Deep Fusion Prior for Multi-Focus Images Super Resolution Fusion

Yuanjie Gu, Student Member, IEEE, Zhibo Xiao, Hailun Wang, Cheng Liu, and Shouyu Wang

Abstract—Multi-focus image fusion (MFIF) and super-resolution (SR) are the inverse problem of imaging model, purposes of MFIF and SR are obtaining all-in-focus and high-resolution 2D mapping of targets. Though various MFIF and SR methods have been designed; almost all the them deal with MFIF and SR separately. This paper unifies MFIF and SR problems in the physical perspective as the multi-focus image super resolution fusion (MFISRF), and we propose a novel unified dataset-free unsupervised framework named deep fusion prior (DFP) based on deep image prior (DIP) to address such MFISRF with single model. Experiments have proved that our proposed DFP approaches or even outperforms those state-of-art MFIF and SR method combinations. To our best knowledge, our proposed work is a dataset-free unsupervised method to simultaneously implement the multi-focus fusion and super-resolution task for the first time. Additionally, DFP is a general framework, thus its networks and focus measurement tactics can be continuously updated to further improve the MFISRF performance. DFP codes are open source available at http://github.com/GuYuanjie/DeepFusionPrior.

Index Terms—Multi-focus image fusion, super resolution, unified model, unsupervised learning, dataset-free learning.

I. INTRODUCTION

THE majority of information acquisition, processing and analysis are based on visual perception system, which first records images, then distinguishes, recognizes and extracts targets, and finally analyzes them to provide instructions for system decision and control. Among them, image recording is the key premise for precision decision. But unfortunately, due to depth of field (DoF) and resolution limitations of optical systems, recorded images often suffer from resolution reduction and defocus blur, thus inevitably deteriorating the subsequent image processing and analysis. Therefore, in order to improve image quality, multi-focus image fusion (MFIF) and super-resolution (SR) methods have been adopted to extend the imaging DoF and enhance the imaging resolution, respectively.

Various MFIF and SR methods have been designed; however, almost all the them deal with MFIF and SR separately. However, we find that MFIF and SR share a unified physical model. The imaging process can be described as
\[ i(x, y) = h(x, y) * o(x, y) + n(x, y) \]  
where \( h(x,y) \) is the 2-D point spreading function (PSF), \( o(x,y) \) is the object, \( n(x,y) \) is the additive noise, * is the spatial convolution, and \( i(x,y) \) is the image. For SR, its purpose is to use \( i(x,y) \) to obtain an estimate \( \hat{d}(x,y) \) of the real object \( o(x,y) \). While for MFIF, its model can be described in (2), where \( i(x,y) \) and \( o(x,y) \) are the un fused images focusing on foreground and background, \( m(x,y) \) and \( m_b(x,y) \) are the decision maps determining the foreground and background regions, \( E \) is the identity matrix, and \( u(x,y) \) is the fused image.
\[ u(x, y) = m_f(x, y) \cdot i_f(x, y) + m_b(x, y) \cdot i_b(x, y) \]  
Both \( i_f(x,y) \) and \( i_b(x,y) \) can be represented by (3a) and (3b), in which \( o_f(x,y) \) and \( o_b(x,y) \) are the separated foreground and background objects from \( o(x,y) \), and \( h_f(x,y) \) and \( h_d(x,y) \) are the 2-D focus and defocus PSFs.
\[ i_f(x, y) = h_f(x, y) \cdot o_f(x, y) + n_f(x, y) \]  
\[ i_b(x, y) = h_d(x, y) \cdot o_b(x, y) + n_b(x, y) \]  
(4) can be obtained by substituting (3) into (2a), and it can be further generalized to (5).
\[ u(x, y) = \left\{ \begin{array}{l} m_f(x, y) \cdot h_f(x, y) + m_b(x, y) \cdot h_d(x, y) \cdot o_f(x, y) + n_f(x, y) + \left[ m_f(x, y) \cdot h_f(x, y) + m_b(x, y) \cdot h_d(x, y) \cdot o_b(x, y) + n_b(x, y) \right] \end{array} \right\} \]  
Accoding to (2) to (5), for MFIF, its purpose is to use \( i_f(x,y) \) and \( i_b(x,y) \) to obtain the \( \hat{d}(x,y) \) for image fusion. It reveals that MFIF and SR share the unified physical model. Therefore, MFIF and blind SR tasks can be combined as a multi-focus image super resolution fusion (MFISRF) task.

Here, we propose deep fusion prior (DFP), which implements MFIF and blind SR with unified, unsupervised, dataset-free and robust model to deal with the MFISRF task. DFP consists of SKIPnet network, DoubleReblur focus measurement tactic, decision embedding module and loss functions. Foremost, the main contributions of the proposed DFP are summarized as follows:

- In theory, we unify the MFIF and SR tasks as a MFISRF task via describe physical optical model to propose a new perspective for MFIF and SR.
- In framework, we integrate the physical optical model of MFISRF into deep image prior, and design a unified, dataset-free, unsupervised and robust framework DFP to address the MFISRF task. DFP works with our
A. MFIF Methods

For MFIF, methods can be classified into non-deep learning and deep learning types. In non-deep learning approaches, MFIF can be roughly summarized as an inverse process of extraction, and the key to these approaches lies in two important aspects: focus measurement and fusion rules. MFIF focus measurement can be implemented in either spatial domain or transform domain. There are mainly 3 tactics in focus measurement in spatial domain such as pixel- [2], [3], [4], [5], [6], [7], [8], block-[9] and region-based ones [10]. Although spatial domain methods can obtain focus measurement, they are not generalizable enough, focus measurement approaches in transform domain can improve the performance. There are mainly 5 tactics in focus measurement in transform domain such as sparse representation-based [11], [12], multi-scale-based [13], [14], [15], [16], [17], gradient-based [18], [19], feature-space-based [20] and hybrid ones [21]. However, these transform domain methods are mostly based on Laplacian pyramid, gradient pyramid, discrete wavelet (DWT), discrete cosine (DCT), curvelet transform (CVT) and so on. These handcrafted focus measurement approaches make fusion methods complex, thus intensifying the difficulty of designing fusion rules. Meanwhile, these handcrafted focus measurement approaches also cause edge fragmentation, obvious transition and many other problems. Besides, fusion rules are determined based on focus measurement. The often used rules include maximum, minimum, addition, 11-norm, and so on. However, the limit choices of these handcrafted fusion rules produce a glass ceiling on the performance improvement even in some deep learning-based methods.

Deep learning-based MFIF can often obtain end-to-end fusion, since they use networks act as fusion rules. There are mainly two categories as supervised and unsupervised learning-based approaches. For the supervised category, CNN [22] is the first learning-based model realizing end-to-end MFIF. Subsequently, a series of CNNs are used in MFIF including DRPL [23], ECNN [24], IFCNN [25], MADCNN [26] and PCANet [27]. Although these supervised methods have good performance in MFIF, they highly rely on large handcrafted datasets. Moreover, handcrafted datasets often fail in real-world applications. Not relying on using handcrafted datasets, the unsupervised learning methods including FusionDN [28], GCF [29], SESF [30], MFF-GAN [1], PMGI [31] and U2Fusion [32] have been designed. However, because most of these unsupervised methods often try to sharpen the edges via object enhancement and gradient, the difference between fused image and ground truth is still significant. Besides, almost all of these methods still rely on large datasets to drive.

B. SR Methods

For SR, methods can be classified into interpolation-, example- and deep learning-based types. In these approaches, SR can be summarized as an ill-posed inverse process of imaging. The interpolation-based methods rely on sampling and interpolation, and they include nearest-neighbor, bilinear, bicubic, Sinc, Lanczos and so on. These interpolation-based methods only use the pixel information of low-resolution image itself, and the pixel at each position is interpolated based on the information around such pixel, so the reconstructed images are significantly blurry.

The example-based methods exploit transposed convolution [33], [34], sub-pixel [35] or meta upscale [36] to learn the mapping between the low-resolution and high-resolution patches from internal [37], [38], [39], [40] or external [41] datasets. Although the internal dataset-based methods can improve the quality, it is difficult to learn high-frequency information and large textural variations, and significantly slow to learn.

![Fig. 1. Illustration of the key idea of DFP.](image)

Fig. 1. Illustration of the key idea of DFP. \(i_1\) is the foreground image, \(i_2\) is the background image, \((i_1+i_2)/2\) is the average of the interpolated foreground and background images, decision map can be obtained according to the DoubleReblur focus measurement tactic, and the output is the MFISRF result.
Moreover, the generated reconstruction images often have marked serration. Therefore, external dataset-based methods are used to replenish high frequency from external images. Although these external dataset-based methods are efficient, the feature extraction and mapping functions are handcrafted, both limiting high-frequency information learning.

In order to overcome the limitations of above approaches, deep learning-based approaches are proposed. There are mainly two categories as supervised and unsupervised learning-based approaches. Since Dong et al. [42] pioneered the first supervised deep learning-based single image SR approach, many end-to-end CNNs [42], [43], [44], [45], [46], [47] have been designed to improve the SR quality, and the key to these approaches lies in two important aspects: network backbone [48], [49], [50], [51], [52], [53], [54] and upsampling order [57], [58], [59]. But almost all of these supervised deep learning-based approaches rely on the large datasets consisting of fixed-system low-resolution and high-resolution image pairs. Unsupervised SR approaches are mainly based on adversarial generative networks [60], [61], [62], thus they are not fully unsupervised and still rely on large datasets. Deep-image-prior [64] (DIP) shows that the structure of generator deep neural networks is sufficient to capture a great deal of image statistical prior. It only needs low-resolution image as inputs to reconstruct high-resolution image at any scale. Like TV norm [63], DIP is an effective handcrafted prior. It is fully unsupervised without any external datasets and performs well dealing with textured and focused goals, although, not well for defocus goals.

Different from CF-Net [72] achieving multi-exposure image fusion (MEF) and SR simultaneously; up to now, there is still no unified model that can implement both MFF and SR tasks with only one network. Besides, almost all of the supervised and unsupervised deep learning-based methods require large datasets for training in both MFF and SR tasks. Therefore, according to our unified MFISRF model explained in Eqs. (2) to (5), a unified, dataset-free and unsupervised method for MFISRF task should be considered.

III. DEEP FUSION PRIOR

In this section, we introduce the details of our proposed DFP. Based on our unified MFISRF model, the unsupervised DFP consists of SKIPnet, DoubleReblur, decision embedding and loss functions, which are detailed illustrated in the following.

A. Architecture

It is well-known that the encoder-decoder framework is effective for generative works. In addition, U-Net [65] structures have been proved excellent for tasks such as image generation and segmentation. However, to guarantee the image-to-image translation invariance, there is no padding before convolution in U-Net. Because of this, the shape between input and output is inapparity. Moreover, this inapparity also influences on the feature concatenation. The input in SKIPnet is the average interpolated low-resolution images or even the random noise. Therefore, reflection paddings and 1x1 convolutions are adopted to keep the shape of feature map fixed. Leaky-ReLU activation function is used to improve the SKIPnet. Note that this work focuses on proposing a powerful and flexible unsupervised MFISRF method without any dataset rather than proposing new generator network architecture. As a matter of fact, the similar architectures of autocoder network can be found in [64], [66].

The details of SKIPnet architecture is illustrated in Fig. 2. Based on the encoder-decoder framework, the SKIPnet is divided into the encoder and decoder parts, and they are symmetric about the central feature map $φ_C$. The backbone of the encoder part is composed of $D$ encoder-blocks which extract feature maps in $D$ scales. Each block consists of a reflection padding preparing layer, a 1-stride $n_pφ_C$ convolution extracting layer, a batch normalization processing layer, a leaky-ReLU activating layer, a reflection padding preparing layer, a 2-stride $n_pφ_C$ convolution downsampling layer, a batch normalization processing layer and a leaky-ReLU activating layer, successively. Iden tically, the backbone of the decoder part is composed of $D$ decoder-blocks which extract and fuse feature maps in $D$ scales. Each block consists of a bilinear upsampling layer, a batch normalization processing layer, a reflection padding preparing layer, a 1-stride $n_pφ_C$ convolution extracting layer, a batch normalization processing layer, a leaky-ReLU activating layer, a reflection padding preparing layer, a 1-stride $n_pφ_C$ convolution extracting layer, a batch normalization processing layer and a leaky-ReLU activating layer, successively. For multi-scale feature map fusion, encoder feature maps $φ_C(n)$ are concatenated to decoder feature maps $φ_C^δ(n)$. In the end, a 1x1 convolution is adopted to reduce dimensionality. A sigmoid activation function is adopted to obtain the demanded output format. In SKIPnet architecture, the downsampler with conventional approaches such as Bilinear, Bicubic and Lanczos can be used to obtain the same size output as inputs, and the scale of the SR depends on the scale of the downsampler. Moreover, the depth $D$ and the convolution kernel size $n_p$ are adjustable. For convenient parameter adjustment, $D=5$ and $n_p=5$ are used in DFP, and two 3x3 convolution layers can be used to replace the 5x5 convolution layer for higher efficiency. Although some CNN based methods including ResNet [53], DenseNet [54] and Residual Dense Network [46] have been proved to work better on extracting features than autocoder architecture, they often have poor performance in image prior extraction. However, the image prior relying on handcrafted network architecture such as encoder-decoder still works well.
B. DoubleReblur

As we known, most MFIF methods \cite{4, 5, 6, 7, 8, 10, 15, 22, 23} are based on decision maps, which can be obtained by handcrafted focus measurement. Therefore, combining with computational imaging, morphological image processing and graphics, we propose a DoubleReblur focus measurement model as shown in Fig. 3.

![DoubleReblur scheme](Image)

Fig. 3. DoubleReblur scheme

According to the imaging model in (1), the relation between the foreground input \( g(x,y) \) and the background input \( g(x,y) \) can be represented as (11).

\[
i_b(x,y) = i_f(x,y) * h_b(x,y) + n_b(x,y)
\]

To estimate the spread kernel \( h_b(x,y) \), Fourier transform is implemented on both sides of (11) as follows:

\[
\mathcal{F}\{i_b(x,y)\} = \mathcal{F}\{i_f(x,y) * h_b(x,y) + n_b(x,y)\}
\]

\[
I_b(\xi,\eta) = I_f(\xi,\eta) * H_b(\xi,\eta) + N_b(\xi,\eta)
\]

where \( \mathcal{F} \) is the Fourier transform, \( I_f(\xi,\eta), I_b(\xi,\eta) \) and \( N_b(\xi,\eta) \) are the frequency spectra of \( i_b(x,y) \), \( i_f(x,y) \), \( h_b(x,y) \) and \( n_b(x,y) \), respectively.

\[
\frac{I_b(\xi,\eta)}{I_f(\xi,\eta)} = \frac{Y_b(\xi,\eta) * H_b(\xi,\eta) + N_b(\xi,\eta)}{Y_f(\xi,\eta) * H_b(\xi,\eta)}
\]

To remove the noise term, a low-pass filter \( T \) is performed on (13). In the end, the estimate of spread kernel \( \hat{h}_b(x,y) \) can be obtained by follows:

\[
\hat{h}_b(x,y) = \mathcal{F}^{-1}\left\{T\left\{\frac{I_b(\xi,\eta)}{I_f(\xi,\eta)} - \frac{N_b(\xi,\eta)}{I_f(\xi,\eta)}\right\}\right\}
\]

The first reblur image can be obtained by (15).

\[
\tilde{I}_b(x,y) = i_b(x,y) * \hat{h}_b(x,y)
\]

Then, the Gaussian reblur \( \mathcal{G} \) as the second reblur is used to enhance the sharpness difference as \( s(x,y) \) obtained by (16).

\[
s(x,y) = |\tilde{I}_b(x,y) - \mathcal{G}\{|\tilde{I}_b(x,y)\}|
\]

\( d(x,y) \) can be obtained by using threshold segmentation on sharpness difference image \( s(x,y) \).

\[
d(x,y) = \begin{cases} 1, & s(x,y) > t \\ 0, & s(x,y) \leq t \end{cases}
\]

To eliminate gaps and holes, dilating and eroding are implemented as closed operation

\[
d(x,y) = d(x,y) \bullet E
\]

The decision map \( m(x,y) \) can be finally obtained using the largest region floodfill algorithm.

\[
m(x,y) = C\{d(x,y)\}
\]

There are 5 parameters in DoubleReblur as \([k_b, k_d, k_s, t, f] \): \( k_b \) is the kernel size of Gaussian blur, \( k_d \) is the kernel size of dilating, \( k_s \) is the kernel size of segmentation, and \( f \) is the bool flag control if the largest region filling is used. We transform the image to Y channel and use \([5, 3, 3, 0.01, 1] \) for most images and slightly adjust parameters for different image details.

C. Decision Embedding

Since the decision maps are obtained from handcrafted focus measurement, they suffer from disadvantages such as edge fragmentation and false determination. In order to optimize the handcrafted decision maps, learning-based decision embedding is designed as Fig. 4. The input is obtained by the average of low-resolution inputs but can also be replaced by random noise. The binarized decision map can be generated bySKIPnet. Moreover, we design an optimized loss function \( \mathcal{L}_{opt} \) as (20) to optimize the decision map via minimizing the difference between focus measurement regions and original low-resolution inputs.

\[
\mathcal{L}_{opt} = \frac{1}{H \cdot W} \sum_i \sum_j |\hat{m}_{i,j} - m_{i,j}| + |\hat{m}_{i,j} \cdot I_{fore_{i,j}} - I_{fore_{i,j}}|
\]

where \( H \) and \( W \) are the height and width of the image, \( I_{fore} \) and \( I_{back} \) are the foreground and background low-resolution inputs, \( m \) is the handcrafted decision map obtained by DoubleReblur, \( \hat{m} \) is the optimized decision map generated by SKIPnet. Algorithm 2 briefly demonstrates the decision embedding process. The decision embedding module is not indispensable, but the DFP performance will be better when considering it.

![Decision embedding scheme and examples](Image)
E. Loss Functions

The designed loss functions as (21) are composed of content loss $L_{\text{con}}$, joint gradient loss $L_{\text{grad}}$, and gradient limit loss $L_{\text{grad}}$. The content loss is to constrain the SKIPnet to learn the clear region of each image. The joint gradient loss is to enhance the SKIPnet to learn the high-frequency information and to reduce the dependency on decision maps. The gradient limit loss is to reduce the noise and oscillation effects. $\alpha$, $\beta$, and $\gamma$ in (21) are weighted parameters, and set as 1, 0.5 and 0.1.

$$L = \alpha L_{\text{con}} + \beta L_{\text{grad}} + \gamma L_{\text{grad}}$$  

$$L_{\text{con}} = \frac{1}{H \cdot W} \sum_{i,j} \lambda_i \left| \hat{I}_{\text{MFISRF},i,j} - I_{\text{fore},i,j} \right|$$  

$$L_{\text{grad}} = \lambda_1 \left| \nabla \cdot \left( \lambda_1 \hat{I}_{\text{MFISRF},i,j} - I_{\text{fore},i,j} \right) \right| + \lambda_2 \left| \nabla \cdot \left( \lambda_2 \hat{I}_{\text{MFISRF},i,j} - I_{\text{back},i,j} \right) \right|$$  

$\lambda_1$ and $\lambda_2$ are the weighted parameters both equal to 1. For generated MFISRF image $\hat{I}_{\text{MFISRF}}$, the focus regions in $I_{\text{fore}}$ and $I_{\text{back}}$ are obtained via decision maps. $L_1$ norms between the foreground focus region of $\hat{I}_{\text{MFISRF}}$ and $I_{\text{fore}}$, and between the background focus region of $\hat{I}_{\text{MFISRF}}$ and $I_{\text{back}}$ are used to compute their distances. Rather than $L_1$ norm, $L_2$ norm has stronger penalty for large errors and weaker penalty for small errors, and ignores the effect of the image content itself. It is worth noting the human visual system (HVS) is especially sensitive to brightness and color changes in untextured areas of the image. Moreover, [68] has proved that $L_1$ norm performs better than $L_2$ norm in SR tasks. So $L_1$ norm is adopted here.

$$L_{\text{grad}} = \frac{1}{H \cdot W} \sum_{i,j} \left| \Delta \hat{I}_{\text{MFISRF},i,j} \right|$$  

$\Delta \hat{I}_{\text{MFISRF},i,j}$ in (23) is the joint gradient loss, where $\Delta_L$ is the gradient obtained via Laplacian. The difference between focus and defocus regions is almost the high-frequency information. Moreover, the Laplacian gradient map can well describe high-frequency information. For MFIF, the maximal Laplacian gradient map of unfused inputs almost equals to the fused image. For SR, the Laplacian gradient map can provide more high-frequency details. Therefore, $L_1$ norm between the Laplacian gradient map of generated MFISRF image $\hat{I}_{\text{MFISRF}}$ and the joint maximal Laplacian gradient map of low-resolution inputs $I_{\text{fore}}$ and $I_{\text{back}}$ is used to compute their distance.

$\Delta \hat{I}_{\text{MFISRF},i,j}$ in (24) is the gradient limit loss. Unfortunately, DFP exploits the self-similarity property of input images themselves for reconstruction, thus inducing noise and oscillation. The gradient limit loss is used to reduce these noise and oscillation effects.

$$L_{\text{grad}} = \frac{1}{H \cdot W} \sum_{i,j} \left( \nabla \cdot \hat{I}_{\text{MFISRF},i,j} + \nabla \cdot \hat{I}_{\text{MFISRF},i,j} \right)$$  

IV EXPERIMENTAL RESULTS

A. Qualitative Experimental Results

First, we qualitatively compared our proposed DFP with the combinations of learning-based MFIF (CNN [22], PCANet [27], FusionDN [28], SESF [30], PMGI [31], U2Fusion [32]) and conventional SR (Bicubic/unsupervised SR (DIP [64])/supervised SR (SRCNN [41]). These methods rely on the combinations of MFIF and SR work with two models and learn from large datasets. While our proposed DFP only works with one model and learns without any dataset.

Algorithm 2 Decision Embedding

1. Input: foreground input $I_{\text{fore}}$, background input $I_{\text{back}}$, input $n$ and handcrafted decision map $m$ which obtained by DoubleReblur;
2. for iterations $K=500$ do
3. Forward: $\hat{m} = \text{SKIPnet}(\text{input } n)$;
4. for $x, y$ in $\hat{m}$ do
5. if $\hat{m}(x, y) > 0.5$, $\hat{m}(x, y) = 1$;
6. else $\hat{m}(x, y) = 0$;
7. end for
8. $L_{\text{opt}} = \frac{1}{H \cdot W} \sum_{i,j} \left| m_{ij} \left( \hat{m}_{ij} - I_{\text{fore},ij} \right) \right|$
9. Backward to update SKIPnet parameters $\Theta$; end for

Fig. 5 reveal the visual results of MFISRF x2 and x4 on MF1-WHU benchmark evaluation dataset [1]. In MFIF, CNN and PCANet provide similar results with almost discernible difference, but they all suffer from poor MFIF quality. PMGI improves the MFIF performance. It is compatible with multi-task such as infrared and visible image fusion, multi-exposure image fusion and medical image fusion. But these tasks are often in low contrast cases, so the PMGI results are often color infidelity. While SESF slightly improves the MFIF performance compared to PMGI by increasing sharpness and reducing the blur, still has poor fusion quality. FusionDN and its upgraded U2Fusion both compatible with multi-tasks have better MFIF performance than above methods. However, they still cannot completely solve problems such as dark, blurred and enhanced edge in MFIF. In SR, DIP improves image quality to some extent. But high-frequency details are lost especially in defocus and vista regions, since only self-similarity and low-level statistical priors are used. Although SRCNN performs better than DIP, it relies on training with large datasets which composed of fixed low- and high-resolution image pairs. Generally, it is easy to note that the unsupervised methods (FusionDN, SESF, PMGI) perform better than the supervised methods (CNN, PCANet) in MFIF. But contrarily, the supervised method (SRCNN) performs better than the unsupervised methods (Bicubic, DIP) in SR. Even DFP does not use any datasets for training, it still achieves high-quality results comparable to those obtained via the combinations of the unsupervised MFIF (FusionDN, SESF, PMGI, U2Fusion) and the supervised SR (SRCNN).

B. Quantitative Experimental Results

Besides, quantitative comparisons are performed still using the MF1-WHU benchmark evaluation dataset [1]. The evaluation metrics include mean gradient (MG/AG) [31], [32], [71], edge intensity (EI) [32], information entropy (IE/EN) [1], [28], mean gray value (MGA) and polar edge coherence (ECO) [69], respectively. MG reflects the rate of
Fig. 5. We compare our approach against multiple state-of-the-art supervised and unsupervised (S. & unS.) learning-based MFIF (CNN [22], PCANet [27], FusionDN [28], SESF [30], PMGI [31], U2Fusion [32]) and (S. & unS.) SR (Bicubic, DIP [64] and SRCNN [41]) in ×2 condition on MFI-WHU 26 and ×4 condition on MFI-WHU 08.

contrast change of tiny details in the image. MG\textsubscript{R3} describes the relative MG between ground truth (GT) and MFISRF.

The most basic feature of an image is its edge existing between target and background, so it is one of the most important features to evaluate image fusion. EI is a quantitative coefficient to describe the edge information. EI\textsubscript{R3} describes the relative EI between GT and MFISRF. IE reflects the comprehensive characteristics of gray value at a pixel and its surrounding pixel gray distributions. IE\textsubscript{R3} is the relative value between GT and MFISRF. MGA shown in (25a) is the average level of image gray, which represents the overall brightness level of the image. MGA\textsubscript{R3} describes the relative MGA between GT and MFISRF as demonstrated in (25b).
\[ MGA = \frac{1}{H \cdot W} \sum_{i} \sum_{j} I_{i,j} \]  
\[ MGA_{(R)} = \frac{MGA_{MFISRF}}{2} - \frac{MGA_{GT}^{1} + MGA_{GT}^{2}}{2} \]

ECO reflects the edge similarity of the reconstructed and original images. It is illustrated in (26), where ECO is edge coherence according to [69], and C is a regularisation constant. ECO\(_{(R)}\) describes the relative ECO between GT and MFISRF.

\[ ECO_{(R)} = \frac{2 - \frac{ECO_{MFISRF}}{ECO_{GT}^{1} + C} - \frac{ECO_{MFISRF}}{ECO_{GT}^{2} + C}}{ECO_{GT}^{1} + C} \]  

We still compare our DFP with the combinations of 6 MFIF methods and 3 SR methods the same as above qualitative comparisons. 6 MFIF methods include 6 supervised learning-based methods (CNN, DRPL, ECNN, IFCNN, MADCNN and PCANet) and 6 unsupervised learning-based methods (FusionDN, GCF, SESF, MFF-GAN, bMGI and U2Fusion). 3 SR methods include conventional Bicubic, unsupervised learning-based DIP and supervised learning-based SRCNN. All methods except DIP are trained based on large datasets. However, DFP works without dataset.

Tables 1-3 quantitatively compare the performances using different MFIF and SR combined methods and the proposed MFISRF one according to above mentioned coefficients. Red marks the 1\(^{st}\) best of the performance and bold represents 2\(^{nd}\) and 3\(^{rd}\) best of the performance. Our proposed MFISRF often can achieve high-quality super-resolved multi-focus fused images in both ×2 and ×4 conditions. Especially, even compared to the optimized combinations of unsupervised MFIF methods and supervised SR method, our proposed DFP could still obtain rather low RMG, REI, RIE, RMGA and RECO values very close to or even lower than those obtained by the optimized MFIF and SR combined methods.

\[ \text{Fig. 6. We compare our approach against the combinations of FusionDN [28], MFF-GAN [1], PMGI [31], U2Fusion [32] and bicubic SR in ×2 condition on MFI-WHU [1] and Lytro [67].} \]
decision maps are good or bad. However, high-quality SR performance still depends on correct decision maps. Moreover, we also test the convergence details of DFP and DIP still on the same MIJ-WHU 29 with 3000 iterations. As DFP is multi-image SR while DIP is single-image SR, while the convergence is more focused. According to the results, DFP converges only using ~300 iterations; however, DIP does not converge until ~3000 iterations. Additionally, DFP has a smoother trend with less fluctuations compared to DIP. Results in Fig. 7 demonstrate the proposed DFP is robust in MFISRF tasks.

![Image](image_url)

Fig. 7. We study the influence of decision map on both MFIF and SR performances. The first column images are the foreground and background ground truths. The second column results are obtained by DFP with the DoubleReblur parameter set as [3,5,5,0,0.05,1] and without weighted joint gradient loss. The third, fourth and fifth column images are obtained by DFP but with different DoubleReblur parameter sets as [3,5,5,0,1], [3,5,5,0,01,1] and [3,5,5,0,05,1], respectively. In these columns, the first image is the foreground decision map, the second image is the MFISRF image, the third one is the zoomed-in field-of-interest, and the last one is the pseudo-color image depicting the difference of ground truth and MFISRF result in Y channel.

TABLE I
WE QUANTITATIVELY COMPARE OUR APPROACH AGAINST MUTIPLE STATE-OF-THE-ART LEARNING-BASED MFIF AND BICIRCULAR IN ×2 AND ×4 CONDITIONS. RED, GREEN AND BLUE COLORS MARK THE 1ST PERFORMANCE IN RED, 2ND AND 3RD BEST IN BLACK BOLD.

| Method | MG (R) (%) | El (R) (%) | IE (R) (%) | MGA (R) (%) | ECO (R) (%) |
|--------|------------|------------|------------|-------------|-------------|
|        | x2         | x4         | x2         | x4          |             |
| CNX+  | 0.7207     | 3.6682     | 0.0176     | 0.1692      | 0.0844      |
| BICIRCULAR | 2.5936   | 22.7946    | 0.0598     | 0.1723      | 0.0548      |
| PCA+Net | 0.7444  | 3.8778     | 0.0174     | 0.2240      | 0.0196      |
| BICIRCULAR | 2.6060   | 22.9440    | 0.0612     | 0.2252      | 0.0283      |
| FusionDNSF+ | 0.9637   | 12.3182    | 0.3491     | 16.4175     | 0.0014      |
| BICIRCULAR | 1.7848   | 14.4909    | 0.3251     | 16.3919     | 0.0284      |
| SENS+ | 0.7138     | 3.5747     | 0.0159     | 0.1735      | 0.0214      |
| BICIRCULAR | 2.5856   | 22.7460    | 0.0591     | 0.1866      | 0.0290      |
| PAMG+ | 2.3111     | 18.2135    | 0.3899     | 18.1809     | 0.1422      |
| BICIRCULAR | 2.9701   | 27.0107    | 0.3921     | 18.1691     | 0.1434      |
| U2Fusionx+ | 0.9618   | 15.1490    | 0.2975     | 5.3168      | 0.0955      |
| BICIRCULAR | 1.4366   | 9.3573     | 0.2461     | 5.3213      | 0.1489      |
| DFP (ours) | 0.5840 | 6.3614     | 0.0222     | 0.5324      | 0.0056      |
|        | 1.8111     | 14.8770    | 0.0376     | 0.5536      | 0.0062      |

In this study, we unify the MFIF and blind SR problems as the MFISRF task, and propose a novel unified dataset-free unsupervised framework DFP to address such MFISRF task. To our best knowledge, our proposed work is a dataset-free unsupervised method to jointly implement the multi-focus fusion and super-resolution task for the first time. DFP consists of our designed SKIPnet end-to-end generated network to implement unsupervised learning via deep image prior, DoubleReblur tactic for focus measurement based on estimated PSF and Gaussian kernel convolution, decision embedding learned module for decision map optimization, and loss functions to guarantee high-quality MFISRF results robustly. Compared to 6 MFIF and 3 SR method combinations including both supervised and unsupervised ones, the proposed unsupervised dataset-free DFP approaches or even outperforms these state-of-art MFIF and SR method combinations. Furthermore, DFP is a general framework, thus its networks and focus measurement tactics can be continuously updated to further improve the MFISRF performance. We believe DFP can be potentially used in various computational photography applications.

ACKNOWLEDGMENT
This work was supported by National Natural Science Foundation of China (61705092); Natural Science Foundation of Jiangsu Province of China (BK20170194).

TABLE II
WE QUANTITATIVELY COMPARE OUR APPROACH AGAINST MULTIPLE STATE-OF-THE-ART LEARNING-BASED MFIF AND UNSUPERVISED LEARNING-BASED SR IN ×2 AND ×4 CONDITIONS. RED, GREEN AND BLUE COLORS MARK THE 1ST PERFORMANCE IN RED, 2ND AND 3RD BEST IN BLACK BOLD.

| Method | MG (R) (%) | El (R) (%) | IE (R) (%) | MGA (R) (%) | ECO (R) (%) |
|--------|------------|------------|------------|-------------|-------------|
|        | x2         | x4         | x2         | x4          |             |
| CNX+  | 0.5700     | 2.6387     | 0.0310     | 0.3459      | 0.0063      |
| DIP   | 2.0613     | 17.1756    | 0.0599     | 0.3378      | 0.0052      |
| PCA+Net | 0.6014    | 2.7950     | 0.0268     | 0.3261      | 0.0012      |
| DIP   | 2.0911     | 17.5025    | 0.0559     | 0.2713      | 0.0233      |
| FusionDNSF+ | 1.1171  | 14.3756    | 0.3529     | 16.7372     | 0.0158      |
| DIP   | 1.2079     | 9.5311     | 0.3429     | 16.7444     | 0.0311      |
| SENS+ | 0.5553     | 2.6325     | 0.0275     | 0.3093      | 0.0244      |
| DIP   | 0.0534     | 17.0854    | 0.0474     | 0.3486      | 0.0245      |
| PMG+  | 2.2314     | 19.5217    | 0.4057     | 18.5801     | 0.1329      |
| DIP   | 2.6718     | 24.0668    | 0.4118     | 18.5721     | 0.1468      |
| U2Fusionx+ | 1.1756   | 17.5288    | 0.2929     | 5.2692      | 0.0940      |
| DIP   | 0.3607     | 3.9685     | 0.2739     | 5.3012      | 0.1389      |
| DFP (ours) | 0.5840 | 6.3014     | 0.0222     | 0.5324      | 0.0056      |
|        | 1.8111     | 14.8770    | 0.0376     | 0.5536      | 0.0062      |

TABLE III
WE QUANTITATIVELY COMPARE OUR APPROACH AGAINST MULTIPLE STATE-OF-THE-ART LEARNING-BASED MFIF AND SUPERVISED LEARNING-BASED SR IN ×2 AND ×4 CONDITIONS. RED, GREEN AND BLUE COLORS MARK THE 1ST PERFORMANCE IN RED, 2ND AND 3RD BEST IN BLACK BOLD.
REFERENCES

[1] 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,” Inf. Fusion, vol. 66, no. September, 2020, pp. 40–53, 2021.

[2] J. Tian, L. Chen, L. Ma, and W. Yu, “Multi-focus image fusion using a bilateral gradient-based sharpness criterion,” Opt. Commun., vol. 284, no. 7, pp. 80–87, 2011.

[3] B. K. Shreyamsukaumar, “Image fusion based on pixel significance using cross bilateral filter,” Signal, Image Video Process., vol. 9, no. 5, pp. 1193–1204, 2015.

[4] Y. Liu, S. Liu, and Z. Wang, “Multi-focus image fusion with dense SIFT,” Inf. Fusion, vol. 23, pp. 139–155, 2015.

[5] X. Qiu, M. Li, L. Zhang, and X. Yuan, “Guided filter-based multi-focus image fusion through focus region detection,” Signal Process. Image Commun., vol. 72, no. December, 2018, pp. 35–46, 2019.

[6] S. Li, X. Kang, J. Hu, and B. Yang, “Image matting for fusion of multi-focus images in dynamic scenes,” Inf. Fusion, vol. 14, no. 2, pp. 147–162, 2013.

[7] J. Ma, Z. Zhou, B. Wang, and M. Dong, “Multi-focus image fusion based on multi-scale focus measures and generalized random walk,” Chinese Control Conf. CCC, no. 1, pp. 5464–5468, 2017.

[8] J. Ma, Z. Zhou, B. Wang, L. Miao, and H. Zong, “Multi-focus image fusion using boosted random walks-based algorithm with two-scale focus maps,” Neurocomputing, vol. 335, pp. 9–20, 2019.

[9] X. Bai, Y. Zhang, F. Zhou, and B. Xue, “Quadtree-based multi-focus image fusion using a weighted focus-measure,” Inf. Fusion, vol. 22, pp. 105–118, 2015.

[10] Y. Zhang, X. Bai, and T. Wang, “Boundary finding based multi-focus image fusion through multi-scale morphological focus-measure,” Inf. Fusion, vol. 35, pp. 81–101, 2017.

[11] Y. Liu and Z. Wang, “Simultaneous image fusion and denoising with adaptive sparse representation,” IEEE Image Process., vol. 9, no. 5, pp. 1543–1551, 2015.

[12] Y. Liu, X. Chen, R. K. Ward, Z. J. Wang, and S. Member, “Sparse Representation,” Comput. Vis., vol. 23, no. 12, pp. 748–748, 2014.

[13] M. Amin-Naji and A. Aghagolzadeh, “Multi-Focus Image Fusion in DCT Domain using Variance and Energy of Laplacian and Correlation Coefficient for Visual Sensor Networks,” J. AI Data Min., vol. 6, no. 2, pp. 233–250, 2018.

[14] W. Z. Liu, “Multi-focus image fusion based on wavelet transform and adaptive block,” J. Image Graph., vol. 8, no. 11, pp. 1435–1444, 2013.

[15] S. Li, X. Kang, and J. Hu, “Image fusion with guided filtering,” IEEE Trans. Image Process., vol. 22, no. 7, pp. 2864–2875, 2013.

[16] D. P. Bavirisetti, G. Xia, J. Zhao, R. Dhuli, and G. Liu, “Multi-scale Guided Image and Video Fusion: A Fast and Efficient Approach,” Circuits, Syst. Signal Process., vol. 38, no. 12, pp. 5576–5605, 2019.

[17] M. Amin-Naji, P. Ranjar-Neiye, and A. Aghagolzadeh, “Multi-focus image fusion using Singular Value Decomposition in DCT domain,” Iran. Conf. Mach. Vis. Image Process. MVIP, vol. 2017-November, pp. 45–51, 2018.

[18] S. Paul, I. S. Sevencan, and P. Agathoklis, “Multi-exposure and multi-focus image fusion in gradient domain,” J. Circuits, Syst. Comput., vol. 25, no. 10, pp. 1–18, 2015.

[19] Z. Zhou, S. Li, and B. Wang, “Multi-scale weighted gradient-based fusion for multi-focus images,” Inf. Fusion, vol. 20, no. 1, pp. 60–72, 2014.

[20] H. Li, L. Li, and J. Zhang, “Multi-focus image fusion based on sparse feature matrix decomposition and morphological filtering,” Opt. Commun., vol. 342, no. 1–11, 2015.
Yuanjie Gu (IEEE graduate student member) received the B.E. degree from the School of Petroleum, China University of Petroleum, China, in 2020. He is currently a master student in the Computational Optics Laboratory, School of Science, Jiangnan University, Wuxi. His current research interests include computational photography, unsupervised deep learning, image fusion, super-resolution and single molecule imaging.

Zhibo Xiao received the B.E. degree from the School of Petroleum, China University of Petroleum, China, in 2021. She is currently a master student in the Computational Optics Laboratory, School of Science, Jiangnan University, Wuxi. Her current research interests include computational photography and deep learning.

Hailun Wang received the B.E. degree from the School of Petroleum, China University of Petroleum, China, in 2021. She is currently a master student in the Computational Optics Laboratory, School of Science, Jiangnan University, Wuxi. Her current research interests include computational photography and deep learning.

Cheng Liu received the B.E. and M.E. degrees from Soochow University in 1993 and 1996, respectively, and the Ph.D. degree from Shanghai Institute of Optics and Fine Mechanics in 2003. From 2003 to 2011 he was a postdoctoral associate with Yonsei University, National University of Singapore, Wgangju Institute of Science and Technology, and University of Sheffield. Since 2011, he becomes a full professor in Jiangnan University and leads the Computational Optics Laboratory. He is also affiliated with Shanghai Institute of Optics and Fine Mechanics. He has published more than 150 journal articles and conference papers. His current research interests include computational imaging, optical testing and microscopy.

Shuyang Wang received the B.E. and Ph.D degrees from Nanjing University of Science and Technology in 2010 and 2015. He is currently an associate professor and co-principal investigator in the Computational Optics Laboratory, School of Science, Jiangnan University, Wuxi. He is also affiliated with Single Molecule Nanometry Laboratory in Nanjing Agricultural University. He has published more than 100 journal articles and conference papers. His current research interests include computational sensing and imaging, microscopy and single molecule techniques.