This WACV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Robust Portrait Image Matting and Depth-of-field Synthesis via Multiplane Images

Zhefan Rao*, Tianjia Zhang*, Yuen Fui Lau, Qifeng Chen
Hong Kong University of Science and Technology
{zraoac, tzhangbl, yflauad}@connect.ust.hk, cqf@ust.hk

Abstract

High-quality portrait photography has become an essential function in our daily lives. However, due to the limited aperture and focal length of a smartphone camera, images captured by a smartphone cannot match the same level of bokeh effect by a digital single-lens reflex camera. A typical solution on a smartphone is to simulate out-of-focus effects from an all-in-focus image, where the key is robust depth estimation and portrait matting. To achieve this, we propose a multi-stage, multi-branch matting network to estimate a strand-level portrait alpha mask, which is then used to refine the coarse depth map obtained from the pre-trained model. Combining the input portrait image with the estimated depth map and alpha mask, we propose a learning-free optimization mechanism to construct a multi-plane image (MPI) representation for depth-of-field synthesis. The MPI consists of multiple layers of disk-blurred images with kernel size proportional to the absolute depth distance to the focus layer. Then a depth-aware blurring process is applied to enforce the bokeh effect. Besides, each MPI layer has an alpha channel controlling the visibility according to the corresponding depth. Finally, an image with bokeh is rendered by compositing all MPI layers. We conduct comprehensive experiments to evaluate our method, which demonstrates that our method can generate more accurate alpha masks and more realistic images with bokeh compared to prior work.

1. Introduction

The popularity of mobile phones has made them replace the role of traditional digital SLR cameras in most cases [65]. However, due to the limitations of lens aperture and sensor size, the quality of images captured by mobile phones still needs to be improved from that of digital single-lens reflex cameras. The most significant difference is the missing depth-of-field, which describes the bokeh effects that would blur the defocused area in the pictures. The issue arises when light rays passing through the lens fail to converge precisely on the focal point, leading to blurred areas known as circles of confusion (COC). Many smartphones provide portrait mode, which uses software algorithms and advanced camera technology to take photos with a blurred background and a sharp focus on the subject. The function is usually achieved by combining a clear matte foreground and a blurred environment background. However, canonical blurring filters could not simply implement the bokeh effects since the depth information is not considered. Moreover, the foreground subject is likely to be mixed with the background due to incorrect matte estimation.

To tackle these challenges, researchers have been working on modeling and rendering techniques and achieve high image quality comparable to those captured by DSLRs [1,2,52,54]. According to the imaging principle [16,40,52] of the camera lens, the diameter of the circle of confusion of each light ray is determined by the distance between the object and the focus plane. The principle results in the effect that the depth-of-field blur degree is proportional to the distance from the focus plane. Thus, it is essential to deduce spatial distance information, which is usually embedded in the depth map and could be either directly obtained from the range sensors or computed from stereo cameras. However, adding extra depth sensors to mobile phones is impractical due to due to manufacturing costs and space limitations. Recently, transformer architectures in large models have significantly improved dense image prediction, especially single-image depth estimation [43, 44]. Nevertheless, relying solely on a depth map as a flawless scene representation for image synthesis and rendering tasks is insufficient due to in-continuity and misalignment. It is preferable to establish a more integrated, compact representation [28, 34, 36], enabling better simulation of the imaging process and photo-realistic image rendering. The multi-plane image(MPI) representation [67] has shown great power on few-shot high-quality synthesis. This feature makes depth-of-field rendering more practical since we have only a single image as input. By employing the multi-layer characteristic of MPI,

*These authors contributed equally to this work

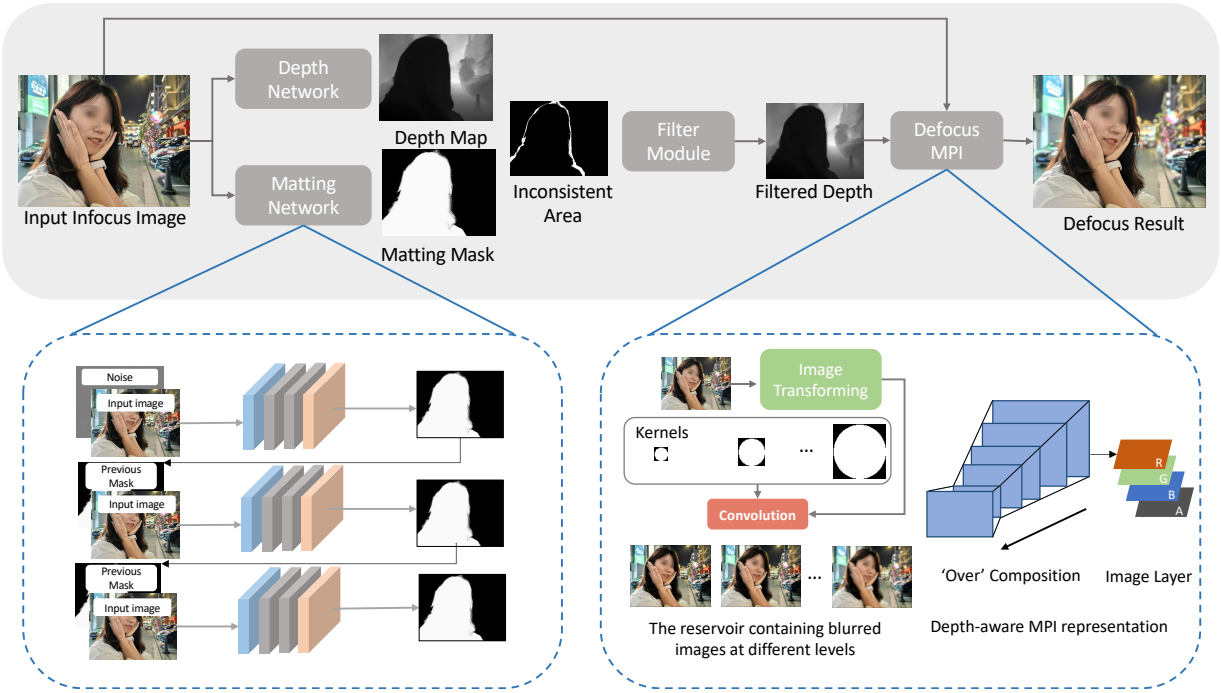


Figure 1. **The framework of the proposed method.** Our study introduces an approach for generating a multi-plane image (MPI) representation of a portrait image from an initial all-in-focus image. A pre-trained transformer-based depth estimation network and a multi-stage matting network are used to obtain a depth map and a foreground mask, respectively. A non-linear filter in the transition area based on the guided filter is applied to align the depth map with the mask. A zero-shot transferring mechanism converts the filtered depth map into an MPI representation consisting of multiple RGBA image layers. The final defocused image is rendered from back to front using the over operator in alpha composition.

most visual artifacts can be removed compared to the classical depth-of-field blur.

For generating blur effects, classical methods [52] view that the depth-of-field blur can be generated by convolving the all-in-focus image with a disk kernel. However, it suffers from the inconsistent fusion of different blurred images. Neural rendering methods [17,37,41,60] can generate more natural blur by learning from image statistics. These methods train neural networks to mimic the rendering process by predicting depth, synthesizing lens blur, and fusing images of multiple forms. However, it usually lacks an accurate mechanism for adjusting the focus distance, hence changing the blur degree according to user interaction after training. Our modified MPI representation could achieve consistent rendering quality comparable to network-based counterparts while retaining interactivity.

Another significant factor in enhancing rendering quality is the separation of the portrait subject from the background, especially in the transition area of tiny edges, such as hair strands and clothes furs. The challenge is commonly addressed by estimating an accurate alpha matte, a gray-scale image that indicates which pixels belong to the foreground or background. Matting-based methods have been widely explored, and some could be integrated into smartphone camera applications [52]. In recent years, as learning-based has dominated vision tasks, many advances in neural image matting have encouraged more detailed segmentation of portrait areas. The significant improvement mechanisms include multi-branch structures to perform global subject location and local detail refinement, adding auxiliary input as guidance, and using attention modules to integrate features at different levels.

Our framework works similarly to [52], yet with a more fine-detailed and realistic quality to reduce the effects of incorrect estimation of depth and alpha mattes. Given an arbitrary all-in-focus portrait image, we apply a pre-trained transformer-based depth estimation network [43] to obtain the depth map. A multi-stage matting network module is introduced to predict an accurate foreground mask by progressively refining the previous resulting mask at each step. We propose a non-linear filter based on the guided filter [14] using the predicted mask as guidance to align the depth map with the mask. Then, a zero-shot transferring mechanism is employed to turn the filtered depth map and the exponentially transformed input image into an MPI scene representation named depth-of-field MPI. The final result image with bokeh effects is rendered using the over operation from back to front. The overall framework is illustrated in Figure 1.

Our contributions to the paper are summarized as fol-

Figure 2. **The illustration of the proposed framework.** The depth and mask are combined to show the inconsistency in the boundary area. After filtering, the gap is minimized. The most blurred layer is obtained by convolving the input image with the largest kernel, which could show how the bokeh style is generated. The final rendered result images present depth-of-field blur effects.

lows:

- We propose a new framework based on the MPI scene representation to synthesize depth-of-field effects for portrait images captured by mobile phones. Unlike previous practices, which usually learn an MPI through a neural network, we adopt a learning-free mechanism to optimize an MPI, which requires fewer data or even no data and presents higher flexibility to adjust parameters.
- We propose a multi-stage matting network that predicts the alpha matte in a coarse-to-fine manner.
- We propose processing strategies, such as mask-based depth filtering and exponential image transforming, to improve the realistic quality of rendered images.

2. Related Work

2.1. Depth-of-Field Synthesis

With the widespread use of smartphones and their increasing imaging quality, people are paying more attention to employing cameras to capture portrait images casually. However, there is still a gap between mobile cameras and DSLRs. A key difference is the lack of depth-of-field (DOF) control, which often occurs when captured with a large-aperture (shallow depth-of-field) lens and makes objects not located in the focal point. This would result in circles of confusion in the out-of-focus area, which is also called Bokeh. The Bokeh effect is helpful for photographers to highlight the object of interest and make the image more artistic. To minimize the imaging gap, researchers and manufacturers have proposed various methods to process the captured images and synthesize the DOF effects.

Classical rendering methods adopt a controllable blur kernel to design the bokeh pattern. The shape and size of the kernel could represent the lens configuration. This kind of method requires the 3D geometry information of the scene, which is usually obtained in object or image space. For object space [21, 56, 63], provided with the complete 3D scene information, the rendering process is achieved by ray-tracing, and then exact results could be obtained. However, this method is not practical for real-world scenes since the scene geometry is difficult to obtain and time-consuming to render. Image-space methods [2, 3, 12, 50, 61] are easier to implement since only a single input image and corresponding depth map are needed. To improve the quality of rendering, some methods [38, 48, 49, 57] integrate multiple modules, such as depth estimation, semantic segmentation, and classical rendering, to construct an automatic rendering system. Recently, most methods [52, 65] decompose the process of producing the final images into multi-layers based on an estimated depth map, then perform the rendering process from back-to-front. However, they often suffer from artifacts due to depth discontinuities.

With the fast development of deep learning, neural rendering methods that learn from statistics to produce realistic bokeh balls have been developed. To improve efficiency and avoid boundary artifacts, neural networks are employed to simulate the rendering process. For example, Deep Shading [35] and DeepFocus [58] train networks to produce a bokeh effect from an all-in-focus image and its corresponding perfect depth map. Other researchers have proposed automatic rendering systems that use depth prediction [8], lens blur, adversarial training [19], multi-image fusion [55] and guided upsampling [11] to generate high-resolution depth-of-field images into shallow DoF images in an end-to-end manner. Moreover, some methods [9, 11] stack multiple network architectures to improve rendering quality by learning different functions separately. However, these methods are limited to learning a large-scale blur and control rendering results to fit different scenes. Nevertheless, neural networks

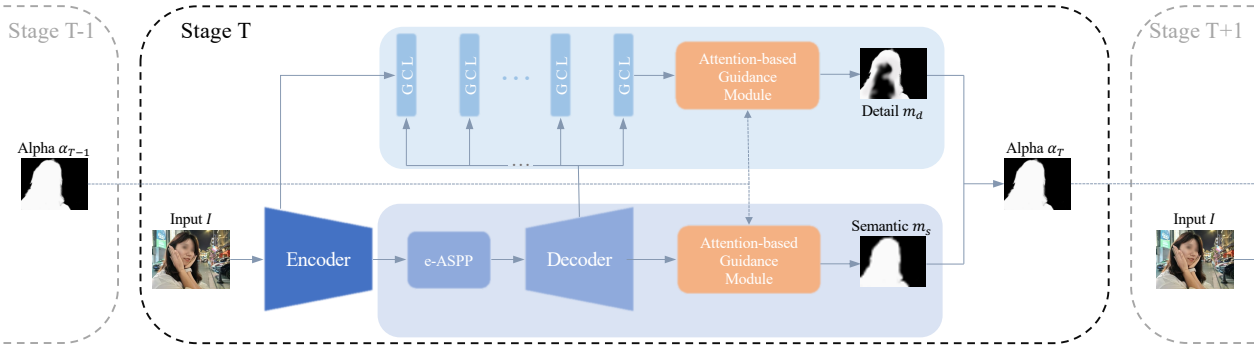


Figure 3. **The pipeline of the matting network.** Given an input image I and a previous alpha map α_{T-1}, the model can predict a detail map m_d and a semantic map m_s with the help of GCL [51], e-ASPP [18], and Attention-based Guidance Module(AGM). For the first stage, the previous alpha map can be set as the constant value.

have shown great power on refocus tasks, which include an extra step to deblur an already defocused image and then perform a similar rendering process to change the focused object. For instance, RefocusGan [45] trains a two-stage GAN to perform refocusing. Recent method [37] integrates the advantages of both classical and neural rendering methods to produce artifact-free and highly controllable bokeh effects. Furthermore, there are more works [7,20,30–33,46] utilize the diffusion model to achieve high performance on generating DOF effect.

2.2. Image Alpha Matting

Over the years, several approaches have been proposed to solve the alpha matting problem. Sampling-based methods and propagation-based methods were the main approaches used in the early days of alpha matting research. Sampling-based methods [10, 13, 53] involve randomly selecting representative samples from the foreground and background regions to estimate the alpha matte of the unknown region. For example, Global Matting [13] utilized all samples available in the image to predict the alpha matte by randomized patch match. Propagation-based methods [5, 22] estimate alpha mattes by propagating the alpha values from the known regions (foreground and background) to the unknown regions based on their similarity. These methods have been effective in producing alpha mattes, but they have limitations in terms of accuracy and computational efficiency.

Recently, the use of deep learning techniques has significantly advanced the state-of-the-art in alpha matting [18, 23–25, 29, 42, 66]. Learning-based methods have become the dominant approach in alpha matting research. These methods utilize deep neural networks to learn the mapping from input images to alpha mattes. The DCNN matting [6] was the first method to introduce a deep neural network into matting, and Deep Matting [59] is a fully neural network model with a large-scale dataset that achieved remarkable results. Besides using a single input image, auxiliary matting methods use some additional information as input to help the matting process. For example, some methods use a trimap [27], which is a user-supplied mask that divides the image into foreground, background, and unknown regions. Other methods use background images [47], coarse annotations [62], or natural language descriptions [26] to provide clues for foreground or background regions.

3. Methodology

Our method takes as input a single portrait image and aims to render the corresponding realistic image with bokeh effects. The overall framework consists of a mask estimation module, a mask-based depth filtering module, and the depth-of-field MPI for rendering. Fig. 2 illustrates the visual results of each module.

3.1. Multi-stage and Multi-branch Refinement Network for Strand-level Matting

The matting network aims to estimate an alpha matte representing the foreground's opacity at each pixel. Fig. 3 illustrates the whole pipeline of the matting network, composed of two components: a high-resolution branch and a low-resolution branch. At the stage T, given the input image $I \in \mathbb{R}^{3\times H\times W}$ and the previous alpha matte $\alpha_{T-1} \in \mathbb{R}^{1\times H\times W}$, the model could predict the alpha matte $\alpha_T \in \mathbb{R}^{1\times H\times W}$ for current stage. More specifically, we utilize the ResNet-50 [15] as our encoder to extract features from the images, $F = \{F_1, F_2, ..., F_5\}$ denotes the extracted features from the n_{th} layer separately. Then the feature F_5 extracted from the last layer would be fed into the low-resolution branch to predict semantic map m_s, and the features F_1, F_2 extracted from the middle layer would be used in the high-resolution branch to generate detail map m_d.

Low-Resolution Branch Without the help of auxiliary input, we introduce the low-resolution branch to estimate

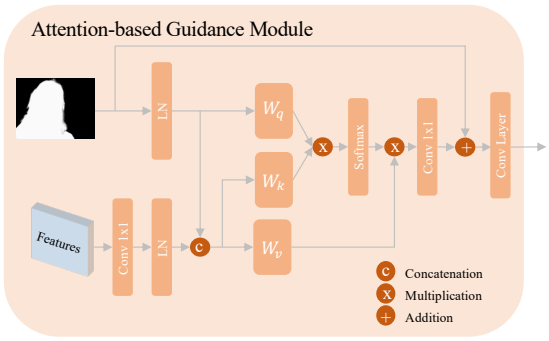


Figure 4. The illustration of Attention-based Guidance Module. One of the inputs is the previous alpha map α_{T-1}, while the other one is the features from the GCL or the decoder.

the semantic map $m_s \in \mathbb{R}^{3\times H\times W}$, which facilitates the model to learn the global context. From the MODNet [18], e-ASPP has shown its efficiency and satisfactory performance compared with ASPP [4]. $A(\cdot)$ denotes the e-ASPP operation block, then $F_A = A(F_5)$ denotes the features after the e-ASPP block and $F_A \in \mathbb{R}^{B\times C\times \frac{H}{32}\times \frac{W}{32}}$. To generate a semantic map of the same size as the input images, F_A is fed into a decoder $S(\cdot)$ consisting of several convolution and upsampling blocks and attention-based guidance modules sequentially. Finally, the output semantic map would be $m_s \in \mathbb{R}^{3\times H\times W}$, which also can be regarded as a trimap.

High-Resolution Branch With the help of the low-resolution branch, the corresponding trimap can be estimated automatically, which can help the high-resolution branch to focus the training on the transition region. The GCL [51] is adopted in the high-resolution branch to fuse the features from encoder the middle layers of the decoder, where $f_k = GCL(f_{k-1}, f'_{k-1})$. As shown in Fig. 3, each GCL block has two inputs, one of them is the previous output f_{k-1} of GCL or generated by the middle features F_1 and F_2 from encoder concatenating and upsampling, where $f_1 \in \mathbb{R}^{C_1\times H\times W}$. The other one $f'_{k-1} \in \mathbb{R}^{C'_{k-1}\times H\times W}$ is generated by the upsampling from the middle feature from the decoder. After the GCL flow, a 1×1 convolution layer is applied on the output to generate the features with the specific number of channels $f_H \in \mathbb{R}^{C'\times H\times W}$. Then the features are fed into the attention-based guidance module to predict detail map $m_d \in \mathbb{R}^{1\times H\times W}$ for the current stage.

Attention-based Guidance Module We introduce the Attention-based Guidance Module(AGM) shown in Fig. 4 to predict a more accurate semantic map and detail map under the guidance of the previous alpha matte. The previous alpha matte α_{T-1} is the query. And the features concatenated by α_{T-1} and the features f from the GCL or decoder are used as key and value. Assume the feature $f \in \mathbb{R}^{C\times H\times W}$, we apply a 1×1 convolution layer to α_{T-1} to generate corresponding $f_\alpha \in \mathbb{R}^{C\times H\times W}$ of the same number of channel as f. Then, we adopt the pixel-wise attention mechanism to help model training focus on the meaningful region with a comparable computational efficiency.

Alpha Matte Fusion The semantic map m_s consists of 3 channels which represent the probability of the foreground, transition and background pixel separately. Specifically, 0_{th}, 1_{st} and 2_{nd} channels denote the foreground pixel, transition pixel and background pixel, respectively. Through the probability, we can get the binary foreground mask m_f and the binary transition mask m_{trans}. Then, the alpha matte α_T can be fused by the following formula:

$$\alpha_T = m_f + m_{trans} \cdot m_d. \tag{1}$$

Loss We utilize three types of losses in total, which consist of semantic map loss $\mathcal{L}_s$, detail map loss $\mathcal{L}_d$, and final alpha matte loss $\mathcal{L}_{fusion}$. Therefore, by combining these three loss functions, we can get the total loss function:

$$\mathcal{L} = w_1\mathcal{L}_s + w_2\mathcal{L}_d + w_3\mathcal{L}_{fusion}, \tag{2}$$

where w_1, w_2, w_3 are set to 0.25, 0.25, and 0.5 in the experiments. The details of these three functions are illustrated in the supplementary materials.

3.2. MPI-based Bokeh Effects Rendering

The MPI representation [67] contains a set of parallel planes with evenly divided depth intervals. Each plane encodes an RGBA image with the same resolution. The depth information for each plane is explicitly embedded in alpha channels. This structure is rather valuable for rendering realistic images from 3D representations. Usually, the MPI is predicted by training neural networks on large datasets [39, 67]. Instead, we explicitly establish the MPI by simulating the lens imaging process. The so-called depth-of-field MPI has fewer parameters and does not require a large dataset to optimize.

A depth-of-field MPI for a single input portrait image is transformed from a depth map and an alpha matte mask. The depth is either predicted from learning models or obtained from real-world sensors. We project the image back to 3D spaces by employing depth information. We view the RGB channels of the depth-of-field MPI as a set of blurred images using disk kernels with different kernel sizes. The kernel size is proportional to the distance between the plane and the focus plane. That means the diameter of the disk kernel would be large for a distant plane while the image in the focus plane is not blurred. The alpha channel represents whether the actual depth of that pixel is consistent with the depth of the plane. If the consistency is met, the pixel in

that plane is visible after rendering and hence different blur effects could be integrated together to produce depth-aware bokeh effects that correspond to real-world observations.

Given a clear portrait image I captured by a small-aperture lens, e.g., a mobile phone camera, we aim to synthesize the depth-ware blur effects. Note that we could infer a depth map D and an alpha matte mask M through a depth estimation model and a matting model, respectively. However, since the depth and matte are not trained on the same dataset, there exists a difference in the foreground estimation. Thus, we filter the depth map to make it semantically aligned with the predicted alpha matte before establishing the synthetic-of-field MPI. We observe that the coarse depth maps have smoother edges and more enlarged foreground areas. We use a fine-detailed mask to regularize the depth. Although edge-preserving filters such as bilateral [2] and guided filters [14] have been proposed for these cases, they cannot cope with largely biased areas. Through regularization and multi-layer image fusion, we could make the intensity variation along the boundary more smooth and natural while preserving the details of the foreground. To achieve this, we first obtain the average depth of the foreground masked areas by $mean(D * M)$. Then, we relax the average to a small range to separate a coarse mask M_{coarse}. we dilate the original mask M with a kernel of size K, resulting in M_{large}. Lastly, we apply the guidance filter [14] on the depth map within the area $M_{coarse} + M_{large} - M$.

We split the range of the filtered depth map into N bins $D_k = [d_k, d_{k+1}], i = 1, ..., N$, and the midpoint depth of each interval is the depth of the corresponding plane, i.e., N planes in MPI. We could apply disk or hexagonal kernels to blur the input image I as the RGB channels of each plane. Fig. 5 illustrates the blurring effects of using two kinds of kernels. Since most DSLR cameras use circular apertures, we only consider disk kernels in the paper, while the idea could be easily extended to hexagonal apertures. The kernel size is determined by:

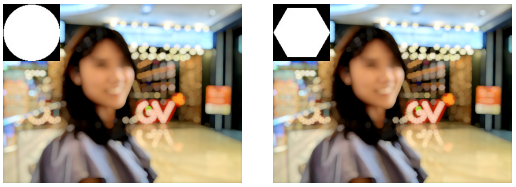
$$r_k = \lfloor |d_i - d_{focus}|/\sigma + \delta \rfloor \tag{3}$$

where σ and δ are the global hyperparameters to control the relative blur degree in different depth planes, and vary with the image resolution.

Simply convolving the input image I in RGB space with a disk kernel is not able to produce salient bokeh effects since this linear transformation over I would reduce the global contrast and light intensity. We transform I into the exponential space and shift the intensity to change the contrast before applying the blur. After the convolution, we transform the image back to RGB space. The whole transformation is defined as:

$$I_k = (Conv_{r_k}((a_1 I_{input} + b_1)^{\alpha_1}))^{1/\alpha_2}/a_2 + b_2 \tag{4}$$

where a_1, a_2, b_1, b_2 are the hyperparameters to control the contrast and light intensity, α_1 and α_2 are used to control bokeh intensity, $Conv_{r_k}$ represent the convolution using a disk kernel with radius r_k.



Disk Kernel Hexagonal Kernel

Figure 5. **The kernels used to generate blur effects.** The aperture size and shape could be simulated by changing the kernel size and shape.

The alpha channel encodes the visibility information, the same as the original MPI. Each pixel in the input image is associated with N depth candidates in MPI, representing different degrees of blurring. We want the plane with the closest depth to the actual depth to be the most visible, and hence we formulate the alpha channel in the k-th layer as follows:

$$\alpha_{i,k} = 1 - \sqrt{(|d_i - (d_k + d_{k+1})/2|)/(d_{max} - d_{min})}, \tag{5}$$

where i is the pixel index, d_i is the actual depth value at pixel i, d_{max} and d_{min} are the maximum and minimum depth in the depth map D, respectively. After processing, the alpha channel is normalized to $[0, 1]$. Then we have established a practical MPI represented by $\{C_j, \alpha_j\}_{j=1}^{N}$, where C_j is the color vector consisting of RGB values and α_j is the alpha vector at the j-th layer, calculated from Eq. 4 and Eq. 5, respectively. The blurred image $I_{defocus}$ is rendered by over-composition from back to front:

$$I_N^{over} = C_N, \alpha_N^{over} = \alpha_N \, , \tag{6}$$

$$\alpha_{j-1}^{over} = \alpha_{j-1} + \alpha_j^{over}(1 - \alpha_{j-1}) \, , \tag{7}$$

$$I_{j-1}^{over} = \frac{C_j \alpha_j + I_j^{over} \alpha_j^{over}(1 - \alpha_j)}{\alpha_{j-1}^{over}} \, , \tag{8}$$

$$I_{defocus} = I_0^{over}. \tag{9}$$

The process of obtaining MPI representation is free of deep neural networks, making it possible to adjust parameters as needed. Usually, we would use the matting equation to paste the clear foreground back to the rendered image to highlight the main body of the portrait character:

$$I_{final} = I_{defocus} * (1 - \alpha_{matte}) + \alpha_{matte} * I_{input}. \tag{10}$$

4. Experiments

4.1. Implementation

Dataset For the matting task, we utilize the P3M-10k dataset [23] to train our model. This dataset comprises 9,421 blurred portrait images in the training set, along with

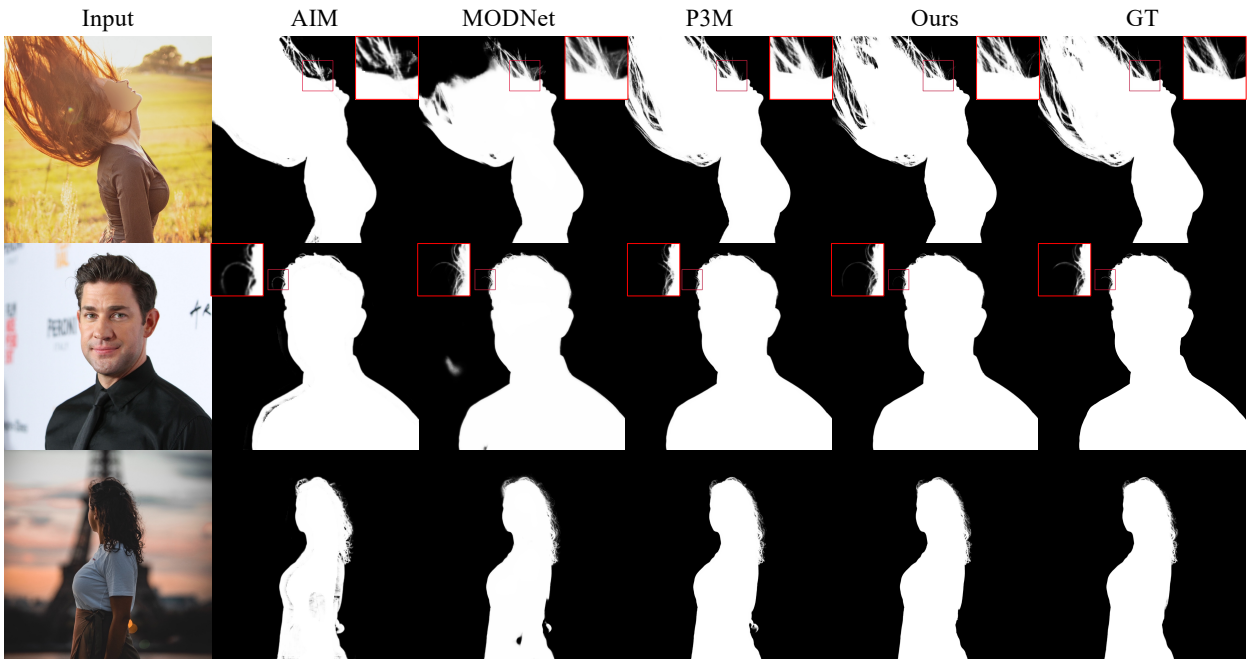


Figure 6. The qualitative results on the P3M-10k dataset [23] of our method compared to the other methods

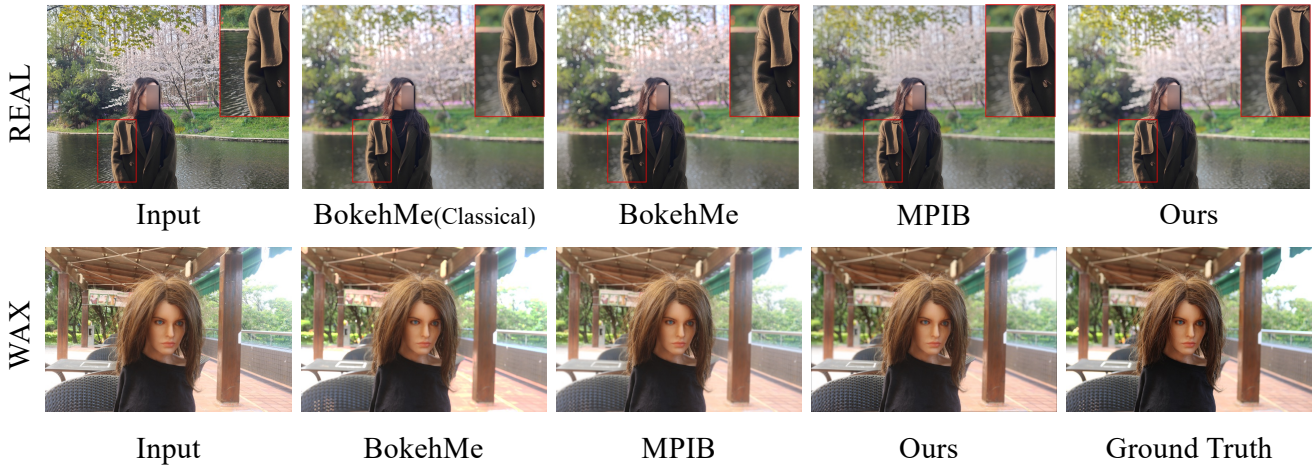


Figure 7. The qualitative results in the REAL(top) and WAX(bottom) datasets of our method compared to baseline methods. Our method, compared to network-based methods, has the ability to eliminate halos in the edge areas between the character and the background environment, resulting in more realistic depth-of-field blur effects.

500 blurred and 500 clear portrait images in the validation set. For the bokeh effect rendering task, we collect two datasets to perform the evaluation. The first, called WAX, is used to perform quantitative optimization and evaluation. It contains 20 photos captured by a Canon EOS 70D camera using wax figures as foreground on different environment backgrounds. The second dataset, called REAL, consists of around 100 real-person images captured using Huawei and iPhone smartphones, and thus they are closer to the expected input data distribution for our task.

Experiment Setting For the matting task, we train the model by the data with the size of 512×512 at the first

Table 1. Quantitative matting results on the P3M-10k dataset [23]. Models are trained on blurry portrait images and tested on both of blurry and clear portrait images.

Methods	Blurry		Non Blurry	
	MAD↓	MSE↓	MAD↓	MSE↓
LF [66]	0.0251	0.0191	0.0178	0.0129
AIM [25]	0.0193	0.0156	0.0165	0.0101
MODNet [18]	0.0089	0.0042	0.0085	0.0041
GFM [24]	0.0082	0.0041	0.0081	0.0041
P3M [29]	0.0046	0.0023	0.0042	0.0019
Ours	**0.0040**	**0.0021**	**0.0038**	**0.0017**

Table 2. The ablation study matting results on the P3M-10k dataset [23].

Methods	Blurry		Non Blurry	
	MAD↓	MSE↓	MAD↓	MSE↓
w/o AGM	0.0052	0.0025	0.0047	0.0023
w/o low-resolution branch	0.0058	0.0036	0.0054	0.0031
w/o high-resolution branch	0.0068	0.0052	0.0079	0.0050
Ours	**0.0040**	**0.0021**	**0.0038**	**0.0017**

stage. Besides, we first randomly crop the images with the size of 768×768 or 1024×1024, then we resize the data to 512×512 during the rest stages. The depth maps of given input portrait images are obtained by pretrained DPT [43]. For estimating the alpha matte, we resize the input image to 1024x1024 to fit the network requirement. The output is bilinearly resized to the original resolution to yield the matte mask. For transferring a depth map into a depth-of-field MPI, we normalize the original depth range to [0, 255] and use $D_k = 32$ depth planes. The choice of other hyper-parameters is presented in the supplementary. Our method is optimized and evaluated on the two customized datasets.

Evaluation Metrics We report the MSE and MAD metrics for image matting, derived from L2 and L1 distance, respectively. For depth-of-field rendering, we adopt the structural similarity (SSIM), the peak signal-to-noise ratio (PSNR), and the perceptual distance LPIPS [64] to test the performance of our method on the WAX dataset.

4.2. Qualitative and Quantitative Evaluation

The matting qualitative and quantitative results are shown in Figure 6 and Table 1. The results show that our matting pipeline yields more accurate matting results, especially in the hair area. Table 2 shows the results of the different architectures.

The qualitative rendering result of self-captured portrait images is shown in Figure 7. The results show that our method could produce more realistic depth-of-field effects than existing state-of-the-art methods. Using multi-layer superposition makes the images appear clear at the near and blurred at the far. Our bokeh effect is more natural and salient than the method [39] using a learned blur kernel. Moreover, our method is better at coping with edge areas since there are fewer halo artifacts. Table 3 reports the quantitative metrics in the WAX dataset. Our method is superior to BokehMe [37] in all metrics and could achieve comparable performance with respect to the learning-based method [39].

The qualitative results of ablation studies to investigate the significance of each proposed module are presented in the supplementary materials. Firstly, we remove the depth filtering module (c) and find that some areas near where the character and the environment intersect are not blurred. This is because the depth estimation network mistakenly considers the part of the environment as the character area, giving a relatively low depth value. Hence, these areas remain clear, which justifies the significance of the depth filtering module. Then, we investigate the usage of intensity-related parameters. We set α_0 and α_1 to different initial values to test their influence. It seems that αs are dominant to highlight the bokeh circle, which would not be synthesized with too small αs. For instance, if they are set to 1, i.e., with no exponential image transforming (e), the rendered result shows only blurred effects, with relatively dark colors and no bokeh style. As αs become large, the bokeh balls are more salient (f). Lastly, we reduce the number of planes (d), and the result cannot obviously show the effect of being gradually blurred from near to far. Also, the incurred in-continuity artifacts indicate that more planes help fuse multiple layers.

Table 3. Quantitative evaluation of our and baseline methods in the WAX dataset.

Methods	PSNR ↑	SSIM ↑	LPIPS ↓
BokehMe [37]	23.12	0.9290	0.3963
	23.22	0.9295	0.3962
	23.15	0.9292	0.3881
MPIB [39]	26.26	0.9157	0.3881
Ours	**26.29**	**0.9386**	**0.3735**

5. Conclusion and Limitation

We have introduced a novel framework for synthesizing depth-of-field in order to create realistic portrait images with stunning bokeh effects, all from a single all-in-focus photo without any additional input. To achieve this goal, we simulated the lens imaging process and divided our framework into three main stages: matting mask estimation, depth filtering, and MPI-based rendering. Our proposed multi-step fine-grained matting network successfully captures intricate details down to the level of individual strands of hair. Notably, the zero-shot mechanism to construct an MPI boasts a significantly reduced number of parameters, making it easily optimized and controllable. However, one of the challenges encountered in our framework is the time-consuming nature of convolutions involving large kernels. This poses a limitation on the applicability of our framework in scenarios that demand real-time processing and ultra-high-definition rendering. Potential strategies for optimization may include implementing more efficient convolution algorithms or exploring hardware acceleration options. Future work will involve optimizing this aspect of our framework to extend its practical use to these high-demand scenarios, thereby enhancing its overall usability and potential impact in the field of image processing and rendering.

References

[1] Soonmin Bae and Frédo Durand. Defocus magnification. In *Computer graphics forum*, volume 26, pages 571–579. Wiley Online Library, 2007. 1

[2] Jonathan T Barron, Andrew Adams, YiChang Shih, and Carlos Hernández. Fast bilateral-space stereo for synthetic defocus. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4466–4474, 2015. 1, 3, 6

[3] Marcelo Bertalmio, Pere Fort, and Daniel Sanchez-Crespo. Real-time, accurate depth of field using anisotropic diffusion and programmable graphics cards. In *Proceedings. 2nd International Symposium on 3D Data Processing, Visualization and Transmission, 2004. 3DPVT 2004.*, pages 767–773. IEEE, 2004. 3

[4] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 5

[5] Qifeng Chen, Dingzeyu Li, and Chi-Keung Tang. Knn matting. *IEEE transactions on pattern analysis and machine intelligence*, 35(9):2175–2188, 2013. 4

[6] Donghyeon Cho, Yu-Wing Tai, and Inso Kweon. Natural image matting using deep convolutional neural networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 626–643. Springer, 2016. 4

[7] Jieren Deng, Xin Zhou, Hao Tian, Zhihong Pan, and Derek Aguiar. Gbsd: Generative bokeh with stage diffusion. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7070–7074. IEEE, 2024. 4

[8] Saikat Dutta. Depth-aware blending of smoothed images for bokeh effect generation. *Journal of Visual Communication and Image Representation*, 77:103089, 2021. 3

[9] Saikat Dutta, Sourya Dipta Das, Nisarg A Shah, and Anil Kumar Tiwari. Stacked deep multi-scale hierarchical network for fast bokeh effect rendering from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2398–2407, 2021. 3

[10] Eduardo SL Gastal and Manuel M Oliveira. Shared sampling for real-time alpha matting. In *Computer Graphics Forum*, volume 29, pages 575–584. Wiley Online Library, 2010. 4

[11] Konstantinos Georgiadis, Albert Saà-Garriga, Mehmet Kerim Yucel, Anastasios Drosou, and Bruno Manganelli. Adaptive mask-based pyramid network for realistic bokeh rendering. In *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part II*, pages 429–444. Springer, 2023. 3

[12] Thomas Hach, Johannes Steurer, Arvind Amruth, and Artur Pappenheim. Cinematic bokeh rendering for real scenes. In *Proceedings of the 12th European Conference on Visual Media Production*, pages 1–10, 2015. 3

[13] Kaiming He, Christoph Rhemann, Carsten Rother, Xiaoou Tang, and Jian Sun. A global sampling method for alpha matting. In *CVPR 2011*, pages 2049–2056. Ieee, 2011. 4

[14] Kaiming He, Jian Sun, and Xiaoou Tang. Guided image filtering. *IEEE transactions on pattern analysis and machine intelligence*, 35(6):1397–1409, 2012. 2, 6

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 4

[16] Robert T Held, Emily A Cooper, James F O'brien, Martin S Banks, et al. Using blur to affect perceived distance and size. *ACM Trans. Graph.*, 29(2):19–1, 2010. 1

[17] Andrey Ignatov, Jagruti Patel, and Radu Timofte. Rendering natural camera bokeh effect with deep learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 418–419, 2020. 2

[18] Zhanghan Ke, Jiayu Sun, Kaican Li, Qiong Yan, and Rynson WH Lau. Modnet: Real-time trimap-free portrait matting via objective decomposition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 1140–1147, 2022. 4, 5, 7

[19] Brian Lee, Fei Lei, Huaijin Chen, and Alexis Baudron. Bokeh-loss gan: multi-stage adversarial training for realistic edge-aware bokeh. In *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part II*, pages 619–634. Springer, 2023. 3

[20] Jonghyun Lee, Hansam Cho, YoungJoon Yoo, Seoung Bum Kim, and Yonghyun Jeong. Compose and conquer: Diffusion-based 3d depth aware composable image synthesis. In *The Twelfth International Conference on Learning Representations*, 2024. 4

[21] Sungkil Lee, Elmar Eisemann, and Hans-Peter Seidel. Real-time lens blur effects and focus control. *ACM Transactions on Graphics (TOG)*, 29(4):1–7, 2010. 3

[22] Anat Levin, Dani Lischinski, and Yair Weiss. A closed-form solution to natural image matting. *IEEE transactions on pattern analysis and machine intelligence*, 30(2):228–242, 2007. 4

[23] Jizhizi Li, Sihan Ma, Jing Zhang, and Dacheng Tao. Privacy-preserving portrait matting. In *Proceedings of the 29th ACM International Conference on Multimedia*, MM '21, page 3501–3509, New York, NY, USA, 2021. Association for Computing Machinery. 4, 6, 7, 8

[24] Jizhizi Li, Jing Zhang, Stephen J Maybank, and Dacheng Tao. Bridging composite and real: towards end-to-end deep image matting. *International Journal of Computer Vision*, 130(2):246–266, 2022. 4, 7

[25] Jizhizi Li, Jing Zhang, and Dacheng Tao. Deep automatic natural image matting. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 800–806. International Joint Conferences on Artificial Intelligence Organization, 8 2021. Main Track. 4, 7

[26] Jizhizi Li, Jing Zhang, and Dacheng Tao. Referring image matting. *arXiv preprint arXiv:2206.05149*, 2022. 4

[27] Yaoyi Li and Hongtao Lu. Natural image matting via guided contextual attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11450–11457, 2020. 4

[28] Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. Neural volumes: learning dynamic renderable volumes from images. *ACM Transactions on Graphics (TOG)*, 38(4):1–14, 2019. 1

[29] Sihan Ma, Jizhizi Li, Jing Zhang, He Zhang, and Dacheng Tao. Rethinking portrait matting with pirvacy preserving. *International Journal of Computer Vision*, 2023. 4, 7

[30] Yue Ma, Yingqing He, Xiaodong Cun, Xintao Wang, Siran Chen, Xiu Li, and Qifeng Chen. Follow your pose: Pose-guided text-to-video generation using pose-free videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 4117–4125, 2024. 4

[31] Yue Ma, Yingqing He, Hongfa Wang, Andong Wang, Chenyang Qi, Chengfei Cai, Xiu Li, Zhifeng Li, Heung-Yeung Shum, Wei Liu, et al. Follow-your-click: Open-domain regional image animation via short prompts. *arXiv preprint arXiv:2403.08268*, 2024. 4

[32] Yue Ma, Hongyu Liu, Hongfa Wang, Heng Pan, Yingqing He, Junkun Yuan, Ailing Zeng, Chengfei Cai, Heung-Yeung Shum, Wei Liu, et al. Follow-your-emoji: Fine-controllable and expressive freestyle portrait animation. *arXiv preprint arXiv:2406.01900*, 2024. 4

[33] Yue Ma, Tianyu Yang, Yin Shan, and Xiu Li. Simvtp: Simple video text pre-training with masked autoencoders. *arXiv preprint arXiv:2212.03490*, 2022. 4

[34] B Mildenhall, PP Srinivasan, M Tancik, JT Barron, R Ramamoorthi, and R Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European conference on computer vision*, 2020. 1

[35] Oliver Nalbach, Elena Arabadzhiyska, Dushyant Mehta, H-P Seidel, and Tobias Ritschel. Deep shading: convolutional neural networks for screen space shading. In *Computer graphics forum*, volume 36, pages 65–78. Wiley Online Library, 2017. 3

[36] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 165–174, 2019. 1

[37] Juewen Peng, Zhiguo Cao, Xianrui Luo, Hao Lu, Ke Xian, and Jianming Zhang. Bokehme: When neural rendering meets classical rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16283–16292, 2022. 2, 4, 8

[38] Juewen Peng, Xianrui Luo, Ke Xian, and Zhiguo Cao. Interactive portrait bokeh rendering system. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 2923–2927. IEEE, 2021. 3

[39] Juewen Peng, Jianming Zhang, Xianrui Luo, Hao Lu, Ke Xian, and Zhiguo Cao. Mpib: An mpi-based bokeh rendering framework for realistic partial occlusion effects. In *The Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 5, 8

[40] Michael Potmesil and Indranil Chakravarty. A lens and aperture camera model for synthetic image generation. *ACM SIGGRAPH Computer Graphics*, 15(3):297–305, 1981. 1

[41] Ming Qian, Congyu Qiao, Jiamin Lin, Zhenyu Guo, Chenghua Li, Cong Leng, and Jian Cheng. Bggan: Bokeh-glass generative adversarial network for rendering realistic bokeh. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 229–244. Springer, 2020. 2

[42] Yu Qiao, Yuhao Liu, Xin Yang, Dongsheng Zhou, Mingliang Xu, Qiang Zhang, and Xiaopeng Wei. Attention-guided hierarchical structure aggregation for image matting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13676–13685, 2020. 4

[43] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12179–12188, 2021. 1, 2, 8

[44] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(3):1623–1637, 2020. 1

[45] Parikshit Sakurikar, Ishit Mehta, Vineeth N Balasubramanian, and PJ Narayanan. Refocusgan: Scene refocusing using a single image. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 497–512, 2018. 4

[46] Tim Seizinger, Marcos V Conde, Manuel Kolmet, Tom E Bishop, and Radu Timofte. Efficient multi-lens bokeh effect rendering and transformation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1633–1642, 2023. 4

[47] Soumyadip Sengupta, Vivek Jayaram, Brian Curless, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Background matting: The world is your green screen. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2291–2300, 2020. 4

[48] Xiaoyong Shen, Aaron Hertzmann, Jiaya Jia, Sylvain Paris, Brian Price, Eli Shechtman, and Ian Sachs. Automatic portrait segmentation for image stylization. In *Computer Graphics Forum*, volume 35, pages 93–102. Wiley Online Library, 2016. 3

[49] Xiaoyong Shen, Xin Tao, Hongyun Gao, Chao Zhou, and Jiaya Jia. Deep automatic portrait matting. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 92–107. Springer, 2016. 3

[50] Cyril Soler, Kartic Subr, Frédo Durand, Nicolas Holzschuch, and François Sillion. Fourier depth of field. *ACM Transactions on Graphics (TOG)*, 28(2):1–12, 2009. 3

[51] Towaki Takikawa, David Acuna, Varun Jampani, and Sanja Fidler. Gated-scnn: Gated shape cnns for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5229–5238, 2019. 4, 5

[52] Neal Wadhwa, Rahul Garg, David E Jacobs, Bryan E Feldman, Nori Kanazawa, Robert Carroll, Yair Movshovitz-Attias, Jonathan T Barron, Yael Pritch, and Marc Levoy. Synthetic depth-of-field with a single-camera mobile phone. *ACM Transactions on Graphics (ToG)*, 37(4):1–13, 2018. 1, 2, 3

[53] Jue Wang and Michael F Cohen. Optimized color sampling for robust matting. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007. 4

[54] Lijun Wang, Xiaohui Shen, Jianming Zhang, Oliver Wang, Zhe Lin, Chih-Yao Hsieh, Sarah Kong, and Huchuan Lu. Deeplens: shallow depth of field from a single image. *arXiv preprint arXiv:1810.08100*, 2018. 1

[55] Zhifeng Wang, Aiwen Jiang, Chunjie Zhang, Hanxi Li, and Bo Liu. Self-supervised multi-scale pyramid fusion networks for realistic bokeh effect rendering. *Journal of Visual Communication and Image Representation*, 87:103580, 2022. 3

[56] Jiaze Wu, Changwen Zheng, Xiaohui Hu, Yang Wang, and Liqiang Zhang. Realistic rendering of bokeh effect based on optical aberrations. *The Visual Computer*, 26:555–563, 2010. 3

[57] Ke Xian, Juewen Peng, Chao Zhang, Hao Lu, and Zhiguo Cao. Ranking-based salient object detection and depth prediction for shallow depth-of-field. *Sensors*, 21(5):1815, 2021. 3

[58] Lei Xiao, Anton Kaplanyan, Alexander Fix, Matt Chapman, and Douglas Lanman. Deepfocus: Learned image synthesis for computational display. In *ACM SIGGRAPH 2018 Talks*, pages 1–2. 2018. 3

[59] Ning Xu, Brian Price, Scott Cohen, and Thomas Huang. Deep image matting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2970–2979, 2017. 4

[60] Xiangyu Xu, Deqing Sun, Sifei Liu, Wenqi Ren, Yu-Jin Zhang, Ming-Hsuan Yang, and Jian Sun. Rendering portraitures from monocular camera and beyond. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 35–50, 2018. 2

[61] Yang Yang, Haiting Lin, Zhan Yu, Sylvain Paris, and Jingyi Yu. Virtual dslr: High quality dynamic depth-of-field synthesis on mobile platforms. *Electronic Imaging*, 28:1–9, 2016. 3

[62] Qihang Yu, Jianming Zhang, He Zhang, Yilin Wang, Zhe Lin, Ning Xu, Yutong Bai, and Alan Yuille. Mask guided matting via progressive refinement network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1154–1163, 2021. 4

[63] Xuan Yu, Rui Wang, and Jingyi Yu. Real-time depth of field rendering via dynamic light field generation and filtering. In *Computer Graphics Forum*, volume 29, pages 2099–2107. Wiley Online Library, 2010. 3

[64] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 8

[65] Xuaner Zhang, Kevin Matzen, Vivien Nguyen, Dillon Yao, You Zhang, and Ren Ng. Synthetic defocus and look-ahead autofocus for casual videography. *arXiv preprint arXiv:1905.06326*, 2019. 1, 3

[66] Yunke Zhang, Lixue Gong, Lubin Fan, Peiran Ren, Qixing Huang, Hujun Bao, and Weiwei Xu. A late fusion cnn for digital matting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7469–7478, 2019. 4, 7

[67] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: learning view synthesis using multiplane images. *ACM Transactions on Graphics (TOG)*, 37(4):1–12, 2018. 1, 5