

DC²: Dual-Camera Defocus Control by Learning to Refocus

Supplementary Material

A. Video Visualization

One major advantage of our method is the fine-grained control we can have on the defocus control. As a result, we can directly simulate changing the focus distance and aperture smoothly just like if we had a DSLR camera with variable focal length and aperture. Please refer to the video provided in the supplementary materials.

B. Detailed Architecture

The model architecture consists of three primary modules: Φ_{ref}^W to refine \mathbf{W} , Φ_{ref}^{UW} to refine \mathbf{UW} , and a fusion model Φ_{fusion} to predict a blending mask to blend the refined outputs. Both Φ_{ref}^W and Φ_{ref}^{UW} use DRBNet architecture [3] that utilize kernels prediction to refine the input. Each refinement module predicts intermediate outputs in a multi-scale setup that can be used to speed up training. Specifically, the model generates refined outputs at the following scales: 8x downsampled, 4x downsampled, 2x downsampled, and the original resolution. To be able to fuse all the multi-scale outputs, Φ_{fusion} consists of several Atrous Spatial Pyramid Pooling (ASPP) convolutions blocks [2] to predict blending mask for each scale. The ASPP blocks for each scale take the refined \mathbf{W} and \mathbf{UW} of the associated scale, as well as an upsampled blending mask from the previous ASPP block with a residual connection of the upsampled mask (except for the first ASPP block since it has no preceding blending mask). There are two hyperparameters associated with the blending block for each scale: (1) atrous rates for the atrous convolutions, and (2) the number of channels each intermediate step of atrous convolutions outputs. In table 1, we include a list of the hyperparameters for the blending block associated with each scale.

One issue with training the model in using cropped patches is that the blur kernel is spatially varying depending on the crop position. To resolve the ambiguity, we follow the solution proposed by Abuolaim *et al.* [1] and concatenate a radial mask to the inputs of all modules where the pixel values of the mask are the distance from the original image center, normalized.

Table 1. **Fusion model (Φ_{fusion}) hyperparameters.** The hyperparameters for the ASPP convolution blocks are the atrous rates for the atrous convolutions, and the channels each layer outputs. The number of atrous convolution layers is the size of the channel list. Note that the final output consists of two channels which correspond to the \mathbf{W} and \mathbf{UW} blending masks.

Blending Block	atrous rates	channels
1/8x scale	1,3,5	16, 32, 2
1/4x scale	1, 3, 6, 12	16, 32, 2
1/2x scale	1, 3, 6, 12, 15	16, 32, 2
1x scale	1,3,6, 12, 15, 18	16, 32, 32, 2

C. Model Analysis

The primary motivation behind our architecture design is, depending on the target defocus map, the network can choose to deblur/blur parts of \mathbf{W} and transfer sharper details from \mathbf{UW} if necessary. Due to our model design, we can directly visualize the intermediate outputs to understand the model behavior. Specifically, we can visualize the refined \mathbf{W} and \mathbf{UW} which are the outputs of Φ_{ref}^W and Φ_{ref}^{UW} , as well as the blending masks predicted by the fusion network Φ_{fusion} . We visualize the intermediate outputs of our method on the task of all-in-focus deblurring in Figures 1 and 2. Note that in both examples, the mask associated with \mathbf{UW} has large values around edges and regions with high frequencies, while the mask for \mathbf{W} has higher values in low frequency regions. This supports our hypothesis of having \mathbf{UW} serve for high frequency details that could be



Aligned UW



Refined UW



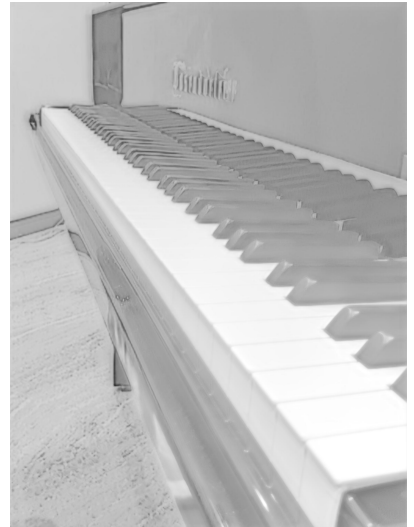
UW blending mask



Ref. W



Refined W



W blending mask

Figure 1. **Intermediate results visualization.** Note that the whitebalance is off in UW, but the refinement module does not get affected by that since it primarily preserves the high frequencies in refined UW. In the refinement of W, we notice that the model deblurs the edges and preserves the low-frequency signals that can be blended with the details from UW

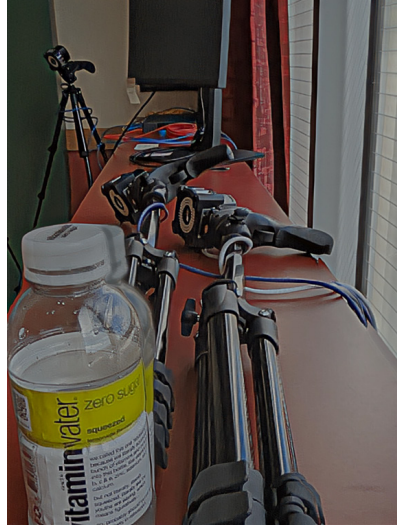
blurry in W, while the W should be used as a reference to preserve the desired colors even in blurry regions. This behavior makes our method robust to color differences in W and UW just like we show in Figure 1 where UW has incorrect white balance, and in Figure 2 we show how the model avoids relying on UW in occluded regions where artifacts may show up in the optical flow alignment.

D. Generalizing to Different Phone Setup

Our method requires only two cameras with different DoFs. This is widely available in modern smartphones since ultra-wide cameras tend to have a deeper DoF due to the small focal length compared to the wide and Telephoto cameras. Our approach that utilizes the defocus map is not specific to a particular device, but rather it can produce fairly good results for any smartphone with a similar UW+W dual-camera setup. To use data captured using an iPhone 14 Pro, we used the iPhone’s portrait mode to obtain a disparity map (shown in Fig. 3), and warped the UW using an optical-flow based alignment. In Fig. 4, we show results of our model on data captured by iPhone 14 Pro *without any finetuning*.



Aligned UW



Refined UW



UW blending mask



Ref. W



Refined W



W blending mask

Figure 2. **Intermediate results visualization.** Note that while the aligned **UW** suffers from an alignment artifact around the bottle, the predicted masks take that into account by setting a low blending value for the occluded region in the **UW** mask and a higher value in the **W** mask.



Figure 3. **Disparity from iPhone.** Using portrait mode, we obtained Dual-camera disparity using an iPhone 14 Pro.



W



UW



Deeper DoF



Reduced DoF

Figure 4. **Results on iPhone 14 Pro.** We ran our model on images from an iPhone 14 Pro, and show that it generalizes with blurring and deblurring despite not finetuning the model on any iPhone data.

References

- [1] 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. [1](#)
- [2] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. [1](#)
- [3] Lingyan Ruan, Bin Chen, Jizhou Li, and Miuling Lam. Learning to deblur using light field generated and real defocus images. In *CVPR*, 2022. [1](#)