Deep fusion prior for plenoptic super-resolution all-in-focus imaging

Yuanjie Gu,a,b,c Yinghan Guan,a,b,d Zhibo Xiao,a,b,c Haoran Dai,e, Cheng Liu,f and Shouyu Wang,a,g,*

aWuxi University, School of Electronics and Information Engineering, OptiX+ Laboratory, Wuxi, China
bHorizonFlow Laboratory, Wuxi, China
cJiangnan University, Computational Optics Laboratory, Wuxi, China
dChongqing Three Gorges University, School of Electronic and Information Engineering, Chongqing, China
eBoston University, Metropolitan College, Department of Computer Science, Boston, Massachusetts, United States
fChinese Academy of Sciences, Shanghai Institute of Optics and Fine Mechanics, Shanghai, China
gNanjing Agricultural University, Single Molecule Nanobiology Laboratory (Sinmolab), Nanjing, China

Abstract. Plenoptic imaging offers not only two-dimensional projections but also adds light array directions, thus supporting single-shot all-in-focus imaging. Its poor spatial resolution becomes an obstacle to high-quality all-in-focus imaging performance. Although various super-resolution (SR) methods have been designed and combined with multifocus image fusion (MFIF), high-quality multifocus fused SR images can be reconstructed for various applications, almost all of them deal with MFIF and SR separately. To our best knowledge, we first unify MFIF and SR problems as the multifocus image SR fusion (MFISRF) in the optical perspective and thus propose a dataset-free unsupervised framework named deep fusion prior (DFP) to address such MFISRF, particularly for plenoptic SR all-in-focus imaging. Both numerical and practical experiments have proved that our proposed DFP approaches or even outperforms those state-of-the-art MFIF and SR method combinations. Therefore, we believe DFP can be potentially used in various computational photography applications. The DFP codes are open source and available at http://github.com/GuYuanjie/DeepFusionPrior. © 2022 Society of Photo-Optical Instrumentation Engineers (SPIE) [DOI: 10.1117/1.OE.61.12.123103]

Keywords: all-in-focus imaging; super-resolution; unsupervised deep learning; dataset-free.

Paper 20221183G received Oct. 12, 2022; accepted for publication Dec. 8, 2022; published online Dec. 24, 2022.

1 Introduction

A majority of information acquisition, processing, and analysis is based on a 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 premise for precision decisions. Common cameras only record two-dimensional (2D) projections but lose depth information. In particular, when dealing with conditions with large depths of view, objects out of the depth of field of the imaging system become blurred and may induce errors in target distinguishing, recognition, and extraction. To break through the limit, all-in-focus imaging provides a solution. It first collects a multifocus image stack and then fuses these multifocus images into an all-in-focus one, where all objects are in focus.1 According to the multifocus image stack collection, there are mainly two tactics for all-in-focus imaging. One relies on the focus scanning of a common camera.2 In other words, sliced images can be captured by moving the imaging system or scanning the focus along the
optical axis. The advantage is that it can have high spatial resolution, but it sacrifices temporal resolution. Despite the use of many devices to accelerate focus scanning, such as electrically tunable liquid crystal lenses,\(^3\) microactuators,\(^4\) and optical diffusers,\(^7\) this method still cannot achieve single-shot all-in-focus imaging. Another tactic is based on numerically reconstructing a multifocus image stack from a single-shot coded-aperture image.\(^6\) Therefore, it completely solves the single-shot all-in-focus imaging problem. For example, plenoptic imaging is a representative technique. It offers not only 2D projections but also adds light array directions. Therefore, it can reconstruct multifocus images in color conditions and large fields of view and performs better than other coded-aperture-based techniques such as spectral focal sweep camera,\(^7\) color-filtered aperture,\(^8\) and lattice-focal lens.\(^9\) Relying on many reported super-resolution (SR) techniques,\(^10\) \(^11\) \(^12\) \(^13\) \(^14\) \(^15\) \(^16\) \(^17\) \(^18\) \(^19\) \(^20\) plenoptic imaging is a promising tool to support SR all-in-focus imaging in both high temporal and spatial resolution. Unfortunately, almost all of them deal with MFIF and SR separately. The separation inevitably induces complicated operations and long processing time.

Here, we propose deep fusion prior (DFP), which implements MFIF and blind SR with a unified, unsupervised, dataset-free, and robust model to deal with the MFSRF task, especially for plenoptic SR all-in-focus imaging. We compare our DFP with the combinations of six MFIF methods and three SR methods. Experiments have proved that DFP approaches or even outperforms those state-of-the-art MFIF and SR method combinations. In addition, DFP has been successfully employed for plenoptic SR all-in-focus imaging. In this manuscript, we first introduce the details of DFP in Sec. 2. Then, in Sec. 3, we compare DFP to other reported works in SR and MFIF. Next, we test the performance of DFP in plenoptic SR all-in-focus imaging in Sec. 4. Finally, we conclude this work in Sec. 5.

### 2 Deep Fusion Prior Method

#### 2.1 Unified Optical Model for MFIF and SR

Various MFIF and SR methods have been designed; however, almost all of them deal with MFIF and SR separately. However, we find that MFIF and SR share a unified optical 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 2D point spread function (PSF) \((h_f(x, y)\) and \(h_d(x, y)\) in the following are the 2D in-focus and defocus PSFs), \(o(x, y)\) is the object, \(n(x, y)\) is the additive noise (containing many kinds of noise such as dark current noise, shot noise, quantization error, CCD impulse noise, and so on), \(*\) is the spatial convolution, and \(i(x, y)\) is the image. Therefore, \(h(x, y)\) and \(n(x, y)\) are the generalized forms describing optical systems and noises. For SR, its purpose is to use \(i(x, y)\) to obtain an estimate \(\hat{o}(x, y)\) of the real object \(o(x, y)\).

While for MFIF, its model can be described in Eq. (2). \(i_f(x, y)\) and \(i_b(x, y)\) are the unfused images focusing on foreground and background. \(m_f(x, y)\) 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.

\[
u(x, y) = m_f(x, y) \cdot i_f(x, y) + m_b(x, y) \cdot i_b(x, y),
\]

\[
m_f(x, y) + m_b(x, y) = E.
\]

Both \(i_f(x, y)\) and \(i_b(x, y)\) can be represented by Eqs. (4) and (5), 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 2D in-focus and defocus PSFs

\[
i_f(x, y) = h_f(x, y) * o_f(x, y) + h_d(x, y) * o_b(x, y) + n_f(x, y),
\]

\[
i_b(x, y) = h_d(x, y) * o_f(x, y) + h_f(x, y) * o_b(x, y) + n_b(x, y).
\]
Equation (6) can be obtained by substituting Eqs. (4) and (5) into Eqs. (2) and (3), and it can be further generalized to Eq. (7). In Eq. (7), $\hat{h}(x, y)$ and $\hat{n}(x, y)$ are the general forms of $h(x, y)$ and $n(x, y)$, and the ideal object $o(x, y)$ is composed of foreground $o_f(x, y)$ and background $o_b(x, y)$. According to Eqs. (2) to (7), for MFIF, its purpose is to use $i_f(x, y)$ and $i_b(x, y)$ to obtain the $\hat{o}(x, y)$ for image fusion.

\[
\begin{align*}
    u(x, y) &= \left[ \left[ m_f(x, y) \cdot h_f(x, y) + m_b(x, y) \cdot h_b(x, y) \right] \ast o_f(x, y) \\
    &\quad + \left[ m_f(x, y) \cdot h_d(x, y) + m_b(x, y) \cdot h_f(x, y) \right] \ast o_b(x, y) \right] \\
    &\quad + m_f(x, y) \cdot n_f(x, y) + m_b(x, y) \cdot n_b(x, y), \\
    u(x, y) &= \hat{h}(x, y) \ast o(x, y) + \hat{n}(x, y).
\end{align*}
\]

(6) \hspace{1cm} (7)

It reveals that MFIF and SR share a unified optical model. Therefore, MFIF and blind SR tasks can be combined as a multifocus image SR fusion (MFISRF) task. Based on this unified optical model, we design the DFP as a novel unified dataset-free unsupervised framework to address the MFISRF task, particularly for plenoptic SR all-in-focus imaging.

### 2.2 Deep Fusion Prior Framework

Based on the unified MFISRF model mentioned above, the unsurprised DFP can obtain the high-quality fused SR result using a single model, and it is trained with only two low-resolution input images without any external dataset. As revealed in Fig. 1, DFP consists of our designed SKIPnet end-to-end generated network, 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.

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 $\phi_C$. The backbone of the encoder part is composed of $D$ encoder blocks that extract feature maps in $D$ scales. Each block consists of a reflection padding preparing layer, a one-stride $n_d \times n_d$ convolution extracting layer, a batch normalization processing layer, a leaky-ReLU activating layer, a reflection padding preparing layer, a two-stride $n_d \times n_d$ convolution downsampling layer, a batch normalization

![DFP framework](image-url)
processing layer, and a leaky-ReLU activating layer successively. Identically, the backbone of the decoder part is composed of \( D \) decoder blocks that 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 one-stride \( n_d \times n_d \) convolution extracting layer, a batch normalization processing layer, a leaky-ReLU activating layer, a reflection padding preparing layer, a one-stride \( n_d \times n_d \) convolution extracting layer, a batch normalization processing layer, and a leaky-ReLU activating layer successively. For multiscale feature map fusion, encoder feature maps \( \phi^{D_d}(n) \) are concatenated to decoder feature maps \( \phi^{U_d}(n) \). In the end, a \( 1 \times 1 \) convolution aims at reducing the dimensionality. A sigmoid activation function is adopted to obtain the demanded output format. In the SKIPnet architecture, the downsampler with conventional approaches such as bilinear, bicubic, and Lanczos can obtain the same size output as inputs, and the scale of the SR depends on the scale of the downsampler. The depth \( D \) and the convolution kernel size \( n_d \) are adjustable. For convenient parameter adjustment, \( D = 5 \) and \( n_d = 5 \) are used in DFP, and two \( 3 \times 3 \) convolution layers can be used to replace the \( 5 \times 5 \) convolution layer for higher efficiency.

Besides the SKIPnet, we also propose DoubleReblur for focus measurement and decision embedding for decision map optimization. Both of their details are briefly mentioned in the Appendix. In addition, the designed loss functions as Eq. (8) are composed of content loss \( L_{\text{con}} \), joint gradient loss \( L_{j} \), 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 Eq. (8) are weighted parameters and set as 1, 0.5, and 0.1. More information on the loss functions can be found in the appendix

\[
L = \alpha L_{\text{con}} + \beta L_{j,\text{grad}} + \gamma L_{\text{grad}}. \tag{8}
\]

In this section, we briefly introduce the DFP principle and framework. In the following, DFP was implemented in both numerical simulations (two-image SR fusion) and practical applications (plenoptic SR all-in-focus imaging).

3 DFP Verified by Two-Image Super-Resolution Fusion

Before using DFP to deal with plenoptic SR all-in-focus imaging, its performance in MFISRF was compared with reported techniques focusing on two-image SR fusion. We qualitatively compared our proposed DFP with the combinations of learning-based MFIF (CNN,21 PCANet,22 FusionDN,23 SESF,24 PMGI,25 U2Fusion26) and conventional SR (Bicubic)/unsupervised SR (DIP27)/supervised SR (SRCNN28). 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 a dataset.

First, the DFP performance in MFISRF was tested. Figure 2 reveals the visual results of MFISRF \( \times 2 \) and \( \times 4 \) on MFI-WHU benchmark evaluation dataset.29 Red arrows marked in Fig. 2 indicate the details of these results. 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, and it is often color infidelity. While SESF slightly improves the MFIF performance compared to PMGI by increasing sharpness and reducing the blur, it still has poor fusion quality. FusionDN and its upgraded U2Fusion both compatible with multitasks 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 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, and U2Fusion) and the supervised SR (SRCNN).

Next, the DFP robustness was tested. Figure 3 lists the results on MFI-WHU 29 with different DoubleReblur parameter sets. Figure 3(a) shows the foreground and background ground truths. Figure 3(b) provides the results obtained by DFP with the DoubleReblur parameter set as [3,5,5,0.05,1] or without weighted joint gradient loss. The result includes (1) the foreground decision map, (2) the MFISRF image, (3) the zoomed-in field of interest, and (4) the pseudocolor image depicting the difference between ground truth and MFISRF result in the luminance component Y channel of YCbCr color space. Identically, the images in Figs. 3(c)–3(e) are obtained by DFP but with different DoubleReblur parameter sets as 0,1,3,5,5, [3,5,5,0.01,1] and [3,5,5,0.05,1], respectively. These results prove that DFP has strong robustness since it can always provide high-quality MFIF results no matter whether the decision maps are good or bad.

Fig. 2 Comparisons of our approach against multiple state-of-the-art supervised and unsupervised (S. and unS) learning-based MFIF (CNN,21 PCANet,22 FusionDN,23 SESF,24 PMGI,25 and U2Fusion26) and (S. and unS.) SR (Bicubic, DIP,27 and SRCNN28) in ×2 condition on MFI-WHU 26 and ×4 condition on MFI-WHU 08. In this figure, S. means supervised, unS. means unsupervised, and GT means ground truth.
Finally, quantitative comparisons are performed using the MFI-WHU benchmark evaluation dataset. The evaluation metrics include mean gradient (MG), edge intensity (EI), information entropy (IE), mean gray value (MGA), and polar edge coherence (ECO). MG reflects the rate of contrast change of tiny details in the image. EI describes the relative EI between the ground truth (GT) and MFISRF. One of the most basic features of an image is its edge existing between target and background, so it is an important feature to evaluate image fusion. IE reflects the comprehensive characteristics of gray value at a pixel and its surrounding pixel gray distributions. MGA is the average level of image gray, and it represents the overall brightness level of the image. ECO reflects the edge similarity of the reconstructed and original images. The ideal values of these metrics are all zero.

Additionally, Tables 1–3 quantitatively compare the performances using different MFIF and SR combined methods and the proposed one according to the above-mentioned coefficients. Our proposed DFP can achieve high-quality super-resolved multifocus fused images. Even compared to the optimized combinations of unsupervised MFIF methods and supervised SR method, our proposed DFP could still obtain low MG, EI, IE, MGA, and ECO values very close to or even lower than those obtained by the optimized MFIF and SR combined methods.

According to the above DFP verification on two-image SR fusion, the proposed unsupervised dataset-free DFP approaches or even outperforms these state-of-the-art MFIF and SR method combinations. Additionally, DFP has strong robustness since it can always provide high-quality MFIF results less influenced by the decision maps. Therefore, these verifications prove that the proposed DFP works well in two-image SR fusion.
Table 1  Comparisons of our approach against multiple state-of-the-art learning-based MFIF (CNN,21 PCANet,22 FusionDN,23 SESF,24 PMGI,25 U2Fusion26) and Bicubic SR in ×2 and ×4 conditions. Italic font marks the first, and black bold font marks the second and third best of the performance.

| Method           | MG_\((R)\) | EI_\((R)\) | IE_\((R)\) | MGA_\((R)\) | ECO_\((R)\) |
|------------------|------------|------------|------------|-------------|--------------|
|                  | ×2         | ×2         | ×2         | ×2          | ×2           |
| CNN+             | 0.7207     | 3.6682     | 0.0176     | 0.1692      | 0.0844       |
| BICUBIC          | 2.5906     | 22.7946    | 0.0598     | 0.1723      | 0.0488       |
| PCANet+          | 0.7444     | 3.8778     | 0.0174     | 0.2240      | 0.0196       |
| BICUBIC          | 2.6040     | 22.9440    | 0.0612     | 0.2252      | 0.0283       |
| FusionDN+        | 0.9637     | 12.3182    | 0.3491     | 16.4175     | 0.0014       |
| BICUBIC          | 1.7848     | 14.4899    | 0.3231     | 16.3919     | 0.0281       |
| SESF+            | 0.7138     | 3.5747     | 0.0159     | 0.1735      | 0.0214       |
| BICUBIC          | 2.5856     | 22.7406    | 0.3899     | 18.1909     | 0.1422       |
| PMGI+            | 2.1311     | 18.4735    | 0.3921     | 18.1691     | 0.1434       |
| BICUBIC          | 2.9701     | 27.0107    | 0.2975     | 5.3168      | 0.0955       |
| U2Fusion+        | 0.9618     | 15.1490    | 0.2461     | 5.3213      | 0.1481       |
| BICUBIC          | 1.3636     | 9.3573     | 0.0222     | 0.5324      | 0.0056       |
| DFP (ours)       | 0.5840     | 6.3614     | 0.0376     | 0.5536      | 0.0062       |
|                  | 1.8111     | 14.8770    |            |              |              |

Table 2  Comparisons of our approach against multiple state-of-the-art learning-based MFIF (CNN,21 PCANet,22 FusionDN,23 SESF,24 PMGI,25 U2Fusion26) and unsupervised learning-based SR (DIP27) in ×2 and ×4 conditions. Italic font marks the first, and black bold font marks the second and third best of the performance.

| Method          | MG_\((R)\) | EI_\((R)\) | IE_\((R)\) | MGA_\((R)\) | ECO_\((R)\) |
|-----------------|------------|------------|------------|-------------|--------------|
|                 | x4         | x4         | x4         | x4          | x4           |
| CNN+            | 0.5780     | 2.6387     | 0.0310     | 0.3439      | 0.0863       |
| DIP             | 2.0613     | 17.1756    | 0.0509     | 0.3378      | 0.0652       |
| PCANet+         | 0.6014     | 2.7950     | 0.0288     | 0.2661      | 0.0192       |
| DIP             | 2.0911     | 17.5025    | 0.0559     | 0.2713      | 0.0233       |
| FusionDN+       | 1.1171     | 14.3756    | 0.3529     | 16.7732     | 0.0158       |
| DIP             | 1.2079     | 9.5311     | 0.3429     | 16.7444     | 0.0317       |
| SESF+           | 0.5593     | 2.6525     | 0.0275     | 0.3503      | 0.0244       |
| DIP             | 2.0534     | 17.0854    | 0.0474     | 0.3486      | 0.0245       |
| PMGI+           | 2.2334     | 19.5217    | 0.4057     | 18.5801     | 0.1329       |
| DIP             | 2.6718     | 24.0968    | 0.4118     | 18.5721     | 0.1468       |
| U2Fusion+       | 1.1756     | 17.2388    | 0.2929     | 5.2692      | 0.0940       |
| DIP             | 0.5697     | 3.9655     | 0.2739     | 5.3012      | 0.1389       |
| DFP (ours)      | 0.5840     | 6.3614     | 0.0222     | 0.5324      | 0.0056       |
|                 | 1.8111     | 14.8770    | 0.0376     | 0.5536      | 0.0062       |
4 DFP Applied for Plenoptic Super-Resolution All-in-Focus Imaging

After verification using two-image SR fusion, our proposed DFP was applied for plenoptic SR all-in-focus imaging. Figure 4 reveals an example from the Lytro first generation dataset. Figure 4(a) is the captured plenoptic image, and Figs. 4(b)–4(e) are the reconstructed images.

| Method            | MG_{(R)}^{x2} | EI_{(R)}^{x2} | IE_{(R)}^{x2} | MGA_{(R)}^{x2} | ECO_{(R)}^{x2} |
|-------------------|---------------|---------------|---------------|----------------|----------------|
| CNN               | 0.7995        | 9.9926        | 0.0237        | 0.9516         | 0.0827         |
| SRCNN             | 1.7145        | 13.6249       | 0.0317        | 0.7879         | 0.0705         |
| PCANet+           | 0.7987        | 9.7746        | 0.0235        | 0.8760         | 0.0133         |
| SRCNN             | 1.7302        | 13.7981       | 0.0341        | 0.7250         | 0.0082         |
| FusionDN+         | 1.8200        | 22.2081       | 0.3592        | 17.1748        | 0.0138         |
| SRCNN             | 0.9588        | 8.5307        | 0.3476        | 17.0702        | 0.0312         |
| PCANet+           | 0.8097        | 10.1118       | 0.0249        | 0.9539         | 0.0153         |
| FusionDN+         | 1.7054        | 13.5246       | 0.0316        | 0.7985         | 0.0095         |
| SESF+             | 1.7301        | 15.1158       | 0.3925        | 19.0188        | 0.1293         |
| SRCNN             | 2.4111        | 21.4495       | 0.3913        | 18.9211        | 0.1447         |
| U2Fusion+         | 2.1945        | 26.6641       | 0.3184        | 5.3172         | 0.0778         |
| SRCNN             | 0.4994        | 6.2283        | 0.2815        | 5.3075         | 0.1156         |
| DFP (ours)        | 0.5840        | 6.3614        | 0.0222        | 0.5324         | 0.0056         |

Fig. 4 Plenoptic SR all-in-focus imaging example from the Lytro first generation dataset. (a) Captured plenoptic image; (b)–(e) Multifocus images, and the zoomed-in fields of interest reveal the in-focus parts; SR all-in-focus images reconstructed using (f) Bicubic and (g) DFP, and the zoomed-in fields of interest reveal the details of the SR all-in-focus images.
focused at different planes via our previously proposed OpenRefocus. Only parts of the image are in focus, as shown in the zoomed-in fields of interest in Figs. 4(b)–4(e); the rest is out of focus. It is because the captured plenoptic image is refocused on different planes. Next, a plenoptic SR all-in-focus image as shown in Fig. 4(f) was reconstructed from these multifocus images via bicubic. Additionally, another one revealed in Fig. 4(g) was also obtained via our proposed DFP. In comparison to these plenoptic SR all-in-focus images in Figs. 4(f) and 4(g), particularly those zoomed-in fields of interest, the proposed DFP not only has higher spatial resolution, but also provides a higher quality all-in-focus image, resulting in a better performance in the MFISRF task.

Moreover, another practical experiment of capturing the complex natural light-field image using Lytro Illum was also provided in Fig. 5. Figure 5(a) is our captured image of leaves at different depths of field. Figures 5(b)–5(e) are the multifocus images reconstructed using OpenRefocus. Through refocus, only parts of the image are in focus, as shown in the zoomed-in fields of interest in Figs. 5(b)–5(e); the rest is out of focus. Figures 5(f) and 5(g) are the plenoptic SR all-in-focus images reconstructed using Bicubic and DFP, respectively. These results demonstrate that both methods can generate all-in-focus images. But DFP, like the above example from the Lytro dataset, can obtain a higher spatial resolution and higher quality all-in-focus image in practice. Both these experiments verify that DFP can be successfully employed for plenoptic SR all-in-focus imaging.

5 Conclusion

We unify the MFIF and blind SR problems as the MFISRF task and propose a unified dataset-free unsupervised framework DFP for plenoptic SR all-in-focus imaging. To the best of our knowledge, our proposed work is the first dataset-free unsupervised method to jointly implement the multifocus fusion and SR task for the first time. The DFP consists of our designed SKIPnet end-to-end generated network, 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 with six MFIF and three SR method combinations, including both supervised and unsupervised ones, the proposed unsupervised dataset-free DFP approaches or even outperforms these state-of-the-art MFIF and SR method combinations. In addition, DFP has been successfully employed for plenoptic SR all-in-focus imaging. Furthermore, DFP is a general framework, meaning that 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.
Appendix

6.1 DFP Training Details

We implemented the proposed DFP based on the PyTorch framework using an NVIDIA GTX 3060 GPU for training. Adam optimizer was adopted, and the learning rate was set to 0.0002. The training procedure of DFP is shown as Algorithm 1. For more multimodel training details, we concatenate model parameters as a new parameter list for the optimizer. It is noted that DFP performs without any external dataset, and the training dataset of DFP is only a pair of input images. We use the fixed inputs in every iteration. Therefore, the testing phase is not used in our DFP. Begin by inputting two low-resolution images, the initial decision map can be obtained by executing DoubleReblur. Then, to optimize the decision map, the low-resolution input and initial decision map are inputted into Decision Embedding for optimization (Sec. 6.3). Following that, SKIPnet’s forwarding, losses computing, and backwarding are operational for iterations \( K = 1000 \sim 5000 \). In the end, the output of SKIPnet in the last iteration is the result of SR fusion.

Algorithm 1  Deep fusion prior.

1: Input: foreground input \( i_f \), background input \( i_b \).
Output: \( \hat{I}_{MFISRF} \).
2: \( \text{map} = \text{DoubleReblur}(i_f, i_b) \)
3: \( \text{map} = \text{Decision Embedding}((i_f + i_b)/2, i_f, \text{map}) \)
4: for iterations \( K = 1000 \sim 5000 \) do
5: Forward: \( \hat{I}_{MFISRF} = \text{SKIPnet}(\text{Bilinear}(i_f + i_b)/2, \text{map}); \)
6: Select LANCZOS2 to downsample \( \hat{I}_{MFISRF} \);
7: Computing losses;
8: Backward to update SKIPnet parameters \( \Theta \);
9: end for

6.2 DoubleReblur

Most MFIF methods 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. 6.
Its main idea is that there exists a significant difference between the in-focus scene \((o_{in})\) and its blurred \((o_{in} * h, h \text{ is the convolution kernel})\), but the difference between defocus scene \((o_{d})\) and its blurred \((o_{d} * h)\) is not remarkable. Therefore, we try to enhance this difference to facilitate the in-focus fields decision.

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

\[
i_b(x, y) = i_f(x, y) * h_s(x, y) + n_s(x, y).
\]

To estimate the spread kernel, Fourier transform is implemented on both sides of

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

\[
I_b(\xi, \eta) = I_f(\xi, \eta) \cdot H_s(\xi, \eta) + N_s(\xi, \eta).
\]

In Eqs. (10) and (11), \(F\{\cdot\}\) is the Fourier transform, \(I_b(\xi, \eta), I_f(\xi, \eta), H_s(\xi, \eta),\) and \(N_s(\xi, \eta)\) are the frequency spectra of \(i_b(x, y), i_f(x, y), h_s(x, y),\) and \(n_s(x, y)\), respectively

\[
\frac{I_b(\xi, \eta)}{I_f(\xi, \eta)} = H_s(\xi, \eta) + \frac{N_s(\xi, \eta)}{I_f(\xi, \eta)}.
\]

To remove the noise term, a low-pass filter \(T\) is performed on Eq. (12). In the end, the estimate of spread kernel \(\hat{h}_s(x, y)\) can be obtained as

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

The first reblur image can be obtained as

\[
\hat{i}_b(x, y) = i_b(x, y) * \hat{h}_s(x, y).
\]

Then, the Gaussian reblur \(\Gamma\) as the second reblur is used to enhance the sharpness difference as \(s(x, y)\) obtained as

\[
s(x, y) = |\hat{i}_b(x, y) - G\{\hat{i}_b(x, y)\}|.
\]

In Eq. (16), \(d(x, y)\) 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 \(\cdot\) as

\[
\hat{d}(x, y) = d(x, y) \cdot E.
\]

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

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

There are five parameters in DoubleReblur as \([k_g, k_d, k_e, t, f]\): \(k_g\) is the kernel size of Gaussian blur, \(k_d\) is the kernel size of dilating, \(k_e\) is the kernel size of eroding, \(t\) is the threshold of segmentation, and \(f\) is the bool flag control if the largest region filling is used. We transform the image to the luminance component \(Y\) channel of YCbCr color space and use \([5, 3, 3, 0.01, 1]\) for most images and slightly adjust parameters for different image details.
6.3 Decision Embedding

Since the decision maps are obtained from handcrafted focus measurement, they suffer from disadvantages such as edge fragmentation and false determination. To optimize the handcrafted decision maps, learning-based decision embedding is designed as Fig. 7. The input is obtained by the average of low-resolution inputs but can also be replaced by random noise. The binary segmentation is equal to the process that is described in Eq. (16). In detail, the last layer of SKIPnet is Sigmoid activation function, which maps the output value from 0 to 1, and we set \( t = 0.5 \) to implement binary segmentation. Then, the binarized decision map can be generated by SKIPnet. Furthermore, as Eq. (19), we design an optimized loss function \( L_{\text{opt}} \) to optimize the decision map by minimizing the difference between the focus measurement regions and the original low-resolution inputs

\[
L_{\text{opt}} = \frac{1}{H \cdot W} \sum_i \sum_j \left| \hat{m}_{i,j} - m_{i,j} \right| + |\hat{m}_{i,j} \cdot I_{\text{fore}} - I_{\text{fore}}| + |(1 - \hat{m}_{i,j}) \cdot I_{\text{back}} - I_{\text{back}}|, \tag{19}
\]

where \( H \) and \( W \) are the height and width of the image, \( I_{\text{fore}} \) and \( I_{\text{back}} \) 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.

6.4 Loss Functions

\( L_{\text{con}} \) in Eq. (20) is the content loss, where \( \lambda_1 \) 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. L1 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. Note that the human visual system (HVS) is especially sensitive to brightness and color changes in untextured areas of the image.
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 $\hat{I}_{MIFS}$ and the joint maximal Laplacian gradient map of low-resolution inputs $I_{\text{fore}}$ and $I_{\text{back}}$ is used to compute their distance.

\[ L_{\text{con}} = \frac{1}{H \cdot W} \sum_{i} \sum_{j} (\Delta L_{\hat{I}_{MIFS},i,j} - L_{\text{fore},i,j}) + \lambda_{1} (\Delta L_{\hat{I}_{MIFS},i,j} - L_{\text{fore},i,j}) + \lambda_{2} (\Delta L_{\hat{I}_{MIFS},i,j} - L_{\text{back},i,j}) \]

$\Delta L_{\hat{I}_{MIFS}}$ in Eq. (21) 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 $\hat{I}_{MIFS}$ and the joint maximal Laplacian gradient map of low-resolution inputs $I_{\text{fore}}$ and $I_{\text{back}}$ is used to compute their distance.

\[ L_{\text{grad}} = \frac{1}{H \cdot W} \sum_{i} \sum_{j} |\Delta_{L} \hat{I}_{MIFS},i,j - \max(\Delta_{L} I_{\text{fore},i,j}, \Delta_{L} I_{\text{back},i,j})| \]  

$L_{\text{grad}}$ in Eq. (22) 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.

\[ L_{\text{grad}} = \frac{1}{H \cdot W} \sum_{i} \sum_{j} (\nabla_{x} \hat{I}_{MIFS},i,j + \nabla_{y} \hat{I}_{MIFS},i,j) \]  

### 6.5 Ablation Experiments

Finally, the ablation experiments on the loss functions were implemented in Fig. 8 and Table 4. As shown in these results, $L_{\text{con}}$ is the indispensable basic content loss, $L_{\text{grad}}$ increases the convergence of high-frequency details, and $L_{\text{grad}}$ is the noise constraint item. $L_{\text{con}} + L_{\text{grad}}$ achieves the best performance and we also use $L_{\text{grad}}$ to reduce the noise. The values of parameters are 1, 0.5, and 0.1 for default in this work.
Acknowledgments

The work was supported by National Natural Science Foundation of China (61705092) and Natural Science Foundation of Jiangsu Province of China (BK20170194). The authors declare no conflicts of interest.

References

1. M. Yu et al., “REPAID: resolution-enhanced plenoptic all-in-focus imaging using deep neural networks,” Opt. Lett. 46, 2896–2899 (2021).
2. Q. Miao et al., “Resolution improvement in optical projection tomography by the focal scanning method,” Opt. Lett. 35, 3363–3365 (2010).

Table 4  Ablation experimental results.

| Loss                      | MG(R)   | EI(R)   | IE(R)  | MGA(R) |
|---------------------------|---------|---------|--------|--------|
| $L_{\text{con}}$          | 1.3586  | 7.5134  | 0.0633 | 0.8888 |
| $L_{\text{con}} + L_{\text{grad}}$ | 1.2118  | 5.9706  | 0.0607 | 0.4706 |
| $L_{\text{con}} + L_{\text{ij, grad}}$ | 0.3249  | 6.831   | 0.0019 | 0.3956 |
| $L_{\text{con}} + L_{\text{ij, grad}} + L_{\text{grad}}$ | 1.0249  | 5.3648  | 0.0542 | 0.6847 |

Fig. 8 Results of the ablation experiments.

Table 4  Ablation experimental results.
3. H. Li et al., “Focal stack camera in all-in-focus imaging via an electrically tunable liquid crystal lens doped with multi-walled carbon nanotubes,” Opt. Express 26, 12441–12454 (2018).
4. S. Kuthirummal et al., “Flexible depth of field photography,” IEEE Trans. Pattern Anal. Mach. Intell. 33(1), 58–71 (2010).
5. O. Cossairt, C. Zhou, and S. Nayar, “Diffusion coded photography for extended depth of field,” ACM Trans. Graphics 29(4), 31 (2010).
6. M. R. Rai and J. Rosen, “Depth-of-field engineering in coded aperture imaging,” Opt. Express 29(2), 1634–1648 (2021).
7. O. Cossairt and S. Nayar, “Spectral focal sweep: extended depth of field from chromatic aberrations,” in IEEE Int. Conf. on Comput. Photogr. (ICCP), pp. 1–8 (2010).
8. Y. Bando, B. Chen, and T. Nishita, “Extracting depth and matte using a color-filtered aperture,” ACM Trans. Graphics 27(5), 134 (2018).
9. A. Levin et al., “4D frequency analysis of computational cameras for depth of field extension,” ACM Trans. Graphics 28(3), 97 (2009).
10. J. Kim, J. K. Lee, and K. M. Lee, “Deeply-recursive convolutional network for image super-resolution,” in Proc. IEEE Comput. Soc. Conf. Comput. Vision Pattern Recognit., pp. 1637–1645 (2016).
11. W. S. Lai et al., “Deep Laplacian pyramid networks for fast and accurate super-resolution,” in Proc. - 30th IEEE Conf. Comput. Vision Pattern Recognit., pp. 5835–5843 (2017).
12. C. Ledig et al., “Photo-realistic single image super-resolution using a generative adversarial network,” in Proc. - 30th IEEE Conf. Comput. Vision Pattern Recognit., pp. 105–114 (2017).
13. B. Lim et al., “Enhanced deep residual networks for single image super-resolution,” in IEEE Comput. Soc. Conf. Comput. Vision Pattern Recognit. Work., pp. 1132–1140 (2017).
14. Y. Zhang et al., “Residual dense network for image super-resolution,” in Proc. IEEE Comput. Soc. Conf. Comput. Vision Pattern Recognit., pp. 2472–2481 (2018).
15. W. Han et al., “Image super-resolution via dual-state recurrent networks,” in Proc. IEEE Comput. Soc. Conf. Comput. Vision Pattern Recognit., pp. 1654–1663 (2018).
16. J. Tian et al., “Multi-focus image fusion using a bilateral gradient-based sharpness criterion,” Opt. Commun. 284(1), 80–87 (2011).
17. B. K. Shreyamsha Kumar, “Image fusion based on pixel significance using cross bilateral filter,” Signal Image Video Process. 9(5), 1193–1204 (2015).
18. Y. Liu, S. Liu, and Z. Wang, “Multi-focus image fusion with dense SIFT,” Inf. Fusion 23, 139–155 (2015).
19. X. Qiu et al., “Guided filter-based multi-focus image fusion through focus region detection,” Signal Process. Image Commun. 72, 35–46 (2019).
20. S. Li et al., “Image matting for fusion of multi-focus images in dynamic scenes,” Inf. Fusion 14(2), 147–162 (2013).
21. Y. Liu et al., “Multi-focus image fusion with a deep convolutional neural network,” Inf. Fusion 36, 191–207 (2017).
22. X. Song and X. J. Wu, “Multi-focus image fusion with PCA filters of PCANet,” Lect. Notes Comput. Sci. 11377, 1–17 (2019).
23. H. Xu et al., “FusionDN: a unified densely connected network for image fusion,” Proc. AAAI Conf. Artif. Intell. 34(7), 12484–12491 (2020).
24. B. Ma et al., “SESF-fuse: an unsupervised deep model for multi-focus image fusion,” Neural Comput. Appl. 33(11), 5793–5804 (2021).
25. H. Zhang et al., “Rethinking the image fusion: a fast unified image fusion network based on proportional maintenance of gradient and intensity,” AAAI 34(7), 12797–12804 (2020).
26. H. Xu et al., “U2Fusion: a unified unsupervised image fusion network,” IEEE Trans. Pattern Anal. Mach. Intell. 44(1), 502–518 (2022).
27. V. Lempitsky, A. Vedaldi, and D. Ulyanov, “Deep image prior,” in Proc. IEEE Comput. Soc. Conf. Comput. Vision Pattern Recognit., pp. 9446–9454 (2018).
28. C. Dong et al., “Image super-resolution using deep convolutional networks,” IEEE Trans. Pattern Anal. Mach. Intell. 38(2), 295–307 (2016).
29. H. Zhang et al., “MFF-GAN: an unsupervised generative adversarial network with adaptive and gradient joint constraints for multi-focus image fusion,” Inf. Fusion 66, 40–53 (2021).
30. D. Han et al., “Multi-exposure image fusion via deep perceptual enhancement,” *Inf. Fusion* **79**, 248–262 (2022).
31. V. Baroncini et al., “The polar edge coherence: a quasi blind metric for video quality assessment,” in *Proc. of 14th Eur. Signal Process. Conf.*, September, pp. 564–568 (2009).
32. M. Nejati, S. Samavi, and S. Shirani, “Multi-focus image fusion using dictionary-based sparse representation,” *Inf. Fusion* **25**, 72–84 (2015).
33. Y. Gu et al., “OpenRefocus: an open-source Qt-based tool for light field parallel refocusing,” *Opt. Eng.* **61**(6), 063108 (2022).

Biographies of the authors are not available.