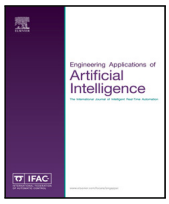
Engineering Applications of Artificial Intelligence 162 (2025) 112667

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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



Research paper

One-shot multi-focus image stack fusion via focal depth regression

Xinzhe Xie [a], Buyu Guo [b,c,*], Shuangyan He [a,b,d], Yanzhen Gu [a,b,d,**], Yanjun Li [b], Peiliang Li [a,b,d]

[a] State Key Laboratory of Ocean Sensing & Ocean College, Zhejiang University, Zhoushan, 316021, PR China
[b] Hainan Institute, Zhejiang University, Sanya, 572025, PR China
[c] Donghai Laboratory, Zhoushan, 316021, PR China
[d] Hainan Provincial Observatory of Ecological Environment and Fishery Resource in Yazhou Bay, Sanya, 572025, PR China

ARTICLE INFO

Keywords:
Multi-focus image fusion
Focus measure
Computational photography
Image stack processing

ABSTRACT

Multi-focus image fusion is a vital computational imaging technique for applications that require an extended depth of field, including medical imaging, microscopy, professional photography, and autonomous driving. While existing methods excel at fusing image pairs, they often suffer from error accumulation that leads to quality degradation, as well as computational inefficiency when applied to large image stacks. To address these challenges, we introduce a one-shot fusion framework that reframes image-stack fusion as a focal-plane depth regression problem. The framework comprises three key stages: intra-layer focus estimation, inter-layer focus estimation, and focus map regression. By employing a differentiable soft regression strategy and using depth maps as proxy supervisory signals, our method enables end-to-end training without requiring manual focus map annotations. Comprehensive experiments on five public datasets demonstrate that our approach achieves state-of-the-art performance with minimal computational overhead. The resulting efficiency and scalability make the proposed framework a compelling solution for real-time deployment in resource-constrained environments and lay the groundwork for broader practical adoption of multi-focus image fusion. The code is available at https://github.com/Xinzhe99/StackMFF-V2.

1. Introduction

Multi-focus image fusion (MFF) represents a crucial computational imaging technology with widespread applications across multiple domains. In the field of professional photography – particularly in landscape, macro, and product work – MFF has become an indispensable tool for overcoming the inherent depth-of-field limitations of conventional optical systems. By combining multiple images captured at different focal planes, the technology enables the production of fully focused composite images that not only extend the capabilities of traditional optical solutions but also reduce the need for specialized equipment such as tilt-shift lenses. Beyond professional imaging, MFF has demonstrated considerable value in scientific research and industrial applications, with notable uses in medical diagnosis (Pei et al., 2021; Pelapur et al., 2014), biological microscopy (Xie et al., 2024b; Liu et al., 2018), and underwater pipeline imaging and inspection. In these contexts, severe defocus blur typically arises in regions outside the optimal focal plane, thereby limiting image interpretability. To mitigate this challenge, researchers capture multiple images at different focal planes and fuse them using MFF algorithms, producing an all-in-focus composite image with enhanced clarity. More recently, the integration of MFF into autonomous driving systems has underscored its broader technological relevance, where it functions as a critical preprocessing step that enhances object detection performance through improved image clarity (Li et al., 2024e).

Although extensive research has been conducted on MFF, both traditional and learning-based methods exhibit inherent limitations when processing image stacks (Xie et al., 2025b). Early approaches (Burt and Adelson, 1985, 1987) employed simple yet effective pipelines, but their fusion quality is generally inferior to more recent techniques (Wang et al., 2024b; Li et al., 2024b). Contemporary traditional algorithms, such as the multi-dictionary linear sparse representation with region fusion model (Wang et al., 2024b), achieve performance comparable to advanced learning-based methods (Xie et al., 2024a; Jiang and Yu, 2025; Ouyang et al., 2025); however, their designs are tailored to two-image fusion and scale poorly to image stacks. Consequently, practitioners in domains such as medical imaging, biological

* Corresponding author at: Donghai Laboratory, Zhoushan, 316021, PR China.
** Corresponding author at: State Key Laboratory of Ocean Sensing & Ocean College, Zhejiang University, Zhoushan, 316021, PR China.
E-mail addresses: xiexinzhe@zju.edu.cn (X. Xie), guobuyu@donghailab.com (B. Guo), hesy@zju.edu.cn (S. He), guyanzhen@zju.edu.cn (Y. Gu), yanjun.li@zju.edu.cn (Y. Li), lipeiliang@zju.edu.cn (P. Li).

https://doi.org/10.1016/j.engappai.2025.112667
Received 14 April 2025; Received in revised form 25 August 2025; Accepted 5 October 2025
Available online 13 October 2025

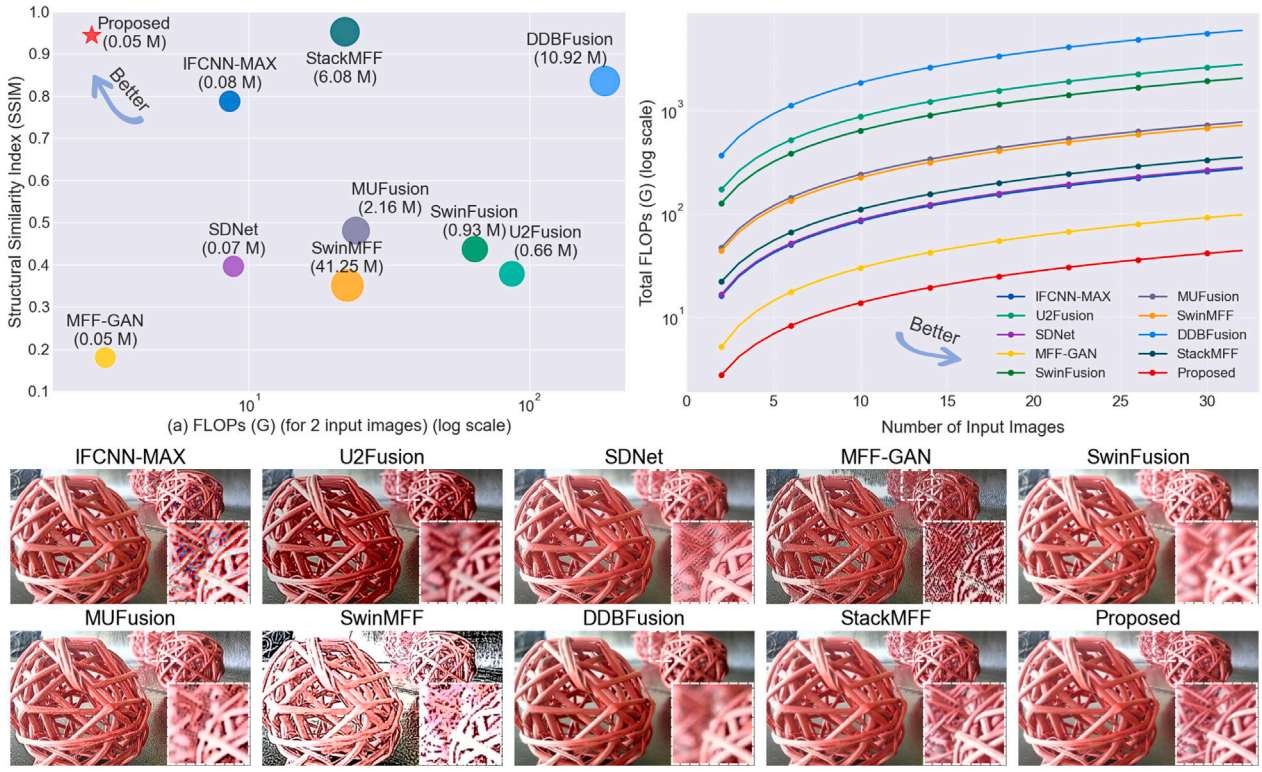


Fig. 1. Comparative analysis of different learning-based multi-focus image fusion methods on a 25-layer image stack. The top-left subplot presents a performance and efficiency comparison, where the bubble size in the scatter plot is scaled logarithmically to represent the number of model parameters. The top-right subplot illustrates the trend of increasing computational cost with the number of input layers. The bottom subplots provide a qualitative comparison of the fusion results for a 25-layer image stack. Note that only the proposed method and StackMFF (Xie et al., 2025b) perform fusion in a single pass; other methods rely on pairwise iterative approaches, which are prone to error accumulation and quality degradation.

research, and industrial inspection often rely on iterative pairwise fusion (e.g., sequential fusion), a process that complicates workflows and substantially increases computation time, limiting applicability in large-scale scenarios.

Recent advances in deep learning have facilitated the development of numerous learning-based MFF methods; however, most focus on image pairs, while stack-level processing remains underexplored. Sequential pairwise fusion is commonly assumed sufficient (Ma et al., 2021; Guo et al., 2019), yet our experiments reveal that iterative application of popular networks (e.g., IFCNN Zhang et al., 2020b, U2Fusion Xu et al., 2020a, MFF-GAN Zhang et al., 2021) leads to error accumulation and severe performance degradation (Fig. 1). The resulting artifacts – including edge distortions, noise, color shifts, and blurring – are particularly pronounced in end-to-end networks that directly infer pixel values. Moreover, learning-based approaches suffer from order-dependent bias during stack fusion, producing unpredictable results sensitive to fusion order. Although recent state-of-the-art models such as SwinMFF (Xie et al., 2024a) and DDBFusion (Zhang et al., 2025) achieve strong pairwise performance, their heavy computational overhead undermines scalability, thereby limiting their practicality in large-scale applications.

The recent method StackMFF (Xie et al., 2025b) mitigates degradation from iterative fusion by using an end-to-end 3D convolutional neural network to process the entire multi-focus image stack. However, its reliance on 3D convolutions leads to high computational complexity, hindering practical use in real-time scientific imaging where MFF is widely applied. Moreover, as an end-to-end network that directly infers pixel values rather than sampling from source images, StackMFF lacks interpretability and may compromise fusion fidelity. Such limitations are unacceptable in scientific imaging, where accuracy and transparency are essential. A more effective strategy for preserving quality would be to sample pixels from the source stack according to their measured focus levels at each spatial location – akin to traditional spatial-domain methods – rather than directly inferring pixel values. Nevertheless, developing such methods faces two obstacles: (1) obtaining annotated focus maps for training, as manual labeling is impractical; and (2) the discrete nature of focus maps, which hinders gradient-based optimization.

Inspired by the data generation process in StackMFF (Xie et al., 2025b), we observe that when scene depth maps are uniformly partitioned and layered, and defocus simulation is applied to synthesize multi-focus image stacks, the normalized depth values become numerically equivalent to normalized focus maps. This equivalence provides more fine-grained information due to its continuous nature. Furthermore, motivated by the soft regression strategy in stereo depth estimation networks (Kendall et al., 2017), which reformulates discrete disparity estimation into a continuous regression task to enable end-to-end training via differentiable cost aggregation, we propose to formulate focus map prediction as a continuous regression problem. This key insight enables the use of readily available depth maps as proxy supervision signals for training focus map prediction networks, thereby eliminating the need for explicit focus map annotations while maintaining high fusion fidelity through direct pixel sampling from the source image stack based on the predicted focus maps. Notably, our method does not assume linear changes in focal planes during inference; rather, it relies on the ordinal relationships among input layers to estimate a focal plane index for each pixel. This design allows our model to perform effectively even under non-linear or abrupt focus changes, as long as the sequential order of focal planes is preserved.

The main contributions of this work are threefold:

1. We propose a novel, ultra-lightweight learning-based framework capable of performing one-shot multi-focus image stack fusion.

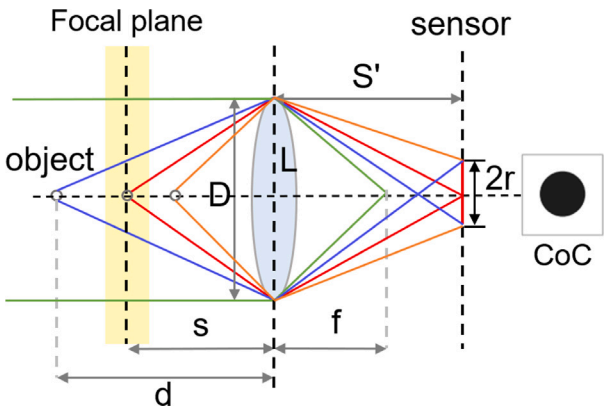


Fig. 2. Schematic of the optical imaging system illustrating defocus blur formation.

This framework incorporates a soft-regression strategy that computes the final focus map by pixel-wise weighting of layer indices with their corresponding focus probabilities, ensuring the differentiability and trainability of the entire pipeline.

2. We introduce an innovative supervision strategy that uniformly partitions scene depth maps and employs layered defocus simulation to synthesize multi-focus image stacks for training. This approach establishes a normalized numerical mapping between depth maps and focus maps, enabling higher-granularity depth maps to serve as proxy supervision signals for focus map regression. This facilitates more precise focus estimation while eliminating the need for explicit focus map annotations.
3. Comprehensive experiments on multiple datasets demonstrate that the proposed fusion framework achieves state-of-the-art performance with minimal model parameters and the shortest inference time. Furthermore, among learning-based methods, our approach exhibits the most efficient computational scaling with respect to input stack size.

To distinguish our approach from conventional image pair fusion, we define multi-focus image stack fusion as the task of synthesizing an all-in-focus image from more than two images captured at different focal planes of a scene.

The remainder of this paper is organized as follows. Section 2 introduces the defocus phenomenon and reviews related works on MFF methods. Section 3 presents the proposed fusion framework in detail. Section 4 reports experimental results and comparisons with state-of-the-art methods. Section 5 discusses the limitations of our approach and outlines future research directions. Finally, Section 6 concludes the paper by summarizing our contributions.

2. Related works

This section presents a literature review on multi-focus image fusion. We begin by analyzing the physical principles of defocus blur to inform our synthetic data generation. We then discuss the limitations of existing pair-based fusion methods when extended to complete image stacks. Finally, we review recent one-shot stack fusion techniques and highlight their relevance to our work.

2.1. Causes of defocus

The formation of defocus blur in optical systems can be analyzed using the thin-lens imaging model, as illustrated in Fig. 2. When an object point at distance d is not perfectly focused, it is projected onto the sensor as a blurred circle of confusion (CoC). The CoC radius r is given by:

$$r(d) = \frac{|d - s| f^2}{2d(s - f)F} \tag{1}$$

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

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

$$s_i = \min(d) + k_i(\max(d) - \min(d)) \tag{2}$$

Here, $k_i \in [0, 1]$ controls the focal plane position, with $k_i = 0$ corresponding to the nearest depth and $k_i = 1$ to the farthest depth. Uniform sampling of k_i values ensures comprehensive scene coverage.

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

$$K_i(x, y) = \frac{1}{2\pi\sigma_i^2(x, y)} \exp\left(-\frac{x^2 + y^2}{2\sigma_i^2(x, y)}\right) \tag{3}$$

where (x, y) denotes the pixel position. Based on the thin-lens model in Fig. 2 and assuming $\sigma_i = 2k \cdot r$ with $0 < k \leq 0.5$, we have:

$$\sigma_i = \frac{|d - s_i| k f^2}{d(s_i - f)F} \tag{4}$$

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

$$I_i = I^{gt} * K_i \tag{5}$$

where $*$ denotes the convolution operation.

Based on this theoretical analysis of defocus formation, multi-focus image stacks can be synthesized by partitioning scene depth maps into intervals and applying controlled blur according to the distance between each depth layer and its corresponding focal plane. This principled approach forms the basis of our training strategy, which is discussed in later sections.

2.2. Multi-focus image pair fusion

Owing to the prevalence of image pair examples in public datasets such as Lytro (Nejati et al., 2015), MFFW (Xu et al., 2020b), MFI-WHU (Zhang et al., 2021), and Real-MFF (Zhang et al., 2020a), coupled with the common assumption (see Fig. 3) that stack fusion can be realized through iterative pairwise fusion, most existing studies – including both traditional and learning-based approaches – have predominantly focused on two-image fusion.

Traditional MFF methods can be broadly categorized into transform domain-based and spatial domain-based approaches. Transform domain-based methods (Yang et al., 2007; Li et al., 2013, 2024c, 2025) exploit various transforms (e.g., wavelet, contourlet, curvelet, Fourier) to analyze source images across different frequency bands or orientations. These methods employ specific fusion rules on the transformed coefficients to preserve salient features from each input image in the fused output. While effective in handling complex scenarios and generally robust to misregistration, they encounter challenges in stack fusion due to computational overhead, particularly when processing a large number of input images.

Spatial domain-based methods (Guo et al., 2015; Liu et al., 2015; Nejati et al., 2015) operate directly on pixel- or patch-level features to assess focus levels and identify in-focus regions. Although these approaches offer intuitive focus measurement mechanisms, they often struggle with complex scenes and suffer from pronounced edge artifacts, which are further exacerbated in stack fusion scenarios. A central limitation lies in their inherently localized processing and fusion logic, typically designed to combine information from only two input images at a time. This localized nature, combined with the need for

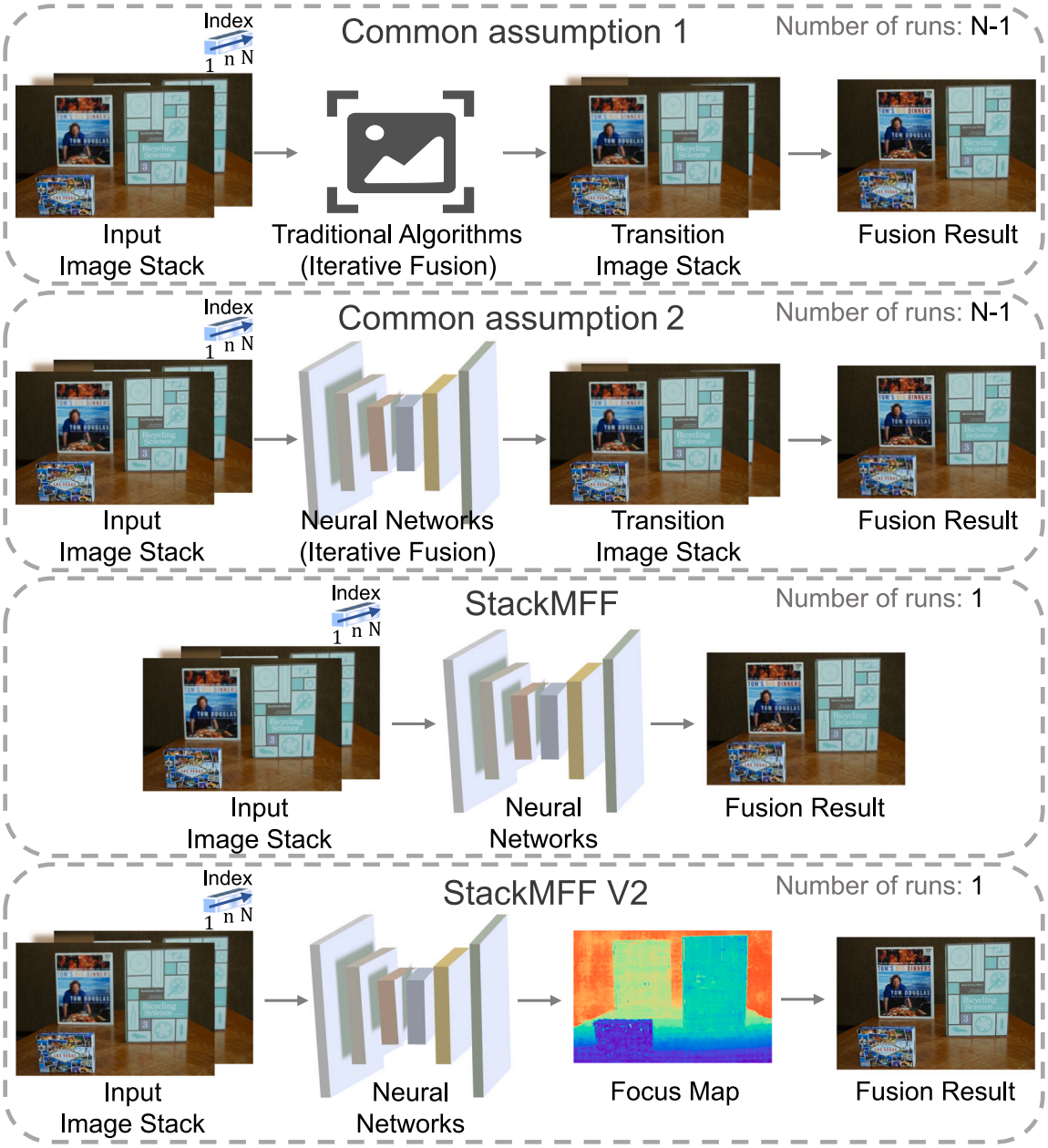


Fig. 3. Multi-focus image stack fusion paradigms: iterative vs. one-shot methods.

complex post-processing to mitigate artifacts and ensure spatial consistency, hinders the extension of these methods to direct, one-shot fusion of an entire image stack. Consequently, even recent state-of-the-art traditional methods (Li et al., 2024a), despite enhancements in pairwise fusion quality, remain constrained by two-image fusion pipelines, necessitating iterative processing for stack fusion.

Similar limitations apply to deep learning-based approaches, which typically require fixed input dimensions. When multi-focus image stacks are processed through iterative fusion, CNN-based methods (Zhang et al., 2020b; Zhang and Ma, 2021) suffer from quality degradation due to error accumulation. GAN-based methods (Zhang et al., 2021; Huang et al., 2020), although effective for image pair fusion, introduce cumulative noise artifacts in stack fusion. Other advanced architectures – including Transformer-based approaches (Xie et al., 2024a; Jiang and Yu, 2025), hybrid CNN-Transformer models (Ma et al., 2022; Duan et al., 2024), and diffusion-based methods (Li et al., 2024d) – while promising for image pairs, often incur substantial computational overhead when applied to stacks, limiting their practicality for real-time or efficient processing. Notably, ZMFF (Hu et al., 2023) provides an unsupervised approach for image pairs by leveraging deep image priors; however, its per-image optimization introduces significant computational bottlenecks, making it inefficient for multi-focus image stacks. Therefore, despite considerable progress in multi-focus image fusion, the efficient and high-quality processing of image stacks remains a major challenge for both traditional and deep learning-based methods.

2.3. Multi-focus image stack fusion

Compared to the extensive research on multi-focus image pair fusion, studies on multi-focus image stack fusion remain limited (e.g., StackMFF Xie et al., 2025b). Nevertheless, the increasing prevalence of

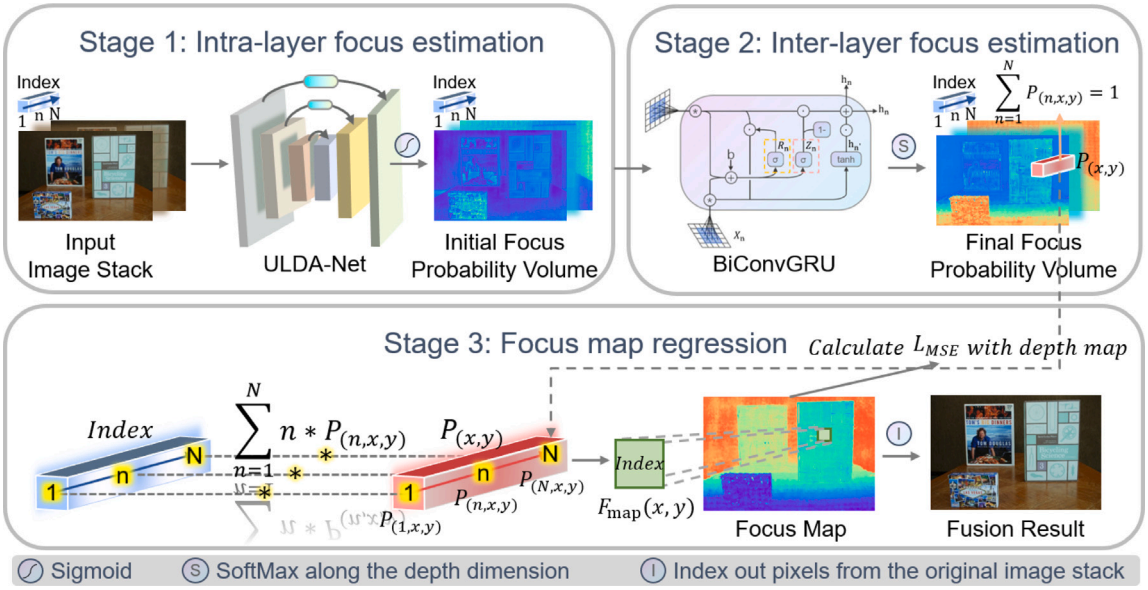


Fig. 4. Overview of the proposed multi-focus image stack fusion framework, StackMFF V2.

multi-focus image stacks underscores the growing need for dedicated fusion research targeting this data type, a requirement largely unmet by existing work.

In parallel developments, the field of depth from focus (DFF) has investigated image stack processing, where all-in-focus images frequently appear as byproducts or intermediate outputs (Maximov et al., 2020; Wang et al., 2021; Si et al., 2023). However, these studies primarily emphasize depth estimation accuracy rather than fusion quality, often resulting in suboptimal fusion outputs that may contain artifacts or inconsistencies.

Motivated by the differentiable soft-regression mechanism (Kendall et al., 2017) in stereo depth estimation and the innovative image stack modeling strategy (Yang et al., 2022) in DFF, we propose a novel approach enabling one-shot fusion of entire image stacks, while simultaneously addressing computational efficiency, fusion quality, and model interpretability.

3. Methods

In this section, we introduce our proposed one-shot multi-focus image stack fusion framework. We first provide an overview of the overall architecture, then describe the functionality of each stage, followed by the proposed proxy supervision strategy, and finally present the loss function used for training.

3.1. Method overview

Fig. 4 illustrates the architecture of the proposed multi-focus image stack fusion framework, StackMFF V2, which comprises three primary stages: intra-layer focus estimation, inter-layer focus estimation, and focus map regression.

Given a multi-focus image stack with continuously varying focal planes, we first employ an ultra-lightweight defocus-level adaptive intra-layer focus estimation network, termed ULDA-Net, to independently extract features from each layer and estimate the initial in-focus probability for each pixel within its respective layer. This process produces an initial focus probability volume, where each layer is represented independently.

Next, a single-layer bidirectional convolutional gated recurrent neural network (BiConvGRU) is applied to model cross-layer dependencies within this initial focus probability volume, enabling information exchange and aggregation along the depth dimension. A Softmax operation is then applied along the depth dimension to normalize the focus probabilities, resulting in the final focus probability volume. In this volume, each element represents the normalized in-focus probability of a pixel at a specific spatial location across different layers, with probabilities summing to 1 along the depth dimension.

Finally, by weighting these probabilities with their corresponding layer indices, a focus map is generated, where each value indicates the layer containing the sharpest pixel at the corresponding spatial location. The fused image is then produced by directly sampling pixels from the original image stack based on this focus map.

The proposed fusion framework assigns distinct tasks and objectives to each stage, enhancing the network's interpretability and facilitating both understanding and optimization. Moreover, this modular design accommodates input image stacks of varying sizes, providing greater flexibility.

3.2. Intra-layer focus estimation

We propose an ultra-lightweight defocus-level adaptive intra-layer focus estimation network (ULDA-Net) for efficient pixel-wise focus probability estimation within individual layers, as illustrated in Fig. 5.

ULDA-Net is designed for efficient and robust intra-layer focus estimation. It begins by extracting low-level features that capture defocus blur patterns while suppressing noise. To accommodate multi-scale defocus characteristics, the network employs parallel depthwise separable convolutions with varying kernel sizes, inspired by Inception architectures (Yu et al., 2024), enabling efficient capture of blur patterns guided by the Point Spread Function (PSF). Focus-aware feature learning is further enhanced through dual attention modules (Ruan et al., 2022), where the spatial attention block (SAB) identifies sharp, in-focus regions, and the channel attention block (CAB) adaptively emphasizes features most relevant for focus discrimination. ULDA-Net adopts a lightweight U-Net-like encoder–decoder architecture with skip connections, facilitating hierarchical feature learning with minimal computational cost. This compact design contains only 0.03M parameters – approximately 1% of MobileNetV3-Small (Howard et al., 2019) – while maintaining robust focus estimation performance.

For an input stack $X \in \mathbb{R}^{B\times N\times H\times W}$ (B: batch size, N: number of layers, H: height, W: width), the network reshapes it to $X \in$

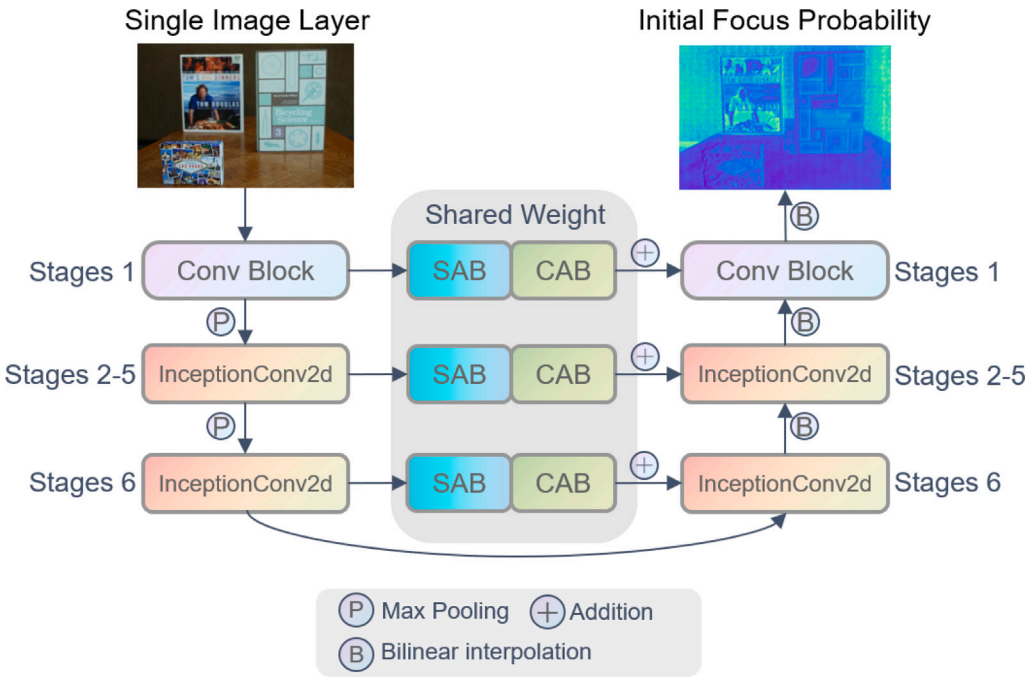


Fig. 5. The proposed ultra-lightweight defocus-level adaptive intra-layer focus estimation network (ULDA-Net).

$\mathbb{R}^{BN\times1\times H\times W}$ for parallel processing. The encoder transforms the input to $X \in \mathbb{R}^{BN\times64\times H/32\times W/32}$, followed by decoder upsampling to $X \in \mathbb{R}^{BN\times8\times H\times W}$. After reshaping to $X \in \mathbb{R}^{B\times N\times8\times H\times W}$, a Sigmoid operation generates the initial focus probability volume, representing the relative in-focus probability for each pixel within its respective focal plane.

Notably, the intra-layer focus estimation processes each focal plane independently, disregarding inter-layer dependencies. This layer-wise strategy enables accurate identification of sharp regions within each plane and provides a reliable foundation for subsequent inter-layer focus modeling across the entire image stack.

3.3. Inter-layer focus estimation

Existing focus estimation approaches often employ a winner-takes-all strategy to directly convert per-layer outputs into a single focus map (Ali et al., 2023). While efficient, this approach ignores a crucial point: after normalizing the outputs of each layer, the network values represent relative rather than absolute focus probabilities. Specifically, the in-focus probabilities within each layer are normalized independently, meaning that a higher probability in one layer does not guarantee sharper focus than a lower probability in another layer. For example, a pixel with 65% in-focus probability in Layer 1 and 60% in Layer 2 cannot be directly compared, as each probability is relative to its layer. Therefore, inter-layer modeling is essential for accurate focus estimation across the entire image stack.

To exploit the sequential and continuous nature of focus variation, we employ a single-layer Bidirectional Convolutional Gated Recurrent Unit (BiConvGRU) to refine the initial focus probability volume. This recurrent module efficiently aggregates information along the depth dimension, enabling accurate inter-layer comparison while maintaining a minimal parameter count (0.015M). The bidirectional design allows information flow from both forward and backward directions of the stack, aligning with the refinement task, as the initial probability volume already provides reliable estimates. Compared to heavier alternatives such as 3D CNNs (Xie et al., 2025b) or Transformers (Kang et al., 2023), this lightweight design achieves effective inter-layer modeling with low computational overhead.

The initial focus probability volume $X \in \mathbb{R}^{B\times D\times8\times H\times W}$ is upsampled to $X \in \mathbb{R}^{B\times D\times16\times H\times W}$ to match the 16-channel hidden state used in the BiConvGRU, allowing rich feature representation of the initial probability volume. A subsequent max-pooling operation reduces the 16 channels back to a single channel, yielding $X \in \mathbb{R}^{B\times D\times1\times H\times W}$. This single-channel representation facilitates the subsequent weighting operations. Finally, redundant dimensions are removed to obtain the final inter-layer-refined probability volume with shape $X \in \mathbb{R}^{B\times D\times H\times W}$.

After inter-layer modeling, the focus values at each pixel location encode inter-layer-aware confidence scores rather than normalized probabilities. Applying a Softmax operation along the depth dimension, as formulated in Eq. (6), converts these scores into relative in-focus probabilities across layers, ensuring that the sum of probabilities equals 1.

$$P(n,x,y) = \frac{e^{c(n,x,y)}}{\sum_{n=1}^{N} e^{c(n,x,y)}} \tag{6}$$

where $c(n,x,y)$ represents the focus confidence score at position (x,y) in the nth layer after BiConvGRU modeling, and $P(n,x,y)$ denotes the normalized probability that the pixel at position (x,y) is in focus in the nth layer. Through BiConvGRU modeling and Softmax operations, we transform the implication of each element within the initial probability volume. This transformation shifts from representing the in-focus probability relative to the current layer to expressing the in-focus probability relative to all pixels at the same horizontal spatial position across different layers.

3.4. Focus map regression

Focus map estimation can be treated as a discrete mapping problem, in which each pixel is assigned a specific focal plane index. However, such discrete assignments are non-differentiable, making them incompatible with end-to-end training via backpropagation. To overcome this limitation, we adopt a soft-regression approach that converts discrete plane selection into a continuous weighted summation. Specifically, after intra-layer and inter-layer focus estimation, the elements in the final probability volume precisely indicate the normalized in-focus probabilities of pixels across different depth planes at each spatial

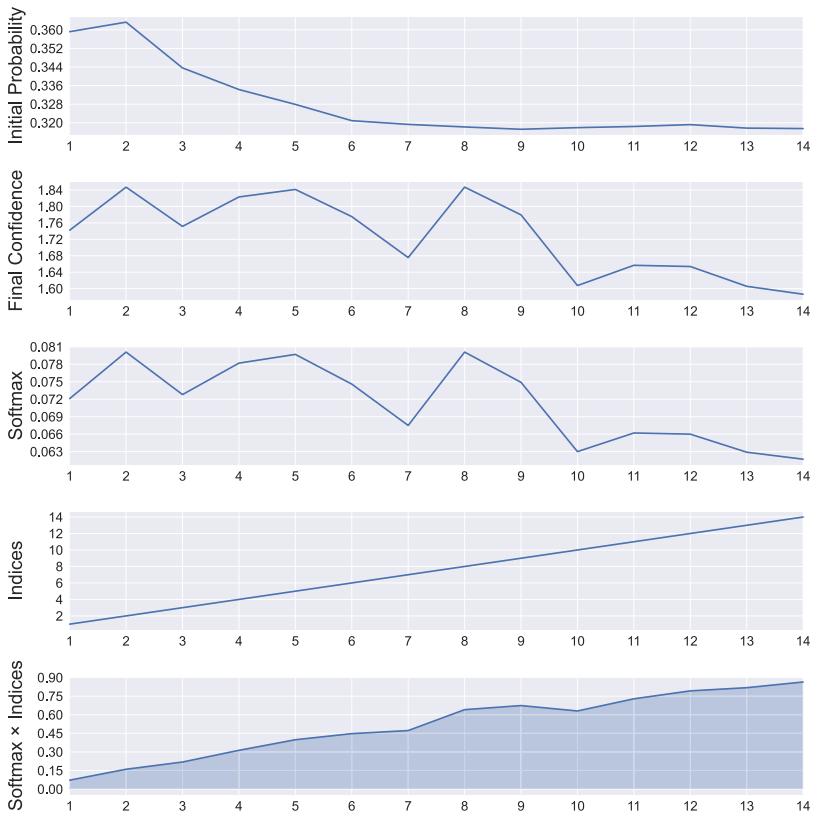


Fig. 6. Focal plane depth regression visualized for a single spatial location: from initial probabilities to the final focal plane index.

location, with the probabilities summing to 1. The focus map F_{map} is then computed through soft regression, formulated as the expectation of layer indices with respect to their corresponding probability distributions:

$$F_{map}(x,y) = \mathbb{E}[n|P(n,x,y)] = \sum_{n=1}^{N} n \cdot P(n,x,y) \tag{7}$$

Here, $\mathbb{E}[n|P(n,x,y)]$ represents the expected layer index n under the probability distribution $P(n,x,y)$ at spatial location (x,y). This differentiable formulation enables end-to-end training while preserving the ability to generate continuous focus values. Subsequently, this probabilistic regression process is applied to every pixel in the fused image, with each resulting focus map value determining which layer to sample from when constructing the all-in-focus image:

$$I_{fused}(x,y) = I_{stack}(F_{map}(x,y),x,y) \tag{8}$$

Fig. 6 illustrates the step-by-step process of computing the focus map value for a pixel at a specific spatial location through five sequential line graphs. The topmost graph shows the initial in-focus probabilities across different layers obtained from ULDA-Net. The second graph illustrates the focus confidence after inter-layer feature modeling. Note that sigmoid normalization is omitted at this stage due to the subsequent softmax operation. The third graph depicts the normalized probability distribution obtained by applying softmax along the depth dimension, as formulated in Eq. (6). The fourth graph shows the layer indices, while the fifth graph represents their element-wise product with the softmax probabilities. The final focus map value, visualized as the shaded area in the figure, is computed by aggregating these products across all layers, as defined in Eq. (7). It is important to note that although the above explanation focuses on a single spatial location for clarity, the actual implementation employs fully vectorized operations across the entire probability volume. Specifically, given the final probability volume, our method uses matrix operations to compute focus map values for all spatial locations simultaneously. This vectorized implementation enables efficient end-to-end training and inference in a single forward pass.

3.5. Depth maps as focus map proxies

The proposed framework requires focus maps as supervision signals, where values ranging from 1 to N denote the corresponding focal plane indices in the source image stack. As observed in Section 2.1, when scene depth maps are uniformly partitioned and layered, defocus simulation can be used to synthesize multi-focus image stacks, and the normalized depth values become numerically equivalent to the normalized focus maps. This key observation allows us to utilize readily available depth maps as proxy supervision signals for training focus map prediction networks.

To illustrate this proxy supervision mechanism, we present the data synthesis process in Fig. 7. During the synthesis of training data, we first strictly partition the scene depth range into N focal planes according to Eq. (2), with each focal plane corresponding to a specific depth interval, as shown in the top row of Fig. 7, where yellow-highlighted regions indicate the depth range of the current focal plane. The object-space distance s_i for each focal plane is determined by uniformly sampling the scene's depth range, creating a series of focal planes that systematically cover the entire scene depth. This uniform partitioning strategy ensures that the focus map values directly correspond to the normalized depth values, establishing a clear mapping between depth information and focal plane indices.

Notably, depth maps provide more fine-grained supervision signals, as multiple depth levels can correspond to a single depth-of-field region in the focus maps. We leverage this advantage by using Depth Anything V2 (Yang et al., 2024) to generate high-quality depth maps for arbitrary images, thereby avoiding the need for manual focus map annotations. These depth maps also facilitate the synthesis of multi-focus image

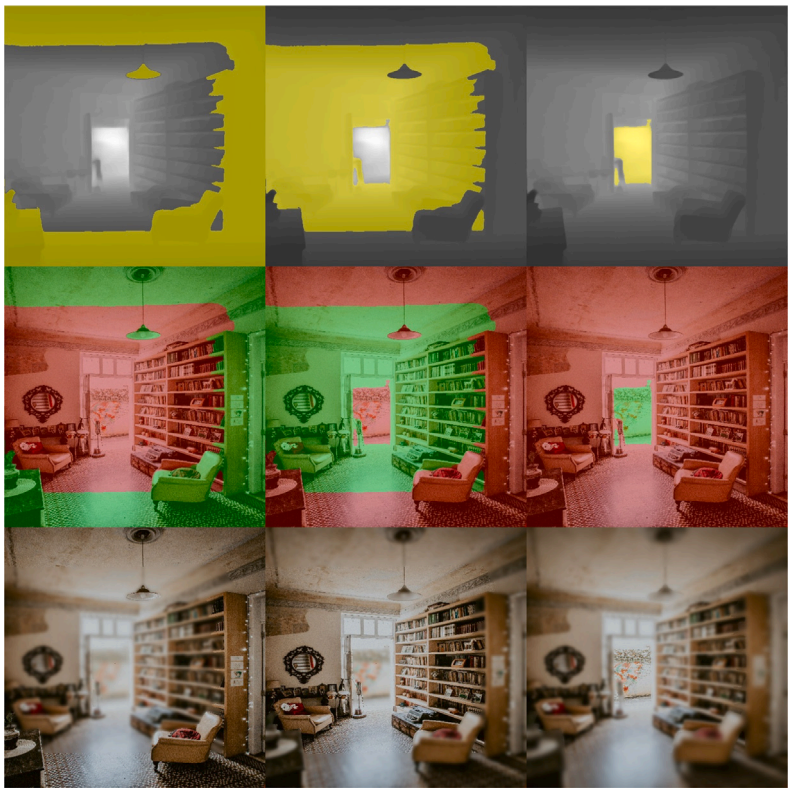


Fig. 7. Visualization of the training data synthesis process (simplified example with three focal planes). For each scene, we illustrate: (top row) depth maps estimated by Depth Anything V2 (Yang et al., 2024) and their corresponding focal plane partitions, with yellow highlights indicating the depth range for the current focal plane; (middle row) visualization of in-focus (green) and out-of-focus (red) regions across different focal planes; (bottom row) synthesized multi-focus image stack frames.

Algorithm 1 : Synthesis procedure of the multi-focus image stack dataset for training.

Require:
I: Input all-in-focus image
D: Input depth map
N: Number of focal planes
Ensure:
$\{I_k\}_{k=1}^{N}$: Synthesized multi-focus image stack
$\{K_i\}_{i=1}^{N} \leftarrow \{2i+1\}_{i=1}^{N}$ ▷ Generate N Gaussian kernels of increasing sizes
$\{B_i\}_{i=1}^{N} \leftarrow \{GaussianBlur(I, K_i)\}_{i=1}^{N}$ ▷ Generate N blurred versions of the input image
$D_{norm} \leftarrow Normalize(D, 0, 255)$ ▷ Normalize depth map to [0, 255]
$R \leftarrow \{r_i\}_{i=0}^{N} \leftarrow LinSpace(0, 255, N+1)$ ▷ Define N+1 reference points for depth quantization
$Q \leftarrow Quantize(D_{norm}, R)$ ▷ Assign each pixel to one of N depth regions
for $k \leftarrow 1$ to N **do** ▷ Generate the image corresponding to focal plane k
    $I_k \leftarrow ZerosLike(I)$ ▷ Initialize the output image
    for $i \leftarrow 1$ to N **do** ▷ Iterate over all depth regions
        $M_i \leftarrow (Q == i)$ ▷ Mask pixels belonging to depth region i
        $j \leftarrow |i - k|$ ▷ Determine blur level according to distance from current focal plane
        $I_k[M_i] \leftarrow B_j[M_i]$ ▷ Assign blurred pixels to the output image
    end for
end for
return $\{I_k\}_{k=1}^{N}$ ▷ Return the synthesized multi-focus image stack

stacks through the layered blur simulation strategy proposed in the previous work, StackMFF (Xie et al., 2025b). We summarize this synthesis procedure in Algorithm 1, which efficiently generates the multi-focus image stack by combining progressive Gaussian blur with depth-based blending. As illustrated in the middle row of Fig. 7, we can accurately identify which regions should be sharp (shown in green) or blurred (shown in red) for each focal plane based on the depth information. Note that this visualization is simplified for clarity; in practice, continuous transitions of blur levels are created according to the distance from each focal plane, with objects gradually becoming more blurred as their distances from the focal plane increase. The final synthesized image stack frames are displayed in the bottom row of Fig. 7.

3.6. Loss function

In our multi-focus image stack synthesis process, each focal plane corresponds to a specific depth range, as defined in Eq. (2). Since the focal planes are uniformly partitioned according to the scene's depth range, a natural mapping exists between depth maps and focus maps. By normalizing both the focus map (originally ranging from 1 to N, where N is the number of focal planes) and the depth map to the range [0, 1], we establish a direct correspondence between these two representations. This normalization allows for direct comparison between the predicted focus values and the normalized depth values during training.

Based on this principle, we employ the Mean Squared Error (MSE) as the loss function to quantify the discrepancy between the normalized predicted focus map F_{map} and the normalized depth map D_{map}:

$$L_{MSE} = \frac{1}{H \times W} \sum_{x=1}^{H} \sum_{y=1}^{W} (F_{map}(x, y) - D_{map}(x, y))^2 \tag{9}$$

Here, H and W denote the height and width of the image, respectively. Both F_{map} and D_{map} are linearly normalized to the interval from 0 to 1, with 0 corresponding to the nearest focal plane or depth

and 1 to the farthest. This depth-based supervision strategy not only eliminates the need for manual focus map annotations but also provides fine-grained supervision signals for training the network.

4. Experiments

In this section, we present a comprehensive experimental evaluation to demonstrate the effectiveness and practical potential of our proposed method. We first describe the implementation details, including the training strategy, hyperparameter settings, and the datasets used for both training and evaluation. Next, we provide both qualitative and quantitative comparisons against several state-of-the-art methods. Finally, we conduct a detailed analysis of key aspects, including model efficiency, statistical significance, comparison with commercial software, and performance in specific corner cases.

4.1. Implementation details

4.1.1. Training strategy

In the previous work, StackMFF (Xie et al., 2025b), only fixed-size image stacks were used, and the training was conducted exclusively on synthetic data derived from the Open Images Dataset V7 (Kuznetsova et al., 2020). This approach may limit generalization because it can cause the network to learn patterns specific to those particular dimensions and dataset, thereby restricting its ability to effectively fuse image stacks of other sizes or types. Therefore, in this work, we propose two corresponding optimization strategies to overcome these limitations:

- **Multi-dataset mixed training.** During training, we randomly sample images from five distinct datasets and combine them into batches, which are then fed into the network. This strategy allows our model to learn patterns common across multiple datasets, thereby improving its performance and generalization.
- **Multi-depth mixed training.** To increase the model's robustness when processing image stacks with varying numbers of layers, we implemented a multi-depth mixed training strategy. Specifically, we used five different image stack configurations during training, consisting of 8, 12, 16, 20, and 24 layers, respectively. Stacks with the same number of layers were batched together, and we alternated between stacks of different depths to reduce the model's sensitivity to input stack depth.

Our image stack fusion method was implemented in PyTorch 2.4.1. To enhance the model's generalization and robustness to real-world complexities, we diversified the training data synthesis process by introducing varying PSF (Point Spread Function) widths, spatially varying defocus, and imaging perturbations, including noise, exposure variations, and color changes. All images were resized to 384 × 384 pixels, and standard data augmentation techniques, such as random horizontal and vertical flipping, were applied.

For model training, we employed a high-performance computing platform with dual NVIDIA A6000 GPUs and an Intel(R) Xeon(R) Platinum 8375C CPU. The AdamW optimizer was used with a batch size of 12. The initial learning rate was set to 1×10^{-3}, with an exponential decay factor of 0.9. The training took approximately 24 h over 30 epochs.

4.1.2. Datasets for training

All of our training data are sourced from publicly available datasets:

- **DUTS** (Wang et al., 2017): A salient object segmentation dataset containing 10,553 training and 5019 test images from diverse environments.
- **NYU Depth V2** (Silberman et al., 2012): An RGB-D dataset comprising 1449 densely labeled pairs across 464 indoor scenes. The high-quality, post-processed depth maps enable direct supervision to generate 1349 training and 100 test stacks.
- **DIODE** (Vasiljevic et al., 2019): An RGB-D dataset spanning indoor and outdoor scenes. We use Depth Anything V2 (Yang et al., 2024) estimates instead of the original LiDAR-based depth maps to avoid missing values, creating 25,458 training and 771 validation stacks.
- **Cityscapes** (Cordts et al., 2016): An urban scene understanding dataset covering 50 cities, yielding 2975 training and 500 validation stacks.
- **ADE** (Zhou et al., 2019): A comprehensive scene understanding dataset spanning indoor and outdoor environments, from which we synthesize 20,210 training and 2000 validation stacks.

For all datasets except NYU Depth V2 (Silberman et al., 2012), scene depth maps are obtained using Depth Anything V2 (Yang et al., 2024). Following the layered blur simulation approach proposed in StackMFF (Xie et al., 2025b), multi-focus image stacks are synthesized from single all-in-focus images as training inputs. The depth maps estimated by Depth Anything V2 serve as proxy supervision signals for end-to-end focus map regression training, as discussed in Section 3.5.

4.1.3. Datasets for evaluation

Traditional MFF test datasets, including Lytro (Nejati et al., 2015), MFFW (Xu et al., 2020b), MFI-WHU (Zhang et al., 2021), and Real-MFF (Zhang et al., 2020a), are primarily designed for image pair fusion evaluation and are not suitable for assessing multi-focus image stack fusion performance. Therefore, we turn to the field of depth from focus, which provides multi-focus image stack datasets that meet our requirements. We extensively evaluate our method on three well-established depth-from-focus benchmarks: FlyingThings3D (Mayer et al., 2016), Middlebury Stereo (Scharstein et al., 2014), and Mobile Depth (Suwajanakorn et al., 2015). Furthermore, following the same synthesis approach used for our training data, we construct two supplementary test sets: one from NYU Depth V2 (Silberman et al., 2012) using its provided depth maps, and another from Road-MF (Li et al., 2024e) using depth maps estimated by Depth Anything V2. Detailed specifications of these test datasets are provided below:

- **Mobile Depth** (Suwajanakorn et al., 2015): 15 real-world stacks (12–33 images per stack) with resolutions ranging from 774 × 518 to 640 × 360 pixels, featuring natural defocus effects.
- **Middlebury Stereo** (Scharstein et al., 2014): 13 real image stacks (15 images per stack) with resolutions from 741 × 497 to 347 × 277 pixels, employing disparity-based synthetic defocus.
- **FlyingThings3D** (Mayer et al., 2016): 120 synthetic stacks (15 images per stack) at 960 × 540 pixels with disparity-rendered defocus effects.
- **NYU Depth V2** (Silberman et al., 2012): 100 stacks (8–24 images per stack) at 620 × 460 pixels, simulated using depth-based layered defocus from real scenes.
- **Road-MF** (Li et al., 2024e): 80 stacks (8–24 images per stack) at 800 × 600 pixels, generated through depth-based layered defocus from real environments.

We conduct comprehensive quantitative and qualitative evaluations across all five datasets to thoroughly assess our method's performance.

4.1.4. Methods for comparison and evaluation metrics

We benchmark the proposed fusion framework against fourteen state-of-the-art methods, including five traditional approaches: CVT (Guo et al., 2012), DWT (Li et al., 1995), DCT (Haghighat et al., 2011), DTCWT (Hill et al., 2002), and NSCT (Yang et al., 2007). We also compare against nine learning-based methods: IFCNN (Zhang et al., 2020b), U2Fusion (Xu et al., 2020a), SDNet (Zhang and Ma, 2021), MFF-GAN (Zhang et al., 2021), SwinFusion (Ma et al., 2022), MUFusion (Cheng et al., 2023), SwinMFF (Xie et al., 2024a), DDBFusion (Zhang et al., 2025), and StackMFF (Xie et al., 2025b). Notably, only StackMFF (Xie et al., 2025b) and the proposed method support

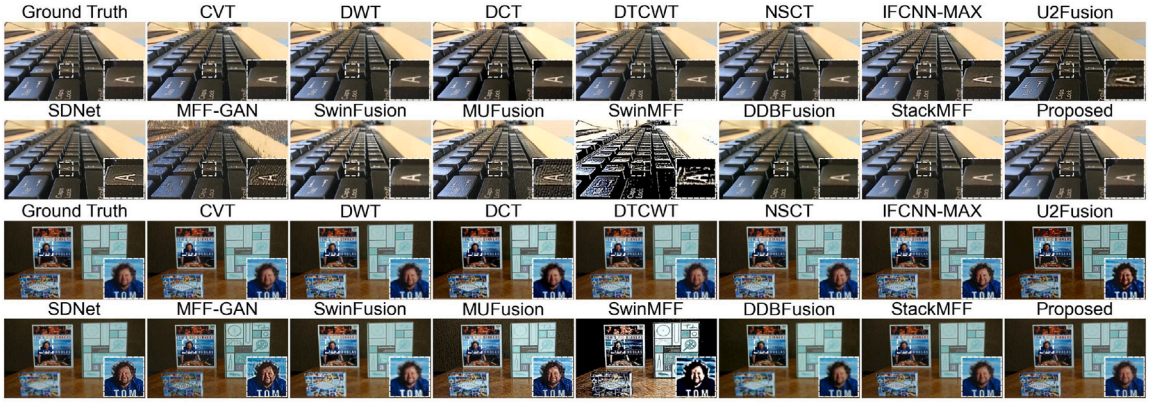


Fig. 8. Comparison of fusion results produced by different methods on the Mobile Depth dataset (Suwajanakorn et al., 2015).

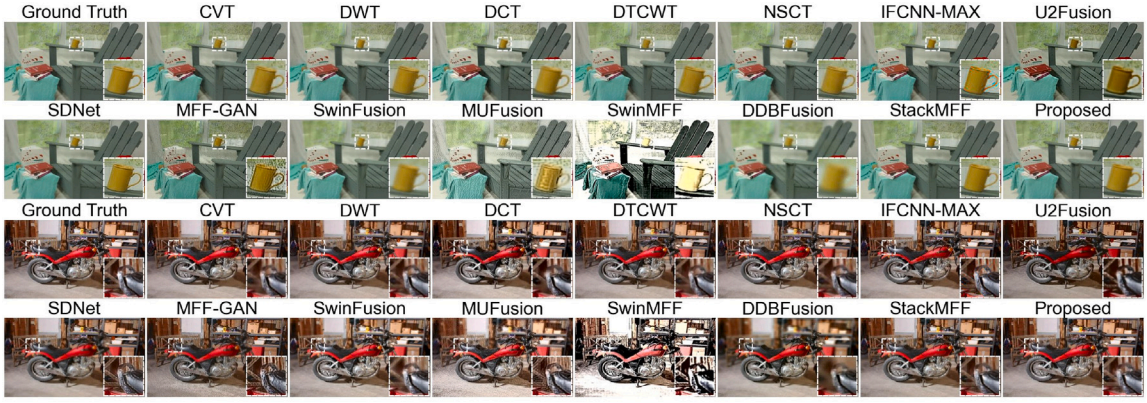


Fig. 9. Comparison of fusion results produced by different methods on the Middlebury Stereo dataset (Scharstein et al., 2014).

one-shot stack fusion, whereas all other methods require iterative pairwise fusion. For traditional methods, we use the default parameters provided in their public implementations. For learning-based methods, we employ their publicly available pre-trained models. For methods that do not support direct stack fusion, we perform iterative pairwise fusion following the default order of the image stack, requiring $N-1$ fusion operations for an N-layer stack.

To quantitatively assess fusion quality, we employ standard full-reference metrics, namely the Structural Similarity Index (SSIM) and the Peak Signal-to-Noise Ratio (PSNR). These metrics evaluate structural similarity and pixel-level fidelity relative to the ground-truth all-in-focus images available in our test datasets.

4.2. Qualitative comparison

To comprehensively evaluate the effectiveness of our proposed method, we conduct extensive experiments across five diverse datasets. These datasets encompass a wide range of scenarios, from controlled indoor scenes to challenging outdoor environments.

Qualitative results, presented in Figs. 8–12, demonstrate several key findings. Conventional methods exhibit consistent and robust fusion performance across diverse scenarios, effectively preserving structural details and overall image quality. In contrast, most learning-based approaches suffer from various artifacts, including blur, noise, and structural distortions. Notably, among all evaluated methods, only StackMFF (Xie et al., 2025b) and the proposed approach achieve superior fusion quality that consistently matches or even exceeds the performance of traditional methods.

The effectiveness of our approach is further demonstrated through focus map visualization, as shown in Figs. 13 and 14. While the ground-truth depth maps are represented with pixel values ranging from 0 to 255, our focus maps are discretized into N levels (where N denotes the number of focal planes), which is considerably smaller than 255. Although this difference in value ranges may affect visualization, it does not compromise estimation accuracy. Our network accurately estimates the depth of field across diverse scenes, effectively capturing transitions between focal planes. Such precise focus estimation directly contributes to the high-quality fusion results observed in indoor environments (NYU Depth V2), outdoor traffic scenes (Road-MF), and synthetic datasets (FlyingThings3D), highlighting the strong generalization capability of our method across different domains.

4.3. Quantitative comparison

Table 1 presents quantitative comparisons against state-of-the-art methods across five diverse datasets. Experimental results indicate that the proposed approach generally outperforms existing methods on most datasets. Specifically, on the Mobile Depth dataset, our method achieves the highest SSIM/PSNR of 0.9508 / 35.1017, outperforming both StackMFF and DTCWT. On the Middlebury and FlyingThings3D datasets, the proposed approach also demonstrates consistent performance advantages, achieving 0.9444/32.1810 and 0.9508/32.7506, respectively. While traditional methods such as DTCWT and NSCT exhibit strong performance on the Road-MF dataset, our method maintains competitive results (0.9808/36.0976) and significantly outperforms recent learning-based approaches. For the NYU Depth V2 dataset, the proposed method achieves the highest SSIM/PSNR of 0.9918/42.8195; however, since a large portion of this dataset was used during training, these results should be interpreted as reference values rather than strict indicators of generalization.

Overall, the experimental evaluation demonstrates that our approach exhibits strong generalization capability across diverse

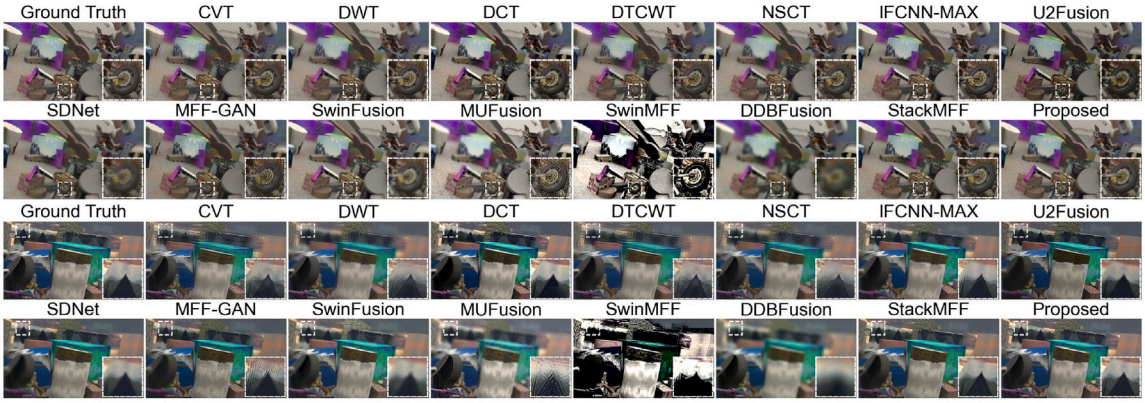


Fig. 10. Comparison of fusion results produced by different methods on the FlyingThings3D dataset (Mayer et al., 2016).

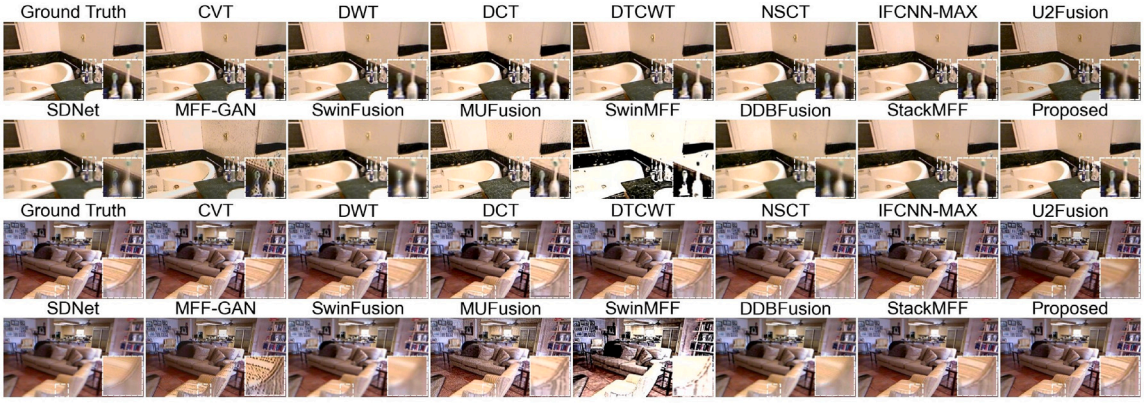


Fig. 11. Comparison of fusion results produced by different methods on the NYU Depth V2 dataset (Silberman et al., 2012).

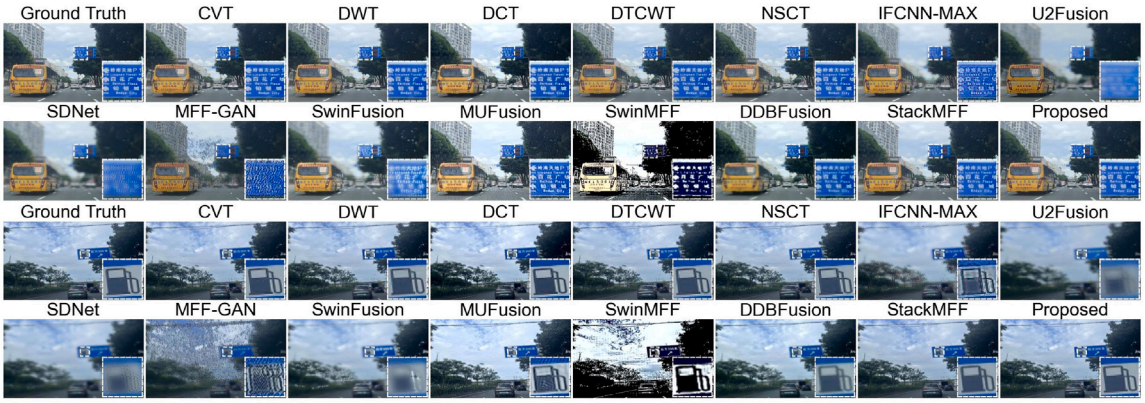


Fig. 12. Comparison of fusion results produced by different methods on the Road-MF dataset (Li et al., 2024e).

scenarios – including real-world images, depth-based synthetic stacks, and artificially defocused samples – highlighting its potential for robust multi-focus image stack fusion.

4.4. More analysis

4.4.1. Model efficiency comparison

Table 2 presents the runtime comparison of different fusion methods, evaluated on both CPU and GPU platforms. The runtime of MFF algorithms is crucial for real-time applications. The proposed method achieves the shortest processing time across all tested datasets. Notably, our method significantly outperforms other GPU-accelerated approaches, including StackMFF – the second-fastest method – by achieving nearly 100% faster execution. In contrast, traditional CPU-based methods, while often effective in terms of fusion quality, require substantially longer runtimes, ranging from several seconds to several hundred seconds.

Table 3 summarizes the model size and computational cost (evaluated on 256×256 images) of learning-based fusion methods. The proposed method has the smallest model size (0.05M parameters, matching MFF-GAN) and requires only 2.75 × (N-1) GFLOPs for N-layer image

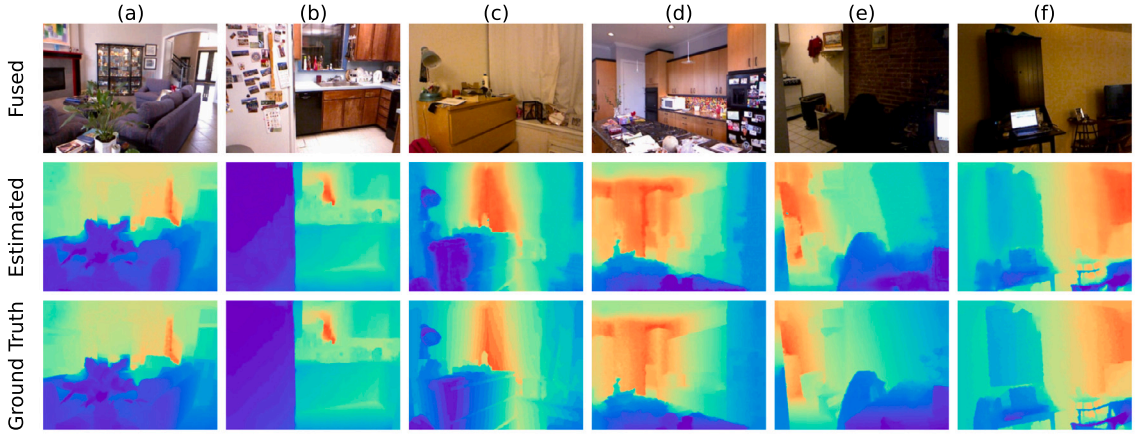


Fig. 13. Qualitative comparison between network-estimated focus maps and ground-truth depth maps on the NYU Depth V2 dataset (Silberman et al., 2012). Both maps are normalized and visualized with pseudo-coloring.

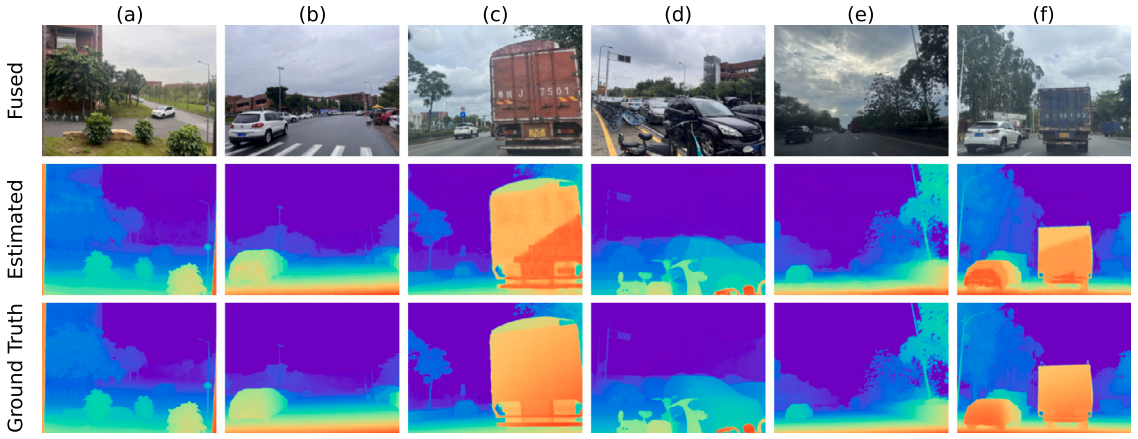


Fig. 14. Qualitative comparison between network-estimated focus maps and ground-truth depth maps on the Road-MF dataset (Li et al., 2024e). Both maps are normalized and visualized with pseudo-coloring.

Table 1
Quantitative evaluation of fusion performance across five test datasets.

Datasets	Mobile Depth		Middlebury		FlyingThings3D		Road-MF		NYU Depth V2	
Method	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑
CVT (Guo et al., 2012)	0.9368	32.6158	0.8893	29.3426	0.9157	30.0917	0.9777	36.0578	0.9717	38.8186
DWT (Li et al., 1995)	0.9340	32.1651	0.8850	29.1761	0.9123	30.0074	0.9309	30.3456	0.9594	35.8626
DCT (Haghighat et al., 2011)	0.4720	17.2719	0.4520	13.9972	0.4603	15.0949	0.4856	16.9598	0.4802	14.2216
DTCWT (Hill et al., 2002)	0.9412	32.7641	0.8938	29.3763	0.9203	30.1512	**0.9826**	36.7138	0.9743	39.1475
NSCT (Yang et al., 2007)	0.9340	32.1651	0.8850	29.1761	0.9123	30.0074	0.9813	**37.0137**	0.9707	38.7653
IFCNN (Zhang et al., 2020b)	0.7882	24.9863	0.9014	29.2064	0.9236	31.3069	0.8952	27.6907	0.9364	34.3915
U2Fusion (Xu et al., 2020a)	0.3788	10.0482	0.3980	10.1318	0.4242	11.4382	0.3811	10.8764	0.3869	10.7027
SDNet (Zhang and Ma, 2021)	0.3961	12.1659	0.4399	14.0048	0.4457	14.5929	0.4144	13.0182	0.4212	14.2688
MFF-GAN (Zhang et al., 2021)	0.1797	7.1264	0.2962	10.1180	0.3006	11.9173	0.2559	9.3437	0.2755	10.5829
SwinFusion (Ma et al., 2022)	0.4381	12.4597	0.4254	13.4794	0.4313	14.1286	0.3945	11.9315	0.4114	13.6265
MUFusion (Cheng et al., 2023)	0.4819	18.7311	0.5809	19.7779	0.4762	19.8073	0.6821	19.6156	0.5891	21.0372
SwinMFF (Xie et al., 2024a)	0.3511	10.8676	0.4215	11.8564	0.3238	12.2809	0.4795	13.2869	0.3983	13.1620
DDBFusion (Zhang et al., 2025)	0.8365	26.3713	0.7181	23.7650	0.6984	23.0223	0.8065	24.4036	0.8786	28.7440
StackMFF (Xie et al., 2025b)	**0.9536**	32.6798	0.9284	31.0764	0.9483	32.5062	0.9692	33.0138	0.9812	37.7552
Proposed	0.9508	**35.1017**	**0.9444**	**32.1810**	**0.9508**	**32.7506**	0.9808	36.0976	**0.9918**	**42.8195**
Enhancement (%)	−0.29%	+7.13%	+1.72%	+3.55%	+0.26%	+0.75%	−0.18%	−2.48%	+1.08%	+9.38%

stacks, while supporting one-shot fusion. StackMFF also allows one-shot fusion but with substantially higher computational cost. Despite its low parameter count, MFF-GAN is often impractical due to the high noise it introduces. The reduction in both parameters and computational cost demonstrates the efficiency of the proposed method and its suitability for resource-constrained applications, such as embedded systems.

Table 2
Comparison of computational efficiency (seconds) across various methods and datasets.

Method	Device	Mobile Depth	Middlebury	FlyingThings3D	Road-MF	NYU Depth V2
CVT (Guo et al., 2012)	CPU	48.00	31.37	37.87	78.14	56.20
DWT (Li et al., 1995)	CPU	5.34	8.62	6.75	4.62	3.45
DCT (Haghighat et al., 2011)	CPU	4.97	3.30	6.04	4.97	9.16
DTCWT (Hill et al., 2002)	CPU	11.44	9.40	14.70	12.82	10.06
NSCT (Yang et al., 2007)	CPU	231.84	165.13	133.84	217.03	152.05
IFCNN (Zhang et al., 2020b)	GPU	0.55	0.50	0.78	0.55	0.44
U2Fusion (Xu et al., 2020a)	CPU	41.04	35.90	45.10	104.96	41.93
SDNet (Zhang and Ma, 2021)	CPU	9.68	5.26	14.04	8.18	6.64
MFF-GAN (Zhang et al., 2021)	CPU	6.40	8.88	10.06	12.67	6.98
SwinFusion (Ma et al., 2022)	GPU	28.21	19.53	32.33	30.19	67.52
MUFusion (Cheng et al., 2023)	GPU	40.40	21.98	55.02	45.79	31.02
SwinMFF (Xie et al., 2024a)	GPU	27.97	18.23	34.05	55.04	24.47
DDBFusion (Zhang et al., 2025)	GPU	33.89	30.06	41.98	35.57	17.35
StackMFF (Xie et al., 2025b)	GPU	0.22	0.19	0.24	0.22	0.20
Proposed	GPU	**0.14**	**0.08**	**0.11**	**0.11**	**0.07**
Reduction (%)	–	36.36%	57.89%	54.17%	50.00%	65.00%

Table 3
Comparison of model size (M), computational cost for fusing N-layer image stacks (FLOPs, in G), and one-shot fusion capability among different learning-based approaches.

Method	Model size (M)	FLOPs (G)	One-shot fusion
IFCNN (Zhang et al., 2020b)	0.08	8.54 × (N-1)	×
U2Fusion (Xu et al., 2020a)	0.66	86.4 × (N-1)	×
SDNet (Zhang and Ma, 2021)	0.07	8.81 × (N-1)	×
MFF-GAN (Zhang et al., 2021)	**0.05**	3.08 × (N-1)	×
SwinFusion (Ma et al., 2022)	0.93	63.73 × (N-1)	×
MUFusion (Cheng et al., 2023)	2.16	24.07 × (N-1)	×
SwinMFF (Xie et al., 2024a)	41.25	22.38 × (N-1)	×
DDBFusion (Zhang et al., 2025)	10.92	184.9 × (N-1)	×
StackMFF (Xie et al., 2025b)	6.08	21.98 × (N-1)	✓
Proposed	**0.05**	**2.75 × (N-1)**	✓
Reduction (%)	–	10.71%	–

Table 4
Ranking of different methods based on quantitative evaluation metrics across four test datasets.

Datasets	Mobile Depth			Middlebury			FlyingThings3D			Road-MF			Overall
Method	SSIM	PSNR	Avg.	SSIM	PSNR	Avg.	SSIM	PSNR	Avg.	SSIM	PSNR	Avg.	Avg.
MFF-GAN (Zhang et al., 2021)	15	15	15.0	15	15	15.0	15	14	14.5	15	15	15.0	14.9
U2Fusion (Xu et al., 2020a)	13	14	13.5	14	14	14.0	13	15	14.0	14	14	14.0	13.9
SwinMFF (Xie et al., 2024a)	14	13	13.5	13	13	13.0	14	13	13.5	11	11	11.0	12.8
SwinFusion (Ma et al., 2022)	11	11	11.0	12	12	12.0	12	12	12.0	13	13	13.0	12.0
SDNet (Zhang and Ma, 2021)	12	12	12.0	11	10	10.5	11	11	11.0	12	12	12.0	11.4
DCT (Haghighat et al., 2011)	10	10	10.0	10	11	10.5	10	10	10.0	10	10	10.0	10.1
MUFusion (Cheng et al., 2023)	9	9	9.0	9	9	9.0	9	9	9.0	9	9	9.0	9.0
DDBFusion (Zhang et al., 2025)	7	7	7.0	8	8	8.0	8	8	8.0	8	8	8.0	7.8
DWT (Li et al., 1995)	5	5	5.0	6	6	6.0	6	6	6.0	6	6	6.0	5.8
IFCNN-MAX (Zhang et al., 2020b)	8	8	8.0	3	5	4.0	3	3	3.0	7	7	7.0	5.5
NSCT (Yang et al., 2007)	5	5	5.0	6	6	6.0	6	6	6.0	2	1	1.5	4.6
CVT (Guo et al., 2012)	4	4	4.0	5	4	4.5	5	5	5.0	4	4	4.0	4.4
DTCWT (Hill et al., 2002)	3	2	2.5	4	3	3.5	4	4	4.0	1	2	1.5	2.9
StackMFF (Xie et al., 2025b)	1	3	2.0	2	2	2.0	2	2	2.0	5	5	5.0	2.8
Proposed	2	1	1.5	1	1	1.0	1	1	1.0	3	3	3.0	1.6

The top-right subplot of Fig. 1 compares the computational cost (FLOPs) of different fusion methods with respect to the number of input layers. All methods show a linear increase, but with different rates: DDBFusion rises by 184.93 GFLOPs per layer (5732.83 GFLOPs for 32 layers), U2Fusion by 86.40 GFLOPs (2678.40 GFLOPs), SwinFusion by 63.73 GFLOPs (1975.36 GFLOPs), and StackMFF from 21.98 to 351.60 GFLOPs. The proposed method increases minimally at 1.375 GFLOPs per layer, scaling from 2.75 to 44.01 GFLOPs, demonstrating a clear advantage in processing large-scale image stacks.

4.4.2. Statistical significance analysis

Table 4 presents the ranking results of different MFF algorithms across four benchmark datasets based on SSIM and PSNR. To maintain experimental rigor and avoid potential training bias, the NYU Depth V2 dataset was excluded from this analysis. The statistical significance of performance differences was evaluated using the Nemenyi post-hoc test, as shown in Fig. 15, where non-overlapping confidence intervals indicate statistically significant differences between methods.

The statistical analysis yields several key findings: (1) The proposed method achieves the top overall ranking of 1.6, showing consistently strong performance across three datasets while remaining competitive on Road-MF. (2) Several traditional fusion methods perform consistently better than many recent deep learning-based approaches, likely due to their established image processing principles, stable fusion quality independent of training data, and robust performance across diverse scenarios. (3) Cross-dataset ranking highlights the varying generalization capabilities of different architectures across diverse image characteristics and fusion scenarios. (4) The results indicate a notable

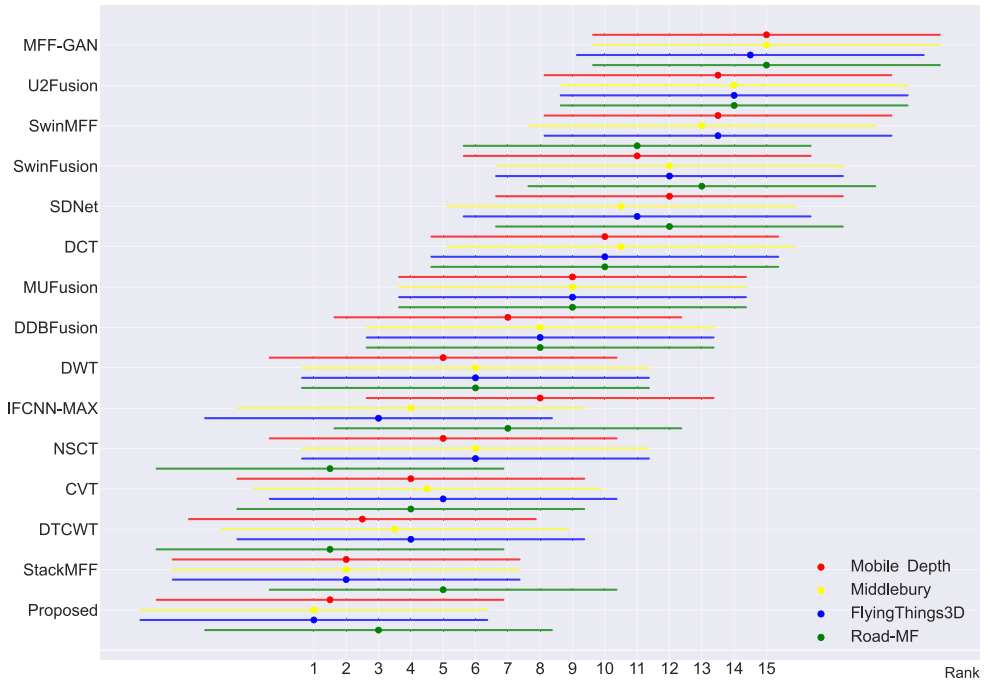


Fig. 15. Results of the Nemenyi post hoc test for multiple method comparisons.

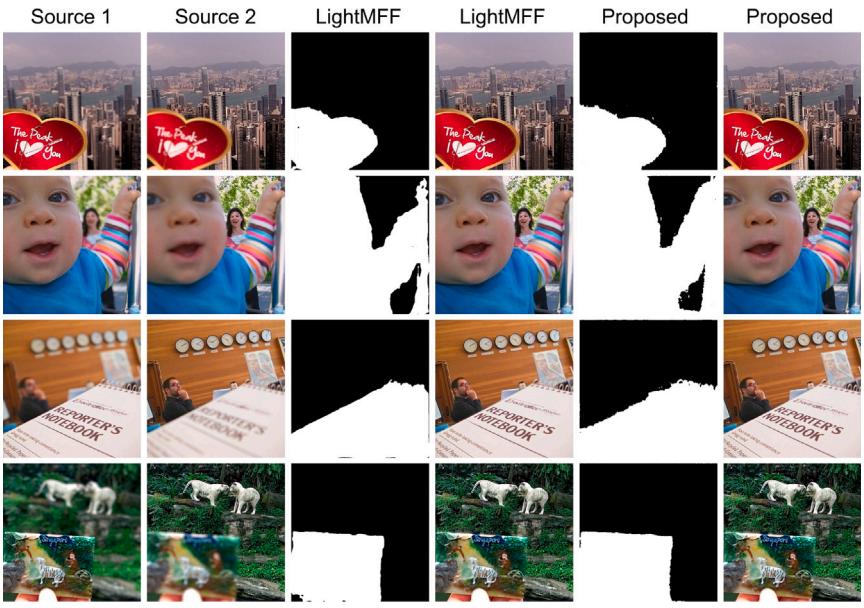


Fig. 16. Qualitative comparison of the proposed method with LightMFF (Xie et al., 2025a), a network specialized for image pair fusion.

performance gap between existing deep learning architectures and the proposed framework, supporting the effectiveness of the novel design.

4.4.3. Performance of small stack fusion

Although the proposed method was specifically designed and trained for medium-to-large multi-focus image stacks, it can also be applied to fusion tasks for image pairs and smaller stacks without requiring additional modifications or post-processing. To evaluate its performance on image pair fusion, we conducted a qualitative comparison of the focus maps and fused images generated by our method on the Lytro dataset (Nejati et al., 2015) with those produced by LightMFF (Xie et al., 2025a), one of the most recent and advanced hybrid approaches that integrates classical operators with deep learning. As shown in Fig. 16, although our network was not trained on image pairs, the inferred focus maps (binarized here for visualization) demonstrate its robustness and strong performance on the image pair fusion task.

Furthermore, we demonstrate the effectiveness of our proposed method in fusing small image stacks. As shown in Fig. 17, we selected image stacks from the Lytro (Nejati et al., 2015) dataset, each containing only three images (representing foreground, midground, and background). The results show that our method still achieves robust performance in this setting.

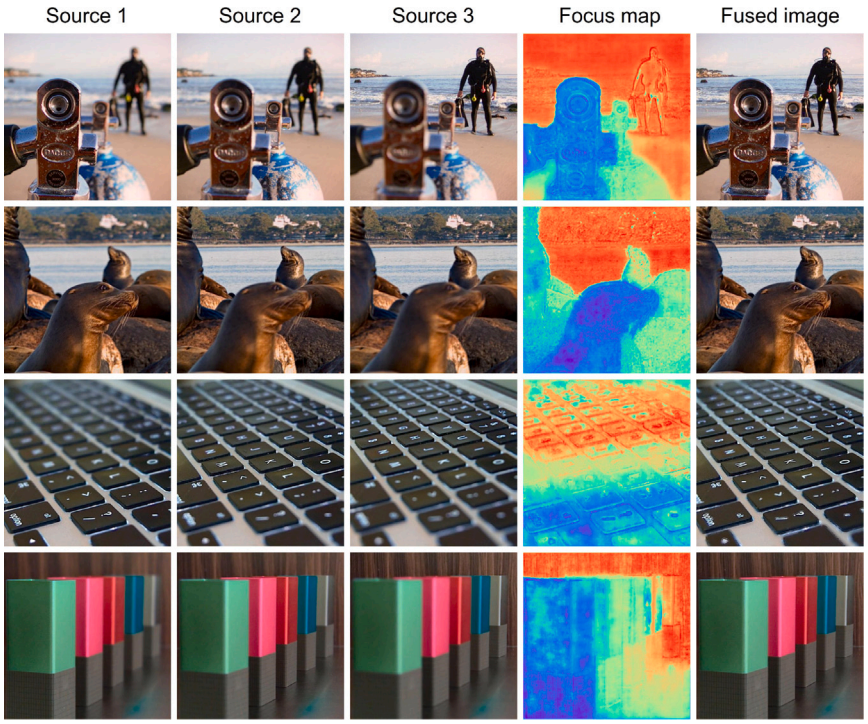


Fig. 17. Effect of the proposed method on fusing small image stacks.

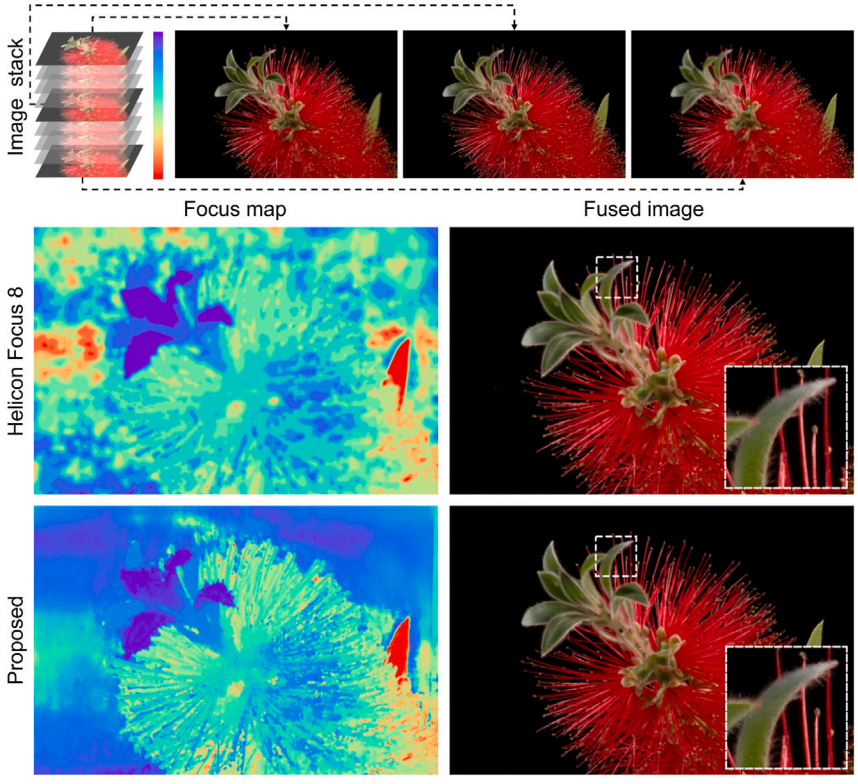


Fig. 18. Comparison with the commercial multi-focus image stack fusion software Helicon Focus 8.

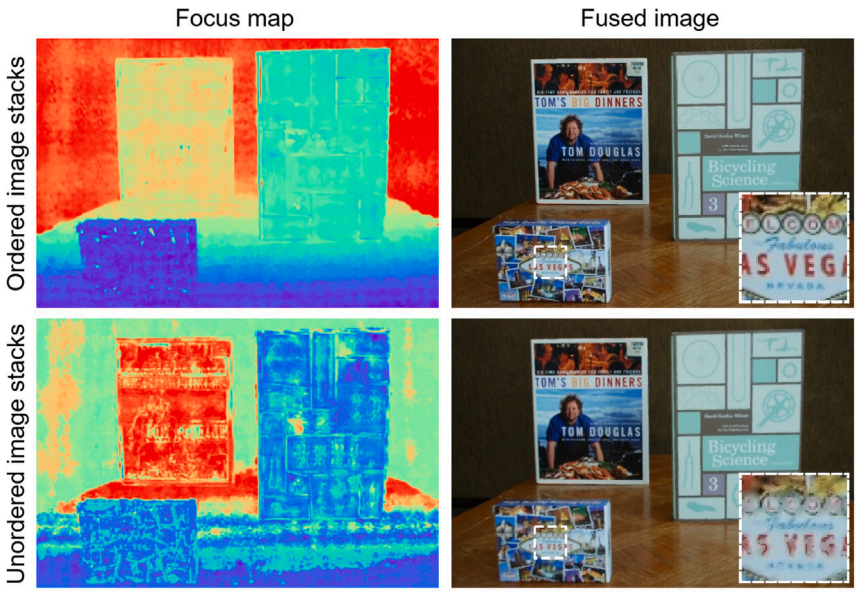


Fig. 19. Focus maps and fused images from ordered vs. unordered image stacks.

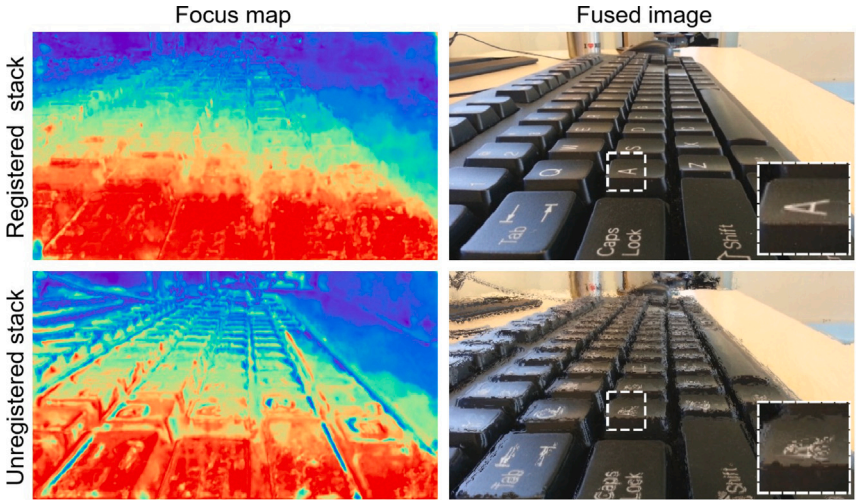


Fig. 20. Comparison of focus maps and fused images from registered vs. unregistered stacks.

Therefore, compared to many existing image fusion networks, our algorithm has a broader range of applications and consistently delivers strong results, highlighting its potential for practical engineering applications.

4.4.4. Comparison with commercial software

To assess the practical potential of our method, we compared it with the commercial focus stacking software Helicon Focus 8. We used a stack of nine pre-registered macro photographs with varying focal planes from a public repository https://github.com/Lackmann1994/FocusStacking. This scenario represents a common challenge in macro photography, where the limited depth of field necessitates the use of focus stacking techniques.

As shown in Fig. 18, our open-source method performs comparably to the commercial software. Moreover, our approach produces a noticeably smoother and more detailed focus map. This map guides the image fusion more effectively, resulting in a sharper fused image.

5. Discussion

A core principle of our model is its dependence on the sequential order of focal planes. The network estimates a probability distribution for each pixel and derives a continuous focal plane value via a differentiable soft-argmax. This approach remains effective under nonlinear or abrupt focus changes, provided the sequential order of focal planes is preserved. However, if the input stack contains discontinuous patterns that violate this ordinal relationship, the results may become unreliable, as they no longer correspond to valid focus plane indices. As shown in Fig. 19, maintaining sequential order yields accurate focus maps and clear fused images (top row), whereas processing the same images in an arbitrary sequence produces chaotic focus maps and less coherent fused images (bottom row). Notably, this reliance on layer order aligns well with many real-world applications, such as PCB inspection, where images are captured sequentially with preset focal ranges.

Another limitation is the network's dependency on preprocessing. Our method is designed for one-shot fusion on a pre-aligned input stack, treating image registration as an upstream, interchangeable module. While our training and testing data are perfectly aligned, real-world stacks often require preprocessing to correct spatial misalignments caused by factors such as camera shake or magnification changes. Fig. 20 illustrates the effect: a properly registered stack produces a clear fused image and well-defined focus map (top row), whereas an unregistered stack introduces artifacts (bottom row). Reliable alignment can be achieved efficiently through a coarse-to-fine pipeline: (1) coarse alignment via SIFT feature detection and matching; (2) sub-pixel refinement using sparse optical flow tracking and motion estimation; and (3) affine transformation to unify all layers within a reference coordinate frame. Example preprocessing scripts are provided in our code repository to facilitate practical deployment.

To address these limitations, future research could explore two promising directions. One avenue is developing advanced fusion techniques capable of handling discontinuous focus patterns that violate ordinal relationships, while the other involves creating a unified framework that integrates preprocessing modules with the fusion network to ensure robust performance in real-world scenarios. In addition, hybrid approaches are particularly promising. Our framework exemplifies this by combining classical depth-map-based fusion (Kozub and Shapoval, 2019) with learning: hand-crafted sharpness metrics are replaced by learnable focus estimation, and hard selection by differentiable regression, enabling end-to-end training. This design retains the fidelity of traditional methods while improving focus estimation. Recent works such as DIMF (Wang et al., 2024a) highlight the potential of incorporating classical priors, which we also consider a promising direction for enhancing robustness and applicability.

6. Conclusion

This paper addresses the limitations of existing multi-focus image fusion methods, which rely on iterative pairwise fusion for image stacks and often suffer from error accumulation and degraded quality. To overcome this, we propose a learning-based, ultra-lightweight one-shot fusion framework that reformulates the task as focal-plane depth regression, preserving pixel fidelity by directly sampling from source images guided by generated focus maps. Our contributions are: (1) a soft-regression-based image stack fusion framework enabling end-to-end training; (2) a depth-map-driven supervision strategy eliminating the need for manually annotated focus maps; and (3) an ultra-lightweight architecture achieving high performance with minimal computational cost.

Extensive experiments on benchmark datasets demonstrate significant improvements in both fusion quality and computational efficiency. The framework may have potential for practical applications in domains such as medical imaging, microscopy, autonomous driving, and industrial inspection, providing a foundation for advancing multi-focus image fusion toward broader adoption in safety-critical and resource-constrained environments, and potentially impacting a wide range of real-world imaging applications.

CRediT authorship contribution statement

Xinzhe Xie: Writing – review & editing, Writing – original draft, Software, Methodology, Conceptualization. **Buyu Guo:** Writing – review & editing, Methodology, Investigation, Funding acquisition, Conceptualization. **Shuangyan He:** Writing – review & editing, Resources, Formal analysis. **Yanzhen Gu:** Writing – review & editing, Visualization, Validation, Data curation. **Yanjun Li:** Writing – review & editing, Resources, Funding acquisition. **Peiliang Li:** Validation, Supervision, Project administration, Conceptualization.

Declaration of competing interest

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

Acknowledgments

This work was supported by the Key R&D Program of "Jianbing Lingyan+X" in Zhejiang Province (No. 2024SSYS0089), the Innovational Fund for Scientific and Technological Personnel of Hainan Province (No. KJRC2023D19), the Government in Guidance of Local Science and Technology Development (No. 2025ZY01111), the Project of Sanya Yazhou Bay Science and Technology City (No. SCKJ-JYRC-2023-59 and No. SCKJ-JYRC-2023-57), and the CNOOC Marine Environmental and Ecological Protection Public Welfare Foundation (No. CF-MEEC/TR/2025-2). Thanks are also due to the Hainan Observation and Research Station of Ecological Environment and Fishery Resource in Yazhou Bay for their support.

Data availability

Data will be made available on request.

References

Ali, U., Lee, B., Lee, H., 2023. 3D shape reconstruction using focus clue from defocus blur detection. In: 2023 2nd International Conference on Emerging Trends in Electrical, Control, and Telecommunication Engineering. ETECTE, IEEE, pp. 1–5.

Burt, P.J., Adelson, E.H., 1985. Merging images through pattern decomposition. In: Tescher, A.G. (Ed.), Applications of Digital Image Processing VIII. Vol. 0575, International Society for Optics and Photonics. SPIE, pp. 173–181. http://dx.doi.org/10.1117/12.966501.

Burt, P.J., Adelson, E.H., 1987. The Laplacian pyramid as a compact image code. In: Readings in Computer Vision. Elsevier, pp. 671–679.

Cheng, C., Xu, T., Wu, X.-J., 2023. MUFusion: A general unsupervised image fusion network based on memory unit. Inf. Fusion 92, 80–92.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B., 2016. The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3213–3223.

Duan, Z., Luo, X., Zhang, T., 2024. Combining transformers with CNN for multi-focus image fusion. Expert Syst. Appl. 235, 121156.

Guo, L., Dai, M., Zhu, M., 2012. Multifocus color image fusion based on quaternion curvelet transform. Opt. Express 20 (17), 18846–18860.

Guo, X., Nie, R., Cao, J., Zhou, D., Mei, L., He, K., 2019. FuseGAN: Learning to fuse multi-focus image via conditional generative adversarial network. IEEE Trans. Multimed. 21 (8), 1982–1996.

Guo, D., Yan, J., Qu, X., 2015. High quality multi-focus image fusion using self-similarity and depth information. Opt. Commun. 338, 138–144.

Haghighat, M.B.A., Aghagolzadeh, A., Seyedarabi, H., 2011. Multi-focus image fusion for visual sensor networks in DCT domain. Comput. Electr. Eng. 37 (5), 789–797. http://dx.doi.org/10.1016/j.compeleceng.2011.04.016, Special Issue on Image Processing.

Hill, P.R., Canagarajah, C.N., Bull, D.R., et al., 2002. Image fusion using complex wavelets. In: BMVC. Citeseer, pp. 1–10.

Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., et al., 2019. Searching for mobilenetv3. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1314–1324.

Hu, X., Jiang, J., Liu, X., Ma, J., 2023. ZMFF: Zero-shot multi-focus image fusion. Inf. Fusion 92, 127–138.

Huang, J., Le, Z., Ma, Y., Mei, X., Fan, F., 2020. A generative adversarial network with adaptive constraints for multi-focus image fusion. Neural Comput. Appl. 32, 15119–15129.

Jiang, S., Yu, S., 2025. Refined multi-focus image fusion using multi-scale neural network with SpSwin autoencoder-based matting. Expert Syst. Appl. 276, 126980. http://dx.doi.org/10.1016/j.eswa.2025.126980, URL: https://www.sciencedirect.com/science/article/pii/S0957417425006025.

Kang, X., Han, F., Fayjie, A., Gong, D., 2023. FocDepthFormer: Transformer with LSTM for depth estimation from focus. arXiv preprint arXiv:2310.11178.

Kendall, A., Martirosyan, H., Dasgupta, S., Henry, P., Kennedy, R., Bachrach, A., Bry, A., 2017. 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.

Kozub, D., Shapoval, I., 2019. Focus stacking of captured images. US Patent 10, 389, 936.

Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., Duerig, T., Ferrari, V., 2020. The open images dataset V4. Int. J. Comput. Vis. 128 (7), 1956–1981. http://dx.doi.org/10.1007/s11263-020-01316-z.

Li, S., Kang, X., Hu, J., 2013. Image fusion with guided filtering. IEEE Trans. Image Process. 22 (7), 2864–2875.

Li, J., Li, X., Li, X., Han, D., Tan, H., Hou, Z., Yi, P., 2024a. Multi-focus image fusion based on multiscale fuzzy quality assessment. Digit. Signal Process. 153, 104592.

Li, X., Li, X., Tan, H., Li, J., 2024e. 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.

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

Li, M., Pei, R., Zheng, T., Zhang, Y., Fu, W., 2024d. FusionDiff: Multi-focus image fusion using denoising diffusion probabilistic models. Expert Syst. Appl. 238, 121664.

Li, L., Song, S., Lv, M., Jia, Z., Ma, H., 2025. Multi-focus image fusion based on fractal dimension and parameter adaptive unit-linking dual-channel PCNN in curvelet transform domain. Fractal Fract. 9 (3), 157.

Li, L., Zhao, X., Hou, H., Zhang, X., Lv, M., Jia, Z., Ma, H., 2024b. Fractal dimension-based multi-focus image fusion via coupled neural p systems in NSCT domain. Fract. Fract. 8 (10), http://dx.doi.org/10.3390/fractalfract8100554, URL: https://www.mdpi.com/2504-3110/8/10/554.

Li, L., Zhao, X., Hou, H., Zhang, X., Lv, M., Jia, Z., Ma, H., 2024c. Fractal dimension-based multi-focus image fusion via coupled neural P systems in NSCT domain. Fract. Fract. 8 (10), 554.

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

Liu, M., 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.

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.

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.

Maximov, M., Galim, K., Leal-Taixé, L., 2020. Focus on defocus: bridging the synthetic to real domain gap for depth estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1071–1080.

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.

Nejati, M., Samavi, S., Shirani, S., 2015. Multi-focus image fusion using dictionary-based sparse representation. Inf. Fusion 25, 72–84.

Ouyang, Y., Zhai, H., Hu, H., Li, X., Zeng, Z., 2025. FusionGCN: Multi-focus image fusion using superpixel features generation GCN and pixel-level feature reconstruction CNN. Expert Syst. Appl. 262, 125665. http://dx.doi.org/10.1016/j.eswa.2024.125665, URL: https://www.sciencedirect.com/science/article/pii/S0957417424025326.

Pei, R., Fu, W., Yao, K., Zheng, T., Ding, S., Zhang, H., Zhang, Y., 2021. Real-time multi-focus biomedical microscopic image fusion based on m-SegNet. IEEE Photonics J. 13 (3), 1–18. http://dx.doi.org/10.1109/JPHOT.2021.3073022.

Pelapur, R., Prasath, V.B.S., Bunyak, F., Glinskii, O.V., Glinsky, V.V., Huxley, V.H., Palaniappan, K., 2014. Multi-focus image fusion using epifluorescence microscopy for robust vascular segmentation. In: 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. pp. 4735–4738. http://dx.doi.org/10.1109/EMBC.2014.6944682.

Ruan, J., Xiang, S., Xie, M., Liu, T., Fu, Y., 2022. MALUNet: A multi-attention and light-weight unet for skin lesion segmentation. In: 2022 IEEE International Conference on Bioinformatics and Biomedicine. BIBM, IEEE, pp. 1150–1156.

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.

Si, H., Zhao, B., Wang, D., Gao, Y., Chen, M., Wang, Z., Li, X., 2023. Fully self-supervised depth estimation from defocus clue. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9140–9149.

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.

Suwajanakorn, S., Hernandez, C., Seitz, S.M., 2015. Depth from focus with your mobile phone. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3497–3506.

Vasiljevic, I., Kolkin, N., Zhang, S., Luo, R., Wang, H., Dai, F.Z., Daniele, A.F., Mostajabi, M., Basart, S., Walter, M.R., et al., 2019. Diode: A dense indoor and outdoor depth dataset. arXiv preprint arXiv:1908.00463.

Wang, L., Lu, H., Wang, Y., Feng, M., Wang, D., Yin, B., Ruan, X., 2017. Learning to detect salient objects with image-level supervision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 136–145.

Wang, J., Qu, H., Zhang, Z., Xie, M., 2024b. 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.

Wang, N.-H., Wang, R., Liu, Y.-L., Huang, Y.-H., Chang, Y.-L., Chen, C.-P., Jou, K., 2021. Bridging unsupervised and supervised depth from focus via all-in-focus supervision. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12621–12631.

Wang, C., Zang, Y., Zhou, D., Mei, J., Nie, R., Zhou, L., 2024a. Robust multi-focus image fusion using focus property detection and deep image matting. Expert Syst. Appl. 237, 121389.

Xie, X., Guo, B., Li, P., He, S., Zhou, S., 2024a. SwinMFF: toward high-fidelity end-to-end multi-focus image fusion via swin transformer-based network. Vis. Comput. 1–24.

Xie, X., Guo, B., Li, P., Jiang, Q., 2024b. Underwater three-dimensional microscope for marine benthic organism monitoring. In: OCEANS 2024-Singapore. IEEE, pp. 1–4.

Xie, X., Lin, Z., Guo, B., He, S., Gu, Y., Bai, Y., Li, P., 2025a. LightMFF: A simple and efficient ultra-lightweight multi-focus image fusion network. Appl. Sci. 15 (13), 7500.

Xie, X., Qingyan, J., Chen, D., Guo, B., Li, P., Zhou, S., 2025b. StackMFF: end-to-end multi-focus image stack fusion network. Appl. Intell. 55 (6), 503. http://dx.doi.org/10.1007/s10489-025-06383-8.

Xu, H., Ma, J., Jiang, J., Guo, X., Ling, H., 2020a. U2fusion: A unified unsupervised image fusion network. IEEE Trans. Pattern Anal. Mach. Intell. 44 (1), 502–518.

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

Yang, F., Huang, X., Zhou, Z., 2022. Deep depth from focus with differential focus volume. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12642–12651.

Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H., 2024. Depth anything V2. arXiv preprint arXiv:2406.09414.

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.

Yu, W., Zhou, P., Yan, S., Wang, X., 2024. Inceptionnext: When inception meets convnext. In: Proceedings of the IEEE/Cvf Conference on Computer Vision and Pattern Recognition. pp. 5672–5683.

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.

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.

Zhang, J., Liao, Q., Liu, S., Ma, H., Yang, W., Xue, J.-H., 2020a. Real-MFF: A large realistic multi-focus image dataset with ground truth. Pattern Recognit. Lett. 138, 370–377.

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

Zhang, H., Ma, J., 2021. SDNet: A versatile squeeze-and-decomposition network for real-time image fusion. Int. J. Comput. Vis. 129 (10), 2761–2785.

Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A., Torralba, A., 2019. Semantic understanding of scenes through the ade20k dataset. Int. J. Comput. Vis. 127, 302–321.

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 pursuing a Ph.D. degree at Ocean College, Zhejiang University. His current research interests include multi-focus image fusion, image pattern recognition under blurred conditions, and the development of underwater imaging 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 Reseaecher Fellow with Ocean College, Zhejiang University, Hangzhou, China. His current research interests include learning-enabled smart sensors and advanced imaging techniques.

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.

Yanzhen Gu received the B.S. degree in marine science and the M.S. degree in physical oceanography from the Ocean University of China, Qingdao, China, in 2006 and 2009, respectively, and the Ph.D. degree in Earth system and geoinformation science from The Chinese University of Hong Kong, Hong Kong, in 2013. He is currently an Associate Professor with Ocean College, Zhejiang University, Hangzhou, China. His research interests include remote sensing with focus on ocean color.

Yanjun Li received the B.S. degree in mechanical engineering and automation, the M.S. degree in mechanical engineering from Nanjing University of Science and Technology, Nanjing, China, in 2010 and 2013, respectively, and the Ph.D. degree from Florida Atlantic University, Boca Raton, FL, USA, in 2016. He is currently a senior engineer with Hainan Institute of Zhejiang University, Sanya, China. His research interests include marine unmanned platforms development and marine sensor development.

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.