applied sciences

MDPI

Article

LightMFF: A Simple and Efficient Ultra-Lightweight Multi-Focus Image Fusion Network

Xinzhe Xie [1,†], Zijian Lin [1,2,†], Buyu Guo [2,3,*], Shuangyan He [1,2], Yanzhen Gu [1], Yefei Bai [1,2,*] and Peiliang Li [1,2,4]

[1] State Key Laboratory of Ocean Sensing & Ocean College, Zhejiang University, Zhoushan 316021, China; xiexinzhe@zju.edu.cn (X.X.); linzj@zju.edu.cn (Z.L.); hesy@zju.edu.cn (S.H.); guyanzhen@zju.edu.cn (Y.G.); lipeiliang@zju.edu.cn (P.L.)
[2] Hainan Institute, Zhejiang University, Sanya 572025, China
[3] Donghai Laboratory, Zhoushan 316021, China
[4] Hainan Observation and Research Station of Ecological Environment and Fishery Resource in Yazhou Bay, Sanya 572024, China
* Correspondence: guobuyuwork@163.com (B.G.); yfbai@zju.edu.cn (Y.B.)
† These authors contributed equally to this work.

Abstract

In recent years, deep learning-based multi-focus image fusion (MFF) methods have demonstrated remarkable performance. However, their reliance on complex network architectures often demands substantial computational resources, limiting practical applications. To address this, we propose LightMFF, an ultra-lightweight fusion network that achieves superior performance with minimal computational overhead. Our core insight is to reformulate the multi-focus fusion problem from a classification perspective to a refinement perspective, where coarse initial decision maps and explicit edge information are leveraged to guide the final decision map generation. This novel formulation enables a significantly simplified architecture, requiring only 0.02 M parameters while maintaining state-of-the-art fusion quality. Extensive experiments demonstrate that LightMFF achieves real-time performance at 0.02 s per image pair with merely 0.06 G FLOPs, representing a 98.05% reduction in computational cost compared to prior approaches. Crucially, LightMFF consistently surpasses existing methods across standard fusion quality metrics.

Keywords: deep learning; multi-focus image fusion; focus property detection; lightweight network

Academic Editor: Pedro Alexandre Mogadouro do Couto

Received: 11 June 2025
Revised: 28 June 2025
Accepted: 2 July 2025
Published: 3 July 2025

Citation: Xie, X.; Lin, Z.; Guo, B.; He, S.; Gu, Y.; Bai, Y.; Li, P. LightMFF: A Simple and Efficient Ultra-Lightweight Multi-Focus Image Fusion Network. *Appl. Sci.* **2025**, *15*, 7500. https://doi.org/10.3390/app15137500

1. Introduction

Multi-focus image fusion (MFF) is a crucial technique in optical imaging that addresses the limited depth of field by integrating multiple source images, each focused at a different depth, into a single all-in-focus representation. By extracting and combining the optimally focused regions, MFF algorithms produce a visually coherent and comprehensive fused image. This approach has broad applications, including microscopy [1], textile quality inspection [2], and autonomous driving [3].

Traditional MFF methods are categorized into two main groups: transform domain-based and spatial domain-based approaches. Transform domain methods, such as multi-scale decomposition and sparse representation, project images into alternative domains for processing [4–9]. While effective, these methods often suffer from intensity inconsistencies [10] and noise sensitivity [11]. Spatial domain methods directly process pixels using

techniques like block-based [12,13], region-based [14,15], and pixel-level strategies [16,17]. However, these methods typically rely on handcrafted features that lack expressiveness, especially in homogeneous regions [18], and can introduce artifacts along object boundaries [19]. More recent advancements, such as MDLSR_RFM [20], have improved traditional methods by using multi-dictionary linear sparse representation for focus detection and edge-weighting matting for boundary refinement. PADCDTNP [21] introduced a self-adjusting, dual-channel neural system for more accurate focused region identification. Despite these improvements, the increasing complexity of traditional methods leads to greater computational demands.

The advent of deep learning has revolutionized MFF. By leveraging the parallel processing capabilities of modern GPUs, deep learning-based methods have surpassed many traditional approaches in both fusion quality and processing efficiency. These methods learn from large datasets, automating feature extraction and fusion and eliminating the need for complex, handcrafted schemes.

Early deep learning methods framed MFF as a classification task. They used convolutional neural networks (CNNs) to generate a decision map for subsequent fusion [10,19,22,23]. By guiding network optimization with annotated decision maps, these approaches significantly reduced task complexity and improved fusion results within the same computational budget. This fusion paradigm remains a popular research topic due to its interpretability, staged optimization capabilities, and synergy with traditional image processing techniques. More recently, a representative and innovative study introduced ZMFF [11], an unsupervised and untrained deep learning model for MFF that uses a deep image prior and a deep mask prior to effectively fuse images without requiring training data. However, this method requires multiple optimization iterations, leading to very low processing efficiency and limited practical utility.

Subsequently, the focus has shifted to developing general end-to-end MFF frameworks [24–28]. These methods aim to perform the entire fusion task—from feature extraction to reconstruction—within a single network, which eliminates the need for decision map post-processing and better handles complex scenes. However, to enhance fusion quality, some of these newer methods have sacrificed computational efficiency. For example, SwinMFF [26] achieves high pixel-value fidelity but has high computational requirements due to its self-attention mechanism. The multi-step sampling of diffusion model-based methods like FusionDiff [29] places significant strain on the hardware. The resulting high computational demands, energy consumption, and hardware costs severely limit the practical deployment of these models in resource-constrained environments, such as autonomous driving systems and edge devices.

To address the high complexity and computational overhead of end-to-end models, which can cause latency in real-time systems, we propose an ultra-lightweight MFF network with minimal computational overhead. We adopt a decision-map-based fusion paradigm to achieve this goal. Our major contributions are summarized as follows:

1. We introduce a novel task transformation paradigm that reframes MFF from direct focus classification to a decision map refinement process. By providing an initial coarse decision map and explicit edge maps as prior guidance and boundary cues, our approach simplifies network optimization and improves fusion quality.
2. We present ULU-Net, a novel U-Net variant designed specifically for MFF. This architecture achieves competitive performance with an ultra-low parameter count, significantly reducing model complexity.
3. We propose LightMFF, an ultra-lightweight MFF framework that combines our novel task transformation paradigm with the ULU-Net architecture. Extensive quantitative and qualitative experiments demonstrate that LightMFF achieves state-of-the-art

fusion quality while requiring significantly fewer parameters and computational resources than existing methods.

The remainder of this paper is organized as follows: Section 2 reviews related works, covering traditional, deep learning-based, and lightweight MFF methods. Section 3 details our proposed LightMFF method, including its theoretical foundations, network architecture, and training strategy. Section 4 presents comprehensive experimental results and comparisons with state-of-the-art methods. Section 5 discusses the limitations and future directions of our work. Finally, Section 6 concludes the paper.

2. Related Works

2.1. Decision Map-Based Deep Learning Methods

Decision map-based deep learning methods have significantly advanced MFF by providing interpretable and precise fusion guidance. This approach typically involves a neural network learning to identify focused or defocused regions, which then informs the final fusion. A notable advantage of this approach is that the models are generally smaller than end-to-end fusion networks. This is because they only need to learn to classify or regress a decision map for each pixel or patch, which is a simpler task than learning to directly synthesize a complete fused image. Consequently, these networks are more efficient and easier to train.

A pioneering work by Liu et al. in 2017 [22] introduced a CNN-based method to generate decision maps through binary patch classification, laying the groundwork for deep learning in MFF. Subsequent efforts enhanced this foundation: ECNN [23] utilizes an ensemble of CNNs for more robust focus detection, while DRPL [19] reframes focus detection as a regression task to learn continuous focus measurements, significantly improving accuracy in boundary regions.

Recent years have seen further innovations in this paradigm. Unsupervised learning frameworks like SESF [30] emerged, eliminating the need for labeled data. Generative adversarial networks (GANs) were leveraged in MFIF-GAN [31] to improve decision map quality. MSFIN [10] introduced multi-scale feature interaction for more precise fusion. ZMFF [11] proposed a zero-shot learning framework using a deep image prior, enhancing generalization to unseen scenes.

Most recently, MFIF-STCU-Net [32] proposed synthesizing training datasets using a depth estimation model and achieved state-of-the-art performance with a multi-level feature fusion (MFF) model that uses a hybrid U-Net and Transformer architecture. DMANet [33] introduced a novel framework that integrates explicit defocus blur modeling into the MFF process, improving both interpretability and performance. LSKN-MFIF [34] is a new MFF network that uses a dynamically adjusting "large selective kernel" module to capture both global and local features, leading to more accurate fusion results that outperform existing methods.

2.2. End-to-End Deep Learning Methods

End-to-end deep learning methods for MFF directly learn the mapping from source images to the fused result, bypassing the explicit generation of a decision map. This paradigm gained significant traction around 2020 with the introduction of IFCNN [24], a general image fusion framework, and U2Fusion [25], a unified unsupervised framework that enhanced training flexibility.

Following these foundational works, subsequent advancements have introduced diverse architectures and learning strategies. In 2021, SDNet [35] achieved real-time performance through a squeeze-and-decomposition network, while MFF-GAN [36] leveraged unsupervised generative adversarial networks to improve visual quality. The year 2022 saw

the integration of Transformer-based architectures with SwinFusion [27], which effectively captured long-range dependencies and significantly enhanced fusion performance for complex scenes.

More recent developments have pushed the boundaries of this field. MUFusion [37] introduced a self-evolutionary training formula with a novel memory unit, leveraging intermediate fusion results for supervision. Further explorations include SwinMFF [26] for enhanced pixel-level fidelity and FusionDiff [29], which applies denoising diffusion probabilistic models to MFF. The latest works address specific limitations in the field: DDBFusion [28] improves feature extraction with a dual decomposition approach, and StackMFF [38] extends MFF to image stacks using 3D CNNs and synthesized training data. Additionally, MMAE [39] introduced a mask attention mechanism to filter redundant information, and LFDT-Fusion [40] proposed an efficient latent feature-guided diffusion model that operates in a compressed latent space.

However, a core challenge for end-to-end methods is their inherent demand for complex networks and substantial computational resources. Unlike decision map-based approaches that decompose the fusion problem, these models must simultaneously perform both feature extraction and image reconstruction, a significantly more complex task. This makes them less suitable for lightweight applications. For instance, Transformer-based approaches like SwinMFF [26] and SwinFusion [27] often require tens of millions of parameters. Similarly, diffusion-based methods such as FusionDiff [29] typically demand extensive computational power due to their iterative nature. Even relatively compact architectures like IFCNN [24] require a considerable number of parameters to maintain fusion quality. This fundamental limitation stems directly from the end-to-end paradigm, where generating a pixel-perfect fusion result from source images is an inherently complex task, posing significant challenges for ultra-lightweight deployments. For resource-constrained scenarios, decision map-based methods offer a more promising direction by decomposing the fusion problem into more manageable sub-tasks.

2.3. Lightweight Multi-Focus Image Fusion Networks

While deep learning-based MFF approaches have shown great promise, their often complex architectures can be computationally demanding, hindering practical application. This has led to an increasing focus on developing lightweight MFF networks. For example, Jin et al. [41] proposed a lightweight scheme that leverages the Laplacian pyramid transform and an adaptive pulse-coupled neural network with local spatial frequency. This approach aimed to improve fusion efficiency and performance by processing fewer sub-images. Similarly, Zhou et al. [42] introduced LNMF. By classifying in-focus and out-of-focus regions, LNMF significantly reduces computational complexity and memory consumption. More recently, Nie et al. [43] presented MLNet, a multi-domain lightweight network. MLNet employs discrete cosine transform-based and local binary pattern-based convolutions to effectively capture both frequency and spatial domain features with a minimal parameter count.

In the field of multi-modal image fusion, which is closely related to the field of MFF, there are also works focusing on lightweight fusion networks. For example, Wu et al. [44] presented a lightweight image fusion algorithm specifically designed for merging visible light and infrared images by utilizing Depthwise Separable Convolution. In the field of medical image fusion, MACTFusion [45] proposed a novel lightweight cross Transformer based on a cross multi-axis attention mechanism. They mainly perform as lightweight on the network, and the proposed method not only performs as lightweight on the network but also indirectly allows the use of a lighter network through the transformation of the fusion paradigm.

Building upon these efforts, our proposed LightMFF also focuses on an ultra-lightweight architecture. Figure 1 presents a comparison between our proposed method and several prominent learning-based existing MFF methods, evaluating both fusion quality and processing efficiency. To the best of our knowledge, the proposed LightMFF is currently the best method with the lowest computational requirements while maintaining good overall performance.

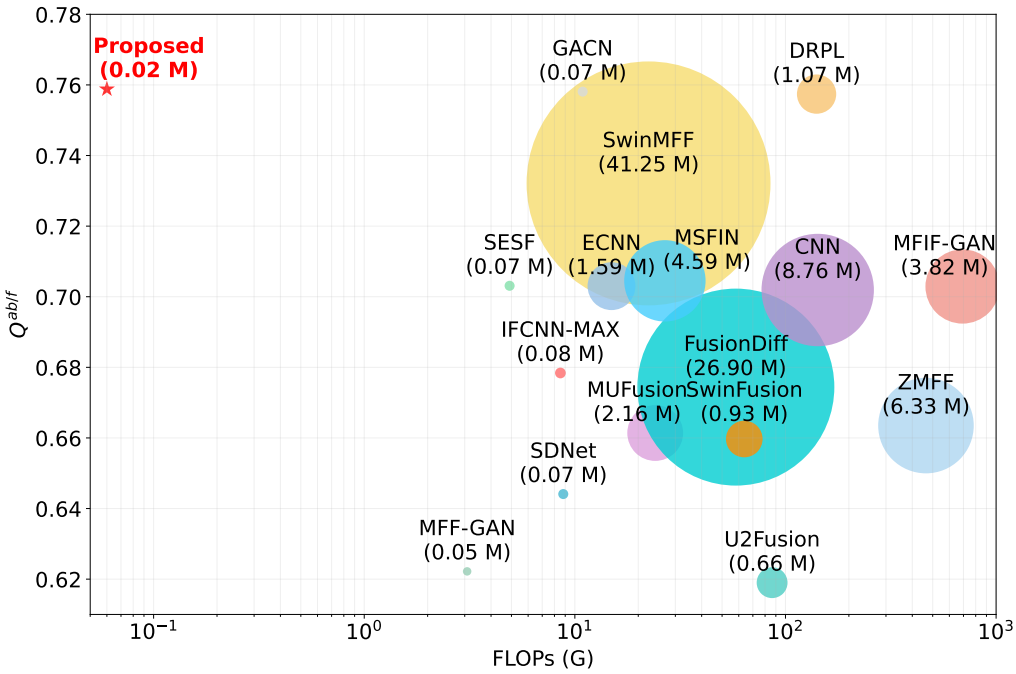

Figure 1. Quantitative comparison of model size (indicated by bubble size), computational complexity (FLOPs), and fusion quality ($Q^{ab/f}$) among different deep learning-based MFF methods on the Lytro dataset [46].

3. Method

3.1. Motivation

The primary objective of decision map-based MFF is to generate a binary decision map that accurately reflects the focus properties of source images. Conventionally, networks are trained to distinguish between focused and defocused regions through loss optimization. We propose a novel paradigm that transforms this learning objective by introducing an initial coarse decision map as prior guidance. This approach shifts the network's task from direct focus/defocus classification to decision map refinement, where the network learns to optimize an imperfect decision map using the source images as a reference. This task transformation significantly reduces the learning complexity and, consequently, the required model capacity. Furthermore, by incorporating explicit edge maps as additional input, we provide crucial boundary cues that further simplify the optimization process, enabling an even more compact network architecture. The proposed fusion framework is illustrated in Figure 2.

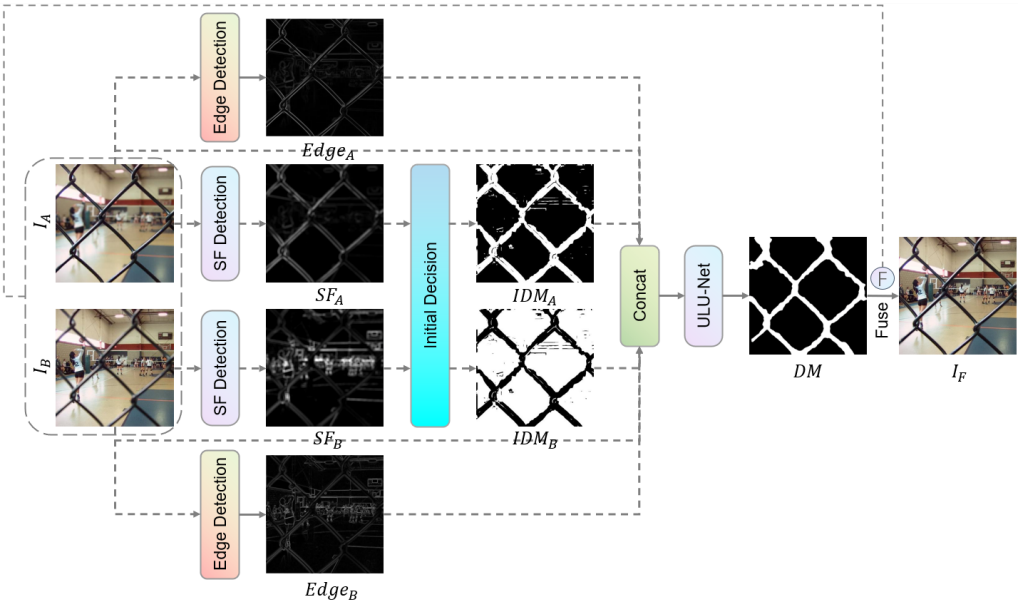

Figure 2. Architecture of the proposed method.

3.2. Focus Property and Edge Detection

To provide prior guidance through an initial coarse decision map, we adopt the pixel-level spatial frequency gradient, as proposed in SESF-Fuse [30], as our initial focus measure. The computational formulation is as follows:

$$\mathrm{RF}_{i(x,y)} = \sqrt{\sum_{a=-r}^{r}\sum_{b=-r}^{r}\left[I_{i(x+a,y+b)} - I_{i(x+a,y+b-1)}\right]^2} \tag{1}$$

$$\mathrm{CF}_{i(x,y)} = \sqrt{\sum_{a=-r}^{r}\sum_{b=-r}^{r}\left[I_{i(x+a,y+b)} - I_{i(x+a-1,y+b)}\right]^2} \tag{2}$$

$$\mathrm{SF}_{i(x,y)} = \sqrt{\frac{\left(\mathrm{CF}_{i(x,y)}\right)^2 + \left(\mathrm{RF}_{i(x,y)}\right)^2}{(2r+1)^2}} \tag{3}$$

where RF and CF represent the row and column frequency vectors, respectively, and r denotes the kernel radius. By computing the pixel-level spatial frequency for source images I_i, where $i \in \{A, B\}$, we obtain their respective spatial frequency maps SF_i. The initial decision maps IDM_A and IDM_B are then generated through pixel-wise comparison of SF_A and SF_B:

$$\mathrm{IDM}_{A(x,y)} = \begin{cases} 1, & \text{if } \mathrm{SF}_{A(x,y)} \geq \mathrm{SF}_{B(x,y)} \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

$$\mathrm{IDM}_{B(x,y)} = 1 - \mathrm{IDM}_{A(x,y)} \tag{5}$$

Additionally, to provide explicit boundary information and facilitate the refinement of decision maps, especially in edge regions, we employ the Sobel operators [47] for edge detection. For source images I_A and I_B, the edge maps Edge_A and Edge_B are computed using horizontal and vertical Sobel kernels:

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

$$\text{Edge}_i = \sqrt{(I_i * S_x)^2 + (I_i * S_y)^2}, \quad i \in \{A, B\} \tag{7}$$

where $*$ denotes the convolution operation. The edge maps highlight structural boundaries in the source images by computing the magnitude of gradients in both horizontal and vertical directions, providing crucial guidance for decision map refinement in regions where focus properties transition.

3.3. Network Architecture

We propose ULU-Net, an ultra-lightweight U-Net variant specifically designed for MFF, as shown in Figure 3. The network employs an encoder-decoder architecture that takes source images, initial decision maps, and edge maps as input to generate refined decision maps. Unlike conventional U-Net architectures that typically use large channel numbers (e.g., 64→128→256→512), we carefully design a minimal channel progression (8→16→24→32→48→64) that significantly reduces parameters while maintaining essential feature extraction capability. The encoder consists of six stages with this progressive channel expansion, while the decoder has five stages for feature reconstruction. To further enhance parameter efficiency, we employ element-wise addition instead of channel concatenation for skip connections, preventing channel number expansion and thus reducing the parameters in subsequent layers.

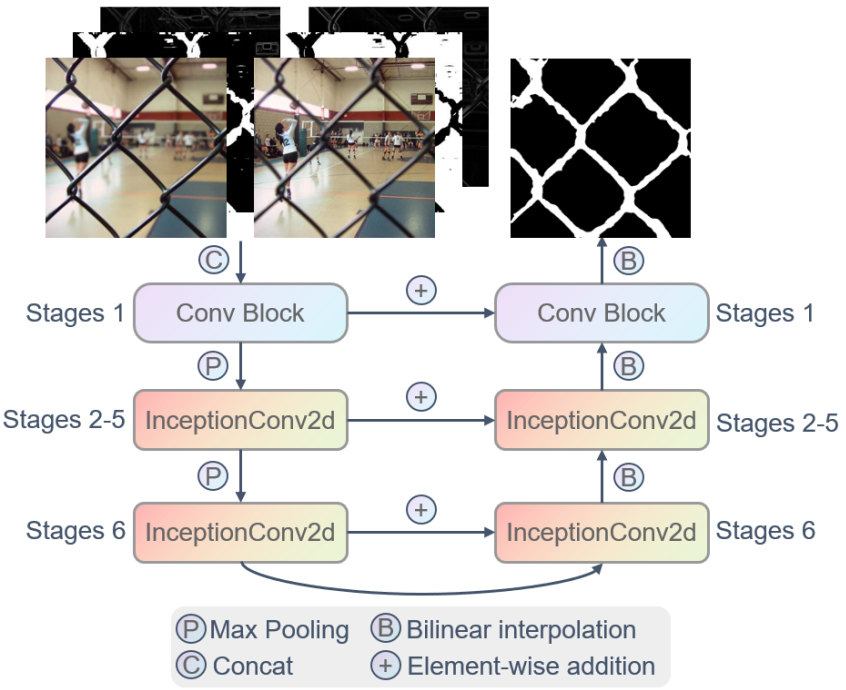

Figure 3. The designed U-shaped ultra-light network ULU-Net.

The input layer concatenates source images (I_A, I_B), initial decision maps (IDM_A, IDM_B), and edge maps (Edge_A, Edge_B) into a multi-channel tensor:

$$X_{\text{input}} = \text{Concat}(I_A, I_B, \text{IDM}_A, \text{IDM}_B, \text{Edge}_A, \text{Edge}_B) \tag{8}$$

To achieve parameter efficiency while maintaining feature extraction capability, we introduce InceptionDWConv2d [48], a lightweight module inspired by the Inception archi-

tecture. The module splits input channels with a branching ratio of 0.125 and processes them through parallel pathways:

$$Y_{\text{out}} = \text{Concat}(X_{\text{identity}}, DWConv_{3\times3}(X_{\text{branch1}}), DWConv_{1\times11}(X_{\text{branch2}}), DWConv_{11\times1}(X_{\text{branch3}})) \tag{9}$$

where $DWConv$ represents depthwise convolution with specified kernel sizes. This design effectively captures multi-scale and directional features while minimizing computational cost.

The entire network contains only 15K parameters, which is merely 1/200 of MobileNetV3's parameter count [49], making it remarkably lightweight compared to existing methods. This ultra-compact design is achieved without sacrificing performance through our careful architectural choices and the effective utilization of focus property and edge information as prior guidance.

3.4. Loss Function

To effectively train our lightweight network for decision map refinement, we propose a combined loss function that leverages two complementary components: Binary Focal Loss (BFL) [50] and Binary Dice Loss (BDL) [51]. The total loss $\mathcal{L}_{\text{total}}$ is defined as:

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{BFL}} + \lambda_2 \mathcal{L}_{\text{BDL}} \tag{10}$$

where $\lambda_1 = 15$ and $\lambda_2 = 1$ are the weighting coefficients that balance the contribution of each loss term. In our context, BFL helps the network focus on hard-to-classify regions. Formally, BFL is defined as:

$$\mathcal{L}_{\text{BFL}} = -\alpha y(1-p)^\gamma \log(p) - (1-\alpha)(1-y)p^\gamma \log(1-p) \tag{11}$$

where p represents the predicted probability, y is the ground truth label, α is the balancing factor (set to 0.25), and γ is the focusing parameter (set to 2.0). The focal loss introduces modulating factors $(1-p)^\gamma$ for positive samples and p^γ for negative samples. This automatically down-weights the contribution of easy examples and focuses on hard examples during training. This property is particularly beneficial for our task, as it helps the network concentrate on refining the decision boundaries where the focus state is ambiguous.

To complement the focal loss, we incorporate the Binary Dice Loss, which is particularly effective for handling imbalanced segmentation tasks. BDL measures the overlap between predicted and ground truth regions:

$$\mathcal{L}_{\text{BDL}} = 1 - \frac{2\sum_{i=1}^{N} p_i y_i + \epsilon}{\sum_{i=1}^{N} p_i^2 + \sum_{i=1}^{N} y_i^2 + \epsilon} \tag{12}$$

where p_i and y_i are the predicted and ground truth values at position i, N is the total number of pixels, and ϵ is a small constant to ensure numerical stability. The Binary Dice Loss is particularly advantageous for our decision map refinement task, as it naturally optimizes for spatial overlap between predicted and ground truth regions.

3.5. Fusion Scheme

Figure 4 illustrates the complete pipeline from the input image pair to the final fused result. Given a pair of source images I_A and I_B, along with their coarse initial decision maps and edge maps as inputs, the proposed ULU-Net first generates a continuous-valued decision map, Output. To obtain more robust fusion results, we process this raw output through several stages:

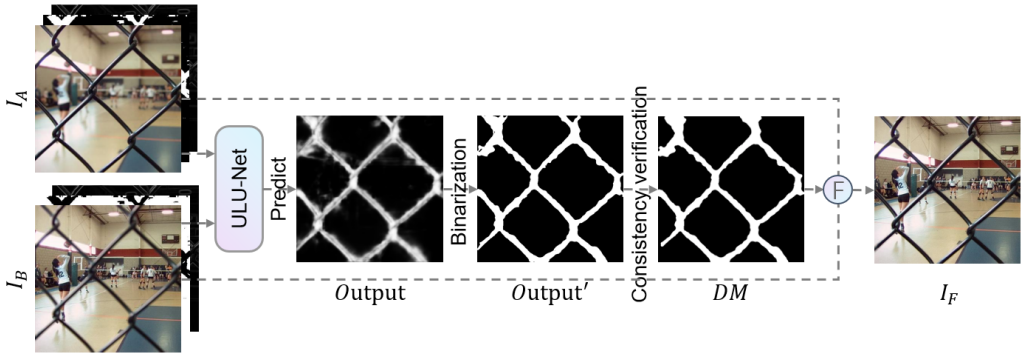

Figure 4. The adopted fusion scheme.

First, we apply a binarization operation to convert the continuous-valued Output to a binary decision map, Output':

$$\text{Output}'(x,y) = \begin{cases} 1, & \text{if Output}(x,y) \geq T \\ 0, & \text{otherwise} \end{cases} \tag{13}$$

where T is the binarization threshold, empirically set to 0.5.

Next, to ensure spatial consistency and eliminate isolated noise, we perform morphological operations on Output' to obtain the final decision map, DM:

$$\text{DM} = \delta(\epsilon(\text{Output}')) \tag{14}$$

where $\epsilon(\cdot)$ and $\delta(\cdot)$ denote the morphological erosion and dilation operations, respectively, with a 5×5 structuring element.

Finally, the fused image I_F is generated by combining the source images according to the refined decision map:

$$I_F(x,y) = \text{DM}(x,y) \cdot I_A(x,y) + (1 - \text{DM}(x,y)) \cdot I_B(x,y) \tag{15}$$

4. Experiments

4.1. Experimental Setup

Training Strategy. We use the DUTS dataset (https://saliencydetection.net/duts/) accessed on 2 June 2025 [52], which comprises 15,572 carefully curated images. This dataset is divided into 10,553 training samples and 5019 validation samples, each resized to a resolution of 256 × 256. To generate the training pairs, we first convert the ground truth annotations into binary masks. These masks then guide the application of variable Gaussian blur to specific regions within the source images, using kernel sizes that range from 3 to 21. This process yields realistic multi-focus image pairs. The detailed procedure for generating these training data can be found in SwinMFF [26]. For network optimization, we utilized the AdamW optimizer, initialized with a learning rate of 1×10^{-3}. A CosineAnnealingLR scheduler was implemented to progressively reduce the learning rate over the course of training. The model was trained for 200 epochs with a batch size of 32. All training was performed on a high-performance computing platform equipped with dual Nvidia A6000 GPUs, supported by an Intel(R) Xeon(R) Platinum 8375C CPU operating at 2.90 GHz. The entire framework was developed in PyTorch (version 2.4.1).

Datasets for Evaluation. To rigorously assess LightMFF's performance, we conducted extensive experiments using three widely recognized MFF benchmarks: Lytro [46], MFFW [53], and MFI-WHU [36]. The Lytro dataset consists of 20 image pairs captured with a Lytro light-field camera and was utilized for both qualitative and quantitative evaluation.

The MFFW dataset includes 13 image pairs notable for their pronounced defocus spread effects, allowing for qualitative evaluation under challenging focus conditions. Extending our evaluation scope, the MFI-WHU dataset provides 120 synthetically generated image pairs, created using Gaussian blur, which encompass diverse scenarios, including high-contrast scenes, and was also used for qualitative assessment.

Methods for Comparison. To thoroughly evaluate the efficacy of the proposed LightMFF method, we performed extensive comparisons against state-of-the-art image fusion approaches across four primary categories. For transform domain methods, our comparisons included DWT [7], DTCWT [54], NSCT [8], CVT [55], GFF [56], SR [57], ASR [58], MWGF [59], ICA [60], and NSCT-SR [61]. In the spatial domain category, we evaluated SSSDI [62], QUADTREE [63], DSIFT [61], SRCF [46], GFDF [64], BRW [65], MISF [66], and MDLSR_RFM [20]. Regarding end-to-end deep learning approaches, we included IFCNN-MAX [24], U2Fusion [25], SDNet [35], MFF-GAN [36], SwinFusion [27], MUFusion [37], FusionDiff [29], SwinMFF [26], and DDBFusion [28]. Additionally, we compared with decision map-based deep learning methods such as CNN [22], ECNN [23], DRPL [19], SESF [30], MFIF-GAN [31], MSFIN [10], GACN [67], and ZMFF [11]. For traditional methods, we either employed the default parameters as specified in their original implementations or directly used the fusion results made publicly available by their respective authors. For deep learning-based methods, we utilized the pre-trained models provided by their authors to ensure a fair comparison. This comprehensive comparison of 37 methods across four distinct categories allows for a thorough assessment of the proposed method's capabilities. It is important to note that while some related work discusses lightweight networks, we could not include them in our comparisons due to the unavailability of their codebases. Nevertheless, many highly compact networks, such as MFF-GAN [36] and SESF [30], are already featured in our evaluation.

Evaluation metrics. To comprehensively evaluate the performance of our method, we employ six complementary metrics that assess different aspects of the fusion results: Edge-based Similarity Metric $Q^{AB/F}$ [68], Normalized Mutual Information Metric Q_{MI} [69], Phase Congruency-based Metric Q^{P} [70], Structural Similarity Metrics Q_W and Q_E [71], and Human Perception-inspired Metric Q_{CB} [72]. For all metrics, higher values (↑) indicate better fusion performance. These metrics collectively provide a comprehensive evaluation framework, covering edge preservation, information content, structural similarity, and perceptual quality of the fusion results.

4.2. Experimental Results

Qualitative comparison. The decision maps offer more direct insights into a network's performance than the fused images themselves. For qualitative comparisons of decision maps, we exclusively present the initial decision maps generated by each network, which helps us avoid variations that might come from different post-processing techniques. We have visualized these decision maps, which are responsible for generating the fusion results, in Figure 5. Visual inspection suggests that LightMFF produces the most visually compelling decision maps among all the approaches we compared.

Compared to the Lytro dataset [46], the MFFW dataset [53] offers more challenging scenarios due to its more pronounced defocus dispersion effects. To further validate our approach under these demanding conditions, we conducted additional visual comparisons on this dataset, as shown in Figure 6. The comparison reveals that most methods, with the exceptions of our proposed LightMFF and MFIF-GAN [31], show noticeably degraded performance on this dataset when compared to their results on the Lytro dataset. Several approaches struggle with large-area misclassification, which severely impacts the quality of

their fusion results. In contrast, LightMFF maintains exceptional performance, consistently producing the most accurate decision maps among all compared methods.

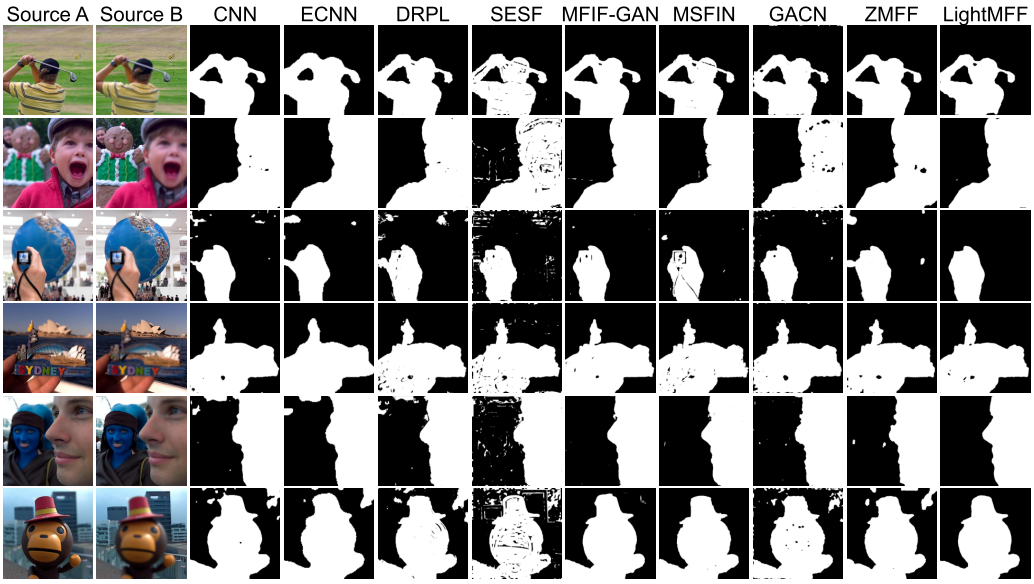

Figure 5. Comparison of decision maps generated by different methods on the Lytro dataset [46].

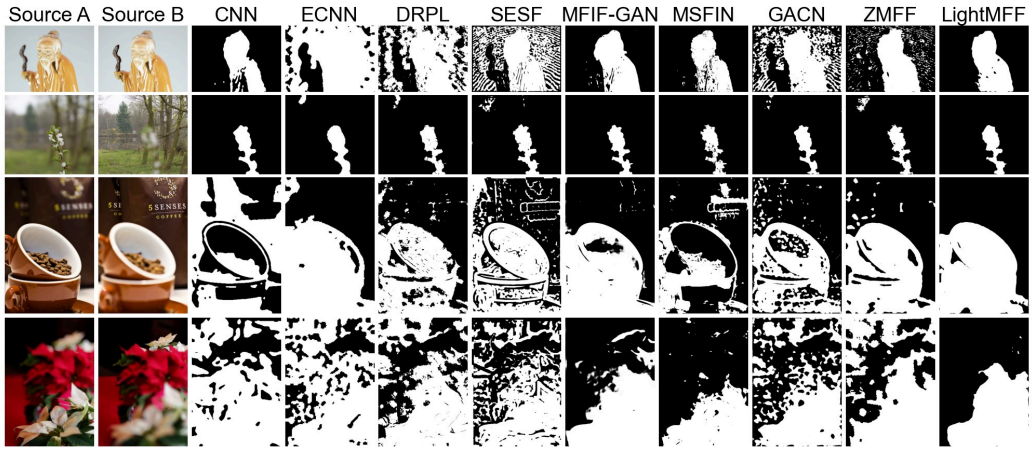

Figure 6. Comparison of decision maps generated by different methods on the MFFW dataset [53].

In Figure 7, we compared the fusion results of different methods using the "coffee cup" example from the MFFW dataset. As the figure illustrates, particularly in the boundary regions of the coffee cup, only ECNN [23] and our proposed LightMFF accurately differentiate between focused and defocused areas along the cup edges. Both ECNN and LightMFF achieve clean fusion results without notable artifacts. In contrast, other approaches exhibit visible fusion artifacts in these challenging regions. This robust performance demonstrates LightMFF's superior capability in handling scenes with strong defocus effects.

Figure 7. Fusion results of various SOTA methods on the "Coffee cup" example from the MFFW dataset [53].

Additionally, we conducted qualitative comparisons on representative image pairs from the MFI-WHU dataset [36]. The corresponding decision maps are shown in Figure 8. The selected cases represent three challenging scenarios: small object detection (first row), high-contrast scenes (second row), and large homogeneous regions (third and fourth rows). The proposed LightMFF consistently outperforms all competing methods across all test cases, demonstrating superior accuracy in every scenario.

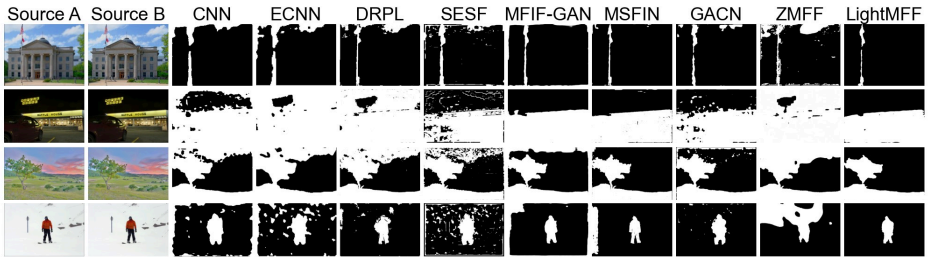

Figure 8. Comparison of decision maps generated by different methods on the MFI-WHU dataset [36].

Quantitative comparison. We conduct comprehensive quantitative evaluations on the widely-used Lytro dataset [46] to assess our method's performance. Table 1 thoroughly compares our approach against 36 other methods, highlighting the best-performing methods in bold and the second-best in underline. For each metric, an upward arrow indicates that a higher value is better. Methods highlighted with a background color represent our proposed approach or the complete proposed model.As shown in Table 1, the proposed LightMFF achieves state-of-the-art performance across most metrics. Specifically, LightMFF obtains the highest scores in five out of six metrics: $Q^{AB/F}$ (0.7588), Q_{MI} (1.1462), Q^{P} (0.8450), Q_E (0.9061), and Q_{CB} (0.8067). For the remaining metric Q_W, our method achieves the second-best score (0.9400), only marginally lower than GACN (0.9405) [67] by 0.0005.

Table 1. Quantitative comparison of different MFF methods on the Lytro dataset [46].

Method	$Q^{AB/F}\uparrow$	$Q_{MI}\uparrow$	$Q^{P}\uparrow$	$Q_W\uparrow$	$Q_E\uparrow$	$Q_{CB}\uparrow$
Methods based on image transform domain						
DWT [7]	0.6850	0.8677	0.2878	0.8977	0.8356	0.6117
DTCWT [54]	0.6929	0.8992	0.2925	0.8987	0.8408	0.6234
NSCT [8]	0.6901	0.9039	0.2928	0.9030	0.8413	0.6174
CVT [55]	0.7243	0.8968	0.7966	0.9388	0.9023	0.7277
DCT [7]	0.7031	0.9383	0.7825	0.9093	0.8073	0.6624
GFF [56]	0.6998	1.0020	0.2952	0.8982	0.8351	0.6518
SR [57]	0.6944	1.0003	0.2921	0.8984	0.8309	0.6406
ASR [58]	0.6951	1.0024	0.2926	0.8986	0.8308	0.6413
MWGF [59]	0.7037	1.0545	0.3176	0.8913	0.8107	0.6758
ICA [60]	0.6766	0.8687	0.2964	0.9084	0.8219	0.5956
NSCT-SR [61]	0.6995	1.0189	0.2949	0.9000	0.8385	0.6501
Methods based on image spatial domain						
SSSDI [62]	0.6966	1.0351	0.2915	0.8961	0.8279	0.6558
QUADTREE [63]	0.7027	1.0630	0.2940	0.8962	0.8265	0.6681
DSIFT [61]	0.7046	1.0642	0.2954	0.8977	0.8354	0.6675
SRCF [46]	0.7036	1.0590	0.2954	0.8978	0.8369	0.6669
GFDF [64]	0.7049	1.0524	0.2974	0.8989	0.8399	0.6657
BRW [65]	0.7040	1.0516	0.2964	0.8984	0.8371	0.6650
MISF [66]	0.6984	1.0391	0.2945	0.8929	0.8063	0.6607
MDLSR_RFM [20]	0.7518	1.1233	0.8294	0.9394	0.9021	<u>0.8064</u>

Table 1. *Cont.*

Method	$Q^{AB/F}$ ↑	Q_{MI} ↑	Q^{P} ↑	Q_{W} ↑	Q_{E} ↑	Q_{CB} ↑
End-to-end methods based on deep learning						
IFCNN-MAX [24]	0.6784	0.8863	0.2962	0.9013	0.8324	0.5986
U2Fusion [25]	0.6190	0.7803	0.2994	0.8909	0.7108	0.5159
SDNet [35]	0.6441	0.8464	0.3072	0.8934	0.7464	0.5739
MFF-GAN [36]	0.6222	0.7930	0.2840	0.8887	0.7660	0.5399
SwinFusion [27]	0.6597	0.8404	0.3117	0.9011	0.7460	0.5745
MUFusion [37]	0.6614	0.8030	0.7160	0.9089	0.8036	0.6758
FusionDiff [29]	0.6744	0.8692	0.2900	0.8980	0.8261	0.5747
SwinMFF [26]	0.7321	0.9605	0.8222	0.9390	0.8986	0.7543
DDBFusion [28]	0.5026	0.8152	0.5610	0.8391	0.4947	0.6057
Decision map-based methods using deep learning						
CNN [22]	0.7019	1.0424	0.2968	0.8976	0.8311	0.6628
ECNN [23]	0.7030	1.0723	0.2945	0.8946	0.8169	0.6698
DRPL [19]	0.7574	1.1405	0.8435	0.9397	0.9060	0.8035
SESF [30]	0.7031	1.0524	0.2950	0.8977	0.8353	0.6657
MFIF-GAN [31]	0.7029	1.0618	0.2960	0.8982	0.8395	0.6660
MSFIN [10]	0.7045	1.0601	0.2973	0.8990	0.8436	0.6664
GACN [67]	0.7581	1.1334	0.8443	**0.9405**	0.9013	0.8024
ZMFF [11]	0.6635	0.8694	0.2890	0.8951	0.8253	0.6136
LightMFF	**0.7588**	**1.1462**	**0.8450**	0.9400	**0.9061**	**0.8067**

Efficiency Analysis. We conduct a comprehensive efficiency analysis comparing LightMFF against state-of-the-art methods across model size, computational complexity (FLOPs), and inference time, as detailed in Table 2. The inference time is measured on the MFI-WHU dataset [36], representing the average time required to process an image pair.

Table 2. Comparison of computational efficiency across different learning-based MFF methods.

Method	Model Size (M)	FLOPs (G)	Time (s)
End-to-end methods based on deep learning			
IFCNN-MAX [24]	0.08	8.54	0.09
U2Fusion [25]	0.66	86.40	0.16
SDNet [35]	0.07	8.81	0.10
MFF-GAN [36]	0.05	3.08	0.06
SwinFusion [27]	0.93	63.73	1.79
MUFusion [37]	2.16	24.07	0.72
FusionDiff [29]	26.90	58.13	81.47
SwinMFF [26]	41.25	22.38	0.46
DDBFusion [28]	10.92	184.93	1.69
Decision map-based methods using deep learning			
CNN [22]	8.76	142.23	0.06
ECNN [23]	1.59	14.93	125.53
DRPL [19]	1.07	140.49	0.22
SESF [30]	0.07	4.90	0.26
MFIF-GAN [31]	3.82	693.03	0.32
MSFIN [10]	4.59	26.76	1.10
GACN [67]	0.07	10.89	0.16
ZMFF [11]	6.33	464.53	165.38
LightMFF	**0.02**	**0.06**	**0.02**
Reduction (%)	60.00%	98.05%	66.67%

The proposed LightMFF achieves remarkable efficiency improvements across all metrics. With only 0.02 M parameters, LightMFF is approximately 2.5× smaller than its closest end-to-end competitor (MFF-GAN [36] with 0.05M parameters) and 3.5× smaller than the most compact decision map-based method (SESF [30]/GACN [67] with 0.07 M parameters). In terms of computational complexity, LightMFF requires merely 0.06 G FLOPs, representing a 98.05% reduction in computational cost compared to the most compact prior method (MFF-GAN [36] with 3.08 G FLOPs). The practical benefits of our efficient design are evident in the inference time measurements, where LightMFF executes in 0.02 s on GPU, achieving a 3× speedup over MFF-GAN (0.06 s) [36] and an 8× improvement over GACN (0.16 s) [67].

4.3. Performance Under Extreme Conditions

To validate the robustness of the proposed method under extreme conditions where input image pairs consist of completely sharp and completely blurred images, we conducted a challenging experiment using images from the MFI-WHU dataset [36], as illustrated in Figure 9. For each image, we synthesized a corresponding completely blurred version using Gaussian blur kernels.

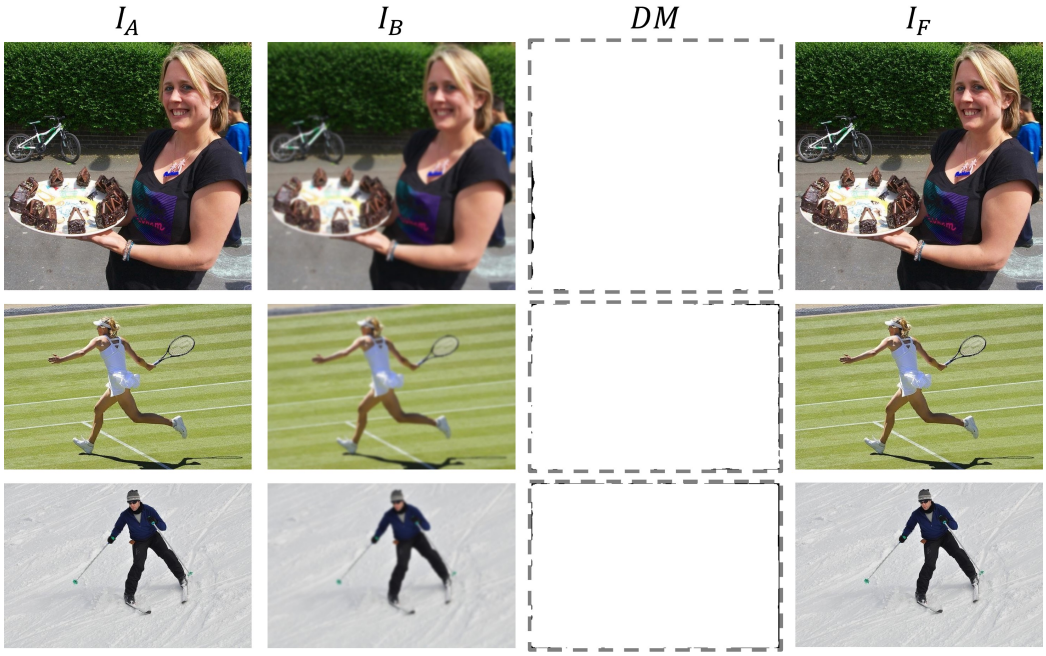

Figure 9. Performance evaluation under extreme conditions. The proposed method demonstrates robust performance when processing image pairs consisting of completely sharp and completely blurred images. From left to right: original sharp image, synthetically blurred counterpart, generated decision map, and final fusion result. The model successfully distinguishes between focused and defocused regions even in these challenging scenarios.

The paired images, comprising the original sharp image and its completely blurred counterpart, were then fed into the proposed model to generate the decision maps and fusion results shown in the figure. Remarkably, the proposed model maintains excellent performance even under these extreme conditions, successfully distinguishing between defocused and focused regions while producing accurate decision maps for effective image fusion. This demonstrates the model's robustness and reliability across diverse imaging conditions.

4.4. Performance in Real-World Scenarios

Motivated by findings in SAMF [3] that MFF can improve object detection accuracy in autonomous driving systems, we evaluated our LightMFF network on the Road-MF dataset [3]. This dataset contains 80 pairs of real-world multi-focus images captured in diverse road scenes, making it an ideal choice for assessing our method's practical applicability.

As shown in Figure 10, our proposed method performs better than DRPL [19] in real-world road scenes. LightMFF notably shows fewer misclassifications in both sky regions and with objects closer to the ground. Even with the inherent challenges of real-world scenes, LightMFF achieves exceptional performance while remaining computationally efficient. This dual advantage highlights the practical value of our lightweight architecture, especially for real-world applications like autonomous driving systems where both accuracy and real-time processing are crucial.

Figure 10. Qualitative comparison of the proposed LightMFF and DRPL on the Road-MF dataset [3]. FR and DM denote the final fusion results and decision maps (after post-processing), respectively.

4.5. Performance Under Image Degradation

To investigate the performance of the proposed model when input images suffer from various degradation conditions, we present a qualitative comparison in Figure 11 using the same input image pair under different scenarios, including variations in contrast and brightness, as well as corruption by Gaussian noise and salt-and-pepper noise.

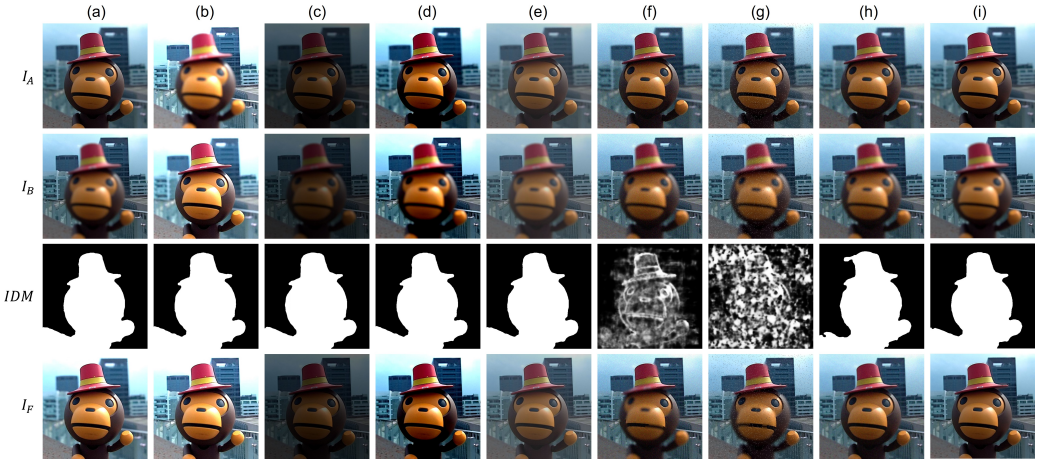

Figure 11. Performance comparison of the model under different image degradation conditions. (**a**) Original image; (**b**) High brightness; (**c**) Low brightness; (**d**) High contrast; (**e**) Low contrast; (**f**) Gaussian noise; (**g**) Salt-and-pepper noise; (**h**) Gaussian filtering for Gaussian noise removal; (**i**) Median filtering for salt-and-pepper noise removal.

The experimental results demonstrate that the proposed model exhibits remarkable robustness to brightness and contrast variations, with virtually no impact on fusion performance. However, the proposed method lacks robustness against high-frequency noise. When input images are corrupted by Gaussian noise or salt-and-pepper noise, the high-frequency noise components significantly degrade the quality of the initial decision map, thereby adversely affecting the fusion results. Nevertheless, as illustrated in Figure 11h,i, this issue can be effectively mitigated through simple filtering operations.

4.6. Ablation Study

To validate the effectiveness of our proposed components and design choices, we conduct comprehensive ablation studies on the Lytro dataset [46]. Specifically, we investigate how the initial decision map prior and edge information contribute to the network's performance. We design four experimental settings:

1. Group 1 (Baseline): Only source images are fed into the network, representing the conventional approach where the network directly learns to distinguish focused from defocused regions.
2. Group 2 (w/ Decision Map): Source images and the initial decision map are used as input. This setting evaluates the effectiveness of our proposed paradigm shift from direct classification to decision map refinement.
3. Group 3 (w/ Edge Map): Source images are complemented with edge maps. This setting examines the impact of explicit boundary information without decision map guidance.
4. Group 4 (Full Model): Our complete model, incorporating both the initial decision map and edge information, demonstrates the synergistic effect of all components.

Table 3 and Figure 12 showcase the results of our ablation study. The baseline model (Group 1) performed less effectively, highlighting the inherent difficulties of directly classifying focus/defocus regions with an ultra-lightweight architecture. Adding the initial decision map (Group 2) significantly boosted performance. This confirms our hypothesis that reframing the learning objective to decision map refinement genuinely reduces task complexity. The edge information (Group 3) also brought noticeable improvements over the baseline, affirming that explicit boundary cues help the network better identify focus transitions. These ablation studies confirm that our task transformation approach and targeted guidance signals successfully reduce computational complexity while maintaining superior multi-focus fusion (MFF) performance.

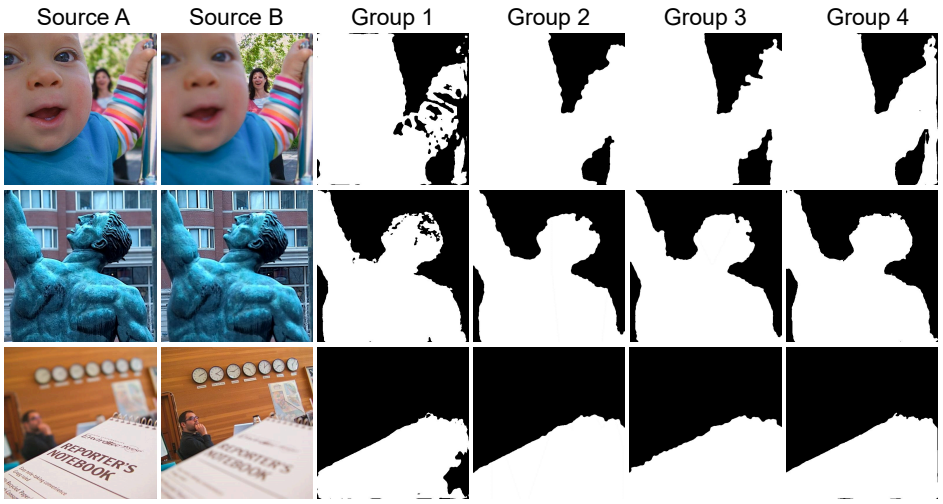

Figure 12. Visualization of ablation study results on the Lytro dataset [46].

Table 3. Ablation quantitative results on the Lytro dataset [46]. Group 1: baseline with only source images; Group 2: with initial decision map; Group 3: with edge maps; Group 4: full model with both components.

Method	$Q^{AB/F} \uparrow$	$Q_{MI} \uparrow$	$Q^{P} \uparrow$	$Q_{W} \uparrow$	$Q_{E} \uparrow$	$Q_{CB} \uparrow$
Group 1	0.7578	1.1448	0.8437	0.9363	0.9021	0.8036
Group 2	0.7587	1.1460	0.8447	0.9389	0.9055	0.8060
Group 3	0.7584	1.1459	0.8443	0.9373	0.9053	0.8057
Group 4	**0.7588**	**1.1462**	**0.8450**	**0.9400**	**0.9061**	**0.8067**

5. Discussion

Despite the demonstrated effectiveness of our proposed method in MFF, we have identified three primary limitations:

First, our decision map-based paradigm, while computationally efficient, might not perform as robustly as end-to-end approaches in handling complex scenes [26].

Second, our current post-processing strategy relies on basic morphological operations to refine the decision map. More advanced techniques, such as image matting-based methods [73], can indeed effectively reduce artifacts along fusion boundaries. However, they introduce substantial computational overhead—typically adding 2-3 s per image pair to processing time. This additional cost conflicts with our goal of maintaining computational efficiency. Therefore, we chose lightweight morphological operations as a compromise. While our current post-processing approach generally improves fusion quality, it may occasionally yield suboptimal results in complex local regions with fine details, as illustrated in Figure 13.

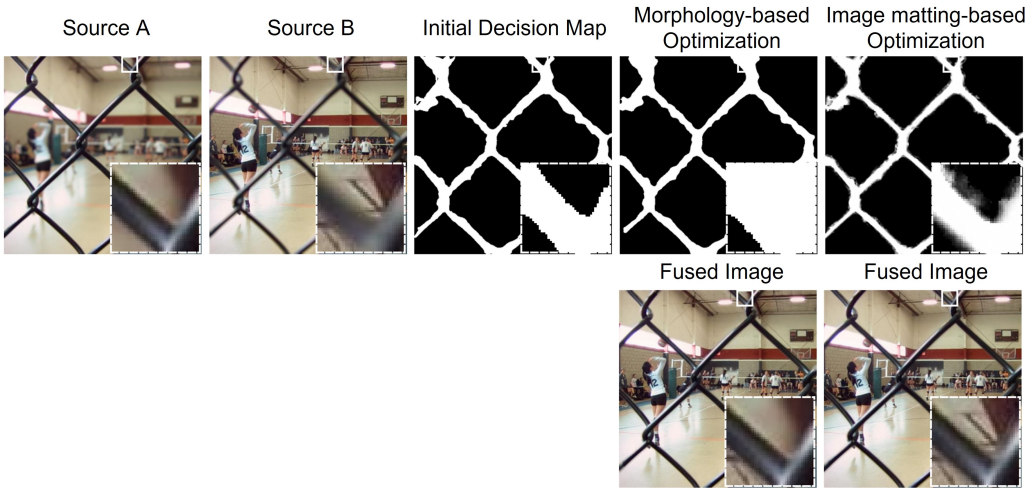

Figure 13. Limitations of current decision map post-processing methods.

Third, our method is susceptible to high-frequency noise. Because the initial decision map is generated based on a focus measure that is sensitive to local variations, the presence of high-frequency noise in the source images can lead to an erroneous initial decision map, thereby further affecting the final decision map. This sensitivity highlights a need for more robust focus measure mechanisms that are less influenced by noise, or for a pre-processing step to denoise the source images.

Our work opens several promising directions for future research. We could investigate more sophisticated yet efficient post-processing techniques to further enhance fusion quality in fine-detail regions. Additionally, developing a more noise-robust focus measure or integrating a denoising module could significantly improve the reliability of the decision map generation. Extending our lightweight architecture design principles to other image fusion tasks (e.g., multi-exposure, infrared-visible) might lead to a more general framework for efficient

image fusion. Finally, exploring the integration of our method with specific downstream tasks like object detection could significantly boost its practical utility in real-world applications.

6. Conclusions

In this paper, we have presented LightMFF, an ultra-lightweight network for MFF. The key contributions are as follows: (1) proposing a novel task transformation paradigm that reformulates multi-focus fusion from focus classification to decision map refinement, which significantly reduces model complexity while maintaining performance; (2) introducing an ultra-lightweight architecture that achieves state-of-the-art performance with only 0.02 M parameters and 0.06 G FLOPs, which represents a 98.05% reduction in computational cost compared to prior methods; and (3) demonstrating through extensive experiments that our method achieves competitive or superior performance on standard benchmarks while maintaining real-time inference capability (0.02 s per image pair).

The proposed LightMFF demonstrates that it is possible to achieve state-of-the-art performance in MFF with significantly reduced computational resources, challenging the conventional wisdom that deeper and more complex networks are necessary for superior performance. Our approach's efficiency makes it particularly suitable for resource-constrained scenarios such as mobile devices, embedded systems, and real-time applications. We believe that this work represents a significant step toward making advanced MFF techniques more accessible and practical for real-world deployment.

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

Funding: This work was supported by the National Key R&D Program of China (No. 2022-34), the 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) and the Research Startup Funding from Hainan Institute of Zhejiang University (NO. 0208-6602-A12204). 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.

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/LightMFF (accessed on 2 June 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. 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]

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. Burt, P.J.; Adelson, E.H. The Laplacian pyramid as a compact image code. In *Readings in Computer Vision*; Elsevier: Amsterdam, The Netherlands, 1987; pp. 671–679.
5. 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.
6. 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]
7. Li, H.; Manjunath, B.; Mitra, S.K. Multisensor image fusion using the wavelet transform. *Graph. Model. Image Process.* **1995**, *57*, 235–245. [CrossRef]
8. 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.
9. Zhang, Q.; Guo, B.L. Multifocus image fusion using the nonsubsampled contourlet transform. *Signal Process.* **2009**, *89*, 1334–1346. [CrossRef]
10. 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]
11. Hu, X.; Jiang, J.; Liu, X.; Ma, J. ZMFF: Zero-shot multi-focus image fusion. *Inf. Fusion* **2023**, *92*, 127–138. [CrossRef]
12. Sujatha, K.; Shalini Punithavathani, D. Optimized ensemble decision-based multi-focus imagefusion using binary genetic Grey-Wolf optimizer in camera sensor networks. *Multimed. Tools Appl.* **2018**, *77*, 1735–1759. [CrossRef]
13. Kausar, N.; Majid, A.; Javed, S.G. A novel ensemble approach using individual features for multi-focus image fusion. *Comput. Electr. Eng.* **2016**, *54*, 393–405. [CrossRef]
14. Huang, Y.; Li, W.; Gao, M.; Liu, Z. Algebraic multi-grid based multi-focus image fusion using watershed algorithm. *IEEE Access* **2018**, *6*, 47082–47091. [CrossRef]
15. Duan, J.; Chen, L.; Chen, C.P. Multifocus image fusion with enhanced linear spectral clustering and fast depth map estimation. *Neurocomputing* **2018**, *318*, 43–54. [CrossRef]
16. Jagtap, N.; Thepade, S. High-quality image multi-focus fusion to address ringing and blurring artifacts without loss of information. *Vis. Comput.* **2022**, *38*, 4353–4371. [CrossRef]
17. Kong, W.; Lei, Y. Multi-focus image fusion using biochemical ion exchange model. *Appl. Soft Comput.* **2017**, *51*, 314–327. [CrossRef]
18. Duan, Z.; Luo, X.; Zhang, T. Combining transformers with CNN for multi-focus image fusion. *Expert Syst. Appl.* **2024**, *235*, 121156. [CrossRef]
19. 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]
20. 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]
21. Li, B.; Zhang, L.; Liu, J.; Peng, H.; Wang, Q.; Liu, J. Multi-focus image fusion with parameter adaptive dual channel dynamic threshold neural P systems. *Neural Netw.* **2024**, *179*, 106603. [CrossRef]
22. 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]
23. Amin-Naji, M.; Aghagolzadeh, A.; Ezoji, M. Ensemble of CNN for multi-focus image fusion. *Inf. Fusion* **2019**, *51*, 201–214. [CrossRef]
24. 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]
25. 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]
26. 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*, 1–24. [CrossRef]
27. 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]
28. 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]
29. 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]

30. 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]
31. 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]
32. 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]
33. 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]
34. 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]
35. Zhang, H.; Ma, J. SDNet: A versatile squeeze-and-decomposition network for real-time image fusion. *Int. J. Comput. Vis.* **2021**, *129*, 2761–2785. [CrossRef]
36. 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]
37. Cheng, C.; Xu, T.; Wu, X.J. MUFusion: A general unsupervised image fusion network based on memory unit. *Inf. Fusion* **2023**, *92*, 80–92. [CrossRef]
38. 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]
39. 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]
40. 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]
41. Jin, X.; Hou, J.; Nie, R.; Yao, S.; Zhou, D.; Jiang, Q.; He, K. A lightweight scheme for multi-focus image fusion. *Multimed. Tools Appl.* **2018**, *77*, 23501–23527. [CrossRef]
42. Zhou, Y.; Liu, K.; Dou, Q.; Liu, Z.; Jeon, G.; Yang, X. LNMF: Lightweight network for multi-focus image fusion. *Multimed. Tools Appl.* **2022**, *81*, 22335–22353. [CrossRef]
43. Nie, X.; Hu, B.; Gao, X. MLNet: A multi-domain lightweight network for multi-focus image fusion. *IEEE Trans. Multimed.* **2022**, *25*, 5565–5579. [CrossRef]
44. Wu, Z.; Chen, J.; Tan, L.; Gong, H.; Zhou, Y.; Shi, G. A lightweight GAN-based image fusion algorithm for visible and infrared images. In Proceedings of the 2024 4th International Conference on Computer Science and Blockchain (CCSB), Shenzhen, China, 6–8 September 2024; pp. 466–470.
45. Xie, X.; Zhang, X.; Tang, X.; Zhao, J.; Xiong, D.; Ouyang, L.; Yang, B.; Zhou, H.; Ling, B.W.K.; Teo, K.L. MACTFusion: Lightweight cross transformer for adaptive multimodal medical image fusion. *IEEE J. Biomed. Health Inform.* **2024**, *29*, 3317–3328. [CrossRef]
46. Nejati, M.; Samavi, S.; Shirani, S. Multi-focus image fusion using dictionary-based sparse representation. *Inf. Fusion* **2015**, *25*, 72–84. [CrossRef]
47. Sobel, I.; Feldman, G. A 3x3 isotropic gradient operator for image processing. *Stanf. Artif. Intell. Proj. (Sail)* **1968**, *1968*, 271–272.
48. Yu, W.; Zhou, P.; Yan, S.; Wang, X. Inceptionnext: When inception meets convnext. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 5672–5683.
49. Howard, A.; Sandler, M.; Chu, G.; Chen, L.C.; Chen, B.; Tan, M.; Wang, W.; Zhu, Y.; Pang, R.; Vasudevan, V.; et al. Searching for mobilenetv3. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1314–1324.
50. Ross, T.Y.; Dollár, G. Focal loss for dense object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2980–2988.
51. Sudre, C.H.; Li, W.; Vercauteren, T.; Ourselin, S.; Jorge Cardoso, M. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Proceedings of the Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, 14 September 2017*; Proceedings 3; Springer: Berlin/Heidelberg, Germany, 2017; pp. 240–248.
52. 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.
53. Xu, S.; Wei, X.; Zhang, C.; Liu, J.; Zhang, J. MFFW: A new dataset for multi-focus image fusion. *arXiv* **2020**, arXiv:2002.04780.
54. 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.

55. 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]
56. Li, S.; Kang, X.; Hu, J. Image fusion with guided filtering. *IEEE Trans. Image Process.* **2013**, *22*, 2864–2875.
57. 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] [PubMed]
58. Liu, Y.; Wang, Z. Simultaneous image fusion and denoising with adaptive sparse representation. *IET Image Process.* **2015**, *9*, 347–357. [CrossRef]
59. Zhou, Z.; Li, S.; Wang, B. Multi-scale weighted gradient-based fusion for multi-focus images. *Inf. Fusion* **2014**, *20*, 60–72. [CrossRef]
60. 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]
61. 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]
62. 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]
63. 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]
64. 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]
65. 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]
66. Zhan, K.; Kong, L.; Liu, B.; He, Y. Multimodal image seamless fusion. *J. Electron. Imaging* **2019**, *28*, 023027. [CrossRef]
67. 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]
68. Xydeas, C.S.; Petrovic, V. Objective image fusion performance measure. *Electron. Lett.* **2000**, *36*, 308–309. [CrossRef]
69. Qu, G.; Zhang, D.; Yan, P. Information measure for performance of image fusion. *Electron. Lett.* **2002**, *38*, 1. [CrossRef]
70. Zhao, J.; Laganiere, R.; Liu, Z. Performance assessment of combinative pixel-level image fusion based on an absolute feature measurement. *Int. J. Innov. Comput. Inf. Control* **2007**, *3*, 1433–1447.
71. Piella, G.; Heijmans, H. A new quality metric for image fusion. In Proceedings of the 2003 International Conference on Image Processing (Cat. No. 03CH37429), Barcelona, Spain, 14–17 September 2003; Volume 3, pp. 3–173.
72. Chen, Y.; Blum, R.S. A new automated quality assessment algorithm for image fusion. *Image Vis. Comput.* **2009**, *27*, 1421–1432. [CrossRef]
73. Levin, A.; Lischinski, D.; Weiss, Y. A closed-form solution to natural image matting. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *30*, 228–242. [CrossRef]