Focus Stacking with High Fidelity and Superior Visual Effects

Bo Liu, Bin Hu, Xiuli Bi*, Weisheng Li, Bin Xiao*

Department of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065, China
{boliu, bixl, liws, xiaobin}@cqupt.edu.cn, s210201036@stu.cqupt.edu.cn

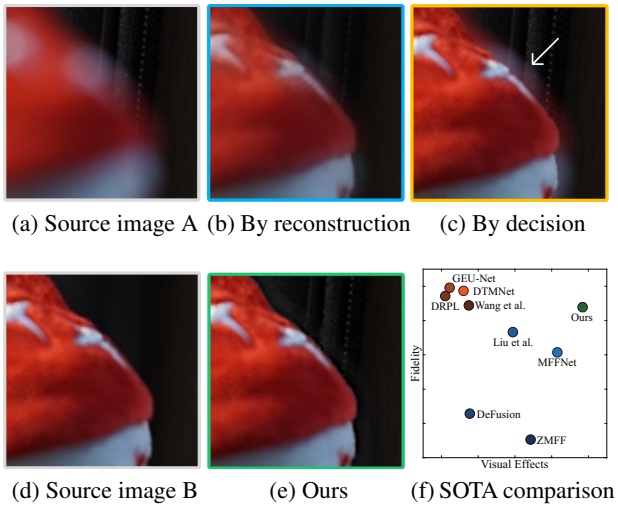
Abstract

Focus stacking is a technique in computational photography, and it synthesizes a single all-in-focus image from different focal plane images. It is difficult for previous works to produce a high-quality all-in-focus image that meets two goals: high-fidelity to its source images and good visual effects without defects or abnormalities. This paper proposes a novel method based on optical imaging process analysis and modeling. Based on a foreground segmentation - diffusion elimination architecture, the foreground segmentation makes most of the areas in full-focus images heritage information from the source images to achieve high fidelity; diffusion elimination models the physical imaging process and is specially used to solve the transition region (TR) problem that is a long-term neglected issue and degrades visual effects of synthesized images. Based on extensive experiments on simulated dataset, existing realistic dataset and our proposed BetaFusion dataset, the results show that our proposed method can generate high-quality all-in-focus images by achieving two goals simultaneously, especially can successfully solve the TR problem and eliminate the visual effect degradation of synthesized images caused by the TR problem.

Introduction

Subject to optical limitations, cameras cannot simultaneously capture objects at different focal planes in focus. As a result, the foreground and background objects in one image cannot be clear at the same time. This problem arises frequently in microscopic imaging and general photography. With the help of focus stacking technique in computational photography, an all-in-focus image can be synthesized using several images in different focusing planes of a same scene.

Users expect synthetic all-in-focus images will not change the information of source images and will be similar to them as much as possible, namely achieving high fidelity. Besides, the all-in-focus images should maintain superior visual effects without showing visual defects or abnormalities, such as color bias (Fig.1b) and the TR problem (Fig.1c). Many focus stacking algorithms have been proposed to achieve the above two goals, and can be divided into two approaches: the reconstruction-based approaches aim at improving visual effects, and the decision-based approaches pursue high fidelity.



(a) Source image A (b) By reconstruction (c) By decision

(d) Source image B (e) Ours (f) SOTA comparison

Figure 1: Focus stacking from source images (a) and (d) by a reconstruction-based method in (b), which shows color bias; by a decision-based method in (c), which produces visual defects indicated by white arrows; by ours in (e). Comparisons of methods in high fidelity (Q^{MI}) and visual effects (Q^{CV}) are shown in (f).

The reconstruction-based approaches (Li et al. 2019) (Deng and Dragotti 2021) (Liang et al. 2022) convert source images to a transform domain and synthesize an all-in-focus image by inversely transforming the synthesized features. The advantage of this strategy is that the processing of source images is continuous, thus generated images usually have better visual effects. The disadvantage is that most of the undisputed focused area will be affected by the corresponding defocused area, producing a certain degree of color bias as shown in Fig.1b. In some scenarios such as microscopic imaging, color bias is not acceptable.

The decision-based approaches (Wang et al. 2022) (Liu et al. 2022) (Hu et al. 2023) generate a binary decision map using preset criteria to indicate which pixels, blocks, or regions of source images will be included in a synthetic image.

*Corresponding author

Since all pixels in a full-in-focus image are directly derived from its source images, decision-based approaches can avoid color bias problems. However, as we will analyze in later section, regardless of the accuracy of focus and defocusing regions, decision-based approaches cannot achieve satisfactory visual effects at the TR where the foreground and background adjoin, yielding visual defects such as the halo indicated by the white arrow in Fig.1c. Although decision-based approaches avoid color bias, these visual defects are left in the photo. Since the above two categories of methods pursue one goal of focus stacking, the synthetic all-in-focus images are unsatisfactory when considering both high fidelity and visual effects, as shown in Fig.1f.

If we carefully analyze the decision-based approaches, the critical problem is overemphasizing binary selection. The diffusion component of the foreground in the far-focus image will spread and superimpose at the TR, resulting in the unsatisfactory visual effect of the region no matter which pixels from source images are selected. Therefore, for decision-based approaches, the decision in TR is a dilemma: if the pixels in the TR are selected from a near-focus image, the diffusion component is retained; if selected from a far-focus image, the background details will be lost.

Existing decision-based approaches ignore this issue because the arbitrary pixel selection in TR does not necessarily lower the fidelity evaluation metrics. However, it will seriously affect the visual effects of synthetic all-in-focus images. By contrast, balancing high fidelity and visual effects near TR is more advisable. The feasible and relatively simple route is to make decision first, inheriting maximum information from source images to maintain high fidelity. Then, we perform diffusion elimination in TR to synthesize high-quality all-in-focus images: a two-stage foreground segmentation - diffusion elimination architecture.

The proposed two-stage architecture is based on the analysis and modeling the optical imaging process of cameras. The diffusion component in TR is modeled as the smoothed color difference between the foreground and background near their boundary. To estimate the diffusion component, a neural network named *Foreground Segment Branch* has been constructed to segment the foreground and obtain its boundary. Simultaneously, the *Kernel Estimation Branch* can restore the point spread function (PSF). Consequently, the diffusion component can be accurately inferred. By removing the diffusion component from initial synthetic images, we can obtain synthetic images with both high fidelity and superior visual effect.

The contribution of our work can be summarized as:

1. We first analyze and model the optical imaging process of focus stacking images, paving ways to solve the long-standing and neglected TR problem.
2. Based on the model, we can further eliminate the diffusion component on TR and improves the final synthesis quality, achieving high fidelity and superior visual effects.
3. We propose a *real-world* stacking dataset "BetaFusion", in which the image pairs are *well registered*, providing a measure for solving the TR problem.

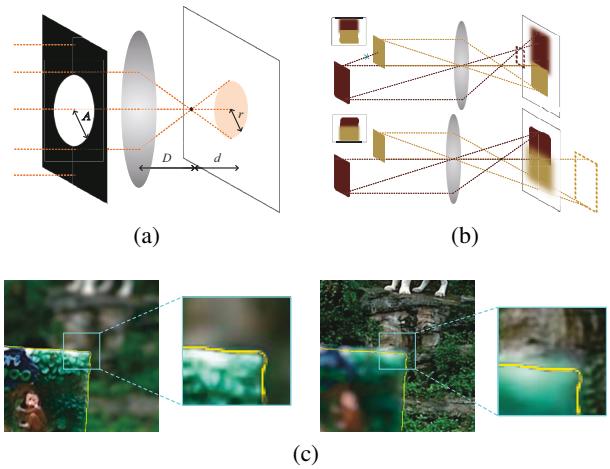
Related Work

Reconstruction-based Methods

Good reconstruction result needs to be supported by reliable features. Therefore, many reconstruction-based methods focus on feature extraction and analysis. Li *et al.* (Li, Yuan, and Fan 2019) utilize wavelet transform to extract multi-scale features, then build high-frequency and low-frequency networks to encode and reconstruct different frequency information of synthetic images. The structure of the network itself can also extract information at different scales. For example, Zhao *et al.* (Zhao, Wang, and Lu 2019) design a multi-level feature extraction structure, which can extract low-frequency content and high-frequency details separately, and supervises the multi-level features during training to improve the synthesis effect. Mustafa *et al.* (Mustafa, Yang, and Zareapoor 2019) extract features in parallel by convolution kernels of different sizes and combine this basic structure with a Siamese network to strengthen the network's ability to extract features. Li *et al.* (Li et al. 2019) take full advantage of U-Net's success in multi-scale feature extraction capabilities to capture and reconstruct information at different frequencies. Starting from the characteristics of the source images, Deng *et al.* (Deng and Dragotti 2021) believe that the source image pairs are composed of the union components and their unique components, so they design the CU-Net to decompose the source image pairs into different components, and then reconstruct all the unique features and common features to obtain synthetic images. DenseNet has been used by (Mustafa, Zareapoor, and Yang 2020) to ensure that the network can extract sufficient features, and the dense connection effectively reduces the loss of features and ensures the quality of reconstruction. Generative adversarial networks (GAN) have also been introduced to focus stacking. As a representative, MFF-GAN proposed by Zhang *et al.* (Zhang et al. 2021) reaches good synthetic results by supervising the generator with real all-focus features. Liang *et al.* (Liang et al. 2022) designed a model named DeFusion using a similar idea to CU-Net. However, DeFusion can be trained without any paired data and complicated losses. Based on the deep image prior, Hu *et al.* (Hu et al. 2023) introduce zero-shot learning to focus stacking task, and their ZMFF successfully avoids the domain shifting problem caused by handmade training data.

Decision-based Methods

Decision-based methods focus on how to generate a more precise decision map. To measure more accurately, the network constructed by Tang *et al.* (Tang et al. 2018) uses neighborhood information to perform pixel-level focus evaluation. This learnable focus measurement method is more efficient than traditional methods and has higher accuracy. Based on U-Net (Ronneberger, Fischer, and Brox 2015), Xiao *et al.* (Xiao et al. 2021) propose global feature encoding U-Net (GEU-Net) and establish a global feature pyramid extraction module and a global attention connection upsampling module to decide from a global perspective. SESF proposed by Ma *et al.* (Ma et al. 2021) uses an encoder-decoder structure to learn high-level features through extensive train-



(a) (b)

(c)

Figure 2: Imaging process of near- and far-focused images. (a) Ideal PSF of imaging device. (b)Near- and far-focused images. (c) Realistic near- and far-focused image pair.

ing. In the testing phase, encoded features are used to make judgments according to spatial frequency. DRPL proposed by Li *et al.* (Li et al. 2020) outputs a mask for each source image, and the two masks should be complementary for a perfect decision. The complementary constraint greatly improves synthesis performance. Xiao *et al.* (Xiao, Wu, and Bi 2021) utilize discrete Tchebichef moments to extract focus features, and their DTMNet applies a lightweight network to generate high-quality synthetic results. To deal with data limitation, Wang *et al.* (Wang et al. 2022) train a residual feature extractor on a super-resolution dataset and then apply the extractor on focus stacking task to generate a decision map. Liu *et al.* (Liu et al. 2022) transform source images and initial synthetic image into feature space, and obtain decision maps by comparing the similarity of the features.

Methodology

This section first analyzes the optical basis for the formation of the TR problem. Then, our focus staking imaging model will be deduced by introducing the form of PSF. We estimate the key parameters in the imaging model using elaborate neural networks. Finally, by eliminating diffusion components, the satisfactory synthetic results can be obtained.

Defocus and PSF

The core of the TR problem is the diffusion of off-focused objects by an optical lens. The camera lens is composed of a series of optical lenses. In order to simplify the analysis, we consider a thin spherical lens model, meanwhile ignoring the influence caused by aberrations. When the target is far enough away from the imaging device, light reflected off the object's surface enters the thin lens approximately in parallel and forms an image on the sensor. For each abstract point $\boldsymbol{p}$ on the surface of the target object, the reflected light entering the lens at location $(\boldsymbol{u}, \boldsymbol{v})$, and $(\boldsymbol{u}, \boldsymbol{v})$ satisfies $\boldsymbol{u}^2+\boldsymbol{v}^2 \leq \boldsymbol{A}^2$ because of the aperture whose radius is $\boldsymbol{A}$. When $\boldsymbol{p}$ is defocused, the reflected light of $\boldsymbol{p}$ will intersect before the sensor, letting the distance between the intersection and sensor be $\boldsymbol{d}$ and $\boldsymbol{D}$ for the distance between the intersection and the lens, as Fig.2a shows. The image of $\boldsymbol{p}$ on sensor forms a set of coordinates $\boldsymbol{P} = \{(\boldsymbol{x}, \boldsymbol{y})\}$:

$$\boldsymbol{P} = \{(\frac{\boldsymbol{d}}{\boldsymbol{D}}\boldsymbol{u}, \frac{\boldsymbol{d}}{\boldsymbol{D}}\boldsymbol{v}) | \boldsymbol{u}^2 + \boldsymbol{v}^2 \leq \boldsymbol{A}^2\}. \tag{1}$$

According to (Lai, Fu, and Chang 1992) and (Zhuo and Sim 2011), when $\boldsymbol{p}$ is a unit-impulse function, the image of $\boldsymbol{p}$ is the lens's point spread function, and the PSF $\boldsymbol{K}$ should obey the 2-D normal distribution like

$$\boldsymbol{K} = \mathbb{N}(0, 0, \sigma_x, \sigma_y). \tag{2}$$

To make the situation simpler, we assume that the point spread function is isotropic, i.e., $\sigma_x = \sigma_y = \sigma$.

Focus Stacking: a Non-binary Problem

Most decision-based approaches consider focus stacking as a binary selection problem, i.e., the focusing properties of pixels at the same position in a source image pair are exactly the opposite: the focused area in one image must be defocused in another image, and vice versa. This assumption is appropriate in most cases. However, as Fig.2b and Fig.2c shows, there is an easily overlooked non-binary phenomenon near TR, which we call foreground diffusion. In this part, we will build a mathematical model to illustrate the formation of this diffusion phenomenon.

When focusing on the foreground $\boldsymbol{f}$, the clear foreground will not be affected by the blurred background $\boldsymbol{b}$, and the near-focus image $\boldsymbol{N}$ can be described as

$$\boldsymbol{N} = (\boldsymbol{f} \cdot \boldsymbol{M}) * \boldsymbol{K}_1 + \bar{\boldsymbol{M}} \cdot (\boldsymbol{b} * \boldsymbol{K}_2). \tag{3}$$

Here $\boldsymbol{M}$ is an accurate mask to indicate the foreground area, and the complementary $\bar{\boldsymbol{M}}$ indicates the background area, $\boldsymbol{K}_1$ and $\boldsymbol{K}_2$ are the PSF of clear foreground and blurred background, respectively. When focused on the background, clear background $\boldsymbol{b}$ will be covered by the blurred foreground near TR, and forms a semi-transmitted state. Taking the influence of blurred foreground into account, the far-focused image $\boldsymbol{F}$ can be represented as

$$\boldsymbol{F} = (\boldsymbol{f} \cdot \boldsymbol{M}) * \boldsymbol{K}_3 + (\bar{\boldsymbol{M}} * \boldsymbol{K}_3) \cdot (\boldsymbol{b} * \boldsymbol{K}_4). \tag{4}$$

Corresponding to near-focused image, $\boldsymbol{K}_3$ stands for the PSF of blurred foreground and $\boldsymbol{K}_4$ is the PSF of clear background. It is worth noting that in both Eq.3 and 4, the visibility of foreground and background is represented by $\boldsymbol{M}$ and $\bar{\boldsymbol{M}}$, that means, $\boldsymbol{b}$ is a complete image as if $\boldsymbol{f}$ does not exist. Since the blur is not obvious when a region is focused, $\boldsymbol{K}_1$ and $\boldsymbol{K}_4$ can be regarded as unit impulses.

For a near-focus and a far-focus image, the diffusion component S covered on background can be modeled as

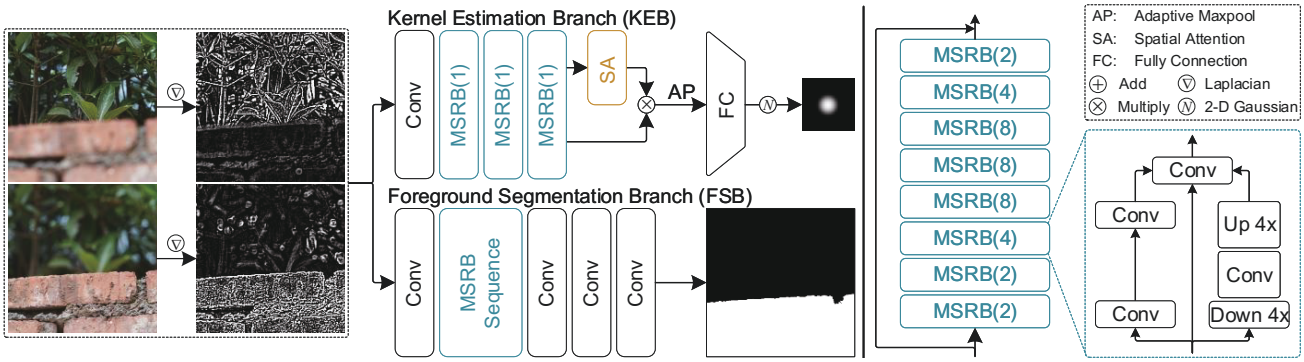
$$\boldsymbol{S} = \bar{\boldsymbol{M}} \cdot (\boldsymbol{F} - \boldsymbol{N}), \tag{5}$$

and can be further expanded as

$$\boldsymbol{S} = \bar{\boldsymbol{M}} \cdot [(\boldsymbol{f} \cdot \boldsymbol{M}) * \boldsymbol{K}_3 + (\bar{\boldsymbol{M}} * \boldsymbol{K}_3) \cdot \boldsymbol{b} - \bar{\boldsymbol{M}} \cdot (\boldsymbol{b} * \boldsymbol{K}_2)]. \tag{6}$$

By manually adding an opposite item pair $\pm(\boldsymbol{b} \cdot \boldsymbol{M}) * \boldsymbol{K_3}$, Eq.6 can be arranged as

$$\begin{aligned} \boldsymbol{S} = {} & \bar{\boldsymbol{M}}[(\boldsymbol{f} \cdot \boldsymbol{M} - \boldsymbol{b} \cdot \boldsymbol{M}) * \boldsymbol{K}_3] - \bar{\boldsymbol{M}} \cdot (\boldsymbol{b} * \boldsymbol{K}_2) \\ & + \bar{\boldsymbol{M}}[(\boldsymbol{b} \cdot \boldsymbol{M}) * \boldsymbol{K}_3 + (\bar{\boldsymbol{M}} * \boldsymbol{K}_3) \cdot \boldsymbol{b}]. \end{aligned} \tag{7}$$



(a) KEB and FSB (b) MSRB Sequence and MSRB

Figure 3: Network designation. (a) Overall structure of Kernel Estimation Branch and Foreground Segmentation Branch. (b) Feature extracting backbone of FSB and internal structure of multi-scale residual block (MSRB).

In the second item of Eq.7, $(\boldsymbol{b}\cdot\boldsymbol{M})*\boldsymbol{K}_3+(\bar{\boldsymbol{M}}*\boldsymbol{K}_3)\cdot\boldsymbol{b}$ constitute a complete image except some details are lost, and $\bar{\boldsymbol{M}}\cdot(\boldsymbol{b}*\boldsymbol{K}_2)$ is the smoothed background, thus this item can be ignored when details are not significant as diffusion component, and we get

$$\boldsymbol{S}=\bar{\boldsymbol{M}}[(\boldsymbol{f}\cdot\boldsymbol{M}-\boldsymbol{b}\cdot\boldsymbol{M})*\boldsymbol{K}_3], \tag{8}$$

showing that diffusion component $\boldsymbol{S}$ is the smooth of color difference between foreground and background near TR with PSF $\boldsymbol{K}_3$, and only exists in background area.

When estimating the diffusion component with Eq.8, $\boldsymbol{M}$ and $\boldsymbol{K}_3$ should be solved first. Considering that the two parameters are irrelevant, we can solve $\boldsymbol{M}$ and $\boldsymbol{K}_3$ separately using two branches of neural networks in next subsection.

Proposed Network

Foreground Segmentation Branch (FSB) Foreground segmentation of near-focus images requires both the integrity of the mask and the accuracy near TR. The former requires the network to have sufficient multi-scale feature extraction capability, while the latter requires the ability to preserve low-level features. Inspired by ResNet (He et al. 2016), we designed the multi-scale residual block (MSRB). In MSRB, residual connections enable the deep network to obtain enough low-level features to ensure the accuracy of boundary segmentation. In order to extract sufficient multi-scale features, in addition to the common path with only two basic convolutions, the multi-scale path contains an additional $down-conv-up$ operation between the two basic convolutions, and features of different scales can be extracted with different downsampling scales.

MSRB can greatly increase the depth of the network, and the deeper network can further strengthen the multi-scale feature extraction ability. In FSB, 8 MSRBs are used after input convolution, and the last MSRB output is sent to 3 continuous convolution layers to obtain the mask. As the network goes deeper, low-level features are unavoidably lost, so we append skip-connection for every two continuous MSRBs and use a long-distance connection to keep boundary segmentation accuracy. Conditional random field (Lafferty, McCallum, and Pereira 2001) is used to ensure boundary segmentation accuracy further.

With the predicted foreground mask $\boldsymbol{M}_{pd}$, the initial synthetic result $\boldsymbol{R}_{init}$ can be obtained by

$$\boldsymbol{R}_{init}=\boldsymbol{M}_{pd}\cdot\boldsymbol{f}+\bar{\boldsymbol{M}}_{pd}\cdot\boldsymbol{b}. \tag{9}$$

Given the ground truth (GT) $\boldsymbol{M}_{gt}$ and $\boldsymbol{R}_{gt}=\boldsymbol{M}_{gt}\cdot\boldsymbol{f}+\bar{\boldsymbol{M}}_{gt}\cdot\boldsymbol{b}$, foreground segmentation branch are trained with the Laplacian gradient loss and map loss, which has the following forms

$$\mathcal{L}_g=\|\nabla\boldsymbol{R}_{gt}-\nabla\boldsymbol{R}_{init}\|_2^2,\ \mathcal{L}_m=\|\boldsymbol{M}_{gt}-\boldsymbol{M}_{pd}\|_1 \tag{10}$$

The Laplacian gradient loss L_g is used to force the network to notice high-frequency components, and the map loss L_m supervises to generate a high-quality mask. The weighted combination of the above terms is the total loss

$$\mathcal{L}_{FSB}=\lambda_1\mathcal{L}_g+\lambda_2\mathcal{L}_m \tag{11}$$

Kernel Estimation Branch (KEB) With Eq.2, estimating PSF $\boldsymbol{K}_3$ when the foreground is defocused can be simplified to estimate a single parameter, *i.e.*, the radius. Any kernel estimation algorithms can be used in our method. Inspired by Luo *et al.* (Luo et al. 2022), we hope that the KEB can accurately estimate this continuous parameter as a regression task, thus we solve it through a fully connected network.

The KEB contains only two MSRBs to extract key features of the input because it is a low-level rather than a high-level feature. After the extraction is completed, a spatial attention (**SA**) module is used to suppress the features of the background region, thus reducing the impact of the background features on the estimation. Further, features that pass through spatial attention will be globally pooled from $\mathbb{R}^{C\times H\times W}$ to a vector of $\mathbb{R}^{C\times 1\times 1}$. The vector will serve as the input of three fully connected layers, and its output σ_{pd} will be activated by the 2-D Gaussian function to obtain the estimated PSF $\boldsymbol{K}_{pd}$.

The KEB is supervised by PSF gorundtruth $\boldsymbol{\sigma}_{gt}$ using

$$\mathcal{L}_{KEB}=\|\boldsymbol{\sigma}_{gt}-\boldsymbol{\sigma}_{pd}\|_1, \tag{12}$$

Diffusion Elimination

When estimating the diffusion component, according to Eq.8, since the background area blocked by the foreground cannot be known from source images, the diffusion component $\boldsymbol{S}_{pd}$ can only be estimated by substitute $\boldsymbol{b} \cdot \boldsymbol{M}_{pd}$ with unblocked background near boundary using

$$\boldsymbol{b} \cdot \boldsymbol{M}_{pd} \approx \frac{(\boldsymbol{b} \cdot \bar{\boldsymbol{M}}_{pd}) * \boldsymbol{K}_{pd}}{\bar{\boldsymbol{M}}_{pd} * \boldsymbol{K}_{pd} + \epsilon}. \tag{13}$$

Combined with Eq.8, the diffusion component is

$$\boldsymbol{S}_{pd} = \bar{\boldsymbol{M}}_{pd}[(\boldsymbol{f} \cdot \boldsymbol{M}_{pd} - \frac{(\boldsymbol{b} \cdot \bar{\boldsymbol{M}}_{pd}) * \boldsymbol{K}_{pd}}{\bar{\boldsymbol{M}}_{pd} * \boldsymbol{K}_{pd} + \epsilon}) * \boldsymbol{K}_{pd}]. \tag{14}$$

The final synthetic result R is obtained by removing the diffusion component from the initial synthetic result

$$\boldsymbol{R} = \boldsymbol{R}_{init} - \boldsymbol{S}_{pd}. \tag{15}$$

Experiments

Experiments Settings

Dataset Due to the scarcity of focus stacking datasets, we used a simulated dataset for training and testing, where the foregrounds were taken from the AM2K (Li et al. 2022) dataset, and the backgrounds were taken from the DIV2K (Agustsson and Timofte 2017) dataset. We used a similar generation tactic as Ma *et al.* (Ma et al. 2020). By combining data in a loop, the whole simulated dataset contains 21.6k focus stacking pairs (19.8k for training and 1.8k for testing) with corresponding GTs, masks, and PSFs, and the image size is 256×256. Additionally, we used two real-world focus stacking datasets in experiments: LytroDataset (Nejati, Samavi, and Shirani 2015) and our BetaFusion dataset, both contain 20 image pairs. We manually annotated the precise foreground masks, which will be used in the testing stage.

Training Details The KEB and the FSB were trained with almost identical settings. The batch size is 32, and the initial learning rate of the AdamW optimizer is set to 1e-4 and will be halved every 25 epochs. The whole training process lasts for about 400 epochs. Parameter λ_1 and λ_2 in Eq.11 are set to 2 and 8, respectively. ϵ in Eq.13 is set to 1e-10. The training was with Nvidia RTX A6000 with 48GB video memory.

Comparison Methods In the comparative experiment, several classical and SOTA methods are compared, including DeFusion (Liang et al. 2022), Liu *et al.* (Liu et al. 2022), MFFNet (Ma et al. 2020), ZMFF (Hu et al. 2023), DRPL (Li et al. 2020), GEU-Net (Xiao et al. 2021), DTMNet (Xiao, Wu, and Bi 2021) and Wang *et al.* (Wang et al. 2022). For fairness, all comparison methods were trained using official code, and all models were well-trained.

Metrics Considering that information fidelity is an important pursuit goal of the focus stacking, we introduce two indicators based on information theory and structure consistency, i.e., Q^{MI} (Hossny, Nahavandi, and Creighton 2008) and Q^{Y} (Yang et al. 2008) as evaluating metrics in comparison experiments. To provide a more comprehensive evaluation of our method, more metrics has been used , such as $Q^{AB/F}$ (Xydeas and Petrović 2000) and mIOU.

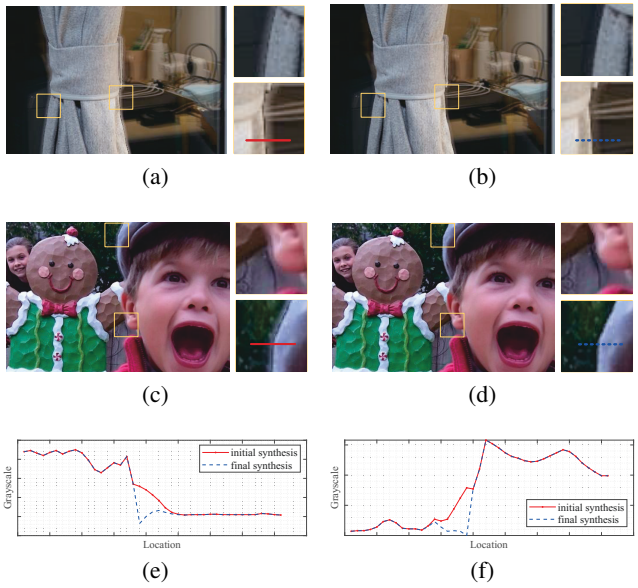


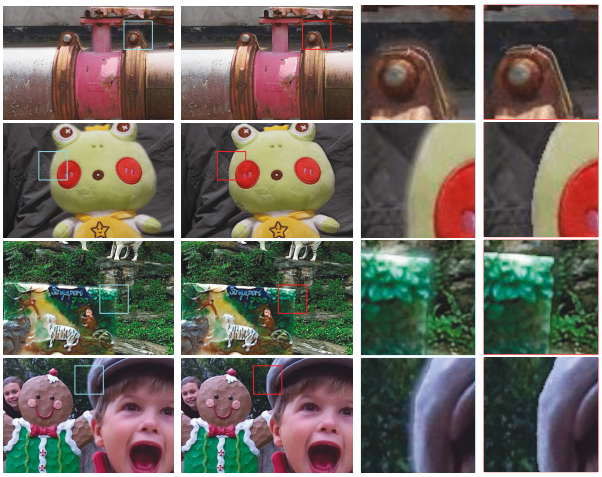
Figure 4: Comparison between initial synthetic image and final synthetic image. (a)(c) and (b)(d) are initial and final synthetic images, respectively. (e) is the 1-D visualization of the marked pixels in (a) and (b), and (f) for (c)(d).

Ablation Study

Pipeline Effectiveness We set an ablation experiment to verify the proposed diffusion elimination operation's effectiveness. The synthetic results without diffusion elimination operation are called the initial synthesis. If diffusion elimination is used, the images are called final synthesis.

A set of samples are shown in Fig.4 to demonstrate the effectiveness of diffusion elimination. Although the foreground segmentation network can segment the foreground accurately, the diffusion components are retained in initial synthesis, covering part of the details of the background area. The diffusion component is effectively removed after applying the diffusion elimination operation, and the final synthesis has better visual effects. Thanks to the accurate modeling of the causes of the diffusion components, details obscured by the diffusion components are effectively maintained during the elimination process. We pick out the same line of pixels respectively from Fig.4a and 4b and form a 1-D visualization in Fig.4e, by the same operation on Fig.4b and 4d we get Fig.4f. From the 1-D visualization, it is clear that the initial and final synthesis have an evident difference in TR. The slope of the red line is the diffusion stack on the background. After removing diffusion components, the slope vanishes in the blue line, proving the success of the diffusion removal operation.

KEB Accuracy The following experiment verifies the accuracy of the kernel estimation. Firstly, a Gaussian convolution kernel K_{gt} with random variance is used to blur the sample I_clear from DIV2K to obtain I_1. Then, KEB is used to estimate the convolution kernel of I_1 to obtain K_{pd}. Fi-



(a) (b) (c) (d)

Figure 5: Universality verification. (a) synthetic results of DRPL without diffusion elimination, (b) results after applying diffusion elimination. (c)(d) enlarged view of (a)(b).

nally, K_{pd} is used to blur I_{clear} to obtain I_2, and the similarity between I_1 and I_2 is calculated. After conducting such an experiment on the whole DIV2K, we use SSIM and PSNR to measure the similarity between I_1 and I_2, and the scores are $SSIM(I_1, I_2) = 0.9571, PSNR(I_1, I_2) = 35.46$. Generally, when $SSIM > 0.9$ or $PSNR > 30$ can prove the high consistency of the two images, thus our KEB is dependable.

Universality of Diffusion Elimination

The proposed method can work as general architecture. Our KEB is independent of segmentation networks and only requires an accurate mask, which means our method can be ported to the other segmentation structure. To verify this, we combine DRPL and our KEB, and try to optimize the initial synthetic results of DRPL. Qualitative comparisons with and without our diffusion elimination operation are shown in Fig.5. Before eliminating the diffusion component, the transition band is obvious, as Fig.5a and 5c show. After applying diffusion elimination to Fig.5a, the TR problem has been well addressed, and the synthetic results look more natural.

Our diffusion elimination operation can process the initial synthetic results of most decision-based methods. In practice, only when there is a precise mask, especially near TR, can we improve their synthetic results.

Discussion: Process with Inaccurate Masks

Obtaining accurate masks is the premise for decision-based methods. However, since our modeling of the diffusion component is based on the color difference between the foreground and the background, deterioration caused by inaccurate segmentation is not severe. There are two types of inaccurate segmentation: regions with various colors and regions with sparse colors. If a monotonous region is poorly segmented, the colors near the mask's boundary have tiny differences. The estimated diffusion component tends to be 0, and the synthetic image is almost unchanged. In another case, an inaccurate diffusion component will be computed, leading to unexpected damage to the synthetic image. Fortunately, regions with rich colors are usually full of details, and such inaccurate segmentation rarely happens.

	Method	Q^{MI}	$Q^{AB/F}$	Q^{Y}	mIOU	PSNR
Recons.	DeF.	0.813	0.607	0.777	-	28.6
	Liu's	1.237	<u>0.904</u>	**0.899**	-	<u>36.3</u>
	MFF.	1.205	0.900	0.893	-	35.6
	ZMFF	-	-	-	-	-
Decis.	DRPL	<u>1.238</u>	0.901	0.892	<u>0.955</u>	35.5
	GEU.	1.204	0.885	0.886	0.856	34.5
	DTM.	1.230	0.898	0.889	0.907	35.2
	Wang's	1.235	0.899	0.890	0.845	35.4
	Ours	**1.238**	**0.906**	<u>0.898</u>	**0.966**	**36.9**

Table 1: Quantitative comparison on Simulated dataset.

Comparison on Simulated Dataset

Since the real-world focus stacking datasets lack GT, evaluating the quality of synthetic images near TR is difficult. Therefore, a simulated dataset containing GT was adopted for testing. Considering the existence of GT, we set the reference object of all the metrics to be GT images. For example, Q^Y, which used to refer to two different original images, will now refer to GT images directly. In addition, we temporarily add a metric *PSNR* to measure the fidelity of the synthetic image with GT. Due to the long inference time of ZMFF, we did not provide corresponding results.

The objective results of this experiment are shown in Tab.1, where it can be seen that on the testing set of 1.8k samples, the synthesized images of our method are closer to GT images than other methods, demonstrating the high fidelity of our method. For subjective results of this experiment, please refer to the supporting material.

Comparison on Real-world Datasets

In order to verify the effectiveness of our method in real scenarios, we conducted a comparative experiment on the Beta-Fusion dataset and LytroDataset. We provide one sample in each dataset in Fig.6, and more samples are placed in the supporting material. In the given samples, it can be clearly seen that our method can obtain better synthesis results on TR and effectively avoid the TR problem in DRPL, DTM-Net, GEU-Net, or the detail loss problems in Wang *et al.*'s method. By subtracting the synthetic image from source image A and following appropriate linear mapping, the residual image can be obtained to reveal the color bias problem in the synthesis process. From the residual images, it can be seen that our method avoids the color bias problem that is common in reconstruction methods.

Quantitative comparisons are shown in Fig.7. This figure evaluates the synthetic image from two aspects. The vertical axis evaluates the information fidelity of the methods by Q^{MI}, and the horizontal axis evaluates the consistency between the synthetic image and the visual perception preference by Q^{CV}. The Q^{CV} score is the probability that the

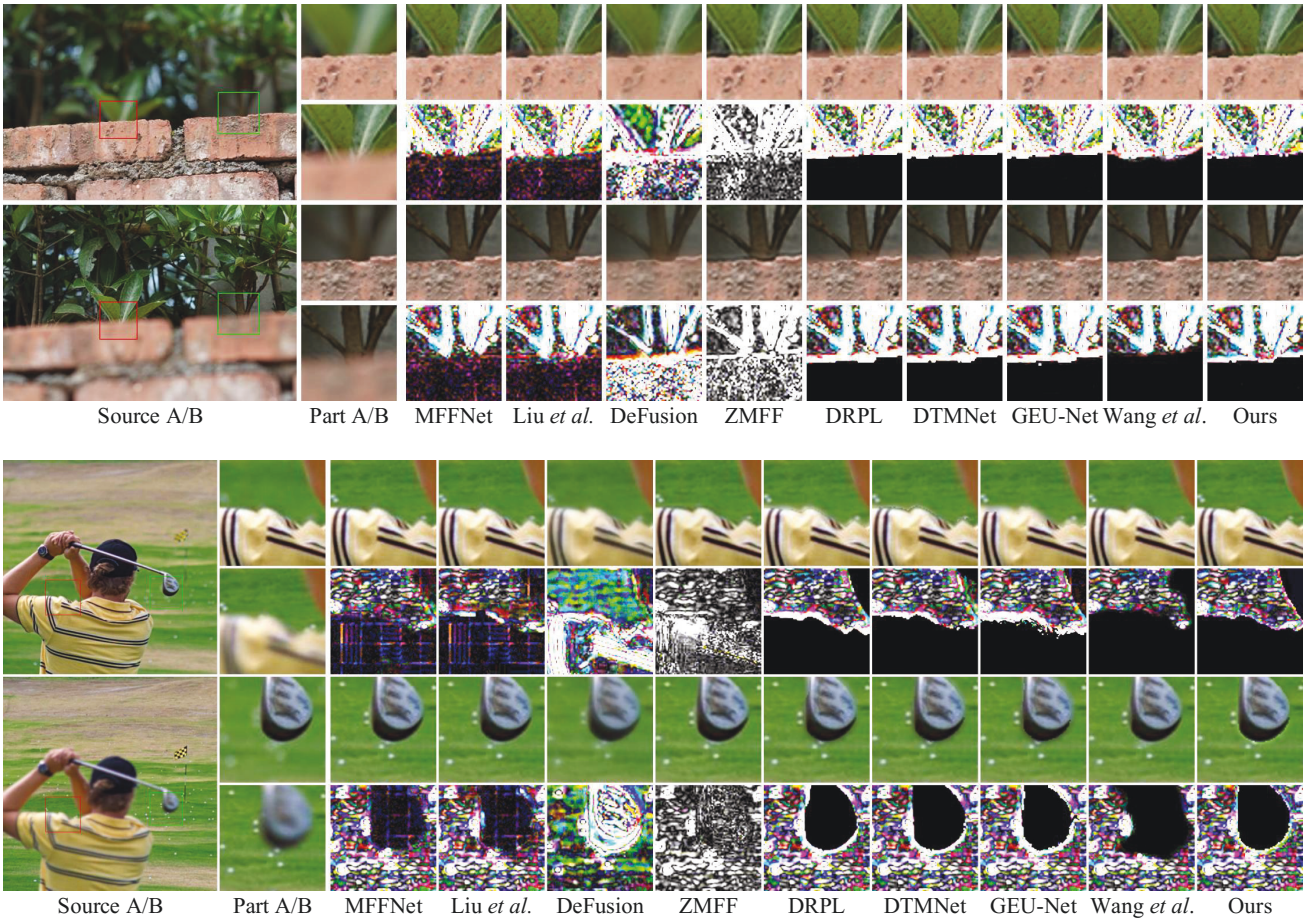
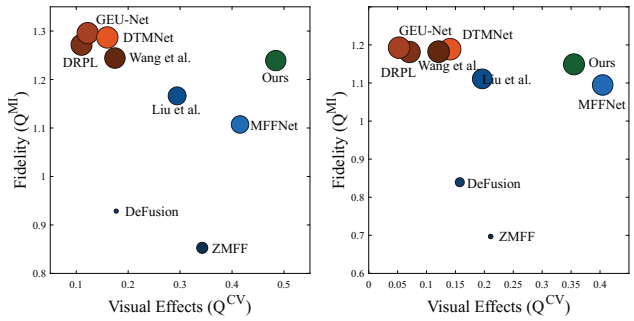


Figure 6: Qualitative comparison with SOTA methods. Notice the boundary between foreground and background.



(a) BetaFusion (b) LytroDataset

Figure 7: Quantitative comparison on BetaFusion dataset and LytroDataset. The radius of circle stands for Q^Y score.

sample is classified as a "high-quality synthesis" by the network. See supporting material for detail. To better evaluate these methods, we introduce another metric revealing the information fidelity of synthetic results, Q^Y, into the plots as the circle radius. Obviously, our approach strikes a good balance between high fidelity and superior visual effects.

Conclusion

This paper analyzes the formation of the diffusion phenomenon and builds an optical model for the diffusion component. We first put forward a foreground segmentation network that can precisely segment the foreground of a near-focus image and obtain an initial synthetic image. A point spread function estimation network is then proposed, with which the diffusion component of the initial result can be accurately eliminated. Further, we conducted qualitative and quantitative experiments with SOTA methods, and the results prove the superiority of the proposed method.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62376046, Grant 62172067 and Grant 61976031, in part by the Natural Science Foundation of Chongqing for Distinguished Young Scholars under Grant CSTB2022NSCQ-JQX0001, the Science and Technology Research Program of Chongqing Municipal Education Commission (Grant No. KJQN202200635), and CSTB2023NSCQ-MSX0341.

References

Agustsson, E.; and Timofte, R. 2017. NTIRE 2017 Challenge on Single Image Super-Resolution: Dataset and Study. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 1122–1131. Honolulu, HI, USA: IEEE. ISBN 978-1-5386-0733-6.

Deng, X.; and Dragotti, P. L. 2021. Deep Convolutional Neural Network for Multi-Modal Image Restoration and Fusion. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(10): 3333–3348.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778. Las Vegas, NV, USA: IEEE. ISBN 978-1-4673-8851-1.

Hossny, M.; Nahavandi, S.; and Creighton, D. 2008. Comments on 'Information measure for performance of image fusion'. *Electron. Lett.*, 44(18): 1066.

Hu, X.; Jiang, J.; Liu, X.; and Ma, J. 2023. ZMFF: Zero-shot multi-focus image fusion. *Information Fusion*, 92: 127–138.

Lafferty, J. D.; McCallum, A.; and Pereira, F. C. N. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, 282–289. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. ISBN 1-55860-778-1.

Lai, S.-H.; Fu, C.-W.; and Chang, S. 1992. A generalized depth estimation algorithm with a single image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(4): 405–411.

Li, H.; Nie, R.; Cao, J.; Guo, X.; Zhou, D.; and He, K. 2019. Multi-Focus Image Fusion Using U-Shaped Networks With a Hybrid Objective. *IEEE Sensors J.*, 19(21): 9755–9765.

Li, J.; Guo, X.; Lu, G.; Zhang, B.; Xu, Y.; Wu, F.; and Zhang, D. 2020. DRPL: Deep Regression Pair Learning for Multi-Focus Image Fusion. *IEEE Trans. on Image Process.*, 29: 4816–4831.

Li, J.; Yuan, G.; and Fan, H. 2019. Multifocus Image Fusion Using Wavelet-Domain-Based Deep CNN. *Computational Intelligence and Neuroscience*, 2019: 1–23.

Li, J.; Zhang, J.; Maybank, S. J.; and Tao, D. 2022. Bridging Composite and Real: Towards End-to-End Deep Image Matting. *Int J Comput Vis*, 130(2): 246–266.

Liang, P.; Jiang, J.; Liu, X.; and Ma, J. 2022. Fusion from Decomposition: A Self-Supervised Decomposition Approach for Image Fusion. In *Computer Vision – ECCV 2022*, volume 13678, 719–735. Cham: Springer Nature Switzerland.

Liu, Y.; Wang, L.; Li, H.; and Chen, X. 2022. Multi-focus image fusion with deep residual learning and focus property detection. *Information Fusion*, 86-87: 1–16.

Luo, Z.; Huang, H.; Yu, L.; Li, Y.; Fan, H.; and Liu, S. 2022. Deep Constrained Least Squares for Blind Image Super-Resolution. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 17621–17631. New Orleans, LA, USA: IEEE. ISBN 978-1-66546-946-3.

Ma, B.; Zhu, Y.; Yin, X.; Ban, X.; Huang, H.; and Mukeshimana, M. 2021. SESF-Fuse: an unsupervised deep model for multi-focus image fusion. *Neural Comput & Applic*, 33(11): 5793–5804.

Ma, H.; Liao, Q.; Zhang, J.; Liu, S.; and Xue, J.-H. 2020. An α-Matte Boundary Defocus Model-Based Cascaded Network for Multi-Focus Image Fusion. *IEEE Trans. on Image Process.*, 29: 8668–8679.

Mustafa, H. T.; Yang, J.; and Zareapoor, M. 2019. Multi-scale convolutional neural network for multi-focus image fusion. *Image and Vision Computing*, 85: 26–35.

Mustafa, H. T.; Zareapoor, M.; and Yang, J. 2020. MLDNet: Multi-level dense network for multi-focus image fusion. *Signal Processing: Image Communication*, 85: 115864.

Nejati, M.; Samavi, S.; and Shirani, S. 2015. Multi-focus image fusion using dictionary-based sparse representation. *Information Fusion*, 25: 72–84.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, volume 9351, 234–241. Cham: Springer International Publishing.

Tang, H.; Xiao, B.; Li, W.; and Wang, G. 2018. Pixel convolutional neural network for multi-focus image fusion. *Information Sciences*, 433-434: 125–141.

Wang, Z.; Li, X.; Duan, H.; and Zhang, X. 2022. A Self-Supervised Residual Feature Learning Model for Multifocus Image Fusion. *IEEE Trans. on Image Process.*, 31: 4527–4542.

Xiao, B.; Wu, H.; and Bi, X. 2021. DTMNet: A Discrete Tchebichef Moments-Based Deep Neural Network for Multi-Focus Image Fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 43–51.

Xiao, B.; Xu, B.; Bi, X.; and Li, W. 2021. Global-Feature Encoding U-Net (GEU-Net) for Multi-Focus Image Fusion. *IEEE Trans. on Image Process.*, 30: 163–175.

Xydeas, C.; and Petrović, V. 2000. Objective image fusion performance measure. *Electron. Lett.*, 36(4): 308.

Yang, C.; Zhang, J.-Q.; Wang, X.-R.; and Liu, X. 2008. A novel similarity based quality metric for image fusion. *Information Fusion*, 9(2): 156–160.

Zhang, H.; Le, Z.; Shao, Z.; Xu, H.; and Ma, J. 2021. MFF-GAN: An unsupervised generative adversarial network with adaptive and gradient joint constraints for multi-focus image fusion. *Information Fusion*, 66: 40–53.

Zhao, W.; Wang, D.; and Lu, H. 2019. Multi-Focus Image Fusion With a Natural Enhancement via a Joint Multi-Level Deeply Supervised Convolutional Neural Network. *IEEE Trans. Circuits Syst. Video Technol.*, 29(4): 1102–1115.

Zhuo, S.; and Sim, T. 2011. Defocus map estimation from a single image. *Pattern Recognition*, 44(9): 1852–1858. Computer Analysis of Images and Patterns.