity in depth estimation from monocular cues. More recently, stereo-based approaches have emerged as a promising direction. Busam et al. [36] introduced StereFo, which leverages stereo matching for accurate depth estimation and implements physically based defocus rendering. While requiring two views, this approach achieves more reliable depth estimation than single-image methods do, leading to improved refocusing quality. Recent advances in learning-based monocular depth estimation [37, 38], trained on large-scale datasets, have opened new possibilities for single-image refocusing. These methods can produce high-quality depth maps from a single image without relying on traditional depth cues or multiple views, enabling more robust and flexible refocusing applications. This work leverages such advances to develop a physically motivated refocusing pipeline that can generate realistic multi-focus image stacks from arbitrary single images.

2.3 3D convolutional neural network

3D CNNs extend traditional 2D architectures to process volumetric data, employing 3D convolution kernels and pooling operations to extract spatiotemporal features. Originally introduced in C3D [39] for video analysis, 3D CNNs have demonstrated remarkable success in various domains, including action recognition [39], medical image analysis [40], and scene reconstruction [41]. The key insight of this work stems from the observation that multi-focus image stacks share fundamental characteristics with volumetric medical data, where critical information is distributed across different slices. In multi-focus stacks, optimal focus typically exists in specific layers for each spatial location, making the ability to aggregate information along the depth dimension crucial for effective fusion. While 2D CNNs excel at spatial feature extraction, they inherently lack the capacity to model such depthwise relationships. This paper leverages 3D CNNs to directly model and process entire focal stacks end-to-end. The proposed framework captures complex spatial-focal relationships across arbitrarily sized image sequences, eliminating the need for iterative processing and its associated error accumulation. This approach not only enables flexible handling of variable-sized inputs but also enhances fusion quality by leveraging comprehensive stackwise feature learning.

3 Methods

In this section, we first present the network architecture StackMFF (Fig. 2). We subsequently delineate the loss function group employed for training StackMFF. Finally, we describe the generation of multi-focus image stack datasets through the proposed refocusing technique.

3.1 Network structure

3D CNNs naturally align with multi-focus image stack fusion by processing volumetric data directly. We adopt Unet3D [40] as our encoder for its multiscale architecture, which captures both fine details and global context, skip connections that preserve high-frequency information critical for focus discrimination, and inherent 3D operations that effectively model focus variations across the image stack. This design enables simultaneous processing of spatial and depth information, facilitating optimal fusion of features from different focal planes.

The encoder employs a two-stage architecture for feature transformation. First, it compresses the input $X \in R^{C\times D\times H\times W}$ into a compact latent representation $X \in R^{256\times D\times H/16\times W/16}$, where C, D, H, and W denote the input channels, stack depth, height, and width, respectively. These features are subsequently upsampled to $X \in R^{16\times D\times H\times W}$, enabling the network to capture high-level abstractions while preserving spatial fidelity. This design effectively standardizes variable-depth image stacks into a unified feature space while maintaining the spatial dimensions of the input.

To enhance spatial awareness and facilitate cross-layer feature aggregation, we extend the original coordinate attention (CA) [42] to 3D space (named 3DCA), enabling explicit modeling of positional relationships and capturing long-range dependencies. The resulting feature maps are then compressed into a single-channel focus representation through a 3D convolution layer, producing a focus volume that aligns with the original input dimensions. Subsequently, max pooling along the depth dimension transforms the features from $X \in R^{16\times D\times H\times W}$ to $X \in R^{C\times 1\times H\times W}$, effectively mapping the sharpest pixels across different focal planes onto a unified 2D representation. Finally, we employ a single convolutional layer for noise suppression, followed by sigmoid activation to normalize the output to [0, 1], yielding the final fused image.

3.2 Loss function

We propose a multi-component loss function to effectively preserve high-frequency information during multi-focus image fusion:

$$L = \alpha L_{MAE} + \beta L_{SSIM} + \lambda L_{Gra} + \gamma L_{SF} \tag{1}$$

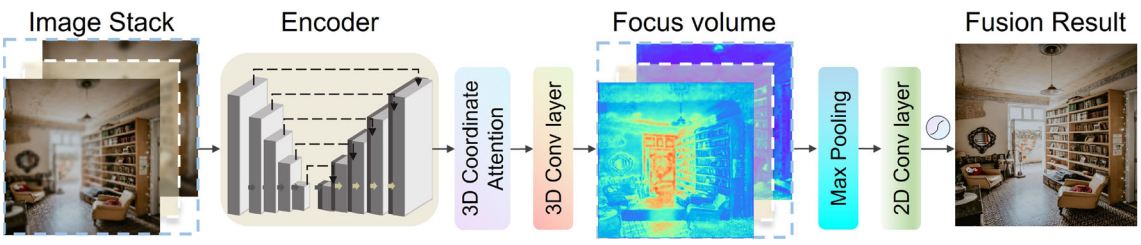

Fig. 2 Framework of the proposed StackMFF

The mean absolute error (MAE) and structural similarity index (SSIM) losses are defined as:

$$L_{MAE} = \frac{1}{HW}\sum_{i=1}^{HW}|I^i_{pred} - I^i_{gt}| \tag{2}$$

$$L_{SSIM} = 1 - \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \tag{3}$$

where I_{pred} and I_{gt} denote the predicted and ground truth images, respectively, with spatial dimensions $H \times W$. For L_{SSIM}, μ and σ represent the mean and standard deviation of the image patches, respectively, with c_1 and c_2 as stabilizing constants.

To enhance focus detection and detail preservation, we introduce two complementary losses. The gradient loss employs the Laplacian operator to preserve high-frequency details:

$$L_{Gra} = \frac{1}{HW}\sum_{i=1}^{HW}|\nabla_1 I^i_{pred} - \nabla_1 I^i_{gt}| \tag{4}$$

where ∇_1 represents the Laplacian operator that captures local intensity variations. Additionally, we incorporate spatial frequency loss to maintain the overall frequency characteristics of the fused image:

$$L_{SF} = \frac{1}{HW}\sum_{i=1}^{HW}|\nabla_2 I^i_{pred} - \nabla_2 I^i_{gt}| \tag{5}$$

where ∇_2 denotes the spatial frequency operator that measures the overall activity level in an image by analyzing row and column frequency changes. The combination of these four loss components ensures comprehensive supervision of both structural and frequency domain characteristics during the fusion process, with the weighting parameters α, β, λ, and γ controlling their relative contributions.

3.3 Image refocusing

Obtaining ground truths for real multi-focus images remains a fundamental challenge in MFF research, as it requires the capture of multiple images of the same scene with different focal settings-a process that is time-consuming and prone to alignment errors. To address this challenge, we propose leveraging image refocusing techniques to synthesize multi-focus image stacks from all-in-focus images. Since the appearance of defocus blur in images is inherently related to scene depth, we introduce a novel refocusing pipeline that uses state-of-the-art monocular depth estimation networks Metric3D [37] to obtain depth maps and simulate depth-of-field variations in optical systems based on scene depth.

First, we introduce the causes of defocusing. The formation of defocus blur in optical systems can be understood through the thin lens imaging model [43], as shown in Fig. 3. When an object point at distance d is not in perfect focus, it projects onto the sensor as a blurred circle of confusion (CoC). The CoC radius r can be calculated as:

$$r(d) = \frac{|d - s|f^2}{2d(s - f)F} \tag{6}$$

where s represents the object–space distance, f denotes the lens focal length, and F is the f-number that determines the aperture size.

For a multi-focus image stack I_i $(i \in 1, 2, ..., N)$, the different object-space distances s_i for each image in the stack are set based on the ground-truth depth map d:

$$s_i = \min(d) + k_i(\max(d) - \min(d)) \tag{7}$$

Here, $k_i \in [0, 1]$ serves as a control parameter for focal plane positioning. The focal plane coincides with the nearest depth when $k_i = 0$ and with the farthest depth when $k_i = 1$. We achieve comprehensive scene coverage by sampling k_i values uniformly between 0 and 1.

The defocus blur in each image is characterized by a point spread function (PSF) modeled as a Gaussian kernel:

$$K_i(x, y) = \frac{1}{2\pi\sigma_i^2(x, y)}\exp\left(-\frac{x^2 + y^2}{2\sigma_i^2(x, y)}\right) \tag{8}$$

where (x, y) indicates the pixel position. Given the thin-lens model illustrated in Fig. 3 and assuming the standard deviation $\sigma_i = 2k \cdot r$, where $0 < k \leq 0.5$, we derive:

$$\sigma_i = \frac{|d - s_i|kf^2}{\mathrm{d}(s_i - f)F} \tag{9}$$

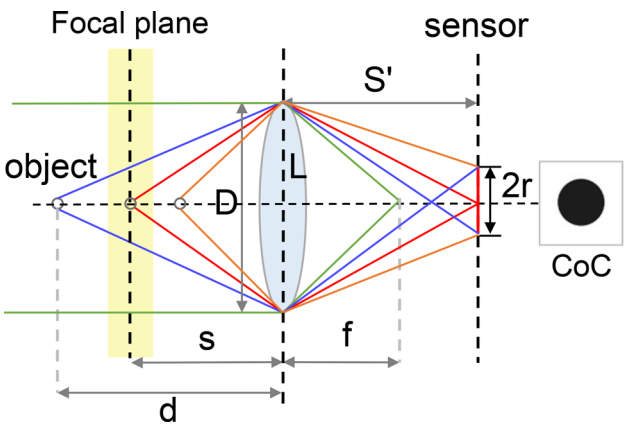

Fig. 3 A thin-lens diagram showing how defocus blur emerges on the sensor when subjects are displaced from the focal plane

The defocused image I_i is then generated by convolving the all-in-focus image I^{gt} with the corresponding Gaussian kernel:

$$I_i = I^{gt} * K_i \tag{10}$$

where $*$ represents the convolution operation.

Figure 4 illustrates the proposed refocusing technology for synthesizing realistic multi-focus image stacks from all-in-focus images. Given an all-in-focus input image, we first estimate its depth map via Metric3D [37]. Given the depth values, we partition the scene into N regions (N=16 in our implementation) to simulate distinct depth-of-field zones. This partitioning generates N binary masks, where pixels with a value of 1 represent in-focus regions within the depth of field, whereas 0 indicates out-of-focus areas. For each mask, we synthesize a defocused image by computing the blur magnitude required for each pixel based on its depth difference from the current in-focus region. Specifically, pixels farther from the in-focus region receive stronger Gaussian blur, whereas those closer to the in-focus region experience weaker blur, resulting in a blur parameter matrix. Applying these spatially varying blur parameters to the original image generates a synthetic defocused image where only regions marked as 1 in the mask remain perfectly sharp, with other regions exhibiting gradual blur. This process is repeated for each mask to create a complete multi-focus image stack. For efficient implementation, we apply N different levels of Gaussian blur to the source image, producing N images with varying blur intensities. The final defocused images are then composed by selecting pixels from these precomputed blur levels according to the index of each mask.

The proposed refocusing technology offers significant advantages over existing approaches for multi-focus image stack synthesis. Leveraging monocular depth estimation enables the generation of multi-focus image stacks from arbitrary all-in-focus images, eliminating the need for specialized hardware (e.g., light field cameras) or multiple views (e.g., stereo pairs). This flexibility dramatically expands the potential training data sources, allowing the synthesis of large-scale, diverse datasets from readily available all-in-focus images. As demonstrated in Fig. 5, the proposed pipeline can effectively simulate focus variations across different scene depths.

4 Experiments

4.1 Implementation details

We trained our model on the Open Images Dataset V7 validation subset, a large-scale collection covering diverse real-world scenes and objects, comprising 37,458 training images and 4,162 validation images (384×384 pixels). The proposed StackMFF network was implemented in PyTorch and trained end-to-end on dual NVIDIA A6000 GPUs. We employed SGD optimization with a momentum of 0.9, a weight decay of $1e^{-4}$, and an initial learning rate of $5e^{-3}$ with exponential decay (gamma=0.9). The loss weights α, β, λ, γ were set to 0.1, 0.7, 0.1, 0.1. The network was trained for 20 epochs with a batch size of 16.

4.2 Experimental setting

4.2.1 Datasets for evaluation

Several widely used datasets for evaluating multi-focus image fusion performance have been established, including Lytro [14] (comprising 20 image pairs and 3 sets of 3-source images), MFI-WHU [8] (containing 120 image pairs), and MFFW [44] (featuring 13 image pairs). While these benchmark datasets in the field of MFF have significantly advanced the development and assessment of multi-focus image pair fusion algorithms, a notable gap remains in suitable datasets for evaluating the performance and robustness of algorithms on multi-focus image stacks.

To conduct a comprehensive evaluation of multi-focus image stack fusion performance, we utilized datasets originally designed for depth estimation in the field of depth from focus. These datasets have been widely adopted in the domain and include the 4D-Light-Field Dataset [45], FlyingThings3D [46], Middlebury Stereo [47], and Mobile Depth [48]. All four datasets are employed in our quantitative evaluation, whereas for qualitative analysis, we focus on

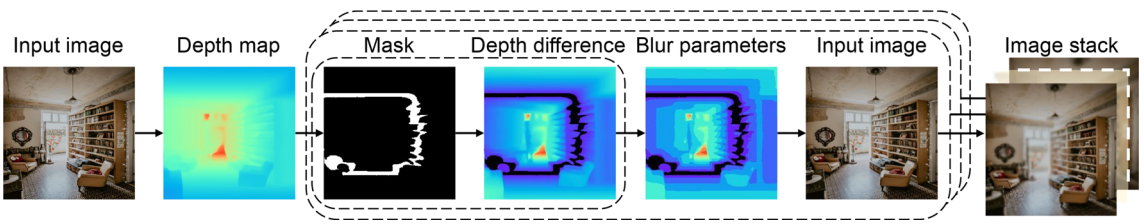

Fig. 4 Schematic diagram of the proposed refocusing pipeline

Fig. 5 Example of a synthetic multi-focus image stack illustrating the gradual shift of in-focus areas from near to far

real-world scenarios using the 4D-Light-Field Dataset, Middlebury Stereo, and Mobile Depth datasets, as they better reflect the challenges encountered in practical applications. These datasets vary significantly in complexity and size:

- 4D-Light-Field dataset: Each multi-focus image stack contains 10 images of 512 × 512 resolution.
- FlyingThings3D dataset: Each stack comprises 15 images of 960 × 540 resolution.
- Middlebury dataset: Each stack includes 15 images of varying resolutions (maximum: 741x497, minimum: 347x277).
- Mobile Depth dataset: Each stack contains approximately 30 images of varying resolutions (maximum: 774x518, minimum: 640x360).

All image stacks were aligned, and more detailed information on these datasets is provided in Table 1. By leveraging these established datasets, we aim to provide a thorough assessment of our fusion algorithms across a diverse range of scenes and imaging conditions. These datasets allow for a robust comparison with existing methods and facilitate a more comprehensive understanding of our algorithm's performance in various scenarios.

4.2.2 Methods for comparison

To comprehensively evaluate the performance of our proposed end-to-end multi-focus image stack fusion method, StackMFF, we conducted comparative analyses against fourteen state-of-the-art approaches. These include six traditional methods: CVT (Curvelet transform) [49], DWT (Discrete wavelet transform) [16], DCT (Discrete cosine transform) [50], DSIFT (Dense scale-invariant feature transform) [24], DTCWT (Dual-tree complex wavelet transform)) [17], and NSCT (Nonsubsampled contourlet transform) [18]. The eight deep learning-based methods include IFCNN [5], U2Fusion [9], SDNet [6], MFF-GAN [8], SwinFusion [7], MUFusion [11], SwinMFF[10] and DDBFusion [12]. While the official implementations of these methods were originally designed to fuse only two images at a time, we have modified their code to enable the fusion of multi-focus image stacks containing more than two images in an iterative manner. We will also release the modified code to facilitate direct use by researchers.

4.2.3 Evaluation metrics

Given that ground truth images are available for all the evaluation datasets, we can employ reference-based evaluation metrics to assess the fusion quality. These metrics include the mean-absolute error (MAE), mean-squared error (MSE), root-mean-squared error (RMSE), and log root mean-squared error (logRMS). For the quantitative evaluation presented in the following sections, we consider only single-channel grayscale images.

4.3 Qualitative comparison

Figs. 6 and 7 present qualitative results on the "sideboard" scene and "table" scene from the 4D-Light-Field [45] dataset. Traditional approaches (CVT, DWT, DCT, DTCWT, DSIFT, and NSCT) demonstrate reliable performance in handling the image stack fusion task. However,

Table 1 Summary of the evaluation datasets

Dataset	Image source	Cause of defocus	Ground Truth
4D-Light-Field [45]	Synthetic	Light-field composition	✓
FlyingThings3D [46]	Synthetic	Disparity rendering	✓
Middlebury Stereo [47]	Real	Disparity rendering	✓
Mobile Depth [48]	Real	Real	✓

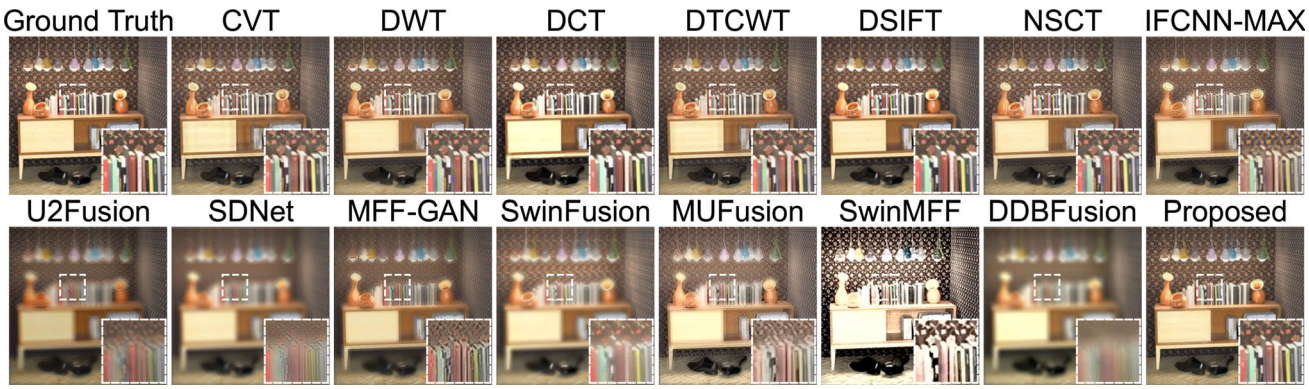

Fig. 6 Visual comparison of multi-focus image fusion methods on the 4D-Light-Field dataset [45]. The example shown is the "sideboard" scene

most learning-based methods encounter significant challenges. U2Fusion, SDNet, SwinFusion, and DDBFusion produce severely blurred results across the entire image, failing to effectively fuse the multi-focus image stack. While IFCNN-MAX shows relatively better performance than these methods do, the fusion result still exhibits noticeable blur. MFF-GAN and MUFusion not only suffer from blurriness but also introduce considerable noise in the fused images. Among all existing learning-based methods, SwinMFF achieves the sharpest fusion results but introduces significant color distortion. These image degradation issues can be attributed to error accumulation during the fusion process. In contrast, our proposed StackMFF successfully overcomes these limitations, producing fusion results that closely match the ground truth in terms of both sharpness and color fidelity.

The Middlebury Stereo dataset [47] provides real-world scenarios that enable comprehensive evaluation of the robustness and generalization capabilities of fusion methods. Figure 8 illustrates the comparative performance on a challenging case featuring intricate text details in the motorcycle logo region. While DSIFT achieves notable performance in preserving the original in-focus information, other conventional approaches (CVT, DWT, DTCWT, and NSCT) struggle to maintain clarity in the highlighted logo area. Among the learning-based methods, most fail to deliver satisfactory results. U2Fusion completely fails to preserve text details, rendering the letters illegible. While SDNet shows marginal improvement, it still introduces noticeable blur artifacts. MFF-GAN and MUFusion generate severe noise that significantly degrades image quality, making text identification impossible. SwinFusion, while avoiding noise artifacts, has two critical issues: inadequate fusion of in-focus areas resulting in blur and notable color distortion that compromises visual authenticity. DDBFusion results in severe degradation with global blurriness, indicating substantial fusion failure. The IFCNN produces relatively better results with discernible text, although it still suffers from visible noise artifacts. In contrast, the proposed StackMFF, along with the traditional DSIFT method, achieves superior performance, accurately preserving fine details and maintaining the fidelity of in-focus regions. These results demonstrate the robust fusion capability of our approach across diverse real-world scenarios, particularly in preserving critical high-frequency details while avoiding common artifacts such as noise and color distortion.

The Mobile Depth dataset [48] provides real-world focal stacks captured via consumer mobile devices and cameras,

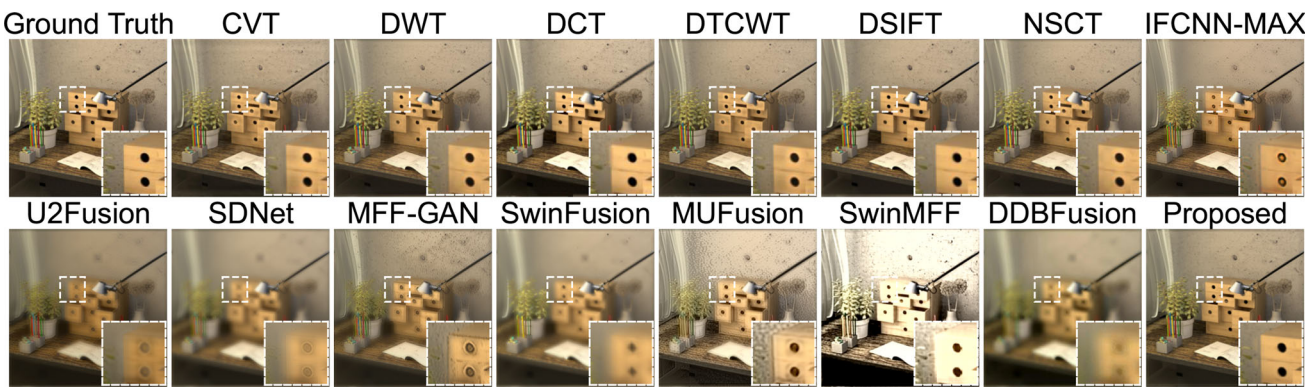

Fig. 7 Visual comparison of multi-focus image fusion methods on the 4D-Light-Field dataset [45]. The example shown is the "table" scene

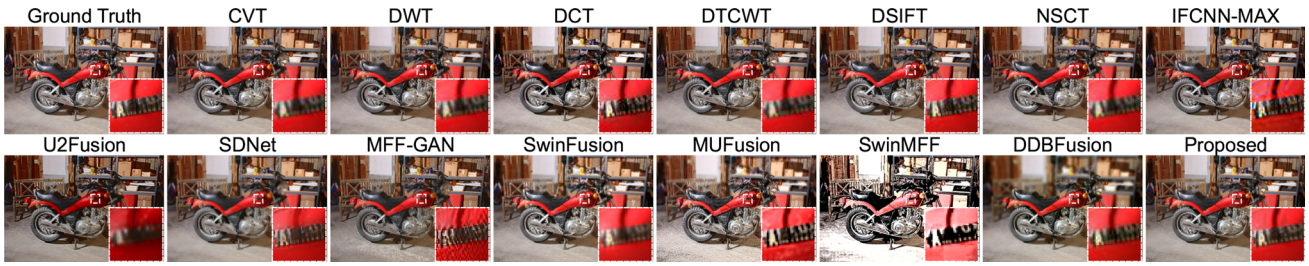

Fig. 8 Visual comparison of the Middlebury Stereo datasets [47]. The example shown is the "Motorcycle" scene

offering an excellent testbed for evaluating fusion methods under practical conditions. Figures 9 and 10 present qualitative comparisons of the challenging "balls" and "keyboard" scenes, respectively. In the "balls" scene, traditional methods consistently demonstrate robust fusion performance. However, learning-based methods exhibit various degradation issues. IFCNN, MFF-GAN, SwinMFF, and MUFusion introduce significant noise that severely compromises the visual quality of the fused results. U2Fusion, DDBFusion, and SwinFusion fail to integrate effectively in focused regions. The "keyboard" scene further highlights the performance gap between different approaches. MFF-GAN and SwinMFF yield severe degradation artifacts, significantly compromising the fusion quality. While the IFCNN demonstrates relatively good performance in this scene, other methods still struggle with various issues: MUFusion introduces distinct stripe-pattern noise artifacts, and SDNet exhibits noticeable noise points throughout the fused image. In contrast, our proposed StackMFF consistently achieves superior performance across both scenarios, successfully preserving sharp details while avoiding common artifacts such as noise and fusion failures. The results demonstrate our method's robust generalization ability in handling diverse real-world scenes captured by consumer devices. These comprehensive comparisons across different scene types underscore the effectiveness of our approach in maintaining consistent fusion quality while avoiding the degradation issues that plague existing learning-based methods.

4.4 Quantitative comparison

Table 2 presents extensive quantitative evaluations across four benchmark datasets: 4D-Light-Field, FlyingThings3D, Middlebury, and Mobile Depth. We adopt four widely used metrics (MSE, MAE, RMSE, and logRMS) for comprehensive performance assessment, where lower values indicate better performance.

The experimental results demonstrate that our proposed StackMFF achieves state-of-the-art performance across multiple datasets. Specifically, StackMFF outperforms all competing methods on both the FlyingThings3D and Middlebury datasets across all the metrics. For the 4D-Light-Field dataset, our method achieves optimal performance in three metrics (MAE, RMSE, and logRMS) while maintaining competitive performance in terms of the MSE. For the Mobile Depth dataset, StackMFF consistently ranks second across all the metrics and is only marginally surpassed by DCT.

Notably, our experimental analysis reveals an interesting phenomenon: traditional methods, particularly DCT, exhibit robust performance that surpasses many recent deep learning-based approaches. For example, DCT achieves superior performance on the Mobile Depth dataset and maintains competitive performance across other benchmarks. This observation suggests that the iterative fusion strategy employed by traditional methods may possess inherent advantages in mitigating error accumulation during the fusion process. The superior performance of StackMFF can

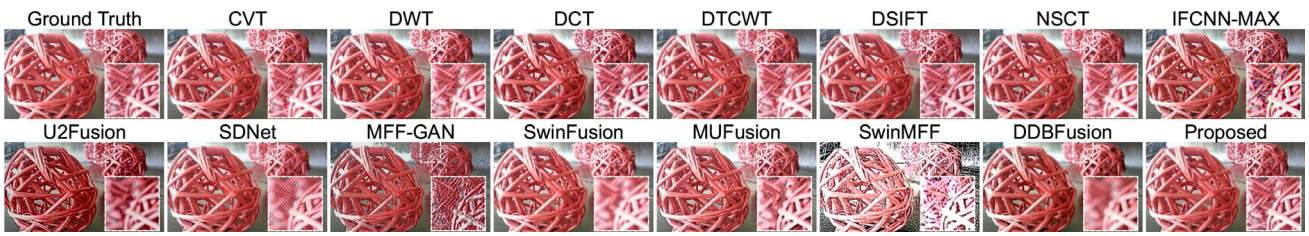

Fig. 9 Visual comparison on the Mobile Depth dataset [48]. The example shown is the "balls" scene

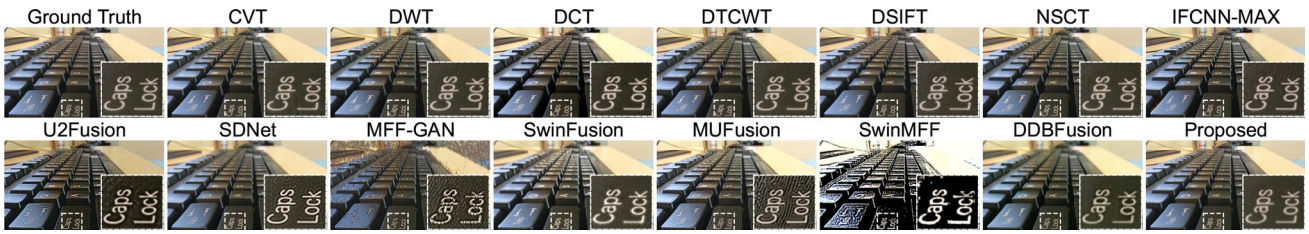

Fig. 10 Visual comparison of the Mobile Depth dataset [48]. The example shown is the "keyboard" scene

be attributed to its novel stackwise modeling approach, which effectively addresses the limitations of iterative fusion strategies. By jointly processing the entire image stack, our method demonstrates enhanced generalization capability and fusion quality across diverse scenarios. These results validate the effectiveness of our proposed architecture and suggest a promising direction for future research in multi-focus image fusion.

Table 2 Quantitative comparisons of the four datasets

Dataset	4D-Light-Field				FlyingThings3D			
Method	*MSE* ↓	*MAE* ↓	*RMSE* ↓	*logRMS* ↓	*MSE* ↓	*MAE* ↓	*RMSE* ↓	*logRMS* ↓
CVT	0.0101	0.0490	0.0613	0.0439	0.0011	0.0199	0.0320	0.0225
DWT	**0.0034**	0.0351	0.0483	0.0353	0.0011	0.0191	0.0324	0.0226
DCT	0.0050	0.0226	0.0304	0.0216	0.0006	0.0072	0.0160	0.0115
DSIFT	0.0102	0.0495	0.0637	0.0455	0.0011	0.0193	0.0318	0.0223
DTCWT	0.0098	0.0445	0.0581	0.0416	0.0007	0.0130	0.0251	0.0180
NSCT	0.0102	0.0491	0.0638	0.0456	0.0011	0.0191	0.0324	0.0226
IFCNN-MAX	0.0102	0.0499	0.0644	0.0459	0.0009	0.0163	0.0286	0.0206
U2Fusion	0.0072	0.0437	0.0526	0.0372	0.0027	0.0294	0.0363	0.0261
SDNet	0.0060	0.0311	0.0417	0.0299	0.0007	0.0095	0.0182	0.0131
MFF-GAN	0.0058	0.0325	0.0418	0.0299	0.0021	0.0231	0.0324	0.0235
SwinFusion	0.0063	0.0324	0.0424	0.0309	0.0008	0.0131	0.0197	0.0144
MUFusion	0.0097	0.0737	0.0903	0.0613	0.0106	0.0813	0.1026	0.0698
SwinMFF	0.0375	0.1617	0.1891	0.1356	0.0602	0.2114	0.2444	0.1756
DBBFusion	0.0093	0.0725	0.0914	0.0656	0.0054	0.0544	0.0720	0.0504
Proposed	0.0048	**0.0221**	**0.0287**	**0.0203**	**0.0004**	**0.0061**	**0.0126**	**0.0090**
Dataset	**Middlebury**				**Mobile Depth**			
Method	*MSE* ↓	*MAE* ↓	*RMSE* ↓	*logRMS* ↓	*MSE* ↓	*MAE* ↓	*RMSE* ↓	*logRMS* ↓
CVT	0.0014	0.0214	0.0361	0.0247	0.0007	0.0149	0.0247	0.0164
DWT	0.0015	0.021	0.0367	0.025	0.0008	0.0152	0.0260	0.0172
DCT	0.0010	0.0095	0.0209	0.0145	**0.0002**	**0.0058**	**0.0102**	**0.0071**
DSIFT	0.0014	0.0209	0.0359	0.0245	0.0007	0.0144	0.0243	0.0161
DTCWT	0.0010	0.0158	0.0294	0.0203	0.0003	0.0113	0.0182	0.0125
NSCT	0.0015	0.0210	0.0367	0.0250	0.0008	0.0152	0.0260	0.0172
IFCNN-MAX	0.0015	0.0224	0.0366	0.0262	0.0037	0.0345	0.0585	0.0406
U2Fusion	0.0050	0.0415	0.0493	0.0346	0.0051	0.0424	0.0500	0.0347
SDNet	0.0010	0.0125	0.0213	0.0148	0.0020	0.0174	0.0310	0.0211
MFF-GAN	0.0050	0.0320	0.0493	0.0348	0.0222	0.0650	0.1016	0.0737
SwinFusion	0.0011	0.0156	0.0231	0.0163	0.0018	0.0179	0.0294	0.0194
MUFusion	0.0107	0.0821	0.1029	0.0690	0.0136	0.0936	0.1161	0.0802
SwinMFF	0.0668	0.2202	0.2569	0.1682	0.0848	0.2526	0.2887	0.1923
DBBFusion	0.0044	0.0447	0.0656	0.0454	0.0025	0.0333	0.0493	0.0350
Proposed	**0.0005**	**0.0083**	**0.0151**	**0.0104**	0.0003	0.0062	0.0120	0.0078

Table 3 Comparison of running times (s) for various methods

Method	4D-Light-Field	FlyingThings3D	Middlebury	Mobile Depth	Device
CVT	17.8	37.87	31.37	31.37	CPU
DWT	1.61	6.75	8.62	5.34	CPU
DCT	1.91	6.04	3.3	4.97	CPU
DSIFT	27.29	75.35	64.97	90.91	CPU
DTCWT	5.1	14.7	9.4	11.44	CPU
NSCT	81.39	133.84	165.13	231.84	CPU
IFCNN-MAX	0.55	0.78	0.5	0.55	GPU
U2Fusion	14.52	45.1	35.9	41.04	CPU
SDNet	3.05	14.04	5.26	9.68	CPU
MFF-GAN	4.09	10.06	8.88	6.4	CPU
SwinFusion	11.02	32.33	19.53	28.21	GPU
MUFusion	13.37	55.02	21.98	40.4	GPU
SwinMFF	9.15	34.05	18.23	27.97	GPU
DBBFusion	17.11	41.98	30.06	33.89	GPU
Proposed	**0.15**	**0.24**	**0.19**	**0.22**	GPU

4.5 More analysis

4.5.1 Model efficiency comparison

Table 3 presents a comprehensive efficiency analysis across multiple datasets: 4D-Light-Field [45], FlyingThings3D [46], Middlebury Stereo [47], and Mobile Depth [48]. The evaluation encompasses traditional methods, recent deep learning approaches, and our proposed StackMFF, with processing times measured in seconds (s) per image stack. Our method demonstrates unprecedented computational efficiency, consistently outperforming both traditional and learning-based approaches across all datasets. Notably, StackMFF achieves processing times ranging from 0.15 to 0.24 s, representing a significant improvement over existing GPU-based methods. Compared with recent transformer-based approaches such as SwinFusion (32.33 s) and MUFusion (55.02 s) on the FlyingThings3D dataset, our method (0.24 s) achieves speed improvements of 135× and 229×, respectively. Compared with the lightweight IFCNN-MAX, StackMFF achieves a 3.3× speedup (0.24 s vs. 0.78 s). The efficiency gains are particularly noteworthy compared with those of CPU-based methods. Taking the 4D-Light-Field dataset as an example, StackMFF (0.15 s) outperforms the fastest CPU method (DWT, 1.61 s) by an order of magnitude. Our method also shows remarkable scalability, with

Table 4 Ranking of different methods on various datasets

Method	4D-Light-Field	FlyingThings3D	Middlebury	Mobile Depth
CVT	11	7	6	4
DWT	1	7	8	6
DCT	3	2	2	1
DSIFT	12	7	6	4
DTCWT	10	3	2	2
NSCT	12	7	8	6
IFCNN-MAX	12	6	8	11
U2Fusion	7	12	12	12
SDNet	5	3	2	9
MFF-GAN	4	11	12	14
SwinFusion	6	5	5	8
MUFusion	9	14	14	13
SwinMFF	15	15	15	15
DBBFusion	8	13	11	10
Proposed	2	1	1	2

(Ranked according to the MSE; the smaller the value is, the higher the ranking.)

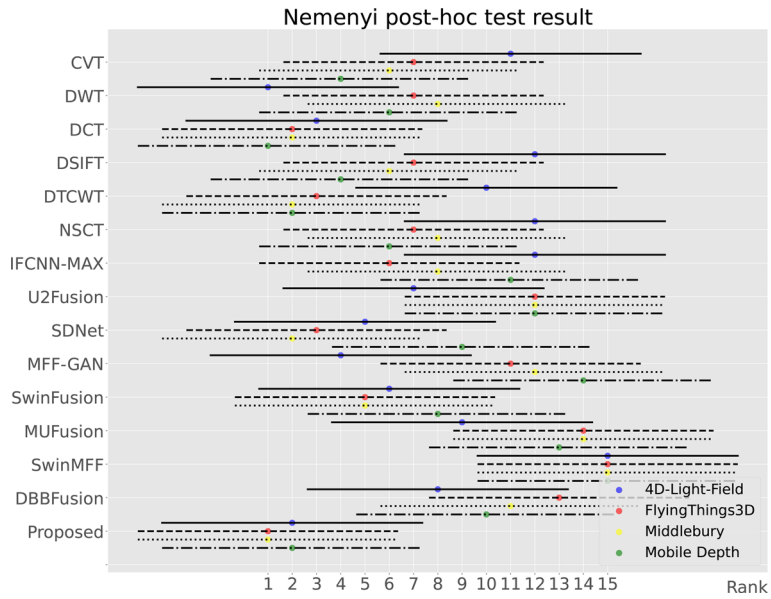

Fig. 11 Results of the Nemenyi post-hoc test

only a minimal increase in processing time between the smallest dataset (4D-Light-Field, 0.15 s) and the largest dataset (FlyingThings3D, 0.24 s), despite significant variations in image complexity and stack size. These results underscore StackMFF's potential for real-world applications. The dramatic reduction in processing time, combined with state-of-the-art fusion quality, makes our method particularly suitable for time-critical applications, including mobile photography [48], autonomous navigation [13], and microscopy imaging [3].

4.5.2 Generalization comparison

Table 4 presents a comprehensive ranking analysis of various image fusion methods across four diverse datasets: 4D-Light-Field, FlyingThings3D, Middlebury Stereo, and Mobile Depth. The rankings are derived from mean squared error (MSE) values, with lower values corresponding to higher rankings. This analysis provides valuable insights into the relative performance and generalization capabilities of different fusion methods across various datasets. Among deep learning methods, StackMFF consistently outperforms other approaches across multiple datasets. This superior performance can be attributed primarily to StackMFF's unique ability to process entire image stacks directly, effectively mitigating the impact of cumulative errors.

To further analyze the performance differences between methods, we conducted the Nemenyi post hoc test, as illustrated in Fig. 11. This test reveals significant contrasts between traditional methods and deep learning approaches in terms of performance consistency. Traditional methods exhibit smaller variations in their performance across datasets, demonstrating greater stability. In contrast, deep learning-based methods show more pronounced fluctuations, highlighting differences in their generalization capabilities. StackMFF stands out in this analysis because it yields relatively consistent performance across multiple datasets. It not only achieves superior quantitative and qualitative performance but also demonstrates strong generalization ability and robustness.

Table 5 Quantitative comparison of network performance under different sequence orders on the Mobile Depth [48] dataset

Group	*MSE* ↓	*MAE* ↓	*RMSE* ↓	*logRMS* ↓
1	0.0003	0.0062	0.0120	0.0078
2	0.0003	0.0061	0.0119	0.0078
3	0.0003	0.0063	0.0123	0.0080

Table 6 Quantitative evaluation results with varying numbers of input images on the MFI-WHU [8] dataset

Number of images	*MSE* ↓	*MAE* ↓	*RMSE* ↓	*logRMS* ↓
32	0.0005	0.0055	0.0083	0.0057
16	**0.0004**	**0.0055**	**0.0062**	**0.0040**
8	0.0005	0.0061	0.0089	0.0061
4	0.0005	0.0063	0.0092	0.0062
2	0.0009	0.0076	0.0113	0.0074

Table 7 Quantitative comparison of different image fusion methods on the Lytro [14] dataset

Method	$EI\uparrow$	$SF\uparrow$	$AVG\uparrow$	$MI\uparrow$	$EN\uparrow$	$VIF\uparrow$	$Q^P\uparrow$	$Q_W\uparrow$	$Q_{CB}\uparrow$
CVT	70.3233	19.2713	6.7897	15.0828	7.5414	1.1044	0.7966	0.9388	0.7276
DWT	70.7942	19.3342	6.8336	15.0872	7.5436	1.1114	0.2878	0.8977	0.6117
DCT	**76.6745**	**21.0541**	**7.3508**	15.1371	7.5685	**1.3511**	0.7825	0.9093	0.6624
DSIFT	70.9808	19.4194	6.8493	15.0688	7.5344	1.1381	0.2954	0.8977	0.6675
DTCWT	70.5666	19.3204	6.8134	15.0791	7.5396	1.1079	0.2925	0.8987	0.6234
NSCT	70.4289	19.2662	6.8027	15.0816	7.5408	1.1249	0.2928	0.9030	0.6174
IFCNN-MAX	70.9193	19.3793	6.8463	15.0722	7.5361	1.1322	0.2962	0.9013	0.5986
U2Fusion	59.8957	14.9334	5.6515	14.6153	7.3077	0.9882	0.2994	0.8909	0.5159
SDNet	60.3437	16.9252	5.8725	14.9332	7.4666	0.9281	0.3072	0.8934	0.5739
MFF-GAN	66.0601	18.4022	6.4089	14.8153	7.4076	1.0084	0.2840	0.8887	0.5399
SwinFusion	62.8130	16.6430	5.9862	15.0476	7.5238	1.0685	0.3117	0.9011	0.5745
MUFusion	75.4977	20.9230	7.2090	**15.2239**	**7.6119**	1.2753	0.7160	0.9089	0.6758
SwinMFF	72.4041	19.7954	6.9734	15.0826	7.5413	1.1810	**0.8222**	**0.9390**	**0.7543**
DBBFusion	48.1600	12.1484	4.5883	15.0663	7.5332	0.8874	0.5610	0.8391	0.6057
StackMFF	59.3624	15.0550	5.6395	15.0591	7.5295	1.0291	0.6763	0.9045	0.6513
StackMFF-E	75.4431	21.0205	7.2660	15.1032	7.5516	1.1613	0.6975	0.9068	0.6779

4.5.3 Sensitivity to sequential order

To investigate whether the proposed network is sensitive to the sequence order of images in the input image stack, we conducted comparative experiments using the Mobile Depth dataset. We chose the Mobile Depth dataset for two primary reasons. First, it comprises real-world captured images, making the findings more practically relevant. Second, the image sequences in this dataset inherently possess a natural order, with DoF positions changing continuously within each image stack. The experimental setup consisted of three groups:

- Group 1. A control group maintaining the original sequence order
- Group 2. A comparison group with a reversed sequence order
- Group 3. A comparison group with a completely randomized sequence order

The results in Table 5 demonstrate that our network exhibits remarkable stability across different input sequence arrangements. Notably, there is virtually no difference in performance between the original order (Group 1) and reversed order (Group 2), which is theoretically expected given the nature of the network's processing. While random shuffling (Group 3) shows a slight degradation in performance, the differences are negligible. This minimal variation in performance metrics across all testing configurations suggests that our network is insensitive to the input sequence order, making it highly robust and practical for real-world applications where the input order may not be strictly controlled.

4.5.4 Sensitivity to the number of images

In practical applications of multi-focus image fusion, the number of input images can vary significantly. To investigate our model's sensitivity to the input image count, we conduct experiments on the MFI-WHU dataset by synthesizing multi-focus image stacks containing 2, 4, 8, 16, and 32 images from 120 all-in-focus ground truth images via the proposed refocusing technology. As shown in Table 6, StackMFF exhibits robust performance across different num-

Fig. 12 Qualitative comparison with varying numbers of input images (2, 4, 8, 16, and 32) on the MFI-WHU [8] dataset. (The depth indicates the number of images in the image stack.)

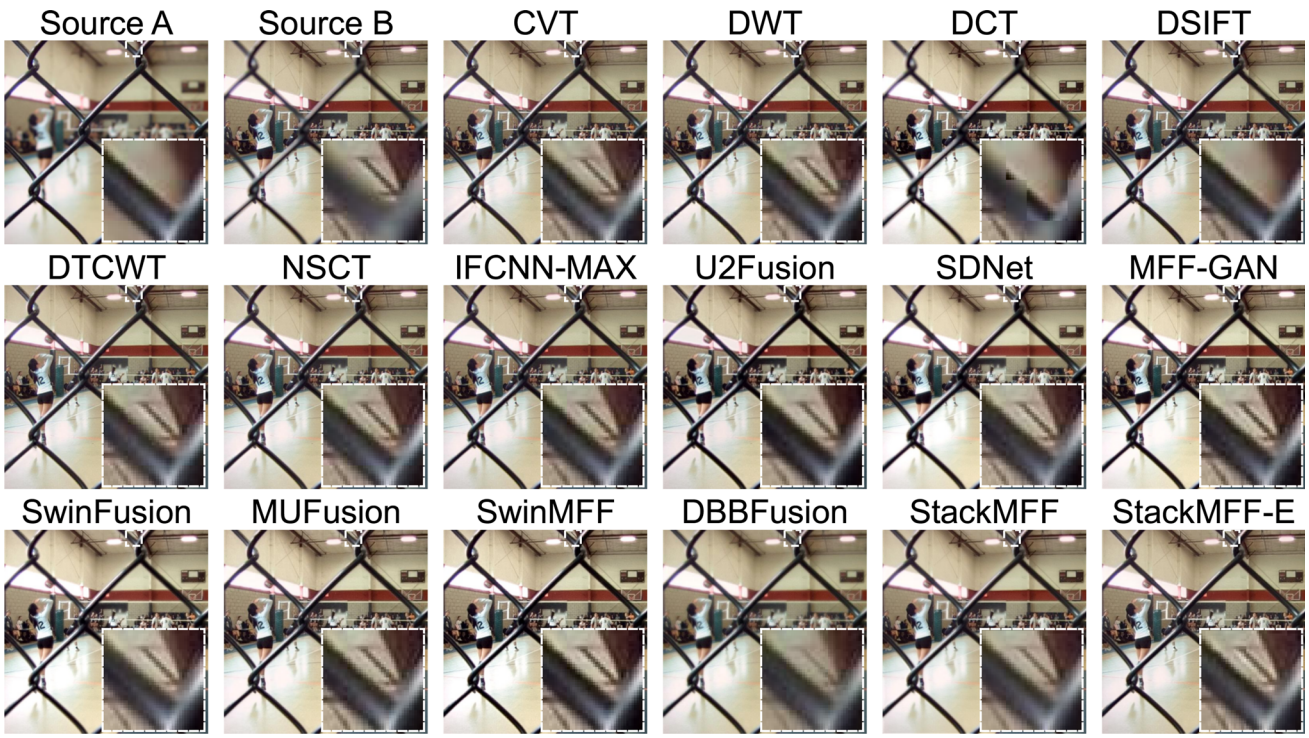

Fig. 13 Qualitative comparison of different image fusion methods on the Lytro [14] dataset

bers of input images, with performance metrics gradually improving as the number of input images increases from 2 to 16, followed by slight degradation when the number of input images increases to 32. This performance variation is likely attributed to our training setup, where the network was specifically trained on 16-image stacks. The qualitative results in Fig. 12 reveal that despite these subtle numerical differences in quantitative comparisons, our method consistently produces visually satisfactory results across all input stack sizes. This visual consistency demonstrates that while quantitative metrics show slight variations, these differences are practically imperceptible to human observers, highlighting the robustness of our approach in real-world scenarios where the number of available input images may vary.

4.5.5 Analysis of image pair fusion

Although our method is specifically designed for multi-focus image stack fusion, it can be directly applied to multi-focus image pair fusion without modification. To validate its effectiveness in this scenario, we conducted comprehensive comparisons (Table 7 and Fig. 13) with state-of-the-art methods on the widely used Lytro dataset. We employed nine widely used metrics for quantitative evaluation: image feature-based

Table 8 Quantitative results using different training datasets (StackMFF-NYU denotes the model trained on multi-focus image stacks synthesized from RGB-D images in NYU Depth V2 dataset [51])

Dataset	4D-Light-Field				FlyingThings3D			
Method	*MSE* ↓	*MAE* ↓	*RMSE* ↓	*logRMS* ↓	*MSE* ↓	*MAE* ↓	*RMSE* ↓	*logRMS* ↓
StackMFF-NYU	**0.0026**	0.0233	0.0322	0.0228	0.0011	0.0134	0.0225	0.0159
StackMFF	0.0048	**0.0221**	**0.0287**	**0.0203**	**0.0004**	**0.0061**	**0.0126**	**0.0090**
Dataset	**Middlebury**				**Mobile Depth**			
Method	*MSE* ↓	*MAE* ↓	*RMSE* ↓	*logRMS* ↓	*MSE* ↓	*MAE* ↓	*RMSE* ↓	*logRMS* ↓
StackMFF-NYU	0.0012	0.0146	0.0232	0.0157	0.0006	0.0110	0.0164	0.0115
StackMFF	**0.0005**	**0.0083**	**0.0151**	**0.0104**	**0.0003**	**0.0062**	**0.0120**	**0.0078**

Fig. 14 Qualitative comparison between StackMFF-NYU and StackMFF

metrics (EI, SF, AVG, Q^P), human perception-inspired metrics (VIF, Q_{CB}), information theory-based metrics (MI, EN), and image structure similarity-based metrics (Q_W). Higher values indicate better performance, with the best results shown in bold and the second-best results underlined.

For StackMFF, we evaluated two versions: direct image pair fusion (StackMFF) and an expanded version (StackMFF-E), where image pairs were artificially replicated to match the training stack size of 16 images. While direct pair fusion with StackMFF showed moderate performance, StackMFF-E achieved second-best results in SF and AVG metrics, demonstrating significant improvement through expansion. Qualitative analysis revealed that StackMFF and StackMFF-E outperformed quantitatively superior methods such as DCT in complex regions, where DCT exhibited notable fusion errors. Moreover, the visual results clearly demonstrate that StackMFF-E outperforms StackMFF, validating the effectiveness of our simple replication strategy. These experimental results confirm that our proposed StackMFF remains competent for multi-focus image pair fusion tasks.

4.5.6 Effects of the proposed image refocusing pipeline

The proposed image refocusing pipeline effectively addresses the training data scarcity issue by synthesizing multi-focus image stacks from all-in-focus images. Furthermore, it enables greater training data diversity by eliminating the dependency on specialized light field or depth cameras that typically capture limited scene types. To investigate the impact of dataset scale and scene diversity on fusion performance, we conduct additional experiments by training StackMFF on synthetic data generated from NYU Depth V2 dataset [51], which contains 1,449 RGB-D images exclusively from indoor scenes. As shown in Table 8, the model trained on this limited dataset (denoted as StackMFF-NYU) demonstrates performance degradation across multiple evaluation datasets. Specifically, on the FlyingThings3D dataset, StackMFF-NYU shows significant performance drops of 175% in MSE and 120% in MAE compared to the original model. Similar trends are observed on the Middlebury and Mobile Depth datasets, with performance gaps ranging from 36.7% to 140% across different metrics. In Fig. 14, we visualize the fusion results on some samples from the 4D-Light-Field dataset [45]. The qualitative comparison demonstrates that StackMFF-NYU yields significantly inferior fusion performance compared to StackMFF. These results strongly suggest that the scale and diversity of the training dataset play crucial roles in the model's generalization capability.

4.5.7 Ablation study

Table 9 presents an ablation study on the 4D-Light-Field dataset, demonstrating the effectiveness of our technical contributions: the extension of coordinate attention to 3D networks (3DCA) and the use of gradient loss L_{Gra} and spatial frequency loss L_{SF}. The baseline configuration employs only basic loss terms (L_{MAE} & L_{SSIM}). As shown in the table, both 3DCA and the additional loss terms individually improve the performance across all the metrics compared with the baseline. When combining these components, our model achieves the best results, with improvements in all evaluation metrics, validating the effectiveness of our proposed modules.

Table 9 Ablation study results for proposed StackMFF method on the 4D-Light-Field dataset [45]

Index	Baseline	3DCA	L_{Gra} & L_{SF}	$MSE \downarrow$	$MAE \downarrow$	$RMSE \downarrow$	$logRMS \downarrow$
1	√			0.0057	0.0294	0.0381	0.0272
2	√		√	0.0053	0.0276	0.0358	0.0256
3	√	√		0.0052	0.0271	0.0352	0.0252
4	√	√	√	**0.0048**	**0.0221**	**0.0287**	**0.0203**

5 Conclusion and discussion

This paper introduces StackMFF, an end-to-end 3D convolutional neural network capable of efficiently processing multi-focus image stacks of arbitrary sizes. Through its specially designed volumetric feature learning architecture, the framework simultaneously models the entire multi-focus stack, fundamentally eliminating the error accumulation and additional computational overhead inherent in iterative fusion approaches. The effectiveness of the framework stems from three key technical contributions. First, we leverage monocular depth estimation to propose a novel refocusing pipeline for generating large-scale realistic multi-focus stacks for end-to-end training. Second, we extend existing coordinate attention mechanisms and apply them to 3D CNNs for effective cross-layer feature aggregation. Third, we introduce a novel loss function group, incorporating additional gradient loss and spatial frequency loss, to facilitate better preservation of high-frequency details in the fused image. To facilitate comprehensive evaluation and future research, we establish a benchmark suite comprising 15 publicly available algorithms and various evaluation datasets. Extensive experiments demonstrate that StackMFF not only achieves superior fusion quality while avoiding common artifacts such as edge degradation, noise amplification, and blurring but also delivers the fastest processing speed among all existing methods. Future research directions include exploring joint depth estimation from defocus cues and extending the framework to broader applications such as multimodal, multiexposure, and medical image fusion.

Acknowledgements Thanks for support provided by Hainan Observation and Research Station of Ecological Environment and Fishery Resource in Yazhou Bay.

Author Contributions Xinzhe Xie: Conceptualization, Methodology, Software, Experiment, Writing - original draft, Writing - review & editing. Qingyan Jiang: Formal analysis, Investigation, Methodology, Resources, Validation. Dong Chen: Conceptualization, Formal analysis, Resources, Visualization. Buyu Guo: Conceptualization, Methodology, Investigation, Writing - review & editing, Funding acquisition. Peiliang Li: Conceptualization, Validation, Project administration, Supervision, Funding acquisition. Sangjun Zhou: Visualization, Validation, Data curation, Software.

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 StackMFF is available at https://github.com/Xinzhe99/StackMFF. Researchers are welcome to access the code, reproduce the results, and build upon this work for further research and applications.

Declarations

Ethical and informed consent for data used Not applicable.

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

References

1. Bacus JW, Grace LJ (1987) Optical microscope system for standardized cell measurements and analyses. Appl Opt 26(16):3280–3293
2. Liu M, Wang X, Zhang H (2018) Taxonomy of multi-focal nematode image stacks by a cnn based image fusion approach. Comput Methods Programs Biomed 156:209–215
3. Xie X, Guo B, Li P, Jiang Q (2024) Underwater three-dimensional microscope for marine benthic organism monitoring. In: OCEANS 2024-Singapore. IEEE, pp 1–4
4. Li J, Guo X, Lu G, Zhang B, Xu Y, Wu F, Zhang D (2020) Drpl: deep regression pair learning for multi-focus image fusion. IEEE Trans Image Process 29:4816–4831
5. Zhang Y, Liu Y, Sun P, Yan H, Zhao X, Zhang L (2020) Ifcnn: a general image fusion framework based on convolutional neural network. Inf Fusion 54:99–118
6. Zhang H, Ma J (2021) Sdnet: a versatile squeeze-and-decomposition network for real-time image fusion. Int J Comput Vision 129(10):2761–2785
7. Ma J, Tang L, Fan F, Huang J, Mei X, Ma Y (2022) Swinfusion: cross-domain long-range learning for general image fusion via swin transformer. IEEE/CAA J Autom Sin 9(7):1200–1217
8. Zhang H, Le Z, Shao Z, Xu H, Ma J (2021) Mff-gan: an unsupervised generative adversarial network with adaptive and gradient joint constraints for multi-focus image fusion. Inf Fusion 66:40–53
9. Xu H, Ma J, Jiang J, Guo X, Ling H (2020) U2fusion: a unified unsupervised image fusion network. IEEE Trans Pattern Anal Mach Intell 44(1):502–518
10. Xie X, Guo B, Li P, He S, Zhou S (2024) Swinmff: toward high-fidelity end-to-end multi-focus image fusion via swin transformer-based network. Vis Comput, 1–24
11. Cheng C, Xu T, Wu X-J (2023) Mufusion: a general unsupervised image fusion network based on memory unit. Inf Fusion 92:80–92
12. Zhang Z, Li H, Xu T, Wu X-J, Kittler J (2025) Ddbfusion: an unified image decomposition and fusion framework based on dual decomposition and bézier curves. Inf Fusion 114:102655
13. Li X, Li X, Tan H, Li J (2024) 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). IEEE, pp 3845–3849
14. Nejati M, Samavi S, Shirani S (2015) Multi-focus image fusion using dictionary-based sparse representation. Inf Fusion 25:72–84
15. Burt PJ, Adelson EH (1987) The laplacian pyramid as a compact image code. In: Readings in computer vision. Elsevier, ???, pp 671–679
16. Li H, Manjunath B, Mitra SK (1995) Multisensor image fusion using the wavelet transform. Graph Models Image Processing 57(3):235–245

17. Hill PR, Canagarajah CN, Bull DR et al (2002) Image fusion using complex wavelets. In: BMVC. Citeseer, pp 1–10
18. Yang B, Li S, Sun F (2007) Image fusion using nonsubsampled contourlet transform. In: Fourth International Conference on Image and Graphics (ICIG 2007). IEEE, pp 719–724
19. Yang B, Li S (2010) Multifocus image fusion and restoration with sparse representation. IEEE Trans Instrum Meas 59(4):884–892. https://doi.org/10.1109/TIM.2009.2026612
20. Xiao B, Xu B, Bi X, Li W (2020) Global-feature encoding u-net (geu-net) for multi-focus image fusion. IEEE Trans Image Process 30:163–175
21. Li S, Kwok JT, Wang Y (2001) Combination of images with diverse focuses using the spatial frequency. Inf Fusion 2(3):169–176. https://doi.org/10.1016/S1566-2535(01)00038-0
22. Li S, Kang X, Hu J (2013) Image fusion with guided filtering. IEEE Trans Image Process 22(7):2864–2875
23. Zhou Z, Li S, Wang B (2014) Multi-scale weighted gradient-based fusion for multi-focus images. Inf Fusion 20:60–72
24. Liu Y, Liu S, Wang Z (2015) Multi-focus image fusion with dense sift. Inf Fusion 23:139–155. https://doi.org/10.1016/j.inffus.2014.05.004
25. Duan Z, Luo X, Zhang T (2024) Combining transformers with cnn for multi-focus image fusion. Expert Syst Appl 235:121156
26. Liu Y, Chen X, Peng H, Wang Z (2017) Multi-focus image fusion with a deep convolutional neural network. Inf Fusion 36:191–207
27. Liu Y, Wang L, Cheng J, Chen X (2021) Multiscale feature interactive network for multifocus image fusion. IEEE Trans Instrum Meas 70:1–16
28. Ma B, Zhu Y, Yin X, Ban X, Huang H, Mukeshimana M (2021) Sesf-fuse: an unsupervised deep model for multi-focus image fusion. Neural Comput Appl 33:5793–5804
29. Ma J, Le Z, Tian X, Jiang J (2021) Smfuse: multi-focus image fusion via self-supervised mask-optimization. IEEE Trans Comput Imaging 7:309–320
30. Wang Y, Xu S, Liu J, Zhao Z, Zhang C, Zhang J (2021) Mfif-gan: a new generative adversarial network for multi-focus image fusion. Signal Process Image Commun 96:116295
31. Hu X, Jiang J, Liu X, Ma J (2023) Zmff: zero-shot multi-focus image fusion. Inf Fusion 92:127–138
32. Hu X, Jiang J, Wang C, Liu X, Ma J (2024) Incrementally adapting pretrained model using network prior for multi-focus image fusion. IEEE Trans Image Process
33. Li M, Pei R, Zheng T, Zhang Y, Fu W (2024) Fusiondiff: multi-focus image fusion using denoising diffusion probabilistic models. Expert Syst Appl 238:121664
34. Bando Y, Nishita T (2007) Towards digital refocusing from a single photograph. In: 15th Pacific conference on computer graphics and applications (PG'07). IEEE, pp 363–372
35. Zhang W, Cham W-K (2011) Single-image refocusing and defocusing. IEEE Trans Image Process 21(2):873–882
36. Busam B, Hog M, McDonagh S, Slabaugh G (2019) Sterefo: efficient image refocusing with stereo vision. In: Proceedings of the IEEE/CVF international conference on computer vision workshops, pp 0–0
37. Yin W, Zhang C, Chen H, Cai Z, Yu G, Wang K, Chen X, Shen C (2023) Metric3d: towards zero-shot metric 3d prediction from a single image. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 9043–9053
38. Yang L, Kang B, Huang Z, Xu X, Feng J, Zhao H (2024) Depth anything: unleashing the power of large-scale unlabeled data. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 10371–10381
39. Tran D, Bourdev L, Fergus R, Torresani L, Paluri M (2015) Learning spatiotemporal features with 3d convolutional networks. In: Proceedings of the IEEE international conference on computer vision, pp 4489–4497
40. Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O (2016) 3d u-net: learning dense volumetric segmentation from sparse annotation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19. Springer, pp 424–432
41. Murez Z, Van As T, Bartolozzi J, Sinha A, Badrinarayanan V, Rabinovich A (2020) Atlas: end-to-end 3d scene reconstruction from posed images. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16. Springer, pp 414–431
42. Hou Q, Zhou D, Feng J (2021) Coordinate attention for efficient mobile network design. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 13713–13722
43. Zhuo S, Sim T (2011) Defocus map estimation from a single image. Pattern Recogn 44(9):1852–1858
44. Xu S, Wei X, Zhang C, Liu J, Zhang J (2020) Mffw: a new dataset for multi-focus image fusion. arXiv preprint arXiv:2002.04780
45. Honauer K, Johannsen O, Kondermann D, Goldluecke B (2017) A dataset and evaluation methodology for depth estimation on 4d light fields. In: Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part III 13. Springer, pp 19–34
46. Mayer N, Ilg E, Hausser P, Fischer P, Cremers D, Dosovitskiy A, Brox T (2016) A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4040–4048
47. Scharstein D, Hirschmüller H, Kitajima Y, Krathwohl G, Nešić N, Wang X, Westling P (2014) High-resolution stereo datasets with subpixel-accurate ground truth. In: Pattern Recognition: 36th German Conference, GCPR 2014, Münster, Germany, September 2-5, 2014, Proceedings 36. Springer, pp 31–42
48. Suwajanakorn S, Hernandez C, Seitz SM (2015) Depth from focus with your mobile phone. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3497–3506
49. Guo L, Dai M, Zhu M (2012) Multifocus color image fusion based on quaternion curvelet transform. Opt Express 20(17):18846–18860
50. Haghighat MBA, Aghagolzadeh A, Seyedarabi H (2011) Multi-focus image fusion for visual sensor networks in dct domain. Comput Electr Eng 37(5):789–797. https://doi.org/10.1016/j.compeleceng.2011.04.016. Special Issue on Image Processing
51. Silberman N, Hoiem D, Kohli P, Fergus R (2012) Indoor segmentation and support inference from rgbd images. In: Computer Vision–ECCV 2012: 12th European conference on computer vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V 12. Springer, pp 746–760

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.

Jiang Qingyan received a Bachelor's degree in Industrial Design in 2012 and a Master's degree in Mechanical Engineering in 2015 from Qingdao University of Science and Technology, Qingdao, China. He is dedicated to mechanical structure design, with a particular focus on the design and development of marine instruments and equipment.

Dong Chen received the B.S. degree in communication engineering from the QingDao Agricultural University, QingDao, china, in 2012, and the M.S. degree in signal and information processing from the Ocean University of China, Qingdao, China, in 2015. He is currently an Middle Engineer with Ocean College, Zhejiang University, ZhouShan, China. His research interests include integration of underwater observation system and sensor R&D integration.

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 Reseaecher 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 oceangraphic information sensing.

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.