Multi-Focus Image Fusion via Explicit Defocus Blur Modelling

Yuhui Quan[1], Xi Wan[1], Zitao Tang[1], Jinxiu Liang[2#], Hui Ji[3]

[1]South China University of Technology [2]Peking University [3]National University of Singapore

yhquan@scut.edu.cn, csxwan@mail.scut.edu.cn, zitaotang007@gmail.com, cssherryliang@pku.edu.cn, matjh@nus.edu.sg

Abstract

Multi-focus image fusion (MFIF) enhances depth of field in photography by generating an all-in-focus image from multiple images captured at different focal lengths. While deep learning has shown promise in MFIF, most existing methods overlooked the physical properties of defocus blurring in their network design, limiting their interoperability and generalization. This paper introduces a novel framework that integrates explicit defocus blur modeling into the MFIF process, improving both interpretability and performance. Using an atom-based spatially-varying parameterized defocus blurring model, our approach calculates pixel-wise defocus descriptors and initial focused images from multi-focus source images in a scale-recurrent manner to estimate soft decision maps. Fusion is then performed using masks derived from these decision maps, with special treatment for pixels likely defocused in all source images or near boundaries of defocused/focused regions. The model is trained with a fusion loss and a cross-scale defocus estimation loss. Extensive experiments on benchmark datasets demonstrated the effectiveness of our approach. Our code is avalible on https://github.com/Tangzitao/DMANet.

Introduction

Defocus blur is a prevalent challenge in photography, often resulting from a shallow depth of field (DoF) when capturing scenes with varying depths. This type of blur occurs when objects in an image fall outside the camera's focal plane, yielding a loss of sharpness and detail in the image.

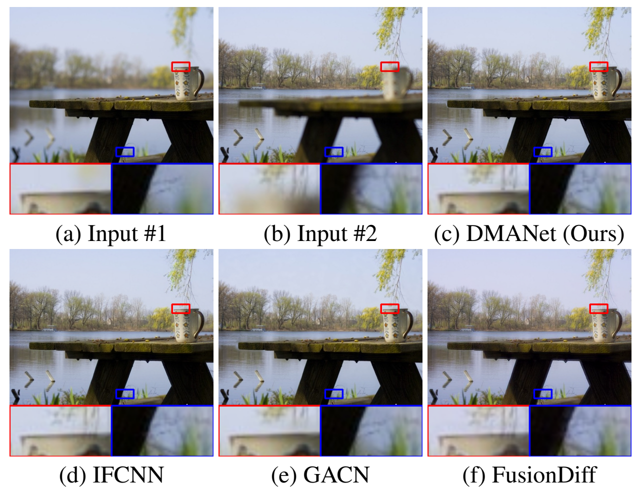
Formally, we can express the object field Z as a sum over weighted impulses:

$$Z(x,y) = \int\int Z(u,v)\delta(u-x,y-y)dudv, \quad (1)$$

where δ denotes the impulse function, corresponding to a Dirac delta PSF. Then, the image plane field can be calculated as a superposition over weighted point spread functions (PSFs) in the image plane using the same weighting function as in the object plane. As a result, the defocused image Y with blurring effects can be expressed as

$$Y(x,y) = \int\int Z(u,v)D_{u,v}(x-u,y-v)dudv, \quad (2)$$

[#]Corresponding author.



(a) Input #1 (b) Input #2 (c) DMANet (Ours)

(d) IFCNN (e) GACN (f) FusionDiff

Figure 1: (a)& (b): Input image pair. (c)-(f) MFIF results obtained from our approach and several existing methods. Below each image, two zoomed-in regions are shown, corresponding to the focused regions of the two input images. Our approach recovers more details around the boundaries between defocused and focused regions.

where $D_{u,v}$ represents the per-pixel defocus PSFs (also called defocus kernels) determined by scene depths.

The challenge of maintaining sharpness across varying depths is particularly significant in fields such as macro photography (Gallo, Muzzupappa, and Bruno 2014), medical imaging (Basak, Kundu, and Sarkar 2022), and microscopy (Zhou et al. 2022), where precision and clarity are crucial. Traditional methods for extending DoF, such as using a smaller aperture or employing MFIF, have inherent trade-offs. While smaller apertures can increase DoF, they also limit light-gathering ability, introducing noise and diffraction artifacts (Horn 1968). Multi-focus image fusion (MFIF) offers an alternative by merging multiple photographs captured at varying focal lengths into an all-in-focus (AIF) image with enhanced clarity and detail, particularly applicable to complex scenes where objects are at various focal distances.

In recent years, deep learning has become a dominant approach for MFIF, which can be classified into decision-based methods or reconstruction-based methods. Decision-based methods cast MFIF as a pixel-wise binary classification problem, producing a decision map indicating "focused" or "defocused" labels on image pixels; see *e.g.* Liu et al. (2017); Wang et al. (2023); Xiao, Wu, and Bi (2021); Ma et al. (2020a); Li et al. (2020). Reconstruction-based methods perform fusion on features extracted from input images and then reconstruct the AIF image from the fused features, typically done by an end-to-end deep neural network (DNN); see *e.g.* Li et al. (2019); Zhang et al. (2020); Li et al. (2024); Wang et al. (2021a). For both types of methods, extraction of defocus-blur-related features plays a critical role.

By capturing the underlying physics, the physical model of defocus blur accurately represents defocus blur and encodes its essential features. For instance, a focused region corresponds to a Dirac delta PSF while a defocused region with severe blur corresponds to a large-support defocus PSF (Aslantas and Toprak 2017; Ma et al. 2020b; Chen et al. 2024). However, this crucial model is ignored in the DNNs of most existing deep learning-based MFIF methods, limiting their interpretability and generalization performance. Therefore, we are motivated to explicitly incorporate the defocus blurring model (2) into our DNN design for MFIF.

In this paper, we propose an end-to-end MFIF DNN called Defocus Model Aware Network (DMANet). Levering a parameterized linear combination model of per-pixel defocus PSFs to address computational issues, the DMANet first estimates pixel-wise defocus descriptors as well as initial focused images from the input multi-focus images. This defocus blur estimator is constructed as a coarse-to-fine module to exploit multi-scale analysis and cross-scale similarity of defocus blur for improvement. The estimated defocus descriptors capture essential cues of defocus blur and the initial focused images partially mitigate the blur effects in input images, both benefiting the fusion process. Built upon these estimation results, a decision map estimator forms decision maps that indicate pixel-wise focus level on input images. Finally, built upon the decision maps, image fusion is performed using masks produced by an uncertainty-aware fusion module, giving a specific treatment to pixels that exhibit indefinite focus properties in the decision maps. These pixels probably correspond to the ones that are defocused in all source images or around boundaries between defocused and focused regions. For these pixels, the fusion utilizes the estimated initial focused images.

Experimental results on benchmark datasets have demonstrated the superior performance of our DMANet over state-of-the-art techniques. See Fig. 1 for an illustration. In summary, this paper makes the following contributions:

- A physical-model-driven cross-scale defocus blur estimator produces defocus descriptors and initial focused images from multi-focus images, enhancing decision map estimation and image fusion.
- An uncertainty-aware fusion module that gives a separate treatment to uncertain pixels in decision maps that are probably defocused in all source images or around boundaries between defocused and focused regions.

Related Work

There are plenty works on MFIF. Below, we only review those most relevant to our study, while a discussion of traditional methods is provided in the supplemental materials. A comprehensive overview can be referred to Liu et al. (2020); Bhat and Koundal (2021); Zhang (2022a).

Decision-based MFIF This type of methods treats MFIF as a pixel-wise classification task, training DNNs to predict decision maps that segment focused and defocused regions, guiding the fusion process; see *e.g.* Duan, Luo, and Zhang (2023); Wang et al. (2024, 2023); Wu et al. (2023); Zhao et al. (2023); Bouzos, Andreadis, and Mitianoudis (2023); Duan, Luo, and Zhang (2024). Existing works typically use convolutional network architectures that have shown success in classification and segmentation tasks (Liu et al. 2017; Tang et al. 2018; Guo et al. 2019; Ma et al. 2020a; Xu et al. 2020a; Ma et al. 2022; Amin-Naji, Aghagolzadeh, and Ezoji 2019; Xu et al. 2020b; Huang et al. 2020). For enhancement, multi-scale features (Li et al. 2020; Wang et al. 2021b; Liu et al. 2021) and global context (Xiao, Wu, and Bi 2021; Xiao et al. 2021; Nie, Hu, and Gao 2022; Chen et al. 2024; Li et al. 2023) have been exploited.

Reconstruction-based MFIF This kind of methods learns an end-to-end DNN that maps source images to a fused image, typically using an encoder-decoder architecture that sequentially performs feature extraction, feature fusion, and image reconstruction; see *e.g.* Zhao, Wang, and Lu (2019); Li et al. (2019); Zhang et al. (2020); Wang et al. (2021a); Marivani et al. (2022); Li et al. (2024). Pre-training on large datasets using pre-text tasks such as image reconstruction (Li and Wu 2019; Mustafa, Yang, and Zareapoor 2019; Luo et al. 2023a; Zhang et al. 2021b; Cheng et al. 2021) and block-masking (Liang et al. 2022) significantly benefits these methods. Recently, generative MFIF methods based on generative adversarial network (GAN) (Huang et al. 2020) and diffusion models (Li et al. 2024) have shown superior performance.

Joint methods Both types of methods above have their strengths and limitations. Decision-based methods are more interpretable and better at preserving sharpness, but they often introduce artifacts at boundaries between focused and defocused areas and are not good at handling regions that are not in focus through all input images. Reconstruction-based methods alleviate these issues by directly regressing AIF images, producing more natural boundary effects. However, they often result in fidelity loss in regions where focused parts are available in input images. To marry the merits of both types of methods, Liu et al. (2022) proposed a two-stage DNN, first reconstructing an initial fused image and then refining it via a decision-based approach. Zhang et al. (2024) constructed a dual-branch DNN with parallel reconstruction and decision branches, fusing their results for final output.

All aforementioned methods have not utilized an explicit defocus blurring model in their DNN design. In comparison, our DMANet has an explicit integration of a parameterized defocus blurring model, leading to not only higher interpretability but also better performance, even without utilizing pre-trained models on large datasets.

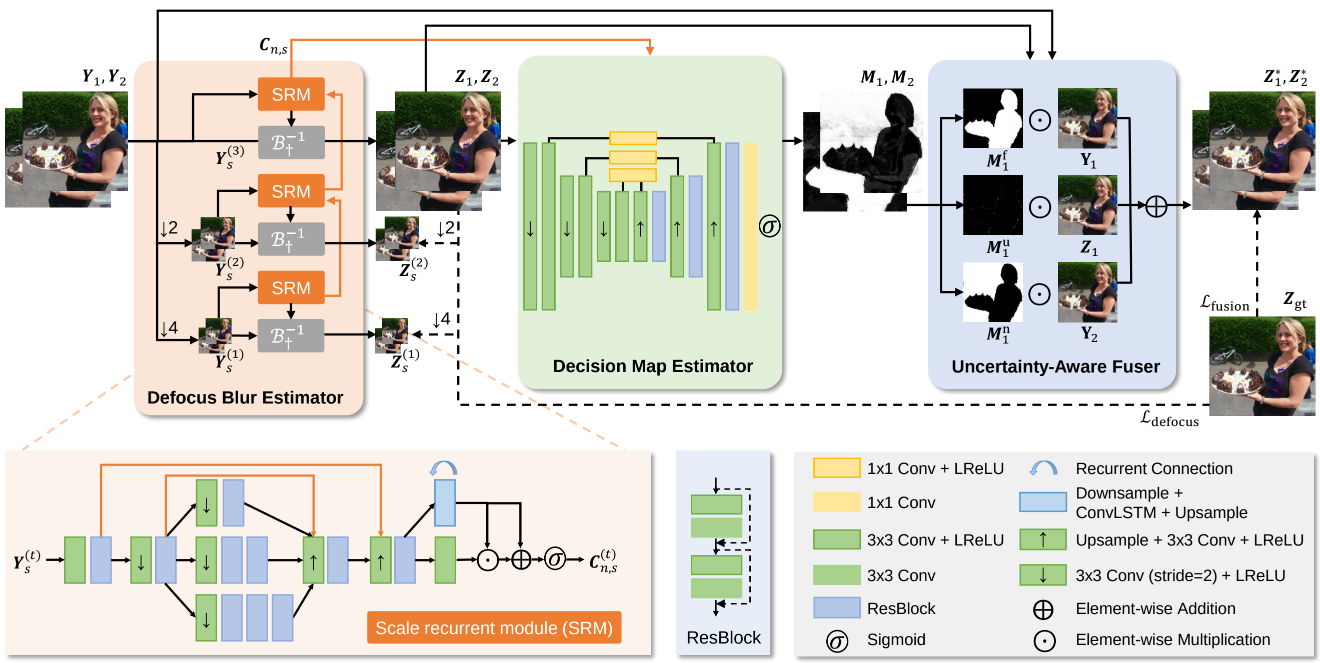


Figure 2: Overview of DMANet for MFIF. The process begins with feeding the multi-scale versions of the multi-focus source image pair $(\boldsymbol{Y}_1, \boldsymbol{Y}_2)$ into the Defocus Blur Estimator, generating defocus descriptors $(\{\boldsymbol{C}_{n,1}\}_n, \{\boldsymbol{C}_{n,2}\}_n)$ and initial focused images $(\boldsymbol{Z}_1, \boldsymbol{Z}_2)$, operating in a multi-scale fashion. Using these estimation results, the Decision Map Estimator predicts decision maps $(\boldsymbol{M}_1, \boldsymbol{M}_2)$ that distinguish between focused and defocused regions on source images. Eventually, the Uncertainty-Aware Fuser composites the source images and the initial focused images using the decision maps to obtain the AIF images $\boldsymbol{Z}_1^*, \boldsymbol{Z}_2^*$.

Methodology

MFIF aims to form an AIF image $\boldsymbol{Z} \in \mathbb{R}^{H\times W\times C}$ by merging S source images $\{\boldsymbol{Y}_s \in \mathbb{R}^{H\times W\times C}\}_{s=1}^S$, each captured with a different focal plane. In accordance with the existing literature, we assume $S = 2$. A more general formulation is provided in supplemental materials. The overall framework of our proposed DMANet is depicted in Figure 2, consisting of three parts: Defocus Blur Estimator (DBE), Decision Map Estimator (DME), and Uncertainty-Aware Fuser (UAF). Firstly, the input source images are processed through the DBE, outputting defocus blur descriptors and initial focused images. Secondly, based on DBE's estimation results, the DME predicts decision maps that indicate pixel-wise focus levels on the source images. Finally, the UAF generates the fusion result by utilizing both the DME's and the DBE's outputs, as well as the source images.

Defocus Blur Modeling and Estimation

The defocus blur modeling is based on the spatially-varying convolution model often see in single-image defocus deblurring (e.g., (Quan, Wu, and Ji 2021, 2023; Quan, Yao, and Ji 2023; Quan et al. 2024)). Let $\boldsymbol{Y}, \boldsymbol{Z}$ denote the defocused and AIF images in the discrete case. We can express (2) as

$$\boldsymbol{Y}(p,q) = \sum_{i,j} \boldsymbol{Z}(i,j)\boldsymbol{D}_{i,j}(p-i, q-j), \tag{3}$$

where $\boldsymbol{D}_{i,j}$ denotes the pixel-wise defocus PSFs. As $\boldsymbol{D}_{i,j}$ can be sufficiently large, the spatially-varying convolution process (3) is time-costly and involves many parameters encoded in $\{\boldsymbol{D}_{i,j}\}_{i,j}$ to estimate. To alleviate these issues, the spatially-varying PSF $\boldsymbol{D}_{i,j}$ is expressed using a set of spatially-invariant PSF atoms $\{\boldsymbol{A}_n\}_{n=1}^N$ as:

$$\boldsymbol{D}_{i,j} = \sum_{n=1}^{N} c_n^{i,j} \boldsymbol{A}_n, \tag{4}$$

where $c_n^{i,j}$ are the linear combination coefficients. Substituting (4) into (3), we have

$$\boldsymbol{Y}(p,q) = \sum_{i,j} \boldsymbol{Z}(i,j) \sum_{n=1}^{N} c_n^{i,j} \boldsymbol{A}_n = \sum_{n=1}^{N} (\boldsymbol{C}_n \odot \boldsymbol{Z}) \otimes \boldsymbol{A}_n, \tag{5}$$

where $\boldsymbol{C}_n = [c_n^{i,j}]_{i,j} \in \mathbb{R}^{H\times W}$ denotes the coefficient matrix w.r.t. PSF atom $\boldsymbol{A}_n$, $\odot$ the element-wise product, and $\otimes$ the discrete convolution operation. We can see that the original spatially-varying convolution process in (3) now turns to a simple form that involves element-wise product and convolution operations. The parameters to estimate are reduced to N coefficients matrices. As N is set to a relatively small number compared to the size of $\boldsymbol{D}_{i,j}$, the computational cost and number of parameters are reduced.

Based on (5), we define the defocus blur operator $\mathcal{B}$ by

$$\mathcal{B} : \boldsymbol{Z} \to \sum_{n=1}^{N} (\boldsymbol{C}_n \odot \boldsymbol{Z}) \otimes \boldsymbol{A}_n. \tag{6}$$

Then we can obtain an initial focused image using the 1st-order expansion of $\mathcal{B}^{-1}: \boldsymbol{Y} \rightarrow \boldsymbol{Z}$ as

$$\mathcal{B}^{-1} \approx \mathcal{E} + (\mathcal{E} - \mathcal{B}), \tag{7}$$

where $\mathcal{E}$ denotes an identity map. Let $\mathcal{B}_{\dagger}^{-1}$ denote the approximate inverse operator, *i.e.*,

$$\mathcal{B}_{\dagger}^{-1}: \boldsymbol{Y} \rightarrow \boldsymbol{Y} + (\boldsymbol{Y} - \sum_{n=1}^{N} (\boldsymbol{C}_n \odot \boldsymbol{Y}) \otimes \boldsymbol{A}_n). \tag{8}$$

By applying $\mathcal{B}_{\dagger}^{-1}$, one can have an initial focused image.

Structure of DBE The coefficient matrices $\{\boldsymbol{C}_n\}_n$ above are deemed as the defocus blur descriptors to estimate in our approach. The DBE is devoted to estimating these descriptors from multiple source images, as well as producing initial focused images by utilizing (8):

$$\text{DBE}: \{\boldsymbol{Y}_s\}_s \rightarrow \{\boldsymbol{C}_{n,s}, \boldsymbol{Z}_s\}_{n,s}, \tag{9}$$

where $\boldsymbol{C}_{n,s} \in \mathbb{R}^{H \times W}, \boldsymbol{Z}_s \in \mathbb{R}^{W \times H \times C}$ denote the defocus blur descriptors and the initial focused image on the s-th source image, respectively. For enhanced efficiency, the DBE employs a coarse-to-fine scale-recurrent structure. Firstly, multi-scale versions of source images, denoted by $\boldsymbol{Y}_s^{(1)}, \cdots, \boldsymbol{Y}_s^{(T)}$, are formed via downsampling with factors 2^{T-1}, ..., 2^0, respectively. At each scale t, the DBE estimates the defocus descriptors $\{\boldsymbol{C}_{n,s}^{(t)}\}_{n,s}$ and the initial focused images $\{\boldsymbol{Z}_s^{(t)}\}_s$, utilizing the results from the previous scale as well as the input images at the current scale.

Since defocus blur has strong cross-scale similarity in blur shape, $\boldsymbol{C}_{n,s}^{(t)}$ is estimated by a scale-recurrent module (SRM) implemented based on ConvLSTM, as shown in Figure 2. Afterward, $\boldsymbol{Z}_s^{(t)}$ is calculated based on (8). This process can be expressed as: for $t = 1, \cdots, T$,

$$\{\boldsymbol{C}_{n,s}^{(t)}\}_{n,s} = \text{SRM}(\{\boldsymbol{Y}_s^{(t)}, \boldsymbol{C}_{n,s}^{t-1}\uparrow_2, \boldsymbol{Z}_s^{t-1}\uparrow_2\}_{n,s}), \tag{10}$$

$$\boldsymbol{Z}_s^{(t)} = \mathcal{B}_{\dagger}^{-1}(\boldsymbol{Y}_s^{(t)}; \{\boldsymbol{C}_{s,n}^{(t)}\}_n, \{\boldsymbol{A}_n\}_n) \tag{11}$$

$$= 2\boldsymbol{Y}_s^{(t)} - \sum_n \left(\boldsymbol{C}_{s,n}^{(t)} \odot \boldsymbol{Y}_s^{(t)}\right) \otimes \boldsymbol{A}_n, \tag{12}$$

where $\boldsymbol{Z}_s^{(t)} = \boldsymbol{Y}_s^{(1)}$, and $\uparrow_2$ denotes the upsampling operation by factor 2. Regarding the kernel atoms, we define $\boldsymbol{A}_n$ to be a Gaussian kernel with size $n \times n$. The standard deviation parameters $\{\sigma_n\}_n$ of the Gaussian kernels are defined as learnable parameters, initialized as $\sigma_n = 0.5(n-1)$. The results at the finest scale T are used as the output of DBE:

$$\boldsymbol{C}_{n,s} = \boldsymbol{C}_{n,s}^{(T)}, \;\; \boldsymbol{Z}_s = \boldsymbol{Z}_s^{(T)}, \;\; \forall n, s. \tag{13}$$

The detailed SRM is provided in the supplemental material.

Decision Map Estimation and Image Fusion

Structure of DME Using the estimated defocus blur descriptors $\{\boldsymbol{C}_{n,s}\}_{n,s}$ and initial focused images $\{\boldsymbol{Z}_s\}_s$ as input, the DME generates a soft decision map $\boldsymbol{M}_s \in [0,1]^{H \times W}$ for each source image, measuring focus level on each pixel location in the source image:

$$\text{DME}: \{\boldsymbol{Z}_s, \boldsymbol{C}_{n,s}\}_{n,s} \rightarrow \{\boldsymbol{M}_s\}_s. \tag{14}$$

As shown in Figure 2, the DME first concatenates its inputs and then processes them using an encoder-decoder structure. Concretely, the DME sequentially contains three convolutional encoder blocks (EBs) with downsampling and three convolutional decoder blocks (DBs) with upsampling. Each EB and its corresponding DB are also connected to a skip-connection block composed of convolutional layers. Finally, a 1×1 Conv layer with a Sigmoid activation is applied to obtain the decision maps.

Structure of UAF The UAF fuses the source images $\{\boldsymbol{Y}_s\}_s$ and the initial focused results $\{\boldsymbol{Z}_s\}_s$ by utilizing the decision maps $\{\boldsymbol{M}_s\}_s$, resulting in an AIF image $\boldsymbol{Z}$:

$$\text{UAF}: \{(\boldsymbol{Y}_s, \boldsymbol{Z}_s, \boldsymbol{M}_s)\}_s \rightarrow \boldsymbol{Z}. \tag{15}$$

Given $\boldsymbol{M}_s$, we form three masks $\boldsymbol{M}_s^{\text{f}}, \boldsymbol{M}_s^{\text{n}}, \boldsymbol{M}_s^{\text{u}}$ as follows:

$$\boldsymbol{M}_s^{\text{f}}(i,j) = \mathcal{I}(\boldsymbol{M}_s(i,j) \leq 0.5 - \gamma), \tag{16}$$

$$\boldsymbol{M}_s^{\text{n}}(i,j) = \mathcal{I}(\boldsymbol{M}_s(i,j) \geq 0.5 + \gamma), \tag{17}$$

$$\boldsymbol{M}_s^{\text{u}}(i,j) = \mathcal{I}(\boldsymbol{M}_s(i,j) \in (0.5 - \gamma, 0.5 + \gamma)), \tag{18}$$

where $\mathcal{I}$ denotes an indicator function outputting 1 if the condition holds and 0 otherwise, and $\gamma \in (0, 0.5)$ is a threshold. Based on $\boldsymbol{M}_s$, the three masks identify three types of regions: focused regions in $\boldsymbol{Y}_1$, focused regions in $\boldsymbol{Y}_2$, and uncertain regions that are challenging to distinguish whether focused or defocused in $\boldsymbol{Y}_1$ or $\boldsymbol{Y}_2$. The uncertainty is measured by the confidence level in the decision map, *i.e.*, whether $\boldsymbol{M}_s(i,j)$ falls into the range $[0.5 - \gamma, 0, 5 + \gamma]$.

Afterward, the UAF calculates the AIF result $\boldsymbol{Z}_s^*$:

$$\boldsymbol{Z}_s^* = \boldsymbol{M}_s^{\text{f}} \odot \boldsymbol{Y}_1 + \boldsymbol{M}_s^{\text{n}} \odot \boldsymbol{Y}_2 + \boldsymbol{M}_s^{\text{u}} \odot \boldsymbol{Z}_s, \tag{19}$$

where $\odot$ represents element-wise multiplication. That is, for regions being focused in source images with high confidence, we directly utilize the source image pixels for fusion. Otherwise, we leverage the initial focused image $\boldsymbol{Z}_s$ to enhance the fusion result. The reason is, that those uncertain pixels are likely to be defocused on all source images or lie around the boundaries between focused/defocused images. Intuitively, using the initial focused images for these pixels is a better choice. During training, supervision is imposed on $\boldsymbol{Z}_s^*$. In inference, we can take either $\boldsymbol{Z}_1^*$ or $\boldsymbol{Z}_2^*$, or directly average them, as the final output. In practice, we found not noticeable performance effects on these schemes. For a faster speed, we define $\boldsymbol{Z} = \boldsymbol{Z}_1^*$ as the final AIF result.

Loss Function

The training loss is defined by

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{fusion}} + \beta \mathcal{L}_{\text{defocus}}, \tag{20}$$

where $\beta \in \mathbb{R}^+$ is a weight balancing the fusion loss $\mathcal{L}_{\text{fusion}}$ and the defocus estimation loss $\mathcal{L}_{\text{defocus}}$. The fusion loss $\mathcal{L}_{\text{fusion}}$ measures the fusion accuracy by comparing the ground truth (GT) AIF image $\boldsymbol{Z}_{\text{gt}}$ with the predictions $\{\boldsymbol{Z}_s^*\}_s$:

$$\mathcal{L}_{\text{fusion}} := \sum_{s=1}^{S} \|\boldsymbol{Z}_s^* - \boldsymbol{Z}_{\text{gt}}\|_{\text{F}}^2. \tag{21}$$

The defocus estimation loss $\mathcal{L}_{\text{defocus}}$ measures the accuracy of DBE by measuring the difference between the predicted

Table 1: Quantitative comparison on MFI-WHU dataset and computational complexity comparison.

Method	PSNR(dB)	SSIM	LPIPS	NMI	Q_G	Q_M	MI	Q_Y	ARank	Size(M)	Time(s)
IFCNN	37.155	0.9911	0.0078	0.8993	0.6624	0.7920	6.5535	0.9416	5.50	0.084	0.02
GACN	36.623	0.9962	0.0099	**1.1954**	**0.7404**	**2.4860**	**8.7133**	**0.9875**	**3.13**	0.069	0.15
GRFusion	37.062	**0.9969**	0.0078	1.1633	0.7182	2.3681	8.4824	0.9723	3.25	9.857	0.97
FusionDiff	37.309	0.9952	**0.0049**	1.0316	0.6927	1.3589	7.5189	0.9623	4.13	26.915	135.91
DB-MFIF	**38.202**	0.9968	0.0067	1.0711	0.6799	1.6495	7.8116	0.9472	3.75	2.700	0.01
DMANet	**42.898**	**0.9980**	**0.0042**	**1.2225**	**0.7387**	**2.5281**	**8.9069**	**0.9895**	**1.13**	4.065	0.03

initial focused image $Z_s^{(t)}$ and the GT across scales. Let $Z_{\text{gt}}^{(t)}$ denote GT's downsampled version at scale t. We define

$$\mathcal{L}_{\text{defocus}} := \sum_{s=1}^{S}\sum_{t=1}^{T}\lambda^{(t)}\|Z_s^{(t)} - Z_{\text{gt}}^{(t)}\|_{\text{F}}^2. \tag{22}$$

For simplicity, we set the weight $\lambda^{(t)} = 1$ for all t.

Experiments

Experimental Setup

Datasets Following Wang et al. (2023), model training is done on 10000 synthesized image pairs sourced from Agustsson and Timofte (2017) and Wang et al. (2019). The performance is then evaluated on the following five benchmark datasets: MFI-WHU (Zhang et al. 2021a), Lytro (Nejati, Samavi, and Shirani 2015), SIMIF (also termed TSAI) (Tsai 2024), MFFW (Xu et al. 2020c), and OR-PAM (Zhou et al. 2022). The former four are natural image datasets, and the last one is a biological image dataset.

Metrics When GTs are unavailable, we adopt five GT-independent metrics from Wang et al. (2023): Normalized Mutual Information (NMI) (Hossny, Nahavandi, and Creighton 2008), Mutual Information (MI) (Qu, Zhang, and Yan 2002), Q_G (Xydeas, Petrovic et al. 2000), Q_M (Wang and Liu 2008), and Q_Y (Li, Hong, and Wu 2008), measuring fusion performance from different aspects. NMI and MI rely on information theory, Q_M on feature similarity, and Q_Y (Li, Hong, and Wu 2008) is based on structural similarity. When GTs are available, we also calculate three full-reference performance metrics: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS) defined on AlexNet. For all metrics, higher scores indicate better performance. In addition, we compute the rank of a method among all compared methods for each metric, and the average rank score over all metrics, denoted as ARank. In the tables on performance comparison, we use **RED** and **BLUE** to denote the best and second-best results, respectively.

Compared methods Six supervised MFIF techniques are chosen for experimental comparison: IFCNN (Zhang et al. 2020), GACN (Ma et al. 2022), EAY-Net (Wang et al. 2023), GRFusion (Li et al. 2023), FusionDiff (Li et al. 2024), and DB-MFIF (Zhang et al. 2024). Due to space limit, comparisons with traditional spatial domain-based methods (DSIFT (Liu, Liu, and Wang 2015) and MFF (Li, Li, and Zhang 2015)) and transform domain-based ones (MWGF (Zhou, Li, and Wang 2014) and GFF (Li, Kang, and Hu 2013)) are shown in the supplementary materials. As the original training data vary across these methods, we retrain GACN, GRFusion, and FusionDiff using the same data as ours for a fair comparison. For IFCNN and DB-MFIF, we use their published models to obtain results, as no training code is available. For EAY-Net, we quote its results from literature whenever possible; otherwise, we leave its results blank as neither the trained model nor the code is available.

Table 2: Quantitative comparison on Lytro dataset.

Method	NMI	Q_G	Q_M	MI	Q_Y	ARank
IFCNN	0.9374	0.6634	0.9484	6.9018	0.9471	6.2
GACN	1.1668	0.7258	2.4589	8.6112	0.9776	4.0
EAY-Net	1.1853	**0.7271**	2.5376	8.7483	0.9781	2.8
GRFusion	**1.1879**	0.7263	**2.6397**	**8.9203**	**0.9866**	**2**
FusionDiff	0.9223	0.6555	0.7601	6.9166	0.9461	6.8
DB-MFIF	1.0589	0.6908	1.7154	7.9532	0.9646	5.0
DMANet	**1.1897**	**0.7278**	**2.6251**	**8.9338**	**0.9870**	**1.2**

Performance Evaluation

Quantitative comparison Table 1 presents the quantitative results on MFI-WHU which includes GTs. Our DMANet achieves the best ARank and excels across most metrics, particularly showcasing a significantly higher PSNR compared to other methods. Tables 2,4,3, and 5 display the results on Lytro, MFFW, SIMIF, and OR-PAM, respectively, where GTs are unavailable. Our DMANet consistently ranks first in terms of ARank across all datasets, demonstrating its robust performance. Furthermore, it consistently ranks first in a majority of individual metrics and at least second in others, indicating its ability to strike a superior balance across diverse image quality aspects. Notably, DMANet excels on SIMIF which comprises high-resolution images, achieving the highest scores across five metrics. Interestingly, while DMANet demonstrates consistent top performance, the next best-performing method varies across datasets: GRFusion on Lytro, EAY-Net on SIMIF and MFFW, and GACN on OR-PAM. This observation underscores the diverse challenges posed by different datasets and further highlights the robustness and generalization of our approach.

Qualitative comparison Figure 3 visually validates our method's superiority on the MFI-WHU dataset. DMANet produces cleaner and more accurate object boundaries, evident in the minimal errors observed in the difference maps compared to the GT. Its advantage is evident in the accurate

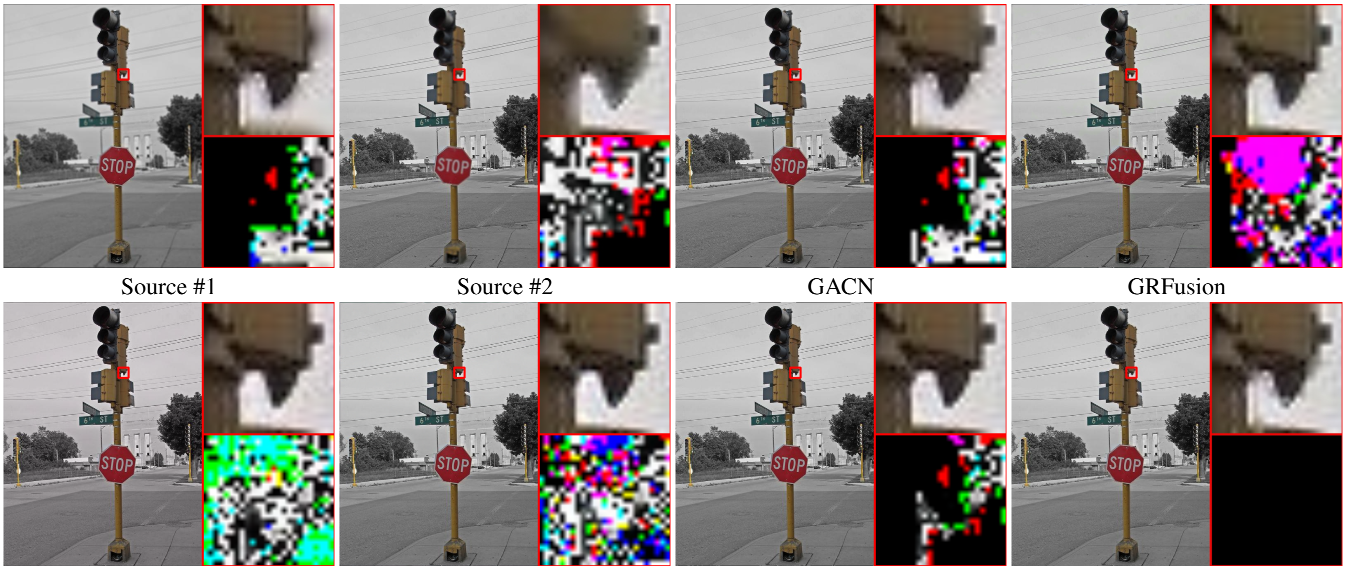


Figure 3: Fused images by different methods on a sample from MFI-WHU. Behind each image, a zoomed-region and its difference from GT are provided.

Table 3: Quantitative comparison on SIMIF dataset.

Method	NMI	Q_G	Q_M	MI	Q_Y	ARank
IFCNN	1.0382	0.6748	0.9423	7.7098	0.9304	6.2
GACN	1.2930	0.7568	2.5497	9.5533	0.9724	3.8
EAY-Net	**1.3042**	0.7570	2.5669	**9.6268**	0.9739	**2.6**
GRFusion	1.2899	**0.7581**	**2.6120**	9.5785	**0.9770**	**2.6**
FusionDiff	1.0353	0.6771	0.8508	7.7313	0.9405	6
DB-MFIF	1.1470	0.6728	1.7421	8.5749	0.9145	5.8
DMANet	**1.3149**	**0.7622**	**2.6432**	**9.7668**	**0.9840**	**1**

Table 4: Quantitative comparison on MFFW dataset.

Method	NMI	Q_G	Q_M	MI	Q_Y	ARank
IFCNN	0.8205	0.5890	0.6577	5.5283	0.8777	6.0
GACN	1.0820	0.6722	2.4375	7.3923	0.9333	3.4
EAY-Net	**1.1777**	**0.6983**	**2.5238**	**7.9861**	**0.9824**	**1.6**
GRFusion	1.0823	0.6590	2.3715	7.7050	0.9073	3.6
FusionDiff	0.8164	0.5848	0.6012	5.7824	0.8795	6.4
DB-MFIF	1.0589	0.6908	1.7154	7.9532	0.9646	5.6
DMANet	**1.1494**	**0.6987**	**2.5271**	**8.2002**	**0.9578**	**1.4**

fusion of the focused traffic pole (source image #1) with the in-focus background (source image #2), even revealing finer details than individual source images. Figure 4 further demonstrates our method's robustness across diverse datasets lacking GTs. The difference maps, generated against the zoom-in regions of partially focused source image #2, illustrate our approach's ability to accurately identify and merge in-focus regions, even in challenging edge cases. For example, the first two examples showcase scenarios with intertwined focused/defocused regions: the doll's hair (defocused) and

Table 5: Quantitative comparison on OR-PAM dataset.

Method	NMI	Q_G	Q_M	MI	Q_Y	ARank
IFCNN	0.4057	0.5272	0.2719	2.2872	0.8170	4.6
GACN	**0.8434**	**0.6826**	**2.3648**	**4.7866**	**0.9640**	**1.8**
EAY-Net	n/a	n/a	n/a	n/a	n/a	n/a
GRFusion	0.8089	0.6626	2.1754	4.6187	0.9367	3.0
FusionDiff	0.4346	0.5147	0.2099	2.4587	0.8129	5.0
DB-MFIF	0.3928	0.4952	0.2492	2.2438	0.8208	5.4
DMANet	**0.9406**	**0.6936**	**2.4111**	**5.3699**	**0.9533**	**1.2**

hand (focused) in the first, and the closer red flowers (focused) and farther purple flowers (defocused) in the second. Despite these challenges, our method accurately identifies the in-focus areas, as evidenced by the black regions in the difference maps. In all examples, the proposed DMANet consistently produces sharper boundaries (*e.g.*, the doll's hand and clear hair contours in the first sample), more realistic edges (*e.g.*, detailed flower petals in the second sample), and fewer artifacts (*e.g.*, the cleaner rendering of the biological tissue in the third sample) compared to other methods. The minimal artifacts and noise in our difference maps highlight our approach's robustness.

Computational complexity comparison Table 1 also reports the computational complexity in terms of the number of model parameters and the inference time as well as the number of floating-point operations (FLOPs) in processing an image pair of size 512×512. Our DMANet model is much smaller than GRFusion and FusionDiff, and its speed is much smaller than GACN, GRFusion, and FusionDiff. These results indicate that the performance gains of our proposed approach mainly comes from its architecture design, rather

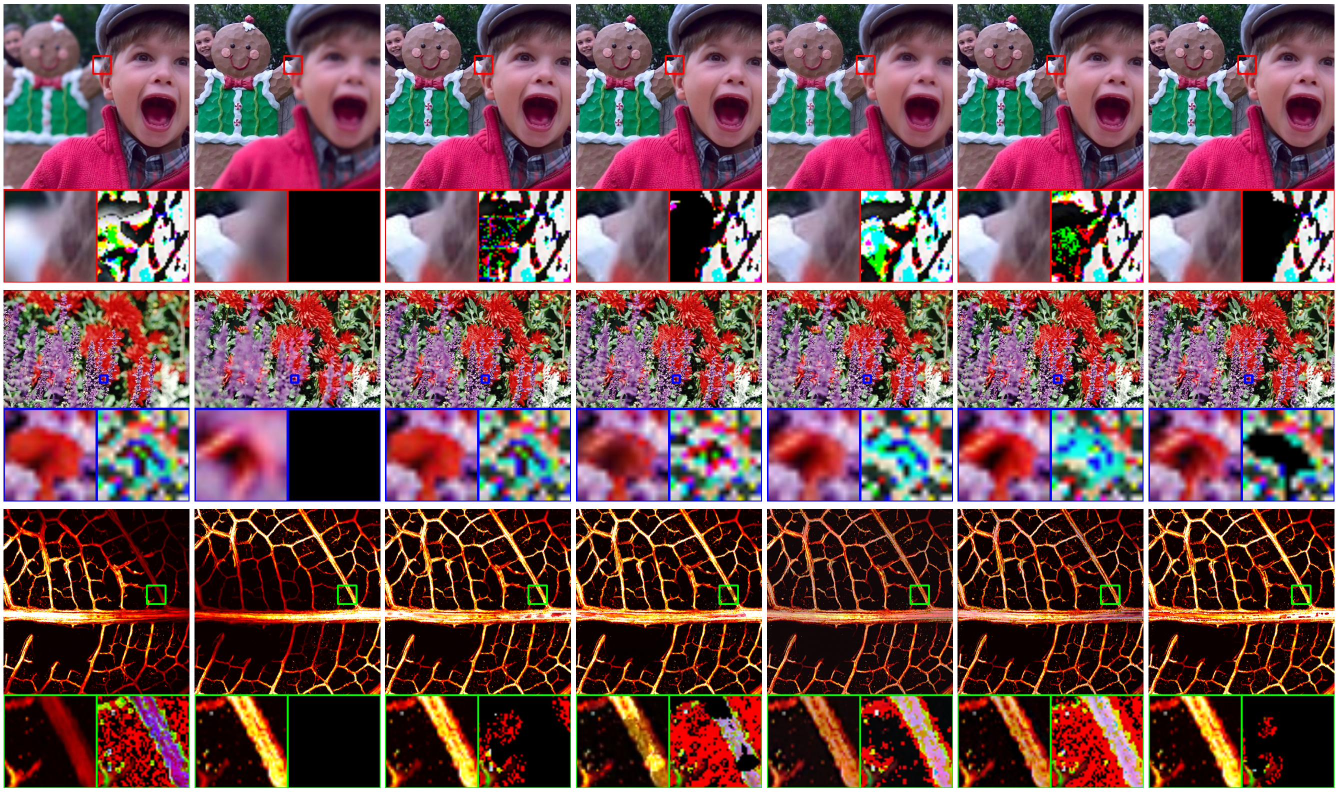


Figure 4: Fused images by different methods on the samples from on Lytro, MFFW, and OR-PAM datasets (from top to bottom). Below each image, a zoomed-region and its difference from the focused region in Source #2 are provided.

Table 6: Results of ablation studies on Lytro dataset.

	NMI	Q_G	Q_M	MI	Q_Y	ARank
1-scale	1.1879	0.7273	2.6276	8.9198	0.9865	2
w/o DBE	1.1862	0.7251	2.5479	8.9070	0.9858	4.8
w/o SR	1.1865	0.7270	2.6174	8.9099	0.9856	4.2
w/o $\mathcal{L}_{\text{defocus}}$	1.1878	0.7272	**2.6306**	8.9189	0.9858	2.6
Original	**1.1897**	**0.7278**	2.6251	**8.9338**	**0.9870**	**1.4**

than significantly increasing model complexity.

Ablation Studies

We constructed several baseline methods to evaluate the contributions of different components in our approach: (a) 1-scale: only one scale is used the DBE, with the ConvLSTM replaced a convolutional block with a similar size; (b) w/o DBE: remove the DBE from DMANet and the DME accepts the source images as input. (c) w/o SR: discard the self-recurrent (SR) structure by replacing the ConLSTM in SRM with a convolutional block of a similar size. (d) w/o $\mathcal{L}_{\text{defocus}}$: remove the defocus estimation loss. The results of these baselines on the Lytro dataset are listed in Table 6, where all baselines show a noticeable performance decrease. These observations confirm that each component in our approach contributes noticeably to the performance.

Conclusion

In this work, we proposed DMANet, a deep learning-based MFIF method that integrates explicit defocus blur modeling into its network design. DMANet consists of three main components: the DBE employs a coarse-to-fine scale-recurrent structure to generate defocus blur descriptors and initial focused images, done by introducing a parameterized defocus blurring model; the DME utilizes the DBE's results to produce soft decision maps; and the UAF generates final fusion result by identifying focused regions in the source images, with a specific treatment on the uncertain regions utilizing the focused images produced by DME. Our approach achieved state-of-the-art results on five benchmark datasets.

Acknowledgments This work was supported in part by National Natural Science Foundation of China (Nos. 62372186 and 62302019), Natural Science Foundation of Guangdong Province (Nos. 2022A1515011755 and 2023A1515012841), Fundamental Research Funds for the Central Universities (No. x2jsD2230220), National Key Research and Development Program of China (No. 2024YFE0105400), China Postdoctoral Science Foundation (No. 2022M720236), and Singapore MOE Academic Research Fund (AcRF) Tier 1 Research Grant (No. A-8000981-00-00).

References

Agarwala, A.; Dontcheva, M.; Agrawala, M.; Drucker, S.; Colburn, A.; Curless, B.; Salesin, D.; and Cohen, M. 2004. Interactive Digital Photomontage. *ACM Transactions on Graphics*, 23(3): 294–302.

Agustsson, E.; and Timofte, R. 2017. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 126–135.

Amin-Naji, M.; Aghagolzadeh, A.; and Ezoji, M. 2019. Ensemble of CNN for Multi-Focus Image Fusion. *Information Fusion*, 51: 201–214.

Aslantas, V.; and Toprak, A. N. 2017. Multi-focus image fusion based on optimal defocus estimation. *Computers & Electrical Engineering*, 62: 302–318.

Basak, H.; Kundu, R.; and Sarkar, R. 2022. MFSNet: A Multi Focus Segmentation Network for Skin Lesion Segmentation. *Pattern Recognition*, 128: 108673.

Bhat, S.; and Koundal, D. 2021. Multi-focus image fusion techniques: a survey. *Artificial Intelligence Review*, 54: 5735–5787.

Bouzos, O.; Andreadis, I.; and Mitianoudis, N. 2023. A Convolutional Neural Network-based Conditional Random Field model for Structured Multi-Focus Image Fusion Robust to Noise. *IEEE Transactions on Image Processing*.

Burt, P. J.; and Adelson, E. H. 1985. Merging Images Through Pattern Decomposition. In *Applications of Digital Image Processing VIII*, volume 0575, 173–181.

Chen, P.; Jiang, J.; Li, L.; and Yao, J. 2024. A defocus and similarity attention-based cascaded network for multi-focus and misaligned image fusion. *Information Fusion*, 103: 102125.

Cheng, C.; Wu, X.-J.; Xu, T.; and Chen, G. 2021. Unifusion: A lightweight unified image fusion network. *IEEE Transactions on Instrumentation and Measurement*, 70: 1–14.

Duan, Z.; Luo, X.; and Zhang, T. 2023. Multi-focus image fusion via gradient guidance progressive network. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, 2159–2164. IEEE.

Duan, Z.; Luo, X.; and Zhang, T. 2024. Combining transformers with CNN for multi-focus image fusion. *Expert Systems with Applications*, 235: 121156.

Gallo, A.; Muzzupappa, M.; and Bruno, F. 2014. 3D Reconstruction of Small Sized Objects from a Sequence of Multi-Focused Images. *Journal of Cultural Heritage*, 15(2): 173–182.

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

Hasinoff, S. W.; Kutulakos, K. N.; Durand, F.; and Freeman, W. T. 2009. Time-Constrained Photography. In *Proceedings of the IEEE International Conference on Computer Vision*, 333–340.

Hill, P. R.; Canagarajah, C. N.; Bull, D. R.; et al. 2002. Image Fusion Using Complex Wavelets. In *BMVC*, 1–10. Citeseer.

Horn, B. K. P. 1968. Focusing. *Artificial Intelligence Memo*.

Hossny, M.; Nahavandi, S.; and Creighton, D. 2008. Comments on'Information measure for performance of image fusion'. *Electronics Letters*.

Huang, J.; Le, Z.; Ma, Y.; Mei, X.; and Fan, F. 2020. A Generative Adversarial Network with Adaptive Constraints for Multi-Focus Image Fusion. *Neural Computing and Applications*, 32(18): 15119–15129.

Kutulakos, K. N.; and Hasinoff, S. W. 2009. Focal Stack Photography: High-Performance Photography with a Conventional Camera. *MVA*.

Lee, M.; and Tai, Y.-W. 2016. Robust All-in-Focus Super-Resolution for Focal Stack Photography. *IEEE Transactions on Image Processing*, 25(4): 1887–1897.

Levoy, M.; Ng, R.; Adams, A.; Footer, M.; and Horowitz, M. 2006. Light Field Microscopy. In *Proceedings of the ACM SIGGRAPH Conference*, SIGGRAPH '06, 924–934. New York, NY, USA. ISBN 978-1-59593-364-5.

Li, H.; Li, L.; and Zhang, J. 2015. Multi-focus image fusion based on sparse feature matrix decomposition and morphological filtering. *Optics Communications*, 342: 1–11.

Li, H.; Nie, R.; Cao, J.; Guo, X.; Zhou, D.; and He, K. 2019. Multi-focus image fusion using u-shaped networks with a hybrid objective. *IEEE Sensors Journal*, 19(21): 9755–9765.

Li, H.; Wang, D.; Huang, Y.; Zhang, Y.; and Yu, Z. 2023. Generation and recombination for multifocus image fusion with free number of inputs. *IEEE Transactions on Circuits and Systems for Video Technology*.

Li, H.; and Wu, X.-J. 2019. DenseFuse: A fusion approach to infrared and visible images. *IEEE Transactions on Image Processing*, 28(5): 2614–2623.

Li, J.; Guo, X.; Lu, G.; Zhang, B.; Xu, Y.; Wu, F.; and Zhang, D. 2020. DRPL: Deep regression pair learning for multi-focus image fusion. *IEEE Transactions on Image Processing*, 29: 4816–4831.

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

Li, S.; Hong, R.; and Wu, X. 2008. A novel similarity based quality metric for image fusion. In *Proceedings of International Conference on Audio, Language and Image Processing*, 167–172. IEEE.

Li, S.; Kang, X.; and Hu, J. 2013. Image Fusion With Guided Filtering. *IEEE Transactions on Image Processing*, 22(7): 2864–2875.

Liang, P.; Jiang, J.; Liu, X.; and Ma, J. 2022. Fusion from decomposition: A self-supervised decomposition approach for image fusion. In *Proceedings of the European Conference on Computer Vision*, 719–735. Springer.

Liu, Y.; Chen, X.; Peng, H.; and Wang, Z. 2017. Multi-focus image fusion with a deep convolutional neural network. *Information Fusion*, 36: 191–207.

Liu, Y.; Liu, S.; and Wang, Z. 2015. Multi-focus image fusion with dense SIFT. *Information Fusion*, 23: 139–155.

Liu, Y.; Wang, L.; Cheng, J.; and Chen, X. 2021. Multiscale feature interactive network for multifocus image fusion. *IEEE Transactions on Instrumentation and Measurement*, 70: 1–16.

Liu, Y.; Wang, L.; Cheng, J.; Li, C.; and Chen, X. 2020. Multi-focus image fusion: A survey of the state of the art. *Information Fusion*, 64: 71–91.

Liu, Y.; Wang, L.; Li, H.; and Chen, X. 2022. Multi-focus image fusion with deep residual learning and focus property detection. *Information Fusion*, 86: 1–16.

Luo, X.; Gao, Y.; Wang, A.; Zhang, Z.; and Wu, X.-J. 2023a. IFSepR: A general framework for image fusion based on separate representation learning. *IEEE Transactions on Multimedia*.

Luo, X.; Peng, J.; Zhao, W.; Xian, K.; Lu, H.; and Cao, Z. 2023b. Point-and-Shoot All-in-Focus Photo Synthesis From Smartphone Camera Pair. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(5): 2088–2101.

Ma, B.; Yin, X.; Wu, D.; Shen, H.; Ban, X.; and Wang, Y. 2022. End-to-end learning for simultaneously generating decision map and multi-focus image fusion result. *Neurocomputing*, 470: 204–216.

Ma, H.; Liao, Q.; Zhang, J.; Liu, S.; and Xue, J.-H. 2020a. An α-matte boundary defocus model-based cascaded network for multi-focus image fusion. *IEEE Transactions on Image Processing*, 29: 8668–8679.

Ma, H.; Liao, Q.; Zhang, J.; Liu, S.; and Xue, J.-H. 2020b. An α-matte boundary defocus model-based cascaded network for multi-focus image fusion. *IEEE Transactions on Image Processing*, 29: 8668–8679.

Marivani, I.; Tsiligianni, E.; Cornelis, B.; and Deligiannis, N. 2022. Designing CNNs for Multimodal Image Restoration and Fusion via Unfolding the Method of Multipliers. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(9): 5830–5845.

Mustafa, H. T.; Yang, J.; and Zareapoor, M. 2019. Multi-scale convolutional neural network for multi-focus image fusion. *Image and Vision Computing*, 85: 26–35.

Nejati, M.; Samavi, S.; and Shirani, S. 2015. Multi-focus image fusion using dictionary-based sparse representation. *Information Fusion*, 25: 72–84.

Nie, X.; Hu, B.; and Gao, X. 2022. MLNet: A Multi-domain Lightweight Network for Multi-focus Image Fusion. *IEEE Transactions on Multimedia*.

Qu, G.; Zhang, D.; and Yan, P. 2002. Information measure for performance of image fusion. *Electronics letters*, 38(7): 1.

Quan, Y.; Wu, Z.; and Ji, H. 2021. Gaussian Kernel Mixture Network for Single Image Defocus Deblurring. In *Advances in Neural Information Processing Systems*, volume 34, 20812–20824.

Quan, Y.; Wu, Z.; and Ji, H. 2023. Neumann Network With Recursive Kernels for Single Image Defocus Deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5754–5763.

Quan, Y.; Wu, Z.; Xu, R.; and Ji, H. 2024. Deep Single Image Defocus Deblurring via Gaussian Kernel Mixture Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(12): 11361–11377.

Quan, Y.; Yao, X.; and Ji, H. 2023. Single Image Defocus Deblurring via Implicit Neural Inverse Kernels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 12600–12610.

Rockinger, O. 1997. Image sequence fusion using a shift-invariant wavelet transform. In *Proceedings of international conference on image processing*, volume 3, 288–291. IEEE.

Tang, H.; Xiao, B.; Li, W.; and Wang, G. 2018. Pixel convolutional neural network for multi-focus image fusion. *Information Sciences*, 433: 125–141.

Tsai, C.-C. 2024. Standard Images for Multifocus Image Fusion. https://www.mathworks.com/matlabcentral/fileexchange/45992-standard-images-for-multifocus-image-fusion.

Wang, C.; Zang, Y.; Zhou, D.; Mei, J.; Nie, R.; and Zhou, L. 2024. Robust multi-focus image fusion using focus property detection and deep image matting. *Expert Systems with Applications*, 237: 121389.

Wang, C.; Zhou, D.; Zang, Y.; Nie, R.; and Guo, Y. 2021a. A deep and supervised atrous convolutional model for multi-focus image fusion. *IEEE Sensors Journal*, 21(20): 23069–23084.

Wang, P.-w.; and Liu, B. 2008. A novel image fusion metric based on multi-scale analysis. In *Proceedings of International Conference on Signal Processing*, 965–968. IEEE.

Wang, Y.; Wang, L.; Yang, J.; An, W.; and Guo, Y. 2019. Flickr1024: A large-scale dataset for stereo image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 0–0.

Wang, Y.; Xu, S.; Liu, J.; Zhao, Z.; Zhang, C.; and Zhang, J. 2021b. MFIF-GAN: A new generative adversarial network for multi-focus image fusion. *Signal Processing: Image Communication*, 96: 116295.

Wang, Z.; Li, X.; Zhao, L.; Duan, H.; Wang, S.; Liu, H.; and Zhang, X. 2023. When Multi-Focus Image Fusion Networks Meet Traditional Edge-Preservation Technology. *International Journal of Computer Vision*, 1–24.

Wu, P.; Jiang, L.; Hua, Z.; and Li, J. 2023. Multi-focus image fusion: Transformer and shallow feature attention matters. *Displays*, 76: 102353.

Xiao, B.; Wu, H.; and Bi, X. 2021. DTMNet: a discrete tchebichef moments-based deep neural network for multi-focus image fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 43–51.

Xiao, B.; Xu, B.; Bi, X.; and Li, W. 2021. Global-feature encoding U-Net (GEU-Net) for multi-focus image fusion. *IEEE Transactions on Image Processing*, 30: 163–175.

Xu, S.; Ji, L.; Wang, Z.; Li, P.; Sun, K.; Zhang, C.; and Zhang, J. 2020a. Towards reducing severe defocus spread effects for multi-focus image fusion via an optimization based strategy. *IEEE Transactions on Computational Imaging*, 6: 1561–1570.

Xu, S.; Ji, L.; Wang, Z.; Li, P.; Sun, K.; Zhang, C.; and Zhang, J. 2020b. Towards Reducing Severe Defocus Spread Effects for Multi-Focus Image Fusion via an Optimization Based Strategy. *IEEE Transactions on Computational Imaging*, 6: 1561–1570.

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

Xydeas, C. S.; Petrovic, V.; et al. 2000. Objective image fusion performance measure. *Electronics letters*, 36(4): 308–309.

Zhan, K.; Kong, L.; Liu, B.; and He, Y. 2019. Multimodal image seamless fusion. *Journal of Electronic Imaging*, 28(2): 023027–023027.

Zhang, H.; Le, Z.; Shao, Z.; Xu, H.; and Ma, J. 2021a. MFF-GAN: An unsupervised generative adversarial network with adaptive and gradient joint constraints for multi-focus image fusion. *Information Fusion*, 66: 40–53.

Zhang, J.; Liao, Q.; Ma, H.; Xue, J.-H.; Yang, W.; and Liu, S. 2024. Exploit the Best of Both End-to-End and Map-Based Methods for Multi-Focus Image Fusion. *IEEE Transactions on Multimedia*.

Zhang, J.; Shao, J.; Chen, J.; Yang, D.; and Liang, B. 2021b. Polarization image fusion with self-learned fusion strategy. *Pattern Recognition*, 118: 108045.

Zhang, X. 2022a. Deep learning-based multi-focus image fusion: A survey and a comparative study. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9): 4819–4838.

Zhang, X. 2022b. Deep Learning-Based Multi-Focus Image Fusion: A Survey and a Comparative Study. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9): 4819–4838.

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

Zhao, F.; Zhao, W.; Lu, H.; Liu, Y.; Yao, L.; and Liu, Y. 2023. Depth-distilled multi-focus image fusion. *IEEE Transactions on Multimedia*.

Zhao, W.; Wang, D.; and Lu, H. 2019. Multi-focus image fusion with a natural enhancement via a joint multi-level deeply supervised convolutional neural network. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(4): 1102–1115.

Zhou, W.; He, J.; Li, Y.; Sun, Z.; Chen, J.; Wang, L.; Hui, H.; and Chen, X. 2022. Multi-focus image fusion with enhancement filtering for robust vascular quantification using photoacoustic microscopy. *Optics Letters*, 47(15): 3732–3735.

Zhou, Z.; Li, S.; and Wang, B. 2014. Multi-scale weighted gradient-based fusion for multi-focus images. *Information Fusion*, 20: 60–72.

Multi-Focus Image Fusion via Explicit Defocus Blur Modelling Supplementary Material

Yuhui Quan[1], Xi Wan[1], Zitao Tang[1], Jinxiu Liang[2]Corresponding author., Hui Ji[3]

[1]South China University of Technology [2]Peking University [3]National University of Singapore

yhquan@scut.edu.cn, csxwan@mail.scut.edu.cn, zitaotang007@gmail.com, cssherryliang@pku.edu.cn, matjh@nus.edu.sg

Related Traditional Methods

Traditional approaches to extending DoF face inherent trade-offs. While smaller apertures can increase DoF, they reduce light gathering capability and introduce noise and diffraction artifacts (Horn 1968). Early methods based on focus stacking required specialized hardware configurations and densely sampled images, leading to increased capture time and computational overhead (Burt and Adelson 1985; Kutulakos and Hasinoff 2009; Levoy et al. 2006; Agarwala et al. 2004; Hasinoff et al. 2009). Recent work has explored optimizing focus distance selection to reduce data requirements (Lee and Tai 2016; Luo et al. 2023b; Zhang 2022b).

Pioneering MFIF methods generally fall into two categories: Transform domain methods decompose source images in alternative feature spaces before fusion. While these approaches achieve smooth transitions near focus boundaries, they often sacrifice sharpness in regions distant from these boundaries (Rockinger 1997; Hill et al. 2002; Li, Kang, and Hu 2013; Zhou, Li, and Wang 2014). Spatial domain methods analyze local features at pixel, block, or region levels to generate focus decision maps (Zhan et al. 2019; Li, Li, and Zhang 2015; Nejati, Samavi, and Shirani 2015; Liu, Liu, and Wang 2015). Though these methods demonstrate strong performance overall, they struggle to handle transitions between focused and defocused regions effectively. Traditional MFIF methods to extend the DoF, such as using a smaller aperture or focus stacking, come with trade-offs. Smaller apertures reduce light gathering ability, introducing noise and diffraction artifacts (Horn 1968). MFIF offers a more sophisticated solution but often relies on densely sampled images, increasing capture time, data storage, and computational costs (Burt and Adelson 1985; Kutulakos and Hasinoff 2009; Levoy et al. 2006; Agarwala et al. 2004; Hasinoff et al. 2009). In the last decades, research explore how to capture fewer images at strategically chosen focus distances in order to reduce data volume and processing overhead (Lee and Tai 2016; Luo et al. 2023b; Zhang 2022b). Traditional MFIF methods are primarily categorized into transform domain and spatial domain approaches. Transform domain methods first transform the source images into a different feature domain, where decomposition takes place. The decomposed coefficients are then merged based on specific fusion criteria, and the fused image is reconstructed using an inverse transform (Rockinger 1997; Hill et al. 2002; Li, Kang, and Hu 2013; Zhou, Li, and Wang 2014). They typically produce smooth and natural transitions near focus-defocus boundaries, though their results tend to lose sharpness in areas farther from these boundaries compared to the original input images. Spatial domain methods, by contrast, rely on spatial features and typically divide the source images into pixels, blocks, or regions to measure activity levels, which are then used to generate decision maps that indicate focus properties (Zhan et al. 2019; Li, Li, and Zhang 2015; Nejati, Samavi, and Shirani 2015; Liu, Liu, and Wang 2015). Although these advanced methods show strong performance, managing the intersection between focused and defocused regions remains a challenge for most spatial domain techniques.

Generalized Formulation

Given S source images $\{\boldsymbol{Y}_s \in \mathbb{R}^{H\times W\times C}\}_{s=1}^S$ captured with different focal planes, our goal is to form an AIF image $\boldsymbol{Z} \in \mathbb{R}^{H\times W\times C}$. For each pixel position (i,j), we have S decision maps $\{\boldsymbol{M}_s(i,j)\}_{s=1}^S$ where $\sum_{s=1}^S \boldsymbol{M}_s(i,j) = 1$ (which can be achieved by using a softmax operation). Firstly, We analyze the confidence of focus decisions using a variance-based uncertainty measure. We use $\boldsymbol{M}^{\mathrm{u}}$ to identify pixels with low-confidence decisions

$$\boldsymbol{M}^{\mathrm{u}}(i,j) = \mathcal{I}(\frac{1}{S}\sum_{s=1}^{S}(\boldsymbol{M}_s(i,j) - \frac{1}{S})^2 \leq \gamma), \tag{23}$$

where γ denotes a variance threshold. For pixels in certain regions ($\boldsymbol{M}^{\mathrm{u}}(i,j) = 0$):

$$\boldsymbol{Z}^*(i,j) = \boldsymbol{Y}_{s^*(i,j)}(i,j) \tag{24}$$

where $s^*(i,j) = \mathrm{argmax}_s\, \boldsymbol{M}_s(i,j)$ selects the source image with highest confidence. For pixels in uncertain regions ($\boldsymbol{M}^{\mathrm{u}}(i,j) = 1$), we use a weighted fusion approach:

$$\boldsymbol{Z}^*(i,j) = \sum_{s=1}^{S} w_s(i,j)\boldsymbol{Z}_s(i,j) \tag{25}$$

where

$$w_s(i,j) = \frac{\exp(\tau \boldsymbol{M}_s(i,j))}{\sum_{k=1}^{S}\exp(\tau \boldsymbol{M}_k(i,j))} \tag{26}$$

Here, τ is a temperature parameter controlling the sharpness of the weight distribution, and $\{\boldsymbol{Z}_s\}_{s=1}^S$ are the initial focused predictions. This formulation naturally reduces to the dual-image case when $S = 2$, where the variance measure simplifies to a function of the difference between two decision maps.

Supplemental Details

Details of SRM

As illustrated in the bottom left of Figure 2 in the main text, the SRM comprises an encoder-decoder backbone CNN for feature extraction and a modified ConvLSTM for scale-recurrent processing.

Encoder-decoder backbone The progressive encoder-decoder backbone CNN consists of an input convolutional block, two encoder blocks, a progressive convolutional block, and two decoder blocks.

- **Input convolutional block:** This block contains a convolutional layer followed by a residual block ("ResBlock" in the bottom middle of Figure 2 of in the main text) with two convolutional layers and residual connections.
- **Encoder/Decoder blocks:** Each block comprises a convolutional layer with downsampling/upsampling, respectively, followed by a residual block. Downsampling is achieved using a convolutional layer with a stride of 2.
- **Progressive convolutional block:** This block gradually increases the number of residual blocks as the scale increases. This design enhances the network's expressive power by exploring both common and unique features across different scales, leading to better predictions. Specifically, at the smallest scale, the block has one convolutional layer. An additional convolutional layer is added for each doubling of the input scale.

Modified ConvLSTM The ConvLSTM is adopted with modifications, containing two branches. One branch mainly consists of a ConvLSTM layer, with downsampling before and upsampling after to introduce local smoothness into the predicted defocus coefficient maps. The ConvLSTM captures dependencies among blur amounts at different scales. Its hidden states progressively refine the estimation of defocus coefficient maps as the scale increases by aggregating information from various scales. The other branch utilizes a single convolutional layer without downsampling or upsampling to preserve fine details. The final defocus coefficient map is obtained by summing the outputs from both paths.

Additional Implementation Details

Our approach is implemented using PyTorch and run on an NVIDIA GTX 1080Ti GPU in an end-to-end manner. We use $T = 3$ scales and $N = 20$ kernel atoms in DBE, $\gamma = 0.1$ in UAF, and $\beta = 0.5$ in the loss function. The model weights are initialized by Xavier. The training employs Adam with 40 epochs, batch size 4, and learning rate 2×10^{-4}.

The summation in $\mathcal{B}_{\dagger}^{-1}$ is implemented using a 1×1 convolution. For color images, the same atom $\boldsymbol{A}_n$ is applied independently to the R, G, B channels.

Details of Compared Methods

Due to the high resolution of images in the SIMIF dataset, GRFusion encounters an "Out of memory" error when run directly, even on a 4090 GPU with 24GB memory. To obtain the reported results, the input image is uniformly divided into four parts, processed separately, and the results are merged into a single image.

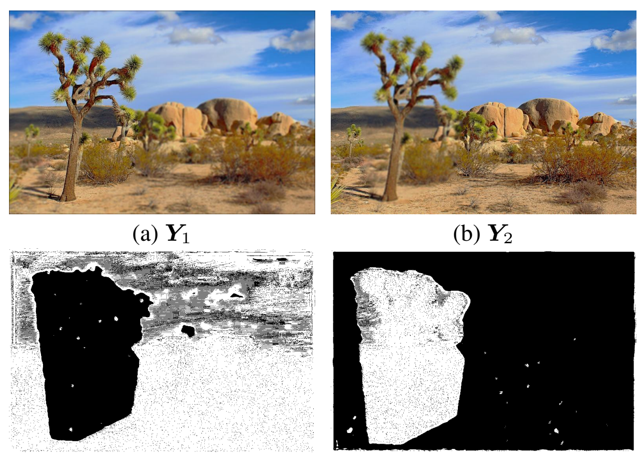
Details of Datasets

The MFI-WHU dataset is a synthetic dataset comprising 120 pairs of out-of-focus images with manually generated decision maps. This dataset provides GT data for evaluation. The Lytro dataset contains 20 image pairs, each of size 520×520 pixels, captured with a Lytro camera. The MFFW dataset consists of 13 image pairs sourced from the Internet and various devices, characterized by bokeh effects. The SIMIF (also called TSAI) dataset includes 12 pairs of high-resolution multi-focus images. The OR-PAM dataset is a biological image dataset containing microvascular images of leaf phantoms, mouse liver, and brain captured using optical-resolution photoacoustic microscopy.

Supplemental Results

Visualization of the Coefficient

Figure 5 illustrates the SRM outputs C_n for an image pair from the MFI-WHU dataset. For sharp pixels, the absolute values of C_n ($n \in 1, \ldots, N$) are relatively small (10^{-3} to 10^{-5}), while blurred pixels exhibit larger absolute values (0.1 to 1). This pattern aligns with DBE's residual learning mechanism, where near-zero C_n values indicate minimal network adjustments. The visualization demonstrates DBE's effectiveness in capturing defocus blur information and enhancing decision map estimation.



(a) $\boldsymbol{Y}_1$ (b) $\boldsymbol{Y}_2$ (c) $\boldsymbol{C}_{n,1}$ (d) $\boldsymbol{C}_{n,2}$

Figure 5: Visualization of SRM coefficients $\boldsymbol{C}_n$. (a) and (b) denote near-focus image $\boldsymbol{Y}_1$ and far-focus image $\boldsymbol{Y}_2$, respectively, while (c) and (d) denote their corresponding SRM output $\boldsymbol{C}_{n,1}$ and $\boldsymbol{C}_{n,2}$. The grayscale images in (c) and (d) represent the sum of absolute $\boldsymbol{C}_n$ values for each pixel. Brighter regions indicate larger absolute coefficient values, corresponding to areas with more significant defocus blur.

Visualization of the PSFs

Figure 6 illustrates the PSFs $\boldsymbol{D}$ for both blurred and clear image points. The PSF of blurred points exhibits radial symmetry with larger spatial extent and higher intensity values, while the PSF of clear points remains concentrated near the

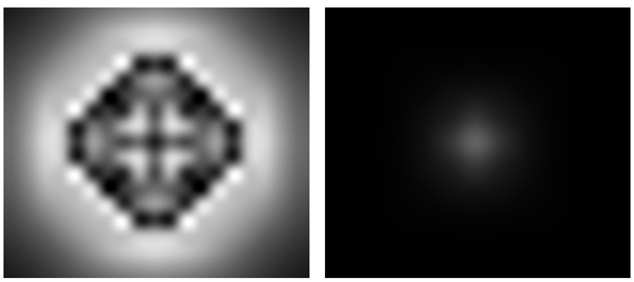
Table 7: Ablation study of dual mask prediction during training on the MFI-WHU dataset.

Method	PSNR(dB)	SSIM	LPIPS	NMI	Q_G	Q_M
Ours	**42.898**	**0.9980**	**0.0042**	**1.2225**	**0.7387**	**2.5281**
w/o M_2	42.347	0.9969	0.1554	1.2219	0.7382	2.5242

Table 8: Quantitative comparison with traditional methods on MFI-WHU dataset.

Method	PSNR(dB)	SSIM	LPIPS	NMI	Q_G	Q_M	MI	Q_Y
DMANet	**42.898**	**0.9980**	**0.0042**	**1.2225**	**0.7387**	**2.5281**	**8.9069**	**0.9895**
DSIFT	**39.396**	**0.9962**	**0.0057**	**1.2208**	**0.7398**	**2.5434**	**8.8962**	**0.9887**
MFF	27.716	0.9539	0.1770	0.7373	0.6114	0.3321	5.3697	0.9152
MWGF	38.700	0.9904	0.1603	1.1489	0.7277	2.3029	8.3713	0.9862
GFF	36.018	0.9951	0.0108	1.1213	0.7250	1.9123	8.1736	0.9759

center with minimal spread. For visualization clarity, we applied uniform scaling to both PSF representations.



(a) Blur point (b) Clear point

Figure 6: Visualization of PSFs $\boldsymbol{D}$ computed as weighted (C_n) combinations of Gaussian kernels (A_n) for (a) a blurred point and (b) clear point, respectively.

Ablation Studies

To analyze the effectiveness of UAF, we decrease its threshold γ from 0.1 to 0. Note that when $\gamma = 0$, the fusion becomes using binary masks without uncertainty awareness The corresponding PSNR results on the MFI-WHU dataset are shown in Table 9. As γ decays, the PSNR performance decreases, demonstrating the effectiveness of UAF.

Table 9: PSNR result w.r.t. varied γ on MFI-WHU dataset.

γ	0.1	0.08	0.06	0.04	0.02	0
PSNR(dB)	**42.898**	42.885	42.863	42.833	42.798	42.781

Table 7 evaluates the effectiveness of dual mask prediction ($\boldsymbol{M}_1$ and $\boldsymbol{M}_2$) during training, while maintaining $\boldsymbol{Z}_1^*$ as the final prediction target. The results show that incorporating M_2 prediction yields a significant 0.55 dB improvement in PSNR. This enhancement can be attributed to the network's improved discrimination between defocused and focused regions when leveraging both input and output information. For computational efficiency, only M_1 is utilized during the testing phase.

Table 10: Effect of PSF atom parameter N on PSNR performance using the MFI-WHU dataset.

N	14	20	26
Param.(M)	4.026	4.065	4.137
PSNR(dB)	42.219	42.898	42.904

Table 10 analyzes the impact of PSF atom parameter N on model performance. While increasing N shows marginal improvements in performance, with peak results at $N = 26$, we selected $N = 20$ as the optimal value to maintain an effective balance between computational complexity and fusion quality.

Figure 7 showcases results from our proposed method with and without uncertainty awareness (UA) in the UAF module, alongside the estimated uncertainty map. Incorporating UA improves the quality of fusion, particularly in challenging boundary regions where focused and defocused areas are difficult to distinguish. For instance, observe the left contour of the brown desk: the method with UA produces a sharper and more accurate boundary compared to the version without UA. This highlights how uncertainty-aware fusion contributes to preserving fine details and enhancing the overall visual fidelity of the fused images.

Quantitative Comparisons with Traditional Methods

Table 8 presents the quantitative results on the MFI-WHU dataset. Our method outperforms all evaluated traditional approaches, including both transform domain methods (MWGF and GFF) and spatial domain methods (DSIFT and MFF). These results demonstrate the effectiveness of our approach, which achieves superior performance compared to both traditional and deep learning-based methods. Specifically, our method shows improvements in terms of both structural preservation and detail retention, as evidenced by the higher scores across all evaluation metrics.

Source #1 Source #2 GT Uncertainty map w/o UA w/ UA (Ours)

Figure 7: Visual comparison to demonstrate the impact of uncertainty awareness (UA) in our UAF module. It showcases results from our proposed method with and without UA, alongside the estimated uncertainty map in the fourth column.

Visual Comparisons

Figures 8 and 9 present additional visual comparison results, showcasing our method's performance against four supervised MFIF techniques: IFCNN, GACN, FusionDiff, and DB-MFIF. As shown in Fig. 8, both our method and GACN successfully identify the focused regions, such as the closer flowers in the first example and the background in the second example. However, our method demonstrates higher accuracy, evident in the larger correctly identified regions within the difference maps compared to the GT. Fig. 9 illustrates results across a diverse set of datasets. Notably, our method consistently and accurately identifies the focused regions in Source #1 across all examples, as seen in the postcard (Row#1), the back of the doll (Row#2), the red flowers (Row#3), the cup with blue covers in the front (Row#4), and the textured wall (Row#5). While GACN partially identifies the focused regions in Rows#2 and #4, it falls short of the consistent accuracy achieved by our method. These results further emphasize the robustness and superior performance of our approach across different challenging scenarios.

Limitations

Despite the promising results achieved, our approach cannot consistently perform the best across all GT-independent metrics. This is indeed a limitation of existing methods. It comes from that, without utilizing GTs, those metrics measure MFIF accuracy in diverse aspects. As a result, it is challenging to strike a balance among these metrics. We will investigate effective approaches to win the trade-off.

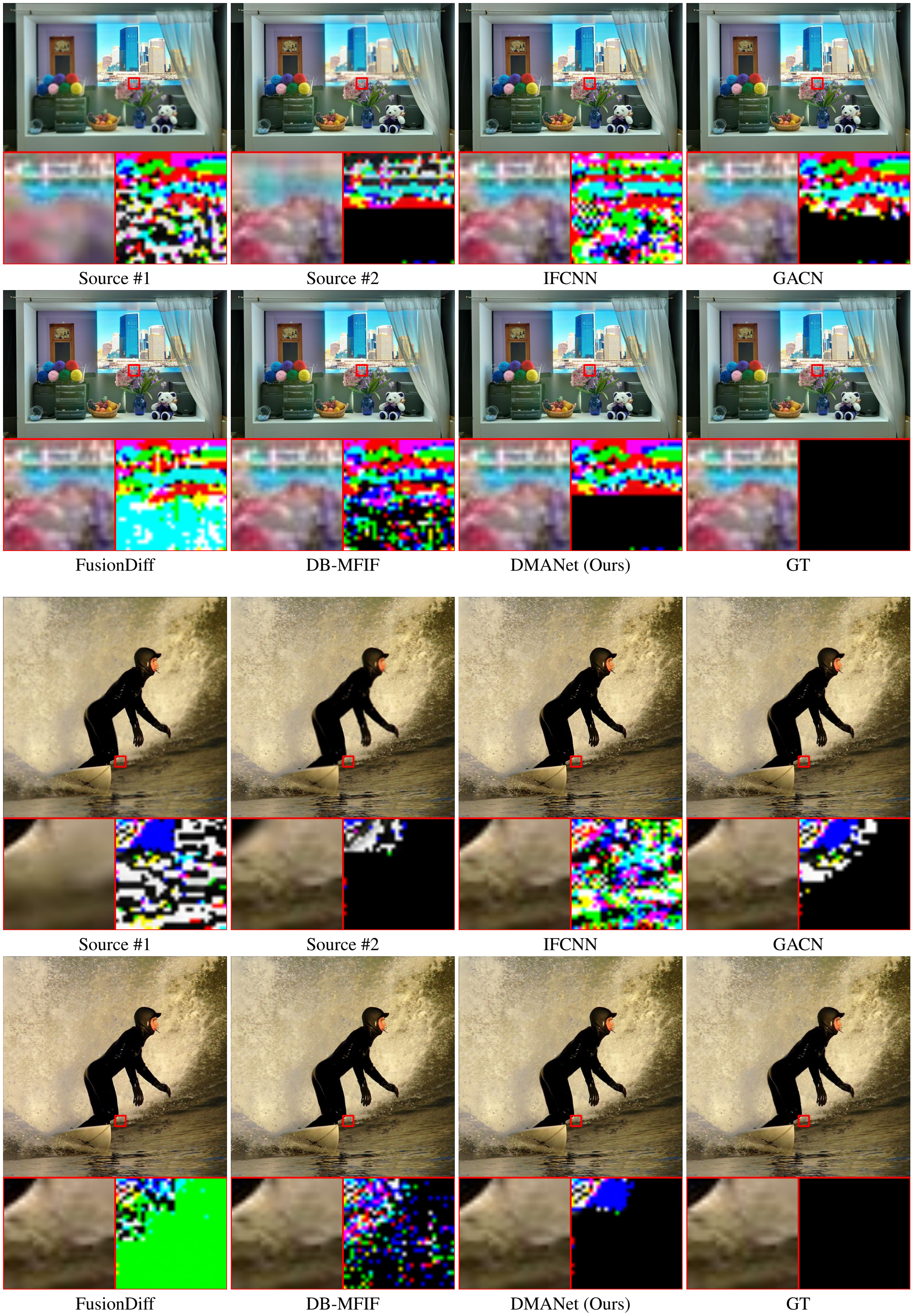


Figure 8: Fused images by different methods on two samples from the MFI-WHU dataset. Behind each image, a zoomed region and its difference from GT are provided.

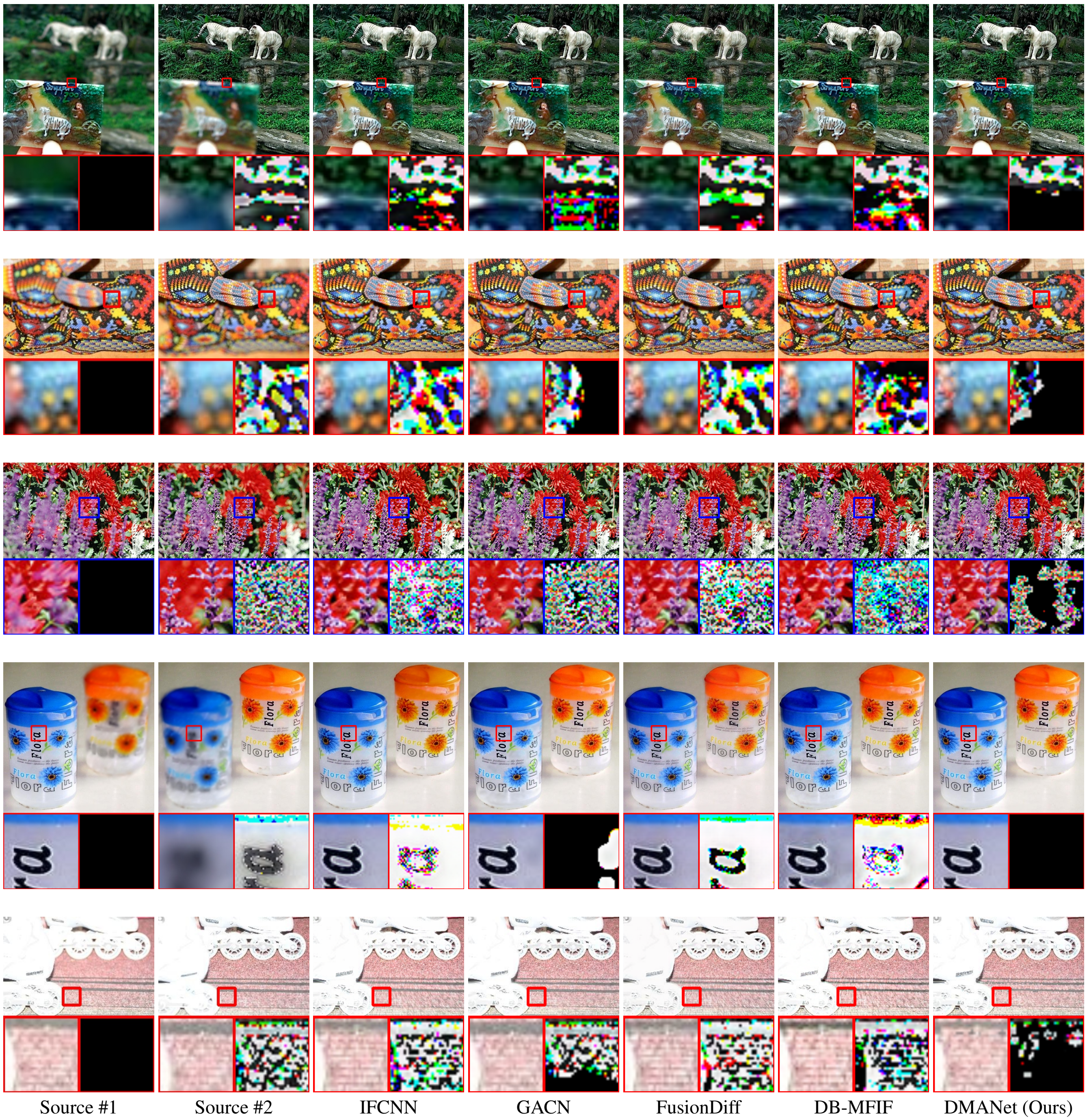


Figure 9: Fused images by different methods on the samples from the Lytro (Row#1), MFFW (Row#2-#3), and SIMIF (Row#3-#4) datasets. Behind each image, a zoomed region and its difference from Source #1 are provided.

Reproducibility Checklist

Unless specified otherwise, please answer "yes" to each question if the relevant information is described either in the paper itself or in a technical appendix with an explicit reference from the main paper. If you wish to explain an answer further, please do so in a section titled "Reproducibility Checklist" at the end of the technical appendix.
This paper:

- Includes a conceptual outline and/or pseudocode description of AI methods introduced (yes)
- Clearly delineates statements that are opinions, hypothesis, and speculation from objective facts and results (yes)
- Provides well marked pedagogical references for less-familiare readers to gain background necessary to replicate the paper (yes)

Does this paper make theoretical contributions? (no)
If yes, please complete the list below.

- All assumptions and restrictions are stated clearly and formally. (yes/partial/no)
- All novel claims are stated formally (e.g., in theorem statements). (yes/partial/no)
- Proofs of all novel claims are included. (yes/partial/no)
- Proof sketches or intuitions are given for complex and/or novel results. (yes/partial/no)
- Appropriate citations to theoretical tools used are given. (yes/partial/no)
- All theoretical claims are demonstrated empirically to hold. (yes/partial/no/NA)
- All experimental code used to eliminate or disprove claims is included. (yes/no/NA)

Does this paper rely on one or more datasets? (yes)
If yes, please complete the list below.

- A motivation is given for why the experiments are conducted on the selected datasets (yes)
- All novel datasets introduced in this paper are included in a data appendix. (NA)
- All novel datasets introduced in this paper will be made publicly available upon publication of the paper with a license that allows free usage for research purposes. (NA)
- All datasets drawn from the existing literature (potentially including authors' own previously published work) are accompanied by appropriate citations. (yes)
- All datasets drawn from the existing literature (potentially including authors' own previously published work) are publicly available. (yes)
- All datasets that are not publicly available are described in detail, with explanation why publicly available alternatives are not scientifically satisficing. (NA)

Does this paper include computational experiments? (yes)
If yes, please complete the list below.

- Any code required for pre-processing data is included in the appendix. (yes).
- All source code required for conducting and analyzing the experiments is included in a code appendix. (yes)
- All source code required for conducting and analyzing the experiments will be made publicly available upon publication of the paper with a license that allows free usage for research purposes. (yes)
- All source code implementing new methods have comments detailing the implementation, with references to the paper where each step comes from (yes)
- If an algorithm depends on randomness, then the method used for setting seeds is described in a way sufficient to allow replication of results. (yes)
- This paper specifies the computing infrastructure used for running experiments (hardware and software), including GPU/CPU models; amount of memory; operating system; names and versions of relevant software libraries and frameworks. (yes)
- This paper formally describes evaluation metrics used and explains the motivation for choosing these metrics. (yes)
- This paper states the number of algorithm runs used to compute each reported result. (yes)
- Analysis of experiments goes beyond single-dimensional summaries of performance (e.g., average; median) to include measures of variation, confidence, or other distributional information. (yes)
- The significance of any improvement or decrease in performance is judged using appropriate statistical tests (e.g., Wilcoxon signed-rank). (yes)
- This paper lists all final (hyper-)parameters used for each model/algorithm in the paper's experiments. (yes)
- This paper states the number and range of values tried per (hyper-) parameter during development of the paper, along with the criterion used for selecting the final parameter setting. (NA)