Rendering Portraitures from Monocular Camera and Beyond

Xiangyu Xu[1,2](✉), Deqing Sun[3], Sifei Liu[3], Wenqi Ren[4], Yu-Jin Zhang[1], Ming-Hsuan Yang[5,6], and Jian Sun[7]

[1] Tsinghua University, Beijing, China
xuxiangyu2014@gmail.com
[2] SenseTime, Beijing, China
[3] Nvidia, Santa Clara, USA
[4] Tencent AI Lab, Bellevue, USA
[5] UC Merced, Merced, USA
[6] Google, Menlo Park, USA
[7] Face++, Beijing, China

Abstract. Shallow Depth-of-Field (DoF) is a desirable effect in photography which renders artistic photos. Usually, it requires single-lens reflex cameras and certain photography skills to generate such effects. Recently, dual-lens on cellphones is used to estimate scene depth and simulate DoF effects for portrait shots. However, this technique cannot be applied to photos already taken and does not work well for whole-body scenes where the subject is at a distance from the cameras. In this work, we introduce an automatic system that achieves portrait DoF rendering for monocular cameras. Specifically, we first exploit Convolutional Neural Networks to estimate the relative depth and portrait segmentation maps from a single input image. Since these initial estimates from a single input are usually coarse and lack fine details, we further learn pixel affinities to refine the coarse estimation maps. With the refined estimation, we conduct depth and segmentation-aware blur rendering to the input image with a Conditional Random Field and image matting. In addition, we train a spatially-variant Recursive Neural Network to learn and accelerate this rendering process. We show that the proposed algorithm can effectively generate portraitures with realistic DoF effects using one single input. Experimental results also demonstrate that our depth and segmentation estimation modules perform favorably against the state-of-the-art methods both quantitatively and qualitatively.

1 Introduction

Shallow Depth of Field (DoF) shooting can enhance photos and render artistic images in which the region containing the main object at a certain distance to

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-01240-3_3) contains supplementary material, which is available to authorized users.

V. Ferrari et al. (Eds.): ECCV 2018, LNCS 11213, pp. 36–51, 2018.
https://doi.org/10.1007/978-3-030-01240-3_3

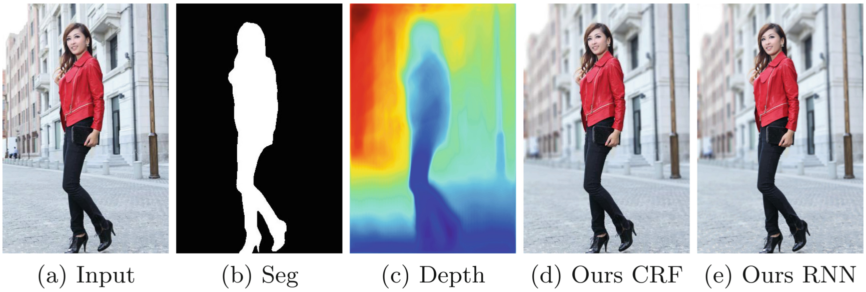


Fig. 1. The proposed method generates realistic DoF effects for whole-body portrait using a single RGB image (a) captured from a monocular camera. (b) and (c) are the segmentation and depth estimates of (a). (d) and (e) are our DoF results generated by the CRF based rendering system and the learned RNN filter, respectively.

the camera is well-focused, while other pixels are blurred [4]. Usually, a single-lens reflex (SLR) camera with a large aperture and certain photography skills are needed to render portraitures.

The portrait mode, which allows users to take DoF photos, is a major feature of the latest smart phones, *e.g.*, iPhone7+ and Google Pixel 2. Unlike SLR cameras, mobile phone cameras have a small, fixed-size aperture, which generates pictures with everything more or less in focus (Fig. 1(a)). Thus, generating DoF effects requires depth information, which has been obtained via specialized hardware in high-end phones. For example, iPhone 7+ relies on dual-lens to estimate depth, and Google Pixel2 uses Phase-Detect Auto-Focus (PDAF), which can also be regarded as two lenses on the left and right sides.

However, existing systems using specialized hardware have several limitations. First, they do not perform well for whole-body portraits which are at a relatively large distance to the lens. As the baseline between two lenses is small, it is challenging to estimate large depth fields. Second, it is impractical to implement these hardware solutions other than high-end phones. More importantly, there are billions of photos already taken that these systems cannot process.

In this paper, we introduce an automatic system that achieves DoF rendering for monocular cameras. Specifically, we use deep neural networks to estimate depth and segment portrait from a single image. While deep learning based methods have made significant progress in single image depth prediction and portrait segmentation, the results by state-of-the-art methods [7,9–11,20,23,25] are still too coarse for DoF rendering. To obtain more precise depth and segmentation, we improve the initial estimates using the Spatial Propagation Networks (SPN) [22]. With the refined depth and segmentation, our system applies depth and segmentation aware blurring to the background with Conditional Random Field (CRF) and image matting. Experimental results show that our system can achieve realistic DoF effects on a variety of half and full-body portrait images. To further accelerate this rendering process, we train a spatially-variant Recursive Neural Network (RNN) [21] filter with guidance from the depth and segmentation to learn to generate the DoF effects. Since it is extremely difficult to capture image pairs with and without DoF effects for the same scene, we use the generated results from the CRF-based system as our training samples. We show that the proposed network could effectively and efficiently approximate the CRF-based system and generate high-quality DoF results.

The main contributions of this work are summarized as follows. First, we propose an automatic system that achieves realistic DoF rendering for single portrait images. While some components of this system are known in the field, it requires meticulous algorithmic design and efforts to achieve the state-of-the-art results. Second, we train a depth and segmentation guided RNN model to approximate and accelerate the rendering process, which outperforms previous deep learning based filtering methods. In addition, we achieve the state-of-the-art performance on portrait segmentation using a SPN. We also demonstrate that sparse depth labels can be used for training a SPN, and that depth estimation can be improved by using additional portrait segmentation data.

2 Related Work

Portrait Segmentation. Deep learning achieves promising results on many applications [19,20,22,32,33]. For semantic segmentation, many recent works are based on CNNs. Long *et al.* [23] introduce fully convolutional neural network (FCNN), which convolutionalizes the classification networks, such as VGG [26], to directly output segmentation maps. Numerous segmentation methods have subsequently been developed. In particular, Shen *et al.* [25] adapt the FCNN to selfie portrait segmentation by using additional position and shape channels. Liu *et al.* [20] extend the FCNN by adding recurrent modules and use it on foreground segmentation. However, FCNN based methods do not explicitly model the pairwise relations (*i.e.*affinity) of pixels and their segmentation maps lack details and subtle structures. To remedy this problem, Chen *et al.* [5] and Zheng *et al.* [35] apply a dense CRF to model the affinity and refine the segmentation maps predicted by FCNNs. Liu *et al.* [22] propose the spatial propagation network with 2D propagation modules to learn pixel affinities in an end-to-end manner. As the fine structures and accurate segmentation boundaries are critical for rendering realistic DoF images, we apply SPNs to segment portraits and achieve the state-of-the-art results on a portrait segmentation dataset.

Depth Estimation with Single Image. Deep learning based models have been used to learn depth from a single image, both in supervised and unsupervised ways. For supervised depth learning, Eigen *et al.* [10] propose a CNN architecture that integrates coarse-scale depth prediction with fine-scale prediction. Furthermore, Eigen *et al.* [9] use a pre-trained classification network to improve depth accuracy, such as the AlexNet [16] and VGG [26] models. Recently, Laina *et al.* [17] use a ResNet-based encoder-decoder architecture to generate dense depth maps. These supervised-learning methods need densely-labeled RGB-D images which are limited to indoor scenes (*e.g.*, NYU dataset [24]).

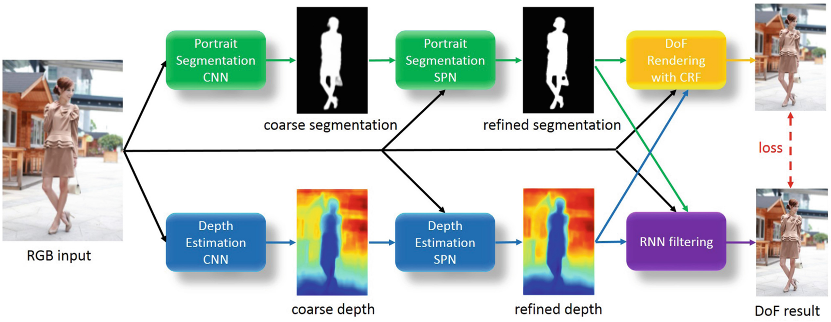


Fig. 2. Overview of the proposed algorithm. We first use off-the-shelf models for single image depth estimation and portrait segmentation. Then we further train SPNs to learn image affinities for refining the depth and segmentation. Finally, we generate the DoF result by exploiting both the refined depth and segmentation map, and learn a spatially-variant RNN to accelerate the rendering process.

On the other hand, several methods [11,13,30] learn depth map prediction in an unsupervised way using an image alignment loss that enforces left-right consistency of the training stereo pairs. However, these methods are still limited to specific scenarios (e.g., scenes in the KITTI [12] and Cityscape [8] datasets) and cannot handle portraits taken by cellphones in everyday life.

Chen *et al.* [7] propose the Depth in the Wild (DIW) dataset which consists of everyday images with relative depth labels between sparsely sampled point pairs. We show that SPN can be trained by sparse labels from the DIW dataset for accurate depth estimation. Moreover, the additional portrait segmentation dataset helps improve depth estimation for portraits, as we can enforce that the depth of different locations on the human body should be consistent.

DoF Rendering. The DoF effect is an important characteristic for realistic image synthesis in computer graphics. A number of DoF rendering methods have been proposed for image synthesis, such as rendering lightfield [27,34] and tracing rays [18,28]. All these image synthesis methods assume that the 3D information of the scene is known.

In contrast, generating DoF effects for RGB images captured from monocular cameras is more challenging. Some methods [14,36] rely on 3D cameras to capture the depth map as well as the RGB image, and generate DoF effects with the obtained depth. Barron *et al.* [2] recover depth with stereo pairs to render defocus images. Bae *et al.* [1] achieve desired DoF effects without using depth information by detecting and magnifying depth blur in a single image. However, their method needs the input images to have mild depth blur at first, which is not always accessible in real scenarios, such as small apertures of cellphones. Shen *et al.* [25] also generate DoF effects for single images by portrait segmentation. But their method is designed for selfies and cannot be used for whole-body images. In addition, the uniform blur kernel they use can bring boundary effects, as shown in Fig. 3(c). Different from the aforementioned methods, our method does not need special input or shooting devices such as 3D cameras. Instead, we use deep neural networks to obtain accurate depth and segmentation with fine details. Then we adopt a CRF model to split image layers using the estimated depth and generate the DoF effect for whole-body portraits by exploiting both the depth and segmentation information. In addition, we propose segmentation and depth guided RNN to accelerate and approximate the rendering process.

3 Proposed Algorithm

As it is extremely difficult to capture image pairs with and without DoF effect for the same scene, we do not take the elegant end-to-end approach for DoF rendering. Instead, we propose to integrate both learning-based and traditional vision algorithms into a novel system that does not require such a training set. Similar to Google Pixel2, our system simulates the real imaging process and applies depth-dependent blurring to an input image. While Google Pixel2 relies on hardware and lacks technical details, our software-based system works with any type of cellphone and can also process existing photos.

An overview of our system is shown in Fig. 2. Specifically, we first use off-the-shelf models for single image depth estimation [7] and portrait segmentation [20] to bootstrap our system. Since the initial estimation maps are coarse, we further train SPNs [22] to learn image affinity for refining the depth estimation and segmentation. With the refined depth and segmentation map, we split the background into layers of different depth using a CRF model and then perform segmentation and depth aware blur rendering to generate the DoF result. In the meanwhile, a spatially-variant RNN filter is learned with segmentation and depth as guidance map and the aforementioned DoF result as ground truth to accelerate the rendering process.

3.1 Portrait Segmentation

Spatial Propagation Network. The SPN [22] model consists of a deep CNN that learns the affinity entities of an input image I, and a spatial linear propagation module that refines a coarse mask M. The coarse mask is refined under the guidance of affinities, *i.e.*, learned pairwise relationships for any pixel pairs. All modules are differentiable and can be jointly trained using backpropagation.

In this work, we adopt an encoder-decoder architecture with concatenation skip connections as the guidance network, where we use the VGG-16 [26] pre-trained network from the conv1 to pool5 as the downsampling part. The upsampling part has the exactly symmetric architecture and is learned from scratch. With the weights generated by the guidance network, the propagation module takes a coarse mask as input, and propagates the coarse information in four directions, *i.e.*, left-to-right, top-to-bottom, and the other two with the reverse directions.

Loss Function. For portrait segmentation, the coarse mask of SPN for image I is generated by the foreground segmentation model [20]. We denote the output of SPN as v, and the final segmentation map is generated by a sigmoid function: $m = 1/(1 + \exp(-v))$. We use a pixel-wise cross-entropy loss for training, which is defined as:

$$L_1(m) = -\sum_{i \in \mathcal{F}} \log m_i - \sum_{j \in \mathcal{B}} \log(1 - m_j), \tag{1}$$

where the sets $\mathcal{F}$ and $\mathcal{B}$ contain pixels in the foreground and background masks of the ground truth, respectively.

3.2 Depth Estimation

The initial depth predicted by [7] is also refined by a SPN, which has the same network architecture as the one for segmentation. We use the Depth in the Wild dataset [7] that contains images from different scenes. As the images of this dataset are only sparsely annotated with relative depth between pairs of random point pairs, we use the ranking loss [7] for training. Consider a training image I and its annotation $\{i, j, \gamma\}$ where i and j are the two annotated points, and $\gamma \in \{+1, -1\}$ is the ground-truth depth relation between i and j: $\gamma = 1$ if i is further than j, and $\gamma = -1$ vice versa. Let z be the predicted depth map and z_i, z_j be the depths at point i and j. The ranking loss is defined as:

$$L_2(z) = \begin{cases} \log(1 + \exp(-z_i + z_j)), \gamma = +1, \\ \log(1 + \exp(z_i - z_j)), \gamma = -1, \end{cases} \tag{2}$$

which encourages the predicted depth difference between z_i and z_j to be consistent with the ground-truth ordinal relation.

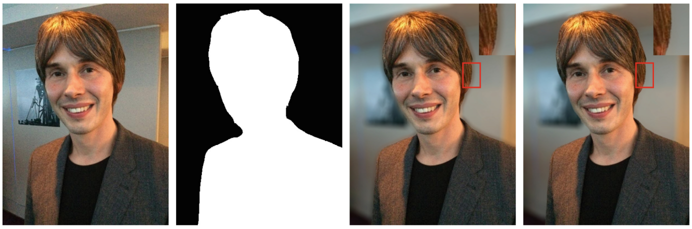
In addition to the dataset with depth annotation, we also exploit the segmentation labels in the portrait segmentation dataset for better depth estimation of portrait images. As pixels at different locations of the portrait should have similar depth values, we use a loss function:

$$L_3(z) = \sum_{i,j \in \mathcal{F}} \max\{0, (z_i - z_j)^2 - \delta\}, \tag{3}$$

where $i, j \in \mathcal{F}$ are the pixels on the human body. As the depth values at different parts of the human body are not exactly the same, we adopt a soft constraint that allows small depth differences and only punishes the depth differences larger than a margin δ.

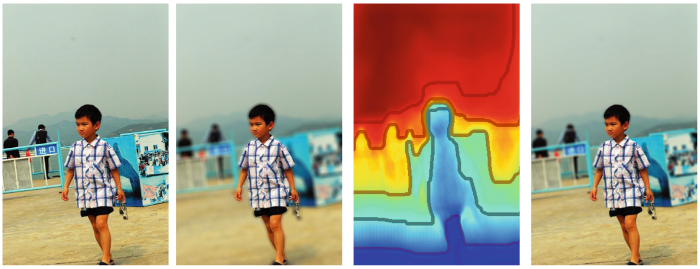
3.3 DoF Rendering

Most smartphones have two shooting modes that use the front and rear cameras respectively. For selfie images captured by a front camera, the background is always further than the person. To generate the DoF effect, we can simply blur



(a) Input (b) Mask (c) Uniform (d) Guided

Fig. 3. Effectiveness of segmentation-guided blur kernel. A uniform blur causes boundary artifacts (c) while our method generates DoF effects with sharper boundaries (d).



(a) Input (b) w/o depth (c) Depth layers (d) with depth

Fig. 4. Comparison of the whole-body DoF results with and without using depth information. We generate more realistic result (d) by splitting depth layers (c).

the background with a disk blur kernel and keep the foreground clear. The blur process is formulated as:

$$B_i = m_i I_i + (1 - m_i) \sum_j w_{ij} I_j, \tag{4}$$

where I, B are the clear image and blurred result respectively; and m represents the portrait segmentation mask. The disk blur kernel w is defined as:

$$w_{ij} = \begin{cases} 1/C, \|p_i - p_j\| < r, \\ 0, \text{otherwise}, \end{cases} \tag{5}$$

where p_i is the coordinate of pixel i, and r is the radius of the disk blur kernel. The blur kernel is normalized by a constant C.

However, a uniform kernel may contaminate the background pixels with foreground pixels in the blurring process, and lead to boundary effect as shown in Fig. 3(c). To address this issue, we propose a new blur kernel $\hat{w}_{ij}$ which is guided by the segmentation mask m. The guided blur kernel is defined as:

$$\hat{w}_{ij}(m) = w_{ij}(1 - m_j) / \sum_j w_{ij}(1 - m_j), \tag{6}$$

where only the background pixels are used during the blurring process. Our method effectively removes the boundary effect as shown in Fig. 3(d).

Whole-Body Portraits. For whole-body portraits taken by a rear camera, naively blurring the background without considering the depth information cannot generate realistic results. As shown in Fig. 1(b), some parts of the background have similar depth with the human body and should also be kept clear. Thus, we exploit the depth estimation to generate better blurred portraitures.

As shown in Fig. 4(c), even with SPN refinement, the depth estimation from a single image is still imperfect and noisy. Thus we split the image into different depth layers using a CRF model which encourages depth smoothness in neighboring regions. The energy function for our depth labeling problem is formulated as:

$$E(l|z) = \sum_i u(l_i|z_i) + \lambda \sum_{(i,j)\in\mathcal{N}, i<j} e(l_i, l_j|z_i, z_j), \tag{7}$$

where $\mathcal{N}$ is the 4-nearest-neighborhood system on pixels. In addition, λ is a hyper-parameter that balances the unary term $u(l_i|z_i)$ and the pairwise term $e(l_i, l_j|z_i, z_j)$. We derive the function $u(l_i|z_i)$ from the estimated depth z_i to measure the cost of assigning the layer label $l_i \in \{1, 2, ..., K\}$ to the pixel i. Specifically, we first find K clusters for the depth values using the K-means algorithm. We assume that the depth value in each cluster follows a Gaussian distribution, and $u(l_i|z_i)$ can be defined as the negative log-likelihood of the pixel i belonging to each cluster l_i:

$$u(l_i|z_i) = \|z_i - C_{l_i}\|/\sigma^2_{l_i}, \tag{8}$$

where C_{l_i} and $\sigma^2_{l_i}$ are the cluster center and variance of the cluster l_i.

The pairwise term $e(l_i, l_j|z_i, z_j)$ measures the cost of assigning the labels l_i, l_j to the adjacent pixels i, j and imposes spatial smoothness:

$$e(l_i, l_j|z_i, z_j) = \mathbb{1}(l_i \neq l_j)\exp(-\|f_S(z)_{i\to j}\|), \tag{9}$$

where $\mathbb{1}(\cdot)$ is the indicator function, and f_S is a Sobel operator which detects depth variations between the pixels i and j. We use the Graph Cut algorithm [15] to minimize the energy function $E(l|z)$.

After splitting the image I into K depth layers, we blur each layer l with a unique blur kernel $\hat{w}^l$ with different disk radius r^l. We assume that the human body should be kept clear and do not consider foreground blur. Thus, we set further layer with larger kernel size while closer layer with smaller one. The final result can be rendered by:

$$B_i = m_i I_i + (1 - m_i) \sum_{l=1}^{K} \sum_t g_{it} \mathbb{1}(l_t = l) \sum_j \hat{w}^l_{ij}(m) I_j, \tag{10}$$

where g_{it} is a Gaussian kernel centered at pixel i which feathers and combines layers of different depth. $\sum_t g_{it}\mathbb{1}(l_t = l)$ measures to what degree pixel i belongs to layer l. Figure 4(d) shows a rendered DoF result.

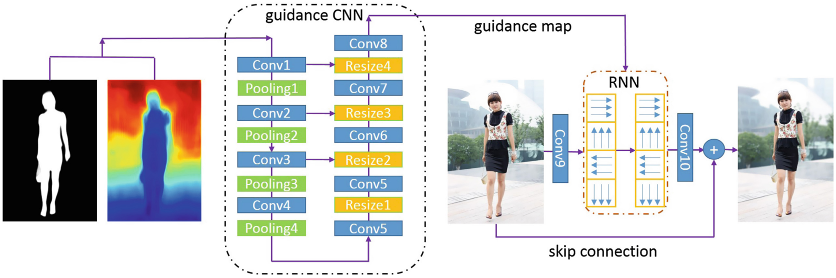


Fig. 5. Illustration of our spatially-variant RNN model. The proposed network contains two groups of RNNs for image filtering and a deep CNN to learn the guidance map with our refined depth and segmentation estimation. To simplify the network training, we add a skip connection from the input to output and learn the residual map instead of the RGB image.

3.4 RNN Filter Learning

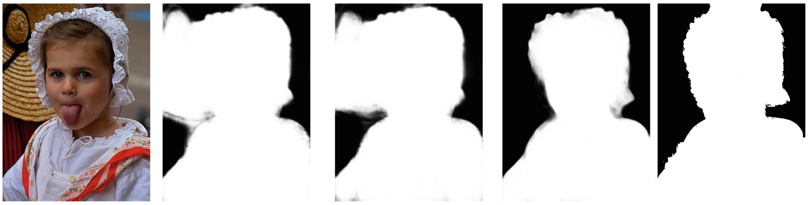
While effective at generating high-quality DoF images, the CRF-based method is computationally expensive because of the CRF optimization, image matting and guided blurring. To reduce the computational cost, we train a deep neural network to approximate the rendering process. Since the DoF blur is spatially-variant, we adopt the RNN filters [21] instead of using a CNN which has the same convolutional kernel at different spatial locations. However, the original method [21] cannot be directly applied for our task, because it learns the the guidance map from RGB images and does not explicitly consider segmentation and depth information. To address this issue, we propose to use the refined segmentation and depth estimation to generate guidance for approximating DoF effects. To simplify the network training, we add a skip connection from the clear image input to the RNN output, because the generated DoF results resemble the original inputs. We use an encoder-decoder CNN to generate the guidance map for the following RNN which combines two groups of recursive filters in a cascaded scheme. The pipeline of our RNN model is shown in Fig. 5.

4 Experimental Results

We show the main results in this section and present more analysis and evaluations in the supplementary material.

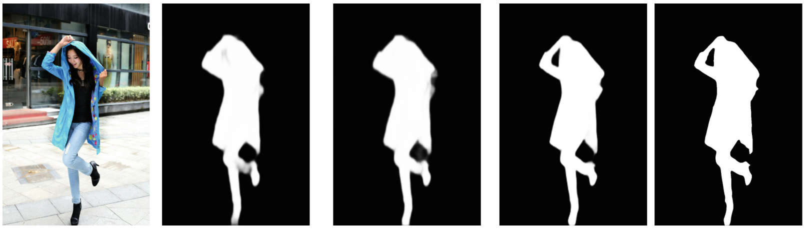
4.1 Implementation Details

Network Training. To train the segmentation network for front camera, we use the selfie image dataset from [25] which is composed of 1428 training and 283 test images. For rear cameras, we use the Baidu human segmentation dataset which has 5387 densely labeled images [29] of which 500 are used for testing and the



(a) Input (b) FCNN [23] (c) PFCN+ [25] (d) Ours (e) GT

Fig. 6. Visual comparison of different segmentation methods on a selfie image.



(a) Input (b) FCNN [23] (c) DHSnet [20] (d) Ours (e) GT

Fig. 7. Visual comparison of different segmentation methods on a whole-body image.

rest for training. During training, we randomly change the aspect ratio and flip the image for data augmentation. To train the depth network, we use the Depth in the Wild [7] dataset which consists of 421K training and 74K test images. For the RNN training, we conduct our CRF-based DoF rendering method on the Baidu human dataset to generate 5K training image pairs and collect another 100 portrait images for evaluation. For all the networks, we use the standard SGD for training with momentum as 0.9 and learning rate as 0.0001.

DoF Rendering. Instead of using the segmentation map from SPN directly, we use the KNN matting method [6] to composite clear foreground and blurry background images. Our estimated segmentation result provides a good trimap initialization for image matting. We generate a trimap by setting the pixels within a 10-pixel radius of the segmentation boundary as the "unknown". This matting scheme performs well as our segmentation provides accurate initial segmentation boundaries. For the CRF model, we empirically split the image into $K = 6$ layers and set the hyper-parameter in (7) as $\lambda = 10$.

4.2 Results of Portrait Segmentation

We quantitatively evaluate our segmentation results on the selfie image dataset [25] and Baidu human segmentation dataset [29]. The segmentation performance is measured by the Interaction-over-Union (IoU) metric. As shown in Table 1, our algorithm achieves the state-of-the-art on selfies. For whole-body images [29], the proposed method achieves an IoU of 93.22 which outperforms

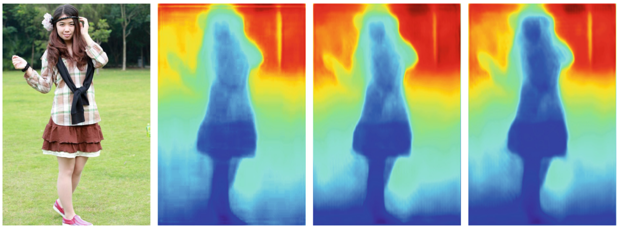

(a) Input (b) Chen [7] (c) Ours (d) Ours+seg

Fig. 8. Visual example of our depth estimation. (c) represents the SPN trained with (2). (d) is further trained with (3) on additional segmentation data.

the 91.43 of the finetuned model of [20]. In addition, we show two examples in Figs. 6 and 7 for qualitative evaluation. The segmentation maps of our method have fine details and small structures, thereby providing accurate foreground information for generating good DoF effects.

4.3 Results of Depth Estimation

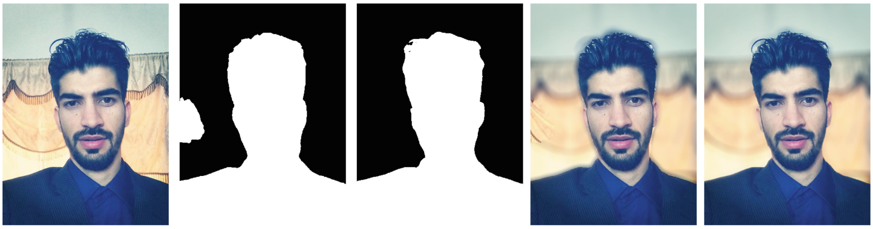
Similar to [7], we use the Weighted Human Disagreement Rate (WHDR) between the predicted ordinal relations and ground-truth ordinal relations to evaluate our method on the DIW test set. Although our result 14.35 on the DIW dataset is only slightly better than the 14.39 of [7], it shows that our SPN can estimate depth properly, and training SPN with sparse labels is effective. As the WHDR only measures ordinal relations of sparse point pairs (one pair each image), it does not evaluate the depth estimation performance well. We present visual examples for qualitative comparison in Fig. 8. Refinement with SPN removes noise in the background and generates sharper boundaries (*e.g.*, the background on the left side in Fig. 8(c)). As shown in Fig. 8(d), using additional segmentation data and our new depth loss (3) further improves the depth consistency on the human body and leads to better depth estimation for portraits.

4.4 Results of CRF-Based DoF Rendering

User Study. Since there is no ground truth for DoF images to perform a quantitative evaluation, we conduct the following user study on the generated DoF

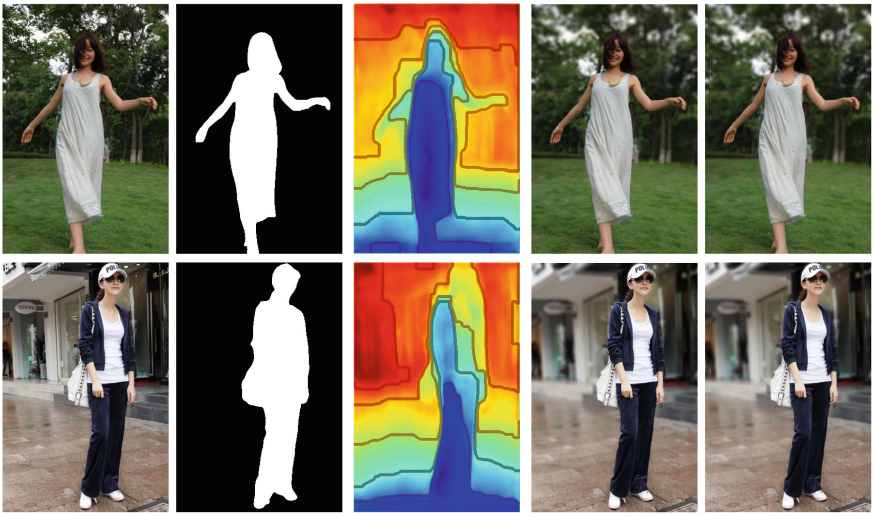
Table 1. Quantitative comparison of different segmentation methods on the selfie image dataset [25]. GC represents graph cut. FCNN has been finetuned on the selfie training set for fair comparisons.

Methods	GC [3]	FCNN [23]	PFCN+ [25]	Ours
Mean IoU (%)	80.02	94.97	95.52	**96.40**



(a) Input (b) Mask of [25] (c) Our mask (d) DoF of [25] (e) Our DoF

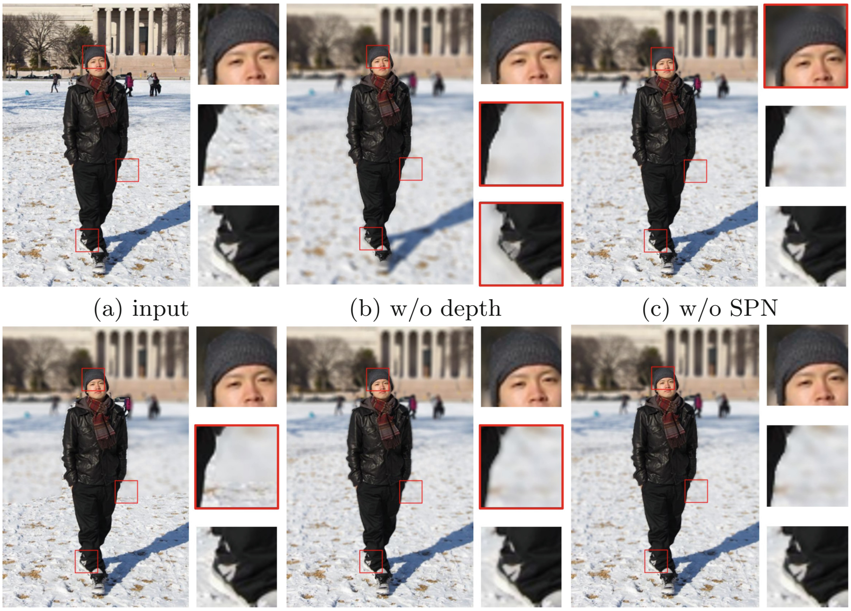
Fig. 9. DoF results on a selfie image. Our method generates better segmentation mask and DoF result without boundary effect (note the glowing boundary in (d)).



(a) Input (b) Our mask (c) Depth layers (d) Ours RNN (e) Ours DoF

Fig. 10. Visual example of our DoF results on whole-body images. Our method generates realistic DoF results.

results. This study uses 30 whole-body portrait images where: (a) 10 images are captured by single-lens reflex (SLR) camera, (b) 10 are generated by our algorithm, and (c) 10 images are generated by naively blurring the background without considering depth. These images are presented in a random order to 22 subjects, who are asked to decide if a presented image is generated by a computer or captured by a real SLR. 79.1% users regard (b) as real captured, while the numbers are 81.8% for (a) and 13.2% for (c). The user study shows that the proposed method can effectively generate realistic DoF results, while naively blurring without considering depth cannot generate convincing results. We show several visual examples for selfie and whole-body images in Figs. 9 and 10.


(a) input (b) w/o depth (c) w/o SPN

(d) w/o CRF (e) w/o guided kernel (f) final result

Fig. 11. Ablation study of each component in our CRF-based method.

Ablation Study. As introduced in Sect. 3.3, our DoF rendering system is composed of different components, *i.e.*, SPN, guided blur kernel, depth aware filtering and CRF model. Figure 11 shows an ablation study of each component of our system in rendering DoF images. First, without depth information, uniform blur is applied to the background in Fig. 11(b), and the closer region such as the ground near the human foot is over-blurred. Second, without using the SPN, the coarse segmentation map leads to incorrectly blurred foreground regions such as the top part of the hat in Fig. 11(c). Third, using a naive thresholding scheme to split depth layers instead of the CRF model generates unrealistic boundaries between the clear and blurry regions as shown in the middle part of Fig. 11(d). In addition, removing the guided blur kernel results in noticeable boundary artifacts around the trousers in Fig. 11(e). By contrast, our system effectively integrates different components and generates high-quality DoF results (Fig. 11(e)).

4.5 Results of RNN Filter

We evaluate the proposed RNN filter against the state-of-the-art deep filtering approaches [31] and [21]. We also train a CNN network with refined depth and segmentation maps as additional input to compare with our spatially-variant RNN design. This CNN has the same encoder-decoder structure as the guidance network in Sect. 3.4. For fair comparisons, we use the same settings and training data for all these methods as introduced in Sect. 4.1. As shown in Table 2,

Table 2. Quantitative comparison of different deep networks for learning DoF effects on the Baidu human test set [29].

Methods	Xu *et al.* [31]	Liu *et al.* [21]	Ours CNN	Ours RNN
PSNR (dB)	31.55	33.55	37.35	**40.74**
SSIM	0.9235	0.9432	0.9723	**0.9868**

the proposed filtering algorithm outperforms state-of-the-art methods for approximating DoF effects in terms of PSNR and SSIM. For qualitative evaluation, we show a visual example in Fig. 12. The CNN-based methods (Fig. 12(b) and (d)) incorrectly blur the foreground, such as the textures of the clothes, because CNN uses uniform kernel at different spatial locations and cannot well handle the spatially-variant DoF case. The result by Liu *et al.* contains significant artifacts on the background due to the lack of effective guidance. In contrast, the proposed RNN model explicitly uses the depth and segmentation as guidance to learn a spatially-variant image filter. Thus, we can effectively approximate the CRF-based rendering system and generate realistic DoF results (Fig. 12(e)).

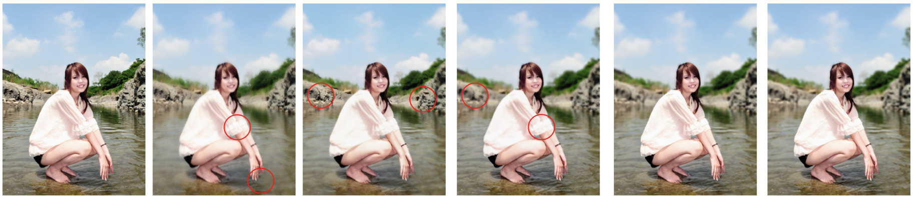

(a) Input (b) Xu [31] (c) Liu [21] (d) Ours CNN (e) Ours RNN (f) Ours CRF

Fig. 12. Visual example of our RNN filtering result. Our method generates realistic DoF result while others wrongly blur the foreground or contain significant artifacts.

Running Time. We implement the proposed algorithm on a desktop with an Intel i7 CPU, 8 GB RAM and an Nvidia GTX 1060 GPU. It takes about 8 s for the CRF-based method to process a 500 × 300 image. By contrast, the learned RNN filter takes only 1.12 s, which significantly accelerates the rendering process and makes it more practical for real applications.

5 Conclusions

In this work, we propose a deep learning and CRF based system that can automatically render realistic DoF results for single portrait images. A spatially-variant RNN filter is trained to accelerate the rendering process with guidance from depth and segmentation. In addition, we achieve the state-of-the-art performance on portrait segmentation using SPN. Furthermore, we demonstrate that sparse depth labels can be used for SPN training. We also show that depth estimation can be improved by enforcing depth consistency on human body with additional portrait segmentation data.

Acknowledgement. This work is supported in part by National Nature Science Foundation of P.R. China (No. 611711184, 61673234, U1636124), the NSF CAREER Grant (No. 1149783), and gifts from Adobe and Nvidia.

References

1. Bae, S., Durand, F.: Defocus magnification. Comput. Graph. Forum **26**, 571–579 (2007)
2. Barron, J.T., Adams, A., Shih, Y., Hernández, C.: Fast bilateral-space stereo for synthetic defocus. In: CVPR (2015)
3. Boykov, Y., Jolly, M.P.: Interactive graph cuts for optimal boundary & region segmentation of objects in N-D images. In: ICCV (2001)
4. Campbell, F.: The depth of field of the human eye. Optica Acta: Int. J. Opt. **4**, 157–164 (1957)
5. Chen, L., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Semantic image segmentation with deep convolutional nets and fully connected CRFs. In: ICLR (2015)
6. Chen, Q., Li, D., Tang, C.: KNN matting. PAMI **25**, 2175–2188 (2013)
7. Chen, W., Fu, Z., Yang, D., Deng, J.: Single-image depth perception in the wild. In: NIPS (2016)
8. Cordts, M., et al.: The cityscapes dataset for semantic urban scene understanding. In: CVPR (2016)
9. Eigen, D., Fergus, R.: Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: ICCV (2015)
10. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. In: NIPS (2014)
11. Garg, R., B.G., V.K., Carneiro, G., Reid, I.: Unsupervised CNN for single view depth estimation: geometry to the rescue. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016 Part VIII. LNCS, vol. 9912, pp. 740–756. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46484-8_45
12. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? The KITTI vision benchmark suite. In: CVPR (2012)
13. Godard, C., Aodha, O.M., Brostow, G.J.: Unsupervised monocular depth estimation with left-right consistency. In: CVPR (2017)
14. Huhle, B., Schairer, T., Jenke, P., Straßer, W.: Realistic depth blur for images with range data. In: Kolb, A., Koch, R. (eds.) Dyn3D 2009. LNCS, vol. 5742, pp. 84–95. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-03778-8_7
15. Kolmogorov, V., Zabih, R.: What energy functions can be minimized via graph cuts? In: ECCV (2002)
16. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: NIPS (2012)
17. Laina, I., Rupprecht, C., Belagiannis, V., Tombari, F., Navab, N.: Deeper depth prediction with fully convolutional residual networks. In: 3DV (2016)
18. Lee, S., Eisemann, E., Seidel, H.: Real-time lens blur effects and focus control. ACM Trans. Graph. (SIGGRAPH) **29**, 1–7 (2010)

19. Liu, C., Xu, X., Zhang, Y.J.: Temporal attention network for action proposal. In: ICIP (2018)
20. Liu, N., Han, J.: DHSNet: deep hierarchical saliency network for salient object detection. In: CVPR (2016)
21. Liu, S., Pan, J., Yang, M.: Learning recursive filters for low-level vision via a hybrid neural network. In: ECCV (2016)
22. Liu, S., De Mello, S., Gu, J., Zhong, G., Yang, M.-H., Kautz, J.: Learning affinity via spatial propagation networks. In: NIPS (2017)
23. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR (2015)
24. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and Support inference from RGBD images. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012 Part V. LNCS, vol. 7576, pp. 746–760. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33715-4_54
25. Shen, X., et al.: Automatic portrait segmentation for image stylization. Comput. Graph. Forum (Eurographics) **35**, 93–102 (2016)
26. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR (2015)
27. Soler, C., Subr, K., Durand, F., Holzschuch, N., Sillion, F.X.: Fourier depth of field. ACM Trans. Graph. **28**, 18 (2009)
28. Wu, J., Zheng, C., Hu, X., Wang, Y., Zhang, L.: Realistic rendering of bokeh effect based on optical aberrations. Vis. Comput. **26**, 555–563 (2010)
29. Wu, Z., Huang, Y., Yu, Y., Wang, L., Tan, T.: Early hierarchical contexts learned by convolutional networks for image segmentation. In: ICPR (2014)
30. Xie, J., Girshick, R.B., Farhadi, A.: Deep3D: fully automatic 2D-to-3D video conversion with deep convolutional neural networks. In: ECCV (2016)
31. Xu, L., Ren, J., Yan, Q., Liao, R., Jia, J.: Deep edge-aware filters. In: ICML (2015)
32. Xu, X., Pan, J., Zhang, Y.J., Yang, M.H.: Motion blur kernel estimation via deep learning. TIP **27**, 194–205 (2018)
33. Xu, X., Sun, D., Pan, J., Zhang, Y., Pfister, H., Yang, M.H.: Learning to super-resolve blurry face and text images. In: ICCV (2017)
34. Yu, X., Wang, R., Yu, J.: Real-time depth of field rendering via dynamic light field generation and filtering. Comput. Graph. Forum **29**, 2099–2107 (2010)
35. Zheng, S., et al.: Conditional random fields as recurrent neural networks. In: ICCV (2015)
36. Zhou, T., Chen, J.X., Pullen, J.M.: Accurate depth of field simulation in real time. Comput. Graph. Forum **26**, 15–23 (2007)