Faking It:
Simulating Background Blur in Portrait Photography using a Coarse Depth Map Estimation from a Single Image

Nadine Friedrich Oleg Lobachev Michael Guthe
University Bayreuth, AI5: Visual Computing, Universitätsstraße 30, D-95447 Bayreuth, Germany

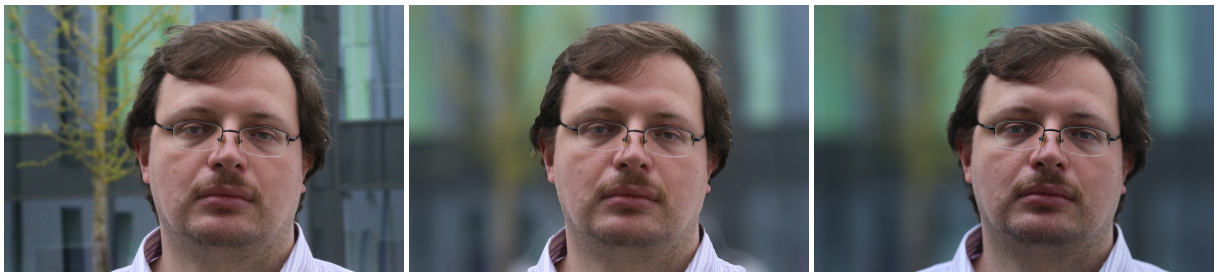


Figure 1: Our approach vs. a real image with bokeh. Left: input image, middle: result of our simulation, right: gold standard image, captured with the same lens as the input image, but with a large aperture, yielding natural background blur.

ABSTRACT

In this work we simulate background blur in photographs through a coarse estimation of a depth map. As our input is a single portrait picture, we constraint our objects to humans first and utilise skin detection. A further extension alleviates this. With auxiliary user input we further refine our depth map estimate to a full-fledged foreground–background segmentation. This enables the computation of the actual blurred image at the very end of our pipeline.

Keywords

bokeh, background blur, depth map, foreground–background segmentation

1 INTRODUCTION

High-quality portrait photography often features a special kind of background blur, called bokeh. Its nature originates from the shape of camera lenses, aperture, distance to background objects, and their distinctive light and shadow patterns. This effect is thus used for artistic purposes, it separates the object the lens is focused on from the background and helps the viewer to concentrate on the foreground object—the actual subject of the photograph.

We do not render a depth-of-field blur in a 3D scene, but pursue a different approach. Our input is a single 2D image without additional data—no depth field, no IR channel, no further views. Of course, a full 3D reconstruction is impossible in this case. But how could additional information help?

We restrict our choice of pictures to portraits of humans (though, Figs. 7 and 8 try out something different). We know, the image has a foreground where typically our human is pictured, and background that we would like to segment out and blur. We detect human skin colour for initialisation and engage further tricks—including user annotations—we detail below to find the watershed between foreground and background.

The central contribution of this work is the way how we combine skin detection, user annotations, and edge-preserving filters to obtain bluring masks, the coarse depth maps from a single image.

The next section handles related work, Section 3 presents our method, Section 4 shows the results, Section 5 presents the discussion, Section 6 concludes.

2 RELATED WORK

One of the first approaches for simulating bokeh effect were Potmesil and Chakravarty [PC81]; Cook [Coo86]. Most typical simulations of camera background blur

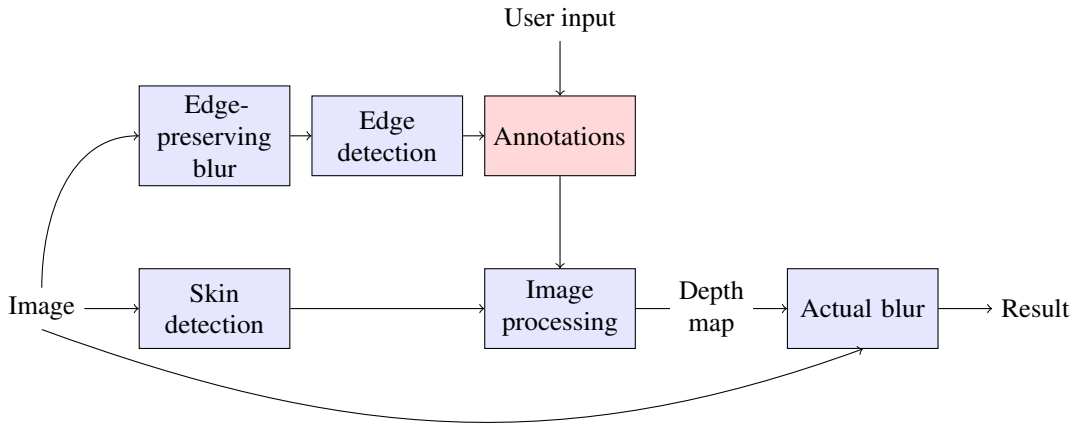


Figure 2: An overview of our approach. Everything that has skin colour is detected as foreground, then we add everything else where the user input matches on an image blurred in an edge-preserving manner. The different results are combined to a single mask. The mask and original input image are the input for bokeh simulation.

base on a full-fledged 3D scene, some of more recent methods are Wu et al. [Wu+12]; Moersch and Hamilton [MH14]. Yu [Yu04]; Liu and Rokne [LR12]; McIntosh, Riecke, and DiPaola [MRD12] discuss bokeh effect as a post-processing technique in rendering. This is different from our approach.

Nasse [Nas10] provides a nice technical overview of the bokeh effect. Sivokon and Thorpe [ST14] are concerned with bokeh effects in aspheric lenses.

Yan, Tien, and Wu [YTW09] are most similar to our approach, as they are concerned not only with bokeh computation, but also with foreground–background segmentation. They use a technique called "lazy snapping" [Li+04], we discuss the differences to our approach in Section 5.4.

A lot of research focuses on how to compute a realistic bokeh effect, given an image and its depth map, (see, e.g., [BFSC04]) It is in fact wrong to use a Gaussian blur (like [GK07] do) as the resulting image is too soft.

Lanman, Raskar, and Taubin [LRT08] capture the characteristics of bokeh and vignetting using a regular calibration pattern and then apply these data to further images. We rely on McGraw [McG14] in the actual bokeh *computation* from input data and estimated depth maps, which is a much more synthetic method as detailed below. This work actually focuses on obtaining the mask, "what to blur" from a single 2D image.

Bae and Durand [BD07] estimate an existing de-focus effect on images made with small sensors and amplify it to simulate larger sensors. This includes both the estimation of the depth map and the generation of a shallow depth-of-field image. Motivation of this work is very similar to ours, but the method is completely different. They estimate existing small defocus effects from the image and then amplify them using Gaussian blur.

Notably, Zhu et al. [Zhu+13] do the reverse of our approach. We estimate with some assumptions about the images and further inputs the foreground–background segmentation to compute then the depth-of-field effect. Zhu et al. estimate the foreground–background segmentation from shallow depth-of-field images. Works like Zhang and Cham [ZC12] concentrate on "refocusing," i.e., on detecting unsharp areas in a picture and on making the unsharp areas more sharp.

Saxena, Chung, and Ng [SCN07] present a supervised learning approach to the depth map estimation. This is different from our method. Saxena, Chung, and Ng divide the visual clues in the image into relative and absolute depth clues—evidences for difference of depth between the patches or for an "actual" depth. They use then a probabilistic model to integrate the clues to a unified depth image. This work does not focus on the computation of the shallow depth-of-field image. Eigen, Puhrsch, and Fergus [EPF14] use deep learning technique. A sophisticated neural network is trained on existing RGB+D datasets and evaluated on a set of other images from the same datasets. This is radically different from our approach. Aside from the presence of humans in the picture we make no further assumptions and utilize no previously computed knowledge. We have to use some auxiliary user input though. Eigen, Puhrsch, and Fergus [EPF14] also do not focus on the generation of shallow depth-of-field image.

3 METHOD

We chain multiple methods. First, the foreground mask expands to everything in the input image that has a skin colour. This way, we identify hands and other body parts showing skin. We expand the selection by selecting further pixels of the similar colour in the vicinity of already selected ones—we need to select all the skin, not just some especially good illuminated parts.

However, all this does not help with selection of clothes, as it can be of any colour or shape, a further problem is hair. For this sake we have allowed user input for the annotations of definitely foreground and definitely background areas. An attempt to expand the annotation (à la "magic brush" selection in photo-editing software) based on the actual input image would result in too small "cells" on some occasions and hence too much hysteresis—think: canny edge detection. For this reason we apply an edge preserving blur to the image used as input for "magic brush." This ensures higher-quality depth maps, separating the foreground (actual subject) and background. Given the depth map and initial input image, we apply the method of McGraw [McG14] to obtain the actual blurred image.

The "cells" we have mentioned above are actually regions with higher frequency than elsewhere in the image, that is: regions where edge detection would find a lot of edges. We futher discuss this issue in Section 5.3. An overview of our pipeline is in Figure 2.

3.1 Parts of our pipeline

Filtering approaches increase the edge awareness of our estimation. We use egde-preserving filtering [BYA15] as a part of our pipeline. Skin detection [EMH15] was part of our pipeline (see also [Bra98]). The depth maps were also processed with standard methods like erosion and dilation.

3.2 Neighbourhood detection

To detect similar-coloured pixels in the vicinity of pixels already present in the mask, we used the von Neumann neighbourhood (i. e., 4-connected). We used HSV colour space, the folklore solution for human skin detection. A naive implementation evidenced hysteresis: a pixel is deselected as it is deemed as background, but it is selected again because it has a similar colour as foreground. To amend this problem, we utilised canny edge detection on the image after edge-preserving blur. This reduces the number of falsely detected small edges. Now, in the von Neumann neighbourhood computation we check additionally if a pixel or its neighbours are on the edge. It is the case, we exclude these pixels from further processing.

3.3 The pipeline executed (Fig. 3)

Figure 3 demonstrates the processing steps on an example image (a). Fig. (b) shows the result of edge-preserving blur, the edge detection applied to it yields (d). Some parts of the image are already selected via skin detection (c). Basing on edges and user input, a full shape can be selected (e). We do not limit our approach to a single shape and to foreground only, as (f) shows. These intermediate results are then processed with erosion and dilation image filters, yielding (g). This final depth map is then applied to the input image (a) using the method of McGraw [McG14]. The final result is in (h).

4 RESULTS

4.1 Selfies

Our method works best on selfie-like images. Such images typically feature relatively large subject heads, further selfies are mostly captured on a mobile phone, thus they have a large depth-of-field. This fact makes them very suitable for an artistic bokeh simulation that is impossible to achieve with hardware settings in this case.

The input and reference images in Figure 1 were shot on a Canon 6D full-frame camera at 200 mm focal distance. To mimic the large depth-of-field of lesser cameras, the input image was captured at $f/32$, the reference image was captured at $f/4$ to showcase the real bokeh effect. The images were produced with Canon EF 70–200 mm $f/4$L lens. Our method works fine also when the head is relatively smaller in the whole picture (Fig. 4).

Featuring more than one person in a photograph is not a problem for our method, as Fig. 5 shows.

4.2 Multiple depths

Our depth maps facilitate not only a foreground–background segmentation, as showcased in Figs. 3, 6, and 7.

The input for Figure 6 was captured on a mobile phone and because of small sensor size it features a greater depth of field. Porting out application to mobile phones might be a promising way of using it. Fig. 7 also features multiple depth levels, we discuss it below.

5 DISCUSSION

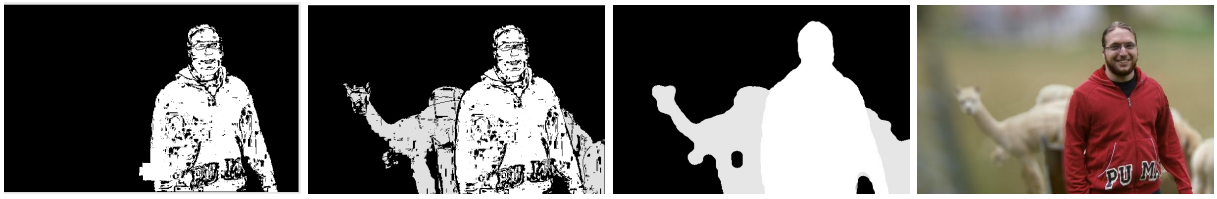
We discuss following issues: how our method performs on non-human subjects of a photograph (Sec. 5.1), the issues with thin locks of hair (Sec. 5.2), we give more details on the cases when edge detection does not perform well (Sec. 5.3). Then we compare our method to "lazy snapping" (Sec. 5.4) and the result of our method to a real photograph with bokeh effect (Sec. 5.5).

5.1 Non-humans

We applied our method to Figs. 7 and 8. Naturally, no skin detection was possible here. The masks were created with user annotations on images after edge-preserving blur with canny edge detection as separator for different kinds of objects.

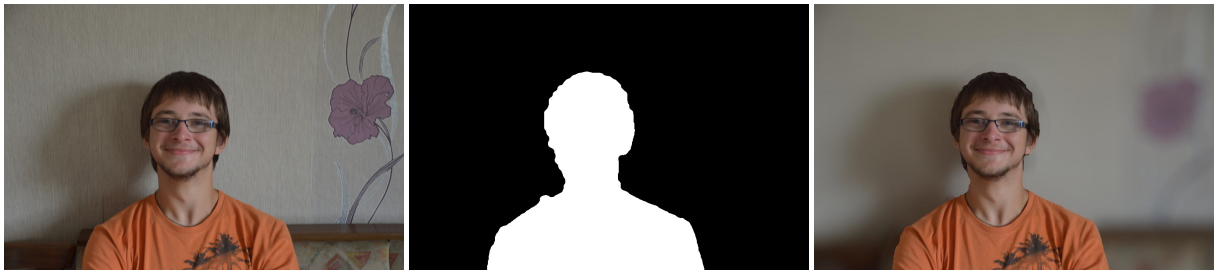
Note that in both examples, in case of the real shallow depth of field image, the table surface (Fig. 7) or soil (Fig. 8) would feature an area that is in-focus, as the focal plane crosses the table top or the ground. This is not the case in our images, as only the relevant objects were selected as foreground. Of course, it would be easy to simulate this realistic bokeh effect using a simple further processing of the depth map.

(a) Input image (b) Result of edge-preserving blur (c) Skin detection (d) Canny edges



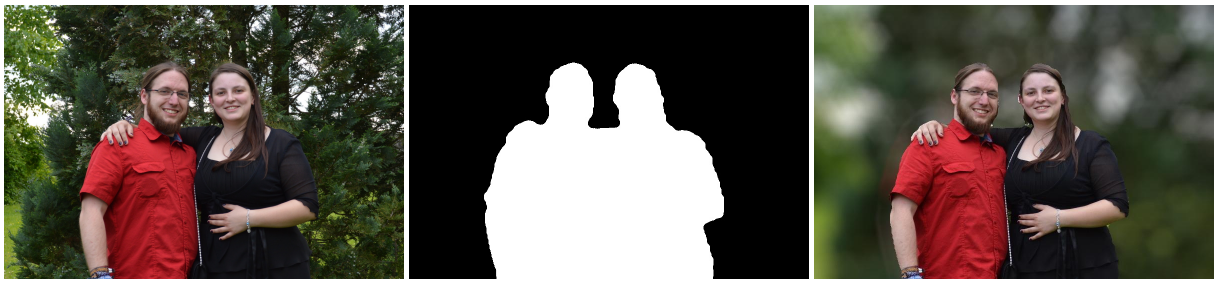
(e) Depth map, an intermediate state (f) Adding a further level to the depth map, an intermediate state (g) Final depth map (h) Final result

Figure 3: Results of various intermediate steps of our pipeline. Input image (a) was captured at 27 mm full-frame equivalent at $f/2.8$ on a compact camera with crop factor 5.5. The binary foreground–background segmentation mask is in Fig. (g), final result with bokeh effect applied is in (h).



(a) Input image (b) Mask (c) Result

Figure 4: Filtering an image with head and shoulders. Input image (a) was captured using 57 mm full-frame equivalent lens at $f/4.5$ with crop factor 1.5.



(a) Input image (b) Mask (c) Result

Figure 5: Two persons in a photograph. Input image was captured at 43 mm focal distance equivalent on a full-frame, $f/5.6$, crop factor 1.5.

5.2 Hair

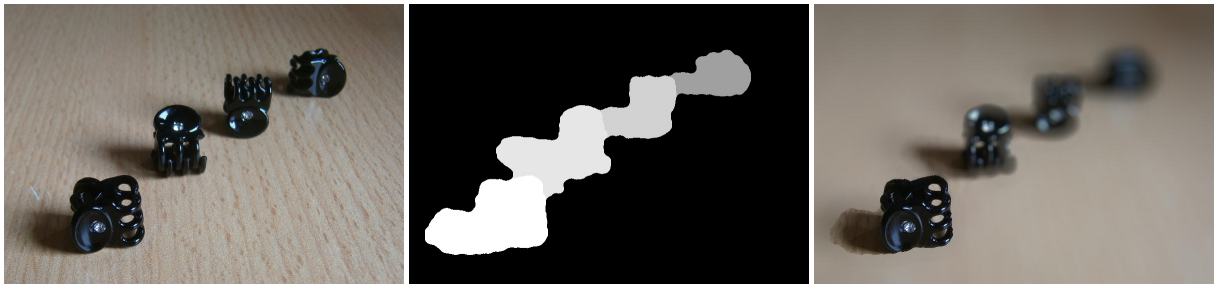
Thin flocks of hair cannot be easily detected, esp. on a nosily background. Automatic or annotation-based selection of such hair parts features a larger problem. Naturally, anything not present in the foreground selection enjoys background treatment during the actual bokeh simulation. One of most prominent visuals for such a side effect is Figure 9, even though some other our examples also showcase this issue.

5.3 Obstacles for edge detection

We use canny edge detection after an edge-preserving blur to separate "meaningful" edges from nonsense ones. This is basically the object segmentation that determines the boundaries of "cells" on which user annotations act. If an image features a lot of contrasts that survive the blur per Badri, Yahia, and Aboutajdine [BYA15], the user would require to perform more interactions than desired, as the intermediate result features too many

(a) Input image (b) Mask (c) Result

Figure 6: Showcasing more than a foreground and background separation. Input image captured on a mobile phone. The big plant on the left has a further depth level assigned.



(a) Input image (b) Mask (c) Result

Figure 7: Showcasing more than a foreground and background separation. This image has no humans on it. Input image (a) was captured at 27 mm full-frame equivalent at $f/2.8$ on a compact camera with crop factor 5.5.

"cells." Figure 10 illustrates this issue. Of course, a fine-tuning of edge-preserving blur parameters would alleviate this problem. However, we did not want to give our user any knobs and handles besides the quite intuitive input method for the "cell" selection, i. e., the annotations as such.

5.4 Comparison to lazy snapping

Yan, Tien, and Wu [YTW09] use lazy snapping [Li+04] and face detection for the segmentation. They typically produce gradients in their depth maps, to alleviate the issue we mentioned above in Section 5.1.

Lazy snapping uses coarse user annotations, graph cut, and fine-grain user editing on the resulting boundaries. In a contrast, we apply skin detection and edge detection on images blurred in an edge-preserving manner. The cells after edge detection are then subject to user annotations. We do not allow fine-grain editing of boundaries and thus drastically reduce the amount of user input, we are basically satisfied with coarse user annotations.

5.5 Comparison to real bokeh

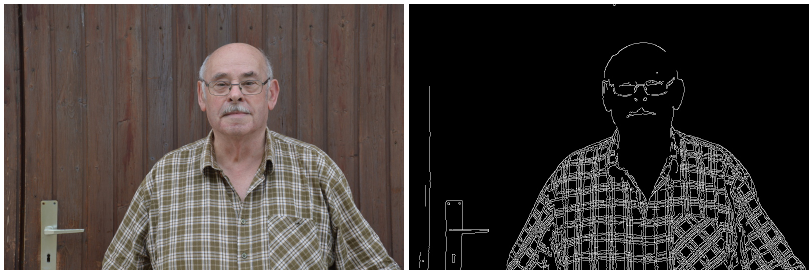
Compare images in the middle (our approach) and on the right hand side (ground truth) of Figure 1. We see a sharper edge in the hair, similarly to the issue discussed above. There is also a strange halo effect around the collar of the shirt. A further refinement and processing of the depth map data could help. Aside from these issues, the bokeh effect itself is represented quite faithfully. In an interesting manner, our synthetic image appears to be more focusing on the subject than the ground truth image. A possible reason is: the whole subject in our version is sharp. The ground truth version focuses on the eyes, but parts of the subject are already unsharp due to a too shallow depth-of-field: see shirt collar or the hair on the left. As our version is based on an image with a large depth-of-field (Fig. 1, left), it does not have these issues.

(a) Input image (b) Mask (c) Result

Figure 8: Applying our method to a photograph of a dog. By definition, no skin detection was possible. Captured on a mobile phone.

(a) Input image (b) Mask (c) Result

Figure 9: Limitation of our method: hair. Notice how some hair locks are missing in the mask and are blurred away. Captured at 69 mm full-frame equivalent at $f/4.8$ with crop factor 1.5.



(a) Input image (b) Canny edges

Figure 10: Limitation of our method: obstacles for edge detection. Input image (a) was captured at 82 mm full-frame equivalent at $f/6.3$ with crop factor 1.5. Note how the plaid shirt forms separate cells after canny edge detection (b), necessitating a larger annotation.

6 CONCLUSIONS

We have combined skin detection with user annotations to facilitate a coarse depth map generation from a single 2D image without additional modalities. The user input was processed on an extra layer after edge-aware blurring. In other words, we have enabled foreground–background separation through image processing and computer vision techniques and minimal user input. The resulting depth maps were then subsequently used to process the input image with a simulation of out-of-focus lens blur. Combined, we create a well-known lens effect ("bokeh") from single-image 2D portraits.

Future work

A mobile phone-based application might be of an interest, considering the selfie boom. Some UI tweaks like a fast preview loop after each user input and general performance improvements might be helpful in this case.

Face detection could be useful in general and for better handling of hair—we would use different parameters in the pipeline around the head, i. e., for hair, than everywhere else. Correct hair selection is probably the best area to further improve our work.

Further, our application benefits from any improvements in skin detection, edge-preserving blur, or bokeh simulation.

7 ACKNOWLEDGEMENTS

We would like to thank the photographers R. Friedrich, J. Kollmer, and K. Wölfel. Both the photographers and the models agreed that their pictures may be used, processed, and copied for free.

We thank T. McGraw, E. S. L. Gastal, M. M. Oliveira, H. Badri, H. Yahia, and D. Aboutajdine for being able to use their code.

REFERENCES

[BD07] S. Bae and F. Durand. Defocus magnification. *Comput. Graph. Forum*, 26(3):571–579, 2007.

[BFSC04] M. Bertalmio, P. Fort, and D. Sanchez-Crespo. Real-time, accurate depth of field using anisotropic diffusion and programmable graphics cards. In *3D data processing, visualization and transmission*, 2004, pages 767–773.

[Bra98] G. R. Bradski. Conputer vision face tracking for use in a perceptual user interface. *Intel technology journal*, 1998.

[BYA15] H. Badri, H. Yahia, and D. Aboutajdine. Fast edge-aware processing via first order proximal approximation. *IEEE T. Vis. Comput. Gr.*, 21(6):743–755, 2015.

[Coo86] R. L. Cook. Stochastic sampling in computer graphics. *ACM T. Graphic.*, 5(1):51–72, 1986.

[EMH15] A. Elgammal, C. Muang, and D. Hu. Skin detection. In, *Encyclopedia of Biometrics*, pages 1407–1414. Springer, 2015.

[EPF14] D. Eigen, C. Puhrsch, and R. Fergus. Depth map prediction from a single image using a multi-scale deep network. In, *Adv. Neur. In.* Volume 27, pages 2366–2374. Curran, 2014.

[GK07] J. Göransson and A. Karlsson. Practical post-process depth of field. *GPU Gems*, 3:583–606, 2007.

[Li+04] Y. Li, J. Sun, C.-K. Tang, and H.-Y. Shum. Lazy snapping. *ACM T. Graphic.*, 23(3):303–308, 2004.

[LR12] X. Liu and J. Rokne. Bokeh rendering with a physical lens. In *PG '12 Short proc.* EG, 2012. ISBN: 978-3-905673-94-4.

[LRT08] D. Lanman, R. Raskar, and G. Taubin. Modeling and synthesis of aperture effects in cameras. In. In COMPAESTH '08. EG, 2008. ISBN: 978-3-905674-08-8.

[McG14] T. McGraw. Fast bokeh effects using low-rank linear filters. *Visual Comput.*, 31(5):601–611, 2014.

[MH14] J. Moersch and H. J. Hamilton. Variable-sized, circular bokeh depth of field effects. In *Graphics Interface '14*. CIPS, 2014, pages 103–107.

[MRD12] L. McIntosh, B. E. Riecke, and S. DiPaola. Efficiently simulating the bokeh of polygonal apertures in a post-process depth of field shader. *Comput. Graph. Forum*, 31(6):1810–1822, 2012.

[Nas10] H. H. Nasse. Depth of field and bokeh. *Carl Zeiss camera lens division report*, 2010.

[PC81] M. Potmesil and I. Chakravarty. A lens and aperture camera model for synthetic image generation. *SIGGRAPH Comput. Graph.*, 15(3):297–305, 1981.

[SCN07] A. Saxena, S. H. Chung, and A. Y. Ng. 3-D depth reconstruction from a single still image. *Int. J. Comput. Vision*, 76(1):53–69, 2007.

[ST14] V. P. Sivokon and M. D. Thorpe. Theory of bokeh image structure in camera lenses with an aspheric surface. *Opt. Eng.*, 53(6):065103, 2014.

[Wu+12] J. Wu, C. Zheng, X. Hu, and F. Xu. Rendering realistic spectral bokeh due to lens stops and aberrations. *Visual Comput.*, 29(1):41–52, 2012.

[YTW09] C.-Y. Yan, M.-C. Tien, and J.-L. Wu. Interactive background blurring. In. In MM '09. ACM, 2009, pages 817–820.

[Yu04] T.-T. Yu. Depth of field implementation with OpenGL. *J. comput. sci. coll.*, 20(1):136–146, 2004. ISSN: 1937-4771.

[ZC12] W. Zhang and W.-K. Cham. Single-image refocusing and defocusing. *IEEE T. Image Process.*, 21(2):873–882, 2012.

[Zhu+13] X. Zhu, S. Cohen, S. Schiller, and P. Milanfar. Estimating spatially varying defocus blur from a single image. *IEEE T. Image Process.*, 22(12):4879–4891, 2013.