Dual-Path Deep Unsupervised Learning for Multi-Focus Image Fusion

Yuhui Quan, Xi Wan, Tianxiang Zheng, Yan Huang and Hui Ji

***Abstract*—Multi-focus image fusion (MFIF) aims at merging multiple images captured at different focal lengths to create an all-in-focus image. This paper introduces a fully unsupervised learning approach for MFIF that uses only pairs of defocused images for end-to-end training, bypassing the need for ground-truths in supervised learning. Unlike existing methods training via a similarity loss between fused and source images, we propose a dual-path learning framework comprising two networks: an image fuser and a mask predictor. The mask predictor is modeled as a self-supervised denoising network on imperfect fusion masks, trained with a blind-spot unsupervised learning scheme. The image fuser, crafted with deep unrolling, leverages the output from the mask predictor to supervise its mask generation at each unrolled step. Moreover, we introduce a fusion consistency loss to ensure the alignment between the image fuser and the mask predictor. In extensive experiments, our proposed approach shows superiority over existing end-to-end unsupervised methods and competitive performance against the supervised ones.**

***Index Terms*—Multi-focus image fusion, End-to-end unsupervised learning, Blind-spot scheme, Deep unrolling.**

I. Introduction

Objects out of focus within an image appear blurry, impairing both human visual perception and computer vision applications. Typically, capturing a single image with every object in focus is difficult due to the inherent limitations in depth of field, the range within which objects appear sharp. Multi-focus image fusion (MFIF) is a technique aiming at overcoming this limitation, by blending multiple images captured at different focal lengths to create a composite image with all regions in focus. MFIF has been applied in various fields, *e.g.*, machine vision, medical imaging, and remote sensing (*e.g.* [1]–[5]).

Conventional MFIF methods (*e.g.* [6]–[13]) typically generate a decision map (also termed fusion mask) to denote pixel focus degree, or an activity level map for feature importance in a transform domain, guiding the fusion of source images or features. However, these methods rely on handcrafted rules or models, which may be too simplistic for real-world images with complex structures [5].

In recent years, deep learning (DL) has become a prominent technique for MFIF, primarily due to the exceptional capacity of deep neural networks (NNs) to capture complex patterns in images [5]. The majority of existing DL-based MFIF methods rely on the supervised learning paradigm, requiring a large number of ground-truth (GT) all-in-focus images for NN training; see *e.g.* [14]–[39]. While supervised methods have indeed shown success in some applications, they heavily rely on a large number of accurately annotated all-in-focus images. In many practical scenarios, obtaining such images that thoroughly cover all variations can be prohibitively expensive, especially when dealing with diverse or complex data, or data with distributions that change over time. For example, in fields like optical imaging in science and engineering, acquiring all-in-focus images may be not practical.

Unsupervised DL methods offer significant advantages in these scenarios by eliminating the need for ground truth data. They are more adaptable to varying data distributions and can be more easily applied to new and diverse datasets without the costly and time-consuming acquisition process of all-in-focus images. Building on these benefits, a variety of unsupervised DL approaches have been introduced for MFIF. The self-supervised pre-training techniques [40]–[50] utilize self-expression or pretext tasks to train feature extractors using standard natural image sets, but they do not fully utilize the rich complementary information provided in paired defocused images for unsupervised training. The zero-shot DL methods [51], [52] apply sample-specific learning to test data, bypassing the need for training datasets, whereas this approach is computationally costly. The end-to-end unsupervised DL methods [44], [53]–[60] train a universal model on defocused image pairs, addressing previous constraints. Their performance, however, leaves substantial room for advancement compared to other types of unsupervised methods.

Aim and Main Idea: This paper aims at developing an effective end-to-end unsupervised DL approach for MFIF with performance boost. To address the absence of GTs during training, we propose a self-supervised dual-path framework utilizing two NNs; see Fig. 1 for an illustration. One NN serves as a mask predictor which estimates the ideal fusion mask from the initial one generated via a simple maximal gradient criterion. The other NN serves as an image fuser, designed via deep unrolling (*i.e.*, unfolding an iterative solver for a

Y. Quan, X. Wan and T. Zheng are with School of Computer Science and Engineering at South China University of Technology, Guangzhou, 510000, China, as well as Pazhou Lab, Guangzhou 510335, China. (email: csyhquan@scut.edu.cn; csxwan@mail.scut.edu.cn; cszhengtx@mail.scut.edu.cn)

Y. Huang is with School of Computer Science and Engineering at South China University of Technology, Guangzhou, 510000, China. (email: aihuangy@scut.edu.cn)

H. Ji is with Department of Mathematics at National University of Singapore, 119076, Singapore. (email: matjh@nus.edu.sg)

Corresponding author: Yan Huang.

This work was supported in part by National Natural Science Foundation of China under Grant 62372186, in part by Science and Technology Plan Project of Guangzhou under Grant 2023A04J1681, in part by Natural Science Foundation of Guangdong Province under Grants 2022A1515011755 and 2023A1515012841, in part by Fundamental Research Funds for the Central Universities under Grant x2jsD2230220, and in part by Singapore MOE AcRF Tier 1 under Grant A-8000981-00-00.

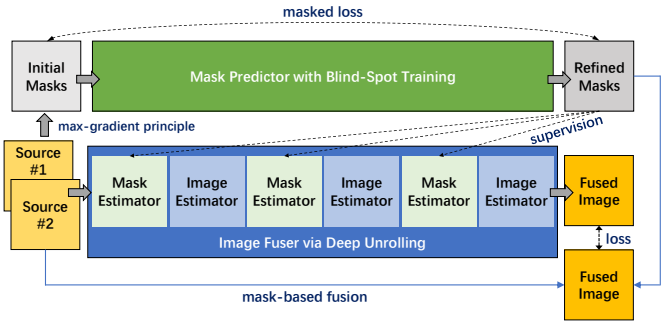

Fig. 1. Illustration of main idea of our proposed approach.

variational model), predicting the fused image from the source images by iteratively alternating a fusion mask estimator and a fused image estimator. This unrolling NN enjoys interpretability, providing implicit regularization from its structure for reducing possible overfitting in unsupervised training. To our knowledge, it is one of the few unrolling NN architecture for MFIF. Different from the existing one [36] that alternates two coefficient estimators via unrolled sparse coding, our unrolling NN alternates the estimations of fusion masks and images, facilitating the unsupervised training scheme.

In our dual-path framework, the two NNs are synergistically trained. We cast the unsupervised mask prediction as a self-supervised denoising task, where the initial masks are considered as the noisy approximations of the ideal fusion masks. Motivated by existing blind-spot un-supervised denoising techniques [61], [62], the mask predictor is trained using a masked self-reconstruction loss. This strategy nudges the predicted masks towards the ideal ones, without explicit GTs. The output of the mask predictor serves two purposes: supervising the mask estimators within the unrolling NN of the image fuser as well as generating a fused image by applying the output masks to the source images. Furthermore, we impose a fusion consistency loss between this fused image and the output of the image fuser. This interconnected training scheme ensures that the dual paths enhance each other, leading to higher accuracy and reliability for MFIF.

Contributions: We conduct extensive performance evaluation on several datasets, with comparisons against a rich set of baseline methods across a variety of metrics. The results show that our approach not only has a noticeable improvement in overall performance over current end-to-end unsupervised DL-based methods, but also compete well against other types of methods, including the supervised ones.

To conclude, the main contributions of this work include:

- A dual-path framework with deep unrolling, tailored for unsupervised learning of MFIF, leveraging interactions between the two paths to address the absence of GTs;
- A blind-spot based un-supervised learning technique for fusion mask prediction;
- A fusion consistency loss for the alignment between the fusion mask predictor and the image fuser;
- An efficient unsupervised MFIF method with SOTA performance.

II. Related Work

There is abundant literature on MFIF. This section only reviews the most pertinent studies. Interested readers are referred to [3]–[5] for more comprehensive surveys,

A. Non-DL Methods for MFIF

Conventional MFIF methods typically generate a decision map for fusion, using a focus measure applied to image pixels/blocks; see *e.g.* [7], [8], [10], [13]. To improve decision maps, the studies in [6], [11], [12] employ dictionary learning on image patches. Different from these spatial methods, transform-based methods (*e.g.* [9]) extract and fuse image features through a transform, usually with an activity level map, and reconstruct the fused image via an inverse transform. The conditional random field optimization model proposed in [63] enjoy the merits of both spatial and transform-based methods. Overall, as conventional methods separately and manually craft focus measurements and fusion rules, they may not work well on real-world images in complex situations [5].

B. Supervised DL Methods for MFIF

Supervised DL-based MFIF methods typically fall into two categories: classification-based and regression-based. Classification-based methods [14]–[31], [39] treat MFIF as a segmentation task. A deep NN is trained to classify pixels/patches as focused or defocused before fusion. These works align well with conventional spatial methods. In contrast, regression-based methods [32]–[38] map source images to fused images, usually using a deep NN with three parts: feature extraction, feature fusion, and image reconstruction, mirroring conventional transform-based methods [4].

Each type of these methods has its own strengths and weaknesses. Classification-based methods are more interpretable and sharpness-preserving, whereas often causing artifacts near edges. Regression-based methods tend to produce more-natural visual effect around boundaries but may reduce the sharpness of the fused image. To harness both the advantages, the two-stage NN proposed in [64] first regresses an initial fused image based on which the decision map is predicted and refined for fusion, and the two-branch network in [65] employs a regression branch and a classification branch.

C. Unsupervised DL Methods for MFIF

Existing unsupervised DL methods for MFIF vary with different configurations of training data, leading to diverse approaches. Based on the configuration in training data, these methods can be categorized into three types.

[Standard image datasets] Self-supervised pre-training of feature extractors: This kind of methods leverages the existing large-scale datasets of natural images (*e.g.* the CoCo dataset [66]) for self-supervised pre-training of feature extractors within MFIF NNs. Most works use an auto-encoder scheme with self-reconstruction for pre-training; see *e.g.* [40]–[46], [50]. Taking a different approach, Liang *et al.* [47] utilize a block-masking-based pre-text task for pre-training. Wang *et al.* [48] define the pre-text task by super-resolution.

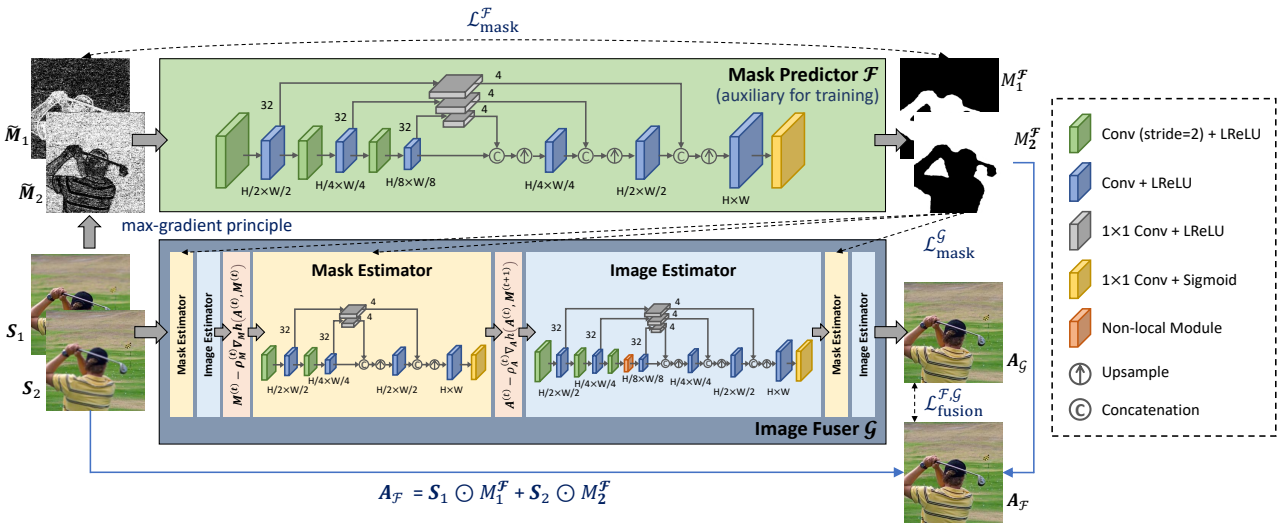

Fig. 2. Architecture of our proposed dual-path framework for unsupervised MFIF.

[Defocused image pairs] End-to-end unsupervised DL: Methods of this kind directly train NNs on defocused image pairs. They rely on various loss functions defined between fused and source images, such as ℓ_2-loss [54], [60], [67], ℓ_1-loss [59], SSIM loss [53], [56], [58]–[60], [68], perceptual loss [56], structure tensor loss [55], gradient-domain loss [56], [58], max-gradient loss [57], [59], [67] and local difference evaluation loss [49]. Our approach foregoes the direct similarity loss between fused and source images, opting instead for unsupervised learning via a dual-path framework with novel loss functions. Unlike [57], which uses a smoothed ℓ_1-loss between initial and predicted decision maps, we adopt blind-spot training with a masked loss for self-supervised fusion mask denoising with better performance.

[A test sample itself] Zero-shot self-supervised learning: Zero-shot self-supervised learning is performed only on the test sample itself. Based on the deep image prior [69], *i.e.*, a convolutional NN (CNN) prioritizes fitting image structures over random noise. He *et al.* [51], [52] proposed to parameterize the ideal fusion masks and all-in-focus image by separate untrained CNNs, whose weights are jointly optimized via a reconstruction loss. Early stopping is employed to mitigate errors in both fusion mask and fused image. This kind of methods is free from training data, but its per-sample test-time training can be computationally overwhelming, particularly when processing many samples.

III. Methodology

A. Problem Statement and Overall Framework

The task of MFIF is to synthesize a composite image with all its regions in focus, using source images captured on the same scene but with different focal settings. Let $(\boldsymbol{S}_1, \boldsymbol{S}_2) \subset \mathbb{R}^{H\times W}$ denote a source image pair. The fused image $\boldsymbol{A} \in \mathbb{R}^{H\times W}$ can be composited by

$$\boldsymbol{A} = \boldsymbol{M}_1 \odot \boldsymbol{S}_1 + \boldsymbol{M}_2 \odot \boldsymbol{S}_2, \tag{1}$$

where $\boldsymbol{M}_1, \boldsymbol{M}_2$ denote two fusion masks (*i.e.* soft decision maps), satisfying the following constraints:

$$\boldsymbol{M}_1 + \boldsymbol{M}_2 = \boldsymbol{1} \quad \text{and} \quad \boldsymbol{M}_1, \boldsymbol{M}_2 \in [0,1]^{H\times W}. \tag{2}$$

Here, $\boldsymbol{1}$ denotes a matrix of ones and $\odot$ denotes the element-wise multiplication operation.

Addressing practical demands, our unsupervised DL approach performs end-to-end training with the data that only contains source image pairs, without any GT fused or all-in-focus image involved. See Fig. 2 for an illustration of our approach which employs a dual-path framework to facilitate unsupervised training. In this framework, one path, termed mask predictor, generates initial masks from source images and refines them by a CNN trained with the blind-spot-based unsupervised learning scheme. The other path, termed image fuser, employs a deep unrolling NN to synthesize the fused image from the source images. These two paths are interconnected through loss functions during training. The mask predictor acts as an auxiliary and is only utilized in training, while the image fuser is used in test.

B. Mask Predictor with Blind-Spot Training

The NN of mask predictor, denoted by $\mathcal{F} : [0,1]^{H\times W} \to [0,1]^{H\times W}$, has an encoder-decoder convolutional structure with skip connection blocks. It takes an initial mask of $\boldsymbol{S}_1$ or $\boldsymbol{S}_2$ as input and outputs a refined version with each of its elements constrained within $[0,1]$ by a Sigmoid activation. The initial masks, denoted by $\widetilde{\boldsymbol{M}}_1$ and $\widetilde{\boldsymbol{M}}_2$, are derived using the principle of maximal gradient:

$$\widetilde{\boldsymbol{M}}_k = \operatorname{sgn}\big(\max(|\nabla \boldsymbol{S}_1|, |\nabla \boldsymbol{S}_2|) - |\nabla \boldsymbol{S}_k|\big), \ k = 1, 2, \tag{3}$$

where sgn, max and $|\cdot|$ denote the element-wise sign, maximum, and magnitude operations, respectively. This principle is based on the observation that the gradient of a focused pixel is usually larger than that of a defocused one.

The initial masks can be considered as the noisy observations of the ideal masks with complex statistical characteristics. The refinement of these masks can be recast as a

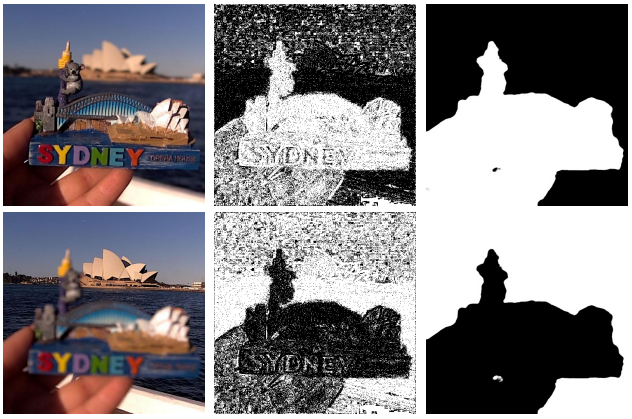

Fig. 3. Illustration of results from the mask predictor $\mathcal{F}$.

denoising problem, and our approach is based on a specific implementation of the blind-spot self-supervised denoising techniques [61], [62]. The process involves obscuring a subset of pixels randomly and reconstructing them from the remaining pixels. For the binary masks $\widetilde{\boldsymbol{M}}_1, \widetilde{\boldsymbol{M}}_2$, we obscure the subset of pixels via

$$\mathcal{M}_{\boldsymbol{B}_k}(\widetilde{\boldsymbol{M}_k}) = \boldsymbol{B}_k \odot \widetilde{\boldsymbol{M}}_k + (\boldsymbol{1} - \boldsymbol{B}_k) \odot \boldsymbol{N}_k, \tag{4}$$

where $\boldsymbol{B}_k$ is a binary mask where $\boldsymbol{B}_k[x, y]$ is drawn from a Bernoulli distribution $\mathcal{B}(p)$ with a success probability of $p = 0.9$, and $\boldsymbol{N}_k$ is also a binary mask with $\boldsymbol{N}_k[x, y] \sim \mathcal{B}(p^k_{x,y})$ where $p^k_{x,y}$ is set to the percentage of values of 1 in the eight neighbors of $\widetilde{\boldsymbol{M}}_k[x, y]$. That is, we select a portion of pixels in $\widetilde{\boldsymbol{M}}_k$ and randomly reset them based on their neighborhood distribution.

Based on above, the loss function for the self-supervised learning of mask refinement is formulated as

$$\begin{aligned}\mathcal{L}^{\mathcal{F}}_{\text{mask}} := &\sum_{k=1,2} \|(\boldsymbol{1} - \boldsymbol{B}_k) \odot \big(\mathcal{F}(\mathcal{M}_{\boldsymbol{B}_k}(\widetilde{\boldsymbol{M}_k})) - \widetilde{\boldsymbol{M}_k}\big)\|^2_{\text{F}} \\ &+ \alpha\|\mathcal{F}(\mathcal{M}_{\boldsymbol{B}_1}(\widetilde{\boldsymbol{M}_1})) + \mathcal{F}(\mathcal{M}_{\boldsymbol{B}_2}(\widetilde{\boldsymbol{M}_2})) - \boldsymbol{1}\|^2_{\text{F}},\end{aligned} \tag{5}$$

with $\alpha \in \mathbb{R}^+$. In the first term of $\mathcal{L}^{\mathcal{F}}_{\text{mask}}$, the NN is trained to reconstruct the original non-perturbed mask pixels from the perturbed ones, with the loss computed solely on the perturbed pixels. This training strategy effectively regularizes the NN, steering it away from trivial solutions such as identity mapping, and towards solid denoising performance, as illustrated in Fig. 3. The second term in $\mathcal{L}^{\mathcal{F}}_{\text{mask}}$ reinforces the complementary nature of the masks, as depicted by the constraint in Eq. (2), bringing additional regularization.

C. Image Fuser via Deep Unrolling

The NN of image fuser, denoted by $\mathcal{G}$: $(\mathbb{R}^{H\times W}, \mathbb{R}^{H\times W}) \to \mathbb{R}^{H\times W}$, is constructed by unrolling the optimization model:

$$\begin{aligned}\min_{\boldsymbol{A},\boldsymbol{M}_1,\boldsymbol{M}_2}\ &\lambda\|\boldsymbol{A} - \boldsymbol{M}_1 \odot \boldsymbol{S}_1 - \boldsymbol{M}_2 \odot \boldsymbol{S}_2\|^2_{\text{F}} + \\ &\|\boldsymbol{1} - \boldsymbol{M}_1 - \boldsymbol{M}_2\|^2_{\text{F}} + \phi(\boldsymbol{A}) + \psi(\boldsymbol{M}_1) + \psi(\boldsymbol{M}_2),\end{aligned} \tag{6}$$

where $\lambda \in \mathbb{R}^+$ is a sufficiently-large weight, and ϕ, ψ denote the regularization terms for the fused image and fusion masks, respectively. Define $\boldsymbol{M} = [\boldsymbol{M}_1; \boldsymbol{M}_2] \in \mathbb{R}^{2H\times W}$, $\boldsymbol{S} = [\boldsymbol{S}_1; \boldsymbol{S}_2] \in \mathbb{R}^{2H\times W}$ and $\boldsymbol{\Sigma} = [\boldsymbol{I}, \boldsymbol{I}] \in \mathbb{R}^{H\times 2H}$, with $\boldsymbol{I}$ being an identity matrix. Then, let

$$h(\boldsymbol{A}, \boldsymbol{M}) = \lambda\|\boldsymbol{A} - \boldsymbol{\Sigma}(\boldsymbol{M} \odot \boldsymbol{S})\|^2_{\text{F}} + \|\boldsymbol{1} - \boldsymbol{\Sigma}\boldsymbol{M}\|^2_{\text{F}}. \tag{7}$$

Then we can rewrite (6) as

$$\min_{\boldsymbol{A},\boldsymbol{M}}\ h(\boldsymbol{A}, \boldsymbol{M}) + \phi(\boldsymbol{A}) + \psi(\boldsymbol{M}). \tag{8}$$

To solve (8), we apply the proximal gradient method, leading to the iteration scheme: for $t = 0, \cdots, T-1$:

$$\begin{cases}\boldsymbol{M}^{(t+1)} = \mathcal{P}_\psi\big(\boldsymbol{M}^{(t)} - \rho^{(t)}_{\boldsymbol{M}}\nabla_{\boldsymbol{M}} h(\boldsymbol{A}^{(t)}, \boldsymbol{M}^{(t)})\big), \\ \boldsymbol{A}^{(t+1)} = \mathcal{P}_\phi\big(\boldsymbol{A}^{(t)} - \rho^{(t)}_{\boldsymbol{A}}\nabla_{\boldsymbol{A}} h(\boldsymbol{A}^{(t)}, \boldsymbol{M}^{(t+1)})\big),\end{cases} \tag{9}$$

where $\mathcal{P}_f(\boldsymbol{Z}) := \operatorname{argmin}_{\boldsymbol{X}} \|\boldsymbol{X} - \boldsymbol{Z}\|^2_{\text{F}} + f(\boldsymbol{X})$ denotes the proximal operator, and $\rho^{(t)}_{\boldsymbol{M}}, \rho^{(t)}_{\boldsymbol{A}} \in \mathbb{R}^+$ are step sizes. The calculation of $\nabla_{\boldsymbol{M}} h(\boldsymbol{A}, \boldsymbol{M})$ and $\nabla_{\boldsymbol{A}} h(\boldsymbol{A}, \boldsymbol{M})$ is given by

$$\begin{aligned}\nabla_{\boldsymbol{M}} h(\boldsymbol{A}, \boldsymbol{M}) = &-2\lambda\boldsymbol{S} \odot \Big(\boldsymbol{\Sigma}^\top\big(\boldsymbol{A} - \boldsymbol{\Sigma}(\boldsymbol{M} \odot \boldsymbol{S})\big)\Big) \\ &- 2\boldsymbol{\Sigma}^\top(\boldsymbol{1} - \boldsymbol{\Sigma}\boldsymbol{M}),\end{aligned} \tag{10}$$

$$\nabla_{\boldsymbol{A}} h(\boldsymbol{A}, \boldsymbol{M}) = 2\lambda\big(\boldsymbol{A} - \boldsymbol{\Sigma}(\boldsymbol{M} \odot \boldsymbol{S})\big), \tag{11}$$

where $h(\boldsymbol{A}, \boldsymbol{M}) = \lambda\|\boldsymbol{A} - \boldsymbol{\Sigma}(\boldsymbol{M} \odot \boldsymbol{S})\|^2_{\text{F}} + \|\boldsymbol{1} - \boldsymbol{\Sigma}\boldsymbol{M}\|^2_{\text{F}}$.

The NN $\mathcal{G}$ is constructed by implementing the proximal operators $\mathcal{P}_\psi$ and $\mathcal{P}_\phi$ through two sub-NNs while treating the step sizes $\rho^{(t)}_{\boldsymbol{M}}$, $\rho^{(t)}_{\boldsymbol{A}}$ as trainable parameters. Consequently, the unrolling NN has T stages, each consisting of alternating connection of the mask and image estimators, as shown in Fig. 2. Analogous to $\mathcal{F}$, both estimators are structured with an encoder-decoder convolutional architecture with skip connection blocks. The mask estimator also uses a Sigmoid activation to fix the range of its estimated masks. In the bottleneck of the image estimator, a standard non-local module [70] is employed.

Remark 1. *The proposed unrolling NN can be directly extended from fusing two inputs to fusing more than two inputs, by including additional variables to the model* (6)*. Let* $(\boldsymbol{S}_k, \boldsymbol{M}_k)_{k=1}^K$ *denote the* K *(*$K \geq 3$*) input images and masks, respectively. The extended model* (6) *is expressed as follows:*

$$\begin{aligned}\min_{\boldsymbol{A},\{\boldsymbol{M}_k\}_{k=1}^K}\ &\lambda\|\boldsymbol{A} - \sum_{k=1}^K \boldsymbol{M}_k \odot \boldsymbol{S}_k\|^2_F + \|\boldsymbol{1} - \sum_{k=1}^K \boldsymbol{M}_k\|^2_F \\ &+ \phi(\boldsymbol{A}) + \sum_{k=1}^K \psi(\boldsymbol{M}_k).\end{aligned} \tag{12}$$

The unrolling is then performed similarly to the method that takes two images as input.

In the absence of GT images to guide training, it is crucial to prevent the unrolling NN from overfitting to trivial solutions. To address this concern, we leverage the mask predictor to supervise the output masks $\boldsymbol{M}^{(t)}$ from the intermediate mask estimators. Let $\boldsymbol{M}^{\mathcal{F}} = [\mathcal{F}(\widetilde{\boldsymbol{M}_1}); \mathcal{F}(\widetilde{\boldsymbol{M}_2})] \in \mathbb{R}^{2H\times W}$ denote the fusion masks generated by the mask predictor for the

source image pair $(\boldsymbol{S}_1, \boldsymbol{S}_2)$. We introduce the following loss function:

$$\mathcal{L}^{\mathcal{G}}_{\text{mask}} := \sum_{t=1}^{T} \omega_t \|\boldsymbol{M}^{(t)} - \boldsymbol{M}^{\mathcal{F}}\|_{\text{F}}^2, \tag{13}$$

where ω_t is set to 2^t, prioritizing the accuracy of predictions in the later unrolling stages. In implementation, $\mathcal{F}$ is detached in this loss, preventing potential adverse effects on the learning process of $\mathcal{F}$ caused by the inaccuracies of $\boldsymbol{M}^{(t)}$ in early unrolling stages.

D. Overall Training Loss and Inference Scheme

In addition to the $\mathcal{L}^{\mathcal{F}}_{\text{mask}}$ for training the mask predictor and the $\mathcal{L}^{\mathcal{G}}_{\text{mask}}$ for training the image fuser, we introduce a fusion consistency loss to the joint training of $\mathcal{F}$ and $\mathcal{G}$. Let

$$\boldsymbol{A}_{\mathcal{F}} = \boldsymbol{M}_1^{\mathcal{F}} \odot \boldsymbol{S}_1 + \boldsymbol{M}_2^{\mathcal{F}} \odot \boldsymbol{S}_2, \quad \boldsymbol{A}_{\mathcal{G}} = \mathcal{G}(\boldsymbol{S}_1, \boldsymbol{S}_2). \tag{14}$$

Here $\boldsymbol{A}_{\mathcal{F}}$ is the fused image composed via (1), using the masks predicted by $\mathcal{F}$, and $\boldsymbol{A}_{\mathcal{G}}$ is the fused image predicted by $\mathcal{G}$. The fusion consistency loss is then defined as

$$\mathcal{L}^{\mathcal{F},\mathcal{G}}_{\text{fusion}} = \|\boldsymbol{A}_{\mathcal{F}} - \boldsymbol{A}_{\mathcal{G}}\|_1. \tag{15}$$

In summary, the total loss function for training is given by

$$\mathcal{L}_{\text{total}} = \mathcal{L}^{\mathcal{F}}_{\text{mask}} + \beta \mathcal{L}^{\mathcal{G}}_{\text{mask}} + \gamma \mathcal{L}^{\mathcal{F},\mathcal{G}}_{\text{fusion}}, \quad \beta, \gamma \in \mathbb{R}^+. \tag{16}$$

After training, the inference is done via (1) using the masks output by the unrolling NN $\mathcal{G}$. This can ensure better fidelity with the source image, compared to directly using the output image of $\mathcal{G}$.

IV. Experiments

The performance evaluation is conducted using two training datasets. One is Lytro [71], a widely-used real dataset without GTs provided, and we compare with existing end-to-end unsupervised methods. The other is Wang *et al.*'s synthetic dataset [28] with GTs, for which we also compare with other types of methods including the supervised ones.

Implementation details: Our approach, called UDPNet (Unsupervised Dual-Path Network), is implemented with PyTorch and run on an NVIDIA GTX 1080Ti GPU. We set $T = 3$ for the image fuser and $(\alpha, \beta, \gamma, \lambda) = (0.4, 0.2, 0.2, 1)$ in the loss functions. The model is trained using Adam with 200 epochs, batch size 16, and learning rate 2×10^{-4}. When dealing with color images, the fusion is done at Y-channel of the YCbCr color space. Our code will be released upon paper's acceptance in GitHub: https://github.com/csxwan/UDPNet.

A. Ours vs. Unsupervised End-to-End Methods

Datasets: The Lytro dataset [71] provides real source image pairs acquired by the Lytro camera. Following [56], [60], the training set contains 66264 cropped image patch pairs of size 60×60, and the test set contains 10 image pairs. In addition, four commonly-used benchmark datasets are also used for test: MFFW [72], MFI-WHU [67], Real-MFF [73], and OR-PAM [2]. The MFFW dataset contains images collected from Internet, acquired by various devices; MFI-WHU is a synthetic dataset with GTs; Real-MFF is a real-world dataset with GTs; and OR-PAM is a biological image dataset.

TABLE I
Quantitative comparison of end-to-end unsupervised MFIF methods on MFI-WHU and Real-MFF datasets.

Method	MFI-WHU			Real-MFF			#Para. (M)	Time (s)
	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS		
PMGI	23.14	0.858	0.136	22.36	0.805	0.060	0.042	0.029
FusionDN	21.29	0.885	0.113	24.29	0.921	0.047	1.163	0.170
MFF-GAN	21.22	0.946	0.026	22.99	0.921	0.029	0.040	0.024
DIFNet	30.56	0.934	0.089	36.03	0.977	0.029	0.048	**0.009**
U2Fusion	27.85	0.946	0.048	27.93	0.943	0.032	0.659	0.090
SwinFusion	34.88	0.992	0.015	39.04	0.985	0.018	0.975	3.338
SMFuse	35.04	**0.994**	**0.011**	38.81	0.983	0.013	**0.038**	0.101
MUFusion	35.25	0.990	0.015	39.74	**0.987**	0.013	0.555	0.439
UDPNet	**36.68**	**0.994**	**0.011**	**40.05**	**0.987**	**0.012**	0.392	0.022

Metrics and baselines: Both MFI-WHU and Real-MFF provide GTs. Thus, on these datasets, we employ full-reference metrics (PSNR, SSIM, LPIPS) as they offer reliable performance evaluation by directly measure the difference between GTs and predictions. See Table I for the results on MFI-WHU and Real-MFF. In the GT-absent case, the performance metrics for MFIF vary in existing works. Following existing works [56], [60] on unsupervised MFIF, we adopted 10 metrics for quantitative performance evaluation in the absence of GTs, which are grouped into four types, including: (a) information theory-based: NMI [74], Q_{NCIE} [75], MI [76]; (b) feature-based: Q_{G} [77], Q_{M} [78], Q_{P} [79]; (c) structural similarity-based: Q_{C} [80], Q_{Y} [81]; (d) and human perception inspired: Q_{CB} [82], VIF [83]. To provide an overall evaluation, we compute the rank of a method among all compared methods for each metric, and report the average rank score, denoted as ARank, over all metrics. Totally seven end-to-end unsupervised MFIF methods are used for performance comparison, including PMGI [54], FusionDN [56], MFF-GAN [67], DIFNet [55], U2Fusion [60], SwinFusion [59], SMFuse [57], and MUFusion [84]. As the training datasets used by these methods differ from each other, for a fair comparison, we retrain all their NNs using our same dataset.

Quantitative analysis: It can be seen that our UDPNet performs the best on both the datasets in all metrics, showing significant PSNR gain over other methods. Particularly, the PSNR gain over the second-best performer on MFI-WHU is over 1dB. Table I also reports the computational complexity in terms of the number of parameters and the average runtime for an image pair of size 512×512. Both the model size and speed of UDPNet are close to most other compared methods, indicating that our effective architecture design leads to performance gain without introducing noticeable model complexity. Table II lists the ten metric scores and ARank on three GT-absent datasets. Our UDPNet ranks the 1st in most metrics and at least 2nd in other metrics, achieving the best ARank over all these datasets. These results have further demonstrated the effectiveness of our approach. See also supplemental material for the results in terms of no-reference metrics.

TABLE II
QUANTITATIVE COMPARISON OF UNSUPERVISED END-TO-END MFIF METHODS ON THREE DATASETS. **BOLD: TOP-1**; UNDERLINE: 2ND BEST.

Dataset	Method	NMI↑	Q_{NCIE}↑	Q_G↑	Q_M↑	Q_P↑	Q_C↑	Q_Y↑	Q_{CB}↑	MI↑	VIF↑	**ARank↓**
Lytro	PMGI	0.8431	0.8252	0.4574	0.3528	0.6711	0.6800	0.7410	0.5981	6.2423	1.0409	8.78
	FusionDN	0.8735	0.8266	0.4821	0.3511	0.7158	0.7353	0.7959	0.6025	6.4681	1.0515	7.44
	MFF-GAN	0.8440	0.8255	0.5589	0.6018	0.7289	0.7274	0.8496	0.5943	6.2809	1.1151	7.00
	DIF-Net	0.8835	0.8273	0.5085	0.3674	0.7429	0.7753	0.8330	0.5984	6.5773	1.0891	6.33
	U2Fusion	0.9111	0.8281	0.5567	0.4217	0.7765	0.7592	0.8491	0.6724	6.6800	1.1216	5.44
	SwinFusion	0.9640	0.8318	0.6687	1.1277	0.8217	0.7944	0.9480	0.7215	7.2063	1.2908	3.33
	SMFuse	1.2066	0.8477	0.7104	2.6299	0.8210	0.7927	0.9717	0.7911	9.0215	1.3589	2.33
	MUFusion	0.9954	0.8331	0.6545	0.9897	0.8185	0.7964	0.9429	0.7467	7.4184	1.2849	3.33
	UDPNet	**1.2166**	**0.8481**	**0.7314**	**2.6576**	**0.8453**	**0.8103**	**0.9900**	**0.8102**	**9.0946**	**1.3867**	**1.00**
MFFW	PMGI	0.7711	0.8194	0.4749	0.3459	0.5716	0.6542	0.7211	0.5740	5.3940	0.9196	8.33
	FusionDN	0.7896	0.8205	0.4837	0.3415	0.5812	0.6775	0.7423	0.5511	5.5869	0.9246	7.78
	MFF-GAN	0.7681	0.8196	0.5226	0.4990	0.5961	0.6791	0.7736	0.5159	5.3966	0.9280	7.22
	DIF-Net	0.8133	0.8213	0.5004	0.3455	0.6262	0.7398	0.8011	0.5647	5.7365	0.9525	6.00
	U2Fusion	0.8199	0.8210	0.5213	0.4010	0.6585	0.7101	0.7892	0.6180	6.6800	0.9543	5.22
	SwinFusion	0.8434	0.8227	0.6219	0.8422	0.7021	0.7495	0.8924	0.6597	5.9581	1.0886	3.44
	SMFuse	1.1574	**0.8400**	0.6582	2.5035	0.6569	0.7261	0.8928	0.6994	8.2442	1.1566	2.67
	MUFusion	0.8622	0.8238	0.6009	0.8231	0.6736	0.7515	0.8842	0.6721	6.0877	1.0870	3.33
	UDPNet	**1.1627**	**0.8400**	**0.7089**	**2.5040**	**0.7574**	**0.7815**	**0.9812**	**0.7548**	**8.2803**	**1.2732**	**1.00**
OR-PAM	PMGI	0.5462	0.8065	0.4505	0.1718	0.6492	0.5551	0.6297	0.5690	2.9448	0.8778	4.89
	FusionDN	0.3658	0.8050	0.3634	0.1496	0.4832	0.3874	0.4272	0.3674	2.3461	0.6599	7.56
	MFF-GAN	0.3486	0.8043	0.4573	0.2811	0.6033	0.5173	0.5860	0.3344	2.0075	0.6998	7.22
	DIF-Net	0.4109	0.8046	0.4122	0.1186	0.7219	0.6602	0.7436	0.5223	2.2107	0.8921	6.11
	U2Fusion	0.4411	0.8046	0.1478	0.1034	0.5821	0.2936	0.2852	0.5523	2.1584	0.6116	7.89
	SwinFusion	0.4828	0.8060	0.5878	0.2250	0.7030	0.7196	0.8990	0.6185	2.7692	0.9446	3.44
	SMFuse	**0.9428**	**0.8167**	0.6524	2.3793	0.6503	0.7127	0.9004	0.6372	**5.3850**	1.0594	2.22
	MUFusion	0.4420	0.8054	0.5340	0.2191	0.6699	**0.7386**	0.8731	0.5828	2.5328	0.8770	4.11
	UDPNet	0.9268	0.8166	**0.7058**	**2.4823**	**0.7589**	0.7358	**0.9827**	**0.7060**	5.2816	**1.3161**	**1.44**

TABLE III
QUANTITATIVE COMPARISON ON LYTRO, MFFW, AND SIMIF DATASETS. **BOLD: TOP-1**; UNDERLINE: TOP-1 IN GT-FREE METHODS.

Dataset	Metric	GT-Free Method								Supervised Method			
		L1F	SESF	U2Fusion	SDNET	SwinFusion	SMFuse	MUFusion	UDPNet	DCNN	IFCNN	GACN	EAY-Net
Lytro	NMI↑	1.0075	1.1554	0.7963	0.8632	0.9292	1.1837	0.8667	**1.1875**	1.1508	0.9374	1.1668	1.1853
	Q_G↑	0.6923	0.7232	0.5798	0.5958	0.6732	0.7155	**0.7283**	0.7253	0.6634	0.5985	0.7258	0.7271
	Q_M↑	1.4223	2.3982	0.4820	0.5698	1.0381	2.6303	0.4786	**2.6314**	2.3960	0.9484	2.4589	2.5376
	MI↑	7.4280	8.5284	5.6679	6.2169	6.9740	8.8895	6.4804	**8.9180**	8.4749	6.9018	8.6112	8.7483
	Q_{CB}↑	0.7651	0.8045	0.6458	0.6537	0.7186	0.7974	0.6469	0.8076	**0.8082**	0.7297	0.8050	0.8077
	Q_Y↑	0.9630	0.9778	0.8756	0.9011	0.9545	0.9753	0.9076	0.9869	**0.9870**	0.9471	0.9776	0.9781
	ARank↓	7.00	4.83	11.83	10.50	8.33	4.17	10.67	**1.50**	4.00	8.67	4.00	2.50
MFFW	NMI↑	0.8727	1.0874	0.7467	0.8044	0.8451	1.1615	0.8068	1.1662	1.1448	0.8205	1.0820	**1.1777**
	Q_G↑	0.6506	0.6819	0.5374	0.5509	0.6192	0.6593	0.5572	**0.7098**	0.6870	0.5890	0.6722	0.6983
	Q_M↑	1.2079	2.3652	0.4045	0.4968	0.7917	2.5286	0.4291	**2.5385**	2.4807	0.6577	2.4375	2.5238
	MI↑	5.9395	7.3945	4.8760	5.2795	5.9832	8.2776	5.6878	**8.3178**	7.7930	5.5283	7.3923	7.9861
	Q_{CB}↑	0.6803	0.7417	0.5948	0.5861	0.6462	0.6874	0.5884	0.7365	0.7464	0.6420	0.7192	**0.7525**
	Q_Y↑	0.9094	0.9618	0.8063	0.8435	0.8966	0.8722	0.8660	0.9682	0.9797	0.8777	0.9333	**0.9824**
	ARank↓	7.00	4.50	11.67	11.00	7.67	4.67	10.17	2.00	3.17	9.00	5.33	**1.83**
SIMIF	NMI↑	1.1518	1.2856	0.8563	0.9859	1.0290	1.3149	0.9923	**1.3155**	1.2779	1.0382	1.2930	1.3042
	Q_G↑	0.7360	0.7545	0.5880	0.6374	0.6762	0.7498	0.6349	**0.7678**	0.7575	0.6748	0.7568	0.7570
	Q_M↑	1.6100	2.5061	0.4306	0.5789	0.9138	2.6502	0.4822	**2.6531**	2.5216	0.9423	2.5497	2.5669
	MI↑	8.5183	9.5089	6.5455	7.2744	7.6753	9.7666	7.3912	**9.7747**	9.4648	7.7098	9.5533	9.6268
	Q_{CB}↑	0.7915	0.8316	0.6144	0.6782	0.7219	0.8116	0.6769	0.8322	0.8353	0.7484	0.8328	**0.8356**
	Q_Y↑	0.9646	0.9714	0.8476	0.8925	0.9189	0.9616	0.8983	**0.9840**	0.9823	0.9304	0.9724	0.9739
	ARank↓	6.83	5.17	12.00	10.50	8.83	4.17	10.50	**1.50**	3.83	8.17	3.83	2.67

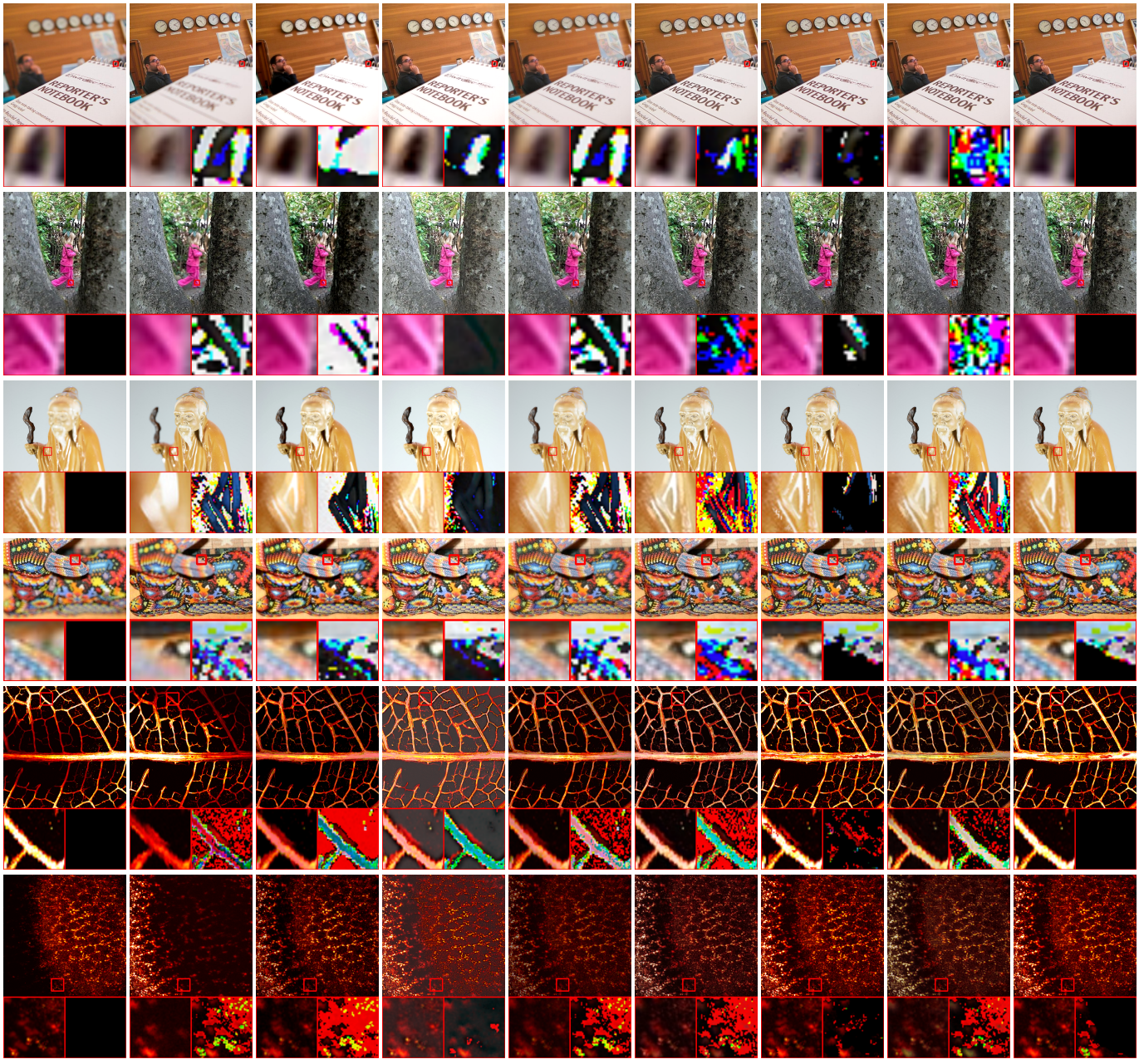

Fig. 4. Fused images and difference images (using Source #1 as reference) by different methods on the samples from Lytro (Row #1-#2), MFFW (Row #3-#4) and OR-PAM (Row #5-#6) datasets (from top to bottom). Below each image, we show a zoomed-in region and its difference from the corresponding focused region of one of the source images.

Qualitative analysis: See Fig. 4 for a qualitative comparison of results from different methods. For better visual inspection, a zoomed-in region of each image, along with its difference compared to the corresponding region of one of the source images, is shown below the image. In a perfect fusion, the difference image should have black segmented regions corresponding to the focused areas of the source image.

Overall, DPNet outperforms other methods in visual quality, such as boundary sharpness, texture detail preservation, and color fidelity. For instance, in the first sample, UDPNet produces sharper boundaries around the notepad, while other methods blur the boundaries (*e.g.*, MFF-GAN and DIF-Net) or introduce artifacts (*e.g.*, SMFuse and MUFusion). In the third sample, UDPNet preserves surrounding texture details and handles central bright stripes better than other methods, which either smooth out the texture (*e.g.*, PMGI, MFF-GAN, and DIF-Net) or produce more and wider bright stripes (*e.g.*, SwinFusion, SMFuse, and MUFusion). In the fifth sample, UDPNet maintains a more consistent color with the source image, while other methods impart a reddish hue to certain regions (*e.g.*, PMGI, MFF-GAN, DIF-Net, and SwinFusion) or cause color bleeding at leaf vein edges (*e.g.*, SwinFusion, SMFuse, and MUFusion).

Additionally, Fig. 6 shows the entire difference images

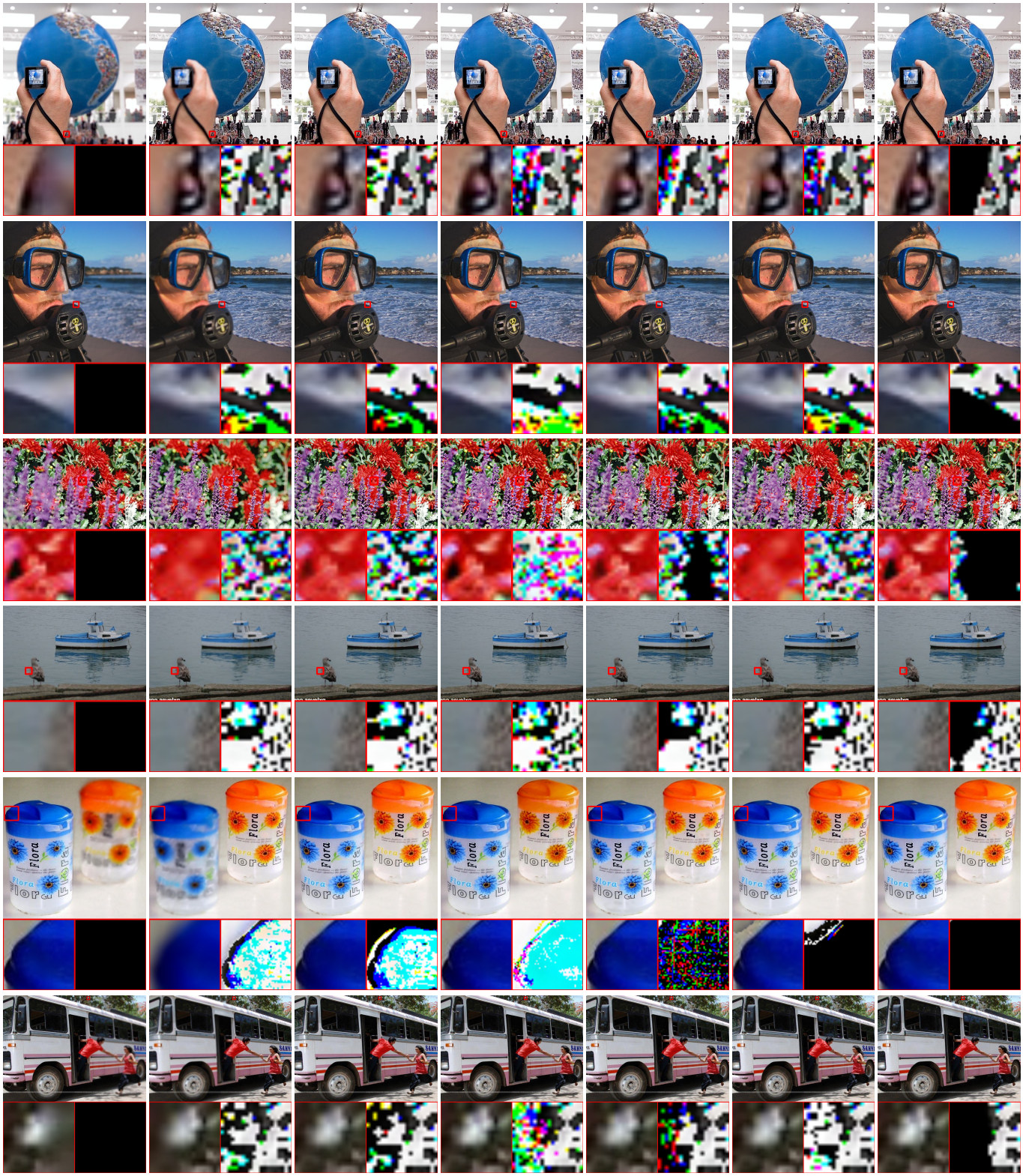

Fig. 5. Fused images and difference images (using Source #1 as reference) by different methods on the samples from on Lytro (Row #1-#2), MFFW (Row #3-#4), and SIMIF (Row #5-#6) datasets (from top to bottom).Below each image, we show a zoomed-in region and its difference from the corresponding focused region of one of the source images.

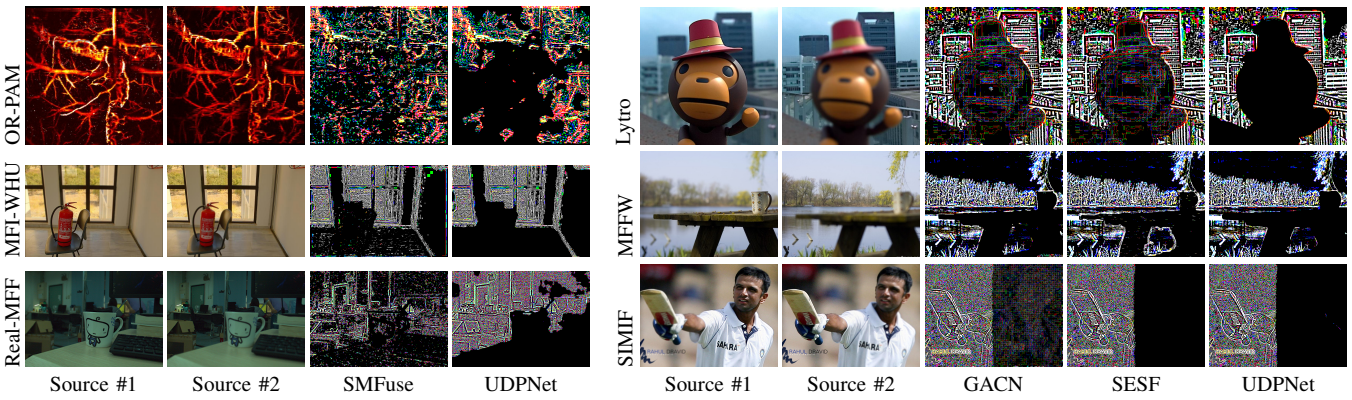

Fig. 6. Difference images (using Source #1 as reference) on images from six datasets.

of UDPNet and its top competitor, SMFuse. Compared to SMFuse, UDPNet more accurately segments focus regions and produces more complete regions in the difference images, demonstrating its superior performance.

B. Comparison to Other Types of Methods

Datasets, metrics and baselines: To avoid potential unfair comparisons due to re-implementation, we quoted the results from existing references whenever possible. Consequently, we followed the configuration used in [28] (a supervised MFIF mehtod). We train on 10000 synthesized source image pairs and test on three datasets: Lytro [71], MFFW [72], and SIMIF (also termed TSAI) [85]. Also, same as [28], we adopted the following six quantitative metrics: NMI [74], MI [76], Q_M [78], Q_Y [81], and Q_{CB} [82]. The ARank over all compared methods is also reported.

Following [28], some recent MFIF methods of distinct types are used for comparison, including L1F [86], SESF [43], U2Fusion [60], SDNET [68], DCNN [14], IFCNN [34], GACN [19], and EAY-Net [28]. Among them, L1F is a non-learning method, SESF uses self pre-training, U2Fusion uses end-to-end unsupervised DL, and the others are supervised methods. The results of these methods are directly quoted from [28]. In addition, we compare with SwinFusion, SMFuse and MUFusion, the three top competitors of unsupervised methods in the previous experiment. SwinFusion and MUFusion are retrained using the same data as ours. As for SMFuse, we call its published model to obtain results that are better than the re-trained one.

Quantitative analysis: See Table III for the quantitative results. Our UDPNet achieves the best ARank on both Lytro and SIMIF, and second best on MFFW. Particularly, it performs the best in terms of four metrics on Lytro. These results demonstrate the competitive performance of our approach over supervised methods. Compared to other GT-free methods, UDPNet performs slightly worse than SESF in only one metric on MFFW, but its ARank scores are noticeably higher, indicating that UDPNet is better at winning the trade-off among various quality aspects reflected by different metrics.

Qualitative analysis: The qualitative comparison shown in Fig. 5 demonstrates the advantages of UDPNet over both supervised and self-supervised methods in terms of visual quality. Compared to the supervised methods DCNN, GACN, and IFCNN, UDPNet produces better results in terms of boundary sharpness and texture detail preservation. For instance, the boundary between the hand and the background in the first sample is sharper with UDPNet than with DCNN, GACN, and IFCNN. Additionally, the texture detail of the red petals in the third sample is better preserved with UDPNet than with DCNN and GACN.

Compared to the results from self-supervised pre-training-based SESF, the results from UDPNet have less artifacts and better preserve texture details. For instance, in the fifth sample, artifacts show in the upper left part of the cup in the result of SESF, and the texture detail of the red petals is removed by SESF in the third sample. See also Fig. 6 for the difference images. Again, UDPNet accurately predicts the focused foreground with more regions, while the other methods miss some parts.

C. Ablation Studies

The following baseline methods are constructed from UDPNet, with their results on Real-MFF and MFFW listed in Tables IV and V, respectively. (a) w/o $\mathcal{F}$: disabling the mask predictor by fixing its output to be the initial masks; (b) w/o unrolling: replacing the unrolling NN by a CNN of similar size, using the structure of the image estimator; (c) w/o BS: replacing $\mathcal{L}^{\mathcal{F}}_{\text{mask}}$ by a common self-reconstruction loss $\|\mathcal{F}(\widetilde{\boldsymbol{M}}_k) - \widetilde{\boldsymbol{M}}_k\|_{\text{F}}^2$ without blind-sport operations; (d) w/o $\mathcal{L}^{\mathcal{G}}_{\text{mask}}$: disabling $\mathcal{L}^{\mathcal{G}}_{\text{mask}}$ by using initial masks for supervision; (e) w/o $\mathcal{G}$: abandoning the image fuser in both training and testing, training mask predictor only using the blind-spot loss, and performs fusion using the refined masks from the mask predictor; (f) w/o $\mathcal{G}^*$: disarding the image fuser in testing while keep it when training the entire UDPNet, and performs fusion using the same way as "w/o $\mathcal{G}$" during inference; and (g) w/o $\mathcal{L}^{\mathcal{F},\mathcal{G}}_{\text{fusion}}$: removing the fusion consistency loss.

We have the following observations from the results. (a) The mask predictor plays an important role for the performance of the image fuser. Without refinement and supervision by the mask predictor, significant performance drop is observed, demonstrating the necessity of our dual-path framework. (b) The unrolling architecture is important to the image fuser, with substantial performance contribution due to the regularization

TABLE IV
ABLATION STUDY ON REAL-MFF DATASET.

Metric	w/o $\mathcal{F}$	w/o unrolling	w/o BS	w/o $\mathcal{L}_{\text{mask}}^{\mathcal{G}}$	w/o $\mathcal{G}$	w/o $\mathcal{G}^*$	w/o $\mathcal{L}_{\text{fusion}}^{\mathcal{F},\mathcal{G}}$	UDPNet
PSNR	34.07	36.14	37.58	34.05	39.71	39.89	39.82	**40.05**
SSIM	0.955	0.976	0.979	0.956	0.982	0.983	0.986	**0.987**
LPIPS	0.045	0.029	0.016	0.048	0.013	0.013	0.013	**0.012**

TABLE V
ABLATION STUDY ON MFFW DATASET.

Metric	w/o $\mathcal{F}$	w/o unrolling	w/o BS	w/o $\mathcal{L}_{\text{mask}}^{\mathcal{G}}$	w/o $\mathcal{G}$	w/o $\mathcal{G}^*$	w/o $\mathcal{L}_{\text{fusion}}^{\mathcal{F},\mathcal{G}}$	UDPNet
NMI↑	1.1568	0.7851	1.0664	1.1501	1.1064	1.1412	0.7954	**1.1627**
Q_{NCIE}↑	0.8395	0.8202	0.8334	0.8389	0.8366	0.8381	0.8206	**0.8400**
Q_G↑	0.7078	0.5102	0.5049	0.7066	0.6208	0.7017	0.4910	**0.7089**
Q_M↑	**2.5264**	0.3929	1.5962	2.5015	2.3322	2.4861	0.3453	2.5040
Q_P↑	0.7571	0.5973	0.5094	**0.7590**	0.6027	0.7530	0.6044	0.7574
Q_C↑	0.7799	0.6737	0.6407	0.7800	0.6929	0.7781	0.7267	**0.7815**
Q_Y↑	0.9791	0.7879	0.7128	0.9797	0.8037	0.9739	0.7863	**0.9812**
Q_{CB}↑	0.7502	0.6467	0.6159	0.7506	0.6542	0.7447	0.5729	**0.7548**
MI↑	8.2385	5.5169	7.6308	8.1906	7.9115	8.1297	5.6083	**8.2803**
VIF↑	1.2740	0.9229	0.8453	**1.2764**	1.0613	1.2755	0.9467	1.2732
ARank↓	2.40	7.00	7.00	2.30	5.20	3.80	6.80	**1.50**

from its interpretable structure. (c) The blind-spot training scheme is critical. Without it, the mask predictor performs poorly, leading to noticeable performance loss, probably due to that the mask predictor tends to learn an identity mapping. (d) The supervision from mask predictor is useful. (e) Without the unrolling NN, the fusion results becomes noticeably worse. Though the mask predictor improves initial masks, potential overfitting remains due to the lack of GT in training. (f) Introducing the unrolling NN into training allows interactions between mask refinement and image sharpening, bringing implicit regularization to the whole UDPNet. As a result, superior performance of "w/o $\mathcal{G}^*$" over "w/o $\mathcal{G}$" is observed. However, $\mathcal{G}^*$" performs worse than UDPNet, verifying the higher quality of the masks predicted by $\mathcal{G}$ over that by $\mathcal{F}$, which is probably due to the higher interoperability and generalizability of $\mathcal{G}$. (g) The fusion consistency benefits further performance gain, though not big. This is probably because the supervision from mask predictor provides certain alignment between two paths.

To demonstrate the effectiveness of the blind-spot mechanism, we replace the blind-spot refiner with other denoising methods, including pre-trained segmentation NN, denoising NN, and corrosion expansion. (a) Segmented by SAM: We input an out-of-focus image into Segment Anything Model (SAM) [87] pre-trained segmentation NN to obtain the segmented regions, and determine whether each region is clear or not by the initial mask. (b) Restormer-based denoising: The initial mask is fed into the pre-trained denoiser Restormer [88] to get the denoising result. (c) Opening/Closing: The initial mask is subjected to opening and closing operations with different window sizes. See Fig. 7 for the results, from which we observe that: (a) Objects frequently exhibit both defocused and in-focus regions, and image segmentation often struggles to distinguish between them. (b) General image denoisers are ineffective as the errors in initial masks possess statistical

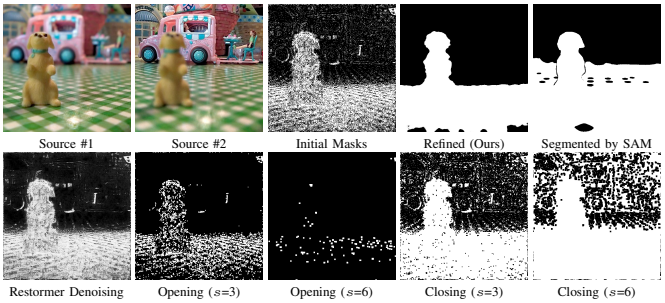

Fig. 7. Masks from various methods. s: window size to operate.

properties different from typical measurement noise, and they overlook inherent properties of masks. (c) Morphological filtering like opening or closing also falls short; they fail to adequately fill small holes or tend to over-expand regions, with sensitivity to the window size setting.

V. CONCLUSION

We presented an effective unsupervised end-to-end learning approach for MFIF. Different from the existing works along this line, our approach foregoes the direct similarity loss between fused and source images. Instead, it leverages a dual-path learning framework with novel loss functions. The dual-path framework is composed of a self-supervised mask predictor trained with a blind-spot scheme and a deep unrolling-based image fuser supervised by the mask predictor. A fusion consistency loss is introduced to enhance the interaction and alignment between two paths during training. Our approach achieved desired results in extensive experiments.

The performance gained by our unrolling network highlights the importance of introducing interpretable designs for MFIF. Our proposed blind-spot learning technique demonstrates the potential of self-supervised denoising techniques in unsupervised MFIF. Moving forward, we aim to further optimize both the unrolling network architecture and the self-supervised denoising mechanism for additional performance improvement.

Our approach, involving mask prediction within the unrolling and unsupervised learning scheme, is specifically tailored for the interaction of multiple measurements in MFIF. This makes it not directly applicable to other image fusion tasks with different interaction schemes. In future, we will also extend our unsupervised learning approach to work with alternative interaction schemes for various image fusion tasks, such as multi-exposure fusion and multi-modal image fusion.

REFERENCES

[1] S. Sun, Y. Wang, X. Lu, J. Sun, and L. Xu, "Multi-focus image reconstruction and fusion for 3d flow visualization using an optimized four-plane ect sensor," *IEEE Transactions on Instrumentation and Measurement*, 2023.
[2] W. Zhou, J. He, Y. Li, Z. Sun, J. Chen, L. Wang, H. Hui, and X. Chen, "Multi-focus image fusion with enhancement filtering for robust vascular quantification using photoacoustic microscopy," *Optics Letters*, vol. 47, no. 15, pp. 3732–3735, 2022.
[3] S. Bhat and D. Koundal, "Multi-focus image fusion techniques: a survey," *Artificial Intelligence Review*, vol. 54, pp. 5735–5787, 2021.
[4] Y. Liu, L. Wang, J. Cheng, C. Li, and X. Chen, "Multi-focus image fusion: A survey of the state of the art," *Information Fusion*, vol. 64, pp. 71–91, 2020.

[5] X. Zhang, "Deep learning-based multi-focus image fusion: A survey and a comparative study," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 9, pp. 4819–4838, 2022.

[6] Q. Zhang, Y. Liu, R. S. Blum, J. Han, and D. Tao, "Sparse representation based multi-sensor image fusion for multi-focus and multi-modality images: A review," *Information Fusion*, vol. 40, pp. 57–75, 2018.

[7] M. S. Farid, A. Mahmood, and S. A. Al-Maadeed, "Multi-focus image fusion using content adaptive blurring," *Information fusion*, vol. 45, pp. 96–112, 2019.

[8] B. Xiao, G. Ou, H. Tang, X. Bi, and W. Li, "Multi-focus image fusion by hessian matrix based decomposition," *IEEE Transactions on Multimedia*, vol. 22, no. 2, pp. 285–297, 2020.

[9] S. Liu, J. Chen, and S. Rahardja, "A new multi-focus image fusion algorithm and its efficient implementation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 5, pp. 1374–1384, 2020.

[10] J. Wang, H. Qu, Y. Wei, M. Xie, J. Xu, and Z. Zhang, "Multi-focus image fusion based on quad-tree decomposition and edge-weighted focus measure," *Signal Processing*, vol. 198, p. 108590, 2022.

[11] Q. Zhang, G. Li, Y. Cao, and J. Han, "Multi-focus image fusion based on non-negative sparse representation and patch-level consistency rectification," *Pattern Recognition*, vol. 104, p. 107325, 2020.

[12] Q. Zhang, F. Wang, Y. Luo, and J. Han, "Exploring a unified low rank representation for multi-focus image fusion," *Pattern Recognition*, vol. 113, p. 107752, 2021.

[13] J. Chen, X. Li, L. Luo, and J. Ma, "Multi-focus image fusion based on multi-scale gradients and image matting," *IEEE Transactions on Multimedia*, vol. 24, pp. 655–667, 2022.

[14] Y. Liu, X. Chen, H. Peng, and Z. Wang, "Multi-focus image fusion with a deep convolutional neural network," *Information Fusion*, vol. 36, pp. 191–207, 2017.

[15] H. Tang, B. Xiao, W. Li, and G. Wang, "Pixel convolutional neural network for multi-focus image fusion," *Information Sciences*, vol. 433, pp. 125–141, 2018.

[16] X. Guo, R. Nie, J. Cao, D. Zhou, L. Mei, and K. He, "Fusegan: Learning to fuse multi-focus image via conditional generative adversarial network," *IEEE Transactions on Multimedia*, vol. 21, no. 8, pp. 1982–1996, 2019.

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

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

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

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

[21] Y. Wang, S. Xu, J. Liu, Z. Zhao, C. Zhang, and J. Zhang, "Mfif-gan: A new generative adversarial network for multi-focus image fusion," *Signal Processing: Image Communication*, vol. 96, p. 116295, 2021.

[22] Y. Liu, L. Wang, J. Cheng, and X. Chen, "Multiscale feature interactive network for multifocus image fusion," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–16, 2021.

[23] B. Xiao, H. Wu, and X. Bi, "Dtmnet: a discrete tchebichef moments-based deep neural network for multi-focus image fusion," in *Proceedings of IEEE/CVF International Conference on Computer Vision*, 2021, pp. 43–51.

[24] B. Xiao, B. Xu, X. Bi, and W. Li, "Global-feature encoding u-net (geu-net) for multi-focus image fusion," *IEEE Transactions on Image Processing*, vol. 30, pp. 163–175, 2021.

[25] X. Nie, B. Hu, and X. Gao, "Mlnet: A multi-domain lightweight network for multi-focus image fusion," *IEEE Transactions on Multimedia*, 2022.

[26] Z. Duan, X. Luo, and T. Zhang, "Multi-focus image fusion via gradient guidance progressive network," in *Proceedings of IEEE International Conference on Multimedia and Expo*. IEEE, 2023, pp. 2159–2164.

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

[28] Z. Wang, X. Li, L. Zhao, H. Duan, S. Wang, H. Liu, and X. Zhang, "When multi-focus image fusion networks meet traditional edge-preservation technology," *International Journal of Computer Vision*, pp. 1–24, 2023.

[29] P. Wu, L. Jiang, Z. Hua, and J. Li, "Multi-focus image fusion: Transformer and shallow feature attention matters," *Displays*, vol. 76, p. 102353, 2023.

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

[31] Z. Duan, X. Luo, and T. Zhang, "Combining transformers with cnn for multi-focus image fusion," *Expert Systems with Applications*, vol. 235, p. 121156, 2024.

[32] W. Zhao, D. Wang, and H. Lu, "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*, vol. 29, no. 4, pp. 1102–1115, 2019.

[33] H. Li, R. Nie, J. Cao, X. Guo, D. Zhou, and K. He, "Multi-focus image fusion using u-shaped networks with a hybrid objective," *IEEE Sensors Journal*, vol. 19, no. 21, pp. 9755–9765, 2019.

[34] Y. Zhang, Y. Liu, P. Sun, H. Yan, X. Zhao, and L. Zhang, "Ifcnn: A general image fusion framework based on convolutional neural network," *Information Fusion*, vol. 54, pp. 99–118, 2020.

[35] C. Wang, D. Zhou, Y. Zang, R. Nie, and Y. Guo, "A deep and supervised atrous convolutional model for multi-focus image fusion," *IEEE Sensors Journal*, vol. 21, no. 20, pp. 23 069–23 084, 2021.

[36] I. Marivani, E. Tsiligianni, B. Cornelis, and N. Deligiannis, "Designing cnns for multimodal image restoration and fusion via unfolding the method of multipliers," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 9, pp. 5830–5845, 2022.

[37] J. Liu, S. Li, H. Liu, R. Dian, and X. Wei, "A lightweight pixel-level unified image fusion network," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[38] M. Li, R. Pei, T. Zheng, Y. Zhang, and W. Fu, "Fusiondiff: Multi-focus image fusion using denoising diffusion probabilistic models," *Expert Systems with Applications*, p. 121664, 2023.

[39] O. Bouzos, I. Andreadis, and N. Mitianoudis, "A convolutional neural network-based conditional random field model for structured multi-focus image fusion robust to noise," *IEEE Transactions on Image Processing*, 2023.

[40] H. Li and X.-J. Wu, "Densefuse: A fusion approach to infrared and visible images," *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2614–2623, 2019.

[41] H. T. Mustafa, J. Yang, and M. Zareapoor, "Multi-scale convolutional neural network for multi-focus image fusion," *Image and Vision Computing*, vol. 85, pp. 26–35, 2019.

[42] X. Luo, Y. Gao, A. Wang, Z. Zhang, and X.-J. Wu, "Ifsepr: A general framework for image fusion based on separate representation learning," *IEEE Transactions on Multimedia*, 2023.

[43] B. Ma, Y. Zhu, X. Yin, X. Ban, H. Huang, and M. Mukeshimana, "Sesf-fuse: An unsupervised deep model for multi-focus image fusion," *Neural Computing and Applications*, vol. 33, pp. 5793–5804, 2021.

[44] J. Zhang, J. Shao, J. Chen, D. Yang, and B. Liang, "Polarization image fusion with self-learned fusion strategy," *Pattern Recognition*, vol. 118, p. 108045, 2021.

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

[46] K. Wu and Y. Mei, "Multi-focus image fusion based on unsupervised learning," *Machine Vision and Applications*, vol. 33, no. 5, p. 75, 2022.

[47] P. Liang, J. Jiang, X. Liu, and J. Ma, "Fusion from decomposition: A self-supervised decomposition approach for image fusion," in *Proceedings of European Conference on Computer Vision*. Springer, 2022, pp. 719–735.

[48] Z. Wang, X. Li, H. Duan, and X. Zhang, "A self-supervised residual feature learning model for multifocus image fusion," *IEEE Transactions on Image Processing*, vol. 31, pp. 4527–4542, 2022.

[49] J. Liu, S. Li, R. Dian, and Z. Song, "Focus relationship perception for unsupervised multi-focus image fusion," *IEEE Transactions on Multimedia*, 2023.

[50] X. Jin, X. Xi, D. Zhou, X. Ren, J. Yang, and Q. Jiang, "An unsupervised multi-focus image fusion method based on transformer and u-net," *IET Image Processing*, vol. 17, no. 3, pp. 733–746, 2023.

[51] X. Hu, J. Jiang, X. Liu, and J. Ma, "Zero-shot multi-focus image fusion," in *Proceedings of IEEE International Conference on Multimedia and Expo*. IEEE, 2021, pp. 1–6.

[52] ——, "Zmff: Zero-shot multi-focus image fusion," *Information Fusion*, vol. 92, pp. 127–138, 2023.

[53] X. Yan, S. Z. Gilani, H. Qin, and A. Mian, "Unsupervised deep multi-focus image fusion," *arXiv preprint arXiv:1806.07272*, 2018.

[54] H. Zhang, H. Xu, Y. Xiao, X. Guo, and J. Ma, "Rethinking the image fusion: A fast unified image fusion network based on proportional maintenance of gradient and intensity," in *Proceedings of AAAI conference on artificial intelligence*, vol. 34, no. 07, 2020, pp. 12 797–12 804.

[55] H. Jung, Y. Kim, H. Jang, N. Ha, and K. Sohn, "Unsupervised deep image fusion with structure tensor representations," *IEEE Transactions on Image Processing*, vol. 29, pp. 3845–3858, 2020.

[56] H. Xu, J. Ma, Z. Le, J. Jiang, and X. Guo, "Fusiondn: A unified densely connected network for image fusion," in *Proceedings of AAAI conference on artificial intelligence*, vol. 34, no. 07, 2020, pp. 12 484–12 491.

[57] J. Ma, Z. Le, X. Tian, and J. Jiang, "Smfuse: Multi-focus image fusion via self-supervised mask-optimization," *IEEE Transactions on Computational Imaging*, vol. 7, pp. 309–320, 2021.

[58] Z. Duan, T. Zhang, X. Luo, and J. Tan, "Dckn: multi-focus image fusion via dynamic convolutional kernel network," *Signal Processing*, vol. 189, p. 108282, 2021.

[59] J. Ma, L. Tang, F. Fan, J. Huang, X. Mei, and Y. Ma, "Swinfusion: Cross-domain long-range learning for general image fusion via swin transformer," *IEEE/CAA Journal of Automatica Sinica*, vol. 9, no. 7, pp. 1200–1217, 2022.

[60] H. Xu, J. Ma, J. Jiang, X. Guo, and H. Ling, "U2fusion: A unified unsupervised image fusion network," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 1, pp. 502–518, 2022.

[61] A. Krull, T.-O. Buchholz, and F. Jug, "Noise2void-learning denoising from single noisy images," in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2129–2137.

[62] J. Batson and L. Royer, "Noise2self: Blind denoising by self-supervision," in *Proceedings of International Conference on Machine Learning*. PMLR, 2019, pp. 524–533.

[63] O. Bouzos, I. Andreadis, and N. Mitianoudis, "Conditional random field model for robust multi-focus image fusion," *IEEE Transactions on Image Processing*, vol. 28, no. 11, pp. 5636–5648, 2019.

[64] Y. Liu, L. Wang, H. Li, and X. Chen, "Multi-focus image fusion with deep residual learning and focus property detection," *Information Fusion*, vol. 86, pp. 1–16, 2022.

[65] J. Zhang, Q. Liao, H. Ma, J.-H. Xue, W. Yang, and S. Liu, "Exploit the best of both end-to-end and map-based methods for multi-focus image fusion," *IEEE Transactions on Multimedia*, 2024.

[66] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Proceedings of European Conference on Computer Vision*. Springer, 2014, pp. 740–755.

[67] H. Zhang, Z. Le, Z. Shao, H. Xu, and J. Ma, "Mff-gan: An unsupervised generative adversarial network with adaptive and gradient joint constraints for multi-focus image fusion," *Information Fusion*, vol. 66, pp. 40–53, 2021.

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

[69] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Deep image prior," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 9446–9454.

[70] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7794–7803.

[71] M. Nejati, S. Samavi, and S. Shirani, "Multi-focus image fusion using dictionary-based sparse representation," *Information Fusion*, vol. 25, pp. 72–84, 2015.

[72] S. Xu, X. Wei, C. Zhang, J. Liu, and J. Zhang, "Mffw: A new dataset for multi-focus image fusion," *arXiv preprint arXiv:2002.04780*, 2020.

[73] J. Zhang, Q. Liao, S. Liu, H. Ma, W. Yang, and J.-H. Xue, "Real-mff: A large realistic multi-focus image dataset with ground truth," *Pattern Recognition Letters*, vol. 138, pp. 370–377, 2020.

[74] M. Hossny, S. Nahavandi, and D. Creighton, "Comments on 'information measure for performance of image fusion'," *Electronics letters*, vol. 44, no. 18, pp. 1066–1067, 2008.

[75] Q. Wang, Y. Shen, and J. Jin, "Performance evaluation of image fusion techniques," *Image fusion: algorithms and applications*, vol. 19, pp. 469–492, 2008.

[76] G. Qu, D. Zhang, and P. Yan, "Information measure for performance of image fusion," *Electronics letters*, vol. 38, no. 7, p. 1, 2002.

[77] C. S. Xydeas, V. Petrovic *et al.*, "Objective image fusion performance measure," *Electronics letters*, vol. 36, no. 4, pp. 308–309, 2000.

[78] P.-w. Wang and B. Liu, "A novel image fusion metric based on multi-scale analysis," in *Proceedings of International Conference on Signal Processing*. IEEE, 2008, pp. 965–968.

[79] J. Zhao, R. Laganiere, and Z. Liu, "Performance assessment of combinative pixel-level image fusion based on an absolute feature measurement," *International Journal of Innovative Computing Information and Control*, vol. 3, no. 6, pp. 1433–1447, 2007.

[80] N. Cvejic, A. Loza, D. Bull, and N. Canagarajah, "A similarity metric for assessment of image fusion algorithms," *International Journal of Signal Processing*, vol. 2, no. 3, pp. 178–182, 2005.

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

[82] Y. Chen and R. S. Blum, "A new automated quality assessment algorithm for image fusion," *Image and vision computing*, vol. 27, no. 10, pp. 1421–1432, 2009.

[83] Y. Han, Y. Cai, Y. Cao, and X. Xu, "A new image fusion performance metric based on visual information fidelity," *Information fusion*, vol. 14, no. 2, pp. 127–135, 2013.

[84] C. Cheng, T. Xu, and X.-J. Wu, "Mufusion: A general unsupervised image fusion network based on memory unit," *Information Fusion*, vol. 92, pp. 80–92, 2023.

[85] C.-C. Tsai, "Standard images for multifocus image fusion," 2023, https://www.mathworks.com/matlabcentral/fileexchange/45992-standard-images-for-multifocus-image-fusion.

[86] S. Yu, X. Li, M. Ma, X. Zhang, and S. Chen, "Multi-focus image fusion based on l1 image transform," *Multimedia Tools and Applications*, vol. 80, pp. 5673–5700, 2021.

[87] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4015–4026.

[88] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Restormer: Efficient transformer for high-resolution image restoration," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 5728–5739.