IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 6 December 2022, accepted 11 December 2022, date of publication 14 December 2022,
date of current version 20 December 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3229056

RESEARCH ARTICLE

Guided Image Deblurring by Deep Multi-Modal Image Fusion

YUQI LIU[1], ZEHUA SHENG[1], AND HUI-LIANG SHEN[1,2], (Member, IEEE)
[1]College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310027, China
[2]Key Laboratory of Collaborative Sensing and Autonomous Unmanned Systems of Zhejiang Province, Hangzhou 310015, China
Corresponding author: Hui-Liang Shen (shenhl@zju.edu.cn)

ABSTRACT Estimating sharp images from blurry observations is still a difficult task in the image processing research field. Previous works may produce deblurred images that lose details or contain artifacts. To deal with this problem, a feasible solution is to seek the help of additional images, such as the near-infrared image and the flashlight image, *etc*. In this paper, we propose a fusion framework for image deblurring, called Guided Deblurring Fusion Network (GDFNet), to integrate the multi-modal information for better image deblurring performance. Unlike previous works that directly compute a deblurred image using paired multi-modal degraded and guidance images, GDFNet employs image fusion techniques to obtain a deblurred image. GDFNet can combine the advantages by fusing the pre-deblurred streams of single and guided image deblurring using convolutional neural network (CNN). We adopt a blur/residual image splitting strategy by fusing the residual images to enhance the representation ability of encoders and preserve details. We employ a 2-level coarse-to-fine reconstruction strategy to improve the fusion and deblurring performance by supervising its multi-scale output. Quantitative comparisons on multi-modal image datasets show that our GDFNet can recover correct structures and produce fewer artifacts while preserving more details. The peak signal-to-noise ratio (PSNR) of GDFNet evaluated on the blurry/flash dataset is at least 0.9 dB higher than the compared algorithms.

INDEX TERMS Blind image deblurring, guided image deblurring, image deblurring, image fusion, multi-modal image fusion.

I. INTRODUCTION

Image deblurring aims to recover sharp images from their blurred observations degraded by camera/object movements or lens defocus. Under the uniform blur assumption, the blurry image $\mathbf{I}$ can be modeled as the convolution of the sharp latent image $\mathbf{F}$ and a point spread function (PSF) $\mathbf{k}$,

$$\mathbf{I} = \mathbf{F} * \mathbf{k} + \mathbf{n}, \tag{1}$$

where $\mathbf{n}$ represents the noise and $*$ denotes the convolution operator. Generally, the blurring degradation can be categorized into out-of-focus blur and motion blur. Out-of-focus blur occurs when the image plane is away from the ideal reference plane. Motion blur is caused by relative motions between the scene and camera during exposure.

The associate editor coordinating the review of this manuscript and approving it for publication was Arianna Dulizia.

Actually, restoring the sharp image based on one single blurry image is severely ill-posed as it needs to estimate both the point spread function $\mathbf{k}$ and the sharp image $\mathbf{F}$ simultaneously. The results of single blind image deblurring algorithms, such as multi-input multi-output U-Net (MIMO) [1], multi-stage progressively restoration network (MPR) [2], and a deep neural network embedded with residual Fourier transform (DeepRFT) [3], usually lose details during the process of removing motion blur. Guided image deblurring algorithms handle deblurring with the help of the additional structure and texture information provided by different wavelength sensors or camera settings. For example, near-infrared (NIR) sensors are highly adaptable to thick fog and darkness due to different wavelength sensitivity, and flashlight imaging captures a clear picture by changing the environment illumination. Previous work has used aligned multi-modal image pairs such as flash/no-flash image pairs [4], [5], RGB/NIR

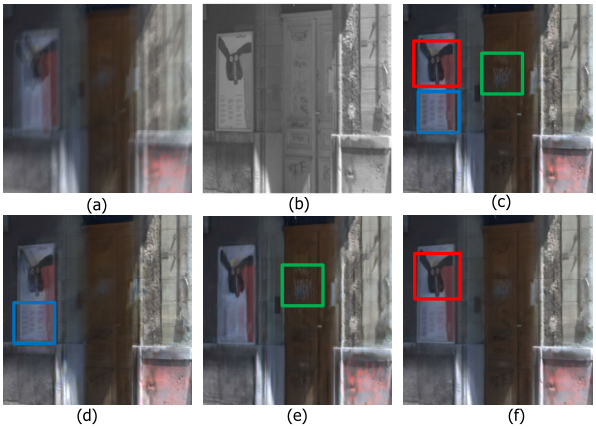

FIGURE 1. A deblurring example of the proposed method. (a) Input blurry RGB image. (b) Input guidance NIR image. (c) The fused deblurred image. (d)-(f) The pre-deblurred images. The best parts of the pre-deblurred images marked with color boxes in (d)-(f) are fused into (f) using GDFNet.

image pairs [6], and blurry/noisy image pairs [7] to relax the illness of the blind image deblurring problem. However, the original visible information like pixel intensity and texture is inaccurate in the guidance images due to structural inconsistency caused by the noise in NIR imaging [8], the reflectance differences, and the object movements. It is the reason why integrating information from guidance images to degraded images sometimes produces artifacts or halos. This observation inspires us to deblur images in a fusion manner.

In this work, we propose a deep fusion network, called Guided Deblurring Fusion Network (GDFNet), to perform joint image deblurring by fusing the pre-deblurred images obtained by multiple image deblurring streams. It is motivated by the fact that the single image deblurring stream cannot effectively recover detailed contents while the guided deblurring stream produces incorrect low-frequency content due to structural inconsistency. In comparison, our proposed GDFNet can effectively address these two problems in an image fusion mechanism. Enlighten by the residual learning [9] and the frequency principle [10], we embed a blur/residual image splitting strategy in GDFNet to estimate the fusion weights of residual images to enhance the representation ability of encoders. We use a coarse-to-fine reconstruction strategy to generate finer fusion weight maps by training the network using multi-scale supervision. Experimental results show that our GDFNet outperforms the competitors including blind deblurring algorithms, cascaded algorithms, and other fusion networks on multi-modal datasets. In summary, the main contributions of this work are as follows:

- We propose a deep fusion framework GDFNet to deal with image deblurring by fusing the pre-deblurred streams of single and guided image deblurring using a multi-modal image pair as input.
- We employ a blur/residual splitting strategy to fuse the pre-deblurred residual images to enhance the representation ability with a coarse-to-fine reconstruction structure trained using multi-scale supervision to improve the deblurring performance by generating finer fusion weights.
- We experimentally show that the GDFNet can fuse single and guided image deblurring streams, and outperforms the existing deblurring algorithms and fusion approaches on multi-modal datasets.

The organization of the paper is as follows. We review the related work in Section II. Section III presents the motivation, framework and details of the proposed method. Section IV illustrates the experimental results, and finally Section V concludes the work.

II. RELATED WORK

In this section, we provide a brief review of the work related to image deblurring and image fusion.

A. IMAGE DEBLURRING

Image deblurring techniques can be coarsely classified into two categories: single image deblurring and guided image deblurring. Single blind image deblurring refers to restoring the latent image from its degraded observation. Early works usually employ an alternative framework to estimate the blur kernel and latent image iteratively based on natural image priors. A heavy-tailed distribution on image gradients regularizes the iterative optimization for deblurring [11]. This regularization term is further improved by fitting the distribution using a hyper-Laplacian function [12]. The work [13] fits the logarithmic density of image gradients by concatenating two piece-wise continuous functions as a prior. L0 distribution is used to approximate the image gradients and intensity in [14] and [15]. Later, Pan et al. [16] and Yan et al. [17] define the dark channel and extreme channel and apply L0 sparse constraint on the intensities. Bai et al. [18] use coarse-to-fine priors and recover the latent image using a multi-scale image pyramid. Levin et al. [19] modify a conventional lens by inserting a patterned disc into the aperture to produce a characteristic distribution of image frequencies that is very sensitive to defocus blur. Zhang et al. [20] estimate the sharp image using multiple blurry observations with a coupled sparse prior.

However, the recovered latent image can be visually poor when the kernel estimation is inaccurate. Recently, CNN has been widely applied in image processing and computer vision tasks. A cascaded network is adopted to estimate the latent image and blur kernel iteratively in [21]. Two generative networks are used to capture the blur kernel and the latent image in [22]. The work [23] uses a scale-recurrent network that shares network weights across scales. Min et al. decompose the low- and high-frequency information using wavelet transform followed by a recursive convolutional neural network to deblur. Further, a Multi-Input Multi-Output (MIMO) U-Net [1] is presented to deblur images in a coarse-to-fine strategy. Ople et al. [24] extract multi-scale features using dilated convolutions with different dilated rates.

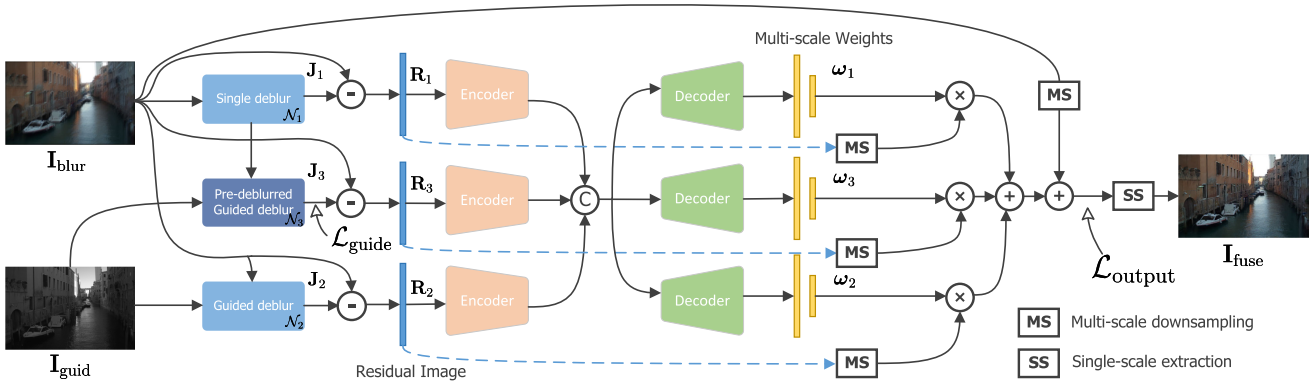

FIGURE 2. The architecture of our proposed Guided Deblurring Fusion Network (GDFNet) using RGB/NIR image pairs as an example. First, the blurry RGB image and the sharp NIR guidance image are fed into three pre-deblurring streams to predict three pre-deblurred images. Then the residual images are computed by subtracting the pre-deblurred image from the blurry RGB image. Three individual encoders are used to extract the features of these residual images. Then the features of three streams are concatenated, followed by two convolutional blocks to aggregate. We use three decoders to estimate the fusion weights of three residual images of the pre-deblurred images in a coarse-to-fine scheme. Finally, the composite residual images are added to the original blurry input to generate the final fused deblurred result.

Liu et al. [25] refine the optimization based deblurred results using an encoder-decoder network. A multi-stage architecture called MPRNet progressively learns restoration functions for the degraded inputs in [2]. A residual fast Fourier transform with convolution block is introduced in DeepRFT [3] to integrate both low- and high-frequency residual information. Saqlain et al. [26] introduce a generative adversarial network (GAN) based approach called DeblurFusedGAN (DFGAN) that fuses a lightweight attention (LSA) mechanism and gradient-based filters in the generator work. Wang et al. [27] introduce Uformer, a transformer-based architecture for image deblurring. It uses a locally-enhanced window (LeWin) transformer block and a learnable multi-scale restoration modulator to capture both local and global dependencies. Tsai et al. [28] construct a transformer-based architecture using intra- and inter-strip tokens to catch blurred patterns with different orientations. Chen et al. [29] introduce an efficient Nonlinear Activation Free Network (NAFNet) that lowers the computational cost and removes unnecessary activation functions. Chu et al. [30] investigate the distribution differences in the features between training and inference and introduce a Test-time Local Converter (TLC).

Guided image deblurring algorithms introduce additional information from the guidance image to facilitate image deblurring. The paired images used in image deblurring are multi-modal, such as RGB/NIR [31], and blurry/flash [32], [33]. However, extraneous artifacts could appear when the guidance and input images are captured in different spectrums or have inconsistent structures. In the pioneering work [4], a robust flash gradient constraint is introduced to solve the flash deblurring problem by performing kernel estimation and non-blind deconvolution iteratively. Guided filtering [32] can be applied to deblur images by calculating the local linear model between two inputs. Further, a CNN-based joint filtering algorithm is designed to deblur images by estimating the coefficients of the spatially variant linear representation model (SVLRM) [5].

B. IMAGE FUSION

Image fusion can be roughly separated into three tasks: feature extraction, fusion rules, and feature reconstruction. Traditional fusion algorithms use domain transform approaches like wavelet transform, Laplacian pyramid decomposition, and guided filtering as feature extraction components. Recently, fusion algorithms based on deep learning have been introduced to improve the ability of feature representation like DenseFuse [34], or directly fuse images in an end-to-end manner [35]. Zhou et al. [36] introduced a fusion algorithm that fuses infrared and visible using L_0 filter, the weighted least squares (WLS) filter, and parallel gradient fusion called target-aware decomposition and parallel gradient fusion (TAD-PGF). U2Fusion [37] solves different fusion problems using one fusion network in an unsupervised manner. In [38], a pair of infrared and visible images are used to fuse the obvious object information based on multi-level Gaussian curvature filtering image decomposition. Tang et al. [39] introduce a semantic-aware image fusion network (SeA-Fusion), which leverages the semantic segmentation task to guide the image fusion with a gradient residual dense block (GRDB). More multi-modal image fusion techniques in medical imaging are discussed in Tirupal et al. [40] and Srinvasu et al. [41].

In this work, we integrate the image deblurring and image fusion techniques and propose a deep learning based image fusion algorithm to deblur images by fusing the pre-deblurred streams.

III. PROPOSED METHOD

A. GDFNet

The idea of deblurring by image fusion is motivated by the following observations. Single image deblurring algorithms usually recover low-frequency information or unreliable textures since the degradation corrupts the details and there are few clues to recover them. On the contrary, guided deblurring

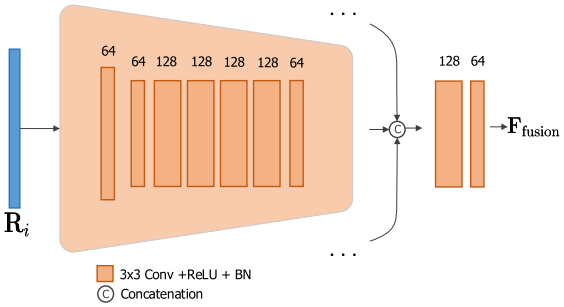

FIGURE 3. Illustration of the encoders and feature aggregation.

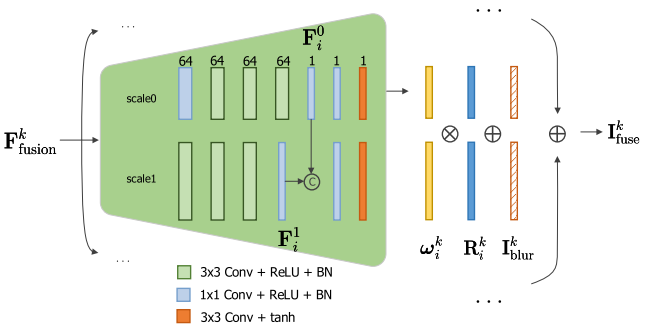

FIGURE 4. Illustration of fusion weights generation.

algorithms preserve more high-frequency components like edges and textures according to the guidance image. However, they produce artifacts due to the structural inconsistency and differences in wavelength sensitivites between multi-modal images. To overcome these drawbacks, we fuse these deblurred images into $\mathbf{I}_{\text{fuse}}$. It is represented as the linear combination of three pre-deblurred image $\mathbf{J}_1$, $\mathbf{J}_2$, and $\mathbf{J}_3$,

$$\mathbf{I}_{\text{fuse}} = \boldsymbol{\omega}_1 \odot \mathbf{J}_1 + \boldsymbol{\omega}_2 \odot \mathbf{J}_2 + \boldsymbol{\omega}_3 \odot \mathbf{J}_3, \tag{2}$$

where $\boldsymbol{\omega}_1, \boldsymbol{\omega}_2, \boldsymbol{\omega}_3$ represents corresponding fusion weights and $\odot$ is the element-wise product. The pre-deblurred images $\mathbf{J}_1$, $\mathbf{J}_2$, and $\mathbf{J}_3$ can be represented as

$$\mathbf{J}_1 = \mathcal{N}_1(\mathbf{I}_{\text{blur}}), \tag{3}$$
$$\mathbf{J}_2 = \mathcal{N}_2(\mathbf{I}_{\text{blur}}, \mathbf{I}_{\text{guid}}), \tag{4}$$

and

$$\mathbf{J}_3 = \mathcal{N}_3(\mathbf{J}_1, \mathbf{I}_{\text{guid}}), \tag{5}$$

where $\mathcal{N}_1$, $\mathcal{N}_2$, and $\mathcal{N}_3$ represent three pre-deblurring networks.

Since the pre-deblurred images computed by multiple image deblurring streams are the estimates of the sharp images, they are similar in low-frequency content. Therefore, small differences in fusion weights can greatly affect the fused details, which makes the image deblurring not robust. As we use CNN to predict the fusion weights, the networks often fit target functions from low frequencies to high frequencies according to the frequency principle claimed by Xu et al. [10]. The success of residual learning [9], [42] inspires us to use an effective blur/residual image splitting strategy in GDFNet to focus on high-frequency component fusion. The blurry image itself can be regarded as the low-frequency component of the estimated sharp latent image, and the difference between a pre-deblurred image and the blurry image is a reasonable initial guess of the high-frequency component. Correspondingly, the residual $\mathbf{R}_i$ is defined in the form

$$\mathbf{R}_i = \mathbf{J}_i - \mathbf{I}_{\text{blur}}. \tag{6}$$

Combining (2) and (6), the fused deblurred image $\mathbf{I}_{\text{fuse}}$ can be expressed as

$$\mathbf{I}_{\text{fuse}} = \boldsymbol{\omega}_1 \odot \mathbf{R}_1 + \boldsymbol{\omega}_2 \odot \mathbf{R}_2 + \boldsymbol{\omega}_3 \odot \mathbf{R}_3 + \sum_{i=1}^{3} \boldsymbol{\omega}_i \odot \mathbf{I}_{\text{blur}}. \tag{7}$$

In this way, we encourage GDFNet to focus on the fusion of high-frequency components which makes the convergence faster and the deblurring performance better.

The overall framework of our GDFNet is illustrated in Fig. 2. The network takes the blurry image $\mathbf{I}_{\text{blur}}$ and the guidance image $\mathbf{I}_{\text{guid}}$ as input to compute a deblurred image $\mathbf{I}_{\text{fuse}}$ through image fusion. We use three deblurring streams to generate three different pre-deblurred images for image fusion. A single deblurring stream takes $\mathbf{I}_{\text{blur}}$ as input and predicts a coarsely deblurred image $\mathbf{J}_1$. A guided deblurring stream takes $\mathbf{I}_{\text{blur}}$ and $\mathbf{I}_{\text{guid}}$ as input and jointly estimates an edge-preserving deblurred image $\mathbf{J}_2$. Another guided deblurring stream takes $\mathbf{I}_{\text{guid}}$ and the output of the single deblurring stream $\mathbf{J}_1$ as input to recover another pre-deblurred image $\mathbf{J}_3$.

Then we compute the corresponding residual images $\mathbf{R}_1$, $\mathbf{R}_2$, and $\mathbf{R}_3$ of three pre-deblurred images $\mathbf{J}_1$, $\mathbf{J}_2$, and $\mathbf{J}_3$, respectively, by subtracting them with the blurry observation $\mathbf{I}_{\text{blur}}$. The features of these residual images are extracted by three individual encoders and aggregated by a feature concatenation layer. Since these aggregated features contain all information of three pre-deblurred images, they can predict the fusion weights $\boldsymbol{\omega}_1$, $\boldsymbol{\omega}_2$, and $\boldsymbol{\omega}_3$ of the three residual images. We compute the weighted summation of the multi-scale residual images using the coarse-to-fine fusion weights $\boldsymbol{\omega}_1$, $\boldsymbol{\omega}_2$, and $\boldsymbol{\omega}_3$ to obtain multi-scale fused residual images. During training process, we add multi-scale blurry input to them for multi-scale supervision to generate finer fusion weights. During the inference, we add the blurry input $\mathbf{I}_{\text{blur}}$ to the original scaled fused residual image using the single-scale extraction step to generate the final fused output image $\mathbf{I}_{\text{fuse}}$.

B. PRE-DEBLURRING STREAMS

To generate the pre-deblurred images for image fusion, we use three pre-deblurring streams, including one single

image deblurring stream and two guided image deblurring streams. All of the streams are replaceable and can be implemented using state-of-the-art approaches.

The single image deblurring stream $\mathcal{N}_1$ provides an estimate $\mathbf{J}_1$, which cannot restore the details completely because most of the high-frequency components are lost during degradation. We use the guided image deblurring network $\mathcal{N}_2$ to recover more details by taking the concatenation of the blurry observation $\mathbf{I}_{\text{blur}}$ and guidance image $\mathbf{I}_{\text{guid}}$ as input. The deblurred image $\mathbf{J}_2$ embeds the information of the guidance image. Since the structure in the blurry input is not reliable, this stream tends to use low-frequency contents in $\mathbf{I}_{\text{guid}}$. However, these contents may be wrong due to object movements and sensor differences. As a result, it causes the pre-deblurred image $\mathbf{J}_2$ to be structurally and color inconsistent with the ground truth, and creates halos and fake shadows. Therefore, we employ another guided image deblurring network $\mathcal{N}_3$ which uses the single deblurring stream $\mathbf{J}_1$ and the guidance image $\mathbf{I}_{\text{guid}}$ to estimate $\mathbf{J}_3$. The pre-deblurred image $\mathbf{J}_1$ is a prediction of the sharp image containing coarse but correct structures. Thus, this guided stream learns to believe the structures in $\mathbf{J}_1$ rather than $\mathbf{I}_{\text{guid}}$, because the structures in $\mathbf{I}_{\text{guid}}$ can be wrong in some cases like object movements. On the other hand, $\mathbf{I}_{\text{guid}}$ brings less impact on corrupting the structure since the inconsistency between the multi-modal images $\mathbf{J}_1$ and $\mathbf{I}_{\text{guid}}$ is reduced. Therefore, the output $\mathbf{J}_3$ provides better estimates and different information compared to the other two streams.

C. FEATURE AGGREGATION

Based on our blur/residual image splitting strategy, the residual images of three pre-deblurred streams are fed into the fusion network and encoded by three independent encoders. We choose to use individual encoders because three pre-deblurred residual images focus on recovering different contents of sharp images. Fig. 3 illustrates the details of the feature aggregation architecture. The encoders contain 6 convolutional blocks and 2 max-pooling layers, which compute features with 64 channels and the spatial size is $\frac{1}{4}$ of the input size. The features of the residual images of the three pre-deblurred residual images are then concatenated, followed by two convolutional blocks to reduce the dimension of the channel. The entire feature aggregation process is represented as

$$\mathbf{F}_{\text{fusion}} = \text{conv}(\mathcal{N}_{E_1}(\mathbf{R}_1) \oplus \mathcal{N}_{E_2}(\mathbf{R}_2) \oplus \mathcal{N}_{E_3}(\mathbf{R}_3)), \quad (8)$$

where $\mathcal{N}_{E_1}$, $\mathcal{N}_{E_2}$, and $\mathcal{N}_{E_3}$ are the corresponding encoders of three residual images which include two downsampling layers, $\oplus$ denotes the concatenate operation, and $\text{conv}(\cdot)$ denotes the convolutional operator that reduces the dimension of the channel to 64. We use these encoders to keep information in three streams for the following fusion weights generation. The blur/residual image splitting strategy enhances the feature representation ability by directly extracting features on the residual images.

D. COARSE-TO-FINE RECONSTRUCTION

We adopt a coarse-to-fine reconstruction strategy that can further improve the quality of fused deblurred images [43]. The architecture of the coarse-to-fine reconstruction and its decoders for fusion weights generation is demonstrated in Fig. 4. We use three individual decoders that contain 5 convolutional blocks and upsampling layers for every level of scale, each of them generates 2-scale features

$$\mathbf{F}_i^k = \mathcal{N}_{D_i}^k(\mathbf{F}_{\text{fusion}}^k), k \in \{0, 1\}, \quad (9)$$

where k represents the scale level and $\mathcal{N}_{D_i}^k$ is the i-th ($1 \leq i \leq 3$) decoder at level k, and $\mathbf{F}_{\text{fusion}}^k$ is the 2^k upsampling of $\mathbf{F}_{\text{fusion}}$.

The fusion weights $\boldsymbol{\omega}_i^k$ of the residual images at scale level k are computed using a tanh activation function instead of a ReLU because we need to preserve both positive and negative values for fusion. The weights are given by

$$\boldsymbol{\omega}_i^k = \begin{cases} \mathcal{N}_{\text{tanh}}^k((\mathbf{F}_i^{k-1})_\uparrow \oplus \mathbf{F}_i^k) & k = 1 \\ \mathcal{N}_{\text{tanh}}^k(\mathbf{F}_i^k) & k = 0, \end{cases} \quad (10)$$

where $\boldsymbol{\omega}_i^k$ is the fusion weights of the i-th residual image at scale level k, $\mathcal{N}_{\text{tanh}}^k$ is the tanh activation function at level k, $(\cdot)_\uparrow$ represents the upsampling operation, and $\oplus$ denotes concatenation. The fused deblurred image is computed by adding the fused residuals layers of three pre-deblurred images using $\boldsymbol{\omega}_i^k$ to the original blurred observation,

$$\mathbf{I}_{\text{fuse}}^k = \boldsymbol{\omega}_1^k \odot \mathbf{R}_1^k + \boldsymbol{\omega}_2^k \odot \mathbf{R}_2^k + \boldsymbol{\omega}_3^k \odot \mathbf{R}_3^k + \mathbf{I}_{\text{blur}}^k, \quad (11)$$

where $\mathbf{R}_i^k, i \in \{1,2,3\}$, represents the corresponding residual image at scale level k downsampled using max-pooling, $\mathbf{I}_{\text{fuse}}^k, k \in \{0,1\}$ represents the 2-level fused deblurred images, and $\odot$ denotes the element-wise product. Both 2-level deblurred images are used for back propagation in training, but only the original scale image is needed in inference.

E. LOSS FUNCTION

We employ L_1, structural similarity index measure (SSIM), learned perceptual image patch similarity (LPIPS) [44], and a frequency-domain loss function on the multi-scale output $\mathbf{I}_{\text{fuse}}^k$ [43] to train GDFNet. The multi-scale LPIPS can be represented as

$$\mathcal{L}_{\text{MS-per}} = \sum_{k=0}^{1} \left\| \mathcal{P}\left(\mathbf{I}_{\text{fuse}}^k\right) - \mathcal{P}\left(\mathbf{I}_{\text{gt}}^k\right) \right\|_1, \quad (12)$$

where k and $\mathcal{P}(\cdot)$ denote the scale level and LPIPS network, respectively. LPIPS is a pre-trained network for evaluating the perceptual similarity between two images. Recent studies show that reducing the frequency-domain discrepancy is essential for restoring the lost high-frequency components [1], [3]. We adopt the multi-scale frequency reconstruction (MSPR) loss function on our multi-scale output,

$$\mathcal{L}_{\text{MS-freq}} = \sum_{k=0}^{K-1} \frac{1}{M^k} \left\| \mathcal{F}\left(\mathbf{I}_{\text{fuse}}^k\right) - \mathcal{F}\left(\mathbf{I}_{\text{gt}}^k\right) \right\|_1, \quad (13)$$

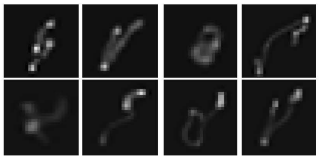

FIGURE 5. The 8 blur kernels used for generating synthetic blurry input on the RGB/NIR dataset.

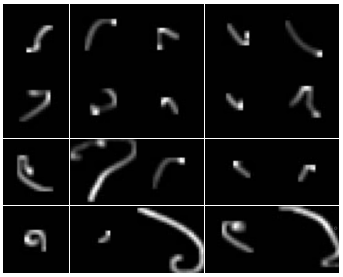

FIGURE 6. The 20 blur kernels used for generating synthetic blurry input on the ambient/flash dataset.

where $\mathcal{F}(\cdot)$ denotes the Fourier transformation and M^k denotes the number of elements at scale k. The total loss function is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{MS-}L_1} + \lambda_1 \mathcal{L}_{\text{MS-SSIM}} + \lambda_2 \mathcal{L}_{\text{MS-per}} + \lambda_3 \mathcal{L}_{\text{MS-freq}}. \tag{14}$$

The balance parameters are empirically set as $\lambda_1 = 0.2$, $\lambda_2 = 0.1$, and $\lambda_3 = 0.1$.

IV. EXPERIMENTS

In this work, the three pre-deblurred streams $\mathcal{N}_1$, $\mathcal{N}_2$, and $\mathcal{N}_3$ in GDFNet are flexible and can be replaced by other deblurring streams. To evaluate our algorithm, we conduct experiments using MIMO [1], MPR [2], and DeepRFT [3] as $\mathcal{N}_1$. As for the choice of the guided deblurring networks $\mathcal{N}_2$ and $\mathcal{N}_3$, we use SVLRM [5] to compute $\mathbf{J}_2$ and $\mathbf{J}_3$. We evaluate GDFNet on popular public multi-modal image datasets including RGB/NIR [31] and flash/ambient [33]. Our method is compared with single image deblurring algorithms MIMO [1], MPR [2], DeepRFT [3], Uformer [27], NAFNet [29], guided image deblurring algorithm SVLRM [5], and the combinations such as MIMO+SVLRM, MPR+SVLRM, and DeepRFT+SVLRM. We also compare our fusion network with DenseFuse [34] and SeAFusion [39] using the same input as GDFNet. All compared networks are retrained on the same datasets for fair comparisons.

A. IMPLEMENTATION DETAILS

The experiments are conducted on an Intel Xeon Silver 4210R CPU @ 2.40GHz with 64GB memory and an NVIDIA Quadro RTX 8000. We implemented our network using PyTorch [45]. It is trained using Adam optimizer [46], and the initial learning rate is set to 1×10^{-4}, with a batch size of 64 and a maximal training epoch of 200. The training image is of size 128×128. We use zero-padding in convolutional layers.

TABLE 1. Quantitative comparisons with image deblurring algorithms on the RGB/NIR datasets in terms of PSNR, SSIM, and LPIPS values. The best ones are in bold.

Method	PSNR↑	SSIM↑	LPIPS↓
NAFNet [29]	23.15	0.7463	0.1430
UFormer [27]	24.63	0.8048	0.1098
SVLRM [5]	25.23	0.7528	0.1698
MIMO [1]	30.72	0.8897	0.0600
MIMO+SVLRM	31.19	0.8980	0.0568
Ours (GDFNet w/ MIMO)	**31.32**	**0.9029**	**0.0541**
MPR [2]	25.14	0.7030	0.1620
MPR+SVLRM	27.03	0.7858	0.1410
Ours (GDFNet w/ MPR)	**27.61**	**0.8059**	**0.1161**
DeepRFT [3]	24.54	0.6716	0.2103
DeepRFT+SVLRM	26.67	0.7739	0.1493
Ours (GDFNet w/ DeepRFT)	**27.09**	**0.7862**	**0.1422**

TABLE 2. Quantitative comparisons with image fusion approaches on the RGB/NIR datasets in terms of PSNR, SSIM, and LPIPS values. The best ones are in bold.

Method	PSNR↑	SSIM↑	LPIPS↓
DenseFuse [34] (w/ MIMO)	29.74	0.8907	0.0631
SeAFusion [39] (w/ MIMO)	31.27	0.9013	0.0542
Ours (GDFNet w/ MIMO)	**31.32**	**0.9029**	**0.0541**
DenseFuse [34] (w/ MPR)	26.71	0.7870	0.1257
SeAFusion [39] (w/ MPR)	27.21	0.7952	0.1207
Ours (GDFNet w/ MPR)	**27.61**	**0.8059**	**0.1161**
DenseFuse [34] (w/ DeepRFT)	26.26	0.7711	**0.1378**
SeAFusion [39] (w/ DeepRFT)	26.96	0.7829	0.1396
Ours (GDFNet w/ DeepRFT)	**27.09**	**0.7862**	0.1422

B. DATASETS

We evaluate GDFNet on two multi-modal image pair datasets. The RGB-NIR scene [31] contains registered RGB and NIR image pairs. They are collected using several digital single-lens reflex (DSLR) cameras with and without infrared blocking filters in separated exposures. Although the cameras are equipped with tripods, there are still small misalignments between the RGB and NIR images. Many algorithms are introduced to register multi-modal image such as [47], [48], and [49]. This dataset applies a feature-based alignment algorithm [50] to register these image pairs. We generate the blurry input by computing the convolution of the sharp RGB images and the ground truth blur kernels in [51]. Fig. 5 shows the blur kernels used in our experiments. For each scene, we randomly select one kernel to blur the RGB image, and take the NIR image as the guidance reference image. We discard the image pairs of large misalignment in the RGB/NIR dataset.

The blurry/flash dataset is generated using ambient and flash illumination image pairs presented in [33]. The ambient and flash dataset is collected using DLSR cameras controlled by a mobile App that sequentially captures the flash and

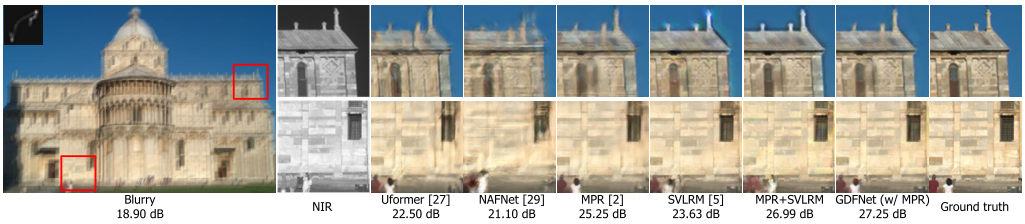

FIGURE 7. The deblurring results and corresponding PSNR of a palace produced by different algorithms on the blurry/flash dataset.

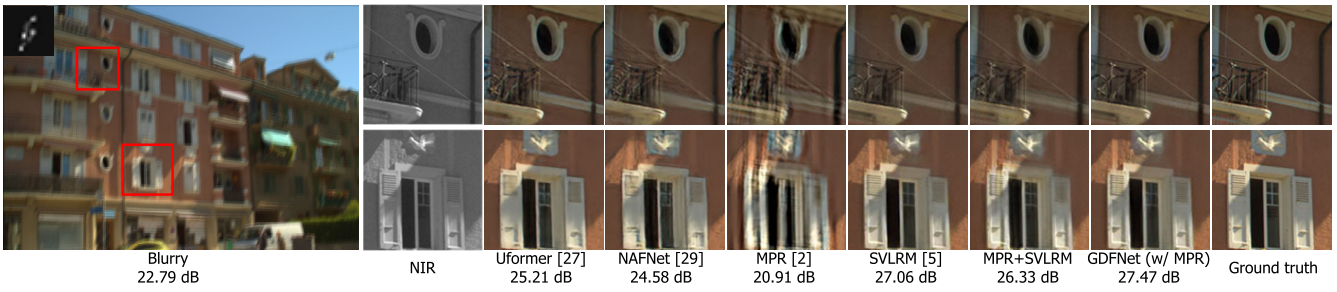

FIGURE 8. The deblurring results and corresponding PSNR of a building produced by different algorithms on the blurry/flash dataset.

TABLE 3. Quantitative evaluations on the RGB/noisy NIR datasets in terms of PSNR, SSIM, and LPIPS values. The best ones are in bold.

Method	PSNR↑	SSIM↑	LPIPS↓
MPR [2]	25.14	0.7030	0.1620
SVLRM [5]	24.03	0.6618	0.3268
MPR+SVLRM	26.27	0.7309	0.2145
DenseFuse [34] (w/ MPR)	25.67	0.7378	0.1715
SeAFusion [39] (w/ MPR)	26.51	0.7449	0.1715
Ours (GDFNet w/ MPR)	**26.71**	**0.7481**	**0.1685**

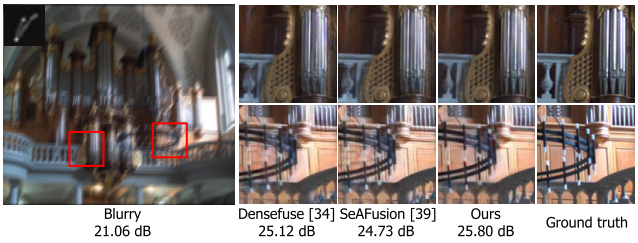

FIGURE 9. The deblurring results and corresponding PSNR of a indoor scene produced by different fusion networks using the same input on the RGB/NIR dataset.

no-flash photographs with a small delay between two exposures. An improved version of the dual inverse compositional alignment algorithm (DIC) [52] is used to correct the misalignment. We blurred the ambient image using 20 ground truth blur kernels provided in [53] as shown in Fig. 6, which are resized to 13×13, 19×19, and 25×25 to increase the diversity of degradation types. For the guidance image, we multiply the intensities of flash images by 1.5 to remove some structures by over-exposure, which actually makes the image deblurring more challenging.

C. EVALUATION ON RGB/NIR DATASET

We evaluate our GDFNet on the RGB/NIR dataset [31], which includes 300 scenes. Deblurring streams $\mathcal{N}_1$ and $\mathcal{N}_2$ takes 40 randomly choosed image pairs to train, and 10 image pairs to validate. We choose the trained network with the best PSNR on the validation dataset. After we determine $\mathcal{N}_1$, $\mathcal{N}_3$ takes the output $\mathbf{J}_1$ and the guidance NIR image $\mathbf{I}_{\text{guid}}$ as input, using another 50 image pairs to train. Then we fix the parameters of these three pre-deblurred streams and train GDFNet using corresponding pre-deblurred images of 50 scenes. The remaining 150 scenes are used for testing.

As shown in Table 1, we compare our method GDFNet with image deblurring algorithms quantitatively on 150 scenes of RGB/NIR dataset. We split the comparisons into three parts based on using MIMO, MPR, and DeepRFT as $\mathcal{N}_1$. Our method (GDFNet with MIMO) outperforms the image deblurring algorithms in terms of PSNR, SSIM, and LPIPS. Our method (GDFNet with MPR) and our method (GDFNet with DeepRFT) also perform favorably against state-of-the-art image deblurring algorithms. Fig. 7 and Fig. 8 show the image deblurring results by the competing algorithms. The single deblurring algorithms MPR [2], Uformer [27], and NAFNet [29] lose textures and generate ringing patterns around large edges. The SVLRM [5] produces color shifts and ghost shadows. The details of textures can be obtained from NIR input but artifacts may appear since the local linear assumption [32] is weak on inconsistent structures. By exploring the use of the single deblurred image and NIR image, MPR+SVLRM reduces the appearance of

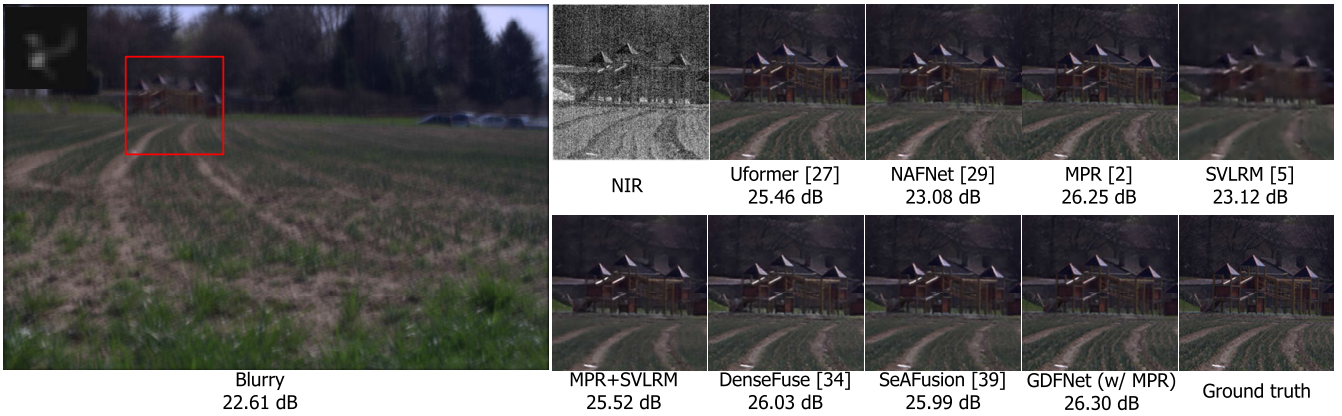

FIGURE 10. The results and corresponding PSNR produced by different algorithms on the RGB/NIR image pair with stripe noise.

artifacts. However, some blurry edges still exist. Instead of using one single or guided deblurring algorithms, our GDFNet fuses three pre-deblurred streams, generating a clearer image with sharper edges.

To validate the superiority of our fusion network, we compare two image fusion approaches with our GDFNet using the same pre-deblurred images as input. We modified SeAFusion so that it can accept three images as input. The results in Table 2 show that the GDFNet outperforms other algorithms. As shown in Fig. 9, DenseFuse [34] and SeAFusion [2] produce chromatic aberrations around the organ pipes and false edges around the chandelier. They are designed for image fusion, but they do not correctly fuse the high-frequency information. Our GDFNet produces fewer artifacts and creates clear edges.

TABLE 4. Quantitative comparisons with image deblurring algorithms on the blurry/flash datasets in terms of PSNR, SSIM, and LPIPS values. The best ones are in bold.

Method	PSNR↑	SSIM↑	LPIPS↓
NAFNet [29]	28.10	0.8432	0.1348
UFormer [27]	26.36	0.7930	0.1575
SVLRM [5]	30.35	0.8578	0.1474
MIMO [1]	28.85	0.8172	**0.1161**
MIMO+SVLRM	30.45	0.8664	0.1379
Ours (GDFNet w/ MIMO)	**31.34**	**0.8709**	0.1255
MPR [2]	29.88	0.8438	0.1133
MPR+SVLRM	31.14	0.8752	0.1229
Ours (GDFNet w/ MPR)	**32.35**	**0.8842**	**0.1080**
DeepRFT [3]	29.77	0.8743	**0.0628**
DeepRFT+SVLRM	30.71	0.8898	0.1054
Ours (GDFNet w/ DeepRFT)	**32.33**	**0.8952**	0.0867

D. EXTENSION TO STRIPE NOISE ON RGB/NIR DATASET

Although our algorithm assumes that the NIR images are noise-free, there is line pattern stripe noise [8] or random noise [54] in case of poor imaging quality. Since we use the NIR images as the guidance image, the stripe noise or random noise may deteriorate the guided image deblurring pre-deblurred images and further corrupt the final fused deblurred image. We compare our GDFNet on RGB/NIR dataset where the NIR images are noisy. We manually corrupt the NIR images using stripe noise and random noise. The stripe noise is simulated in a similar manner to [8], with its intensity varying from $[-10, 10]$. The random noise level is set to 10% following Tai and Lin [54]. Table 3 shows the quantitative evaluation results where our GDFNet still perform better than deblur/fusion approaches. Since GDFNet fuses three streams to compute the deblurred image, it is robust to noise when the NIR images are severely degraded.

We show the deblurring results of a RGB and noisy NIR image pair in Fig. 10. Single image deblurring algorithms such as NAFNet [29], Uformer [27], and MPR [2] produce deblurred images with small blur at the edges. The result of SVLRM [5] is blurry because the structure of the reference NIR image is destroyed by noise. Noise has little effect on MPR+SVLRM because the network learns that the guidance is less reliable and tends to use the output of MPR to regress the restoration result. DenseFuse [34] and SeAFusion [39] produce sharper images but are still affected by noise. Our GDFNet generates sharp result because it can lower the fusion weights of useless deblurring streams.

E. EVALUATION ON BLURRY/FLASH DATASET

We evaluate our GDFNet on the *Object* category in the ambient/flash dataset [33] which includes 578 scenes. Deblurring streams $\mathcal{N}_1$ and $\mathcal{N}_2$ takes 160 scenes to train, and 40 scenes to validate. We choose the trained network with the best PSNR on the validation dataset. Similar to the RGB/NIR dataset, we use 80 scenes and 20 scenes to train and validate $\mathcal{N}_3$. Then we fix three pre-deblurred streams, and train GDFNet using another 40 scenes to train and 10 scenes to validate. The remaining 228 scenes are used for testing.

As shown in Table 4 and Table 5, we compare our GDFNet with image deblurring algorithms and image fusion approaches quantitatively on 228 scenes of the blurry/flash dataset. Our methods perform better

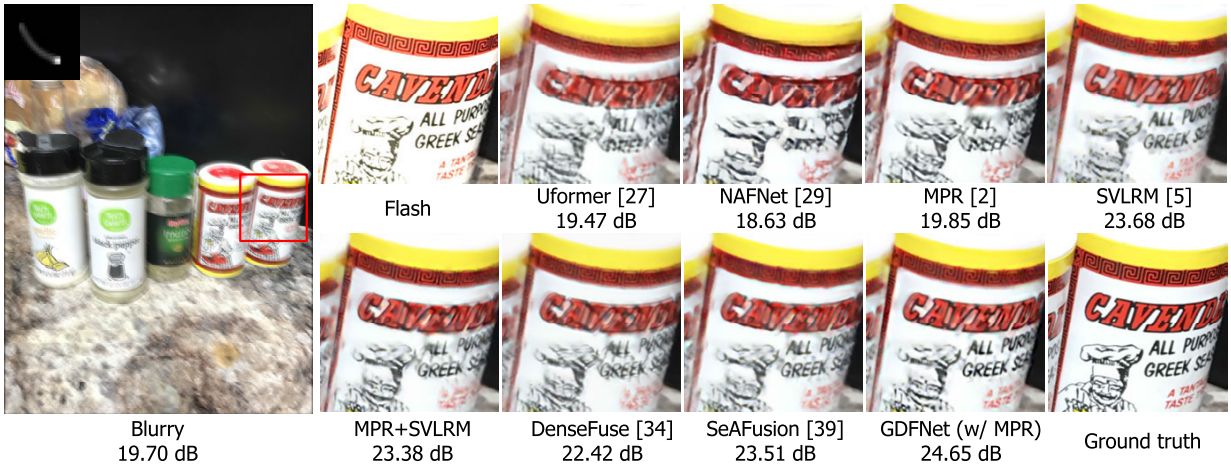

FIGURE 11. The deblurring results and corresponding PSNR of a building produced by different algorithms on the blurry/flash dataset.

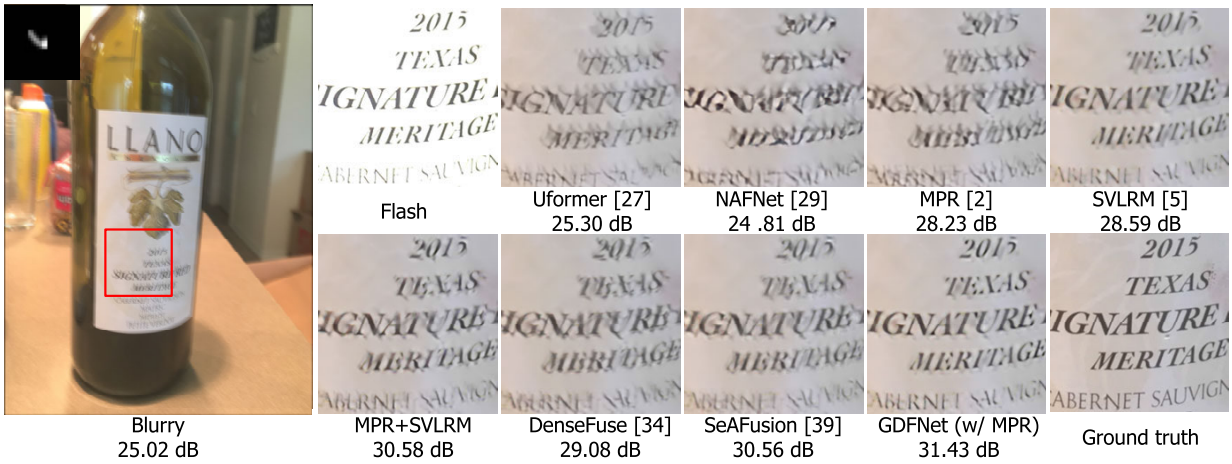

FIGURE 12. The deblurring results and corresponding PSNR of a building produced by different algorithms on the blurry/flash dataset.

TABLE 5. Quantitative comparisons with image fusion approaches on the blurry/flash datasets in terms of PSNR, SSIM, and LPIPS values. The best ones are in bold.

Method	PSNR↑	SSIM↑	LPIPS↓
DenseFuse [34] (w/ MIMO)	29.97	0.8628	0.1496
SeAFusion [39] (w/ MIMO)	30.85	0.8647	0.1366
Ours (GDFNet w/ MIMO)	**31.34**	**0.8709**	**0.1255**
DenseFuse [34] (w/ MPR)	30.45	0.8748	0.1335
SeAFusion [39] (w/ MPR)	31.41	0.8754	0.1238
Ours (GDFNet w/ MPR)	**32.35**	**0.8842**	**0.1080**
DenseFuse [34] (w/ DeepRFT)	31.09	0.8945	0.1085
SeAFusion [39] (w/ DeepRFT)	32.11	0.8938	0.0954
Ours (GDFNet w/ DeepRFT)	**32.33**	**0.8952**	**0.0867**

TABLE 6. Quantitative comparisons on loss functions using GDFNet (w/ MPR) on the blurry/flash dataset. The best metrics are in bold, and the second best ones are underlined.

L1	SSIM	Perceptual	FFT	PSNR↑	SSIM↑	LPIPS↓
✓	✓	✓	✓	**32.35**	<u>0.8842</u>	<u>0.1080</u>
✓	✓	✓	-	32.20	0.8817	**0.0990**
✓	✓	-	-	<u>32.30</u>	**0.8847**	0.1129
✓	-	-	-	32.19	0.8810	0.1179

TABLE 7. Performance comparisons on different ablations of GDFNet (w/ MPR) on the blurry/flash dataset. The best metric values are in bold.

Method	PSNR↑	SSIM↑	LPIPS↓
Ours (w/o residual images)	31.37	0.8726	0.1299
Ours (w/o multi-scale reconstruction)	32.00	0.8779	0.1152
Ours (w/o using $\mathbf{J}_2$)	31.88	0.8752	0.1192
Ours (w/o using $\mathbf{J}_3$)	32.05	0.8791	0.1141
Ours	**32.35**	**0.8842**	**0.1080**

against the state-of-the-art image deblurring algorithms and image fusion approaches. Fig. 11 and Fig. 12 illustrate the image deblurring results by the evaluted image deblurring and image fusion algorithms. In the resultant images produced by MPR [2] and NAFNet [29] the texts are hardly recognizable since the recovered high-frequency information is incorrect. The resultant content of Uformer [27] is barely readable due to ghost shadows. The texts produced by SVLRM [5] and MPR+SVLRM are recognizable but the edges are less sharp. Image fusion approaches DenseFuse and SeAFusion produce texts that are generally clear but still

TABLE 8. Inference times (in seconds) of different algorithms on the RGB/NIR dataset.

	Method	Time	Method	Time	Method	Time
Deblurring	SVLRM [5]	0.040	NAFNet [29]	0.108	Uformer [27]	0.822
	MIMO [1]	0.029	MPR [2]	0.100	DeepRFT [3]	0.173
Fusion	DenseFuse [34]	0.014	SeAFusion [39]	0.018	GDFNet	0.019

contain artifacts. Our GDFNet produces the highest PSNR values and successfully recovers texts with sharp edges.

F. ABLATION STUDIES AND RUNNING TIMES

As shown in (14), we use several loss functions, including L1 loss, SSIM loss, perceptual loss, and frequency loss. To analyze the effect of the loss function on the performance of GDFNet, we train it on the blurry/flash dataset using different versions of loss functions. Quantitative results in Table 6 show that our loss function is suitable when considering all the PSNR, SSIM, and LPIPS metrics.

We conduct ablation study on the framework components, including residual image, multi-scale reconstruction, and multiple guided deblurring streams. As shown in Table 7, our framework works better than the versions with any components removed. The effect of using residual images is significant; it increases the PSNR by 0.98 dB. The multi-scale reconstruction with supervision improves the performance by 0.35 dB in terms of PSNR. The results of without using $\mathbf{J}_2$ and without using $\mathbf{J}_3$ demonstrate that both streams are useful, increasing the PSNR by 0.47 dB and 0.3 dB, respectively.

To quantitatively compare the inference time of the proposed method with image deblurring algorithms and image fusion approaches, we evaluate all algorithms on the RGB/NIR dataset. Table 8 shows that the inference time of our GDFNet is lower than the deblurring algorithms and close to the image fusion approaches.

V. CONCLUSION

This paper proposes a novel guided image deblurring framework based on deep image fusion using multi-modal image pairs, called Guided Deblurring Fusion Network (GDFNet). Previous work on image deblurring has focused on image deblurring, neglecting the alternative of image fusion. We use GDFNet to fuse the pre-deblurred streams of single and guided image deblurring algorithms to aggregate the structures and the sharp details based on fusion weights. In detail, GDFNet employs the blur/residual image splitting strategy and a coarse-to-fine reconstruction module supervised by multi-scale ground truths. The effectiveness of the strategy used in GDFNet is demonstrated by ablation study. Our method can be easily extended by replacing the approaches used as pre-deblurred streams. Quantitative comparisons show that GDFNet outperforms the image deblurring algorithms and image fusion approaches. The average PSNR of GDFNet is at least 0.9 dB higher than existed algorithms evaluated on 228 test scenes from the blurry/flash dataset.

REFERENCES

[1] S.-J. Cho, S.-W. Ji, J.-P. Hong, S.-W. Jung, and S.-J. Ko, "Rethinking coarse-to-fine approach in single image deblurring," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 4641–4650.

[2] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, M.-H. Yang, and L. Shao, "Multi-stage progressive image restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14821–14831.

[3] X. Mao, Y. Liu, F. Liu, Q. Li, W. Shen, and Y. Wang, "Intriguing findings of frequency selection for image deblurring," 2021, *arXiv:2111.11745*.

[4] S. Zhuo, D. Guo, and T. Sim, "Robust flash deblurring," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 2440–2447.

[5] J. Dong, J. Pan, J. S. Ren, L. Lin, J. Tang, and M.-H. Yang, "Learning spatially variant linear representation models for joint filtering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 8355–8370, Nov. 2022.

[6] H. Yamashita, D. Sugimura, and T. Hamamoto, "RGB-NIR imaging with exposure bracketing for joint denoising and deblurring of low-light color images," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 6055–6059.

[7] L. Yuan, J. Sun, L. Quan, and H.-Y. Shum, "Image deblurring with blurred/noisy image pairs," *ACM Trans. Graph.*, vol. 26, no. 3, p. 1, Jul. 2007, doi: 10.1145/1276377.1276379.

[8] S. Cao, H. Fang, L. Chen, W. Zhang, Y. Chang, and L. Yan, "Robust blind deblurring under stripe noise for remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–17, 2022.

[9] H. Gao, X. Tao, X. Shen, and J. Jia, "Dynamic scene deblurring with parameter selective sharing and nested skip connections," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3848–3856.

[10] Z.-Q. J. Xu, Y. Zhang, T. Luo, Y. Xiao, and Z. Ma, "Frequency principle: Fourier analysis sheds light on deep neural networks," 2019, *arXiv:1901.06523*.

[11] R. Fergus, B. Singh, A. Hertzmann, S. T. Roweis, and W. T. Freeman, "Removing camera shake from a single photograph," *ACM Trans. Graph.*, vol. 25, no. 3, pp. 787–794, 2006.

[12] D. Krishnan and R. Fergus, "Fast image deconvolution using hyper-Laplacian priors," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 1033–1041.

[13] Q. Shan, J. Jia, and A. Agarwala, "High-quality motion deblurring from a single image," *ACM Trans. Graph.*, vol. 27, no. 3, p. 73, 2008.

[14] J. Pan and Z. Su, "Fast ℓ^0-regularized kernel estimation for robust motion deblurring," *IEEE Signal Process. Lett.*, vol. 20, no. 9, pp. 841–844, Sep. 2013.

[15] J. Pan, Z. Hu, Z. Su, and M.-H. Yang, "Deblurring text images via L_0-regularized intensity and gradient prior," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2901–2908.

[16] J. Pan, D. Sun, H. Pfister, and M.-H. Yang, "Blind image deblurring using dark channel prior," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1628–1636.

[17] Y. Yan, W. Ren, Y. Guo, R. Wang, and X. Cao, "Image deblurring via extreme channels prior," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4003–4011.

[18] Y. Bai, H. Jia, M. Jiang, X. Liu, X. Xie, and W. Gao, "Single-image blind deblurring using multi-scale latent structure prior," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 7, pp. 2033–2045, Jul. 2020.

[19] A. Levin, R. Fergus, F. Durand, and W. T. Freeman, "Image and depth from a conventional camera with a coded aperture," *ACM Trans. Graph.*, vol. 26, no. 3, p. 70, 2007.

[20] H. Zhang, D. Wipf, and Y. Zhang, "Multi-image blind deblurring using a coupled adaptive sparse prior," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1051–1058.

[21] C. J. Schuler, M. Hirsch, S. Harmeling, and B. Schölkopf, "Learning to deblur," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 7, pp. 1439–1451, Jul. 2016.

[22] D. Ren, K. Zhang, Q. Wang, Q. Hu, and W. Zuo, "Neural blind deconvolution using deep priors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3341–3350.
[23] X. Tao, H. Gao, X. Shen, J. Wang, and J. Jia, "Scale-recurrent network for deep image deblurring," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8174–8182.
[24] J. J. M. Ople, P.-Y. Yeh, S.-W. Sun, I.-T. Tsai, and K.-L. Hua, "Multi-scale neural network with dilated convolutions for image deblurring," *IEEE Access*, vol. 8, pp. 53942–53952, 2020.
[25] K.-H. Liu, C.-H. Yeh, J.-W. Chung, and C.-Y. Chang, "A motion deblur method based on multi-scale high frequency residual image learning," *IEEE Access*, vol. 8, pp. 66025–66036, 2020.
[26] A. S. Saqlain, F. Fang, L.-Y. Wang, T. Ahmad, and Z. U. Abidin, "DFGAN: Image deblurring through fusing light-weight attention and gradient-based filters," in *Proc. IEEE 2nd Int. Conf. Inf. Commun. Softw. Eng. (ICICSE)*, Mar. 2022, pp. 110–114.
[27] Z. Wang, X. Cun, J. Bao, W. Zhou, J. Liu, and H. Li, "Uformer: A general U-shaped transformer for image restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 17683–17693.
[28] F.-J. Tsai, Y.-T. Peng, Y.-Y. Lin, C.-C. Tsai, and C.-W. Lin, "Stripformer: Strip transformer for fast image deblurring," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 1–17.
[29] L. Chen, X. Chu, X. Zhang, and J. Sun, "Simple baselines for image restoration," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 1–15.
[30] X. Chu, L. Chen, C. Chen, and X. Lu, "Improving image restoration by revisiting global information aggregation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 53–71.
[31] M. Brown and S. Susstrunk, "Multi-spectral SIFT for scene category recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2011, pp. 177–184.
[32] K. He, J. Sun, and X. Tang, "Guided image filtering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 6, pp. 1397–1409, Jun. 2013.
[33] Y. Aksoy, C. Kim, P. Kellnhofer, S. Paris, M. Elgharib, M. Pollefeys, and W. Matusik, "A dataset of flash and ambient illumination pairs from the crowd," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 634–649.
[34] H. Li and X.-J. Wu, "DenseFuse: A fusion approach to infrared and visible images," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2614–2623, Dec. 2018.
[35] 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 *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 7, pp. 12797–12804.
[36] Y. Zhou, K. Gao, Z. Dou, Z. Hua, and H. Wang, "Target-aware fusion of infrared and visible images," *IEEE Access*, vol. 6, pp. 79039–79049, 2018.
[37] H. Xu, J. Ma, J. Jiang, X. Guo, and H. Ling, "U2Fusion: A unified unsupervised image fusion network," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 1, pp. 502–518, Jan. 2020.
[38] W. Tan, H. Zhou, J. Song, H. Li, Y. Yu, and J. Du, "Infrared and visible image perceptive fusion through multi-level Gaussian curvature filtering image decomposition," *Appl. Opt.*, vol. 58, no. 12, pp. 3064–3073, Apr. 2019.
[39] L. Tang, J. Yuan, and J. Ma, "Image fusion in the loop of high-level vision tasks: A semantic-aware real-time infrared and visible image fusion network," *Inf. Fusion*, vol. 82, pp. 28–42, Jun. 2022.
[40] T. Tirupal, B. C. Mohan, and S. S. Kumar, "Multimodal medical image fusion techniques—A review," *Current Signal Transduction Therapy*, vol. 16, no. 2, pp. 142–163, 2021.
[41] P. N. Srinivasu, "Performance measurement of various hybridized kernels for noise normalization and enhancement in high-resolution mr images," in *Bio-inspired Neurocomputing*. Cham, Switzerland: Springer, 2021, pp. 1–24.
[42] D. Park, D. U. Kang, J. Kim, and S. Y. Chun, "Multi-temporal recurrent neural networks for progressive non-uniform single image deblurring with incremental temporal training," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 327–343.
[43] S. Nah, T. H. Kim, and K. M. Lee, "Deep multi-scale convolutional neural network for dynamic scene deblurring," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3883–3891.
[44] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 586–595.
[45] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in PyTorch," in *Proc. 31st Conf. Neural Inf. Process. Syst. Workshop*, Long Beach, CA, USA, 2017.
[46] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[47] Y. Zhao, X. Huang, and Z. Zhang, "Deep Lucas–Kanade homography for multimodal image alignment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 15950–15959.
[48] S.-Y. Cao, J. Hu, Z. Sheng, and H.-L. Shen, "Iterative deep homography estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 1879–1888.
[49] H. Le, F. Liu, S. Zhang, and A. Agarwala, "Deep homography estimation for dynamic scenes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7652–7661.
[50] R. Szeliski, "Image alignment and stitching: A tutorial," *Found. Trends Comput. Graph. Vis.*, vol. 2, no. 1, pp. 1–104, 2006.
[51] A. Levin, Y. Weiss, F. Durand, and W. T. Freeman, "Understanding and evaluating blind deconvolution algorithms," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 1964–1971.
[52] A. Bartoli, "Groupwise geometric and photometric direct image registration," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 12, pp. 2098–2108, Dec. 2008.
[53] U. Schmidt, C. Rother, S. Nowozin, J. Jancsary, and S. Roth, "Discriminative non-blind deblurring," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 604–611.
[54] Y.-W. Tai and S. Lin, "Motion-aware noise filtering for deblurring of noisy and blurry images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 17–24.

YUQI LIU received the B.Eng. degree in automation science and technology from Xi'an Jiaotong University, China, in 2015, and the M.Sc. degree in communications and signal processing from the Imperial College London, U.K., in 2016. He is currently pursuing the Ph.D. degree with the College of Information Science and Electronic Engineering, Zhejiang University, China. His research interests include image processing and computer vision.

ZEHUA SHENG received the B.Eng. degree in information engineering from Zhejiang University, China, in 2017, where he is currently pursuing the Ph.D. degree. His research interests include image denoising and multi-modal image processing.

HUI-LIANG SHEN (Member, IEEE) received the B.Eng. and Ph.D. degrees in electronic engineering from Zhejiang University, Hangzhou, China, in 1996 and 2002, respectively. He was a Research Associate and a Research Fellow at The Hong Kong Polytechnic University, Hong Kong, from 2001 to 2005. He is a Professor with the College of Information Science and Electronic Engineering, Zhejiang University. His research interests include multispectral imaging, image processing, computer vision, and machine learning.