Multi-View Motion Synthesis via Applying Rotated Dual-Pixel Blur Kernels

Abdullah Abuolaim Mahmoud Afifi Michael S. Brown
York University
{abuolaim,mafifi,mbrown}@eecs.yorku.ca

Abstract

Portrait mode is widely available on smartphone cameras to provide an enhanced photographic experience. One of the primary effects applied to images captured in portrait mode is a synthetic shallow depth of field (DoF). The synthetic DoF (or bokeh effect) selectively blurs regions in the image to emulate the effect of using a large lens with a wide aperture. In addition, many applications now incorporate a new image motion attribute (NIMAT) to emulate background motion, where the motion is correlated with estimated depth at each pixel. In this work, we follow the trend of rendering the NIMAT effect by introducing a modification on the blur synthesis procedure in portrait mode. In particular, our modification enables a high-quality synthesis of multi-view bokeh from a single image by applying rotated blurring kernels. Given the synthesized multiple views, we can generate aesthetically realistic image motion similar to the NIMAT effect. We validate our approach qualitatively compared to the original NIMAT effect and other similar image motions, like Facebook 3D image. Our image motion demonstrates a smooth image view transition with fewer artifacts around the object boundary.

1. Introduction

Unlike digital single-lens reflex (DSLR) and mirrorless cameras, smartphone cameras cannot produce a natural shallow depth of field (DoF) due to the camera's small aperture and simple optical system. Instead, many smartphones (e.g., iPhone 12, Google Pixel 4, Samsung Galaxy) emulate a shallow DoF via a *portrait mode* setting that processes the image at capture time. These methods typically isolate the subject from the background and then blur the background to emulate the swallow DoF [26]. An example is shown in the first row of Fig. 1.

Most smartphone cameras apply the synthetic bokeh effect using a common image processing framework. This traditional procedure takes an input image with minimal DoF blur and an estimated depth map to determine the blur kernel size at each pixel (i.e., defocus map). In some cases, a segmentation mask is also used to avoid blurring pixels that belong to the people and their accessories. Fig. 2 shows

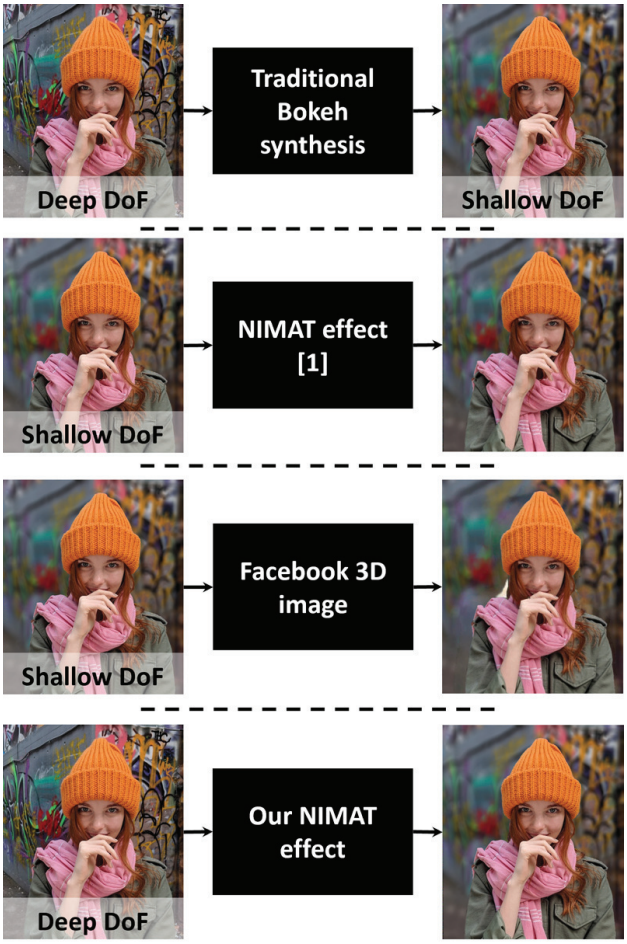


Figure 1: This figure shows a comparison between different image motion effects. We also show the output of the traditional bokeh synthesis. Our approach takes the sharp image (i.e., deep DoF) to generate the image motion. Other approaches start with the blurry input (i.e., shallow DoF) to synthesize the image motion. **Note: this figure is designed to be animated. However, the IEEE PDF eXpress validator does not allow the animation package. Therefore, we provide in-PDF animated figures in our arXiv version.**

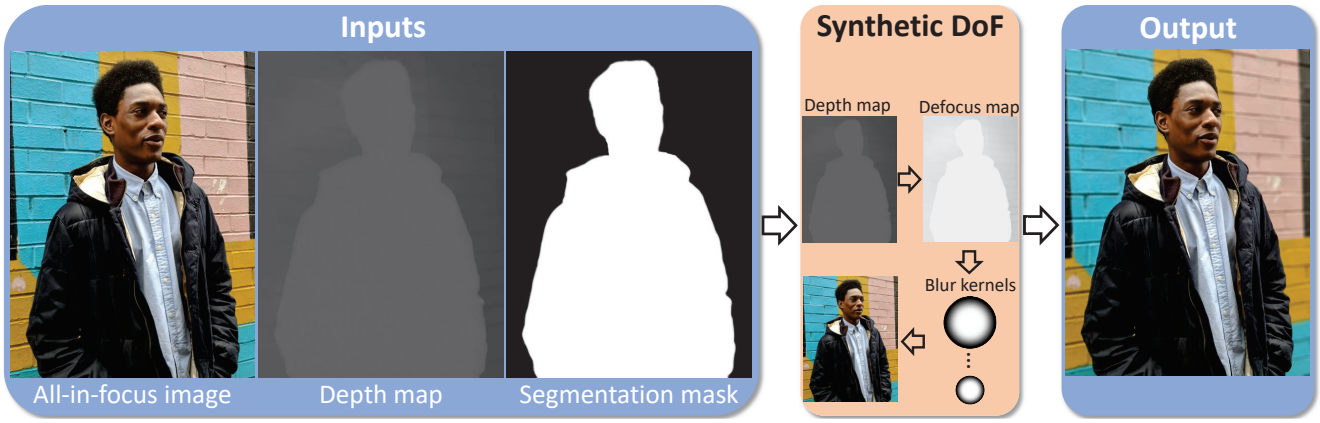


Figure 2: This figure shows a typical synthetic shallow depth of field (DoF) processing framework. This framework takes three inputs: single image, estimated depth map, and segmentation mask. Given the inputs, the synthetic DoF unit produces the desired image. The image, depth map, and segmentation mask are taken from the dataset in [26].

an illustrative example of the common synthetic bokeh framework.

Recently, Abuolaim et al. proposed a new image motion attribute (NIMAT) effect [1] that generates multiple sub-aperture views based on DoF blur and dual-pixel (DP) image formation. Abuolaim et al.'s method produces multiple views from a single input image captured by a DSLR camera and has a natural shallow DoF. Their DP- and DoF-based view synthesis is designed to generate pixel motion correlated to the defocus blur size at each pixel. However, obtaining an image with a natural shallow DoF using a smartphone camera is difficult, as mentioned earlier. Inspired by NIMAT [1], we provide a similar effect by modifying the traditional synthetic bokeh framework. Our modification enables synthesizing shallow DoF along with generating multiple views by applying a rotated blurring kernel. In our proposed framework, the defocus blur kernel shape is determined based on the sub-aperture image formation found in DP sensors. To our knowledge, we are the first to introduce this novel synthetic bokeh and DP-/DoF-based multi-view synthesis. Fig. 1 shows a comparison of different image motion approaches. It also provides the output of the traditional bokeh synthesis in the first row. Recall that other image motion approaches do not synthesize the bokeh effect. As a result, our method combines image motion and synthetic DoF into a single step. As demonstrated in Fig. 1, our image motion exhibits a smooth view transition with fewer artifacts around the object boundary compared to other approaches. Note that Fig. 1 is designed to be animated. However, the IEEE PDF eXpress validator does not allow the animation package. Therefore, we provide in-PDF animated figures in our arXiv version [1].

[1] https://arxiv.org/pdf/2111.07837.pdf

2. Related Work

Synthetic bokeh The bokeh effect in photography is an aesthetic quality of the blur that renders the main subject of the taken photo in focus while the background details fall out of focus. As mentioned earlier, standard smartphone cameras cannot produce such bokeh photographs due to the small size of the aperture and short focal length used in almost all smartphone cameras. Due to this limitation, a large body of work has targeted ways to emulate a shallow DoF image for smartphone cameras (e.g., [12,13,16,24–26,28]).

Prior methods require either up-down translation of the camera (e.g., [13]) or benefits from the parallax caused by accidental handshake during capturing (e.g., [12,28]). However, both strategies may lead to undesirable results as they rely on a specific type of movement that is not always applied in real scenarios. As a result, having low parallax limits these methods' ability to work properly.

Another strategy requires multi-image capturing, or stereo imaging, to estimate image depth from defocus cues extracted from these multiple images, or stereo pairs, of the same scene [9, 11, 24, 25, 27, 30]. However, this strategy results in ghosting effects and cannot work properly with non-static objects.

Instead of relying on multi-image capture, monocular single-image depth estimation methods are adopted to predict depth information using either inverse rendering [7,15] or supervised machine learning [8,14,18,21]. Given the estimated depth map, synthetic rendering of shallow DoF images is then a straightforward process. However, the quality of this synthetic bokeh effect is tied to the accuracy of the estimated depth map. In recent years, learning-based depth estimation methods have achieved impressive results; however, like most deep learning-based techniques, such learning depth estimators often suffer from poor generalization

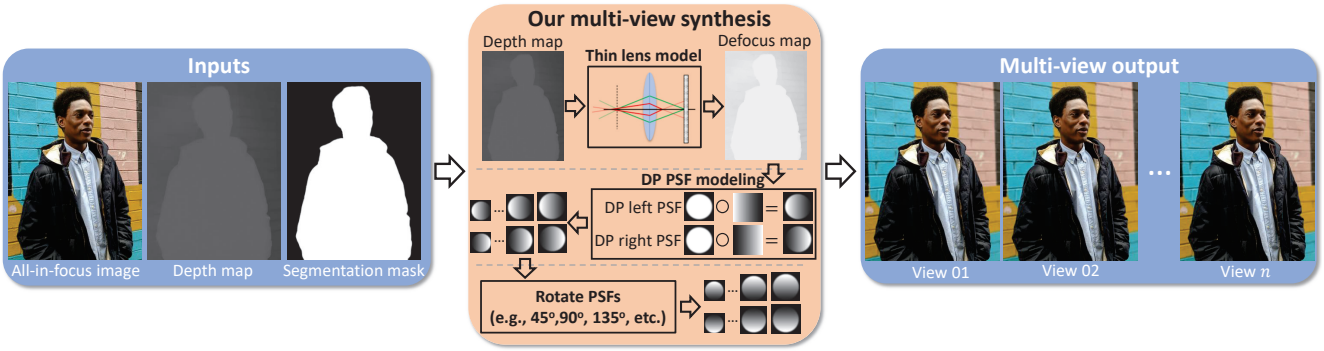


Figure 3: An overview of our proposed framework for multi-view synthesis based on rotated DP blur kernels. This framework takes three inputs: single image, estimated depth map, and segmentation mask. Given the inputs, the multi-view synthesis unit produces n views based on the number of rotated point spread functions (PSFs). The image, depth map, and segmentation mask are taken from the dataset in [26].

to images taken under conditions beyond training examples. Thus, synthesized shallow DoF images could suffer from obvious artifacts around the main object's edges.

To mitigate failure cases in single-image depth estimation, a few methods propose to replace the depth estimation process with some constraints in the scene to improve the results. For example, by dealing only with photos of people against a distant background, bokeh effects can be generated without a need for a depth map estimation [22, 23]. With this reasonable constraint, synthetic shallow DoF can be achieved by first segmenting out the human subject. This is typically performed using a trained convolutional neural network. Next, the background can be blurred using a global blur kernel. While effective, this approach assumes a constant difference in depth between the main subject (i.e., people) and the background. In addition, this approach requires a deep network to segment people from images properly.

Unlike all methods above, in this paper, our goal is to produce an image motion effect similar to the NIMAT effect [1]. A high-quality bokeh synthesis is an extra by-product output.

DP sensor DP sensors were developed as a means to improve the camera's autofocus system. The DP design produces two sub-aperture views of the scene that exhibit differences in phase that are correlated to the amount of defocus blur. Then, the phase difference between the left and right sub-aperture views of the primary lens is calculated to measure the blur amount. The phase information is also used to adjust the camera's lens such that the blur is minimized. While intended for autofocus [3, 5], the DP images have been found useful for other tasks, such as depth map estimation [10, 20, 29], defocus deblurring [2, 4, 6], and synthetic DoF [26].

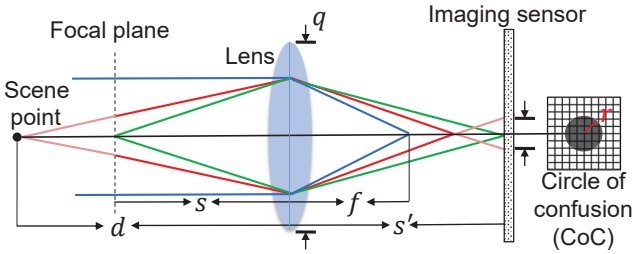


Figure 4: Thin lens model illustration and dual-pixel (DP) image formation. The circle of confusion (CoC) size is calculated for a given scene point using its distance from the lens, camera focal length, and aperture size. Note: we acknowledge that this figure was adapted from [4]

.

3. Defocus-Based Multi-View Synthesis

In this section, we describe our framework for multi-view synthesis based on rotated DP blur kernels. An overview of the proposed framework is shown in Fig. 3. First, we introduce the thin lens model used to determine the blur kernel size at each pixel. Then, the DP point spread function (PSF) is described in Sec. 3.2. Afterward, Sec. 3.3 introduces the defocus blur procedure. Lastly, Sec. 3.4 explains the process of multi-view synthesis via rotated PSFs.

3.1. PSF Size Based on the Thin Lens Model

The size of the PSFs at each pixel in the image can be calculated using the depth map. Therefore, we model camera optics using a thin lens model that assumes negligible lens thickness, helping to simplify optical ray tracing calculations [19]. This model can approximate the circle of confusion (CoC) size for a given point based on its distance from the lens and camera parameters (i.e., focal length, aperture size, and focus distance). This model is illustrated in Fig. 4,

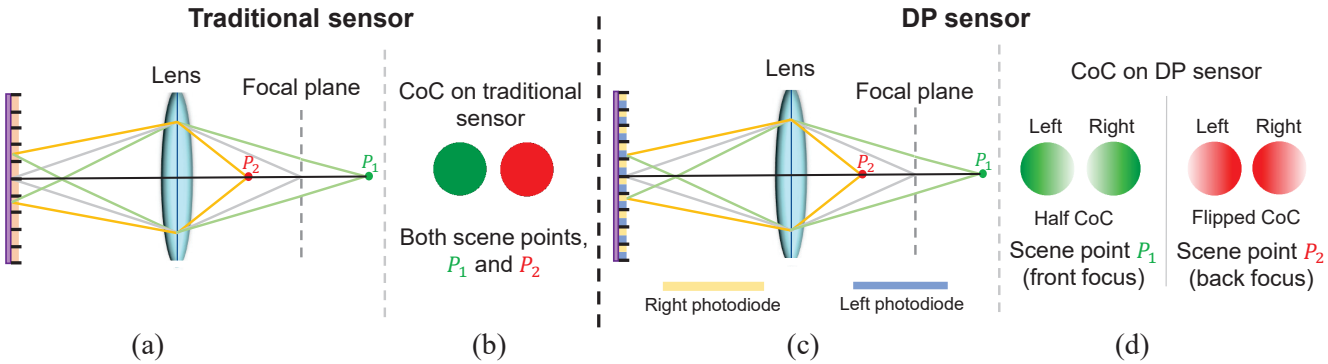


Figure 5: Circle of confusion (CoC) formation in DP sensors. (a) Traditional sensor and (c) DP sensor. (b) and (d) are the CoC formation on the 2D imaging sensor of two scene points, $P1$ and $P2$. On the two DP views, the half-CoC flips direction if the scene point is in front or back of the focal plane. Note: we acknowledge that this figure was adapted from [1].

where f is the focal length, s is the focus distance, and d is the distance between the scene point and camera lens. The distance between the lens and sensor s', and the aperture diameter q are defined as:

$$s' = \frac{f\, s}{s - f}, \tag{1}$$

$$q = \frac{f}{F}, \tag{2}$$

where F is the f-number ratio. Then, the CoC radius r of a scene point located at distance d from the camera is:

$$r = \frac{q}{2} \times \frac{s'}{s} \times \frac{d - s}{d}. \tag{3}$$

3.2. PSF Shape Based on DP Image Formation

Once the radius of the PSF is calculated at each pixel (Sec. 3.1), we need to decide the PSF shape to be applied. In this section, we adopt a DP-based PSF shape for DP view synthesis.

We start with a brief overview of DP sensors. A DP sensor uses two photodiodes at each pixel location with a microlens placed on the top of each pixel site, as shown in Fig. 5-c. This design was developed by Canon to improve camera autofocus by functioning as a simple two-sample light field camera. The two-sample light-field provides two sub-aperture views of the scene and, depending on the sensor's orientation, the views can be referred to as left/right or top/down pairs; we follow the convention of prior papers [2, 20] and refer to them as the left/right pair. The light rays coming from scene points that are within the camera's DoF exhibit little to no difference in phase between the views. On the other hand, light rays coming from scene points outside the camera's DoF exhibit a noticeable defocus disparity in the *left-right* views. The amount of defocus disparity is correlated to the amount of defocus blur.

(a) All-in-focus input (b) Our synthetic bokeh

(c) All-in-focus input (d) Our synthetic bokeh

Figure 6: Our synthetic bokeh results given an input all-in-focus image. The images used in this figure are from the synthetic DoF dataset [26].

Unlike traditional stereo, the difference between the DP views can be modeled as the latent sharp image be-

(a) Our synthetic DP views

(b) Real DP views

Figure 7: Results from our DP-view synthesis framework based on defocus blur in DP sensors. (a) Our synthetic DP views. (b) Real DP views. Our framework can produce DP views that have defocus disparity similar to the one found in real DP sensors. The image on the left is from the synthetic DoF dataset [26]. **Note: the DP views are designed to be animated. We provide in-PDF animated figures in our arXiv version.**

ing blurred in two different directions using a half-circle PSF [20]. This is illustrated in the resultant CoC of Fig. 5-d. The ideal case of a half-circle CoC on real DP sensors is only an approximation due to constraints of the sensor's construction and lens array. These constraints allow a part of the light ray bundle to leak into the other-half dual pixels (see half CoC of left/right views in Fig. 5-d).

Unlike other approaches [4, 20], we provide a simplified model of the DP PSF using a disk $\mathbf{C}$ shape that is element-wise multiplied by a ramp mask as follows:

$$\mathbf{H}_l = \mathbf{C} \circ \mathbf{M}_l, \quad \text{s.t. } \mathbf{H}_l \geq \mathbf{0}, \text{ with} \sum \mathbf{H}_l = 1, \tag{4}$$

where $\circ$ denotes element-wise multiplication, $\mathbf{M}_l$ is a 2D ramp mask with a constant intensity fall-off towards the right direction, and $\mathbf{H}_l$ is the left DP PSF. One interesting property of the DP sensors is that the right DP PSF $\mathbf{H}_r$ is the $\mathbf{H}_l$ that is flipped around the vertical axis – namely, $\mathbf{H}_l^f$:

$$\mathbf{H}_r = \mathbf{H}_l^f. \tag{5}$$

Another interesting property of the DP PSFs is that the orientation of the "half CoC" of each left/right view reveals if the scene point is in front or back of the focal plane [1, 4, 20]. Following the prior work of modeling directional blur using DP image formation, we also select the DP-based "half CoC" PSF model to capture the directional blur in this paper. However, this directional blur PSF does not have to be DP-based and can be any generic PSF that involves blurring and shifting the image content. Therefore, we test other non-DP-based directional PSF in Sec. 4.2.

3.3. Applying Synthetic Defocus Blur

In our framework, we use an estimated depth map to apply synthetic defocus blur in the process of generating a shallow DoF image. To blur an image based on the computed CoC radius r, we first decompose the image into discrete layers according to per-pixel depth values, where the maximum number of layers is set to 500 (similar to [17]). Then, we convolve each layer with the DP PSF (Sec. 4), blurring both the image and mask of the depth layer. Next, we compose the blurred layer images in order of back-to-front, using the blurred masks. For an all-in-focus input image $\mathbf{I}_s$, we generate two images – namely, the *left* $\mathbf{I}_l$ and *right* $\mathbf{I}_r$ sub-aperture DP views – as follows (for simplicity, let $\mathbf{I}_s$ be a patch with all pixels from the same depth layer):

$$\mathbf{I}_l = \mathbf{I}_s * \mathbf{H}_l, \tag{6}$$

$$\mathbf{I}_r = \mathbf{I}_s * \mathbf{H}_r, \tag{7}$$

where $*$ denotes the convolution operation. The final output image $\mathbf{I}_b$ (i.e., synthetic shallow DoF image) that is produced by the traditional *portrait mode* can be obtained as follows:

$$\mathbf{I}_b = \frac{\mathbf{I}_l + \mathbf{I}_r}{2}. \tag{8}$$

Fig. 6 shows the results of the generated synthetic bokeh image $\mathbf{I}_b$ using our proposed framework. Furthermore, our synthetically generated DP views exhibit defocus disparity similar to what we find in real DP data, where the in-focus regions show no disparity and the out-of-focus regions have defocus disparity. We provide in Fig. 7 an animated comparison between our generated DP views and real DP views extracted from a Canon DSLR camera.

(a) Facebook 3D image (b) NIMAT effect [1] (c) Our NIMAT effect

(d) Facebook 3D image (e) NIMAT effect [1] (f) Our NIMAT effect

Figure 8: A comparison between different image motion approaches. This image motion is produced by animating the synthetic output views of each approach. Two cases of scene depth variation are provided: a small depth variation in the first row and a large one in the second row. Our proposed image motion produces a pleasant motion transition and fewer artifacts compared to others. The images used in this figure are from the synthetic DoF dataset [26]. **Note: the synthetic output views are designed to be animated. We provide in-PDF animated figures in our arXiv version.**

3.4. Multi-View Synthesis

The main idea of this work is to generate multiple views from an all-in-focus image with its corresponding depth map. Therefore, we can generate an aesthetically realistic image motion by synthesizing a multi-view version of a given single image. As discussed in Sec. 3.2, the DP two sub-aperture views of the scene depending on the sensor's orientation and, in this work, our formation contain left/right DP pairs, and consequently, our framework synthesizes the horizontal DP disparity as shown in Fig. 7. We can synthesize additional views with different "DP dispar-

ity" by rotating the PSFs during the multi-view synthesis process as shown in Fig. 3. For example, eight views can be generated by performing a 45^o clockwise rotation step three times (i.e., 45^o, 90^o, 135^o). Then, we generate our effect by alternating the output views to produce the image motion.

4. Experiments

4.1. Results Using DP PSF

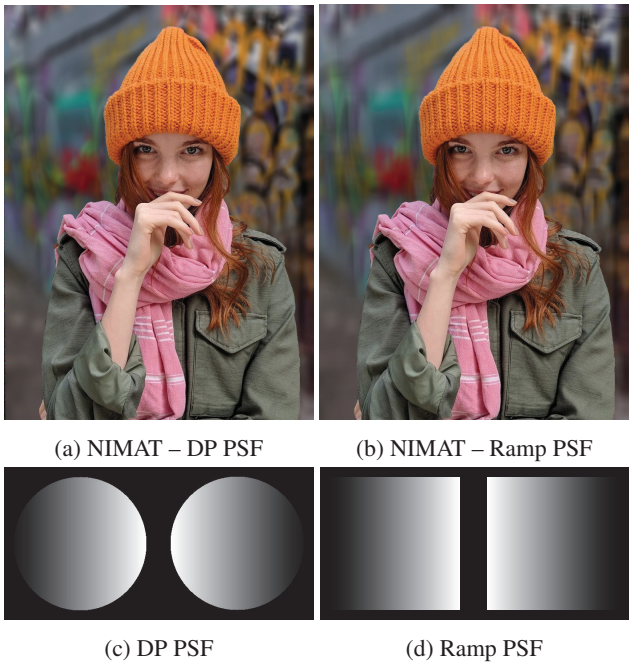
Following the qualitative comparison procedure introduced in [1], we provide the animated image motion (or NIMAT effect) of different approaches in Fig. 8. In particular, we compare ours with the results from [1] and the Facebook 3D image. As mentioned earlier and unlike other approaches, our proposed framework starts with the deep DoF image (i.e., almost all-in-focus) to produce the synthetic bokeh (or synthetic shallow DoF) image and the multiple DoF/DP-based views. Therefore, we provide the synthetic bokeh image as input to other approaches. This section also introduces the NIMAT-like effect from the common Facebook 3D image by uploading a single image and rendering the 3D version. Then, we save multiple frames at different view directions following the circular pixel motion transition found in the NIMAT effect [1].

The results in this section show two cases of scene depth variations — namely, a small depth variation (Fig. 8, first row) and a large one (Fig. 8, second row). While the Facebook 3D image motion is sufficient in the first row, it suffers from few artifacts around the foreground object boundary (e.g., the wall behind the person's head and arm). As for the NIMAT effect results from [1] in the first row, the image motion is barely noticeable in the background due to the small blur size that is a result of the small scene depth variation.

The second row of Fig. 8 shows the large depth variation case, where the blur size varies from small to large. In this case, the Facebook 3D image exhibits noticeable and unpleasing artifacts (e.g., missing pixels). While the NIMAT effect from [1] produces pleasing image motion, we can still spot few artifacts that do not exist in ours. Note that we are aware the Facebook 3D image is not made for the same purpose, but we rendered it with the same motion transition settings of the NIMAT effect for comparison purposes.

4.2. Results Using Other PSFs

As mentioned earlier in Sec. 3.2, the directional PSF used to render the NIMAT effect can be any generic PSF that involves blurring and shifting the image content. In Fig. 9, we show the NIMAT effect rendered using two different PSF shapes – namely, DP-based PSF (Fig. 9, c) and transitional blurring 2D ramp mask with a constant intensity fall-off towards the opposite direction (i.e., Ramp PSF in Fig. 9, d). These results demonstrate that other non-DP-based PSF can be utilized to render the NIMAT effect as long as it satisfies the conditions of having a transnational and blurring operator. Nevertheless, different motion transitions and artifacts can be introduced by using different PSFs as shown in Fig. 9.



(a) NIMAT – DP PSF (b) NIMAT – Ramp PSF

(c) DP PSF (d) Ramp PSF

Figure 9: A comparison between different PSFs used to render the NIMAT effect. The two PSFs (i.e., DP PSF and Ramp PSF) are able to render smooth image motion. However, different motion transitions and artifacts can be introduced by using different PSFs. **Note: the synthetic output views are designed to be animated. We provide in-PDF animated figures in our arXiv version.**

5. Conclusion

In this work, we proposed a modification to the DoF synthesis associated with most smartphones' *portrait mode* feature. This modification can be easily integrated into the traditional DoF synthesis unit and enables the generation of multiple sub-aperture views along with the synthetic bokeh photo. With this modification, we are also able to produce an aesthetic image motion effect similar to the novel NIMAT effect from [1]. For our multi-view synthesis, we introduced the novel idea of convolving the input image with the rotated blurring kernels based on the DoF blur and DP image formation. We validated our approach qualitatively and demonstrated that it produces smooth motion transition in the NIMAT effect with fewer artifacts compared to others. We aim to encourage work in this new research direction that presented a new pleasing effect of image motion.

References

[1] Abdullah Abuolaim, Mahmoud Afifi, and Michael S Brown. Improving single-image defocus deblurring: How dual-pixel images help through multi-task learning. In *WACV*, 2022.

[2] Abdullah Abuolaim and Michael S Brown. Defocus deblurring using dual-pixel data. In *ECCV*, 2020.

[3] Abdullah Abuolaim and Michael S Brown. Online lens motion smoothing for video autofocus. In *WACV*, 2020.

[4] Abdullah Abuolaim, Mauricio Delbracio, Damien Kelly, Michael S Brown, and Peyman Milanfar. Learning to reduce defocus blur by realistically modeling dual-pixel data. In *ICCV*, 2021.

[5] Abdullah Abuolaim, Abhijith Punnappurath, and Michael S Brown. Revisiting autofocus for smartphone cameras. In *ECCV*, 2018.

[6] Abdullah Abuolaim, Radu Timofte, and Michael S Brown. Ntire 2021 challenge for defocus deblurring using dual-pixel images: Methods and results. In *CVPR Workshops*, 2021.

[7] Jonathan T Barron and Jitendra Malik. Shape, illumination, and reflectance from shading. *TPAMI*, 37(8):1670–1687, 2014.

[8] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *arXiv preprint arXiv:1406.2283*, 2014.

[9] Ravi Garg, Vijay Kumar Bg, Gustavo Carneiro, and Ian Reid. Unsupervised CNN for single view depth estimation: Geometry to the rescue. In *ECCV*, 2016.

[10] Rahul Garg, Neal Wadhwa, Sameer Ansari, and Jonathan T Barron. Learning single camera depth estimation using dual-pixels. In *ICCV*, 2019.

[11] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, 2017.

[12] Hyowon Ha, Sunghoon Im, Jaesik Park, Hae-Gon Jeon, and In So Kweon. High-quality depth from uncalibrated small motion clip. In *CVPR*, 2016.

[13] Carlos Hernández. Lens blur in the new google camera app. `http://research.googleblog.com/2014/04/lens-blur-in-new-google-camera-app.html`, 2014.

[14] Derek Hoiem, Alexei A Efros, and Martial Hebert. Automatic photo pop-up. In *SIGGRAPH*. 2005.

[15] Berthold KP Horn. Obtaining shape from shading information. *The psychology of computer vision*, pages 115–155, 1975.

[16] Andrey Ignatov, Jagruti Patel, and Radu Timofte. Rendering natural camera bokeh effect with deep learning. In *CVPR Workshops*, 2020.

[17] Junyong Lee, Sungkil Lee, Sunghyun Cho, and Seungyong Lee. Deep defocus map estimation using domain adaptation. In *CVPR*, 2019.

[18] Fayao Liu, Chunhua Shen, Guosheng Lin, and Ian Reid. Learning depth from single monocular images using deep convolutional neural fields. *TPAMI*, 38(10):2024–2039, 2015.

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

[20] Abhijith Punnappurath, Abdullah Abuolaim, Mahmoud Afifi, and Michael S Brown. Modeling defocus-disparity in dual-pixel sensors. In *ICCP*, 2020.

[21] Ashutosh Saxena, Min Sun, and Andrew Y Ng. Learning 3-d scene structure from a single still image. In *ICCV*, 2007.

[22] 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.

[23] Xiaoyong Shen, Xin Tao, Hongyun Gao, Chao Zhou, and Jiaya Jia. Deep automatic portrait matting. In *ECCV*, 2016.

[24] Supasorn Suwajanakorn, Carlos Hernandez, and Steven M Seitz. Depth from focus with your mobile phone. In *CVPR*, 2015.

[25] Huixuan Tang, Scott Cohen, Brian Price, Stephen Schiller, and Kiriakos N Kutulakos. Depth from defocus in the wild. In *CVPR*, 2017.

[26] 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*, 37(4):64, 2018.

[27] Junyuan Xie, Ross Girshick, and Ali Farhadi. Deep3d: Fully automatic 2d-to-3d video conversion with deep convolutional neural networks. In *ECCV*, 2016.

[28] Fisher Yu and David Gallup. 3d reconstruction from accidental motion. In *CVPR*, 2014.

[29] Yinda Zhang, Neal Wadhwa, Sergio Orts-Escolano, Christian Häne, Sean Fanello, and Rahul Garg. Du2Net: Learning depth estimation from dual-cameras and dual-pixels. In *ECCV*, 2020.

[30] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *CVPR*, 2017.