J. Vis. Commun. Image R. 87 (2022) 103580

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

journal homepage: www.elsevier.com/locate/jvci

Full Length Article

Self-supervised multi-scale pyramid fusion networks for realistic bokeh effect rendering☆

Zhifeng Wang [a], Aiwen Jiang [a,*], Chunjie Zhang [b], Hanxi Li [a], Bo Liu [c]

[a] School of Computer and Information Engineering, Jiangxi Normal University, No. 99, Ziyang Ave., Nanchang, 330022, Jiangxi, China
[b] School of Computer and Information Technology, Beijing Jiaotong University, No. 3 Shangyuancun, Haidian District, Beijing, 100044, China
[c] Department of Computer Science and Software Engineering, Auburn University, 3101P Shelby Center for Engineering Technology, Auburn, 36849-5347, AL, USA

ARTICLE INFO

Keywords:
Bokeh rendering
Circle of confusion
Self-supervised
Multi-scale fusion
Structure consistency

ABSTRACT

Images with visual pleasing bokeh effect are often unattainable for mobile cameras with compact optics and tiny sensors. To balance the aesthetic requirements on photo quality and expensive high-end SLR cameras, synthetic bokeh effect rendering has emerged as an attractive machine learning topic for engineering applications on imaging systems. However, most of bokeh rendering models either heavily relied on prior knowledge such as scene depth or were topic-irrelevant data-driven networks without task-specific knowledge, which restricted models' training efficiency and testing accuracy. Since bokeh is closely related to a phenomenon called "circle of confusion", therefore, in this paper, following the principle of bokeh generation, a novel self-supervised multi-scale pyramid fusion network has been proposed for bokeh rendering. During the pyramid fusion process, structure consistencies are employed to emphasize the importance of respective bokeh components. Task-specific knowledge which mimics the "circle of confusion" phenomenon through disk blur convolutions is utilized as self-supervised information for network training. The proposed network has been evaluated and compared with several state-of-the-art methods on a public large-scale bokeh dataset- the "EBB!" Dataset. The experiment performance demonstrates that the proposed network has much better processing efficiency and can achieve better realistic bokeh effect with much less parameters size and running time. Related source codes and pre-trained models of the proposed model will be available soon on https://github.com/zfw-cv/MPFNet.

1. Introduction

"Bokeh" is Japanese in origin and refers to a blurry quality. In photography, it is a very recognizable technique, which can lead pleasing visual aesthetic photos, as shown in Fig. 1.

In practice, images with visual pleasing bokeh effect are often produced through professional DLSR camera with large aperture and long focal length. However, they are often unattainable for mobile cameras with compact optics and tiny sensors. To balance the aesthetic requirements on photo quality and expensive high-end SLR cameras, bokeh effect has to be simulated computationally. Therefore, synthetic bokeh effect rendering has emerged as an attractive machine learning technology [1,2] for engineering applications on imaging systems.

During the past years, many methods on synthetic bokeh effect rendering relied heavily on prior knowledge such as scene depth. Among these depth-based methods, some methods adopted to estimate scene depth through utilizing hardwares like the dual-pixel autofocus system [3] on Google Pixel devices, the dual-lens on iPhone7+ and the Time-of-Flight (TOF) lens on Huawei P30+ smartphones. However, since these specialized hardwares are expensive, they are often not supported on low-end commercial systems. Moreover, for images already captured using monocular cameras, the accurate depth information are not available either. Therefore, many other methods proposed to employ pre-trained models such as MegaDepth [4] to estimate the depth.

To some degree, incorporating prior knowledge to simulate realistic bokeh blur has potentials to improve visual effect of the final generated image. However, every thing has the pros and the cons. The typical limitations of prior-based methods are: (1) the depth sensor related hardwares are not always available on mobile devices; (2) preprocessing prior information by software is generally time-consuming; (3) when the prior information estimated does not work, unexpected out-of-focus blurriness conversely deteriorates the quality of ultimate synthetic image.

Witnessed the impressive success on image-to-image translation tasks, in recent years, many researchers started to consider bokeh

☆ This paper has been recommended for acceptance by Zicheng Liu.

* Corresponding author.
E-mail address: jiangaiwen@jxnu.edu.cn (A. Jiang).

https://doi.org/10.1016/j.jvcir.2022.103580
Received 6 December 2021; Received in revised form 4 May 2022; Accepted 26 June 2022
Available online 30 June 2022

Fig. 1. The examples of bokeh-free and bokeh images. The left are bokeh-free images. The right are with bokeh effects which highlight object of interest in focus.

simulation as a subtask of image translation [5–7]. Therefore, routines based end-to-end multi-scale encoder–decoder architecture are commonly adopted as camera-independent solutions. The nonlinear mappings between the low- and high-aperture photos captured with high-end DSLR camera are directly modeled in data-driven way. However, though much progress has been achieved, the majority of these models are topic-irrelevant networks. In other words, it means in these cases, task-specific knowledges like the intrinsic mechanism of bokeh generation are often neglected without fully utilization for more effective solutions.

According to the principle of optical imaging [8,9], bokeh generation is closely related to a phenomenon called "circle of confusion (CoC)". The "CoC" approximately brings different sizes of disk blurs on out-of-focus areas. Therefore, following the task-specific knowledge, we propose a novel self-supervised multi-scale pyramid fusion network for bokeh rendering. In the proposed network, blurred images after disk convolution kernels of different radius provide self-supervised informations for bokeh component learning. The final bokeh image is a weighted combination of the learned factorized bokeh components. Therefore, different from existing heavy prior-dependent algorithms, the proposed network does not have to preprocess time-wasting priors during training and testing in practice. At the same time, in contrary to some topic-irrelevant end-to-end networks, the proposed network employs task-specific knowledge as training guidance. With more clear training purpose, it can effectively boost network training's efficiency and accuracy.

The proposed network has been evaluated and compared with several state-of-the-art methods on a public large-scale bokeh dataset-the "EBB!" Dataset [5]. The experiment performance demonstrates that the proposed network has much better processing efficiency and can achieve better realistic bokeh effect with much less parameters size and running time.

The contributions of this paper are summarized as followings:

- An effective multi-scale pyramid fusion network is proposed for realistic bokeh effect rendering. The proposed network can achieve new state-of-the-art performance on a large-scale bokeh benchmark dataset with relatively small parameter size and real-time processing speed.
- Task-specific knowledge which mimics the "circle of confusion" phenomenon through disk blur convolutions is utilized as self-supervised information for network training. Structure consistencies are employed to emphasize the importance of respective bokeh components. With more clear training purpose, it can effectively boost network training's efficiency and accuracy.

In the following sections, related work will be summarized in Section 2. Details on the proposed network will be described in Section 3. Then experiment results and analysis are demonstrated in Section 4. Finally, conclusions will be given in Section 5.

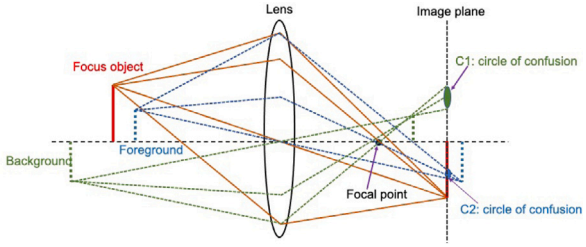


Fig. 2. Illustrations on depth of field.

2. Related work

2.1. Bokeh and depth-of-field

The bokeh effect is optically called "circle of confusion". According to the principle of optical imaging [8,9], as illustrated in Fig. 2, for a specific aperture and focal length, only object points at focus plane (also refers to focus object) can be ideally projected to corresponding points on image plane. Any other object points before or after the ideal plane are out of focus and form circles of confusion when they are projected onto image plane.

The "circle of confusion", such as the C1 and C2 illustrated in Fig. 2, can be in various size, depending on the distances between their respectively captured objects and the focus object. Generally, the nearer the distance, the smaller the size of the circle. If the diameter size of a circle is at the edge of the perception limitation of human eyes, the circle is called "permissible circle of confusion". On image plane, objects with circle of confusion smaller than the "permissible circle of confusion" have similar clarity to the focus object, otherwise they will suffer varying degrees of blurs. In photography, the maximum distance between the objects having "permissible circle of confusion" before and after the focused object is call "depth of field".

Besides the objects distance away from image plane, "depth-of-field" is also closely correlated to aperture and focal length. The larger the aperture or the longer the focal length, the shallower the depth-of-field. Generally, proper shallow depth-of-field is the precondition to obtain good bokeh effect.

2.2. Automatic bokeh effect rendering

Automatic bokeh effect rendering has developed for several years. In this section, we are mainly concentrated on introducing its recent developments on modeling strategies and bokeh content.

2.2.1. Modeling strategies

From the aspect of modeling strategies, in the past, many works on bokeh effect rendering have involved capturing depth information with dual-cameras. Typically, Busam et al. [10] proposed to use high-quality vision disparity map to refocus images through stereo depth estimation. Luo et al. [11] proposed to generate high-definition disparity maps through wavelet synthesis neural network based on a pair of calibrated stereo images. Liu et al. [12] presented a bokeh simulation method based on depth map which was obtained with stereo matching. Jeong et al. [13] presented a real-time bokeh rendering technique that splats pre-computed sprites but takes dynamic visibilities and intrinsic appearances into account at runtime. However, when dealing with monocular camera or post-processing already captured images, these dual-cameras based methods often become invalid.

Therefore, single image based bokeh effect simulation emerges as a machine learning hot-topic in recent years. Typically, Xu et al. [14] exploited both depth estimation network and portrait segmentation network to conduct blur rendering on input image with a conditional

random field [15]. Purohit et al. [16] proposed a depth guided dynamic filtering dense network for bokeh rendering. Pre-trained depth estimation and salient segmentation maps are concatenated with input image along channel dimension before being propagated through their proposed densely connected encoder–decoder network. Dutta [17] proposed to blend original image and different versions of smoothed images to generate bokeh effect with the help of a monocular depth estimation network. Wang et al. [18] proposed a light field refocusing method to improve the imaging quality of camera arrays.

In these depth-map based models, depth estimation [19] is an indispensable step, since different levels of blurring are generally introduced depending on the depth variations in scene content. Therefore, the quality of estimated depth maps is critically important to final rendering effect. Godard et al. [20] proposed a self-supervised learning model to perform monocular depth estimation. Minimum re-projection loss and auto-masking loss were designed to improve the estimation quality of depth maps. Zuo et al. [21] proposed a deep residual dense network to progressively reconstruct high-resolution depth map guided by the intensity image. Song et al. [22] proposed a scene-aware contextualized convolution neural network for intrinsic exploitation of context-dependent depth association, including inner-object continuous depth and inter-object depth change priors nearby.

Witnessed the impressive success on image-to-image translation tasks, such as image deblur [23,24], super-resolution [25,26], style transfer [27,28], image enhancement [29–31], image dehazing [32,33], etc., in recent years, researchers consider bokeh synthetic as a kind of image translation task. Typically, on the AIM 2019 Challenge on bokeh effect synthesis [1], XMU-VIPLab team (Yang et al.) employed selective kernel networks (SKNet) [34] for bokeh effect simulation. Two bokeh-nets were trained to generate local and global features before being concatenated as input into the SKNet. VIDAR team (Xiong et al.) use an ensemble of five U-Net based models with residual attention mechanism to achieve final bokeh image. Qian et al. [6] proposed a GAN-based method solves the synthetic bokeh effect rendering problem. On the AIM 2020 Challenge [2], both CET-CVLab and CET-SP teams use the same U-Net based dilated wavelet CNN model [35] for generating bokeh images. In their networks, standard down-sampling and up-sampling operations are replaced by decomposition based on discrete wavelet transform (DWT) to minimize information loss in these layers. Moreover, Ignatov et al. [5] proposed a multi-scale end-to-end PyNet structure for image rendering. Dutta et al. [7] proposed a deep multi-scale hierarchical network (DMSHN) for bokeh effect rendering. Under the "coarse-to-fine" scheme, their model synthesized bokeh effect by exploiting multi-scale input images at different processing levels. Each lower level acts in the residual manner by contributing its residual image to the higher level. Luo et al. [36] proposed a multi-stage network to learn shallow depth-of-field from a single bokeh-free image through defocus estimation.

In this paper, the proposed model no longer requires to estimating the depth-map of bokeh-free image. It puts forward a lightweight and fast network for efficient end-to-end bokeh effect rendering.

2.2.2. Bokeh contents

From the aspect of image content processed, in early stage, portrait images were mainly considered only. Saliency detection [37–39] is an indispensable step in the portrait-only methods. Typically, Shen et al. [40] proposed to employ fully convolution network for portrait segmentation. Then their model simulated shallow depth-of-field image through uniformly blurring segmented background. Wadhwa et al. [3] combined person segmentation network and dense dual-pixel auto-focus hardware to render a defocused image.

Heavily relying on portrait segmentation often failed to give good rendering effects for generic scenes. Therefore, with the aim of improving the adaptivity and quality of simulated bokeh effects, challenges competitions [1,2] were successively organized as computer vision workshops to gauge and push the state-of-the-art in synthetic shallow depth-of-field rendering. In the challenge workshop, a large scale bokeh dataset -"EBB!" dataset was distributed. The dataset contained more than 10 thousand images collected in the wild on generic scenes.

In this paper, the proposed model does not rely on saliency detection, therefore, has widely applicabilities on scenes with generic contents.

Table 1
The convolution details of sub-blocks in decoder $\mathbf{FE}_i$. All convolutions are with kernel size $k = 3 \times 3$ and padding size $p = 1$. "sk_m_n" means stride size $s = k$, the number of input channels is m, and the number of output channels is n.

	block1	block2	block3	block4
$Conv_1$	s1_3_32	s2_32_64	s2_64_128	s2_128_256
$Conv_2$	s1_32_32	s1_64_64	s1_128_128	s1_256_256
$Conv_3$	s1_32_32	s1_64_64	s1_128_128	s1_256_256

3. Methodology

In this section, we describe the proposed method in details. The architecture of the proposed network is illustrated in Fig. 3. Multi-scale information fusions on three pyramid levels are considered in this network.

Specifically, original image is first pyramidally downsampled into three variants of different resolutions. The downsampling factor is 2. Each variant is encoded by respective feature extraction module $\mathbf{FE}_i, i = \{1, 2, 3\}$, where i is pyramid level index. The encoded features are denoted as $\mathbf{f}_i$.

With the aim to maximumly utilize information from neighboring scales, a kind of cyclic fusion is performed on each pyramid level. Therefore, the corresponding feature $\mathbf{t}_{i,j}$ is then generated, which represents enhancing information at level i by the information from level j. Decoders $\mathbf{Gen}_{i,j}$ are then responsible for generating respective image with bokeh component of certain blur radius from $\mathbf{t}_{i,j}$.

Preprocessing modules $\mathbf{RoughPre}[i], i = \{1, 2, 3\}$ are employed on each pyramid level, mimicking depth-of-field images with different extent of blurriness and of various resolutions. They can supply self-supervision information for effectively boosting network training's efficiency and accuracy. Finally, based the structural similarities between herein self-defined blurred "ground-truths" and the generated image components, adaptive importances are emphasized on each generated components. The ultimate image with expected bokeh effect is then produced through weighted combinations of these generated components.

In the following, the structure details of each proposed modules are described.

3.1. Encoders "$\mathbf{FE}_i$"

Encoder "$\mathbf{FE}_i$" plays a role of feature extraction for original image at each pyramid level i. All "$\mathbf{FE}_i, i = \{1, 2, 3\}$" share the same network structure but with respective parameters. The structure details are shown in Fig. 4. The network of $\mathbf{FE}_i$ consists of four successive sub-blocks with similar inner structures. As shown in Fig. 4, the sub-block is composed of three convolutions in residual way, which is formulated as in Eq. (1). Details of the convolution structures are illustrated in Table 1.

$$\begin{aligned} X_{tmp} &= \mathrm{Conv}_1(X_{in}) \\ X_{out} &= I_{in} + \mathrm{Conv}_3(\mathrm{ReLU}(\mathrm{Conv}_2(X_{tmp}))) \end{aligned} \tag{1}$$

where, X_{in} represents input feature map of the sub-block in $\mathbf{FE}_i$ network. X_{out} represents corresponding output of the sub-block. For simplicity, we omit the index i for each pyramid level.

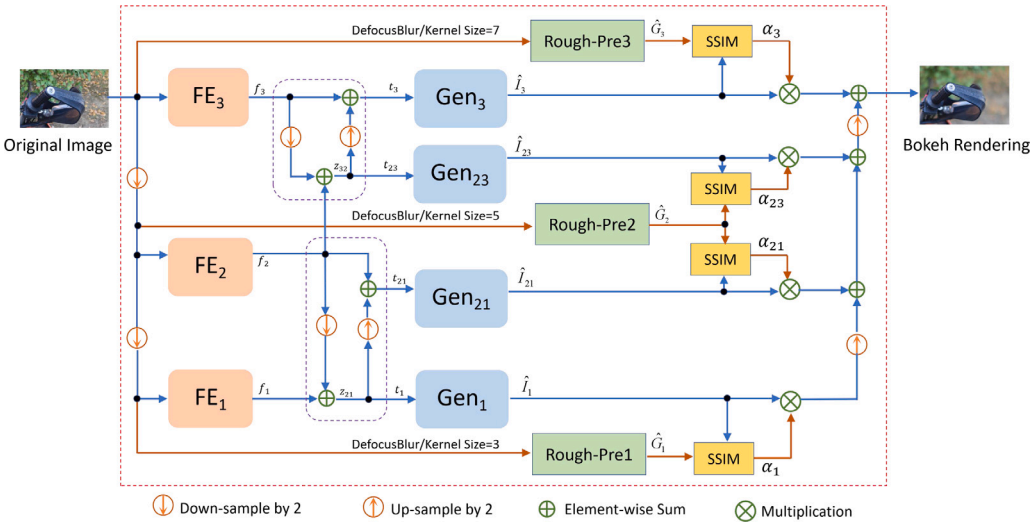


Fig. 3. The architecture of the proposed multi-scale pyramid fusion networks.

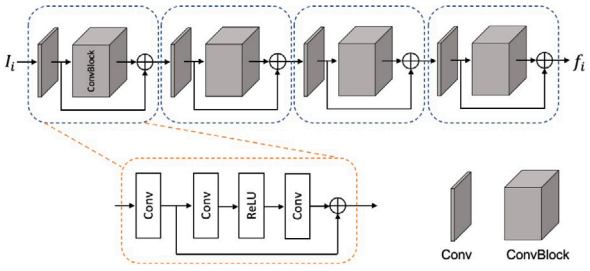


Fig. 4. The structure of the encoder module $\mathbf{FE}_i$.

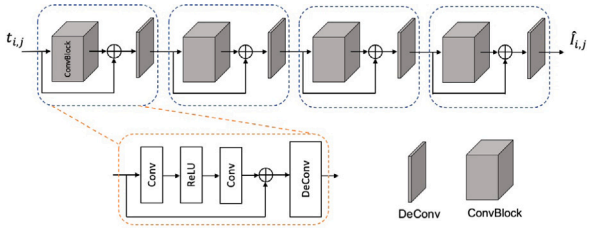


Fig. 5. The structure of the decoder module $\mathbf{Gen}_{i,j}$.

Fig. 6. The visual illustrations of defocus blurs.

Table 2
The convolution details of sub-blocks in decoder $\mathbf{Gen}_{i,j}$. All convolutions are with kernel size $k = 3 \times 3$ and padding size $p = 1$. All DeConvolutions are with kernel size $k = 4 \times 4$ and padding size $p = 1$. "sk_m_n" means stride size $s = k$, the number of input channels is m, and the number of output channels is n.

	block1	block2	block3	block4
$Conv_1$	s1_256_256	s1_128_128	s1_64_64	s1_32_32
$Conv_2$	s1_256_256	s1_128_128	s1_64_64	s1_32_32
DeConv_3	s2_256_128	s2_128_64	s2_64_32	s2_32_3

3.2. Cyclic fusion

Cyclic fusion plays a role of information reutilization and interpolation between neighboring pyramid levels. Just as shown in Eq. (2), $\mathbf{z}_{i,j}$ represents downward information fusion from neighboring scale of higher resolution. $\mathbf{t}_{i,j}$ represents upward fusion from neighboring scale of lower resolution after $\mathbf{z}_{i,j}$ is obtained. Herein, we briefly denote $\mathbf{t}_{i,j}$ as $\mathbf{t}_i$ in case of $i = j$.

$$\begin{aligned} & & t_3 &= f_3 + z_{32}\uparrow \\ z_{32} &= f_3\downarrow + f_2 & t_{23} &= z_{32} \\ z_{21} &= f_2\downarrow + f_1 & t_{21} &= z_{21}\uparrow + f_2 \\ & & t_1 &= z_{21} \end{aligned} \tag{2}$$

Herein, $\downarrow$ means downscale sampling and $\uparrow$ means upscale sampling.

The cyclic manipulations may introduce some information disturbance on sub-pixels through successive downward and upward fusion, which are expected to enhance model's robustness on different scales.

3.3. Decoders "$\mathbf{Gen}_{i,j}$"

$\mathbf{Gen}_{i,j}$ are decoders that generate images with bokeh components of different resolutions from corresponding encoded features $\mathbf{t}_{i,j}$, $\hat{I}_{i,j} = \mathbf{Gen}_{i,j}(t_{i,j})$. The structure of $\mathbf{Gen}_{i,j}$ is illustrated in Fig. 5. It has similar reverse structure when compared with encoders $\mathbf{FE}_i$.

The network of $\mathbf{Gen}_{i,j}$ consists of four successive sub-blocks with similar inner structures. Each sub-block is composed of two convolutions followed by a deconvolution, which is formulated as in Eq. (3).

$$\begin{aligned} Y_{tmp} &= Y_{in} + \text{Conv}_2(\text{ReLU}(\text{Conv}_1(Y_{in}))) \\ \hat{Y}_{out} &= \text{DeConv}_3(Y_{tmp}) \end{aligned} \tag{3}$$

where, Y_{in} represents input feature map of the sub-block in $\mathbf{Gen}_{i,j}$ network. Y_{out} represents corresponding output of the sub-block. For simplicity, we omit the indexes i, j at respective pyramid level.

The inner convolution details of the sub-blocks are described in Table 2.

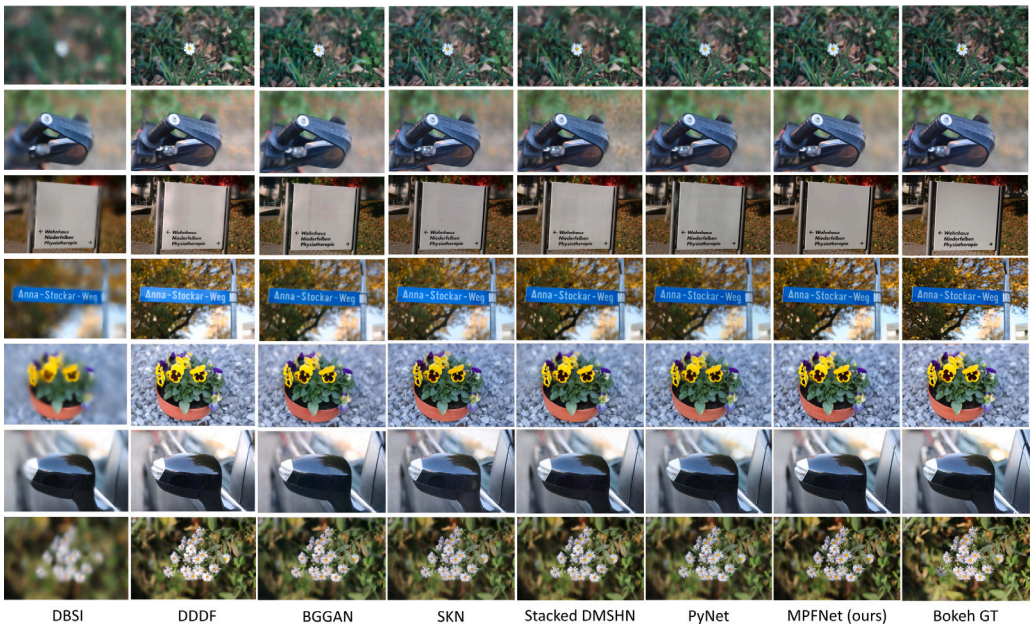


Fig. 7. Samples of visual comparisons among the proposed network and some representative state-of-the-art methods. Images predicted by the proposed methods have more realistic bokeh effects with less artificial defects, such as the road plates in the third row.

3.4. *DefocusBlur* module "**RoughPre**[i]"

"**RoughPre**$[i], i = \{1,2,3\}$" are preprocessing modules for bokeh blurs. They provide self-supervised ground-truth for bokeh components learning.

As explained in Fig. 2, the bokeh blurs are in fact resulted by circles of confusion. In the field of graphics, there are many methods to simulate the bokeh blurs. Typically, circular scatter is the most standard assumption. It is a kind of disk blur. The visual effects of disk blurs are illustrated in Fig. 6.

The bokeh blur algorithm implemented is denoted as *DefocusBlur*. It involves using circular convolution kernel for bokeh simulation. Since the variations of bokeh radius are controlled by the variations of blur radius, **RoughPre**[i] employs different kernel size for *DefocusBlur* on pyramid levels, $\hat{G}_i = DefocusBlur(I_i, k_i)$, where $k_i = \{7,5,3\}$ respectively represents the kernel size of defocus blur at corresponding pyramid level.

In order to supply self-supervision information, structural similarities $\alpha_{i,j} = \text{SSIM}(\hat{G}_i, \hat{I}_{i,j})$ are computed on each ith pyramid level between the blurred "ground-truths" $\hat{G}_i$ and the predicted image components $\hat{I}_i$ from $\mathbf{Gen}_{i,j}$. The similarity $\alpha_{i,j}$ indicates the quality of respective bokeh component learning. We jointly normalized them into range $[0,1]$ to emphasize their importances, as shown in Eq. (4).

$$w_{i,j} = \frac{\alpha_{i,j}}{\sum_m \alpha_{m,j}} \tag{4}$$

where $w_{i,i}$ are briefly denoted as w_i, and $\alpha_{i,i}$ as α_i.

The final predicted bokeh output $\hat{I}_b$ is generated through fusing the bokeh components predicted at each pyramid levels, as formulated in Eq. (5). The jointly normalized similarities w_* adaptively control the importance of information from respective pyramid level during combination.

$$\begin{aligned} P_1 &= w_1 * \hat{I}_1 \\ P_2 &= P_1 \uparrow + w_{21} * \hat{I}_{21} + w_{23} * \hat{I}_{23} \\ \hat{I}_b &= P_2 \uparrow + w_3 * \hat{I}_3 \end{aligned} \tag{5}$$

where $\hat{I}_*$ represents the bokeh image generated from $\mathbf{Gen}_*$. "$\uparrow$" represents upsampling operation. P_* is intermediate map.

3.5. *Training loss*

For comprehensively parameters learning, training losses both on global and component levels are considered.

The first loss $Loss_B$ is globally implemented on output bokeh image. L_1 loss is employed, as shown in Eq. (6). The L_1 loss benefits pixel-wise reconstruction of synthesized bokeh image.

$$Loss_B = |\hat{I}_b - GT|_1 \tag{6}$$

The second loss $Loss_{pyr}$ is locally implemented on pyramid image components. A linear combination of L_1 Loss and SSIM Loss is employed, as shown in Eq. (7). The SSIM loss improves perceptual quality of the generated image components, since it focuses on the similarity of local structures.

$$\begin{aligned} &L_* = |\hat{I}_* - \hat{G}_*|_1 + 0.1 * (1 - \text{SSIM}(\hat{I}_*, \hat{G}_*)), \\ &Loss_{pyr} = w_3 * L_3 + w_{23} * L_{23} + w_{21} * L_{21} + w_1 * L_1 \end{aligned} \tag{7}$$

where, L_* represents loss on respective pyramid component. w_* are the normalized weights the same as the ones defined in Eq. (5)

Therefore, the proposed network is trained with a total loss defined in Eq. (8).

$$Loss = Loss_B + Loss_{pyr} \tag{8}$$

4. Experiment

In this section, we comprehensively describe the experiment settings and results. The proposed network is implemented in PyTorch, and trained on workstation with NVIDIA GeForce RTX 3090 GPU. Adam [41] is employed as optimizer, with initial learning rate set 0.0001. The batch size is set to be 2.

Peak Signal-to-Noise Ratio (PSNR), Structural Similarity (SSIM) [42] and Learned Perceptual Image Patch Similarity metrics (LPIPS) [15] are employed as metrics for performance evaluation. The PSNR and SSIM emphasize objective evaluation on image's pixel quality. The LPIPS focuses more on the perceptual judgments of image quality.

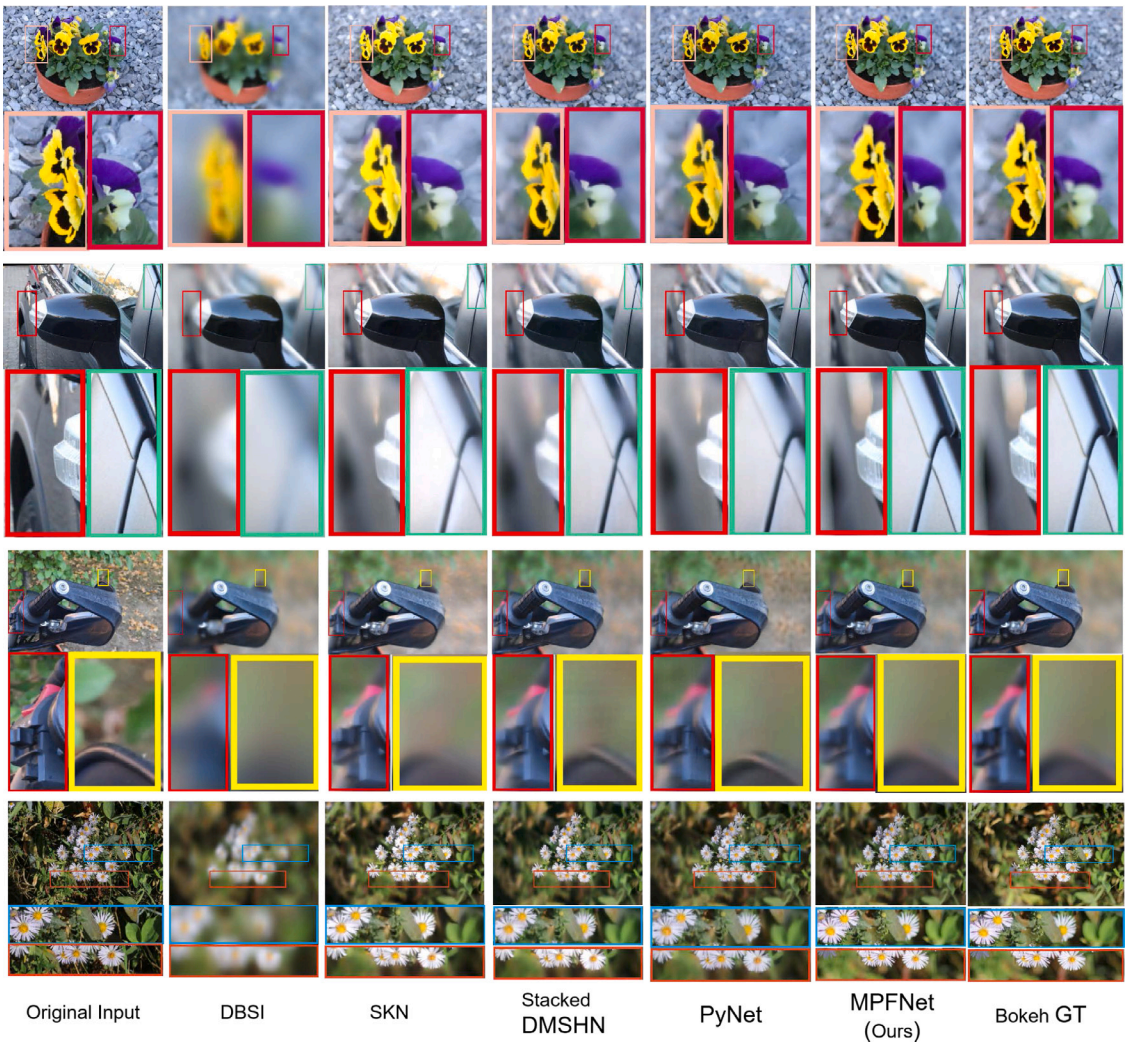


Fig. 8. Finer visual comparisons among the proposed network and some representative state-of-the-art methods. The proposed method achieves clear in-focus objects and structure-preserved out-of-focus backgrounds, as the regions compared in bounding boxes.

4.1. Dataset

"Everything is Better with Bokeh!" (EBB!) dataset [5] is a large-scale dataset that is specially for bokeh effect learning. It contains 5094 pairs of Bokeh-free and Bokeh images which were collected in the wild with the Canon 7D DSLR camera by controlling the aperture size of the lens. In each pair, the normal sharp image was captured with a narrow aperture (f/16), corresponding bokeh image was shot using high aperture (f/1.8). All the captured image pairs are aligned, cropped and downscaled to a final height equal to 1024 pixels. Therefore, the average image resolution is 1024×1536.

In the EBB! Dataset, the available training set consists of 4694 image pairs. Similar to work [17], during experiments, it is divided into two parts, in which 294 pairs are taken for evaluation and the rest 4400 pairs are for training.

4.2. Experiment results and analysis

4.2.1. Ablation study I: The selections of kernel size for DefocusBlur

In order to verify the selection of kernel sizes for *DefocusBlur*, an ablation study is conducted. Since the size selection is a combination problem, we greedily conduct the study as followings. We respectively perform disk blurs on original clear images with different combinations of kernel sizes. SSIM and LPIPS are evaluated between the generated images and bokeh ground-truths. The similarity qualities are demonstrated in Table 3. From the ablation experiments, the combination of {3,5,7} achieves the best results. Therefore, in consideration of performance and computation burden, we select it throughout the experiments in this paper.

Table 3
The similarity qualities of the generated images with different kernel size combinations.

Kernel combination	SSIM↑	LPIPS↓
3, 5, 7	**0.8806**	**0.2255**
3, 5, 9	0.8784	0.2376
3, 5, 11	0.8756	0.2468
3, 5, 15	0.8672	0.2638
3, 7, 9	0.8712	0.2542
3, 7, 11	0.8694	0.2486

4.2.2. Ablation study II: The effectiveness of training losses

To demonstrate the effectiveness of the self-supervised pyramid loss $Loss_{pyr}$, an ablation study is conducted. The experiment results are shown in Table 4. With the help the self-supervised information, the performances achieve great improvements.

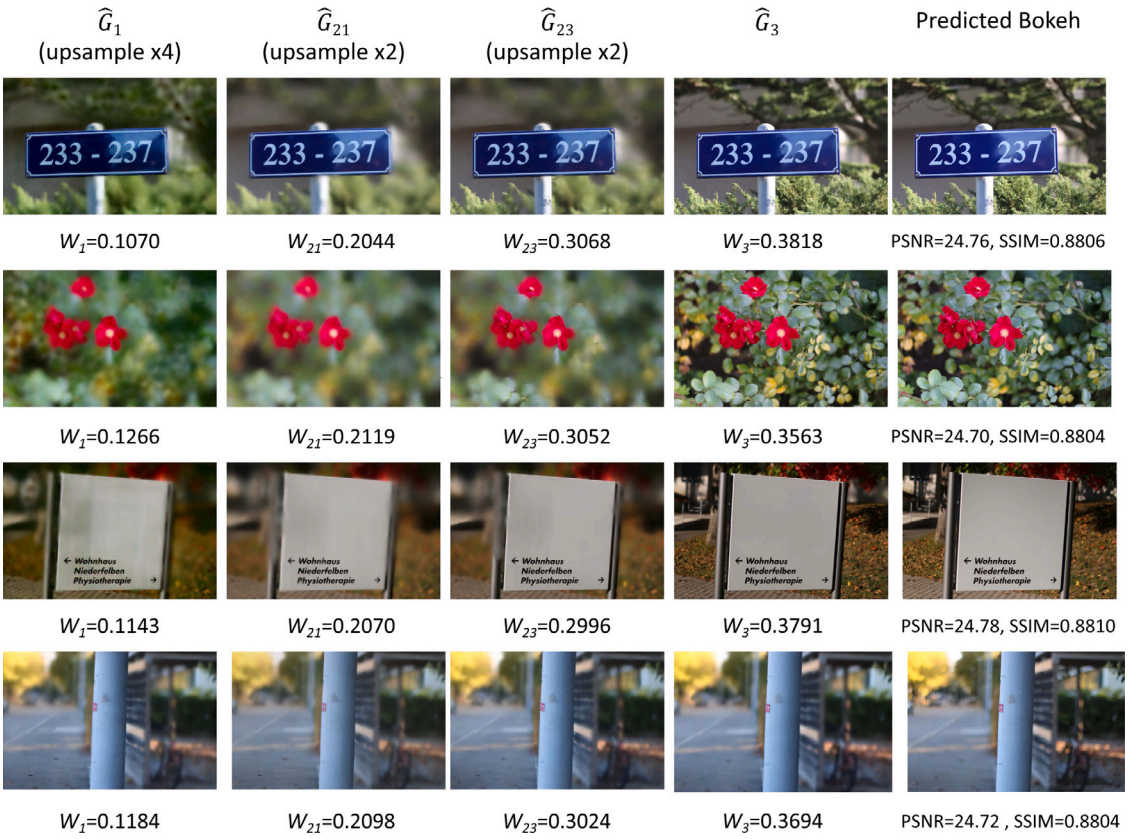


Fig. 9. Visual examples of normalized weights for each image components. Images from $\mathbf{Gen}_1$, $\mathbf{Gen}_{2,1}$, and $\mathbf{Gen}_{2,3}$ are upsampled to the same resolution with $\mathbf{Gen}_3$ for observation convenience.

Table 4
The ablation study on the effectiveness of training losses.

	$Loss_B$	$Loss_B$+$Loss_{pry}$
PSNR↑	24.32	**24.74**
SSIM↑	0.8524	**0.8806**
LPIPS↓	0.2415	**0.2255**

Table 5
The ablation study on pyramid structure and self-supervision training.

	PSNR↑	SSIM↑	LPIPS↓
$\text{MPFNet}_{w/oSS}$	24.18	0.8746	0.2458
MPFNet	24.74	0.8806	0.2255

4.2.3. Ablation study III: Multi-scale pyramid structure vs. Self-supervised augmentation

In order to highlight both the effectiveness of the proposed multi-scale pyramid structure and the promoted self-supervised training augmentation, an ablation study is further conducted. All *DefocusBlur* branches are removed from the training architecture shown in Fig. 3. All the importance weights w_* in Eq. (5) are equally set to be 1. The resulted network is therefore denoted as "$\text{MPFNet}_{w/oSS}$". The experiment results are shown in Table 5.

4.2.4. The comparisons with state-of-the-art methods

The proposed network is compared with several representative state-of-the-art methods. They are Selective Kernel networks (SKN) [1], Stacked Deep Multi-Scale Hierarchical Network(Stacked DMSHN) [7], Depth-guided Dense Dynamic Filtering network(DDDF) [16], Bokeh-Glass Generative Adversarial Network(BGGAN) [6], PyNet [5], Depth-aware Blending of Smoothed Images (DBSI) [17].

Table 6
The comparisons with state-of-the-art methods on "EBB!" Bokeh Dataset. ↑ means the higher the value, the better the performance. ↓ means the smaller the value, the better the performance.

Method	PSNR↑	SSIM↑	LPIPS↓	MOS↑
SKN [1]	24.66	0.8521	0.3323	4.1
Stacked DMSHN [7]	24.72	**0.8793**	0.2271	**4.3**
DDDF [16]	24.14	0.8713	0.2482	3.4
BGGAN [6]	24.39	0.8645	0.2467	3.8
PyNet [5]	**24.93**	0.8788	**0.2219**	4.2
DBSI [17]	23.45	0.8657	0.2463	3.5
MPFNet (ours)	**24.74**	**0.8806**	**0.2255**	**4.5**

Since the evaluation of bokeh effects is subjective, besides the objective metrics like PSNR, SSIM and LPIPS, a user study with MOS (Mean Opinion Scores) metric [2] is conducted to rank perceptual qualities of the predicted images. Specifically, we recruited 30 people with certain photographic knowledge to participate in the evaluation. Participants were asked to rate the image quality by selecting a score of 1–5 levels (5 - comparable perceptual quality, 4 - slightly worse, 3 - notably worse, 2 - poor perceptual quality, 1 - completely corrupted image) in comparison with the original Canon images exhibiting bokeh effect. The expressed preferences are then averaged per each test image and then per each method to obtain the final MOS.

The comparison results on EBB! dataset are shown in Table 6. The comparisons with some SOTA methods on network's parameters size and efficiency are shown in Table 7 based on available open source codes.

From the experiment results in Tables 6 and 7, it is not difficult to observe that the proposed network can obtain more superior performances with much less parameter size and running time. Therefore, the proposed network achieves the best when considering the overall processing efficiency and effectiveness.

epoch	10	50	100	150	200
W_1	0.2797	0.2512	0.2134	0.1936	0.1143
W_{21}	0.2000	0.1898	0.1924	0.1969	0.2070
W_{23}	0.2497	0.2536	0.2637	0.2663	0.2996
W_3	0.2706	0.3054	0.3305	0.3431	0.3791
PSNR	23.53	23.98	25.67	26.4	24.78
SSIM	0.8602	0.8714	0.8762	0.8796	0.8810
Predicted Bokeh Result					

epoch	10	50	100	150	200
W1	0.2714	0.2404	0.2063	0.1768	0.1266
W12	0.1805	0.1837	0.1916	0.2009	0.2119
W23	0.2556	0.2743	0.2766	0.2764	0.3052
W3	0.2925	0.3016	0.3255	0.3459	0.3563
PSNR	23.62	24.01	24.48	24.66	24.70
SSIM	0.8643	0.8748	0.8788	0.8802	0.8804
Predicted Bokeh Result					

epoch	10	50	100	150	200
W1	0.2623	0.2350	0.1431	0.1139	0.1070
W12	0.1865	0.1797	0.2105	0.2158	0.2044
W23	0.2620	0.2507	0.2895	0.3060	0.3068
W3	0.2891	0.3347	0.3569	0.3643	0.3818
PSNR	23.76	24.21	24.66	24.70	24.76
SSIM	0.8675	0.8732	0.8785	0.8798	0.8806
Predicted Bokeh Result					

Fig. 10. The weights, intermediately predicted bokeh results on sampling training epochs.

Table 7
The comparison of network's parameters size (in millions) and running time (in second) for processing single image in 1024 × 1536 resolution.

Method	Parameters (M)	Running time (s)
SKN [1]	5.37	0.055
Stacked DMSHN [7]	10.84	**0.040**
DDDF [16]	N/A	2.5
PyNet [5]	47.5	0.27
DBSI [17]	**5.36**	0.048
MPFNet (ours)	**6.12**	**0.046**

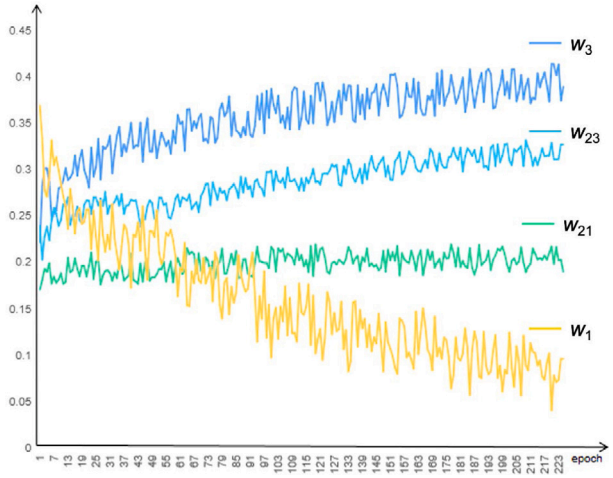


Fig. 11. The training curves of normalized weights.

It should be noted that, the MOS score of the BGGAN we obtained deviates a bit from its behavior performance in AIM2020 [2]. The reasons can be explained in two aspects. The first one owes to different evaluation dataset used, since herein the experiments were performed on val294 subset, not on the test set. The second one is that there exist a certain proportion of images that BGGAN generated with severe artifacts in the val294 dataset, especially in case of large blurs, whose low scores directly lowered the overall average.

For better understanding the advantages of the proposed network on bokeh rendering, samples of visual comparisons are demonstrated in Fig. 7. More finer visual comparisons are illustrated in Fig. 8. From the visual comparisons, we can observe that the proposed network can achieve more visually realistic bokeh effect, with clear in-focus objects and structure-preserved out-of-focus backgrounds.

For better illustrating the contribution of each pyramid branch during bokeh components learning, some of visual examples and corresponding generated pyramid components together with their respective normalized weights are shown in Fig. 9. We can observe each pyramid component contributes final bokeh image in different extent. Moreover, we can also easily find that the weights w_* on respective components stably converge. Even in case of different examples, the weights have similar importance distributions on each components. The details of training curves and intermediately predicted bokeh results for a specific example are shown in Figs. 11 and 10.

5. Conclusion

In this paper, an effective multi-scale pyramid fusion network has been proposed for realistic bokeh effect rendering. Structure consistencies are employed as importance weights for pyramid information fusion. Task-specific knowledge which mimics the "circle of confusion" phenomenon through disk blur convolutions is utilized as self-supervised information for network training. The proposed network has been experimented on a public large-scale bokeh dataset. Compared with state-of-the-art methods, it can achieve more satisfied superior performance with less parameters and with realtime processing speed.

CRediT authorship contribution statement

Zhifeng Wang: Design of this study, Analysis and interpretation of data, Implementation of this methodology and experiments, Preparation the manuscript. **Aiwen Jiang:** Conceptualization and design of this study, Analysis and interpretation of data, Provision of study materials and computing resources, Writing – original draft, Writing – review & editing. **Chunjie Zhang:** Conceptualization, Writing – review & editing. **Hanxi Li:** Conceptualization, Writing – review & editing. **Bo Liu:** Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is supported by National Natural Science Foundation of China under Grand No. 61966018, The open research fund of the State Key Laboratory for Management and Control of Complex Systems under Grant NO. 20220103, and Beijing Natural Science Foundation under Grand No. JQ20022.

References

[1] Andrey Ignatov, Jagruti Patel, Radu Timofte, Bolun Zheng, Xin Ye, Li Huang, Xiang Tian, Saikat Dutta, Kuldeep Purohit, Praveen Kandula, Maitreya Suin, A.N. Rajagopalan, Zhiwei Xiong, Jie Huang, Guanting Dong, Mingde Yao, Dong Liu, Ming Hong, Wenying Lin, Yanyun Qu, Jae-Seok Choi, Woonsung Park, Munchurl Kim, Rui Liu, Xiangyu Mao, Chengxi Yang, Qiong Yan, Wenxiu Sun, Junkai Fang, Meimei Shang, Fei Gao, Sujoy Ghosh, Prasen Kumar Sharma, Arijit Sur, Wenjin Yang, Aim 2019 challenge on bokeh effect synthesis: Methods and results, in: IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), 2019, pp. 3591–3598.

[2] Andrey Ignatov, Radu Timofte, Ming Qian, Congyu Qiao, Jiamin Lin, Zhenyu Guo, Chenghua Li, Cong Leng, Jian Cheng, Juewen Peng, Xianrui Luo, Ke Xian, Zijin Wu, Zhiguo Cao, Densen Puthussery, C V Jiji, P S Hrishikesh, Melvin Kuriakose, Saikat Dutta, Sourya Dipta Das, Nisarg A. Shah, Kuldeep Purohit, Praveen Kandula, Maitreya Suin, Rajagopalan A. N., Saagara M B, Minnu A L, Sanjana A R, Praseeda S, Ge Wu, Xueqin Chen, Tengyao Wang, Max Zheng, Hulk Wong, Jay Zou, Aim 2020 challenge on rendering realistic bokeh, in: Advances in Image Manipulation Workshop and Challenges on Image and Video Manipulation, in Conjunct with European Conference on Computer Vision, 2020.

[3] Wadhwa Neal, Garg Rahul, E. Jacobs David, E. Feldman Bryan, Kanazawa Nori, Carroll Robert, Movshovitz-Attias Yair, T. Barron Jonathan, Pritch Yael, Levoy Marc, Synthetic depth-of-field with a single-camera mobile phone, ACM Trans. Graph. 37 (4) (2018) 1–13.

[4] Zhengqi Li, Noah Snavely, Megadepth: Learning single-view depth prediction from internet photos, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 2041–2050.

[5] Andrey Ignatov, Jagruti Patel, Radu Timofte, Rendering natural camera bokeh effect with deep learning, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2020, pp. 1676–1686.

[6] Ming Qian, Congyu Qiao, Jiamin Lin, Zhenyu Guo, Chenghua Li, Cong Leng, Jian Cheng, Bggan: Bokeh-glass generative adversarial network for rendering realistic bokeh, in: European Conference on Computer Vision, 2020, pp. 229–244.

[7] Saikat Dutta, Sourya Dipta Das, Nisarg A. Shah, Anil Kumar Tiwari, Stacked deep multi-scale hierarchical network for fast bokeh effect rendering from a single image, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2021, pp. 2398–2407.

[8] Robert Kosara, Silvia Miksch, Semantic depth of field, in: Proceedings of the IEEE Symposium on Information Visualization, 2001, pp. 97–104.

[9] E. Bigler, Depth of field and scheimpflug's rule : a minimalist geometrical approach, 2002, https://Galerie-Photo.Com/Profondeur-de-Champ-Scheimpflug-English.Html.

[10] Benjamin Busam, Matthieu Hog, Steven McDonagh, Gregory Slabaugh, Sterefo: Efficient image refocusing with stereo vision, in: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), 2019, pp. 3295–3304.

[11] Chenchi Luo, Yingmao Li, Kaimo Lin, George Chen, Seok-Jun Lee, Jihwan Choi, Youngjun Francis Yoo, Michael O. Polley, Wavelet synthesis net for disparity estimation to synthesize dslr calibre bokeh effect on smartphones, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2404–2412.

[12] Liu Dongwei, Nicolescu Radu, Klette Reinhard, Stereo-based bokeh effects for photography, Mach. Vis. Appl. 27 (8) (2016) 1325–1337.

[13] Yuna Jeong, Seung Youp Baek, Yechan Seok, Gi Beom Lee, Sungkil Lee, Real-time dynamic bokeh rendering with efficient look-up table sampling, IEEE Trans. Vis. Comput. Graphics (2020) 1, http://dx.doi.org/10.1109/TVCG.2020.3014474.

[14] Xiangyu Xu, Deqing Sun, Sifei Liu, Wenqi Ren, Yu-Jin Zhang, Ming-Hsuan Yang, Jian Sun, Rendering portraitures from monocular camera and beyond, in: European Conference on Computer Vision, 2018, pp. 36–51.

[15] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, Oliver Wang, The unreasonable effectiveness of deep features as a perceptual metric, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 586–595.

[16] Kuldeep Purohit, Maitreya Suin, Praveen Kandula, Rajagopalan Ambasamudram, Depth-guided dense dynamic filtering network for bokeh effect rendering, in: IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), 2019, pp. 3417–3426.

[17] Saikat Dutta, Depth-aware blending of smoothed images for bokeh effect generation, J. Vis. Commun. Image Represent. 77 (2021) 103089.

[18] Wang Yingqian, Yang Jungang, Guo Yulan, Xiao Chao, An Wei, Selective light field refocusing for camera arrays using bokeh rendering and superresolution, IEEE Signal Process. Lett. 26 (1) (2019) 204–208.

[19] Tiemi Mizuno Nakamura Angelica, Grassi Valdir, Fernando Wolf Denis, An effective combination of loss gradients for multi-task learning applied on instance segmentation and depth estimation, Eng. Appl. Artif. Intell. 100 (2021) 104205.

[20] Clement Godard, Oisin Mac Aodha, Michael Firman, Gabriel Brostow, Digging into self-supervised monocular depth estimation, in: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3827–3837.

[21] Zuo Yifan, Fang Yuming, Yang Yong, Shang Xiwu, Wang Bin, Residual dense network for intensity-guided depth map enhancement, Inform. Sci. 495 (2019) 52–64.

[22] Wenfeng Song, Shuai Li, Ji Liu, Aimin Hao, Qinping Zhao, Hong Qin, Contextualized CNN for scene-aware depth estimation from single RGB image, IEEE Trans. Multimed. 22 (5) (2020) 1220–1233.

[23] Orest Kupyn, Tetiana Martyniuk, Junru Wu, Zhangyang Wang, DeblurGAN-v2: Deblurring (orders-of-magnitude) faster and better, in: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 8877–8886.

[24] Li Jin, Liu Yanyan, Liu Zilong, Dynamic imaging inversion with double deep learning networks for cameras, Inform. Sci. 536 (2020) 317–331.

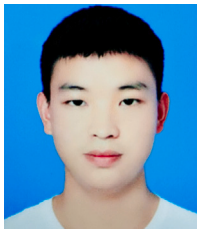
[25] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, Kyoung Mu Lee, Enhanced deep residual networks for single image super-resolution, in: IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2017, pp. 1132–1140.

[26] Zhu Xiaobin, Li Zhuangzi, Li Xianbo, Li Shanshan, Dai Feng, Attention-aware perceptual enhancement nets for low-resolution image classification, Inform. Sci. 515 (2020) 233–247.

[27] Leon A. Gatys, Alexander S. Ecker, Matthias Bethge, Image style transfer using convolutional neural networks, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2414–2423.

[28] Fu Xianping, Yan Yuxiao, Yan Yang, Peng Jinjia, Wang Huibing, Purifying real images with an attention-guided style transfer network for gaze estimation, Eng. Appl. Artif. Intell. 91 (2020) 103609.

[29] Xu Yadong, Yang Cheng, Sun Beibei, Yan Xiaoan, Chen Minglong, A novel multi-scale fusion framework for detail-preserving low-light image enhancement, Inform. Sci. 548 (2021) 378–397.
[30] Wang Wencheng, Chen Zhenxue, Yuan Xiaohui, Wu Xiaojin, Adaptive image enhancement method for correcting low-illumination images, Inform. Sci. 496 (2019) 25–41.
[31] Chen Honggang, He Xiaohai, An Cheolhong, Q. Nguyen Truong, Adaptive image coding efficiency enhancement using deep convolutional neural networks, Inform. Sci. 524 (2020) 298–317.
[32] Zhao Jingming, Zhang Juan, Li Zhi, Hwang Jenq-Neng, Gao Yongbin, Fang Zhijun, Jiang Xiaoyan, Huang Bo, Dd-cyclegan: Unpaired image dehazing via double-discriminator cycle-consistent generative adversarial network, Eng. Appl. Artif. Intell. 82 (2019) 263–271.
[33] Zhang Tianlun, Yang Xi, Wang Xizhao, Wang Ran, Deep joint neural model for single image haze removal and color correction, Inform. Sci. 541 (2020) 16–35.
[34] Xiang Li, Wenhai Wang, Xiaolin Hu, Jian Yang, Selective kernel networks, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 510–519.
[35] Pengju Liu, Hongzhi Zhang, Kai Zhang, Liang Lin, Wangmeng Zuo, Multi-level wavelet-CNN for image restoration, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2018, pp. 886–88609.
[36] Luo Xianrui, Peng Juewen, Ke Xian, Wu Zijin, Cao Zhi-Guo, Bokeh rendering from defocus estimation, in: European Conference on Computer Vision Workshops, 2020, pp. 245–261.
[37] Ji Yuzhu, Zhang Haijun, Zhang Zhao, Liu Ming, Cnn-based encoder-decoder networks for salient object detection: A comprehensive review and recent advances, Inform. Sci. 546 (2021) 835–857.
[38] Noori Mehrdad, Mohammadi Sina, Ghofrani Majelan Sina, Bahri Ali, Havaei Mohammad, Dfnet: Discriminative feature extraction and integration network for salient object detection, Eng. Appl. Artif. Intell. 89 (2020) 103419.
[39] Liu Ze-yu, Liu Jian-wei, Zuo Xin, Hu Ming-fei, Multi-scale iterative refinement network for RGB-d salient object detection, Eng. Appl. Artif. Intell. 106 (2021) 104473.
[40] Xiaoyong Shen, Aaron Hertzmann, Jiaya Jia, Sylvain Paris, Brian Price, Eli Shechtman, Ian Sachs, Automatic portrait segmentation for image stylization, Comput. Graph. Forum 35 (2) (2016) 93–102.
[41] P. Kingma Diederik, Ba Jimmy, Adam: A method for stochastic optimization, in: International Conference on Learning Representations, 2015, pp. 1–13.
[42] Wang Zhou, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Trans. Image Process. 13 (4) (2004) 600–612.



Zhifeng Wang is current a post-graduate student in School of Computer and Information Engineering, Jiangxi Normal University. His research interest is on bokeh rendering.

Aiwen Jiang is full professor and associate director in School of Computer and Information Engineering, Jiangxi Normal University. He received his Ph.D. degree from Institute of Automation, Chinese Academy of Sciences, China, in 2010. He received his bachelor degree from Nanjing University of Post and Telecommunication, China, in 2005. His research interests are computer vision, machine learning.

Chunjie Zhang is full professor in School of Computer and Information Technology, Beijing Jiaotong University. He received his Ph.D. degree from Institute of Automation, Chinese Academy of Sciences, China, in 2011. He serves as associate editors for many famous journals such as Information Sciences, Neurocomputing etc. His research interests focus on computer vision, machine learning, and multimedia information analysis.

Hanxi Li is associate professor in School of Computer and Information Engineering, Jiangxi Normal University. He received his Ph.D. degree from Australia National University in 2011. He received his bachelor degree from Beihang University, China, in 2004. His research interests are computer vision, virtual reality.

Bo Liu is a tenured associate professor in the Department of Computer Science at Auburn University. He obtained his Ph.D. from Autonomous Learning Lab at the University of Massachusetts Amherst, 2016. His research areas cover decisionmaking under uncertainty, human-aided machine learning, symbolic AI, trustworthiness and interpretability in machine learning, and their applications to BIGDATA, autonomous driving, and healthcare informatics.