sensors

MDPI

Article

MLP-MFF: Lightweight Pyramid Fusion MLP for Ultra-Efficient End-to-End Multi-Focus Image Fusion

Yuze Song [1,2,†], Xinzhe Xie [3,†], Buyu Guo [4,5,*], Xiaofei Xiong [6] and Peiliang Li [3,5,7]

[1] National Ocean Technology Center, Tianjin 300112, China; notcsongyuze@163.com
[2] Key Laboratory of Ocean Observation Technology, Ministry of National Resources, Tianjin 300112, China
[3] State Key Laboratory of Ocean Sensing, Ocean College, Zhejiang University, Zhoushan 316021, China; xiexinzhe@zju.edu.cn (X.X.); lipeiliang@zju.edu.cn (P.L.)
[4] Donghai Laboratory, Zhoushan 316021, China
[5] Hainan Institute, Zhejiang University, Sanya 572025, China
[6] South China Sea Ecological Center, Ministry of Natural Resources, Guangzhou 510275, China; xiongxiaofei124@126.com
[7] Hainan Observation and Research Station of Ecological Environment and Fishery Resource in Yazhou Bay, Sanya 572024, China
* Correspondence: guobuyuwork@163.com
† These authors contributed equally to this work.

Abstract

Limited depth of field in modern optical imaging systems often results in partially focused images. Multi-focus image fusion (MFF) addresses this by synthesizing an all-in-focus image from multiple source images captured at different focal planes. While deep learning-based MFF methods have shown promising results, existing approaches face significant challenges. Convolutional Neural Networks (CNNs) often struggle to capture long-range dependencies effectively, while Transformer and Mamba-based architectures, despite their strengths, suffer from high computational costs and rigid input size constraints, frequently necessitating patch-wise fusion during inference—a compromise that undermines the realization of a true global receptive field. To overcome these limitations, we propose MLP-MFF, a novel lightweight, end-to-end MFF network built upon the Pyramid Fusion Multi-Layer Perceptron (PFMLP) architecture. MLP-MFF is specifically designed to handle flexible input scales, efficiently learn multi-scale feature representations, and capture critical long-range dependencies. Furthermore, we introduce a Dual-Path Adaptive Multi-scale Feature-Fusion Module based on Hybrid Attention (DAMFFM-HA), which adaptively integrates hybrid attention mechanisms and allocates weights to optimally fuse multi-scale features, thereby significantly enhancing fusion performance. Extensive experiments on public multi-focus image datasets demonstrate that our proposed MLP-MFF achieves competitive, and often superior, fusion quality compared to current state-of-the-art MFF methods, all while maintaining a lightweight and efficient architecture.

Keywords: deep learning; multi-focus image fusion; lightweight network; end to end; Multi-Layer Perceptron

Academic Editors: Honggang Chen and Xianfeng Ou

Received: 22 July 2025
Revised: 13 August 2025
Accepted: 14 August 2025
Published: 19 August 2025

Citation: Song, Y.; Xie, X.; Guo, B.; Xiong, X.; Li, P. MLP-MFF: Lightweight Pyramid Fusion MLP for Ultra-Efficient End-to-End Multi-Focus Image Fusion. *Sensors* **2025**, *25*, 5146. https://doi.org/10.3390/s25165146

1. Introduction

Modern optical imaging systems inherently face the challenge of limited depth of field (DoF) when capturing 3D scenes. This means that in a single exposure, only objects within a specific focal plane appear sharp, while areas outside this plane become progressively

blurred with increasing defocus. This physical limitation severely constrains the completeness and visual quality of image information, particularly impacting applications with stringent requirements for all-in-focus clarity, such as microscopic imaging [1,2], machine vision [3] and industrial inspection [4]. To overcome the drawback of insufficient DoF in single images, multi-focus image fusion (MFF) technology has emerged. This technique aims to combine a series of partially focused images of the same scene, acquired at different focal settings, into a single image that maintains good focus across the entire spatial range, thereby significantly expanding the information content and usability of the image [5].

To achieve all-in-focus image construction, researchers have proposed various MFF algorithms. These algorithms can be broadly categorized into three types based on their processing principles and image domains: spatial domain-based fusion methods, transform domain-based fusion methods and the more recently developed deep learning-based fusion methods.

Spatial domain methods [6–10] directly perform focus analysis and information selection at the pixel or local region level. These methods typically rely on hand-crafted focus measure operators and directly select or weighted-average pixels from source images based on the evaluation results. Their advantages include simple implementation and lower computational cost. However, they often produce artifacts at focus boundaries and are sensitive to noise and slight registration errors in the source images.

In contrast, transform domain methods [11–18] convert images into specific coefficient domains (e.g., wavelet or Laplacian pyramid domains) for feature extraction and fusion, then inverse transform them back to the spatial domain. This approach better separates high-frequency and low-frequency information, helping to preserve details like textures and edges in the fused image. Despite these advantages, transform domain methods generally involve higher computational complexity and significantly longer fusion times.

With the rapid advancements in deep learning, its data-driven powerful feature-learning capabilities have offered new solutions to the challenges faced by traditional fusion methods. Deep learning-based MFF methods can generally be summarized into two main paradigms. The first category consists of decision map-based deep learning methods. These methods draw inspiration from traditional spatial domain fusion approaches, utilizing deep neural networks to automatically learn and generate a focus decision map, indicating from which source image each pixel or region should be extracted. This approach often yields more interpretable results and can effectively leverage insights from existing traditional methods, for example, by integrating morphological processing or conditional random fields in post-processing to optimize the decision map.

The other paradigm is end-to-end deep learning regression methods. Unlike decision map-based methods, these directly take a series of defocused source images as input and, through an end-to-end trained deep neural network, directly regress to the final all-in-focus fused image. By constructing complex nonlinear mapping relationships, this method automatically learns the entire fusion process from input to output, without explicit focus measure calculations or fusion rule design. This holds the promise of generating higher-quality, more natural fusion results and simplifying the overall fusion workflow. The end-to-end regression paradigm, with its powerful adaptive learning and feature representation capabilities, has demonstrated significant advantages in suppressing artifacts, preserving details, enhancing the visual quality of fused images, and handling complex scenes, making it a current research hot topic and development trend in the MFF field.

However, early deep learning-based end-to-end fusion methods largely relied on Convolutional Neural Networks (CNNs) [19,20]. While CNNs excel at extracting local features, their inherent limitations in capturing long-range dependencies and global contextual information have become increasingly apparent in complex fusion tasks. This

has sparked interest in alternative deep learning architectures. Recently, models such as Transformers [21,22], known for their superior long-range interaction modeling capabilities, and Diffusion models [23], which offer powerful generative capabilities for high-quality image synthesis, are emerging as promising avenues for further advancing MFF. Although these new paradigms aim to overcome the limitations of CNNs and pave the way for more robust and visually compelling fusion results, they also face their own challenges. While Transformer models excel at capturing global dependencies, their core self-attention mechanism leads to substantial computational and memory overhead, especially when processing high-resolution images, which limits their practical application. Furthermore, their efficiency in capturing local image details and textures may be lower compared to CNNs. Diffusion models exhibit extraordinary performance in image generation, but their inference speed is extremely slow, requiring hundreds or even thousands of iterative steps to generate an image, which makes it difficult to meet real-time or near real-time fusion demands. Simultaneously, their training and inference processes consume significant computational resources, and as generative models, their goal is to create content rather than precisely select and integrate information from source images, which may lead to inconsistencies in the fusion process. The recently emerging Mamba model, as a novel architecture based on State Space Models (SSM), has shown great potential in sequence modeling. Mamba, with its linear computational complexity and long-range dependency modeling capability, is expected to overcome the computational efficiency bottlenecks of Transformer models when processing long sequence data, offering a more efficient solution. However, due to architectural limitations, some of the aforementioned models typically only support fixed input image sizes [24,25]. When higher-resolution images need to be fused during the inference phase, they can only be processed by dividing them into patches and then merging them. This approach cannot truly achieve full-image long-range interaction, and such a compromise contradicts the initial purpose for which these architectures were introduced into this field.

Recently, models composed solely of Multi-Layer Perceptrons (MLPs) have shown new potential in the vision domain [26,27], also becoming an emerging trend, particularly with the proposal of some MLP architectures that support flexible input scales, which are expected to solve the aforementioned problems. This paper aims to delve into the application potential of this neural network architecture in the task of MFF. Our research seeks to demonstrate that a carefully designed lightweight MLP-based network can effectively address the challenges faced by MFF (Figure 1), thereby offering a novel perspective and a more competitive alternative to existing methods. The main contributions of this paper are threefold:

1. We propose a lightweight, end-to-end MFF network based on Pyramid Fusion MLP, which can handle flexible input scales, learn multi-scale feature representations, and capture long-range dependencies.
2. We propose a Dual-Path Adaptive Multi-scale Feature-Fusion Module (DAMFFM-HA) that integrates hybrid attention mechanisms and adaptive weight allocation to effectively fuse multi-scale features, thereby enhancing fusion performance.
3. Extensive experiments on publicly available datasets demonstrate the effectiveness and superiority of our proposed method compared to state-of-the-art MFF methods.

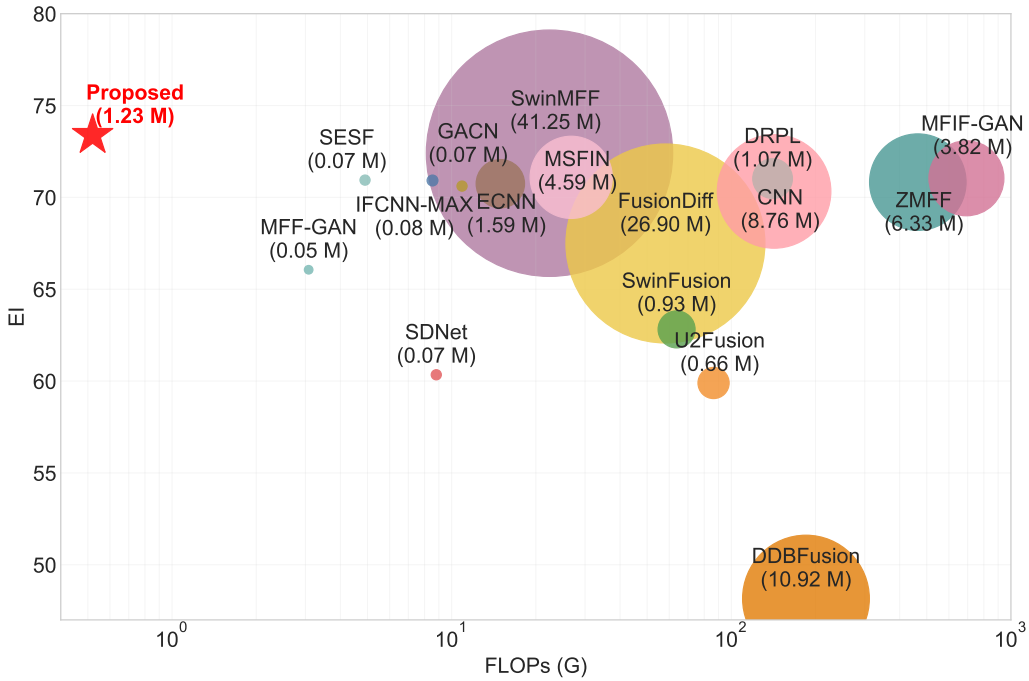

Figure 1. Quantitative comparison of model size (indicated by bubble size), computational complexity (FLOPs), and fusion quality (*EI*) among different deep learning-based MFF methods on the Lytro dataset [28].

The remainder of this paper is organized as follows: Section 2 provides a comprehensive review of related works in MFF, encompassing traditional methods, deep learning-based approaches, existing MLP-based network architecture, and existing MLP-based image-fusion networks. Section 3 presents our proposed method. Section 4 delivers comprehensive experimental results and comparisons with state-of-the-art methods. Section 5 discusses the limitations and potential future directions of our work. Finally, Section 6 concludes the paper.

2. Related Works

2.1. Conventional Image-Fusion Methods

Transform-domain methods. Transform-domain methods include methods utilizing pyramids [11,12], wavelets [13,14], Discrete Cosine Transform (DCT) [15], Non-Subsampled Contourlet Transform (NSCT) [16,29], and Sparse Representation (SR) [17,18]. While these methods offer the distinct advantage of integrating feature information across diverse frequency spectra, they inherently demand the judicious manual selection of appropriate transformations and fusion rules.

Spatial-domain methods. Block-based approaches [6,7] divide images into patches and select blocks with the highest activity measures, providing computational efficiency but potentially introducing blocking artifacts at patch boundaries. Region-based methods [8] perform segmentation first and then apply fusion rules to each region, maintaining spatial coherence but being vulnerable to segmentation errors. Advanced pixel-based techniques using SIFT features or defocus estimation [9,10] provide more sophisticated selection criteria but are computationally intensive and noise-sensitive.

2.2. Deep Learning-Based Image-Fusion Methods

Decision map-based methods. This direction was pioneered by CNN [19], which utilized convolutional networks for this purpose. Subsequent research has advanced performance through various strategies: MSFIN [20] and GEU-Net [30] leveraged multi-scale

features; SESF [31] achieved unsupervised fusion by computing pixel-level spatial frequencies within feature maps; SMFuse [32] introduced a self-supervised learning-based fusion technique; MFIF-GAN [33] significantly improved decision map quality by employing adversarial learning and α-matte modeling. More recent studies, such as [34], have attempted to integrate Transformers for global image modeling to overcome the inherent limitations of Convolutional Neural Networks. ZMFF [35] realized zero-shot fusion by utilizing Deep Image Prior. Furthermore, BridgeMFF [24] proposed a dual-adversarial-based decision map-refinement method. In a notable recent work, Ref. [25] introduced the Wavelet Mamba Module to the MFF field, combining it with deep priors to further enhance the accuracy of decision maps. MFIF-STCU-Net [36] achieved state-of-the-art results by leveraging a hybrid U-Net and Transformer architecture to synthesize training data using a depth estimation model. Building on this, DMANet [37] introduced explicit defocus blur modeling into the MFF process, enhancing both interpretability and performance. Most recently, LSKN-MFIF [38] surpassed existing methods with a dynamically adjusting "large selective kernel" module that captures both global and local features more accurately.

End-to-end methods. IFCNN [39] and U2Fusion [40] were among the first to apply end-to-end methods for MFF. MFF-GAN [41] proposed an unsupervised generative adversarial network with adaptive and gradient joint constraints for MFF. More recently, SwinFusion [22], SwinMFF [21] and FusionDiff [23] have applied Transformers and diffusion models to this domain. Recently, DDBFusion [42] further proposed a unified image decomposition and fusion framework based on dual decomposition and Bézier curves, enhancing the filtering capability for redundant information. StackMFF [5] extended the approach to image stacks using 3D CNNs and synthesized training data. Meanwhile, MMAE [43] introduced a mask attention mechanism to filter out redundant information. A different direction was taken by LFDT-Fusion [44], which proposed an efficient latent feature-guided diffusion model that works in a compressed latent space.

The proposed MLP-MFF adopts an end-to-end approach rather than a decision map-based one. Beyond the aforementioned advantages, this choice streamlines the processing pipeline, significantly reducing error accumulation that might arise from intermediate steps. This simplification contributes to a more user-friendly and effective solution in practical applications, demonstrating greater potential for deployment in production environments.

2.3. MLP-like Architecture in Computer Vision

The MLP-Mixer [45] introduced a novel vision architecture based on Multi-Layer Perceptrons (MLPs), demonstrating performance competitive with CNNs and Transformers in computer vision tasks. This seminal work established that convolutions and attention mechanisms are not indispensable for achieving strong performance. Following this, several variants like ResMLP [46], RepMLPNet [47], EAMLP [48], ViP [49], sMLP-Net [50], and Strip-MLP [51] were proposed. These aim for a better balance between performance and efficiency, yet similar to many Transformer-based networks, they are restricted to fixed-dimension inputs, precluding their direct application in downstream dense image prediction tasks.

Researchers have attempted to address this limitation by substituting spatial MLPs with alternative spatial aggregation operations, such as those found in AS-MLP [52], S2-MLP [53], Shift MLP [54], and CycleMLP [27]. However, these approaches sacrifice global receptive fields and struggle to capture multi-scale information. Furthermore, hybrid architectures combining convolutions and MLPs, including Wave-MLP [55], ATMNet [56], and RaMLP [57], have been proposed to mitigate the aforementioned issues, albeit at the cost of increased network complexity.

In contrast, PFMLP [26] offers a more competitive and efficient pure-MLP architecture, termed Pyramid Fusion MLP, to overcome these limitations. Specifically, each block within PFMLP incorporates multi-scale pooling and fully connected layers to generate a feature pyramid, which is then fused via upsampling layers and additional fully connected layers. Employing diverse downsampling rates enables the acquisition of varying receptive fields, allowing the model to concurrently capture long-range dependencies and fine-grained details. This fully leverages the potential of global contextual information, thereby enhancing the model's spatial representation capabilities. To our knowledge, PFMLP is currently one of the few MLP architectures that simultaneously supports variable input dimensions, maintains a global receptive field, facilitates multi-scale information interaction, and remains purely MLP-based. The backbone network presented in this paper is implemented based on PFMLP.

2.4. MLP-Based Network in Image Fusion

While CNNs and Transformer-based networks have been extensively explored in image fusion, MLP-based networks remain relatively underexplored in this domain. Nevertheless, recent studies have begun to introduce MLP-based architectures into image fusion.

For instance, FusionMLP [58], building upon MAXIM [59], proposes the first MLP-based multi-scene image-fusion network. Similarly, CMFuse [60] introduces a novel infrared and visible image-fusion network that leverages a hybrid CNN and MLP architecture to effectively model long-range dependencies and facilitate cross-modal information exchange.

In contrast to these prior works, our research specifically designs an MLP-based network for MFF. Furthermore, our primary focus is on achieving an optimal balance between model efficiency and fusion performance.

3. Methods

3.1. Overall Architecture

As the backbone of the proposed MLP-MFF, PFMLP is a multi-stage architecture capable of capturing multi-scale context and a global receptive field, all without relying on convolution layers. This allows it to effectively handle variable-sized inputs. As illustrated in Figure 2, the framework first takes two input images of size 3×3 and uses 3×3 convolutions to map them into feature maps of size $8 \times H/2 \times W/2$.

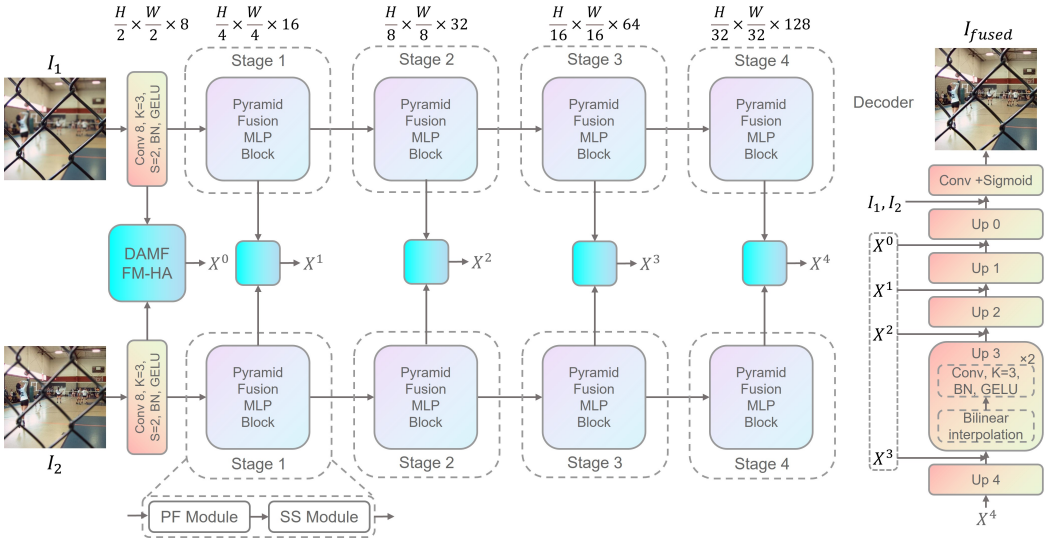

Figure 2. The framework of the proposed MLP-MFF.

The network then proceeds through four stages to capture both local and global spatial information and facilitate inter-channel communication. Each stage begins with a single PFMLP block that incorporates downsampling, reducing the image height and width by half while doubling the channel count. The PFMLP block is utilized to enable information exchange across different spatial locations and channels. As depicted in Figure 3, each PFMLP block consists of a Pyramid Fusion (PF) module and a Single Scale (SS) module, with residual connections applied after each module.

Following each stage, the outputs from both images are fed into the proposed Dual-Path Adaptive Multi-scale Feature-Fusion Module based on Hybrid Attention (DAMFFM-HA) for multi-scale fusion. Finally, the fused image is progressively restored to its original dimensions through four decoders, each comprising bilinear interpolation and convolutional layers.

The proposed MLP-MFF is a MFF network characterized by its global and multi-scale receptive fields, offering flexible handling of various image sizes. Unlike SwinFusion [22] and SwinMFF [21], MLP-MFF does not necessitate patch-wise fusion, thereby genuinely leveraging a global receptive field for MFF.

3.2. Pyramid Fusion MLP Block

The Pyramid Fusion MLP Block comprises a Pyramid Fusion (PF) Module and a Single Scale (SS) Module, as depicted in Figure 3.

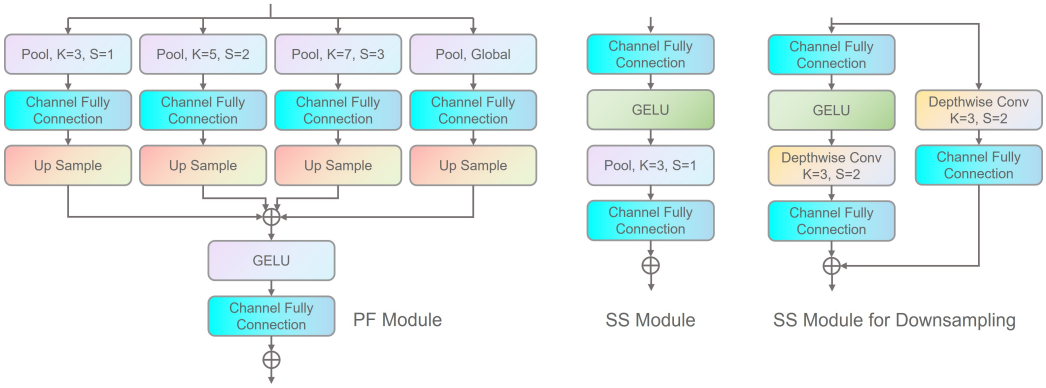

Figure 3. The details of PF Module and SS Module.

Pyramid Fusion Module (PF Module). The PF module generates a feature pyramid using four branches, each employing a pooling layer with a distinct kernel size. Unlike conventional pooling layers, the PF module utilizes learnable pooling layers to more effectively aggregate local visual cues. Following the acquisition of the feature pyramid, a channel fully connected (FC) layer maps the input C channels to C' channels. Subsequently, nearest neighbor sampling is applied to obtain multiple feature maps, which are then summed and passed through a GELU activation function. Finally, another channel FC layer transforms the summation result into the final feature representation.

Single Scale Module (SS Module). The SS module, a variant of the channel MLP, exists in two forms: a naive version and a down-sample version. The naive SS module consists of two-channel MLPs with a GELU activation between them, and a pooling layer (without down-sampling) for information aggregation. The down-sample SS module introduces an additional branch comprising a depth-wise convolution layer and a channel FC layer. Furthermore, the pooling layer in the original branch is replaced with a depth-wise convolution layer. The final output is obtained by summing the results of both branches, thereby minimizing spatial information loss during down-sampling and facilitating more efficient feature extraction.

For a comprehensive description of the Pyramid Fusion MLP Block, readers are referred to PFMLP [26].

3.3. Dual-Path Adaptive Multi-Scale Feature-Fusion Module

To effectively fuse multi-scale features from two multi-focus images, this paper proposes a Dual-Path Adaptive Multi-scale Feature-Fusion Module (DAMFFM-HA), as shown in Figure 4. This module integrates hybrid attention mechanisms and adaptive weight allocation strategies to capture and fuse complementary information between different images, thereby improving fusion quality. The DAMFFM-HA comprises two core components: an attention module responsible for importance assessment and enhancement of input features, and a fusion block that executes adaptive feature-fusion operations.

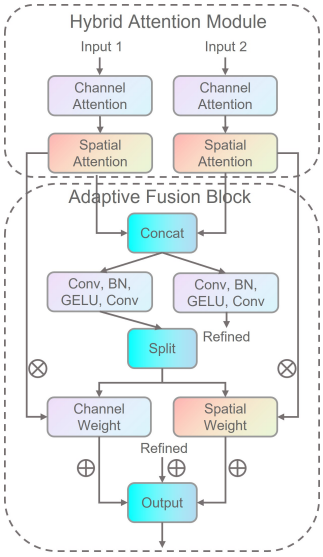

Figure 4. The details of the dual-path adaptive multi-scale feature-fusion module based on hybrid attention (DAMFFM-HA).

Hybrid Attention Module. The attention module employs a hybrid attention mechanism for two-dimensional feature processing, which consists of channel attention for capturing global feature dependencies and spatial attention for emphasizing important spatial locations within feature maps.

The channel attention mechanism is designed to model interdependencies among feature channels and adaptively recalibrate channel-wise feature responses by explicitly modeling channel relationships. For input feature maps $F \in \mathbb{R}^{C\times H\times W}$, global average pooling is applied to obtain $F_{gap} \in \mathbb{R}^{C\times 1\times 1}$, which aggregates spatial information across each channel. This is followed by dimension reduction convolution with reduction ratio of 8, ReLU activation, dimension expansion convolution, and Sigmoid activation to generate channel attention weights $A_c \in \mathbb{R}^{C\times 1\times 1}$.

The spatial attention mechanism focuses on identifying and highlighting informative spatial regions within feature maps to suppress irrelevant background information and enhance focus-relevant areas. The spatial attention process applies 1×1 convolution for input feature dimension reduction, followed by batch normalization and ReLU activation, then another 1×1 convolution to generate single-channel spatial attention maps $A_s \in \mathbb{R}^{1\times H\times W}$, which are finally normalized through Sigmoid function to produce spatial importance weights.

The final attention-enhanced feature representation is obtained by sequentially applying both attention mechanisms: $F_{att} = F \odot A_c \odot A_s$, where $\odot$ denotes element-wise multiplication.

Adaptive Fusion Block. The adaptive fusion block serves as the core component of DAMFFM-HA, responsible for adaptively fusing dual-path features enhanced by attention. The module processes input features F_1 and F_2 through independent attention modules to obtain enhanced features F_{att1} and F_{att2}, ensuring each feature path is optimized according to its inherent characteristics.

The module generates adaptive weights by concatenating two enhanced features along the channel dimension to form $F_{concat} = [F_{att1}, F_{att2}] \in \mathbb{R}^{2C\times H\times W}$, then employs a weight generation network consisting of two 3×3 convolutions, batch normalization, GELU activation, and final Softmax normalization to produce pixel-level fusion weights $W = [W_1, W_2] \in \mathbb{R}^{2\times H\times W}$ where $W_1 + W_2 = 1$.

The weighted fusion and feature-refinement process first performs adaptive weighted fusion $F_{weighted} = W_1 \odot F_{att1} + W_2 \odot F_{att2}$, while simultaneously processing concatenated features through a fusion convolutional network to generate refined features $F_{refined} = \phi(F_{concat})$, where $\phi(\cdot)$ represents a feature-refinement network composed of two 3×3 convolutions, batch normalization, and GELU activation functions, with final fused features obtained through residual connection as $F_{fused} = F_{weighted} + F_{refined}$.

3.4. Decoder

The decoder is responsible for progressively reconstructing the fused image from multi-scale fused features to the original resolution, employing a five-stage upsampling architecture.

Upsampling Block Design. Each decoder stage employs an upsampling block that combines bilinear interpolation with convolutional operations. This module first doubles the spatial dimensions through bilinear interpolation, followed by two consecutive 3×3 convolutions with batch normalization and GELU activation functions, ensuring smooth feature transitions and reducing aliasing effects during upsampling.

For an input feature map $F_{in} \in \mathbb{R}^{C_{in}\times H\times W}$, the upsampling block operates as:

$$F_{up} = \text{Conv}_{3\times3}(\text{BN}(\text{GELU}(\text{Conv}_{3\times3}(\text{BN}(\text{Interpolate}(F_{in}, \text{scale} = 2))))))$$

where $\text{Interpolate}(\cdot)$ represents bilinear interpolation with a scale factor of 2, and the output feature map has dimensions $C_{out} \times 2H \times 2W$.

Progressive Reconstruction with Skip Connections. The decoder incorporates skip connections between corresponding encoder and decoder stages to preserve fine-grained spatial details that may be lost during encoding. The decoder processes multi-scale fused features in the following sequence:

Stage 4 to 3: The deepest fused features X^4 are upsampled through `Decoder4` and element-wise added with X^3: $F_{up4} = \text{Up4}(X^4) + X^3$

Stage 3 to 2: $F_{up3} = \text{Up3}(F_{up4}) + X^2$

Stage 2 to 1: $F_{up2} = \text{Up2}(F_{up3}) + X^1$

Stage 1 to 0: $F_{up1} = \text{Up1}(F_{up2}) + X^0$

Where X^i represents the fused features at stage i obtained from the DAMFFM-HA fusion modules.

Final Reconstruction. The final reconstruction stage comprises a final upsampling block and an output convolutional layer. The upsampled features are enhanced through residual connections with the original input images:

$$F_{final} = \text{Up0}(F_{up1}) + I_1 + I_2$$

where I_1 and I_2 are the original input images. The final output is obtained through a 3×3 convolution followed by a Sigmoid activation function, ensuring output values are constrained within the $[0, 1]$ range:

$$I_{fused} = \sigma(\text{Conv}_{3\times3}(F_{final}))$$

This decoder design ensures the effective reconstruction of high-quality fused images while maintaining both global contextual information captured by the encoder and local spatial details preserved through skip connections with fused multi-scale features.

3.5. Loss Function

To effectively train our MFF network, we design a comprehensive loss function that combines multiple complementary loss terms to optimize the quality of fused images. The overall loss function is formulated as:

$$\mathcal{L}_{total} = \lambda_{mse}\mathcal{L}_{MSE} + \lambda_{ssim}\mathcal{L}_{SSIM} + \lambda_{grad}\mathcal{L}_{grad} \tag{1}$$

where λ_{mse}, λ_{ssim}, and λ_{grad} represent the weighting coefficients for the mean squared error loss, structural similarity loss, and gradient loss, respectively.

The MSE loss measures pixel-wise differences between the fused image and the reference image, ensuring basic reconstruction fidelity:

$$\mathcal{L}_{MSE} = \frac{1}{N}\sum_{i=1}^{N}(I_{fused}(i) - I_{ref}(i))^2 \tag{2}$$

where I_{fused} denotes the fused image, I_{ref} represents the reference image, and N is the total number of pixels.

The Structural Similarity Index Measure (SSIM) better captures perceptual image quality from the perspective of human visual perception. We employ $1 - SSIM$ as our structural similarity loss:

$$\mathcal{L}_{SSIM} = 1 - SSIM(I_{fused}, I_{ref}) \tag{3}$$

The SSIM metric is computed as:

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{4}$$

where μ_x and μ_y are the local means of images x and y, σ_x^2 and σ_y^2 are the local variances, σ_{xy} is the local covariance, and C_1 and C_2 are small constants for numerical stability.

To compute local statistics, we employ an 11×11 Gaussian weighting window with weights determined by:

$$w(x) = \exp\left(-\frac{(x - \mu_w)^2}{2\sigma_w^2}\right) \tag{5}$$

where $\sigma_w = 1.5$ and μ_w corresponds to the window center. The weights are normalized to sum to unity.

The gradient loss preserves edge information and fine-grained textures in the fused image. We utilize Sobel operators to compute image gradients:

$$G_x = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}, \quad G_y = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix} \tag{6}$$

The gradient magnitude is calculated as:

$$\nabla I = |G_x * I| + |G_y * I| \tag{7}$$

where $*$ denotes the convolution operation.

For MFF, the ideal fused image should preserve the strongest gradient information from the input images. Therefore, we define the joint gradient as:

$$\nabla I_{joint} = \max(\nabla I_A, \nabla I_B) \tag{8}$$

The gradient loss is computed using the L1 norm:

$$\mathcal{L}_{grad} = \frac{1}{N}\sum_{i=1}^{N} |\nabla I_{fused}(i) - \nabla I_{joint}(i)| \tag{9}$$

Through extensive empirical evaluation, we set the loss weights as: $\lambda_{mse} = 0.2$, $\lambda_{ssim} = 0.7$, and $\lambda_{grad} = 0.1$. This configuration ensures that structural similarity plays a dominant role in the training process while maintaining appropriate pixel-level constraints and edge preservation capabilities. The emphasis on SSIM aligns with human visual perception, leading to perceptually superior fusion results.

4. Experiments

4.1. Experimental Setup

Training Strategy. We trained our model using the DUTS dataset (accessed on 3 July 2025) [61], which provides 15,572 images. We allocated 10,553 images for training and 5019 for validation, resizing all to 256×256 pixels. To create multi-focus image pairs, we transformed ground truth annotations into binary masks. These masks then guided the application of Gaussian blur, with kernel sizes from 3 to 21, to generate realistic training samples (detailed in SwinMFF [21]). For network optimization, we utilized the AdamW optimizer with an initial learning rate of 1×10^{-3}, β parameters set to $(0.9, 0.999)$, an ϵ of 1×10^{-8}, and a weight decay of 0.0001. A CosineAnnealingLR scheduler was implemented to progressively reduce the learning rate over the course of training. The model was trained for 20 epochs with a batch size of 16 on an Nvidia A6000 GPU system. operating at 2.90 GHz. The entire framework was developed in PyTorch 2.6.0.

Datasets for Evaluation. MLP-MFF's performance was rigorously evaluated against three prominent MFF benchmarks: Lytro [28], MFFW [62], and MFI-WHU [41]. The Lytro dataset, consisting of 20 light-field camera image pairs, facilitated both qualitative and quantitative analysis. MFFW, with its 13 image pairs exhibiting pronounced defocus, was employed for qualitative evaluation. Similarly, the MFI-WHU dataset, offering 120 synthetically generated (via Gaussian blur), was also used for qualitative assessment.

Methods for Comparison. To comprehensively evaluate the performance of our proposed MLP-MFF, we compare it with various state-of-the-art methods from different categories. For traditional methods, we include both spatial domain and transform domain approaches. The spatial domain methods include SSSDI [63], QUADTREE [64], DSIFT [65], SRCF [28], GFDF [66], BRW [67], MISF [68], and MDLSR_RFM [69]. The transform domain methods include DWT [14], DTCWT [70], NSCT [71], CVT [72], GFF [73], SR [74], ASR [75], and ICA [76]. For decision map-based deep learning methods, we compare with CNN [19], ECNN [77], DRPL [78], SESF [31], MFIF-GAN [33], MSFIN [20], GACN [79], and ZMFF [35]. For end-to-end deep learning methods, we compare with IFCNN-MAX [39], U2Fusion [40], MFF-GAN [41], SwinFusion [22], FusionDiff [23], SwinMFF [21], and DDBFusion [42]. These methods represent the current state-of-the-art in MFF, covering different technical approaches and architectural designs. The comparison with these methods allows us to thoroughly evaluate the effectiveness and advantages of our proposed MLP-MFF approach.

Evaluation metrics. To comprehensively evaluate the performance of different fusion methods, we employ eight widely-used metrics that can be categorized based on their

theoretical foundations. Information theory-based metrics include Entropy (EN) and Mutual Information (MI), which measure the information content and transfer in the fused image. Edge and gradient-based metrics consist of Edge Information (EI), Spatial Frequency (SF), Average Gradient (AVG), and $Q^{ab/f}$, which assess the preservation of edge details and image clarity. Structure and visual quality-based metrics include Structural Similarity Index Measure (SSIM) and Visual Information Fidelity (VIF), which evaluate the structural preservation and visual quality of the fused image. These metrics provide a comprehensive evaluation framework that considers different aspects of fusion quality, including information content, edge preservation, structural similarity, and visual quality.

4.2. Experimental Results

Qualitative comparison. In Figure 5, we present a comparative analysis of various MFF methods applied to the Lytro dataset [28]. The figure is structured into four rows, each representing a distinct category of fusion approaches: transform-domain methods, spatial-domain methods, decision map-based deep learning methods, and end-to-end deep learning methods. From the first example, it is evident that transform-domain methods and end-to-end deep learning methods consistently outperform others in preserving intricate details within complex scenes, exhibiting significantly fewer fusion artifacts. Furthermore, the second example clearly highlights the superior performance of our proposed method over existing end-to-end deep learning approaches. Our method effectively suppresses artifacts along object edges, a common challenge for other end-to-end techniques that show varying degrees of artifacts in the provided examples.

Figure 5. The fusion results of various SOTA methods on the Lytro dataset [28].

To provide a more intuitive comparison of the performance of different end-to-end methods, we further used the difference maps between the fusion results and the two source images as a basis for comparison. A greater discrepancy between the two difference maps indicates a superior fusion result [21]. First, we compare our method with decision map-based deep learning approaches in Figure 6. Even though our method is an end-to-end fusion approach where fused image pixel values are inferred by the network rather than sampled from source images, its difference maps are comparable to those produced by decision map-based deep learning methods. This demonstrates that our proposed method, similar to SwinMFF [21], achieves excellent pixel-level fidelity. Next, we compare our method with other end-to-end approaches in Figure 7. Our method, along with IFCNN-MAX and SwinMFF, significantly outperforms other methods in this comparison. While IFCNN-MAX and SwinMFF show slightly superior results to our proposed method, their computational costs are approximately 16 and 43 times higher, respectively. Therefore, our method strikes a favorable balance between computational efficiency and fusion quality.

Figure 6. The difference maps of various SOTA decision map-based MFF methods implemented using deep learning on the Lytro dataset [28].

To further evaluate the performance of different end-to-end methods in scenarios with strong defocus, we conducted an additional comparative analysis on the MFFW dataset [62], as shown in Figure 8. The results indicate that some methods, such as MFF-GAN [41] and DDBFusion [42], exhibit noticeable fusion errors in strongly defocused scenes. In contrast, our proposed method consistently maintains top-tier fusion quality across all examples, with virtually no significant artifacts appearing at the edges. Similarly, we used difference maps for further comparison in Figure 9. It is evident from the difference maps that several methods, including MFF-GAN [41], DDBFusion [42], and SwinFusion [22], produce

two difference maps that are quite similar. This suggests that their fusion results fail to adequately distinguish and fuse the foreground and background in strongly defocused scenes. However, our proposed method, along with SwinMFF [21] and IFCNN-MAX [39], demonstrates significantly superior performance in these challenging conditions.

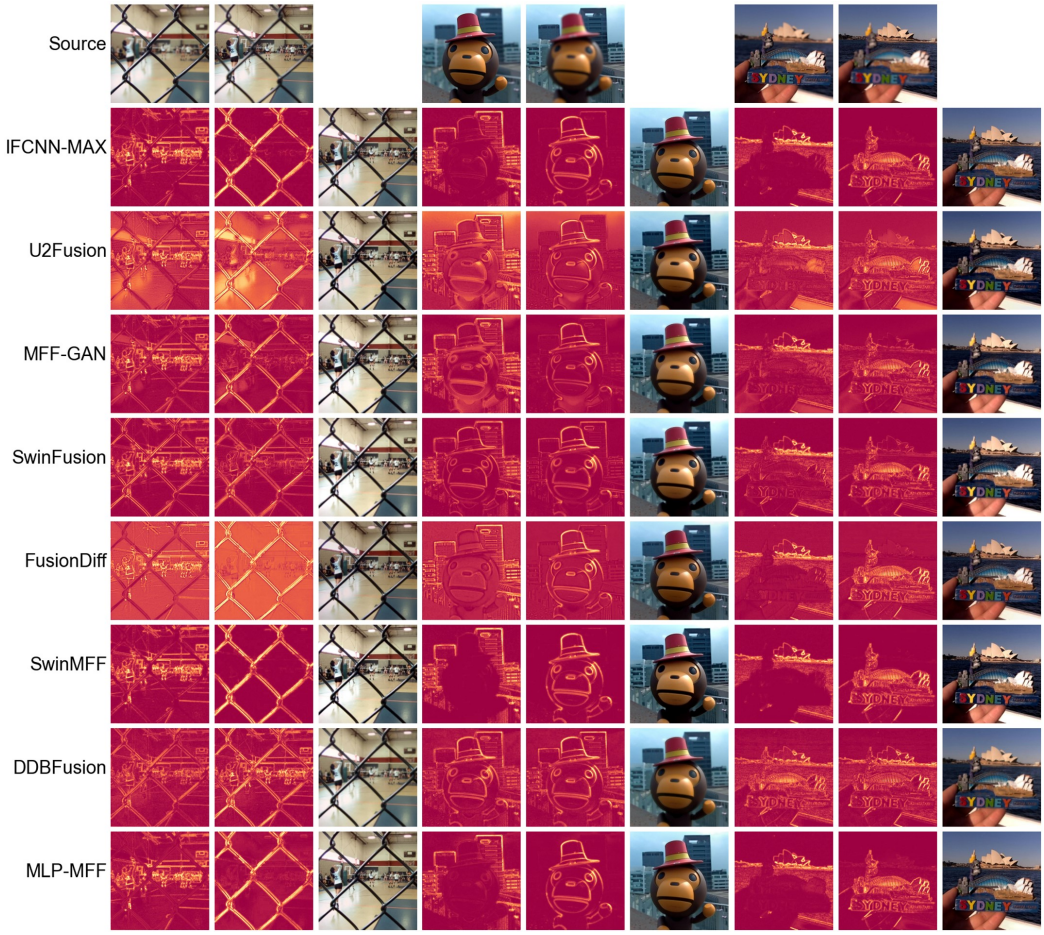

Figure 7. The difference maps of different SOTA end-to-end MFF methods implemented using deep learning on the Lytro dataset [28].

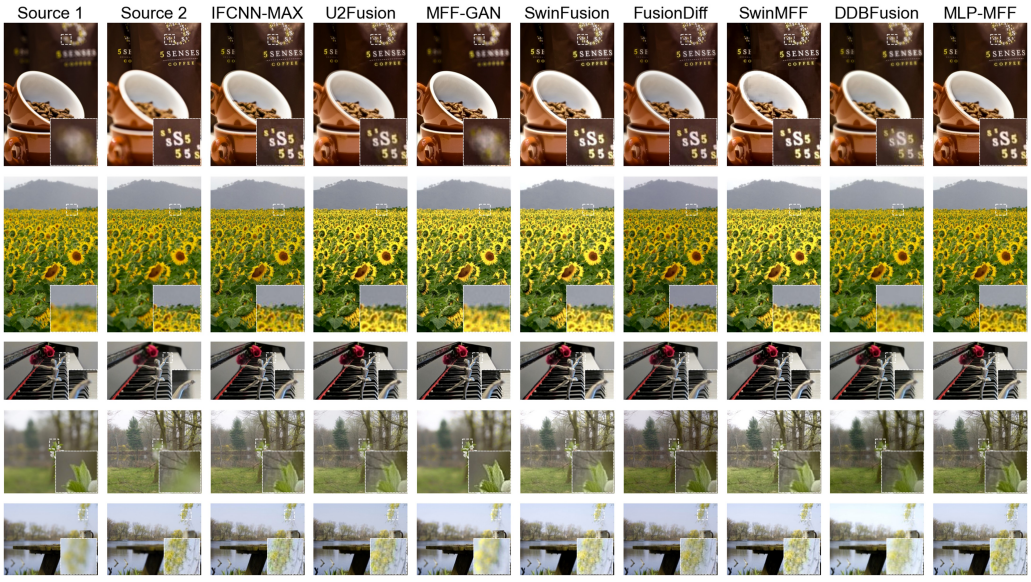

Figure 8. The fusion results of various SOTA methods on the MFFW dataset [62].

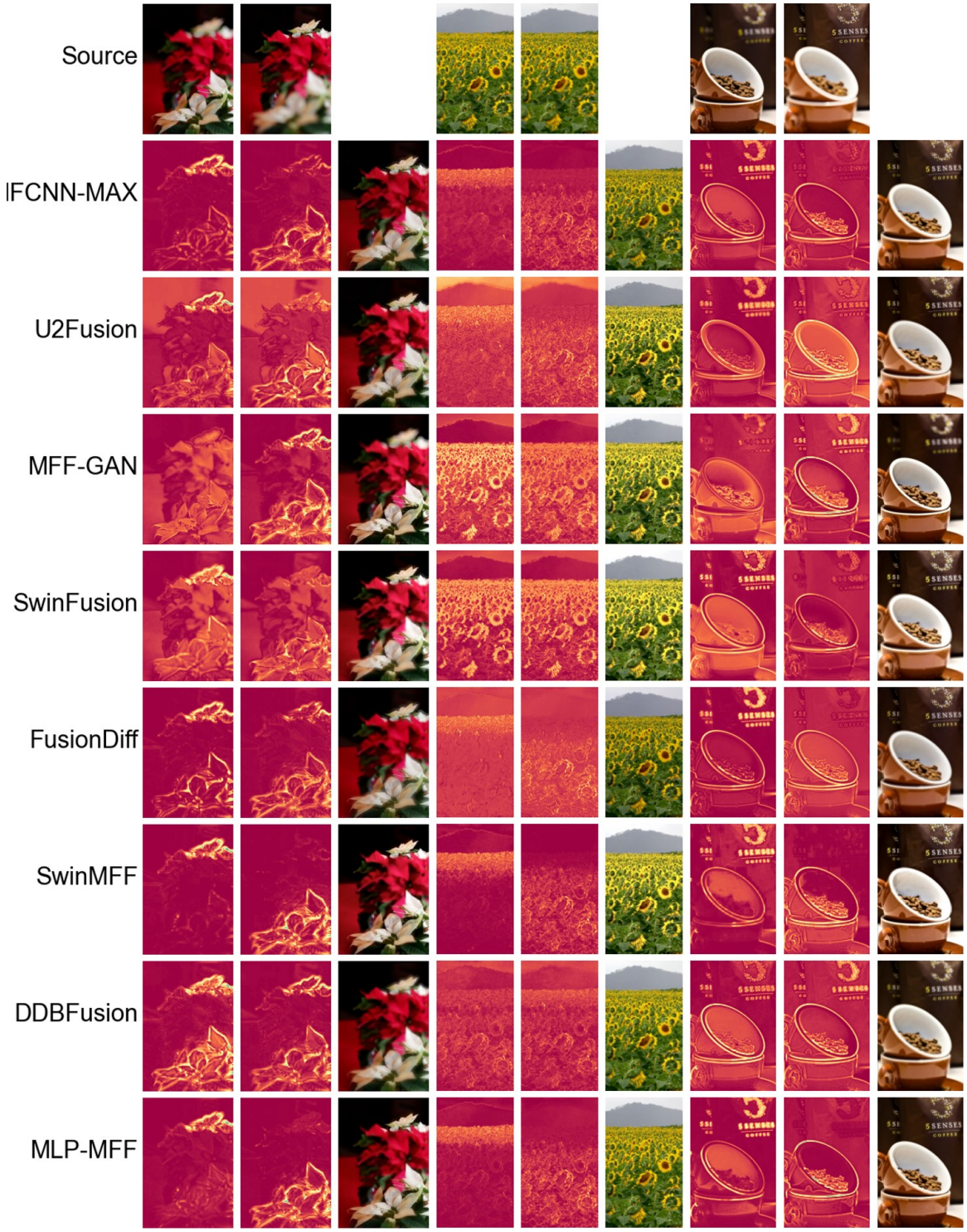

Figure 9. The difference maps of different SOTA end-to-end MFF methods implemented using deep learning on the MFFW dataset [62].

In Figure 10, we further visualize the fusion results of various end-to-end deep learning methods on the MFI-WHU dataset [41], along with their corresponding difference plots. The MFI-WHU dataset offers a more diverse range of examples, including small objects and high-contrast scenes, as depicted in Figure 10. The first example in Figure 10 demonstrates that our proposed method maintains robustness even with small objects, an area where many other methods, such as DDBFusion [42] and U2Fusion [40], are susceptible to noise and artifacts. Interestingly, while the diffusion

model-based FusionDiff [23] generally exhibits poor performance and often produces noticeable color casts in previous examples, it performs exceptionally well in high-contrast scenes.

In terms of comprehensive performance across multiple datasets, our proposed method consistently delivers high-quality fusion results. It effectively handles complex scenarios, preserves fine details, and maintains excellent pixel-level fidelity. Furthermore, it demonstrates a strong ability to distinguish and fuse foreground and background elements, even in challenging conditions like strong defocus.

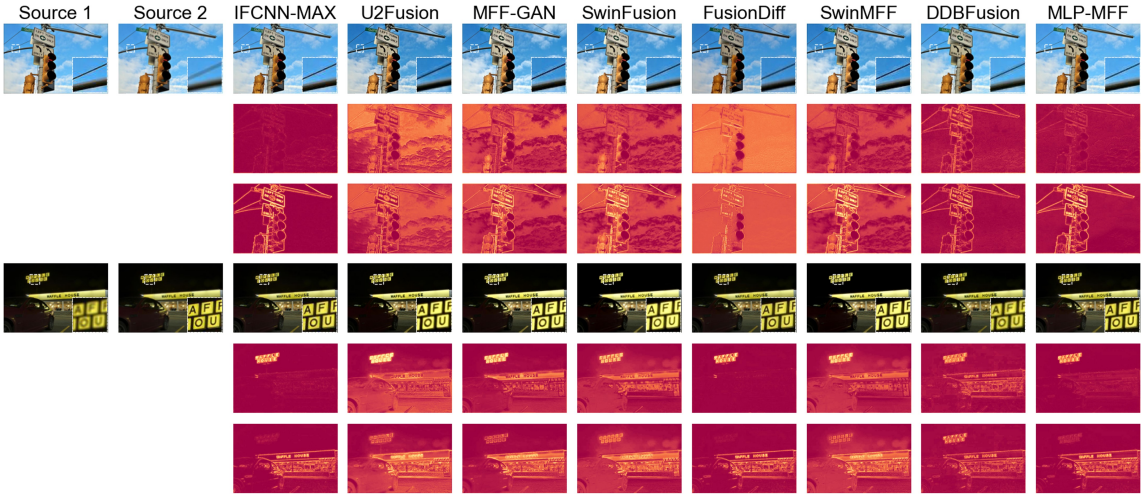

Figure 10. The fusion results of various SOTA methods on the MFI-WHU dataset [41].

Quantitative comparison. Table 1 presents a comprehensive quantitative comparison of different MFF methods on the Lytro dataset [28]. The best-performing method for each metric is shown in bold, while the second-best is underlined. Additionally, a colored background is used to highlight the proposed method. An upward-pointing arrow (↑) next to a metric's name indicates that a higher value is better. This formatting convention is applied consistently across all tables in this paper. The results demonstrate that our proposed MLP-MFF achieves superior performance across multiple evaluation metrics. Specifically, MLP-MFF achieves the highest scores in six out of eight metrics, significantly outperforming other methods in these aspects. End-to-end deep learning methods show varying performance levels. While DDBFusion achieves the highest SSIM score of 0.8661, it performs poorly in other metrics such as EI (48.1600) and SF (12.1484). SwinMFF shows balanced performance across multiple metrics but still falls short of MLP-MFF's comprehensive superiority.

In Figure 11, we present the quantitative metrics for the fusion results of different methods on each image pair within the Lytro dataset [28]. The red line represents our proposed method. The results clearly show that the quantitative advantage of our proposed method extends across nearly the entire dataset, rather than being concentrated in just a few isolated examples that might skew overall metrics. This strong performance across diverse examples indicates that our proposed method possesses good generalization capabilities and superior performance.

In Tables 2 and 3, we provide a further quantitative comparison of various deep learning-based methods on the MFFW dataset [62] and the MFI-WHU dataset [41], respectively. Across both datasets, our proposed method consistently ranks first or second in multiple metrics. This further demonstrates that our method maintains excellent fusion performance across a variety of scenarios.

The comprehensive quantitative experimental results collectively validate the superiority of the proposed MLP-MFF method across multiple evaluation metrics. MLP-MFF consistently demonstrates outstanding performance in terms of information entropy, edge information, structural similarity, and visual information fidelity, significantly outperforming both traditional methods and various existing deep learning approaches. Furthermore, MLP-MFF's stable performance across different public datasets further underscores its strong generalization capability and robustness.

Table 1. Quantitative comparison of different MFF methods on the Lytro dataset [28].

Methods	EI ↑	SF↑	AVG↑	$Q^{ab/f}$↑	SSIM↑	EN↑	MI↑	VIF↑
Methods based on image spatial domain								
SSSDI [63]	70.7102	19.3567	6.8234	0.6966	0.8069	7.5334	15.0668	1.1309
QUADTREE [64]	70.8957	19.4163	6.8412	0.7027	0.8085	7.5342	15.0684	1.1368
DSIFT [65]	70.9808	19.4194	6.8493	0.7046	0.8083	7.5344	15.0688	1.1381
SRCF [28]	71.0810	19.4460	6.8607	0.7036	0.8075	7.5345	15.0690	1.1374
GFDF [66]	70.6258	19.3312	6.8145	0.7049	0.8098	7.5337	15.0674	1.1336
BRW [67]	70.6777	19.3433	6.8200	0.7040	0.8093	7.5337	15.0675	1.1336
MISF [68]	70.4148	19.2203	6.7945	0.6984	0.8084	7.5335	15.0671	1.1222
MDLSR_RFM [69]	70.9078	19.4100	6.8422	0.7518	0.8393	7.5343	15.0686	1.1353
Methods based on image transform domain								
DWT [14]	70.7942	19.3342	6.8336	0.6850	0.8059	7.5436	<u>15.0872</u>	1.1114
DTCWT [70]	70.5666	19.3204	6.8134	0.6929	0.8076	7.5396	15.0791	1.1079
NSCT [71]	70.4289	19.2662	6.8027	0.6901	0.8102	7.5408	15.0816	1.1249
CVT [72]	70.3233	19.2713	6.7897	0.7243	0.8376	<u>7.5414</u>	15.0828	1.1044
GFF [73]	70.5179	19.2947	6.8058	0.6998	0.8088	7.5358	15.0716	1.1277
SR [74]	70.2498	19.2819	6.7818	0.6944	0.8097	7.5325	15.0650	1.1208
ASR [75]	70.3342	19.2818	6.7897	0.6951	0.8093	7.5327	15.0654	1.1201
ICA [76]	68.3180	18.5968	6.6125	0.6766	0.8176	7.5327	15.0655	1.0708
Decision map-based methods using deep learning								
CNN [19]	70.3238	19.2295	6.7860	0.7019	0.8096	7.5331	15.0663	1.1255
ECNN [77]	70.7432	19.3837	6.8261	0.7030	0.8089	7.5338	15.0675	1.1337
DRPL [78]	71.0214	19.4546	6.8531	<u>0.7574</u>	0.8401	7.5342	15.0683	1.1393
SESF [31]	70.9403	19.4158	6.8448	0.7031	0.8086	7.5348	15.0696	1.1395
MFIF-GAN [33]	71.0395	19.4370	6.8560	0.7029	0.8078	7.5345	15.0690	1.1393
MSFIN [20]	71.0914	19.4438	6.8602	0.7045	0.8082	7.5348	15.0695	1.1420
GACN [79]	70.6148	19.3087	6.8101	**0.7581**	<u>0.8413</u>	7.5330	15.0661	1.1304
ZMFF [35]	70.8298	18.9707	6.8045	0.6635	0.8073	7.5368	15.0735	1.1331
End-to-end methods based on deep learning								
IFCNN-MAX [39]	70.9193	19.3793	6.8463	0.6784	0.8111	7.5361	15.0722	1.1322
U2Fusion [40]	59.8957	14.9334	5.6515	0.6190	0.8239	7.3077	14.6153	0.9882
MFF-GAN [41]	66.0601	18.4022	6.4089	0.6222	0.8067	7.4076	14.8153	1.0084
SwinFusion [22]	62.8130	16.6430	5.9862	0.6597	0.8367	7.5238	15.0476	1.0685
FusionDiff [23]	67.4911	18.8483	6.5325	0.6744	0.8071	7.4909	14.9817	1.0448
SwinMFF [21]	<u>72.4041</u>	<u>19.7954</u>	<u>6.9734</u>	0.7321	0.8382	7.5413	15.0826	<u>1.1810</u>
DDBFusion [42]	48.1600	12.1484	4.5883	0.5026	**0.8661**	7.5332	15.0663	0.8874
MLP-MFF	**73.3902**	**19.8181**	**7.0501**	0.7025	0.8344	**7.5741**	**15.1482**	**1.2126**

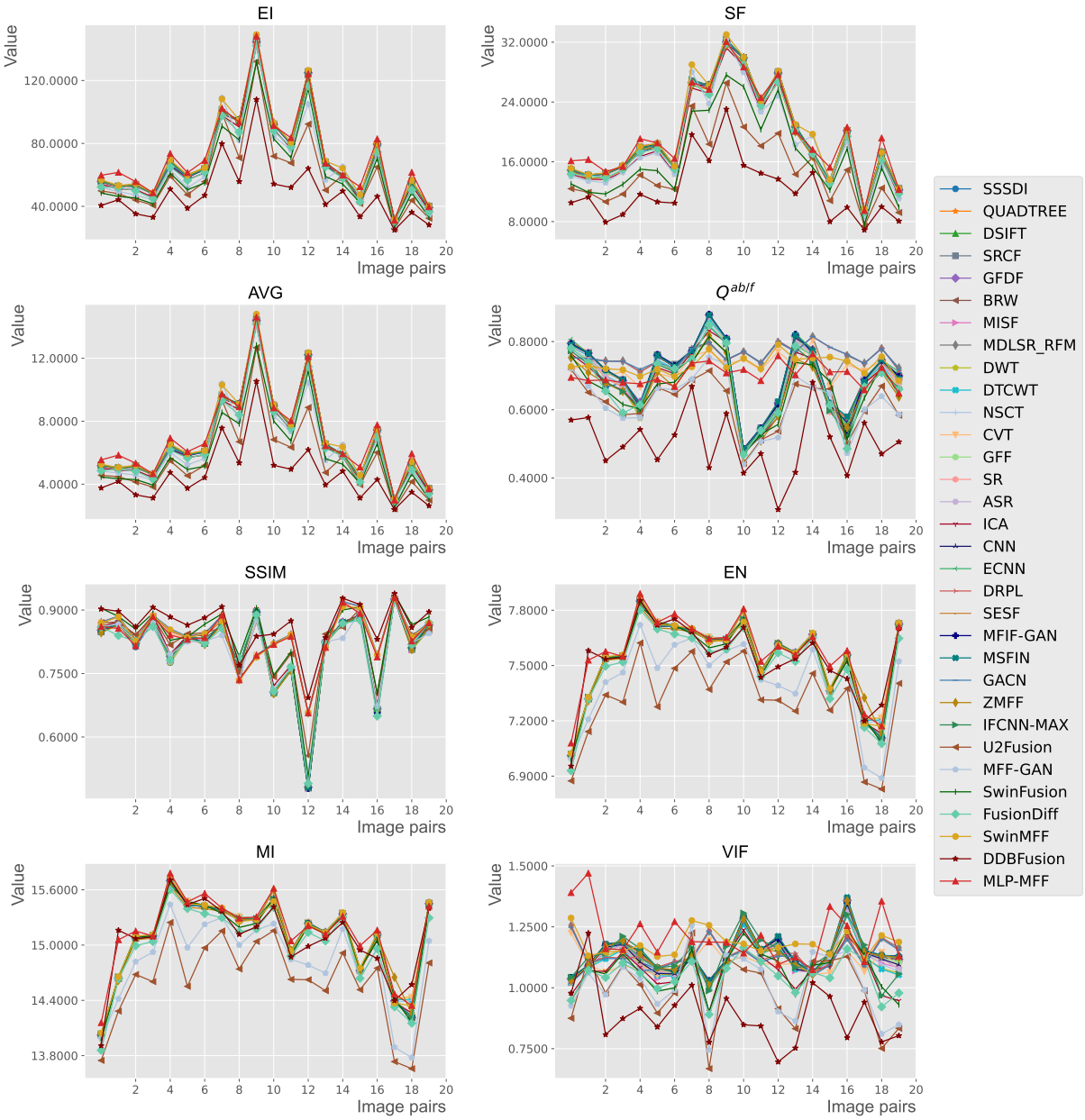

Figure 11. Objective performance of different fusion methods on the Lytro [28] dataset.

Efficiency Analysis. To comprehensively evaluate the computational efficiency of our proposed MLP-MFF, we compare it with various state-of-the-art learning-based MFF methods in terms of model size, computational complexity (FLOPs), and inference time. Note that all FLOPs are calculated on 256 × 256 input images to ensure fair comparison, and the inference time is measured as the average processing time per image on the MFI-WHU dataset [41]. As shown in Table 4, MLP-MFF demonstrates remarkable efficiency advantages across multiple dimensions. In terms of computational complexity, MLP-MFF achieves the lowest FLOPs (0.52G) among all compared methods, representing an 83.12% reduction compared to the previous most efficient method (MFF-GAN with 3.08G FLOPs). Regarding inference speed, MLP-MFF achieves the fastest inference time (0.01s), which is 83.33% faster than the previous fastest method (MFF-GAN with 0.06s). While MLP-MFF's model size (1.23M) is not the smallest among all methods, it remains highly competitive, especially when considering its superior fusion performance demonstrated in Table 1. The model size is significantly smaller than recent Transformer-based methods such as SwinMFF (41.25M), making it more practical for deployment in resource-constrained

environments. The results demonstrate that MLP-MFF successfully achieves an optimal balance between computational efficiency and fusion performance, making it a practical solution for real-world applications.

Table 2. Quantitative comparison of different MFF methods on the MFFW dataset [62].

Methods	EI ↑	SF↑	AVG↑	$Q^{ab/f}$↑	SSIM↑	EN↑	MI↑	VIF↑
Decision map-based deep learning methods								
CNN [19]	73.9624	22.1315	7.4426	0.6216	0.8071	7.1749	14.3497	1.0305
ECNN [77]	75.7644	22.7149	7.6464	**0.7312**	0.8136	7.1904	14.3807	1.0541
DRPL [78]	77.1437	23.2078	7.8088	0.7228	0.8101	7.1944	14.3889	1.0552
SESF [31]	76.9227	23.1558	7.7519	0.6247	0.7947	7.1920	14.3841	1.0542
MFIF-GAN [33]	76.5417	22.9481	7.7225	0.7283	0.8138	7.1570	14.3140	1.0639
MSFIN [20]	75.7969	22.8168	7.6375	0.6183	0.8007	7.1764	14.3528	1.0616
GACN [79]	75.7403	22.5541	7.6327	0.7293	0.8162	7.1951	14.3902	1.0449
ZMFF [35]	77.7055	21.4789	7.6592	0.6541	0.8110	7.1665	14.3329	1.0636
End-to-end methods based on deep learning								
IFCNN-MAX [39]	76.3056	22.1333	7.6334	0.6022	0.8152	7.1710	14.3420	1.0344
U2Fusion [40]	65.7906	16.6017	6.3099	0.5992	0.8178	6.9057	13.8115	0.9189
MFF-GAN [41]	**83.0560**	**28.2025**	**8.4157**	0.4372	0.7482	7.1731	14.3462	**1.2342**
SwinFusion [22]	75.3649	20.5358	7.3528	0.6423	0.8102	7.1419	14.2838	1.1912
FusionDiff [23]	69.6123	21.2969	7.0366	0.6673	0.8198	7.1138	14.2275	0.9052
SwinMFF [21]	80.4903	22.7120	7.9646	0.6636	0.8198	7.1921	14.3843	1.1577
DDBFusion [42]	55.1218	14.7261	5.3471	0.4803	**0.8391**	7.1966	14.3932	0.8952
MLP-MFF	79.5011	22.0943	7.8559	0.6392	0.8222	**7.2152**	**14.4303**	1.1756

Table 3. Quantitative comparison of different MFF methods on the MFI-WHU dataset [41].

Methods	EI ↑	SF↑	AVG↑	$Q^{ab/f}$↑	SSIM↑	EN↑	MI↑	VIF↑
Decision map-based deep learning methods								
CNN [19]	77.0123	26.4975	8.1720	0.7276	0.8310	7.3173	14.6345	1.0959
ECNN [77]	77.9532	26.7520	8.2718	**0.7314**	0.8296	7.3205	14.6411	1.1038
DRPL [78]	78.5301	**26.9109**	8.3340	0.7305	0.8298	7.3222	14.6443	1.1126
SESF [31]	77.6439	26.7527	8.2356	0.7267	0.8293	7.3202	14.6404	1.1078
MFIF-GAN [33]	78.5272	26.9048	8.3274	0.7302	0.8288	7.3247	14.6494	1.1169
MSFIN [20]	77.6764	26.8228	8.2380	0.7273	0.8296	7.3173	14.6345	1.1118
GACN [79]	76.8219	26.5318	8.1374	0.7259	0.8309	7.3141	14.6283	1.1027
ZMFF [35]	**82.0595**	24.9329	**8.3925**	0.6193	0.7947	7.2790	14.5580	1.1742
End-to-end methods based on deep learning								
IFCNN-MAX [39]	79.3862	26.6642	8.3756	0.6936	0.8325	7.3331	14.6662	1.1713
U2Fusion [40]	68.8453	18.1867	6.6806	0.5917	0.8413	7.1315	14.2630	1.1349
SDNet [62]	70.8002	24.2105	7.5039	0.6889	0.8478	7.2619	14.5238	1.0236
MU [62]	76.7037	25.2436	8.0277	0.6496	0.8320	7.2336	14.4673	1.1368
MFF-GAN [41]	68.6117	20.6637	6.9656	0.6777	**0.8620**	7.2894	14.5788	1.1060
SwinFusion [22]	79.1056	21.8127	7.8106	0.5080	0.7185	7.3771	14.7543	1.1136
FusionDiff [23]	72.3067	23.6592	7.5304	0.6762	0.8213	7.2758	14.5516	1.0399
SwinMFF [21]	78.9436	26.3398	8.3138	0.7008	0.8301	7.3274	14.6548	1.1379
DDBFusion [42]	56.5089	17.0766	5.7231	0.5102	0.8510	7.3151	14.6302	1.0139
MLP-MFF	80.9240	25.3620	8.3868	0.6673	0.8291	**7.3804**	**14.7607**	**1.2532**

Table 4. Comparison of computational efficiency across different learning-based MFF methods.

Method	Model Size (M)	FLOPs (G)	Time (s)
Decision map-based methods using deep learning			
CNN [19]	8.76	142.23	0.06
ECNN [77]	1.59	14.93	125.53
DRPL [78]	1.07	140.49	0.22
SESF [31]	0.07	4.90	0.26
MFIF-GAN [33]	3.82	693.03	0.32
MSFIN [20]	4.59	26.76	1.10
GACN [79]	0.07	10.89	0.16
ZMFF [35]	6.33	464.53	165.38
End-to-end methods based on deep learning			
IFCNN-MAX [39]	0.08	8.54	0.09
U2Fusion [40]	0.66	86.40	0.16
MFF-GAN [41]	**0.05**	3.08	0.06
SwinFusion [22]	0.93	63.73	1.79
FusionDiff [23]	26.90	58.13	81.47
SwinMFF [21]	41.25	22.38	0.46
DDBFusion [42]	10.92	184.93	1.69
MLP-MFF	1.23	**0.52**	**0.01**
Reduction (%)	/	83.12%	83.33%

4.3. Performance Under Extreme Situations

To evaluate the robustness of the proposed MLP-MFF model, we performed a challenging experiment using images from the MFI-WHU dataset [41]. We randomly selected three images (as depicted in Figure 12)—and synthesized a corresponding fully blurred version of each using a Gaussian blur filter. This process effectively simulates severe defocus, such as that caused by camera shake or drastically incorrect focus settings. Even when one source image was completely blurred, our model demonstrated a remarkable ability to produce a fully clear output image.

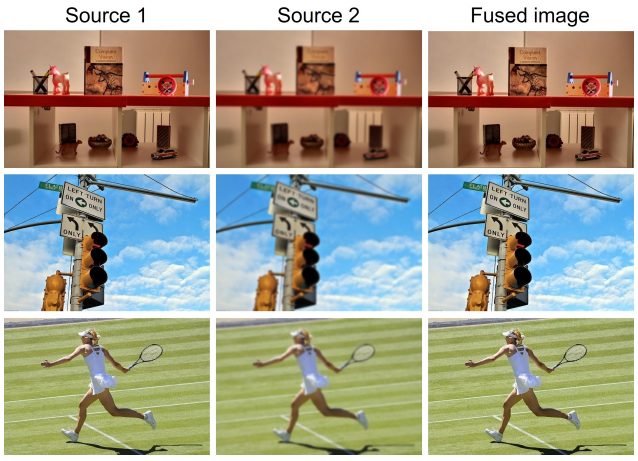

Figure 12. Performance evaluation under extreme conditions. From left to right: original sharp image, synthetically blurred counterpart, and final fusion result.

4.4. Performance on Challenging Scenarios

To further assess our model's robustness, we evaluated its performance on two challenging scenarios: high-contrast scenes and scenes containing small objects. As shown in Figure 13, the first row demonstrates our model's effectiveness on a high-contrast scene, where the dark background and bright sign create significant variations. Our method

successfully merges the focused regions without introducing artifacts or halo effects. The second row illustrates the model's ability to handle small, intricate details, such as the waving flag. The fused image accurately preserves the fine textures of the flag, demonstrating the model's capacity to maintain detail even in complex scenarios. These results highlight the model's general applicability and robustness to diverse image characteristics.

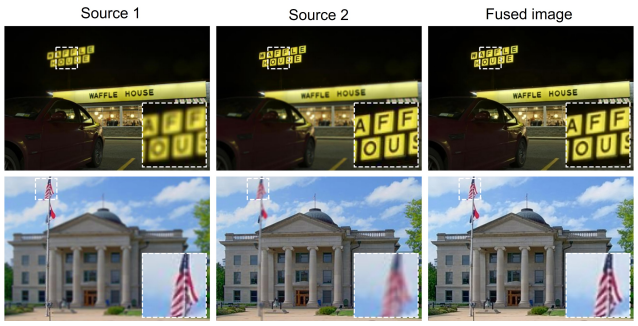

Figure 13. Fusion results on challenging scenarios: the first row shows a high-contrast scene, and the second row shows a scene with small objects.

4.5. Performance in Real-World Scenarios

As demonstrated by the fusion results on the Road-MF dataset [3] in Figure 14, MFF can serve as a crucial preprocessing step to significantly enhance the safety of autonomous driving systems. By fusing multiple images captured with different focal settings, MFF effectively extends the depth of field, ensuring that both near and far objects on the road, such as pedestrians, vehicles, and traffic signs, are simultaneously in sharp focus. This improved clarity provides downstream perception modules with more reliable and detailed visual information. Consequently, this leads to more robust and accurate environmental perception, which is essential for making timely and safe navigation decisions in complex real-world driving conditions. Furthermore, the proposed method can achieve this fusion with a nearly negligible computational overhead and time cost on current automotive-grade processors.

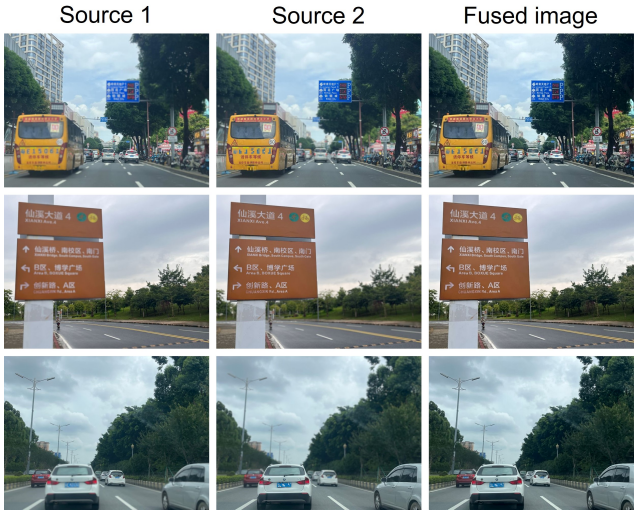

Figure 14. Fusion results on Road-MF [3] dataset.

4.6. Performance of Processing Image Sequences

To demonstrate the scalability of our method for fusing more than two images, we applied it to four image sequences from the Lytro dataset [28]. We adopted a sequential fusion strategy: first, we fused the initial two source images, and then we fused this intermediate result with the third image to generate the final output.

As shown in Figure 15, the fused images successfully preserve all the in-focus regions from the multiple inputs, resulting in a comprehensive, all-in-focus image with excellent visual quality. This experiment confirms that our method can be effectively extended to handle multiple source images by applying it in a sequential manner.

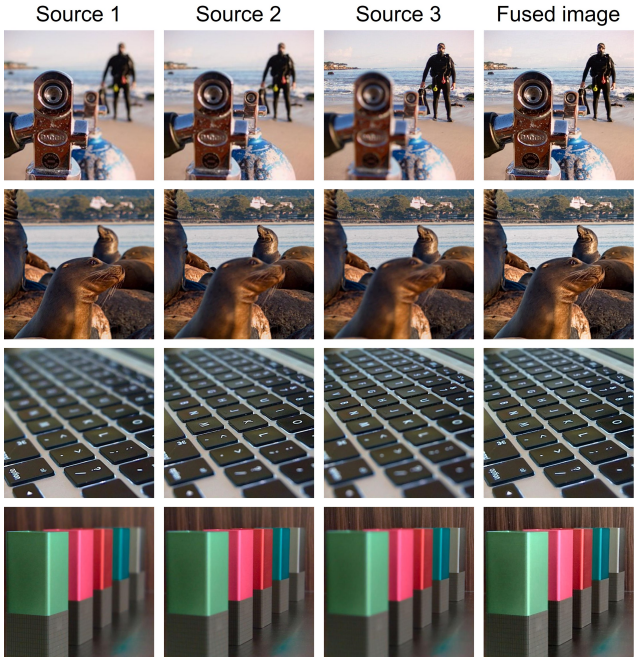

Figure 15. Consecutive multi-focus image-fusion results on Lytro [28] triplet dataset.

4.7. Ablation Study

To validate the effectiveness of our proposed DAMFFM-HA module, we conducted an ablation study on the Lytro dataset, with results presented in Table 5. As observed, when the DAMFFM-HA module is removed and the pixel-wise dot product is directly used as the fusion scheme, the model experiences a performance decline across various metrics. Conversely, incorporating DAMFFM-HA leads to a significant improvement in all metrics, demonstrating the module's substantial role in enhancing the fused image's edge information, structural similarity, and visual information fidelity.

Additionally, we compared the performance of different backbone network architectures for end-to-end MFF. This comparison included UniRepLKNet [80], representing one of the most advanced CNN networks; Swin Transformer [22], representing Transformer-based networks; EVMamba [81], representing vision state space-based networks; and MAXIM [59], another widely used MLP architecture. The results, shown in Table 6, reveal that MLP-MFF, which employs PFMLP as its backbone, achieves optimal results on most metrics compared to other mainstream architectures. This further validates the effectiveness and superiority of PFMLP in MFF tasks.

To further address the architectural design choices, particularly the number of blocks within each stage, we conducted a comparison with various configurations from the original PFMLP paper. This approach, as adopted by the original PFMLP work, is a more common and effective method to analyze the influence of network depth than varying the number of stages. As shown in Table 7, we compare our model ("PFMLP-Ours") with four versions of the PFMLP backbone: PFMLP-N, PFMLP-T, PFMLP-S, and PFMLP-B. The results indicate that while increasing the number of blocks (from N to T, S, and B) leads to a marginal improvement in fusion results, this gain comes with a significant increase in model size (Params) and computational cost (FLOPs). For instance, the PFMLP-B model achieves the highest scores but with a substantial increase in complexity. Our model, which utilizes a

single block per stage, achieves a satisfactory balance between performance and efficiency, demonstrating the rationality of our architectural choice.

In summary, the ablation experiments conclusively demonstrate the significant contribution of both the proposed DAMFFM-HA module and the PFMLP backbone network in enhancing MFF performance.

Table 5. Ablation study on the effectiveness of the proposed DAMFFM-HA module.

Settings	EI ↑	SF↑	AVG↑	Q^{abf}↑	SSIM↑	EN↑	MI↑	VIF↑
w/o DAMFFM-HA	72.5682	19.0216	6.9557	0.6965	0.8253	7.5353	15.0660	1.1772
w DAMFFM-HA	**73.3902**	**19.8181**	**7.0501**	**0.7025**	**0.8344**	**7.5741**	**15.1482**	**1.2126**

Table 6. Comparison of different backbone architectures for end-to-end MFF.

Backbone	EI ↑	SF↑	AVG↑	Q^{abf}↑	SSIM↑	EN↑	MI↑	VIF↑
UniRepLKNet [80]	73.0187	19.5731	6.9837	0.6987	0.8261	7.5381	15.0629	1.1716
Swin Trans. [82]	72.3936	19.5826	6.9265	**0.7215**	0.8288	7.5378	15.0593	1.1783
EVMamba [81]	72.3195	19.4956	6.9474	0.7021	0.8292	7.5351	15.0557	1.1673
MAXIM [59]	72.7428	19.6851	7.0185	0.6985	0.8302	7.5462	15.0926	1.1827
PFMLP-Ours	**73.3902**	**19.8181**	**7.0501**	0.7025	**0.8344**	**7.5741**	**15.1482**	**1.2126**

Table 7. Comparison with several model versions provided in the original PFMLP paper.

Backbone	EI↑	SF↑	AVG↑	Q^{abf}↑	SSIM↑	EN↑	MI↑	VIF↑
PFMLP-Ours	73.3902	19.8181	7.0501	0.7025	0.8344	7.5741	15.1482	1.2126
PFMLP-N	73.3951	19.8223	7.0542	0.7029	0.8348	7.5780	15.1520	1.2130
PFMLP-T	73.3995	19.8265	7.0583	0.7033	0.8352	7.5819	15.1558	1.2134
PFMLP-S	73.4040	19.8307	7.0624	0.7037	0.8356	7.5858	15.1596	1.2138
PFMLP-B	**73.4085**	**19.8349**	**7.0665**	**0.7041**	**0.8360**	**7.5897**	**15.1634**	**1.2142**

5. Discussion

In this study, we explored a MLP-based architecture for MFF. Despite its strengths, our model has two key limitations. First, it relies on synthetically generated defocused images for training, which may not fully represent the complexities of real-world defocus patterns. Second, the architecture is currently optimized for dual-source fusion and would require significant modification to be extended to multi-source or multi-modal scenarios. In the future, we plan to address these limitations by exploring more advanced training techniques to improve performance on real-world data and investigating how to extend this architecture to handle multi-focus image stack fusion.

6. Conclusions

This paper introduces MLP-MFF, a novel lightweight end-to-end MFF network built upon the Pyramid Fusion Multi-Layer Perceptron (PFMLP) architecture. Our method directly addresses common limitations of existing deep learning multi-focus fusion approaches, such as high computational complexity, inflexible input sizes, and restricted global receptive fields. We leverage the inherent strengths of the PFMLP backbone and introduce a Dual-path Adaptive Multi-scale Feature-Fusion Module with Hybrid Attention (DAMFFM-HA), which effectively models both local and global dependencies and adaptively integrates multi-scale features.

Extensive experiments on multiple public MFF datasets demonstrate that MLP-MFF meets or exceeds the fusion quality of current mainstream methods, all while maintaining exceptional efficiency and a lightweight design. Furthermore, MLP-MFF significantly re-

duces computational complexity and inference time, making it highly suitable for practical applications, especially in resource-constrained environments.

Author Contributions: Conceptualization, Y.S., X.X. (Xinzhe Xie) and B.G.; methodology, Y.S., X.X. (Xinzhe Xie) and B.G.; software, Y.S., X.X. (Xinzhe Xie); validation, Y.S., X.X. (Xinzhe Xie), B.G. and X.X. (Xiaofei Xiong); formal analysis, B.G. and X.X. (Xiaofei Xiong); investigation, B.G. and X.X. (Xiaofei Xiong); resources, B.G., X.X. (Xiaofei Xiong) and P.L.; data curation, X.X. (Xinzhe Xie) and X.X. (Xiaofei Xiong); writing—original draft preparation, Y.S. and X.X. (Xinzhe Xie); writing—review and editing, Y.S., X.X. (Xinzhe Xie) and B.G.; visualization, X.X. (Xinzhe Xie); supervision, P.L.; project administration, P.L. and X.X. (Xiaofei Xiong); funding acquisition, B.G. and X.X. (Xiaofei Xiong). All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Science and Technology Development Foundation of South China Sea Bureau, Ministry of Natural Resources (NO. 220101), Project of Sanya Yazhou Bay Science and Technology City (No. SCKJ-JYRC-2023-59), Innovational Fund for Scientific and Technological Personnel of Hainan Province (NO. KJRC2023D19), Research Startup Funding from Hainan Institute of Zhejiang University (NO. 0208-6602-A12204) and the CNOOC Marine Environmental and Ecological Protection Public Welfare Foundation (NO. CF-MEEC/TR/2025-2)

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The code of this study will be publicly accessible in a GitHub repository at https://github.com/Xinzhe99/MLP-MFF (accessed on 3 July 2025).

Acknowledgments: The authors would like to thank Hainan Observation and Research Station of Ecological Environment and Fishery Resource in Yazhou Bay for providing computing resources for this research.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study, in the collection, analyses, or interpretation of data, in the writing of the manuscript, or in the decision to publish the results.

References

1. Xie, X.; Guo, B.; Li, P.; Jiang, Q. Underwater Three-Dimensional Microscope for Marine Benthic Organism Monitoring. In Proceedings of the OCEANS 2024-Singapore, Singapore, 15–18 April 2024; pp. 1–4.
2. Mullen, A.D.; Treibitz, T.; Roberts, P.L.; Kelly, E.L.; Horwitz, R.; Smith, J.E.; Jaffe, J.S. Underwater microscopy for in situ studies of benthic ecosystems. *Nat. Commun.* **2016**, *7*, 12093. [CrossRef]
3. Li, X.; Li, X.; Tan, H.; Li, J. SAMF: Small-area-aware multi-focus image fusion for object detection. In Proceedings of the ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Republic of Korea, 14–19 April 2024; pp. 3845–3849.
4. Chen, Y.; Deng, N.; Xin, B.J.; Xing, W.Y.; Zhang, Z.Y. Structural characterization and measurement of nonwoven fabrics based on multi-focus image fusion. *Measurement* **2019**, *141*, 356–363. [CrossRef]
5. Xie, X.; Qingyan, J.; Chen, D.; Guo, B.; Li, P.; Zhou, S. StackMFF: End-to-end multi-focus image stack fusion network. *Appl. Intell.* **2025**, *55*, 503. [CrossRef]
6. Li, S.; Kwok, J.T.; Wang, Y. Combination of images with diverse focuses using the spatial frequency. *Inf. Fusion* **2001**, *2*, 169–176. [CrossRef]
7. Huang, W.; Jing, Z. Multi-focus image fusion using pulse coupled neural network. *Pattern Recognit. Lett.* **2007**, *28*, 1123–1132. [CrossRef]
8. Li, M.; Cai, W.; Tan, Z. A region-based multi-sensor image fusion scheme using pulse-coupled neural network. *Pattern Recognit. Lett.* **2006**, *27*, 1948–1956. [CrossRef]
9. Liu, Y.; Liu, S.; Wang, Z. Multi-focus image fusion with dense SIFT. *Inf. Fusion* **2015**, *23*, 139–155. [CrossRef]
10. Aslantas, V.; Toprak, A.N. Multi-focus image fusion based on optimal defocus estimation. *Comput. Electr. Eng.* **2017**, *62*, 302–318. [CrossRef]
11. Toet, A. Image fusion by a ratio of low-pass pyramid. *Pattern Recognit. Lett.* **1989**, *9*, 245–253. [CrossRef]

12. Burt, P.J.; Kolczynski, R.J. Enhanced image capture through fusion. In Proceedings of the 1993 (4th) International Conference on Computer Vision, Berlin, Germany, 11–14 May 1993; pp. 173–182.
13. Lewis, J.J.; O'Callaghan, R.J.; Nikolov, S.G.; Bull, D.R.; Canagarajah, N. Pixel-and region-based image fusion with complex wavelets. *Inf. Fusion* **2007**, *8*, 119–130. [CrossRef]
14. Li, H.; Manjunath, B.; Mitra, S.K. Multisensor image fusion using the wavelet transform. *Graph. Model. Image Process.* **1995**, *57*, 235–245. [CrossRef]
15. Wang, M.; Shang, X. A fast image fusion with discrete cosine transform. *IEEE Signal Process. Lett.* **2020**, *27*, 990–994. [CrossRef]
16. Zhang, Q.; Guo, B.l. Multifocus image fusion using the nonsubsampled contourlet transform. *Signal Process.* **2009**, *89*, 1334–1346. [CrossRef]
17. Liu, Z.; Chai, Y.; Yin, H.; Zhou, J.; Zhu, Z. A novel multi-focus image fusion approach based on image decomposition. *Inf. Fusion* **2017**, *35*, 102–116. [CrossRef]
18. Jiang, Y.; Wang, M. Image fusion with morphological component analysis. *Inf. Fusion* **2014**, *18*, 107–118. [CrossRef]
19. Liu, Y.; Chen, X.; Peng, H.; Wang, Z. Multi-focus image fusion with a deep convolutional neural network. *Inf. Fusion* **2017**, *36*, 191–207. [CrossRef]
20. Liu, Y.; Wang, L.; Cheng, J.; Chen, X. Multiscale feature interactive network for multifocus image fusion. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 1–16. [CrossRef]
21. Xie, X.; Guo, B.; Li, P.; He, S.; Zhou, S. SwinMFF: Toward high-fidelity end-to-end multi-focus image fusion via swin transformer-based network. *Vis. Comput.* **2024**, *41*, 3883–3906. [CrossRef]
22. Ma, J.; Tang, L.; Fan, F.; Huang, J.; Mei, X.; Ma, Y. SwinFusion: Cross-domain long-range learning for general image fusion via swin transformer. *IEEE/CAA J. Autom. Sin.* **2022**, *9*, 1200–1217. [CrossRef]
23. Li, M.; Pei, R.; Zheng, T.; Zhang, Y.; Fu, W. FusionDiff: Multi-focus image fusion using denoising diffusion probabilistic models. *Expert Syst. Appl.* **2024**, *238*, 121664. [CrossRef]
24. Xie, X.; Guo, B.; Li, P.; He, S.; Zhou, S. Multi-focus image fusion with visual state space model and dual adversarial learning. *Comput. Electr. Eng.* **2025**, *123*, 110238. [CrossRef]
25. Jin, X.; Zhu, P.; Yu, D.; Wozniak, M.; Jiang, Q.; Wang, P.; Zhou, W. Combining depth and frequency features with Mamba for multi-focus image fusion. *Inf. Fusion* **2025**, 103355. [CrossRef]
26. Huang, Q.; Jie, Z.; Ma, L.; Shen, L.; Lai, S. A Pyramid Fusion MLP for Dense Prediction. *IEEE Trans. Image Process.* **2025**, *34*, 455–467. [CrossRef] [PubMed]
27. Chen, S.; Xie, E.; Ge, C.; Chen, R.; Liang, D.; Luo, P. Cyclemlp: A mlp-like architecture for dense visual predictions. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 14284–14300. [CrossRef]
28. Nejati, M.; Samavi, S.; Shirani, S. Multi-focus image fusion using dictionary-based sparse representation. *Inf. Fusion* **2015**, *25*, 72–84. [CrossRef]
29. Lv, M.; Song, S.; Jia, Z.; Li, L.; Ma, H. Multi-Focus Image Fusion Based on Dual-Channel Rybak Neural Network and Consistency Verification in NSCT Domain. *Fractal Fract.* **2025**, *9*, 432. [CrossRef]
30. Xiao, B.; Xu, B.; Bi, X.; Li, W. Global-feature encoding U-Net (GEU-Net) for multi-focus image fusion. *IEEE Trans. Image Process.* **2020**, *30*, 163–175. [CrossRef]
31. Ma, B.; Zhu, Y.; Yin, X.; Ban, X.; Huang, H.; Mukeshimana, M. Sesf-fuse: An unsupervised deep model for multi-focus image fusion. *Neural Comput. Appl.* **2021**, *33*, 5793–5804. [CrossRef]
32. Ma, J.; Le, Z.; Tian, X.; Jiang, J. SMFuse: Multi-focus image fusion via self-supervised mask-optimization. *IEEE Trans. Comput. Imaging* **2021**, *7*, 309–320. [CrossRef]
33. Wang, Y.; Xu, S.; Liu, J.; Zhao, Z.; Zhang, C.; Zhang, J. MFIF-GAN: A new generative adversarial network for multi-focus image fusion. *Signal Process. Image Commun.* **2021**, *96*, 116295. [CrossRef]
34. Duan, Z.; Luo, X.; Zhang, T. Combining transformers with CNN for multi-focus image fusion. *Expert Syst. Appl.* **2024**, *235*, 121156. [CrossRef]
35. Hu, X.; Jiang, J.; Liu, X.; Ma, J. ZMFF: Zero-shot multi-focus image fusion. *Inf. Fusion* **2023**, *92*, 127–138. [CrossRef]
36. Shao, X.; Jin, X.; Jiang, Q.; Miao, S.; Wang, P.; Chu, X. Multi-focus image fusion based on transformer and depth information learning. *Comput. Electr. Eng.* **2024**, *119*, 109629. [CrossRef]
37. Quan, Y.; Wan, X.; Tang, Z.; Liang, J.; Ji, H. Multi-Focus Image Fusion via Explicit Defocus Blur Modelling. *Proc. AAAI Conf. Artif. Intell.* **2025**, *39*, 6657–6665. [CrossRef]
38. Zhai, H.; Zhang, G.; Zeng, Z.; Xu, Z.; Fang, A. LSKN-MFIF: Large selective kernel network for multi-focus image fusion. *Neurocomputing* **2025**, *635*, 129984. [CrossRef]
39. Zhang, Y.; Liu, Y.; Sun, P.; Yan, H.; Zhao, X.; Zhang, L. IFCNN: A general image fusion framework based on convolutional neural network. *Inf. Fusion* **2020**, *54*, 99–118. [CrossRef]
40. Xu, H.; Ma, J.; Jiang, J.; Guo, X.; Ling, H. U2Fusion: A unified unsupervised image fusion network. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *44*, 502–518. [CrossRef]

41. Zhang, H.; Le, Z.; Shao, Z.; Xu, H.; Ma, J. MFF-GAN: An unsupervised generative adversarial network with adaptive and gradient joint constraints for multi-focus image fusion. *Inf. Fusion* **2021**, *66*, 40–53. [CrossRef]
42. Zhang, Z.; Li, H.; Xu, T.; Wu, X.J.; Kittler, J. DDBFusion: An unified image decomposition and fusion framework based on dual decomposition and Bézier curves. *Inf. Fusion* **2025**, *114*, 102655. [CrossRef]
43. Wang, X.; Fang, L.; Zhao, J.; Pan, Z.; Li, H.; Li, Y. MMAE: A universal image fusion method via mask attention mechanism. *Pattern Recognit.* **2025**, *158*, 111041. [CrossRef]
44. Yang, B.; Jiang, Z.; Pan, D.; Yu, H.; Gui, G.; Gui, W. LFDT-Fusion: A latent feature-guided diffusion Transformer model for general image fusion. *Inf. Fusion* **2025**, *113*, 102639. [CrossRef]
45. Tolstikhin, I.O.; Houlsby, N.; Kolesnikov, A.; Beyer, L.; Zhai, X.; Unterthiner, T.; Yung, J.; Steiner, A.; Keysers, D.; Uszkoreit, J.; et al. Mlp-mixer: An all-mlp architecture for vision. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 24261–24272.
46. Touvron, H.; Bojanowski, P.; Caron, M.; Cord, M.; El-Nouby, A.; Grave, E.; Izacard, G.; Joulin, A.; Synnaeve, G.; Verbeek, J.; et al. Resmlp: Feedforward networks for image classification with data-efficient training. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 5314–5321. [CrossRef] [PubMed]
47. Ding, X.; Chen, H.; Zhang, X.; Han, J.; Ding, G. Repmlpnet: Hierarchical vision mlp with re-parameterized locality. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18– 24 June 2022; pp. 578–587.
48. Guo, M.H.; Liu, Z.N.; Mu, T.J.; Hu, S.M. Beyond self-attention: External attention using two linear layers for visual tasks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 5436–5447. [CrossRef] [PubMed]
49. Hou, Q.; Jiang, Z.; Yuan, L.; Cheng, M.M.; Yan, S.; Feng, J. Vision permutator: A permutable mlp-like architecture for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 1328–1334. [CrossRef]
50. Tang, C.; Zhao, Y.; Wang, G.; Luo, C.; Xie, W.; Zeng, W. Sparse MLP for image recognition: Is self-attention really necessary? *Proc. AAAI Conf. Artif. Intell.* **2022**, *36*, 2344–2351. [CrossRef]
51. Cao, G.; Luo, S.; Huang, W.; Lan, X.; Jiang, D.; Wang, Y.; Zhang, J. Strip-MLP: Efficient token interaction for vision MLP. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023; pp. 1494–1504.
52. Lian, D.; Yu, Z.; Sun, X.; Gao, S. As-mlp: An axial shifted mlp architecture for vision. *arXiv* **2021**, arXiv:2107.08391
53. Yu, T.; Li, X.; Cai, Y.; Sun, M.; Li, P. S2-mlp: Spatial-shift mlp architecture for vision. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2022; pp. 297–306.
54. Wang, G.; Zhao, Y.; Tang, C.; Luo, C.; Zeng, W. When shift operation meets vision transformer: An extremely simple alternative to attention mechanism. *Proc. AAAI Conf. Artif. Intell.* **2022**, *36*, 2423–2430. [CrossRef]
55. Tang, Y.; Han, K.; Guo, J.; Xu, C.; Li, Y.; Xu, C.; Wang, Y. An image patch is a wave: Phase-aware vision mlp. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 10935–10944.
56. Wei, G.; Zhang, Z.; Lan, C.; Lu, Y.; Chen, Z. Active token mixer. *Proc. AAAI Conf. Artif. Intell.* **2023**, *37*, 2759–2767. [CrossRef]
57. Lai, S.; Du, X.; Guo, J.; Zhang, K. RaMLP: Vision MLP via Region-aware Mixing. In Proceedings of the IJCAI, Macao, China, 19–25 August 2023; pp. 999–1007.
58. Liu, S.; Li, S.; Qu, L.; Wang, M.; Song, Z. Fusionmlp: A Mlp-Based Unified Image Fusion Framework. *Available at SSRN 4687800* **2023**.
59. Tu, Z.; Talebi, H.; Zhang, H.; Yang, F.; Milanfar, P.; Bovik, A.; Li, Y. Maxim: Multi-axis mlp for image processing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 5769–5780.
60. Cai, Z.; Ma, Y.; Huang, J.; Mei, X.; Fan, F.; Zhao, Z. CMFuse: Cross-Modal Features Mixing via Convolution and MLP for Infrared and Visible Image Fusion. *IEEE Sensors J.* **2024**, *24*, 24152–24167. [CrossRef]
61. Wang, L.; Lu, H.; Wang, Y.; Feng, M.; Wang, D.; Yin, B.; Ruan, X. Learning to detect salient objects with image-level supervision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 136–145.
62. Xu, S.; Wei, X.; Zhang, C.; Liu, J.; Zhang, J. MFFW: A new dataset for multi-focus image fusion. *arXiv* **2020**, arXiv:2002.04780. [CrossRef]
63. Guo, D.; Yan, J.; Qu, X. High quality multi-focus image fusion using self-similarity and depth information. *Opt. Commun.* **2015**, *338*, 138–144. [CrossRef]
64. De, I.; Chanda, B. Multi-focus image fusion using a morphology-based focus measure in a quad-tree structure. *Inf. Fusion* **2013**, *14*, 136–146. [CrossRef]
65. Liu, Y.; Liu, S.; Wang, Z. A general framework for image fusion based on multi-scale transform and sparse representation. *Inf. Fusion* **2015**, *24*, 147–164. [CrossRef]
66. Qiu, X.; Li, M.; Zhang, L.; Yuan, X. Guided filter-based multi-focus image fusion through focus region detection. *Signal Process. Image Commun.* **2019**, *72*, 35–46. [CrossRef]
67. Ma, J.; Zhou, Z.; Wang, B.; Miao, L.; Zong, H. Multi-focus image fusion using boosted random walks-based algorithm with two-scale focus maps. *Neurocomputing* **2019**, *335*, 9–20. [CrossRef]
68. Zhan, K.; Kong, L.; Liu, B.; He, Y. Multimodal image seamless fusion. *J. Electron. Imaging* **2019**, *28*, 023027. [CrossRef]

69. Wang, J.; Qu, H.; Zhang, Z.; Xie, M. New insights into multi-focus image fusion: A fusion method based on multi-dictionary linear sparse representation and region fusion model. *Inf. Fusion* **2024**, *105*, 102230. [CrossRef]
70. Rockinger, O. Image sequence fusion using a shift-invariant wavelet transform. In Proceedings of the International Conference on Image Processing, Santa Barbara, CA, USA, 26–29 October 1997; Volume 3, pp. 288–291.
71. Yang, B.; Li, S.; Sun, F. Image fusion using nonsubsampled contourlet transform. In Proceedings of the Fourth International Conference on Image and Graphics (ICIG 2007), Chengdu, China, 22–24 August 2007; pp. 719–724.
72. Haghighat, M.B.A.; Aghagolzadeh, A.; Seyedarabi, H. Multi-focus image fusion for visual sensor networks in DCT domain. *Comput. Electr. Eng.* **2011**, *37*, 789–797. [CrossRef]
73. Li, S.; Kang, X.; Hu, J. Image fusion with guided filtering. *IEEE Trans. Image Process.* **2013**, *22*, 2864–2875. [CrossRef]
74. Olshausen, B.A.; Field, D.J. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* **1996**, *381*, 607–609. [CrossRef]
75. Liu, Y.; Wang, Z. Simultaneous image fusion and denoising with adaptive sparse representation. *IET Image Process.* **2015**, *9*, 347–357. [CrossRef]
76. Paul, S.; Sevcenco, I.S.; Agathoklis, P. Multi-exposure and multi-focus image fusion in gradient domain. *J. Circuits, Syst. Comput.* **2016**, *25*, 1650123. [CrossRef]
77. Amin-Naji, M.; Aghagolzadeh, A.; Ezoji, M. Ensemble of CNN for multi-focus image fusion. *Inf. Fusion* **2019**, *51*, 201–214. [CrossRef]
78. Li, J.; Guo, X.; Lu, G.; Zhang, B.; Xu, Y.; Wu, F.; Zhang, D. DRPL: Deep regression pair learning for multi-focus image fusion. *IEEE Trans. Image Process.* **2020**, *29*, 4816–4831. [CrossRef] [PubMed]
79. Ma, B.; Yin, X.; Wu, D.; Shen, H.; Ban, X.; Wang, Y. End-to-end learning for simultaneously generating decision map and multi-focus image fusion result. *Neurocomputing* **2022**, *470*, 204–216. [CrossRef]
80. Ding, X.; Zhang, Y.; Ge, Y.; Zhao, S.; Song, L.; Yue, X.; Shan, Y. Unireplknet: A universal perception large-kernel convnet for audio video point cloud time-series and image recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 17–21 June 2024; pp. 5513–5524.
81. Pei, X.; Huang, T.; Xu, C. Efficientvmamba: Atrous selective scan for light weight visual mamba. In Proceedings of the AAAI Conference on Artificial Intelligence, Philadelphia, PA, USA, 25 February–4 March 2025; Volume 39, pp. 6443–6451.
82. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 10012–10022.