UHD-MFF: Shattering Barriers in Multi-Focus Ultra-High-Definition Image Fusion via Learnable Lookup Tables

Yibing Zhang[1]*, Xunpeng Yi[1]*, Qinglong Yan[1], Yeda Wang[1], Han Xu[3], and Jiayi Ma[1,2]**

[1] Electronic Information School, Wuhan University, Wuhan 430072, China
[2] School of Robotics, Wuhan University, Wuhan 430072, China
[3] School of Automation, Southeast University, Nanjing 210096, China
{zhangyibing,yixunpeng,qinglong_yan,wangyeda}@whu.edu.cn
xu_han@seu.edu.cn, jyma2010@gmail.com

Abstract. With the advancement of imaging technology, ultra-high-definition images have become increasingly essential in modern visual applications. However, existing multi-focus image fusion remains largely confined to low-resolution images and faces three major barriers in UHD scenarios, namely data availability, model adaptability, and deployment feasibility, which severely hinder its practical application. To shatter these barriers, first, we propose the UHD-MFF dataset, the first large-scale ultra-high-resolution multi-focus fusion dataset. Second, we propose a scale-specialized lookup-table framework tailored for ultra-high-resolution images, termed as UMF-LUT. It consists of Coarse-Region Lookup Table (C-LUT) and Detail-Edge Lookup Table (D-LUT). Specifically, C-LUT performs joint queries of multiple gradient cues and semantic cues at low-resolution scales to enable region-level decision-making. Also, D-LUT operates at high-resolution scales, leveraging efficient Laplacian cues to provide complementary edge-level decision information. Such a design makes the model particularly well-suited for ultra-high-resolution multi-focus image fusion. Finally, it offers strong deployability with minimal computational overhead, enabling real-time 4K multi-focus fusion and showing promising potential for smartphone. Extensive experiments demonstrate that it outperforms SOTA methods in both visual fidelity and quantitative metrics. It effectively advances the development of multi-focus image fusion toward ultra-high-resolution imaging scenarios. The code is available at `https://github.com/zyb5/UHD-MFF`.

Keywords: Image Fusion · UHD Image · Multi-Focus Imaging

1 Introduction

Due to the limited depth of field of optical lenses, a single image typically captures only part of a scene in focus, leaving other regions blurred and resulting

* Equal contribution.
** Corresponding author.

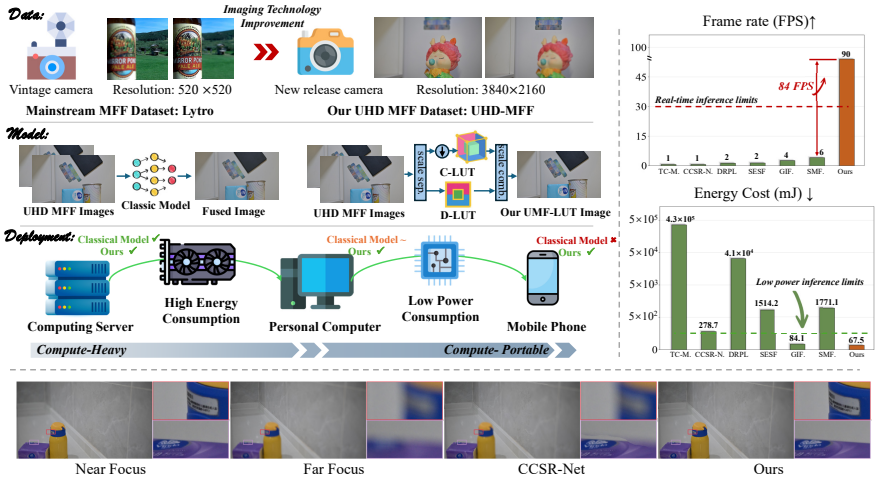

Fig. 1: Three primary barriers in ultra-high-resolution multi-focus image fusion. Our method is compared to state-of-the-art methods in terms of inference frame rate, energy cost, and visual quality.

in information loss. In this background, Multi-Focus Image Fusion (MFIF) has been developed as an effective solution [7]. As an important branch of image fusion, MFIF combines multiple images captured at different focal lengths to generate a single all-in-focus image, thereby enhancing scene clarity and preserving more detail [5, 10]. Owing to its ability to handle complex scenes with varying focal distances, MFIF has been widely applied in areas such as macro digital photography.

MFIF has achieved substantial development in recent years [2,4,20]. However, with the rapid advancement of imaging devices, ultra-high-resolution images have become the prevailing trend. Under this circumstance, the development of multi-focus image fusion faces new bottlenecks. As shown in Fig. 1, we summarize these challenges into three major barriers: **1) Data availability:** Existing mainstream datasets, such as Lytro and Real-MFF, are all limited to resolutions below 1080P. Such relatively low-resolution data are insufficient to support research on ultra-high-resolution image fusion. Moreover, models trained on low-resolution datasets often exhibit suboptimal performance when generalized to ultra-high-resolution scenarios. **2) Model adaptability:** Due to the lack of explicit consideration of ultra-high-resolution image scales, previous methods have often relied on conventional convolutional operations or attention mechanisms in a straightforward manner, without specialized adaptation for UHD images. **3) Deployment feasibility:** The demand for multi-focus image fusion on mobile platforms is substantial. However, existing methods generally do not support deployment on smartphones and, in some cases, are not even feasible for personal computers [30, 31]. Instead, they rely on computation-intensive servers for execution, which severely restricts their practical applicability. These three barriers

collectively impose significant constraints on the advancement of MFIF, hindering its ability to meet the demands of next-generation ultra-high-resolution imaging.

To address these issues, in terms of data availability, to support the advancement of ultra-high-definition multi-focus image fusion and fill the existing data gap in this field, we propose the UHD-MFF dataset, which comprises 1,950 pairs of near- and far-focus images at the resolution of 3840×2160, encompassing a diverse range of indoor and outdoor scenes with varying degrees of focus. In terms of model adaptability, we propose the scale-specialized lookup-table multi-focus image fusion for ultra-high-resolution images, termed as UMF-LUT. It customizes the large-scale challenge of ultra-high-resolution multi-focus images by performing region-wise decisions at low-resolution scales and edge refinement at high-resolution scales. Firstly, at the low-resolution scales, we utilize multiple gradient cues, including gradient difference cues, dilated gradient cues, maximum gradient cues, and learnable regional semantic cues, as lookup elements in the coarse-region lookup table (C-LUT) to perform the region-wise decisions. Later, as a complement, at the high-resolution scales, we employ the Laplacian operator as a residual decision cue in the lightweight detail-edge lookup table (D-LUT) to achieve edge-level decision-making. This design not only ensures clear focus boundaries but also cleverly avoids the substantial memory overhead associated with full-resolution convolutions. It achieves strong multi-focus fusion performance through a scale-specific, customized combination of operations, making it particularly well-suited for high-resolution images. From the perspective of deployment feasibility, UMF-LUT incurs very low computational overhead, enabling real-time fusion of 4K-resolution UHD images on commercial GPUs and demonstrating potential for deployment on mobile smartphones.

Overall, our contributions can be summarized as follows:

- **Data availability:** We propose UHD-MFF, the first large-scale ultra-high-resolution multi-focus image fusion dataset comprising 1,950 diverse image pairs. This fills the critical gap in UHD benchmarks and provides a robust foundation for future data-driven research.
- **Model adaptability:** We introduce UMF-LUT, a scale-decoupled architecture tailored for UHD images. By combining low-resolution regional decisions with native-resolution edge refinement, it ensures accurate focus perception while bypassing the memory bottlenecks of full-resolution convolutions.
- **Deployment feasibility:** Empowered by its special LUT design, UMF-LUT achieves real-time UHD inference with minimal computational overhead. Furthermore, we successfully demonstrate its deployment on mobile smartphones, breaking the hardware barrier of existing methods.

2 Related Work

Traditional Methods. Generally, traditional multi-focus image fusion algorithms can be categorized into two streams [2, 13]: spatial domain-based meth-

ods and transform domain-based methods. Spatial domain methods directly manipulate pixels within source images to generate weight maps, which are then employed to fuse the pixels. In contrast, transform domain methods first transform source images into the frequency domain using specific decomposition rules. The decomposed coefficients are fused based on certain strategies, followed by an inverse transformation to reconstruct the final image. The quadtree-based method [1] utilizes quadtree decomposition and weighted focus measures, which adaptively partitions the image into blocks of varying sizes to precisely extract focused regions. DSIFT [11] innovatively introduces dense SIFT descriptors into the MFIF domain, employing them as activity measures for image patches to generate initial decision maps, which are further refined through feature matching. GFDF [21] leverages focus region detection and the edge-preserving properties of guided filtering to optimize coarse initial decision maps and correct boundary artifacts. Furthermore, the HOSVD and edge-intensity method [14] models higher-order patch correlations. The JSR-based method [18] introduces a fusion framework based on Joint Sparse Representation and optimization theory, decomposing source images into a common component containing redundant information and a component containing focused details.

Deep Learning-based Methods. Deep learning has significantly advanced MFIF [26, 29, 33], with existing methods primarily categorized into regression-based [35] and classification-based approaches [23, 25]. Addressing the lack of ground truth, SMFuse [16] utilizes focus measures as auxiliary signals for self-supervised pixel-level focus detection via iterative optimization. MFF-GAN [32] employs unsupervised adversarial training with adaptive gradient constraints to capture textures while preserving source intensities. To enhance feature representation, MSFIN [12] introduces a multi-scale interaction network to facilitate bidirectional complementarity between deep semantics and shallow details. For generalization, ZMFF [4] adopts a zero-shot framework optimizing loss functions directly from inputs to eliminate train-test distribution shifts. Bridging model- and learning-based methods, CCSR-Net [36] unrolls the iterative solution of coupled convolutional sparse representation into a neural network, combining theoretical interpretability with efficient inference. Several unified frameworks also incorporate the MFIF task. U2Fusion [27] leverages adaptive information preservation metrics to automatically estimate pixel importance and retain critical structures. More recently, TC-MoA [37] introduces a task-customized Mixture of Adapters for general image fusion.

3 Our Approach

3.1 Scale-Specialized Lookup-Table

Coarse-Region Lookup Table (C-LUT): To expand the receptive field while mitigating computational overhead across UHD scales, spatial compression is applied to the Coarse-Region Lookup Table in UMF-LUT, as shown in Fig. 2. Given the source image pair $I_{near}, I_{far} \in \mathbb{R}^{H\times W\times C}$, a down-sampling operator

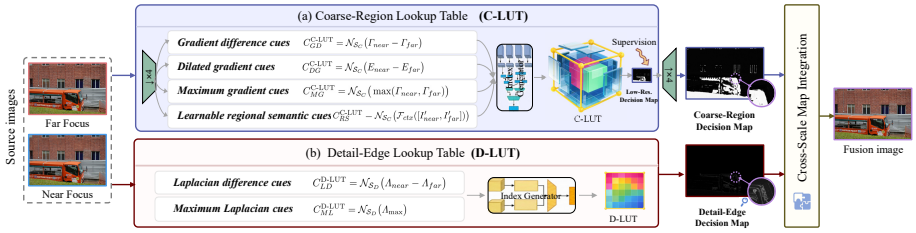

Fig. 2: Overall Structure of the proposed UMF-LUT.

$\mathcal{D}_{\downarrow s}(\cdot)$ with a scale factor s projects the images into a low-resolution space:

$$I'_k = \mathcal{D}_{\downarrow s}(I_k), \quad k \in \{near, far\}. \tag{1}$$

The primary objective of this coarse-scale operation is to establish a reliable regional consensus regarding focus attributes. To ensure computational efficiency, this complex decision-making process is mapped into a lookup table constructed from informative cues. Because focus levels are intrinsically correlated with regional texture variations, first-order Sobel operators are employed in the C-LUT to extract base gradient cues:

$$\Gamma_k = |\mathcal{H}_x * I'_k| + |\mathcal{H}_y * I'_k|, \quad k \in \{near, far\}, \tag{2}$$

where $\mathcal{H}_x$ and $\mathcal{H}_y$ represent the horizontal and vertical Sobel operator kernels, and $*$ denotes the convolution operation.

To strictly map these continuous cues into the discrete index space of the C-LUT while maximizing the utilization of the lookup capacity, we introduce a unified spatial normalization operator $\mathcal{N}_{\mathcal{S}_C}(\cdot)$. This operator applies a nonlinear activation followed by a scale-and-shift transformation, bounding any continuous input within the grid range $[0, \mathcal{S}_C - 1]$. The resulting index group $\mathbf{c} = \{c_1, c_2, c_3, c_4\}$ is formulated as follows:

1) Gradient difference cues: This cue measures the local sharpness difference between the two images and directly reflects which image exhibits stronger structural responses:

$$C_{GD}^{\text{C-LUT}} = \mathcal{N}_{\mathcal{S}_C}\big(\Gamma_{near} - \Gamma_{far}\big). \tag{3}$$

2) Dilated gradient cues: Since pixel-wise gradients are sensitive to noise, we compute the regional energy $E_k(x, y) = \frac{1}{|\Omega|} \sum_{(u,v)\in\Omega(x,y)} \Gamma_k(u, v)$ to incorporate neighborhood statistics and enforce spatial consistency:

$$C_{DG}^{\text{C-LUT}} = \mathcal{N}_{\mathcal{S}_C}\big(E_{near} - E_{far}\big). \tag{4}$$

3) Maximum gradient cues: This cue evaluates the overall gradient magnitude across the two images, capturing local texture strength and reducing ambiguity in low-texture regions:

$$C_{MG}^{\text{C-LUT}} = \mathcal{N}_{\mathcal{S}_C}\big(\max(\Gamma_{near}, \Gamma_{far})\big). \tag{5}$$

4) Learnable regional semantic cues: We employ a lightweight CNN, denoted as $\mathcal{F}_{ctx}$, to extract joint contextual representations from the concatenated images, complementing gradient-based cues with region-level semantic information:

$$C_{RS}^{\text{C-LUT}} = \mathcal{N}_{\mathcal{S}_C}\big(\mathcal{F}_{ctx}([I'_{near}, I'_{far}])\big). \tag{6}$$

To optimize the efficiency of system integration, we implement a strategy that transforms large-scale computational tasks into offline look-up table operations. Building upon the establishment of the four cues, we utilize a four-dimensional LUT, denoted as $\mathcal{L}_{\mathcal{C}}$, to map the continuous coordinate space $\mathbf{c} = \{c_1, c_2, c_3, c_4\} = \{C_{GD}^{\text{C-LUT}}, C_{DG}^{\text{C-LUT}}, C_{MG}^{\text{C-LUT}}, C_{RS}^{\text{C-LUT}}\}$ to the fusion weight:

$$\mathcal{M}_{main}^{LUT}(x, y) = \Phi_{\text{Main-LUT}}(\mathbf{c}), \tag{7}$$

where $\Phi_{\text{Main-LUT}}$ represents the discrete look-up and interpolation operation. The LUT stores fusion weights in a four-dimensional lattice.

Since the calculated cues $\mathbf{c}$ are continuous floating-point values while the LUT grid is discrete, direct indexing is infeasible. Therefore, each coordinate c_i ($i \in \{1, 2, 3, 4\}$) is decomposed into an integer grid index $n_i = \lfloor c_i \rfloor$ and a fractional residual $\alpha_i = c_i - n_i$, where $\lfloor \cdot \rfloor$ denotes the floor function.

To guarantee differentiability for gradient back-propagation and ensure smooth weight transitions, we employ Quadrilinear Interpolation. The output at an arbitrary query point is computed by aggregating the values from the $2^4 = 16$ surrounding vertices of the activated hypercube:

$$\begin{aligned} \Phi_{\text{Main-LUT}}(\mathbf{c}) &= \sum_{\delta_1=0}^{1}\sum_{\delta_2=0}^{1}\sum_{\delta_3=0}^{1}\sum_{\delta_4=0}^{1} \mathcal{W}_{\boldsymbol{\delta}} \cdot \mathcal{L}_{\mathcal{C}}(n'_1, n'_2, n'_3, n'_4), \\ n'_a &= n_a + \delta_a, \quad a \in \{1, 2, 3, 4\}, \end{aligned} \tag{8}$$

where $\boldsymbol{\delta} = (\delta_1, \delta_2, \delta_3, \delta_4)$ represents the binary vertex offset. The interpolation weight $\mathcal{W}_{\boldsymbol{\delta}}$ is derived from the product of linear distances along each dimension:

$$\mathcal{W}_{\boldsymbol{\delta}} = \prod_{j=1}^{4} \left((1 - \alpha_j)^{1-\delta_j} \cdot \alpha_j^{\delta_j} \right). \tag{9}$$

Detail-Edge Lookup Table (D-LUT): The C-LUT captures semantic consistency and performs large-scale regional decisions. However, due to the limited resolution at the coarse scale, its capability to delineate fine-grained focus boundaries is inherently constrained. As a complement, the D-LUT is introduced to function as a high-resolution refiner. Crucially, while the C-LUT relies on first-order Sobel operators to capture broad structural trends and regional energy, the D-LUT utilizes the second-order Laplacian operator. Second-order responses emphasize rapid intensity transitions and zero-crossing structures, making them particularly suitable for boundary localization that first-order gradients often over-smooth. By employing the Laplacian convolution kernel $\mathcal{H}_{Lap}$, we directly extract these high-frequency structural responses Λ_{near} and Λ_{far}:

$$\Lambda_k = |\mathcal{H}_{Lap} * I_k|, \quad k \in \{near, far\}. \tag{10}$$

Using a corresponding normalization operator $\mathcal{N}_{\mathcal{S}_D}(\cdot)$ adapted to the D-LUT bin size $\mathcal{S}_D$, we formulate the coordinate indices $\mathbf{d} = \{d_1, d_2\}$:

1) Laplacian difference cues: This cue computes the pixel-wise difference between second-order Laplacian responses of the two images, directly capturing high-frequency focus transitions and refining boundaries weakened during coarse-scale down-sampling:

$$C_{LD}^{\text{D-LUT}} = \mathcal{N}_{\mathcal{S}_D}\big(\Lambda_{near} - \Lambda_{far}\big). \tag{11}$$

2) Maximum Laplacian cues: This cue evaluates the maximum Laplacian magnitude across the two images, estimating local structural strength and suppressing unstable responses in flat or texture-scarce regions:

$$C_{ML}^{\text{D-LUT}} = \mathcal{N}_{\mathcal{S}_D}\big(\Lambda_{\max}\big), \quad \Lambda_{\max} = \max(\Lambda_{near}, \Lambda_{far}). \tag{12}$$

Building upon these two cues, a two-dimensional LUT, denoted as $\mathcal{L}_D$, maps the continuous coordinate space $\mathbf{d} = \{C_{LD}^{\text{D-LUT}}, C_{ML}^{\text{D-LUT}}\}$ to a high-frequency decision residual:

$$\Delta M = \Phi_{\text{Detail}}(\mathbf{d}). \tag{13}$$

To maintain mathematical rigor, we define a boundary truncation operator $\mathcal{B}_{[a,b]}(z) = \max(a, \min(z, b))$ and decompose the continuous coordinates d_k ($k \in \{1, 2\}$) into integer indices $m_k = \lfloor d_k \rfloor$ and fractional residuals $\beta_k = d_k - m_k$. Given $\delta_k \in \{0, 1\}$ denoting the binary vertex offsets. We determine the valid grid indices for the nearest neighbors as $m'_k = \mathcal{B}_{[0,\mathcal{S}_D-1]}(m_k + \delta_k)$, which explicitly confines the look-up operations within the LUT dimensions. To enable end-to-end training, we employ bilinear interpolation, aggregating the values of these 4 nearest neighbors in the two-dimensional lattice:

$$\Phi_{\text{Detail}}(\mathbf{d}) = \sum_{\delta_1=0}^{1}\sum_{\delta_2=0}^{1} \mathcal{W}_{\boldsymbol{\delta}} \cdot \mathcal{L}_D(m'_1, m'_2), \quad \mathcal{W}_{\boldsymbol{\delta}} = \prod_{j=1}^{2}\Big((1-\beta_j)^{1-\delta_j} \cdot \beta_j^{\delta_j}\Big). \tag{14}$$

Cross-Scale Map Integration. Through the explicit scale-decoupling of UMF-LUT, the final decision map M_{final} is effortlessly obtained by integrating the semantically accurate coarse-region decision map from the C-LUT with the detail-edge decision map from D-LUT:

$$M_{final} = \mathcal{B}_{[0,1]}\big(\mathcal{D}_{\uparrow s}(\mathcal{M}_{main}^{LUT}) + \Delta M\big), \tag{15}$$

where $\mathcal{D}_{\uparrow s}(\cdot)$ represents the bilinear up-sampling operator. The fused image I_f is then generated via pixel-wise weighted aggregation:

$$I_f = M_{final} \cdot I_{near} + (1 - M_{final}) \cdot I_{far}. \tag{16}$$

3.2 Loss Functions

To effectively optimize the scale-decoupled UMF-LUT architecture, we formulate specific learning objectives tailored to the distinct roles of the C-LUT and D-LUT. The respective loss functions are defined as:

$$\mathcal{L}_{\text{C-LUT}} = \mathcal{L}_{rec} + \lambda_1 \mathcal{L}_{bin}, \quad \mathcal{L}_{\text{D-LUT}} = \mathcal{L}_{rec} + \lambda_2 \mathcal{L}_{sparse}, \tag{17}$$

where $\mathcal{L}_{rec} = \lambda_3 \mathcal{L}_{int} + \lambda_4 \mathcal{L}_{grad} + \lambda_5 \mathcal{L}_{ssim}$ denotes the fundamental unsupervised reconstruction loss, comprising standard intensity, gradient, and structural similarity constraints. $\lambda_1 \sim \lambda_5$ are balancing hyperparameters.

To guide the network in achieving optimal scale-decoupling, we introduce two specialized regularization terms:

Binarization Loss. To prevent the C-LUT from outputting ambiguous intermediate values in flat regions and to suppress ghosting artifacts, this term forces the coarse decision map to make decisive selections:

$$\mathcal{L}_{bin} = \mathbb{E}\big[\mathcal{M}_{\text{coarse}} \odot (1 - \mathcal{M}_{\text{coarse}})\big], \tag{18}$$

where $\odot$ denotes element-wise multiplication, and $\mathbb{E}[\cdot]$ calculates the spatial expectation.

Sparsity Loss. To prevent the D-LUT from hallucinating pseudo-textures or amplifying noise, this regularization ensures that the high-frequency residual mask ΔM remains zero-dominant, activating exclusively at significant focal boundaries. We constrain this using the ℓ_1-norm:

$$\mathcal{L}_{sparse} = \|\Delta M\|_1. \tag{19}$$

4 UHD-MFF Dataset

To overcome the data availability barrier, we establish the first ultra-high-definition multi-focus fusion dataset, termed as UHD-MFF, which contains 1,950 pairs of UHD images with diverse and complex scenes. As shown in Tab. 1, UHD-MFF dataset demonstrates strong competitiveness compared with existing mainstream datasets in terms of both resolution and scale. Representative scenes are illustrated in Fig. 3. It consists of UHD-MFF-Real and UHD-MFF-Syn subsets.

UHD-MFF-Real. To validate generalization in practical scenarios, we meticulously captured 150 real-world image pairs at 3840×2160 resolution. Strict pixel-level alignment was ensured during acquisition. Comprehensive capture protocols are deferred to the Supplementary Material.

UHD-MFF-Syn. Utilizing high-quality all-in-focus images from the UHD-LL dataset [6], we adopted a depth-guided variable blur strategy to simulate realistic defocus effects. The simulation process is illustrated in Fig. 4. After quality filtering, the 1,800 synthetic pairs were randomly divided into 1,500 for training and 300 for testing.

Table 1: Comparison of our proposed UHD-MFF with existing multi-focus image fusion datasets. (Resolution corresponds to the average of the dataset, if not constant.)

Dataset	Source	Resolution	Pairs	GT
Classic [13]	Real	512×512	10	✗
Lytro [19]	Real	520×520	20	✗
MFFW2 [28]	Synthetic	600×400	13	✓
MFI-WHU [32]	Synthetic	600×400	120	✓
Real-MFF [34]	Real	625×433	710	✗
UHD-MFF-Real	**Real**	**3840×2160**	**150**	✗
UHD-MFF-Syn	**Synthetic**	**3840×2160**	**1,800**	✓

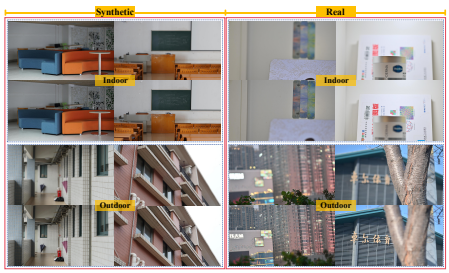

Fig. 3: Samples from UHD-MFF.

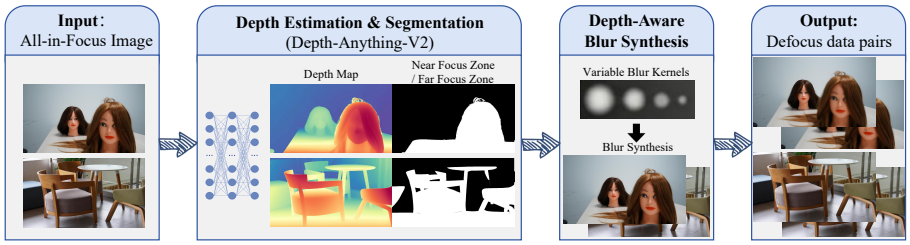

Fig. 4: UHD-MFF-Syn dataset synthesis pipeline.

5 Experiment

5.1 Experiment Settings

Implementation Details. C-LUT and D-LUT are jointly optimized within a unified training framework using the AdamW optimizer. The batch size is set to 8, and the initial learning rate is fixed at 1×10^{-5}. The loss function is formulated as a weighted combination of multiple terms, with coefficients $\lambda_1 = 50$, $\lambda_2 = 0.05$, $\lambda_3 = 20$, $\lambda_4 = 0.7$, and $\lambda_5 = 0.8$. Experiments are conducted in PyTorch on an NVIDIA RTX 3090, with mobile deployment tested on a Huawei P60 Pro.

Datasets. We evaluate our method on our proposed UHD-MFF dataset. To assess zero-shot generalization, the model trained exclusively on UHD-MFF-Syn is directly evaluated on UHD-MFF-Real dataset.

Metrics. We evaluate the performance of the image fusion models using several widely adopted objective metrics, including standard deviation (SD), sum of correlation differences (SCD), mutual information (MI) [22], information entropy (EN) [24], visual information fidelity (VIF), quality of gradient-based fusion ($Q^{AB/F}$) [17], average gradient (AG), and spatial frequency (SF). Higher values of all these metrics indicate higher quality of the fusion image.

SOTA Competitors. To validate the effectiveness of the proposed approach, we performed comprehensive comparisons with multiple state-of-the-art image

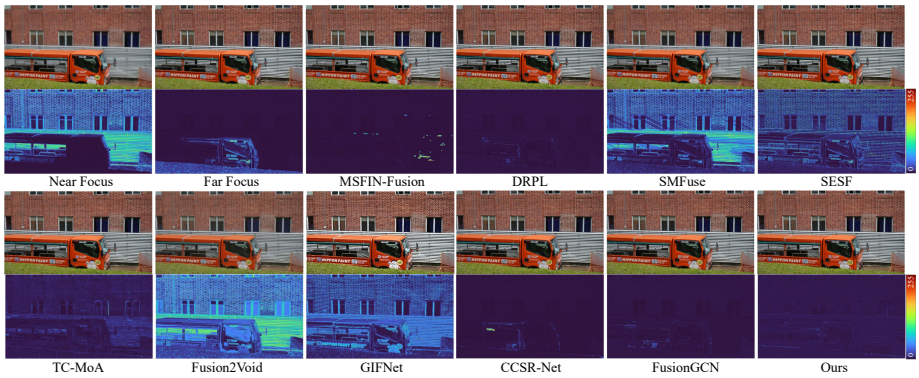

Fig. 5: Qualitative comparison on UHD-MFF-Syn.

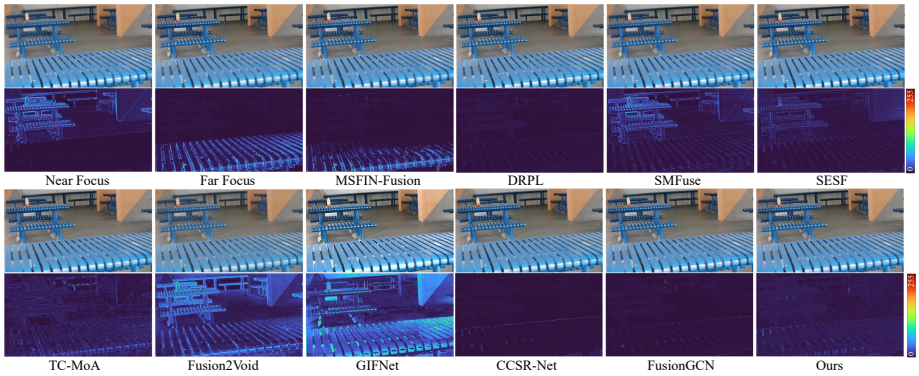

Fig. 6: Qualitative comparison on UHD-MFF-Syn.

fusion methods on multiple datasets. The compared methods include task specific fusion approaches, e.g. SMFuse [16], SESF [15], DRPL [8], MSFIN-Fusion [12], Fusion2Void [9], CCSR-Net [36], FusionGCN [20] and unified image fusion frameworks, e.g. GIFNet [3], TC-MoA [37].

5.2 Qualitative Experiments

We qualitatively compare our method with nine state-of-the-art approaches across UHD-MFF datasets. The visual results are illustrated in Figs. 5- 7. Among them, Fig. 5 and Fig. 6 display the results on the UHD-MFF-Syn dataset. Observing the residual maps, MSFIN-Fusion struggles to preserve sufficient near-focus information, whereas SMFuse and SESF suffer from severe detail loss in far-focus regions. TC-MoA exhibits a relatively uniform information loss across the entire image. Fig. 7 visualizes the results on the UHD-MFF-Real dataset. Although GIFNet retains details reasonably well, its fused results suffer from significant luminance and color distortions. As clearly seen in Fig. 7, GIFNet

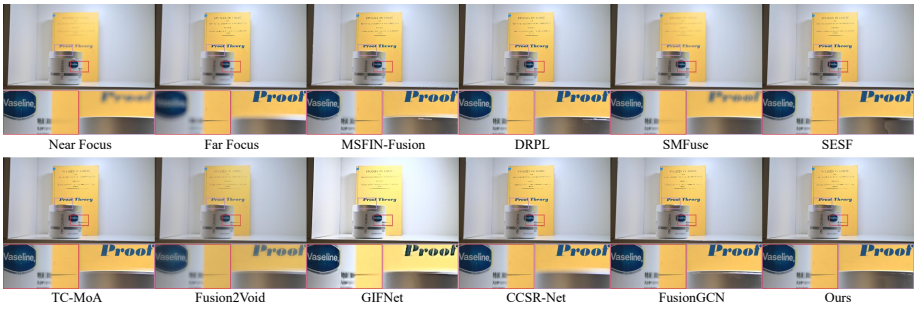

Fig. 7: Qualitative comparison on UHD-MFF-Real.

Table 2: Quantitative comparisons on the UHD-MFF dataset. **Bold** indicates the best, and underline indicates the second-best.

Method	UHD-MFF-Real							UHD-MFF-Syn						
	SD↑	MI↑	EN↑	VIF↑	$Q^{AB/F}$↑	AG↑	SF↑	SD↑	MI↑	EN↑	VIF↑	$Q^{AB/F}$↑	AG↑	SF↑
MSFIN-Fusion	41.0410	4.7667	6.9540	1.3038	0.7145	2.9993	8.6925	49.5343	**6.7902**	7.1517	1.3781	0.6280	1.4852	5.1240
DRPL	41.2584	4.5540	6.8674	1.3187	0.6938	2.9273	8.8130	49.0376	6.1102	7.1611	1.4034	0.7121	1.9383	6.7343
SMFuse	39.2386	3.5286	6.8001	0.8858	0.3739	1.6319	4.7578	48.8134	5.6906	7.1196	1.1449	0.5046	1.1714	4.0319
SESF	41.0118	4.6523	6.7696	1.3628	0.7015	3.0181	8.7894	49.5925	5.3177	7.1522	0.9414	0.3545	1.1705	3.3623
TC-MoA	40.8527	3.8540	6.7344	1.2048	0.6828	2.9395	8.6218	49.8702	5.9282	7.1661	1.4106	0.7491	1.9318	6.7099
Fusion2Void	34.3669	3.4135	6.6375	0.8356	0.3535	1.6709	5.1135	42.9653	5.2050	6.9463	0.9647	0.4350	1.2541	4.5707
GIFNet	**43.6695**	3.0130	6.7784	0.4957	0.1952	2.0124	5.4410	45.5961	4.4482	7.0551	0.8195	0.2674	1.5112	5.1649
CCSR-Net	41.6663	4.5768	6.9491	1.3410	0.7002	2.9764	8.7278	49.8757	6.2169	7.1659	1.4186	0.7448	1.9293	6.6790
FusionGCN	40.6867	4.6535	6.9472	1.3147	0.7078	2.8628	8.6057	49.4332	6.7128	7.1621	1.4036	0.7216	1.9163	6.6971
Ours	41.8366	**4.8029**	**6.9588**	**1.3814**	**0.7290**	**3.0344**	**8.8598**	**49.9062**	6.3999	**7.1672**	**1.4335**	**0.7498**	**1.9416**	**6.7443**

introduces noticeable brightness shifts and color deviations during the fusion process. At the foreground-background boundary around the bottle cap, our method demonstrates the best detail preservation capability, while other methods introduce varying degrees of artifacts or blurring. Furthermore, our approach achieves the best overall texture preservation across the entire image. Comprehensive evaluations across the two datasets indicate that our method exhibits distinct advantages on high-resolution real-world images while maintaining robust performance on low-resolution inputs.

5.3 Quantitative Experiments

The quantitative comparisons between our method and other SOTA approaches across the UHD-MFF dataset are reported in Tab. 2. As shown, our method achieves the best performance in six out of seven evaluated metrics. On the UHD-MFF-Syn, in terms of SD and VIF, our approach outperforms the second-best method, CCSR-Net, by margins of 0.0305 and 0.0149, respectively. For EN and $Q^{AB/F}$, our method surpasses the runner-up, TC-MoA, by 0.0011 and 0.0007. These quantitative results verify that our fusion strategy retains significantly more source image information compared to other competing methods. On the UHD-MFF-Real, our method exhibits a comprehensive advantage over

Fig. 8: Qualitative comparison of ablation experiment on UHD-MFF dataset.

Table 3: Ablation study of different components.

Model Type	SD↑	SCD↑	MI↑	EN↑	VIF↑	$Q^{AB/F}$↑	AG↑	SF↑
w/o $C_{GD}^{\text{C-LUT}}$	49.7349	0.3527	6.1312	7.1586	1.4028	0.7449	1.7748	6.1405
w/o $C_{DG}^{\text{C-LUT}}$	49.2397	0.2412	5.9196	7.1497	1.2564	0.6601	1.5416	5.4037
w/o $C_{MG}^{\text{C-LUT}}$	49.6651	0.3407	6.0500	7.1580	1.3948	0.7436	1.7491	6.0278
w/o $C_{RS}^{\text{C-LUT}}$	49.5990	0.3351	5.9962	7.1564	1.3833	0.7417	1.7202	5.8832
w/o D-LUT	49.8771	**0.3540**	6.3538	7.1636	1.4210	**0.7501**	1.8479	6.4938
Full Model	**49.9062**	0.3414	**6.3999**	**7.1672**	**1.4335**	0.7498	**1.9416**	**6.7443**

the alternatives, further demonstrating our outstanding capability in handling high-resolution fusion tasks.

5.4 Ablation Experiments

We conduct comprehensive ablation studies on the UHD-MFF benchmark to validate our architectural components, training data scale, and supervision paradigms.

Effectiveness of Individual Components. We evaluate the four focus cues and the Detail Branch. As shown in Tab. 3 and Fig. 8, removing any single module degrades fusion quality, causing boundary artifacts and detail loss. Our full model consistently achieves the best quantitative performance, confirming the indispensability of each component.

Impact of Training Data Scale. To evaluate the impact of dataset scale, we conduct an ablation study with three training sizes: 1500 (large), 150 (medium), and 15 (small). As shown in Fig. 9 and Tab. 4, performance exhibits a consistent upward trend with increasing data scale. All quantitative metrics improve progressively as the training set expands, validating large-scale necessity.

Unsupervised vs. Supervised Paradigms. Despite having synthetic ground-truth images, we deliberately employ an unsupervised strategy to prevent the

Fig. 9: Qualitative comparison of different training set sizes.

Table 4: Quantitative comparison of different training set sizes.

Scale of Data	AG↑	SD↑	SCD↑	MI↑	EN↑	SF↑	VIF↑	$Q^{AB/F}$↑
Small	1.4792	49.0784	0.2223	6.1959	7.1415	6.4114	1.1637	0.5367
Medium	1.4904	49.1030	0.2329	6.1716	7.1423	6.3984	1.1726	0.5400
Large	**1.9416**	**49.9062**	**0.3414**	**6.3999**	**7.1672**	**6.7443**	**1.4335**	**0.7498**

Fig. 10: Qualitative comparison between supervised and unsupervised strategies.

Table 5: Quantitative comparison between supervised and unsupervised strategies.

Training Strategy	AG↑	SD↑	SCD↑	MI↑	EN↑	SF↑	VIF↑	$Q^{AB/F}$↑
Supervised	1.9301	49.9301	0.3535	6.3918	7.1666	6.7423	1.4367	0.7491
Unsupervised	1.9416	49.9062	0.3414	6.3999	7.1672	6.7443	1.4335	0.7498

network from overfitting to specific synthetic degradation patterns. Comparing our model with a fully supervised counterpart shown in Fig. 10 and Tab. 5, we observed that while numerical metrics are comparable, the unsupervised paradigm yields significantly more natural luminance and softer transitions. By avoiding the unnatural contrast artifacts occasionally seen in supervised results, our focus-aware unsupervised loss better preserves spectral consistency and generic sharpness discrimination, justifying it as our default setting.

Impact of Training Resolution. To verify the necessity of high-resolution training data, we trained our model on existing low-resolution datasets and evaluated its performance on UHD scenarios. As shown in Fig. 11 and Tab. 6, models trained on low-resolution data suffer from severe detail degradation when applied to 4K inputs. Both visual quality and quantitative metrics confirm that training on native 4K datasets is indispensable for high-fidelity UHD fusion.

Performance Comparison. Superior computational efficiency constitutes the most distinct advantage of our proposed framework. The comparative performance benchmarks are detailed in Tab. 7. Tested on an NVIDIA GeForce RTX 3090 GPU with the UHD-MFF dataset, our method achieves an average in-

Fig. 11: Qualitative comparison of the model performance trained on low-resolution and high-resolution data.

Table 6: Performance comparison on the high resolution test set.

Type of Data	AG↑	SD↑	SCD↑	MI↑	EN↑	SF↑	VIF↑	$Q^{AB/F}$↑
low resolution	1.6032	49.2849	0.2886	5.8601	7.1475	6.2509	1.2233	0.6129
high resolution	**1.9416**	**49.9062**	**0.3414**	**6.3999**	**7.1672**	**6.7443**	**1.4335**	**0.7498**

Table 7: Comparison of computational efficiency and model complexity.

Method	Test Time (ms)	FLOPs (G)	Params (M)	Peak Mem (MB)	Energy Cost (mJ)
TC-MoA	5714.91 ± 13.78	94370.32	340.89	12024.95	434155.48
SMFuse	$\underline{150.35 \pm 20.96}$	637.57	0.038	29651.47	1514.18
SESF	450.85 ± 10.06	**14.814**	$\underline{0.014}$	$\underline{2109.36}$	$\underline{84.07}$
Fusion2Void	883.48 ± 16.49	2161.85	0.26	7374.96	9992.3
MSFIN-Fusion	29453.10 ± 4891.83	1693.34	4.59	10573.32	7837.12
GIFNet	235.27 ± 1.44	381.56	0.82	8015.64	1771.11
DRPL	484.11 ± 21.40	8890.54	1.071	17744.65	40944.24
CCSR-Net	2910.29 ± 409.93	$\underline{50.21}$	0.025	6392.68	278.72
FusionGCN	15382.78 ± 500.96	388.56	0.13	17350.42	1835.17
Ours	$\mathbf{11.03 \pm 0.82}$	67.82	**0.0081**	**1133.15**	**67.45**

ference latency of 11.03 ms per image, which translates to a throughput of 90 frames per second. Notably, our model architecture is the most compact among all competing methods, containing merely 0.0081 million parameters. Regarding resource overhead during inference, our approach requires a peak memory of only 1133.15 MB and consumes a minimal energy of 67.45 mJ.

To validate the practicality of our method in real-world scenarios, we deployed the proposed framework onto a mobile platform. We selected the Huawei P60 Pro as our experimental testbed, which is equipped with the Snapdragon 8+ Gen 1 processor. The performance comparison is presented in Tab. 8. Among the selected comparison methods, only SESF can run on a mobile device, with an average inference latency of 55.68 s, while the other methods fail to operate on mobile devices. In contrast, the average inference latency of our method consistently remains within 1 second, even for high-resolution inputs, demonstrating its substantial advantage for practical deployment on mobile devices.

Table 8: Comparison of performance on mobile devices.

Metrics	TC-MoA	SMFuse	SESF	Fusion2Void	MSFIN-Fusion
Test time (s)	–	–	55.68	–	–
Deployment	✗	✗	✓	✗	✗
Metrics	GIFNet	DRPL	CCSR-Net	FusionGCN	**Ours**
Test time (s)	–	–	–	–	**0.97**
Deployment	✗	✗	✗	✗	✓

6 Conclusion

This paper effectively shatters the three major barriers, including data availability, model adaptability, and deployment feasibility, that hinder multi-focus image fusion in ultra-high-definition scenarios. To bridge the existing data gap, we establish UHD-MFF, the first large-scale UHD resolution multi-focus dataset, providing a robust foundation for future research. Furthermore, we propose UMF-LUT, a highly efficient, scale-specialized lookup-table framework tailored for UHD images. By innovatively decoupling the fusion process into low-resolution region-level decisions via C-LUT and high-resolution edge-level refinements via D-LUT, our method ensures accurate focus perception while successfully bypassing the massive memory overhead of full-resolution convolutions. Extensive experiments confirm that UMF-LUT not only achieves state-of-the-art performance in both visual fidelity and quantitative metrics, but also realizes real-time 4K fusion with minimal computational cost. Ultimately, this work breaks the hardware limitations of existing methods, demonstrating strong feasibility for smartphone deployment and significantly advancing the practical application of multi-focus image fusion in next-generation UHD imaging.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant nos. 625B2135, and 62276192.

References

1. Bai, X., Zhang, Y., Zhou, F., Xue, B.: Quadtree-based multi-focus image fusion using a weighted focus-measure. Information Fusion **22**, 105–118 (2015)
2. Bhat, S., Koundal, D.: Multi-focus image fusion techniques: a survey. Artificial Intelligence Review **54**(8), 5735–5787 (2021)
3. Cheng, C., Xu, T., Feng, Z., Wu, X., Tang, Z., Li, H., Zhang, Z., Atito, S., Awais, M., Kittler, J.: One model for all: Low-level task interaction is a key to task-agnostic image fusion. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 28102–28112 (2025)
4. Hu, X., Jiang, J., Liu, X., Ma, J.: Zmff: Zero-shot multi-focus image fusion. Information Fusion **92**, 127–138 (2023)

5. Huang, Z., Liu, J., Fan, X., Liu, R., Zhong, W., Luo, Z.: Reconet: Recurrent correction network for fast and efficient multi-modality image fusion. In: Proceedings of the European Conference on Computer Vision. pp. 539–555 (2022)
6. Li, C., Guo, C.L., Zhou, M., Liang, Z., Zhou, S., Feng, R., Loy, C.C.: Embedding fourier for ultra-high-definition low-light image enhancement. In: Proceedings of the International Conference on Learning Representations (2023)
7. Li, H., Wu, X.J.: Multi-focus image fusion using dictionary learning and low-rank representation. In: Proceedings of the International Conference on Image and Graphics. pp. 675–686 (2017)
8. 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 Transactions on Image Processing **29**, 4816–4831 (2020)
9. Lin, H., Lin, Y., Xia, J., Fan, L., Li, F., Wang, Y., Ding, X.: Fusion2void: Unsupervised multi-focus image fusion based on image inpainting. IEEE Transactions on Circuits and Systems for Video Technology **35**(4), 3328–3341 (2024)
10. Liu, J., Fan, X., Jiang, J., Liu, R., Luo, Z.: Learning a deep multi-scale feature ensemble and an edge-attention guidance for image fusion. IEEE Transactions on Circuits and Systems for Video Technology **32**(1), 105–119 (2021)
11. Liu, Y., Liu, S., Wang, Z.: Multi-focus image fusion with dense sift. Information Fusion **23**, 139–155 (2015)
12. Liu, Y., Wang, L., Cheng, J., Chen, X.: Multiscale feature interactive network for multifocus image fusion. IEEE Transactions on Instrumentation and Measurement **70**, 1–16 (2021)
13. Liu, Y., Wang, L., Cheng, J., Li, C., Chen, X.: Multi-focus image fusion: A survey of the state of the art. Information Fusion **64**, 71–91 (2020)
14. Luo, X., Zhang, Z., Zhang, C., Wu, X.: Multi-focus image fusion using hosvd and edge intensity. Journal of Visual Communication and Image Representation **45**, 46–61 (2017)
15. Ma, B., Zhu, Y., Yin, X., Ban, X., Huang, H., Mukeshimana, M.: Sesf-fuse: An unsupervised deep model for multi-focus image fusion. Neural Computing and Applications **33**(11), 5793–5804 (2021)
16. Ma, J., Le, Z., Tian, X., Jiang, J.: Smfuse: Multi-focus image fusion via self-supervised mask-optimization. IEEE Transactions on Computational Imaging **7**, 309–320 (2021)
17. Ma, J., Ma, Y., Li, C.: Infrared and visible image fusion methods and applications: A survey. Information Fusion **45**, 153–178 (2019)
18. Ma, X., Hu, S., Liu, S., Fang, J., Xu, S.: Multi-focus image fusion based on joint sparse representation and optimum theory. Signal Processing: Image Communication **78**, 125–134 (2019)
19. Nejati, M., Samavi, S., Shirani, S.: Multi-focus image fusion using dictionary-based sparse representation. Information Fusion **25**, 72–84 (2015)
20. Ouyang, Y., Zhai, H., Hu, H., Li, X., Zeng, Z.: Fusiongcn: Multi-focus image fusion using superpixel features generation gcn and pixel-level feature reconstruction cnn. Expert Systems with Applications **262**, 125665 (2025)
21. Qiu, X., Li, M., Zhang, L., Yuan, X.: Guided filter-based multi-focus image fusion through focus region detection. Signal Processing: Image Communication **72**, 35–46 (2019)
22. Qu, G., Zhang, D., Yan, P.: Information measure for performance of image fusion. Electronics Letters **38**(7), 313–315 (2002)

23. Quan, Y., Wan, X., Tang, Z., Liang, J., Ji, H.: Multi-focus image fusion via explicit defocus blur modelling. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 6657–6665 (2025)
24. Roberts, J.W., Van Aardt, J.A., Ahmed, F.B.: Assessment of image fusion procedures using entropy, image quality, and multispectral classification. Journal of Applied Remote Sensing **2**(1), 023522 (2008)
25. Shao, X., Jin, X., Jiang, Q., Miao, S., Wang, P., Chu, X.: Multi-focus image fusion based on transformer and depth information learning. Computers and Electrical Engineering **119**, 109629 (2024)
26. Wang, X., Fang, L., Zhao, J., Pan, Z., Li, H., Li, Y.: Mmae: A universal image fusion method via mask attention mechanism. Pattern Recognition **158**, 111041 (2025)
27. Xu, H., Ma, J., Jiang, J., Guo, X., Ling, H.: U2fusion: A unified unsupervised image fusion network. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(1), 502–518 (2022)
28. Xu, S., Wei, X., Zhang, C., Liu, J., Zhang, J.: Mffw: A new dataset for multi-focus image fusion. arXiv preprint arXiv:2002.04780 (2020)
29. Yi, X., Ma, Y., Li, Y., Xu, H., Ma, J.: Artificial intelligence facilitates information fusion for perception in complex environments. The Innovation **6**(4) (2025)
30. Yi, X., Tang, L., Zhang, H., Xu, H., Ma, J.: Diff-if: Multi-modality image fusion via diffusion model with fusion knowledge prior. Information Fusion **110**, 102450 (2024)
31. Yi, X., Zhang, Y., Xiang, X., Yan, Q., Xu, H., Ma, J.: Lut-fuse: Towards extremely fast infrared and visible image fusion via distillation to learnable look-up tables. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14559–14568 (2025)
32. 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. Information Fusion **66**, 40–53 (2021)
33. Zhang, H., Xu, H., Tian, X., Jiang, J., Ma, J.: Image fusion meets deep learning: A survey and perspective. Information Fusion **76**, 323–336 (2021)
34. Zhang, J., Liao, Q., Liu, S., Ma, H., Yang, W., Xue, J.H.: Real-mff: A large realistic multi-focus image dataset with ground truth. Pattern Recognition Letters **138**, 370–377 (2020)
35. Zhang, Y., Liu, Y., Sun, P., Yan, H., Zhao, X., Zhang, L.: Ifcnn: A general image fusion framework based on convolutional neural network. Information Fusion **54**, 99–118 (2020)
36. Zheng, K., Cheng, J., Liu, Y.: Unfolding coupled convolutional sparse representation for multi-focus image fusion. Information Fusion **118**, 102974 (2025)
37. Zhu, P., Sun, Y., Cao, B., Hu, Q.: Task-customized mixture of adapters for general image fusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7099–7108 (2024)