General Multi-focus Image Fusion Network

Abstract

Multi-focus image fusion is a computational imaging technique that overcomes the depth-of-field limitation of optical systems by integrating multiple focal planes into a single all-in-focus image. Recently, learning-based approaches for multi-focus image fusion have attracted increasing attention. Among them, a series of stack-based multi-focus fusion models have progressively advanced the research paradigm from image pairs to image stacks. The first-generation model effectively mitigated error accumulation during fusion but failed to preserve the fidelity of the fused image. In contrast, the second-generation model incorporated an ordered-focus prior and provided an open-source solution with performance comparable to that of commercial software. Nevertheless, it assumes ideal inputs, requiring a well-ordered multi-focus image stack free of defocused or invalid layers. To eliminate this constraint and enhance generality, we propose a third-generation model that introduces a redesigned architecture and training strategy. It first employs a *Pyramid Fusion Multilayer Perceptron* to model long-range intra-layer dependencies and estimate layer-wise focus, followed by the proposed *Pixel-wise Cross-Layer Attention* module, which efficiently captures cross-layer relations without relying on focus order. Finally, we formulate focus-map generation as a pixel-wise multi-class classification task to directly predict the focus map for synthesizing the fused image. Extensive experiments demonstrate that the proposed model achieves state-of-the-art performance across diverse benchmarks and real-world applications, highlighting its versatility and generality. The code is available at https://anonymous.4open.science/r/StackMFF-V3-0FCA/.

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

1. Introduction

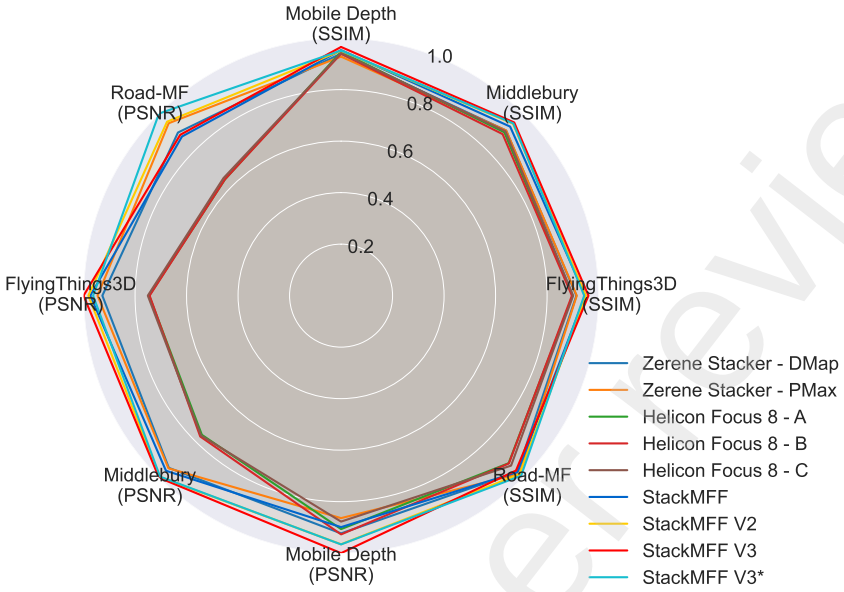

Figure 1: Comparison between the *StackMFF Series* and mainstream commercial-grade focus-stacking software on benchmark datasets (PSNR normalized to the range 0–1 by dividing by the maximum value).

Optical cameras adjust the incident light flux via the lens aperture, which allows a larger aperture to provide a high signal-to-noise ratio under short exposure times, making it suitable for high-speed or low-light imaging. However, the accompanying shallow depth of field (DoF) leads to defocused regions, reducing the overall usable information in the image. While such blur may be desirable in applications like portrait or commercial photography to emphasize the subject, complete focus information is indispensable in scientific imaging and precision industrial inspection. All-in-focus images preserve all relevant scene details, a property that is critical in microscopy (Xie et al., 2024b), biomedical diagnostics (Deng et al., 2025), and industrial chip inspection (Han et al., 2024). Furthermore, synthesized all-in-focus images provide rich visual features that can significantly enhance high-level vision tasks, such as object detection (Li et al., 2024c), semantic segmentation (Jie et al., 2024), and 3D reconstruction (Yan et al., 2020), thereby benefiting complex systems, including autonomous driving and robotic platforms.

One approach for obtaining all-in-focus images is deblurring defocused images. However, defocus blur kernels vary spatially with scene depth and aperture shape,

which makes accurate estimation challenging (Zhang et al., 2022). Traditional methods often adopt a two-stage pipeline: first estimating per-pixel or per-patch defocus kernels based on image priors, and then followed by non-blind deconvolution (Lee et al., 2024). Recently, end-to-end deep learning methods have directly learned the mapping from defocused to all-in-focus images, surpassing traditional approaches in deblurring performance (Li et al., 2025b). Nevertheless, in regions with sparse texture or severe defocus, high-frequency details may still exhibit ringing artifacts, residual blur, or both, leading to suboptimal restoration.

To mitigate the uncertainty of kernel estimation from a single image, multi-focus image stacks (sequences of images captured at different focal planes) can be employed to generate reliable all-in-focus images through fusion algorithms. Traditional multi-focus fusion methods can generally be divided into spatial-domain and transform-domain approaches. Spatial-domain methods compute decision maps based on local activity measures and fuse images accordingly (Liu et al., 2015; Nejati et al., 2015; Li et al., 2023), whereas transform-domain methods first map the images into a transform domain, fuse information across the focal planes, and then reconstruct the all-in-focus image (Li et al., 2024a; Yang et al., 2007; Li et al., 2025a). Recently, deep learning has introduced new avenues for multi-focus fusion, primarily categorized into end-to-end approaches that map the input images directly to an all-in-focus image (Li et al., 2024b; Xie et al., 2024a; Cheng et al., 2023) and decision-map-based approaches that estimate pixel- or region-level focus weights before fusion (Xie et al., 2025b; Quan et al., 2025; Hu et al., 2023).

Historically, particularly after the rise of deep learning, multi-focus image stack fusion has often been treated as an iterative application of pairwise fusion algorithms (Ma et al., 2021; Hu et al., 2023). Moreover, existing benchmark datasets are predominantly composed of image pairs, and corresponding methods are usually designed and evaluated accordingly. However, in practical applications, fusion typically requires processing the entire stack rather than just image pairs. The *StackMFF Series*, through extensive experiments, has demonstrated that iteratively applying pairwise fusion methods to an entire stack leads to unacceptable cumulative errors, significantly degrading the quality of the fused images. This issue is particularly pronounced in deeper stacks or with end-to-end fusion networks.

To address this, *StackMFF* employs a 3D convolutional neural network (3D CNN) to model the entire image stack and map it in an end-to-end manner to an all-in-focus image. While effective, its output pixels are inferred rather than sampled from the input stack, which may reduce image fidelity in scientific imaging and raise reliability concerns. *StackMFF V2* addresses these limitations by leveraging the prior of continuous focus variation in most real-world stacks, introducing depth maps as

a form of proxy supervision, and employing focal-plane soft regression to generate focus maps. Pixels are then indexed from the input stack to synthesize high-fidelity all-in-focus images. Although *StackMFF V2* achieves performance comparable to commercial software on stacks with continuous focus variation, its reliance on ordered focus constrains its applicability. Additionally, *StackMFF V2* assumes ideal input stacks, excluding entirely defocused layers, which must be manually removed. In high-magnification microscopy, where the depth range per focal plane is extremely limited, maintaining this assumption requires highly precise focus stepping; otherwise, fully defocused layers may occur, degrading fusion quality.

To overcome these limitations, this work proposes *StackMFF V3*, the first general multi-focus image stack fusion network. The network not only retains the advantages of previous models—including end-to-end stack processing, lightweight design, support for variable input stack sizes, and full-resolution input—but also eliminates the dependency on the sequential ordering of focal planes in the input stack. It demonstrates strong robustness by producing high-quality fusion results even when the input contains multiple fully defocused layers. Moreover, the method continues the focus-map-based fusion paradigm, ensuring imaging fidelity that is essential for scientific applications. To achieve these objectives, both the network architecture and the training strategy of *StackMFF V3* were redesigned from the ground up. The main contributions of this work are summarized as follows:

1. We reformulate the multi-focus image stack fusion task as a pixel-level multi-class problem and introduce *StackMFF V3*, the first general multi-focus image fusion network designed to overcome the key limitations of its predecessors.
2. We adopt a Pyramid Fusion Multi-Layer Perceptron (PFMLP) to model long-range intra-layer dependencies and propose a *Pixel-wise Cross-layer Attention* module to enable efficient, order-agnostic cross-layer interactions.
3. Extensive qualitative and quantitative experiments demonstrate that *StackMFF V3* consistently outperforms both academic and commercial approaches in terms of fusion quality, scalability, and computational efficiency, across standard benchmarks as well as practical applications.

This study represents the evolution of the *StackMFF Series* from initial exploration to a mature framework, providing not only a novel solution for multi-focus image fusion but also a reference for future research.

The remainder of the paper is organized as follows. Section 2 reviews related work. 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 the proposed approach and outlines potential directions for

future research. Finally, Section 6 concludes the paper by summarizing the main contributions.

2. Related works

This section first reviews computational photography-based approaches for all-in-focus imaging, which typically rely on specialized acquisition procedures or unconventional optical components. Next, we describe the acquisition process of the multi-focus image stacks considered in this work, which are captured via focal-plane scanning. Finally, we discuss post-processing techniques associated with focal-plane scanning, specifically multi-focus image stack fusion.

2.1. All-focus imaging based on computational photography

Conventional optical lenses suffer from a limited DoF, making it difficult to achieve all-in-focus imaging in a single exposure. Over the past few decades, various approaches have been proposed to extend the DoF. One of the most representative methods is focus stacking (Häusler, 1972; Kuthirummal et al., 2010; Choi et al., 2017), which captures a sequence of images at different focal planes by incrementally adjusting the focus along the optical axis and subsequently fuses them in post-processing to produce an all-in-focus image.

Beyond multi-capture approaches, several techniques achieve extended DoF through specialized optical designs. For example, introducing coded apertures or wavefront coding elements into the optical path modifies the point spread function (PSF), enabling computational reconstruction of sharp images by exploiting the distinguishable blur patterns across different depths (Rai and Rosen, 2021). Novel sensor architectures have also opened new possibilities for depth extension. Dual-pixel sensors estimate depth from phase differences to enable refocusing and all-in-focus synthesis (Pyo et al., 2021), while light field cameras employ microlens arrays to record spatial and angular information within a single exposure, facilitating post-capture all-in-focus reconstruction (Sharma et al., 2023). Recent studies further demonstrate that event cameras can leverage the event streams recorded during continuous focal-plane scanning to reconstruct multi-focus stacks and synthesize high-quality all-in-focus images (Lou et al., 2023).

Overall, computational photography-based all-in-focus imaging methods effectively overcome the physical limitations of conventional optics by introducing additional acquisition dimensions or optical encoding during imaging. However, these methods often rely on dedicated hardware or complex capture procedures, which limit their practical applicability. In real-world scenarios, focal-plane scanning combined

with post-fusion of multi-focus image stacks remains the most efficient and widely adopted solution, as it avoids hardware modification while achieving high-quality all-in-focus imaging.

2.2. Acquisition of multi-focus image stacks

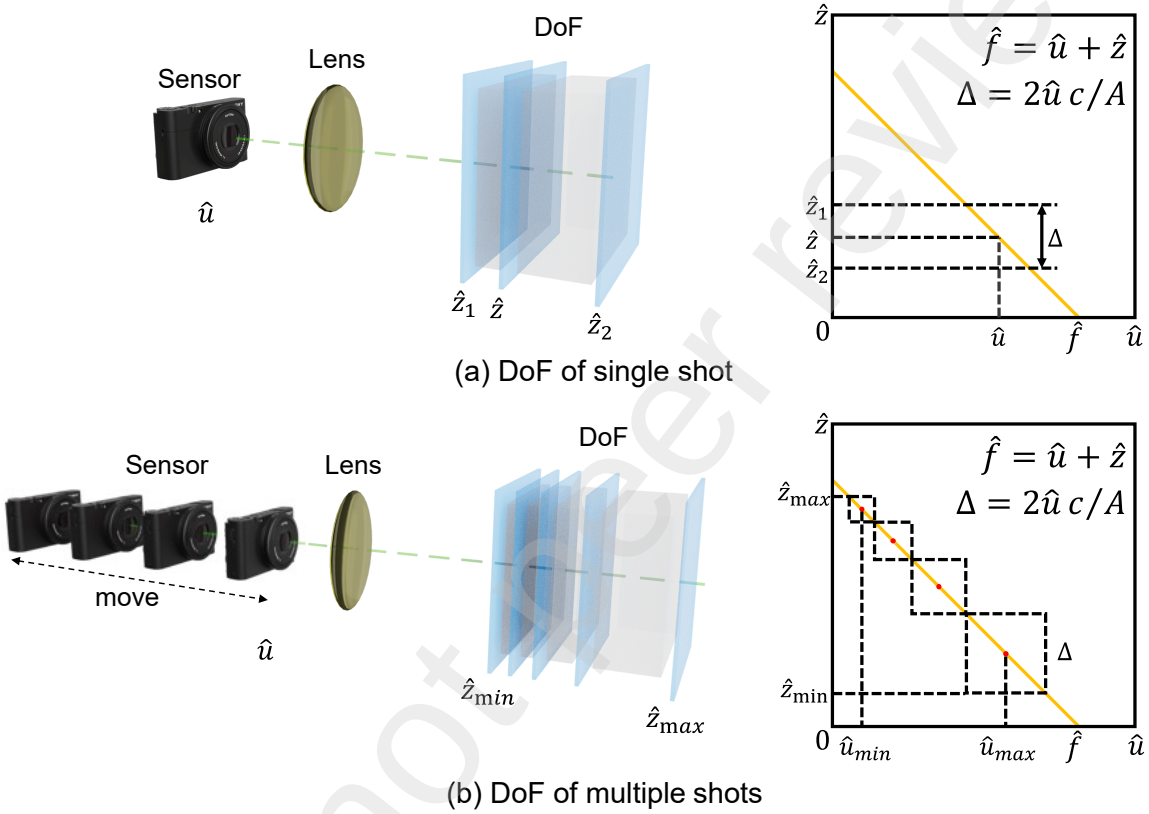

Figure 2: Efficient and complete focus sampling. (a) Left: A geometrical depiction of the depth of field, where objects within $[\hat{z}_1, \hat{z}_2]$ are in focus when $\hat{u}$ and $\hat{z}$ satisfy the thin lens law. Right: The thin lens law is represented as a yellow line in the reciprocal domain, and the corresponding $\hat{z}_1$ and $\hat{z}_2$ can be determined using Eq. (3) and Eq. (4). (b) Efficient and complete focus sampling requires that the DoFs of consecutive sensor positions (e.g., $\hat{v}_{i-1}, \hat{v}_i, \hat{v}_{i+1}$) have neither gaps nor overlaps.

The acquisition of multi-focus image stacks can be broadly categorized into three approaches: mechanical displacement (Zhou et al., 2012), dioptric adjustment (Xie et al., 2024b), and light field imaging (Sharma et al., 2023). In the mechanical displacement approach, the camera or the subject is moved incrementally along the

optical axis to capture in-focus images at different depths, which is suitable for high-precision microscopic imaging and macro photography. The dioptric adjustment method employs electronically controlled liquid lenses or varifocal lenses to rapidly switch the focus. Light field imaging, on the other hand, uses a microlens array to capture the four-dimensional light field of the scene, enabling the generation of multi-focus stacks through digital refocusing.

Taking the mechanical displacement method as an example, to ensure that all scene points are in focus in at least one image of the stack, an ideal sampling strategy should satisfy two criteria (Zhou et al., 2012):

1. Completeness: The collective DoF from all acquired images must encompass the entire target depth range. Given DOF^* as the desired depth span, this condition can be expressed as:

$$DOF_1 \cup DOF_2 \cup DOF_3 \cdots \cup DOF_n \supseteq DOF^*, \tag{1}$$

   where $\cup$ denotes a union operation and $\supseteq$ indicates a superset relationship.
2. Efficiency: To minimize the total number of required images, individual DoFs should not overlap. This can be represented as:

$$DOF_1 \cap DOF_2 \cap DOF_3 \cdots \cap DOF_n = \emptyset, \tag{2}$$

   where $\cap$ denotes the intersection.

We initiate our analysis of the DoF using the fundamental thin lens law, expressed as $1/f = 1/u + 1/z$. In this formula, f denotes the lens's focal length, u represents the distance between the sensor and the lens, and z represents the distance to the object. Conventionally, this equation is transformed into the reciprocal domain, where it becomes $\hat{f} = \hat{u} + \hat{z}$, with $\hat{x}$ defined as $1/x$. In this reciprocal space, the thin lens law takes a linear form. Notably, the reciprocal of the object distance, $\hat{z}$, is frequently expressed in diopters (m^{-1}).

The DoF, bounded by $[z_1, z_2]$, delineates the spatial extent within which the resulting blur radius remains below the critical circle of confusion, c. In this paper, we equate the pixel size to the diameter of the circle of confusion, consistent with standard imaging practices. Given a specific sensor-lens distance, $\hat{u}$, the upper and lower bounds of the DoF in the reciprocal domain, $\hat{z}_1$ and $\hat{z}_2$, are defined as:

$$\hat{z}_1 = \hat{z} + \hat{u} \cdot c/A \tag{3}$$

$$\hat{z}_2 = \hat{z} - \hat{u} \cdot c/A \tag{4}$$

Here, A denotes the diameter of the lens aperture. It is important to note that both the position and the extent of the DoF are influenced by the sensor's location. Fig. 2(a) provides a visual representation of the DoF geometry for various sensor positions u (left) and its manifestation in the reciprocal domain (right). The yellow line in this figure illustrates the thin lens law. Consequently, from Eq. (3) and (4), for any chosen sensor position $\hat{u}$, the effective size of the DoF in the reciprocal domain, Δ, is given by $2 \cdot \hat{u} \cdot c/A$.

To achieve both efficient and exhaustive sampling, it is a prerequisite that no two adjacent DoF regions overlap or leave gaps between them, as depicted in Fig. 2(b). Drawing upon the thin lens law, we establish the following relationships for consecutive sensor positions:

$$|\hat{u}_i - \hat{u}_{i+1}| = |(\hat{f} - \hat{z}_i) - (\hat{f} - \hat{z}_{i+1})| \tag{5}$$

$$|\hat{u}_i - \hat{u}_{i+1}| = |\hat{z}_i - \hat{z}_{i+1}|, \tag{6}$$

Further, by combining Eq. (3) and (4), we derive the relationship for the difference between successive sensor positions:

$$|\hat{u}_i - \hat{u}_{i+1}| = (\hat{u}_i + \hat{u}_{i+1}) \cdot c/A, \tag{7}$$

where $\hat{u}_i$ and $\hat{u}_{i+1}$ denote the sensor locations corresponding to two successive DoF regions.

In the context of consumer photography, the object distance z is significantly larger than the sensor-lens distance u, implying that $\hat{u}_i \approx \hat{f}$. By applying this approximation to Eq. (7), it can be simplified as follows:

$$\hat{u}_i \cdot \hat{u}_{i+1} \cdot |\hat{u}_i - \hat{u}_{i+1}| = \hat{u}_i \cdot \hat{u}_{i+1} \cdot (\hat{u}_i + \hat{u}_{i+1}) \cdot c/A \tag{8}$$

$$|u_{i+1} - u_i| = (u_i + u_{i+1}) \cdot c/A \tag{9}$$

$$\delta u \approx 2 \cdot f \cdot c/A \tag{10}$$

$$\delta u \approx 2 \cdot c \cdot N, \tag{11}$$

where $N = f/A$ denotes the f-number of the lens. Eq. (11) indicates that an effective and comprehensive sampling strategy requires the sensor to advance by a uniform displacement δu between successive image captures. This fixed displacement is primarily determined by the pixel size c and the f-number N. It is important to note that this constant increment occurs in the normal domain rather than in the reciprocal domain. Consequently, if a camera operates at a constant frame rate P, the optimal strategy for sampling the desired depth range involves moving the sensor at a uniform speed:

$$s = \frac{\delta u}{\delta t} = 2 \cdot c \cdot N \cdot P. \tag{12}$$

2.3. Multi-focus image stack fusion

Applying image-pair fusion methods iteratively to an image stack is straightforward in implementation but suffers from cumulative errors, leading to degraded fusion quality as the number of layers increases (Xie et al., 2025c). Recognizing the importance of modeling the entire image stack holistically, several studies have explored global optimization approaches. For instance, some works have computed focus index maps for the whole multi-focus stack using Difference-of-Gaussian (DoG) focus measures and have employed them to synthesize the final fused image (Agarwala et al., 2004; Hasinoff and Kutulakos, 2011). To further improve robustness, pyramid strategies with spatial smoothing optimization have been introduced to enhance performance in textureless regions (Zhou et al., 2012). Rotation-invariant focus measures, which perform well on microscopic multi-focus stacks, have also been proposed (Wu et al., 2024). Moreover, commercial software such as *Helicon Focus* applies multi-stage smoothing to the computed index map, demonstrating robustness on stacks with continuous focus variation (Kozub and Shapoval, 2019).

With the advent of deep learning, researchers have begun leveraging neural networks to process multi-focus image stacks as unified entities. These networks jointly model intra-layer and inter-layer focus relationships. Initially, in the depth-from-focus (DfF) domain, multi-focus stacks were primarily used for depth estimation, and the resulting all-in-focus images were treated merely as byproducts (Wang et al., 2021; Suwajanakorn et al., 2015). This inspired the emergence of learning-based fusion methods that explicitly target multi-focus stack fusion (Xie et al., 2025c; Araujo et al., 2023). The *StackMFF Series* was one of the first to apply deep learning to this task (Xie et al., 2025c). Through extensive experiments and analyses, the *StackMFF Series* revealed the inherent limitations of iterative pairwise fusion and proposed high-performance baseline models. Additionally, benchmark datasets for multi-focus image stack fusion were established, which provide valuable resources for future research.

The original *StackMFF* employs a 3D CNN to map the entire image stack to an all-in-focus image in an end-to-end manner. This effectively addresses the error accumulation problem inherent in iterative fusion. However, because the fused pixels are directly inferred by the network rather than sampled from the input stack, this approach may compromise fidelity and pose potential risks in scientific imaging. To address this issue, *StackMFF V2* reformulated stack fusion as a focal-plane depth regression problem. It leverages the numerical equivalence between depth maps and focus maps under linearly spaced focal depths. Using this equivalence, the network employs a differentiable soft regression strategy that treats depth maps as proxy supervision for focus maps. The network first predicts a focus map, from

which the fused image is obtained via index-based selection, effectively overcoming the limitations of the original *StackMFF*. Furthermore, the framework exploits the inherent continuity of focus variation during focal-plane scanning as a prior. This results in an open-source solution whose performance is on par with commercial software such as *Helicon Focus 8* (Kozub and Shapoval, 2019).

Table 1: Comparison of key characteristics among different versions of *StackMFF Series*.

Characteristic	StackMFF	StackMFF V2	StackMFF V3
Global Receptive Field	×	×	✓
Lightweight	×	✓	✓
Anti-Interference	✓	×	✓
Order-Independent	✓	×	✓
Fidelity	×	✓	✓
Full-Resolution Input	✓	✓	✓
Scalability	✓	✓	✓
One-shot Fusion	✓	✓	✓

Specifically, *StackMFF V3* simultaneously achieves several key capabilities: a global receptive field (intra-layer), a lightweight architecture, anti-interference capability, order-independent input handling, high fidelity, full-resolution input support, scalability to variable stack sizes, and one-shot fusion. Together, these improvements enhance the flexibility and robustness of multi-focus image stack fusion relative to earlier versions.

As with most existing multi-focus image fusion methods, the *StackMFF Series* assumes that the input image stacks are static and have undergone prior spatial registration.

3. Methods

In this section, we present the proposed *StackMFF V3*. We begin with an overview of the overall architecture, followed by detailed descriptions of the functionality of each stage. Finally, we introduce the loss functions and training strategy employed for model optimization.

3.1. Method overview

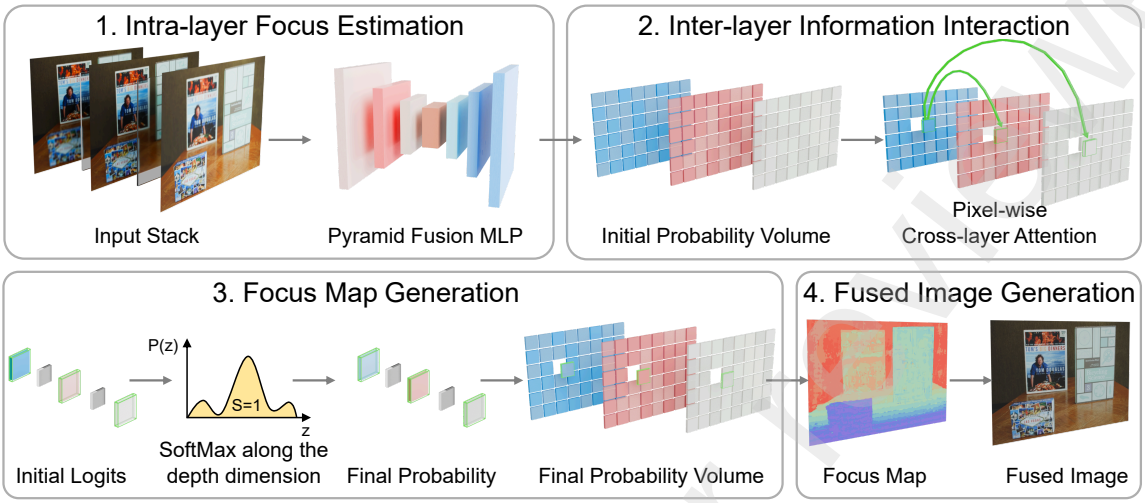

Figure 3: Framework of the proposed general multi-focus image fusion network, StackMFF V3

Fig. 3 illustrates the proposed *StackMFF V3* general multi-focus image fusion framework, which can be roughly divided into four stages: intra-layer focus estimation, inter-layer information interaction, focus map generation, and fused image generation.

Given an input multi-focus image stack, we first employ a Pyramid Fusion MLP (PFMLP)-based visual backbone to independently perform intra-layer focus estimation for each image layer. Pixel-wise in-focus probabilities are obtained in the spatial domain using a Sigmoid function, resulting in an initial probability volume. Next, the proposed *Pixel-wise Cross-layer Attention* module models inter-layer dependencies for pixels located at the same spatial positions across different layers. A SoftMax operation is then applied along the depth dimension to produce the final probability volume. The focus map is subsequently generated by selecting the layer with the highest in-focus probability along the depth dimension. Finally, the fused image is synthesized by directly sampling pixels from the original input stack according to the computed focus map.

To avoid ambiguity, the term *focus map* in this paper refers to a two-dimensional matrix composed of layer indices, where each element indicates the index of the image layer from which the corresponding pixel in the final fused image is sampled.

3.2. Intra-layer focus estimation

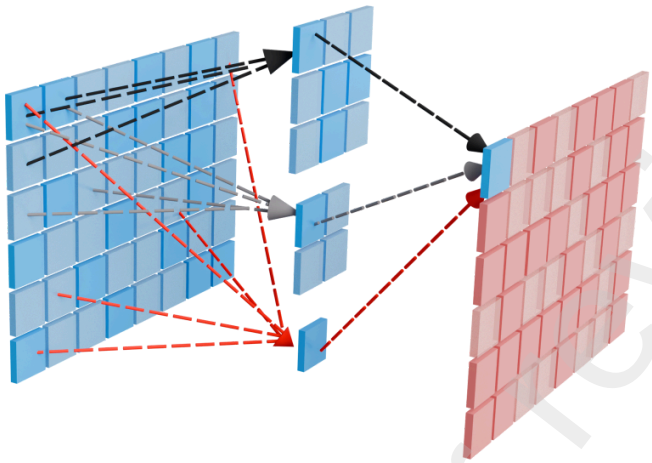

Figure 4: Architecture of the adopted Pyramid Fusion MLP (PFMLP)

StackMFF V2 performs layer-wise focus estimation independently for each image layer, thereby avoiding the substantial and redundant computations associated with the 3D convolutional network used in *StackMFF* (Xie et al., 2025c). We thus adopt the same design principle. However, the inherent receptive field limitations of the convolutional networks employed for intra-layer focus estimation restrict their estimation accuracy. Although recent visual backbones based on Transformer (Xie et al., 2024a) or Mamba (Xie et al., 2025b) models can address this limitation, we adopt a state-of-the-art, efficient MLP-based architecture, referred to as Pyramid Fusion MLP (PFMLP), as the backbone for intra-layer focus estimation to balance receptive field size, computational overhead, and platform compatibility (Fig. 4). Specifically, each block in PFMLP incorporates multi-scale pooling and fully connected layers to construct a feature pyramid, followed by upsampling and additional fully connected layers for feature fusion. By leveraging different downsampling rates, the network captures both long-range dependencies and fine-grained cues and exploits global contextual information to enhance spatial representation. This architecture has been demonstrated to be highly effective for multi-focus image fusion (Song et al., 2025). We refer readers to the original PFMLP paper for more detailed information (Huang et al., 2025).

To efficiently extract intra-layer focus cues, the PFMLP processes each grayscale input image independently through an encoder-decoder architecture. Given a grayscale input image with shape $[B, 1, H, W]$, where B is the batch size and $H \times W$

represents the spatial dimensions, the PFMLP processes the image through a series of transformations that progressively extract hierarchical representations. First, the input image passes through a stem convolution layer that downsamples the image by a factor of 2 while expanding the channel dimension from 1 to C_0, resulting in features with shape $[B, C_0, H/2, W/2]$. This initial feature extraction helps to reduce computational complexity while preserving essential visual information. Next, the features are passed through multiple stages of Pyramid Bottleneck blocks. Each stage progressively increases the channel dimension while reducing the spatial resolution, following a typical encoder structure that downsamples the feature maps from $[B, C_0, H/2, W/2]$ to $[B, C_4, H/32, W/32]$. Each Pyramid Bottleneck block combines spatial and channel-wise feature extraction through parallel pathways, enabling the network to capture both local and global focus characteristics effectively. After the encoder stages, our UPerNet-style (Xiao et al., 2018) decoder fuses multi-scale features to generate pixel-wise focus maps. The decoder employs a Pyramid Pooling Module on the deepest feature map to capture global contextual information, followed by a Feature Pyramid Network that utilizes lateral connections from different encoder stages to build a comprehensive feature representation. The feature fusion process involves progressive upsampling and channel unification operations that transform the multi-scale features back to the original spatial resolution.

The final probability map is generated through a Sigmoid function applied to the decoder output, resulting in features with shape $[B, C_0, H, W]$. This ensures that the probability values are normalized to the range [0, 1], where values closer to 1 indicate a higher likelihood of being in focus. The probability maps of all layers together constitute the initial probability volume.

3.3. Inter-layer information interaction

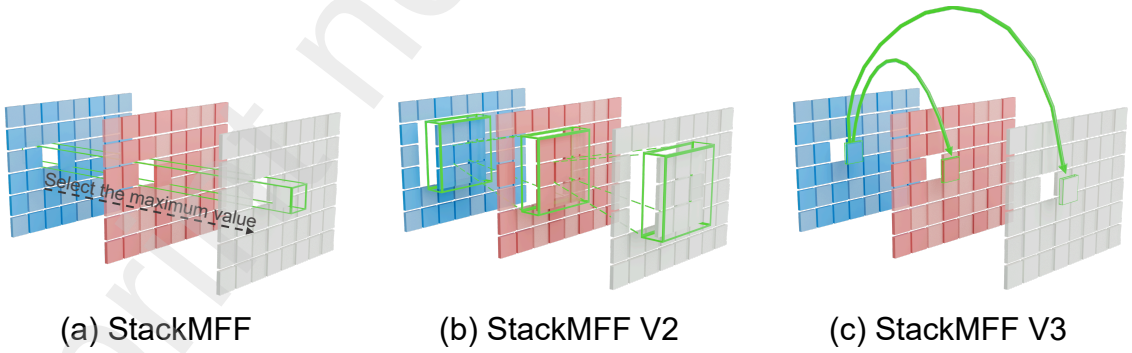

Figure 5: Comparison of inter-layer modeling strategies across the *StackMFF Series*.

As shown in Fig. 5, to aggregate and exploit inter-layer information for improved focus estimation, *StackMFF* (Xie et al., 2025c) employs a combination of 3D convolutional kernels and depth-wise max pooling. Although this approach is simple and effective, it relies on a fixed and manually designed strategy that lacks adaptability. *StackMFF V2*, on the other hand, introduces a bidirectional convolutional gated network for inter-layer modeling, based on the assumption of an ordered focus variation in the input image stack, which imposes constraints on the input sequence. To address these limitations, we propose a novel and more efficient inter-layer modeling approach with a minimal computational cost, termed the *Pixel-wise Cross-layer Attention* module.

Given the feature maps from all focal layers with shape $[B, N, C_0, H, W]$, where B is the batch size, N is the number of focal layers, C_0 denotes the embedded channel dimension, and $H \times W$ represents the spatial dimensions, our *Pixel-wise Cross-layer Attention* module processes these features as follows: First, we reshape the input tensor to enable independent attention computation at each spatial location. Specifically, we permute and reshape the input from $[B, N, C_0, H, W]$ to $[B \times H \times W, N, C_0]$, allowing each spatial position to have its own attention computation across the N focal layers. This design ensures that the relationships between layers can be modeled differently for each spatial location, which is essential for handling spatially varying focus characteristics.

The core of the *Pixel-wise Cross-layer Attention* module is a standard Transformer, which operates along the depth dimension (i.e., across different focal layers) to capture both local and global inter-layer relationships. The reshaped features are then passed through multiple Transformer layers, each consisting of a multi-head self-attention mechanism followed by a feed-forward network. To enhance the positional awareness of our model while maintaining flexibility, we employ Rotary Positional Encoding (RoPE) (Su et al., 2024) to inject positional information into the attention computation. The RoPE encoding is applied along the layer dimension, allowing the model to distinguish between different focal layers effectively.

After processing through the Transformer layers, we reshape the features back to their original spatial structure, resulting in output features with the same shape as the input, i.e., $[B, N, C_0, H, W]$. This design ensures that our layer interaction module can be seamlessly integrated into the overall network architecture while effectively leveraging information along the depth dimension.

One of the key advantages of our approach is its computational efficiency. Although the self-attention mechanism has a theoretical complexity of $\mathcal{O}(N^2)$ with respect to the number of focal layers N, in practice, the number of focal layers in multi-focus image fusion tasks is typically small. Combined with our efficient im-

plementation that uses only two Transformer layers ($L = 2$), our approach achieves high computational efficiency. Importantly, owing to the relatively low computational overhead of our approach, we are able to deploy it at full resolution without a significant performance impact. This is in contrast to many attention-based methods that require downsampling to be computationally feasible.

3.4. Focus map generation

StackMFF (Xie et al., 2025c) directly generates a fused image via end-to-end regression; in contrast, *StackMFF V2* derives the focus map by applying soft regression to the layer indices and subsequently produces the fused image based on the focus map. In this work, we reformulate focus map generation as a pixel-level multi-class classification task, directly predicting for each pixel the index of the most in-focus layer, thereby constructing the focus map to produce the fused image.

The features output by the *Pixel-wise Cross-layer Attention* module are used to generate the final focus map. During training, these features serve as logits for a multi-class classification task, where each spatial location predicts which focal layer is the most in-focus and sharp. The training objective is formulated as a standard cross-entropy loss:

$$\mathcal{L} = \frac{1}{B \cdot H \cdot W} \sum_{b=1}^{B} \sum_{i=1}^{H} \sum_{j=1}^{W} \mathrm{CE}(f_{b,:,i,j}, g_{b,i,j}) \tag{13}$$

where $f \in \mathbb{R}^{B \times N \times H \times W}$ represents the layer interaction features produced by the *Pixel-wise Cross-layer Attention* module, which are reduced to a single channel, and $g \in \mathbb{Z}^{B \times H \times W}$ denotes the ground-truth focus map with values in the range $\{0, 1, \ldots, N-1\}$. Note that the ground-truth indices are 0-based, consistent with PyTorch's cross-entropy implementation.

During training, the logits of the features reduced to a single channel produced by the *Pixel-wise Cross-layer Attention* module are directly used to compute the loss against the ground-truth focus map for model optimization. During inference, a softmax operation along the depth dimension, followed by an argmax, is applied to determine the most in-focus layer at each spatial location, which produces the final probability volume. The probability values in this volume are subsequently used to construct the focus map according to a winner-takes-all strategy applied along the depth dimension.

3.5. Fused image generation

The inference process consists of two steps. First, the *Pixel-wise Cross-layer Attention* module produces features that are reduced to a single channel. A softmax

operation is then applied along the depth dimension to generate the final probability maps, which together form the probability volume:

$$P_{b,k,i,j} = \frac{\exp(f_{b,k,i,j})}{\sum_{l=0}^{N-1} \exp(f_{b,l,i,j})} \tag{14}$$

Then, the argmax operation is further applied along the depth dimension of the final probability volume to determine the most in-focus layer at each spatial location, thereby yielding the predicted focus map:

$$F_{b,i,j} = \arg\max_{k} P_{b,k,i,j}, \quad k \in \{0, 1, \ldots, N-1\} \tag{15}$$

where $P \in \mathbb{R}^{B\times N\times H\times W}$ denotes the final probability volume, and $F \in \mathbb{Z}^{B\times H\times W}$ represents the predicted focus map.

Finally, the fused image is generated by selecting pixels from the input image stack according to the predicted focus map:

$$I^{\text{fused}}_{b,i,j} = I_{b,F_{b,i,j},i,j} \tag{16}$$

where $I \in \mathbb{R}^{B\times N\times H\times W}$ denotes the input image stack, and $I^{\text{fused}} \in \mathbb{R}^{B\times H\times W}$ represents the fused image.

4. Experiments

In this section, we present a comprehensive set of experiments to demonstrate both the effectiveness and practical potential of the proposed method. We first describe the implementation details, including the training strategy, the synthesis method for training datasets, hyperparameter settings, and the datasets used for training and evaluation. Next, we present both qualitative and quantitative comparisons with several state-of-the-art methods. Finally, we conduct a detailed analysis of key aspects, including model efficiency, performance in small-stack fusion tasks, and statistical significance. In addition, we separately evaluate the proposed method in terms of Order-Independence, Anti-Interference, Fidelity, and Generalizability.

4.1. Implementation details

4.1.1. Training strategy

In the previous study, *StackMFF* (Xie et al., 2025c) was trained on image stacks containing a fixed number of layers, which restricts its ability to generalize across stacks of varying sizes. To address this limitation, *StackMFF V2* was trained on multiple fixed stack sizes. In contrast, *StackMFF V3* generalizes this approach by

training on image stacks with a continuously varying number of layers, ranging from 2 to 24. For each training instance, the number of layers is randomly sampled within this range.

All experiments were conducted on a high-performance computing platform equipped with dual NVIDIA A6000 GPUs and an Intel(R) Xeon(R) Platinum 8375C CPU. We used the AdamW optimizer with a batch size of 8 and an initial learning rate of 1×10^{-3}, which was exponentially decayed by a factor of 0.9 per epoch. The training lasted for 50 epochs and required approximately 117 hours.

In addition, we trained a variant denoted as *StackMFF V3**, which adopts the same network architecture as *StackMFF V3* but follows the training strategy of *StackMFF V2*. Owing to the incorporation of focal-order priors, *StackMFF V3** generates smoother focus maps than *StackMFF V3*.

4.1.2. Datasets for training

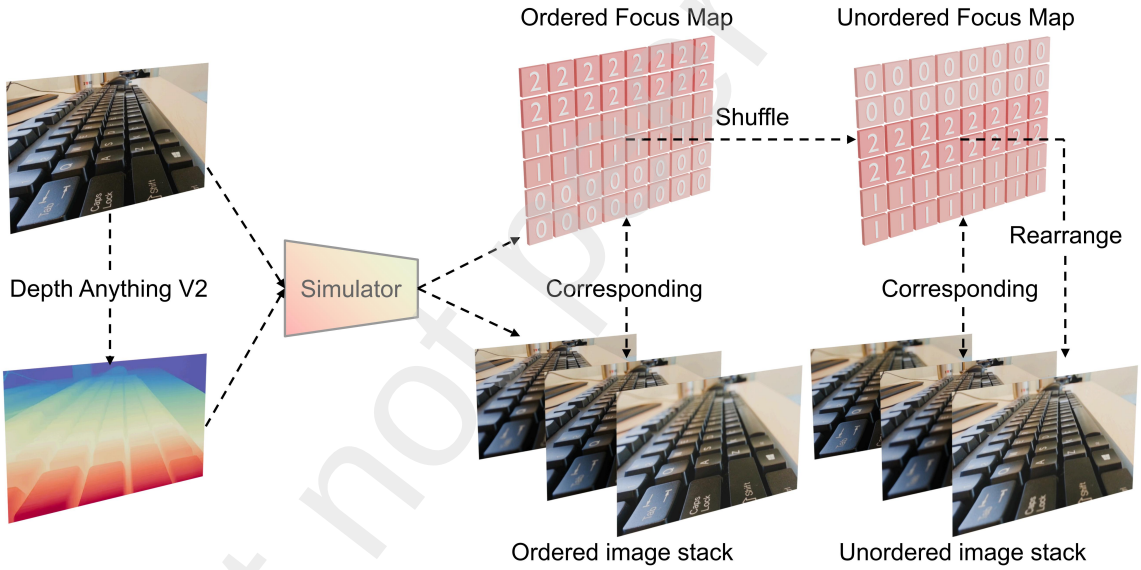

Figure 6: Pipeline for synthesizing multi-focus image stacks and their corresponding focus maps used for model training.

Algorithm 1 Training Data Generation Process

Require:

I: All-in-focus input image

D: Corresponding depth map

N: Number of focal planes (randomly selected between 2 and 24)

Ensure:

$\{L_k\}_{k=1}^{N}$: Synthesized multi-focus image stack

F: Ground-truth focus map

Quantize normalized depth map D into N discrete regions: $Q \leftarrow Quantize(Normalize(D), N)$

for $k \leftarrow 1$ to N **do**

    Create a blurred version B_k of I using a Gaussian kernel with blur radius proportional to k

    Generate layer L_k where pixels in region $Q = k$ are replaced with corresponding pixels from I (sharp),

        and pixels in other regions are replaced with those from $B_{|Q-k|}$ (blurred)

end for

Randomly shuffle layer order: $\{L'_k\} \leftarrow Shuffle(\{L_k\})$

Generate focus map F based on the shuffled order $\{L'_k\}$ **return** $\{L'_k\}_{k=1}^{N}$, F

Our model is trained on several publicly available datasets, including DUTS (Wang et al., 2017), NYU Depth V2 (Silberman et al., 2012), DIODE (Vasiljevic et al., 2019), Cityscapes (Cordts et al., 2016), and ADE20K (Zhou et al., 2019). For all datasets except NYU Depth V2, scene depth maps are estimated using Depth Anything V2 (Yang et al., 2024) to compensate for missing or incomplete depth information.

Following the layered blur synthesis strategy introduced in *StackMFF* (Xie et al., 2025c), multi-focus image stacks are generated from single all-in-focus images using the corresponding depth maps. Specifically, the depth map is normalized, quantized into N discrete regions, and used to simulate images focused at different depths. The resulting layers are then randomly shuffled to remove any focus-order bias, and a corresponding focus map is generated to record the ground-truth focus index for each pixel.

The overall data synthesis procedure is illustrated in Fig. 6 and summarized in Algorithm 1.

4.1.3. Datasets for evaluation

We evaluate the proposed method on the task of multi-focus image stack fusion using several benchmark datasets adopted from *StackMFF V2*, including Mobile Depth (Suwajanakorn et al., 2015), FlyingThings3D (Mayer et al., 2016), Middlebury Stereo (Scharstein et al., 2014), and Road-MF (Li et al., 2024c). Detailed descriptions of these datasets can be found in the original paper.

4.1.4. Methods for comparison and evaluation metrics

To evaluate the performance of the proposed fusion framework, we compare it against 17 representative state-of-the-art methods. These include 5 traditional techniques—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)—as well as 12 learning-based approaches: IFCNN (Zhang et al., 2020), U2Fusion (Xu et al., 2020), 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), CCSR-Net / MCCSR-Net (Zheng et al., 2025), *StackMFF* (Xie et al., 2025c), and its improved version, *StackMFF V2* (Xie et al., 2025a). It is worth noting that only the *StackMFF Series* supports one-shot stack fusion, whereas all other algorithms depend on iterative pairwise fusion.

Additionally, we include two widely used commercial focus-stacking software packages with batch-processing capabilities for comparison: *Helicon Focus 8* and *Zerene Stacker*, which provide three and two fusion methods, respectively.

For the traditional approaches, default parameters from their official implementations are used. Learning-based methods are evaluated using their publicly released pre-trained models. For methods that do not natively support stack fusion, iterative pairwise fusion is performed following the default order of the input stack, which requires $N-1$ fusion operations for an N-image stack. Regarding commercial software, since their built-in alignment procedures may produce misaligned outputs relative to the ground truth and thus bias quantitative evaluation, we re-align the results with the ground truth prior to computing the evaluation metrics.

To quantitatively assess fusion quality, we employ two widely used full-reference metrics: the Structural Similarity Index (SSIM) and the Peak Signal-to-Noise Ratio (PSNR). These metrics evaluate the structural consistency and pixel-level fidelity of the fused images with respect to the ground-truth all-in-focus images provided in our test datasets.

4.2. Qualitative comparison

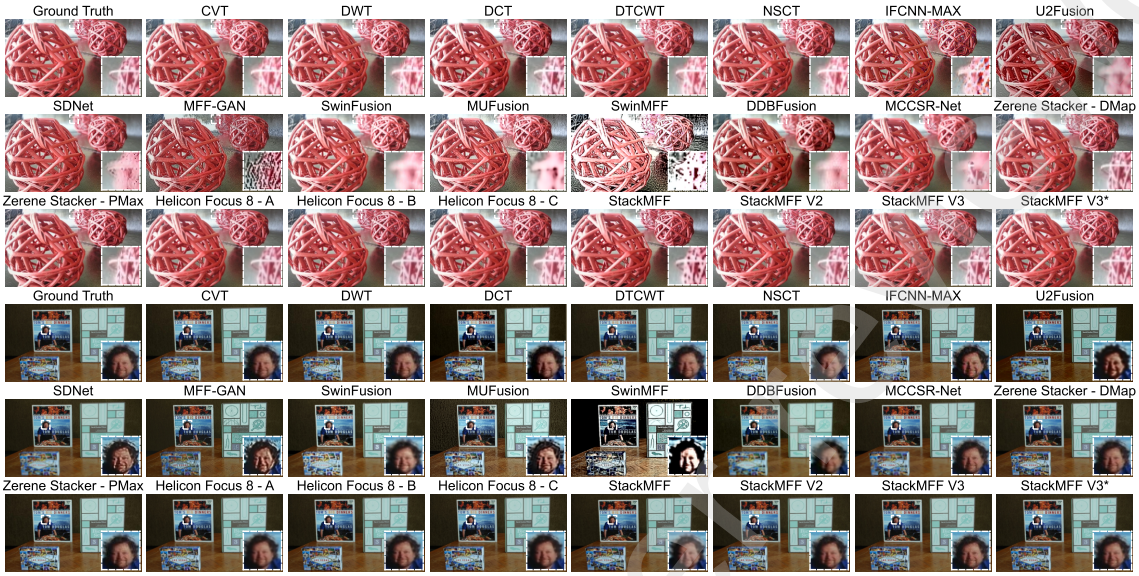

Figure 7: Comparison of fusion results produced by different methods on the Mobile Depth dataset (Suwajanakorn et al., 2015).

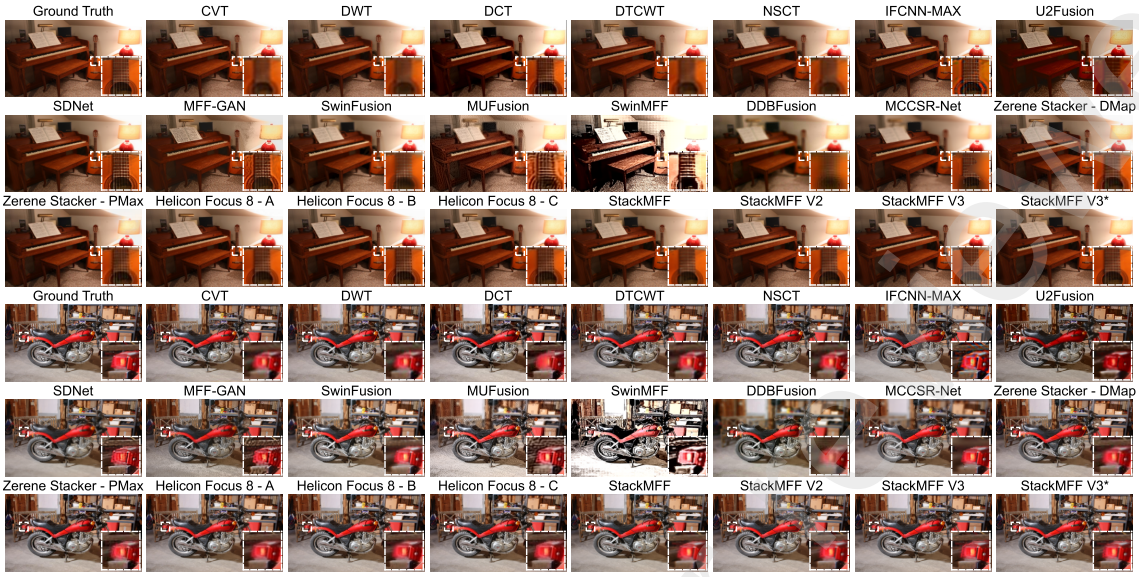

Figure 8: Comparison of fusion results produced by different methods on the Middlebury Stereo dataset (Scharstein et al., 2014).

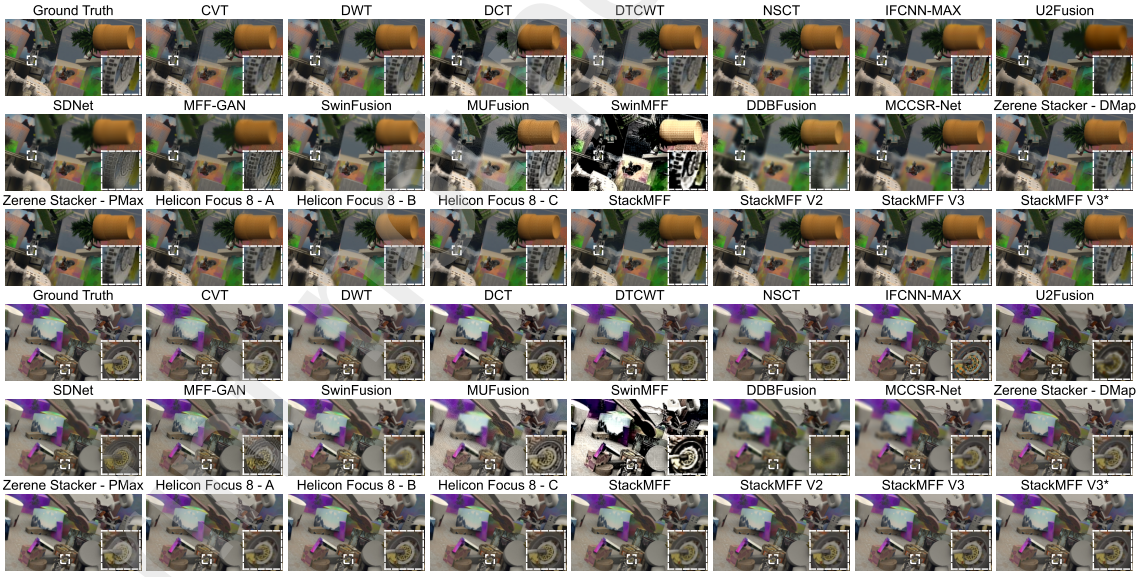

Figure 9: Comparison of fusion results produced by different methods on the FlyingThings3D dataset (Mayer et al., 2016).

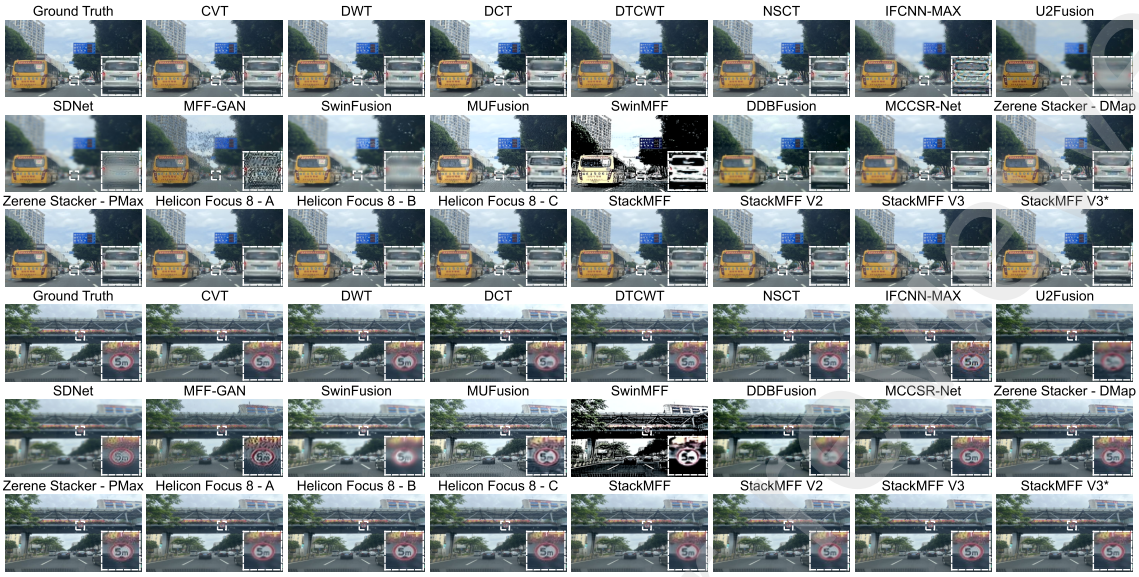

Figure 10: Comparison of fusion results produced by different methods on the Road-MF dataset (Li et al., 2024c).

To comprehensively evaluate the effectiveness of the proposed method, we present the fusion results of several representative approaches across four benchmark datasets, as shown in Fig. 7 through Fig. 10. These datasets encompass a wide range of scenarios, including controlled indoor environments, synthetic imagery, and challenging outdoor scenes.

The comparative results demonstrate that traditional methods, along with commercial focus-stacking software (which is also typically based on traditional algorithms), exhibit consistent and robust fusion performance across various scenarios. They effectively preserve structural details and overall image quality, with almost no fusion failures observed. In contrast, most learning-based approaches—except for the *StackMFF Series*—tend to suffer from accumulated errors in multi-focus image stack fusion. This often results in noticeable artifacts, including blurring, noise, and structural distortions. Notably, among all evaluated deep learning-based methods, only the *StackMFF Series*, particularly our proposed *StackMFF V3*, achieves superior fusion quality across multiple datasets, consistently matching or even surpassing the performance of traditional methods. Moreover, as can be clearly observed in the magnified regions, *StackMFF V3* provides significantly improved fusion quality compared to its predecessors; for instance, the facial details in Fig. 7 and the text on the road sign in Fig. 10 appear distinctly sharper.

The proposed model exhibits strong robustness across diverse scenarios. As an open-source solution, it delivers fusion quality comparable to or exceeding that of the two leading commercial focus-stacking software packages.

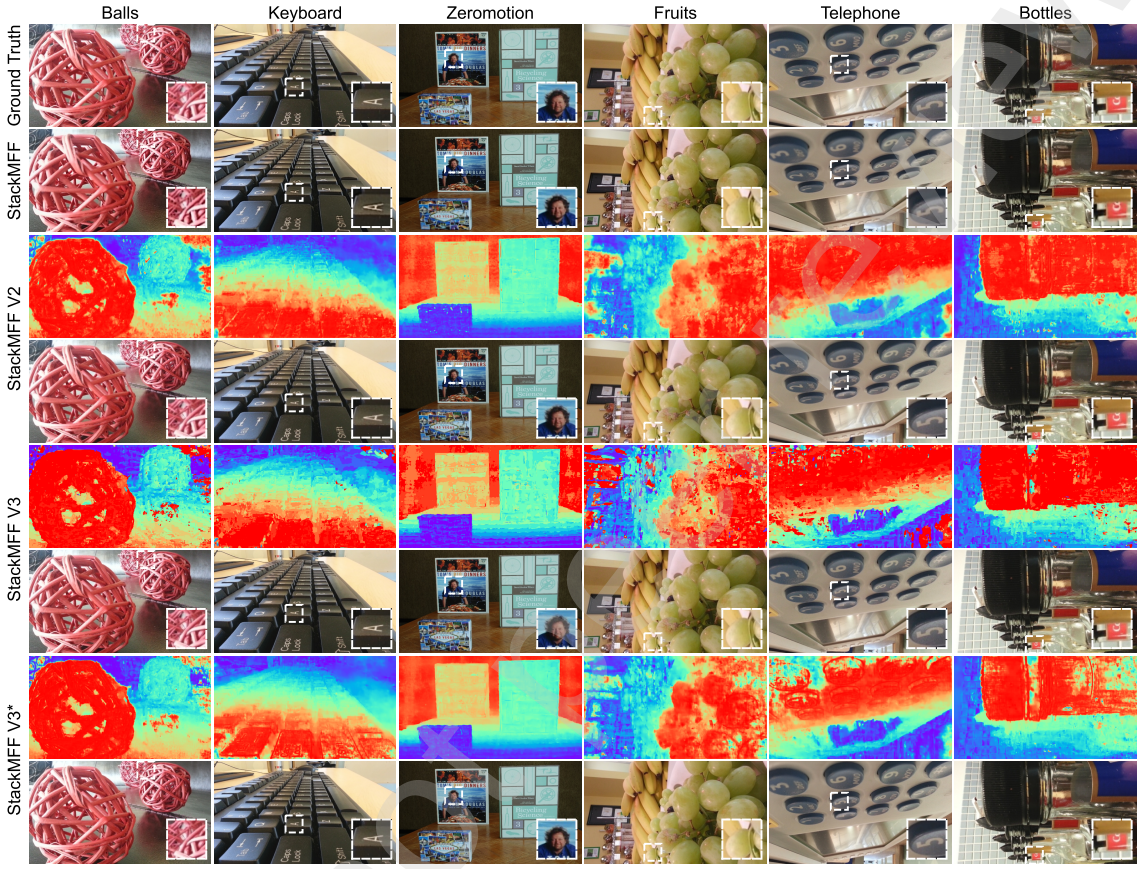

Figure 11: Comparison of fusion results produced by *StackMFF Series* on the Mobile Depth dataset (Suwajanakorn et al., 2015).

We further compare the fusion results and the corresponding focus maps of the *StackMFF Series* in Fig. 11 (*StackMFF* does not provide a focus map). From the fusion results, all three generations of the *StackMFF Series* demonstrate strong overall performance. Only the original *StackMFF* shows a slight disadvantage in fidelity when compared with the ground truth, while both *StackMFF V2* and *StackMFF V3/V3** produce high-quality fusion results.

From the focus maps generated by different models, the proposed *StackMFF V3*

exhibits several regions with noticeable differences from their surroundings, particularly in homogeneous background areas that have little impact on the overall fusion quality, whereas *StackMFF V2* produces smoother focus maps. However, *StackMFF V3* provides sharper and more accurate focus estimation in fine details and edge regions.

This difference mainly arises from their distinct task formulations: *StackMFF V3* formulates focus estimation as a classification problem, while *StackMFF V2/V3** treat it as a regression task. The latter's soft regression strategy introduces a smoothing term, resulting in smoother focus maps. Nevertheless, quantitative evaluations of the final fusion results suggest that the smoothing term does not always lead to positive gains.

4.3. Quantitative comparison

Table 2: Quantitative evaluation of fusion performance across four benchmark datasets.

Datasets	**Mobile Depth**		**Middlebury**		**FlyingThings3D**		**Road-MF**	
Method	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
DWT (Li et al., 1995)	0.9340	32.1651	0.8850	29.1761	0.9123	30.0074	0.9309	30.3456
DCT (Haghighat et al., 2011)	0.4720	17.2719	0.4520	13.9972	0.4603	15.0949	0.4856	16.9598
DTCWT (Hill et al., 2002)	0.9412	32.7641	0.8938	29.3763	0.9203	30.1512	0.9826	36.7138
NSCT (Yang et al., 2007)	0.9340	32.1651	0.8850	29.1761	0.9123	30.0074	0.9813	37.0137
IFCNN-MAX (Zhang et al., 2020)	0.7882	24.9863	0.9014	29.2064	0.9236	31.3069	0.8952	27.6907
U2Fusion (Xu et al., 2020)	0.3788	10.0482	0.3980	10.1318	0.4242	11.4382	0.3811	10.8764
SDNet (Zhang and Ma, 2021)	0.3961	12.1659	0.4399	14.0048	0.4457	14.5929	0.4144	13.0182
MFF-GAN (Zhang et al., 2021)	0.1797	7.1264	0.2962	10.1180	0.3006	11.9173	0.2559	9.3437
SwinFusion (Ma et al., 2022)	0.4381	12.4597	0.4254	13.4794	0.4313	14.1286	0.3945	11.9315
MUFusion (Cheng et al., 2023)	0.4819	18.7311	0.5809	19.7779	0.4762	19.8073	0.6821	19.6156
SwinMFF (Xie et al., 2024a)	0.3511	10.8676	0.4215	11.8564	0.3238	12.2809	0.4795	13.2869
DDBFusion (Zhang et al., 2025)	0.8365	26.3713	0.7181	23.7650	0.6984	23.0223	0.8065	24.4036
CCSR-Net (Zheng et al., 2025)	0.8485	28.3029	0.7207	24.1580	0.6918	24.4370	0.8682	27.5386
MCCSR-Net (Zheng et al., 2025)	0.8750	28.5764	0.8177	26.2944	0.7655	25.7952	0.9090	29.9696
Zerene Stacker - DMap (Undisclosed)	0.9399	33.5643	0.9067	30.5630	0.9139	30.9396	0.9678	33.8450
Zerene Stacker - PMax (Undisclosed)	0.9282	31.4065	0.9068	30.6395	0.9153	31.5620	0.9791	35.7715
Helicon Focus 8 - A (Undisclosed)	0.9469	32.9568	0.8968	24.7029	0.8993	24.8708	0.9203	24.1058
Helicon Focus 8 - B (Kozub and Shapoval, 2019)	0.9394	33.7037	0.8872	24.9958	0.8965	24.8155	0.9216	24.1157
Helicon Focus 8 - C (Undisclosed)	0.9424	31.8975	0.9028	24.8427	0.9012	25.0471	0.9336	24.3949
StackMFF (Xie et al., 2025c)	0.9536	32.6798	0.9284	31.0764	0.9483	32.5062	0.9692	33.0138
StackMFF V2 (Xie et al., 2025a)	0.9508	35.1017	0.9444	32.1810	0.9508	32.7506	0.9808	36.0976
StackMFF V3	**0.9657**	**36.3498**	**0.9510**	**32.3136**	**0.9607**	**33.3734**	0.9607	33.3734
StackMFF V3*	0.9519	35.1195	0.9459	32.1237	0.9441	32.1384	**0.9889**	**37.7828**
Enhancement (%)	+1.27%	+3.56%	+0.70%	+0.41%	+1.04%	+1.90%	+0.64%	+2.08%

Table 2 presents a comprehensive quantitative evaluation of fusion performance for various methods across four benchmark datasets, where ↑ indicates that a higher

value represents better performance.

From the results, traditional methods and commercial focus-stacking software generally achieve competitive performance, consistently maintaining relatively high SSIM and PSNR values across all datasets. Among the learning-based methods, earlier approaches such as U2Fusion, SDNet, MFF-GAN, and SwinFusion exhibit comparatively lower SSIM and PSNR scores, indicating susceptibility to fusion artifacts and less reliable structural preservation. Advanced learning-based methods, including DDBFusion, CCSR-Net, and MCCSR-Net, achieve improved performance but still fall short of the state-of-the-art traditional and hybrid approaches in several datasets.

Notably, the *StackMFF Series* demonstrates a substantial performance advantage over both traditional and learning-based methods. The original *StackMFF* already achieves high SSIM and PSNR values across all datasets, while *StackMFF V2* further improves performance, particularly on Middlebury and FlyingThings3D. The proposed *StackMFF V3* attains the highest scores in most cases, demonstrating superior structural fidelity and pixel-level accuracy. Moreover, the variant *StackMFF V3** provides additional improvements in specific scenarios, achieving the best PSNR on the Road-MF dataset.

Overall, the *StackMFF V3* series consistently outperforms both conventional and state-of-the-art learning-based methods, confirming its robustness and effectiveness in multi-focus image stack fusion across diverse datasets. The improvement percentages listed in the last row of Table 2 quantify the gains of *StackMFF V3* relative to the second-best performing methods across the corresponding datasets and metrics.

4.4. More analysis

4.4.1. Model efficiency comparison

We present a quantitative comparison of the average processing time (in seconds) of various methods across several benchmark datasets in Table 3. As shown in the table, the architectural modifications introduced in *StackMFF V3/V3** result in a modest increase in computational time compared to its predecessors, with processing times ranging from 0.31 s to 0.53 s across different datasets. Despite this increase, *StackMFF V3/V3** remains one of the few methods capable of performing fusion within the sub-second range on a GPU, significantly faster than most traditional CPU-based approaches. This ensures an efficient user experience and highlights its strong potential for deployment in time-critical applications.

Furthermore, we present a comparison of the proposed method with several state-of-the-art learning-based multi-focus image fusion methods in terms of model size, computational cost, one-shot fusion capability, and overall fusion quality based on

Table 3: Comparison of computational efficiency (seconds) across various methods and datasets.

Method	Device	Mobile Depth	Middlebury	FlyingThings3D	Road-MF
CVT (Guo et al., 2012)	CPU	48.00	31.37	37.87	78.14
DWT (Li et al., 1995)	CPU	5.34	8.62	6.75	4.62
DCT (Haghighat et al., 2011)	CPU	4.97	3.30	6.04	4.97
DTCWT (Hill et al., 2002)	CPU	11.44	9.40	14.70	12.82
NSCT (Yang et al., 2007)	CPU	231.84	165.13	133.84	217.03
IFCNN-MAX (Zhang et al., 2020)	GPU	0.55	0.50	0.78	0.55
U2Fusion (Xu et al., 2020)	CPU	41.04	35.90	45.10	104.96
SDNet (Zhang and Ma, 2021)	CPU	9.68	5.26	14.04	8.18
MFF-GAN (Zhang et al., 2021)	CPU	6.40	8.88	10.06	12.67
SwinFusion (Ma et al., 2022)	GPU	28.21	19.53	32.33	30.19
MUFusion (Cheng et al., 2023)	GPU	40.40	21.98	55.02	45.79
SwinMFF (Xie et al., 2024a)	GPU	27.97	18.23	34.05	55.04
DDBFusion (Zhang et al., 2025)	GPU	33.89	30.06	41.98	35.57
CCSR-Net (Zheng et al., 2025)	GPU	1.92	1.87	2.93	2.69
MCCSR-Net (Zheng et al., 2025)	GPU	16.03	10.31	17.22	17.24
Zerene Stacker - DMap (Undisclosed)	CPU	14.64	10.87	10.21	9.23
Zerene Stacker - PMax (Undisclosed)	CPU	6.36	5.13	6.50	7.24
Helicon Focus 8 - A (Undisclosed)	CPU	0.30	0.19	0.19	0.18
Helicon Focus 8 - B (Kozub and Shapoval, 2019)	CPU	0.38	0.26	0.25	0.24
Helicon Focus 8 - C (Undisclosed)	CPU	0.33	0.18	0.19	1.23
StackMFF (Xie et al., 2025c)	GPU	0.22	0.19	0.24	0.22
StackMFF V2 (Xie et al., 2025a)	GPU	0.14	0.08	0.11	0.11
StackMFF V3/V3*	GPU	0.52	0.31	0.52	0.53

Table 4: Comparison of learning-based methods in terms of model size (M), computational cost (FLOPs, in G) for fusing N-layer stacks, one-shot fusion capability, and fusion quality.

Method	Model Size (M)	FLOPs (G)	One-Shot Fusion	Good Results
IFCNN-MAX (Zhang et al., 2020)	0.08	8.54	×	×
U2Fusion (Xu et al., 2020)	0.66	86.4	×	×
SDNet (Zhang and Ma, 2021)	0.07	8.81	×	×
MFF-GAN (Zhang et al., 2021)	0.05	3.08	×	×
SwinFusion (Ma et al., 2022)	0.93	63.73	×	×
MUFusion (Cheng et al., 2023)	2.16	24.07	×	×
SwinMFF (Xie et al., 2024a)	41.25	22.38	×	×
CCSR-Net (Zheng et al., 2025)	0.02	1.5868	×	×
MCCSR-Net (Zheng et al., 2025)	0.04	2.57	×	×
DDBFusion (Zhang et al., 2025)	10.92	184.93	×	×
StackMFF (Xie et al., 2025c)	6.08	21.98	✓	✓
StackMFF V2 (Xie et al., 2025a)	0.05	2.75	✓	✓
StackMFF V3/V3*	2.74	2.04	✓	✓

subjective evaluation consistent with human visual perception, as summarized in Table 4. The computational cost is evaluated under a consistent setting, using image stacks of size 256×256 with two layers. The table also indicates whether each method supports stack-level (one-shot) fusion and whether it achieves satisfactory

fusion results.

The results show that the model size of *StackMFF V3/V3** (2.74 M) falls between that of the original *StackMFF* (6.08 M) and *StackMFF V2* (0.05 M), reflecting a moderate increase compared to *StackMFF V2*. Despite its larger architecture, *StackMFF V3/V3** achieves the lowest FLOPs (2.04 G) among the *StackMFF Series*, indicating superior computational efficiency. Moreover, unlike most other learning-based methods, *StackMFF Series* supports one-shot stack-level fusion while maintaining high-quality fusion results, demonstrating a favorable balance between model complexity, efficiency, and performance.

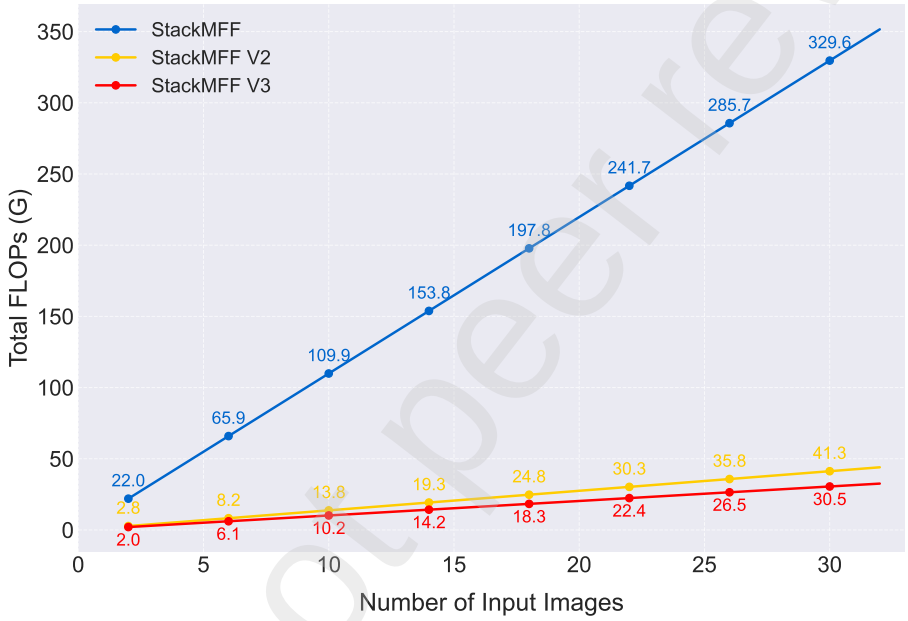

Figure 12: Computational cost of the *StackMFF Series* with increasing number of layers.

Fig. 12 compares the variation in FLOPs (G) among the three generations of the *StackMFF Series* as the number of layers in the input image stack increases. As shown in the figure, the computational cost of all three models grows approximately linearly with the number of layers, while the overall computational overhead decreases progressively across generations. Although the introduced *Pixel-wise Cross-layer Attention* module theoretically exhibits $\mathcal{O}(n^2)$ complexity, the number of layers is small relative to the total number of pixels per layer. Consequently, even when attention is computed along the stack depth at full resolution, its computational overhead remains minor compared to that of the intra-layer focus estimation

stage. Therefore, the proposed inter-layer modeling strategy achieves a more favorable trade-off between computational efficiency and performance compared to *StackMFF* and *StackMFF V2*.

4.4.2. Performance of small stack fusion

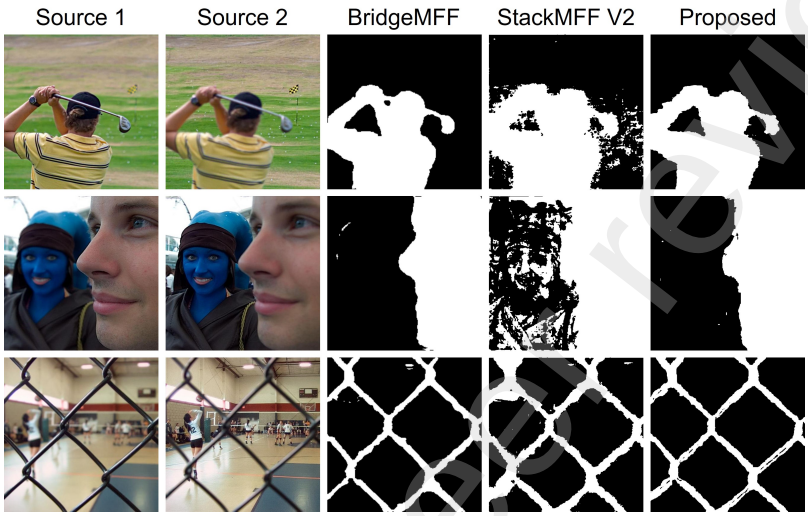

Figure 13: Fusion results of small multi-focus stacks.

Although the proposed *StackMFF V3* is primarily trained on multi-layer image stacks, it is also capable of effectively handling the fusion of small stacks. To demonstrate this capability, Fig. 13 presents the fusion results obtained from two-source multi-focus images selected from the Lytro dataset (Nejati et al., 2015).

In Fig. 13, for a more intuitive comparison, we show only the intermediate fusion result, i.e., the decision map. We compare the proposed *StackMFF V3* with the state-of-the-art pairwise fusion network BridgeMFF (Xie et al., 2025b) and the previous-generation model *StackMFF V2*. It can be observed that *StackMFF V3* achieves performance comparable to networks specifically designed for two-image fusion. Moreover, compared with *StackMFF V2*, the proposed model exhibits superior discrimination between background and foreground regions, producing fewer errors in challenging homogeneous areas.

4.4.3. Order-Independence verification

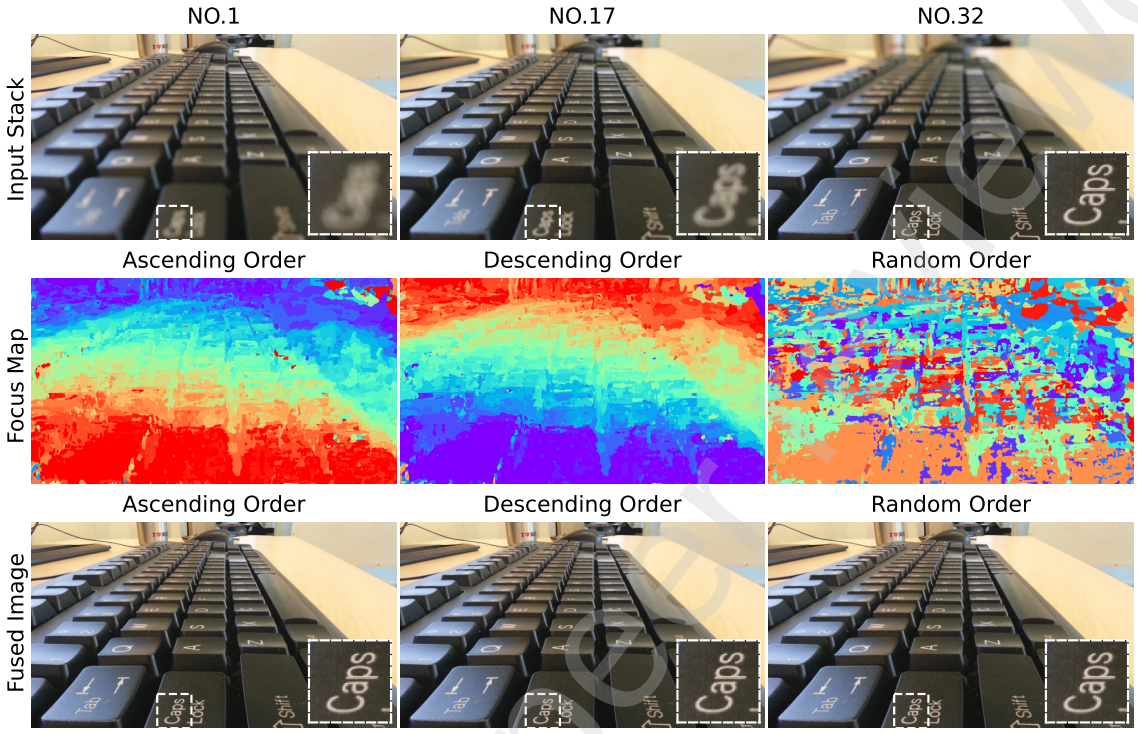

Figure 14: Focus maps and fusion results under different input orders.

StackMFF V2 (Xie et al., 2025a), due to its specific proxy supervision strategy, ensures the fidelity of the fused image but cannot handle unordered image stacks. To improve generality and further reduce the usage barrier, the proposed *StackMFF V3* is designed to be order-independent. As demonstrated in Fig. 14, *StackMFF V3* consistently produces high-quality fused images regardless of whether the input image stack is sequential, reversed, or randomly ordered, showing negligible sensitivity to input ordering. In this experiment, the test input stack consists of 32 images.

4.4.4. Anti-Interference verification

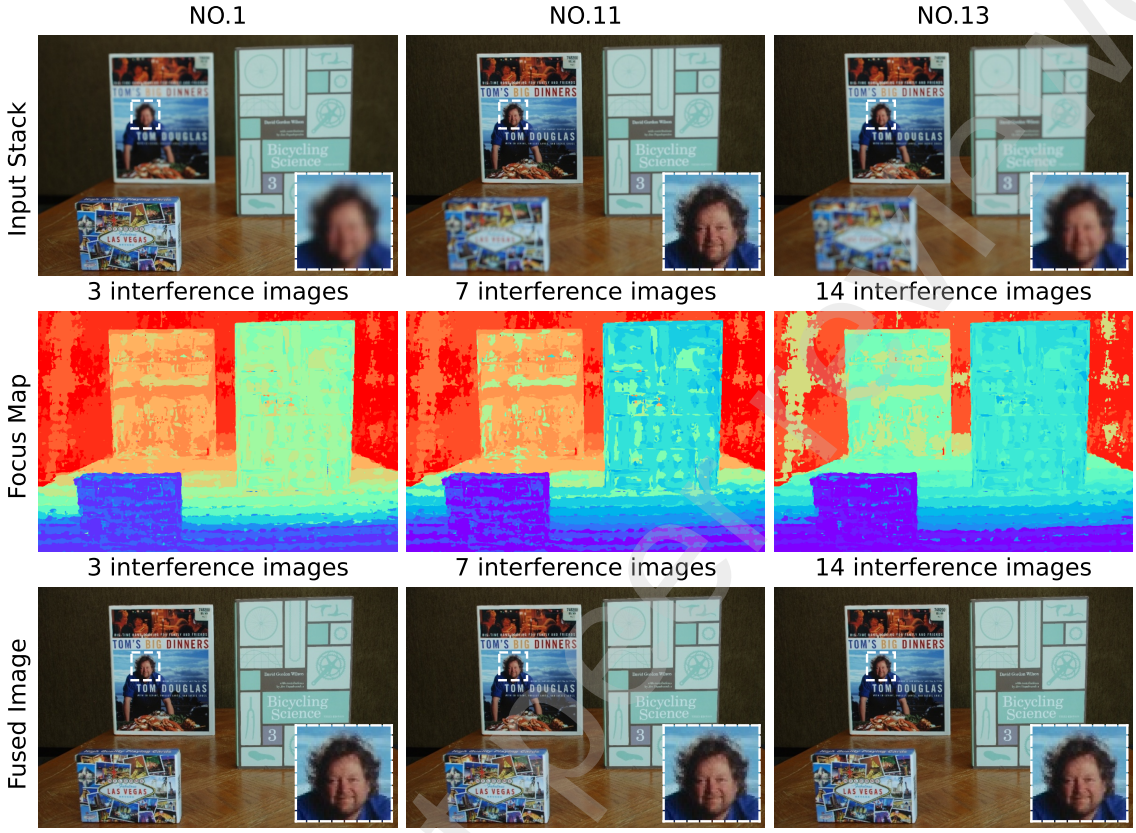

Figure 15: Comparison of fusion results under different interference conditions.

StackMFF V2 (Xie et al., 2025a), due to its specific soft regression strategy for focus maps, cannot effectively handle image stacks containing distracting layers. In contrast, the proposed *StackMFF V3* demonstrates a notable degree of robustness against such interference, similar to *StackMFF* (Xie et al., 2025c). To illustrate this property more intuitively, we construct an ordered image stack in which 3, 7, and 14 fully blurred distracting images are randomly inserted at different positions within a 14-layer stack. The fully blurred images are generated using Gaussian blurring with a kernel size of 31. As shown in Fig. 15, even when the number of distracting layers reaches 14—constituting half of the entire image stack—the model still produces satisfactory fusion results.

4.4.5. Fidelity verification

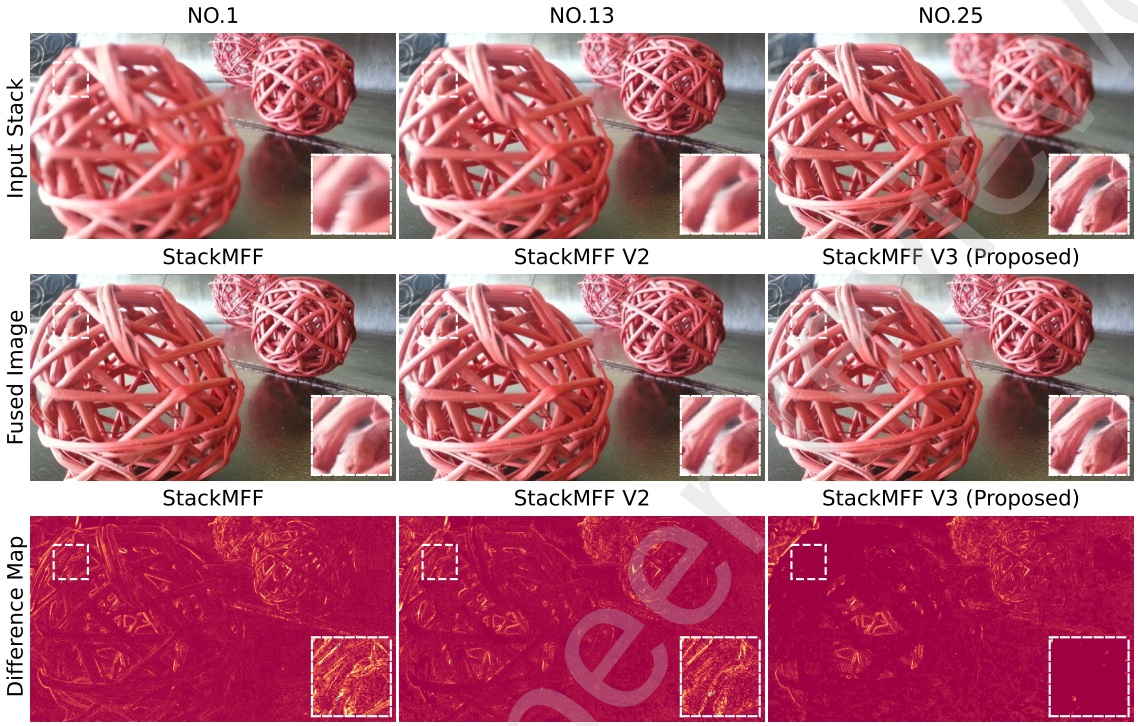

Figure 16: Comparison of fusion results under different interference conditions.

Unlike *StackMFF V2* (Xie et al., 2025a), *StackMFF* (Xie et al., 2025c) demonstrates a degree of robustness against interference and can handle unordered image stack inputs. Nevertheless, as an end-to-end model, the fidelity of its fused images may raise potential safety concerns in practical applications, particularly in scientific imaging contexts. To verify that *StackMFF V3* generates fused images with high fidelity suitable for scientific purposes, Fig. 16 presents difference maps between the fusion results of the *StackMFF Series* and the corresponding ground truth. Regions highlighted in yellow indicate larger pixel-level discrepancies. The results demonstrate that the proposed *StackMFF V3* exhibits the smallest deviations from the ground truth in critical regions, achieving superior pixel-level fidelity compared to its predecessors.

4.4.6. Generalizability verification

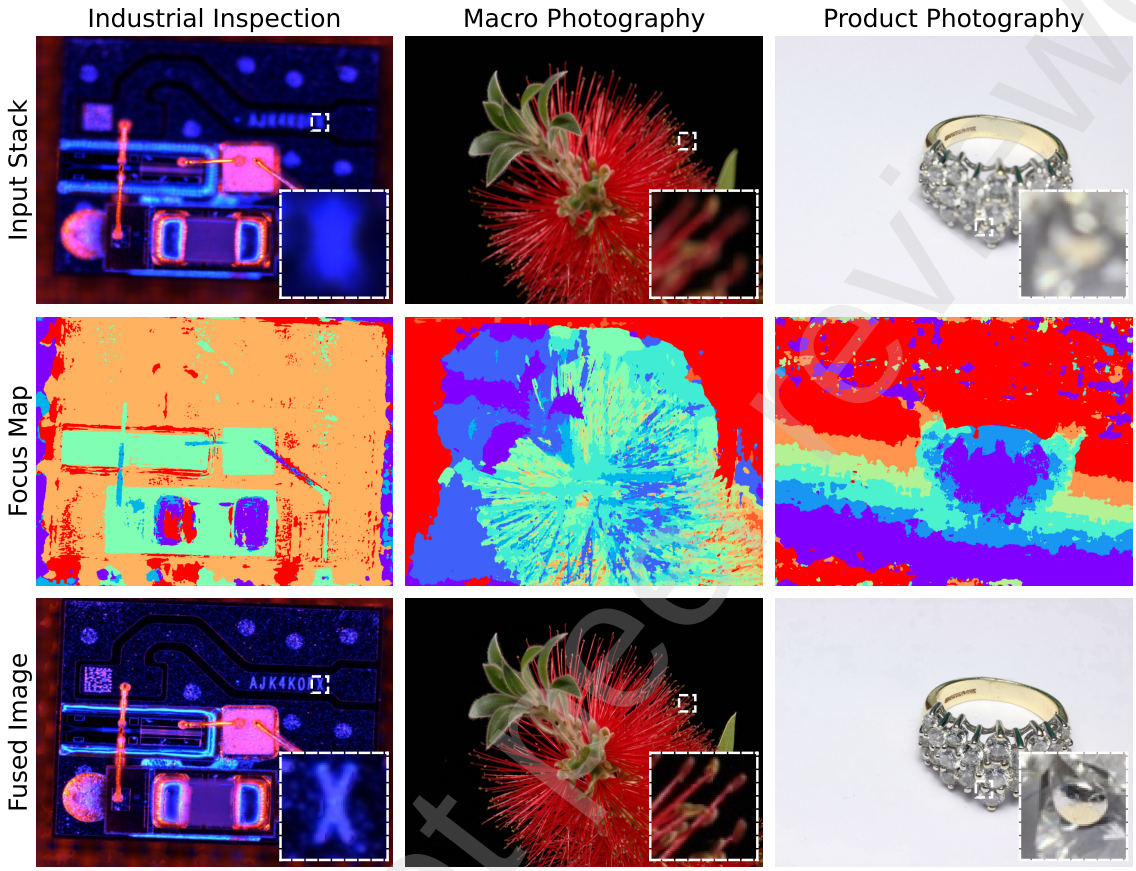

Figure 17: Representative fusion results across diverse application domains.

To validate the generalization capability of the proposed *StackMFF V3* and demonstrate its potential applicability across diverse domains, several representative scenarios are illustrated in Fig. 17, including industrial inspection, macro photography, and product photography. As shown in the figure (with only one image displayed per stack for each scenario), despite being scarcely trained on domain-specific images, the network consistently produces high-quality fusion results across these varied application contexts.

Table 5: Ranking of different methods based on quantitative evaluation metrics across four benchmark 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)	24	24	24.0	24	24	24.0	24	23	23.5	24	24	24.0	23.9
U2Fusion (Xu et al., 2020)	22	23	22.5	23	23	23.0	22	24	23.0	23	23	23.0	22.9
SwinMFF (Xie et al., 2024a)	23	22	22.5	22	22	22.0	23	22	22.5	20	20	20.0	21.8
SwinFusion (Ma et al., 2022)	20	20	20.0	21	21	21.0	21	21	21.0	22	22	22.0	21.0
SDNet (Zhang and Ma, 2021)	21	21	21.0	20	19	19.5	20	20	20.0	21	21	21.0	20.4
DCT (Haghighat et al., 2011)	19	19	19.0	19	20	19.5	19	19	19.0	19	19	19.0	19.1
MUFusion (Cheng et al., 2023)	18	18	18.0	18	18	18.0	18	18	18.0	18	18	18.0	18.0
DDBFusion (Zhang et al., 2025)	16	16	16.0	17	17	17.0	16	17	16.5	17	14	15.5	16.3
CCSR-Net (Zheng et al., 2025)	15	15	15.0	16	16	16.0	17	16	16.5	16	13	14.5	15.5
MCCSR-Net (Zheng et al., 2025)	14	14	14.0	15	12	13.5	15	12	13.5	14	11	12.5	13.4
Helicon Focus 8 - B (Kozub and Shapoval, 2019)	9	4	6.5	12	13	12.5	14	15	14.5	12	16	14.0	11.9
Helicon Focus 8 - A (Undisclosed)	5	6	5.5	9	15	12.0	13	14	13.5	13	17	15.0	11.5
IFCNN-MAX (Zhang et al., 2020)	17	17	17.0	8	9	8.5	5	6	5.5	15	12	13.5	11.1
Helicon Focus 8 - C (Undisclosed)	6	12	9.0	7	14	10.5	12	13	12.5	10	15	12.5	11.1
DWT (Li et al., 1995)	11	10	10.5	13	10	11.5	10	10	10.0	11	10	10.5	10.6
NSCT (Yang et al., 2007)	11	10	10.5	13	10	11.5	10	10	10.0	3	2	2.5	8.6
CVT (Guo et al., 2012)	10	9	9.5	11	8	9.5	7	9	8.0	6	5	5.5	8.1
Zerene Stacker - PMax (Undisclosed)	13	13	13.0	5	5	5.0	8	5	6.5	5	6	5.5	7.5
Zerene Stacker - DMap (Undisclosed)	8	5	6.5	6	6	6.0	9	7	8.0	8	7	7.5	7.0
DTCWT (Hill et al., 2002)	7	7	7.0	10	7	8.5	6	8	7.0	2	3	2.5	6.3
StackMFF (Xie et al., 2025c)	2	8	5.0	4	4	4.0	3	3	3.0	7	9	8.0	5.0
StackMFF V2 (Xie et al., 2025a)	4	3	3.5	3	2	2.5	2	2	2.0	4	4	4.0	3.0
StackMFF V3	1	1	1.0	1	1	1.0	1	1	1.0	9	8	8.5	2.9
StackMFF V3*	3	2	2.5	2	3	2.5	4	4	4.0	1	1	1.0	2.5

4.4.7. Statistical significance analysis

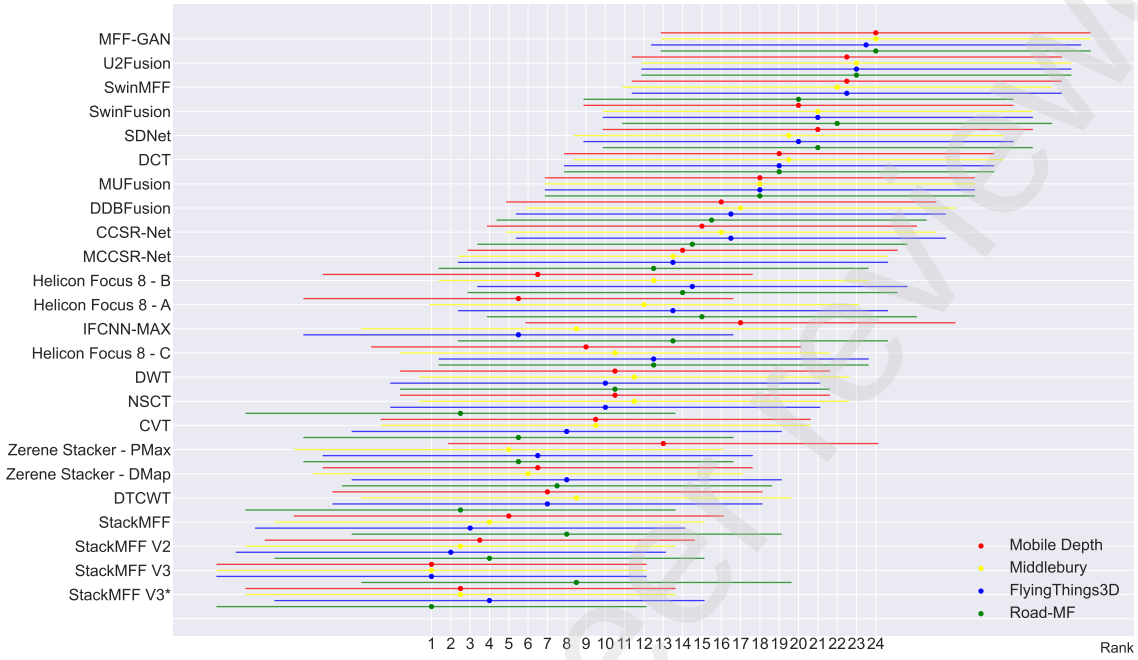

Figure 18: Results of the Nemenyi post-hoc test following the Friedman test for multiple method comparisons. The CD bars represent the critical difference at $\alpha = 0.05$ significance level.

In Table 5, we present the ranking results of various MFF algorithms on four benchmark datasets based on SSIM and PSNR. The statistical significance of performance differences is evaluated using the Nemenyi post-hoc test, as shown in Fig. 18, where non-overlapping confidence intervals indicate statistically significant differences between methods.

The results demonstrate that the proposed *StackMFF V3* achieves the best average rankings of 2.5/2.9 (*StackMFF V3**/*StackMFF V3*) across the benchmark datasets, outperforming the previous two generations of the *StackMFF Series*. This indicates that the current model offers superior performance across multiple aspects, including fusion quality and generalization capability. Moreover, *StackMFF V3** ranks first overall, further confirming the architectural advantages of *StackMFF V3* relative to *StackMFF V2*.

Several additional observations can be drawn from Fig. 18 and Table 5: (1) The majority of deep learning–based MFF methods, including U2Fusion, MFF-GAN,

SwinFusion, and SDNet, consistently achieve relatively low average rankings (18–24) across the four benchmark datasets. This indicates limited generalization and robustness of pairwise or conventional learning-based approaches when applied to multi-layer image stack fusion tasks. Only a few methods, such as IFCNN-MAX and MCCSR-Net, show competitive performance on individual datasets, but their overall stability remains insufficient. (2) Traditional methods such as DTCWT, DWT, and NSCT maintain stable performance across the datasets, with average rankings mostly between 6 and 11. Notably, DTCWT outperforms certain commercial focus-stacking software in multiple datasets, highlighting that classical design principles, such as multi-scale or frequency-domain processing, continue to offer valuable insights for future MFF research. (3) The *Helicon Focus 8* and *Zerene Stacker* achieve varying rankings depending on the dataset, indicating limited robustness across diverse scenarios. While some tools perform well on specific datasets, their overall consistency is inferior to that of the *StackMFF Series*. (4) The *StackMFF Series*, including *StackMFF*, *StackMFF V2*, and *StackMFF V3/V3**, consistently rank at the top across all datasets, with *StackMFF V3** achieving the best overall ranking (2.5) and *StackMFF V3* closely following (2.9). These results indicate that the proposed architecture effectively enhances fusion quality, generalization capability, and robustness, making it the most reliable approach for multi-focus image stack fusion tasks.

5. Discussion and future work

In this work, we propose the latest generation of the *StackMFF Series*—*StackMFF V3*. To the best of our knowledge, this is the first general image stack fusion network, which demonstrates strong applicability across both image-pair and image-stack fusion tasks in various domains. Qualitative and quantitative results indicate that, compared with *StackMFF*, it provides higher pixel-level fidelity, and compared with *StackMFF V2*, it can handle unordered image stacks and exhibits stronger robustness against interference. Furthermore, it inherits the advantages of the *StackMFF Series*, such as one-shot fusion capability, support for a variable number of input layers, and full-resolution input, making it one of the most powerful and versatile image stack fusion methods to date. As of the release of *StackMFF V3*, we believe that the most critical challenges in multi focus image stack fusion have been addressed, representing a breakthrough from "0 to 1." We encourage future researchers to build upon *StackMFF V3*, as its design principles align well with real-world application requirements and may have a positive impact on future production environments.

Despite the impressive potential demonstrated by *StackMFF V3*, we have identified several issues that future work could address: (1) Currently, all multi focus image fusion algorithms, including the *StackMFF Series*, assume that the input image stack has been pre-registered and aligned. Although registration can be achieved through feature matching and simple geometric transformations, this may increase the technical barrier for non-specialist users. To mitigate this, our code repository includes an implementation for image stack registration to help non-specialists quickly get started. Future algorithms could consider integrating the registration process as a learnable component of the network to achieve higher registration quality, thereby improving fusion results and reducing registration-induced artifacts. (2) During the registration step, image cropping is generally required to remove black borders, which inevitably leads to the loss of field of view and reduction of image information. Future research could explore full-frame multi focus image stack fusion, as inspired by recent studies in video stabilization (Peng et al., 2024; Zhao et al., 2023), which may be extended to the multi focus image fusion domain and yield positive outcomes. (3) The current design of the *StackMFF Series* follows a causal inference paradigm—that is, the quality of inter-layer modeling depends on the accuracy of intra-layer focus estimation. However, there is currently no mechanism to re-optimize the intra-layer estimation based on inter-layer modeling results. Future work could consider iterative optimization strategies in a recurrent manner to further improve fusion quality, similar to the methods proposed in (Sun et al., 2021; Wang et al., 2024). (4) The current generation employs PFMLP as the intra-layer focus estimation backbone. Although it maintains a relatively low computational cost, communication latency leads to increased inference time in practice. Future research could investigate more efficient visual backbones to ensure shorter inference delays, even on low-power or embedded platforms.

6. Conclusion

In this work, we proposed *StackMFF V3*, the first general multi-focus image fusion network, which was completely redesigned from the training strategy to the fusion framework. The main contributions of this work are as follows: (1) We introduced a Pyramid Fusion MLP as the visual backbone for intra-layer focus estimation, which models long-range intra-layer dependencies to balance efficiency and fusion quality. (2) We proposed a *Pixel-wise Cross-layer Attention* module that efficiently captures cross-layer relations without relying on focus order, ensuring superior modeling efficiency and robustness. (3) Finally, we reformulated the multi focus image stack fusion task as a pixel-wise multi-class classification problem, enabling the network to

directly predict per-pixel focus maps for generating the fused image while preserving pixel-level fidelity.

The proposed *StackMFF V3* is currently the only image fusion network capable of one-shot processing of full-resolution image stacks of arbitrary order and variable numbers of layers, demonstrating strong robustness against interfering layers. It outperforms existing traditional algorithms, learning-based methods, and commercial focus-stacking software across multiple dimensions, including versatility, generalization capability, fusion quality, efficiency, and application potential. This work marks the maturation of the *StackMFF Series* and opens new possibilities for real-world imaging applications across diverse domains, laying a solid foundation for the development of next-generation general fusion networks.

CRediT authorship contribution statement

Manuscript cannot include author Identifiers. These will be added after the article is accepted.

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.

Acknowledgements

Manuscript cannot include author Identifiers. These will be added after the article is accepted.

Data availability

Data will be made available on request.

References

Agarwala, A., Dontcheva, M., Agrawala, M., Drucker, S., Colburn, A., Curless, B., Salesin, D., Cohen, M., 2004. Interactive digital photomontage, in: ACM SIGGRAPH 2004 Papers, pp. 294–302.

Araujo, A., Ponce, J., Mairal, J., 2023. Towards real-world focus stacking with deep learning. arXiv preprint arXiv:2311.17846 .

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

Choi, D., Pazylbekova, A., Zhou, W., van Beek, P., 2017. Improved image selection for focus stacking in digital photography, in: 2017 IEEE International Conference on Image Processing (ICIP), IEEE. pp. 2761–2765.

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.

Deng, X., Liu, X., Xu, T., Liu, X., Gan, T., Lu, C., Zhou, C., Wang, P., Lei, Y., Ye, X., 2025. Endoscopic depth-of-field expansion via cascaded network with two-streamed multi-scale fusion, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 137–146.

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

Haghighat, M.B.A., Aghagolzadeh, A., Seyedarabi, H., 2011. Multi-focus image fusion for visual sensor networks in dct domain. Computers and Electrical Engineering 37, 789–797. doi:https://doi.org/10.1016/j.compeleceng.2011.04.016. special Issue on Image Processing.

Han, X., Li, R., Wang, B., Lin, Z., 2024. Defect identification of bare printed circuit boards based on bayesian fusion of multi-scale features. PeerJ Computer Science 10, e1900.

Hasinoff, S.W., Kutulakos, K.N., 2011. Light-efficient photography. IEEE Transactions on Pattern Analysis and Machine Intelligence 33, 2203–2214.

Häusler, G., 1972. A method to increase the depth of focus by two step image processing. Optics Communications 6, 38–42.

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

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

Huang, Q., Jie, Z., Ma, L., Shen, L., Lai, S., 2025. A pyramid fusion mlp for dense prediction. IEEE Transactions on Image Processing 34, 455–467. doi:10.1109/TIP.2025.3526054.

Jie, Y., Xu, Y., Li, X., Tan, H., 2024. Tsjnet: A multi-modality target and semantic awareness joint-driven image fusion network. arXiv preprint arXiv:2402.01212 .

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

Kuthirummal, S., Nagahara, H., Zhou, C., Nayar, S.K., 2010. Flexible depth of field photography. IEEE transactions on pattern analysis and machine intelligence 33, 58–71.

Lee, C., Kim, J., Lee, S., Jung, J., Cho, Y., Kim, T., Jo, T., Lee, M., Jang, M., 2024. Blind image deblurring with noise-robust kernel estimation, in: European Conference on Computer Vision, Springer. pp. 188–204.

Li, H., Manjunath, B., Mitra, S.K., 1995. Multisensor image fusion using the wavelet transform. Graphical models and image processing 57, 235–245.

Li, L., Lv, M., Jia, Z., Jin, Q., Liu, M., Chen, L., Ma, H., 2023. An effective infrared and visible image fusion approach via rolling guidance filtering and gradient saliency map. Remote Sensing 15. URL: https://www.mdpi.com/2072-4292/15/10/2486, doi:10.3390/rs15102486.

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

Li, L., Zhao, X., Hou, H., Zhang, X., Lv, M., Jia, Z., Ma, H., 2024a. Fractal dimension-based multi-focus image fusion via coupled neural p systems in nsct domain. Fractal and Fractional 8, 554.

Li, M., Pei, R., Zheng, T., Zhang, Y., Fu, W., 2024b. Fusiondiff: Multi-focus image fusion using denoising diffusion probabilistic models. Expert Systems with Applications 238, 121664.

Li, X., Li, X., Tan, H., Li, J., 2024c. 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, Y., Fang, H., Lei, X., Wang, Q., Hu, G., Dong, J., Li, Z., Lin, J., Liu, Q., Song, X., 2025b. Real-world defocus deblurring via score-based diffusion models. Scientific Reports 15, 22942.

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

Lou, H., Teng, M., Yang, Y., Shi, B., 2023. All-in-focus imaging from event focal stack, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 17366–17375.

Ma, J., Le, Z., Tian, X., Jiang, J., 2021. Smfuse: Multi-focus image fusion via self-supervised mask-optimization. IEEE Transactions on Computational 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 Journal of Automatica Sinica 9, 1200–1217.

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. Information Fusion 25, 72–84.

Peng, Z., Ye, X., Zhao, W., Liu, T., Sun, H., Li, B., Cao, Z., 2024. 3d multi-frame fusion for video stabilization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7507–7516.

Pyo, J., Lee, K., Shim, E.S., Choi, W., Yun, J., Jung, T., Lee, K., Kim, S., Lee, C., Baek, S., et al., 2021. A smart dual pixel technology for accurate and all-directional auto focus in cmos image sensors, in: Proc. of the International Image Sensors Workshop.

Quan, Y., Wan, X., Tang, Z., Liang, J., Ji, H., 2025. Multi-focus image fusion via explicit defocus blur modelling, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 6657–6665.

Rai, M.R., Rosen, J., 2021. Depth-of-field engineering in coded aperture imaging. Optics Express 29, 1634–1648.

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.

Sharma, R., Perry, S., Cheng, E., 2023. Light field view synthesis using the focal stack and all-in-focus image. Sensors 23, 2119.

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.

Song, Y., Xie, X., Guo, B., Xiong, X., Li, P., 2025. Mlp-mff: Lightweight pyramid fusion mlp for ultra-efficient end-to-end multi-focus image fusion. Sensors 25. URL: https://www.mdpi.com/1424-8220/25/16/5146, doi:10.3390/s25165146.

Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., Liu, Y., 2024. Roformer: Enhanced transformer with rotary position embedding. Neurocomputing 568, 127063. URL: https://www.sciencedirect.com/science/article/pii/S0925231223011864, doi:https://doi.org/10.1016/j.neucom.2023.127063.

Sun, J., Shen, Z., Wang, Y., Bao, H., Zhou, X., 2021. Loftr: Detector-free local feature matching with transformers, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 8922–8931.

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, 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, Y., He, X., Peng, S., Tan, D., Zhou, X., 2024. Efficient loftr: Semi-dense local feature matching with sparse-like speed, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 21666–21675.

Wu, A., Lu, R., Li, M., 2024. A novel focus measure algorithm for three-dimensional microscopic vision measurement based on focus stacking. Sensors and Actuators A: Physical 376, 115657.

Xiao, T., Liu, Y., Zhou, B., Jiang, Y., Sun, J., 2018. Unified perceptual parsing for scene understanding, in: Proceedings of the European conference on computer vision (ECCV), pp. 418–434.

Xie, X., Guo, B., He, S., Gu, Y., Li, Y., Li, P., 2025a. One-shot multi-focus image stack fusion via focal depth regression. Engineering Applications of Artificial Intelligence 162, 112667. URL: https://www.sciencedirect.com/science/article/pii/S0952197625026983, doi:https://doi.org/10.1016/j.engappai.2025.112667.

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. The Visual Computer , 1–24.

Xie, X., Guo, B., Li, P., He, S., Zhou, S., 2025b. Multi-focus image fusion with visual state space model and dual adversarial learning. Computers and Electrical Engineering 123, 110238.

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., Qingyan, J., Chen, D., Guo, B., Li, P., Zhou, S., 2025c. Stackmff: end-to-end multi-focus image stack fusion network. Applied Intelligence 55, 503. URL: https://doi.org/10.1007/s10489-025-06383-8, doi:10.1007/s10489-025-06383-8.

Xu, H., Ma, J., Jiang, J., Guo, X., Ling, H., 2020. U2fusion: A unified unsupervised image fusion network. IEEE Transactions on Pattern Analysis and Machine Intelligence 44, 502–518.

Yan, T., Hu, Z., Qian, Y., Qiao, Z., Zhang, L., 2020. 3d shape reconstruction from multifocus image fusion using a multidirectional modified laplacian operator. Pattern Recognition 98, 107065.

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.

Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H., 2024. Depth anything v2. arXiv preprint arXiv:2406.09414 .

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. Information Fusion 66, 40–53.

Zhang, H., Ma, J., 2021. Sdnet: A versatile squeeze-and-decomposition network for real-time image fusion. International Journal of Computer Vision 129, 2761–2785.

Zhang, K., Ren, W., Luo, W., Lai, W.S., Stenger, B., Yang, M.H., Li, H., 2022. Deep image deblurring: A survey. International Journal of Computer Vision 130, 2103–2130.

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

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. Information Fusion 114, 102655.

Zhao, W., Li, X., Peng, Z., Luo, X., Ye, X., Lu, H., Cao, Z., 2023. Fast full-frame video stabilization with iterative optimization, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 23534–23544.

Zheng, K., Cheng, J., Liu, Y., 2025. Unfolding coupled convolutional sparse representation for multi-focus image fusion. Information Fusion 118, 102974. URL: https://www.sciencedirect.com/science/article/pii/S1566253525000478, doi:https://doi.org/10.1016/j.inffus.2025.102974.

Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A., Torralba, A., 2019. Semantic understanding of scenes through the ade20k dataset. International Journal of Computer Vision 127, 302–321.

Zhou, C., Miau, D., Nayar, S.K., 2012. Focal sweep camera for space-time refocusing .