DeepFocus: Learned Image Synthesis for Computational Display

Lei Xiao
Facebook Reality Labs

Anton Kaplanyan
Facebook Reality Labs

Alexander Fix
Facebook Reality Labs

Matt Chapman
Facebook Reality Labs

Douglas Lanman
Facebook Reality Labs

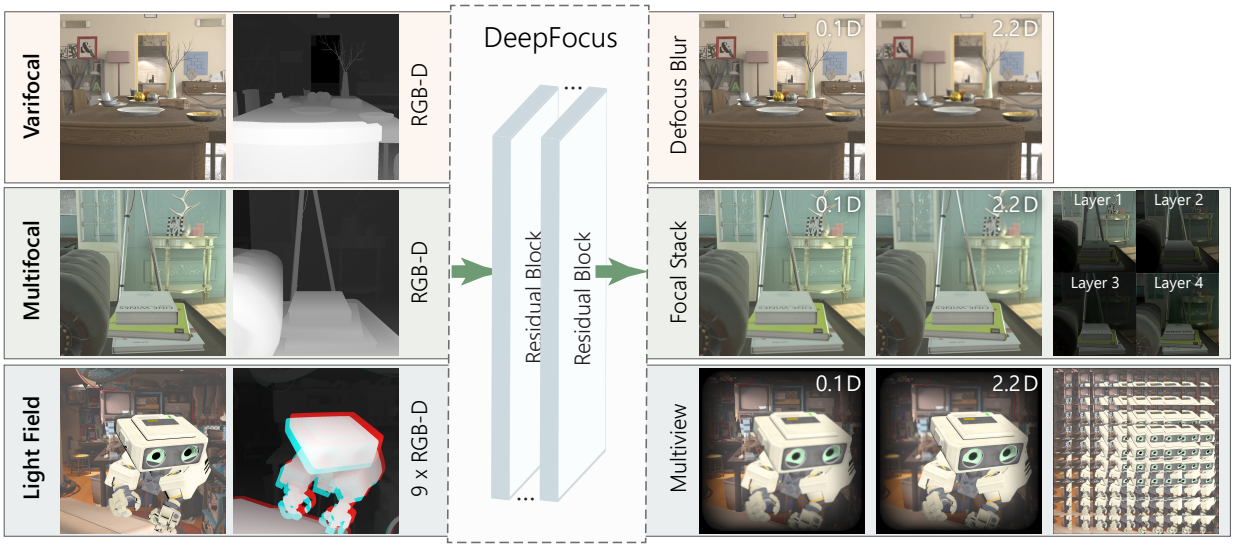

Figure 1: We present *DeepFocus*: a unified rendering and optimization framework, based on convolutional neural networks, that enables real-time operation of accommodation-supporting head-mounted displays. (Top row) For varifocal displays, DeepFocus synthesizes physically-accurate defocus blur from a single RGB-D input, as shown in the simulated retinal images on the right. (Middle row) For multifocal displays, the network solves the inverse problem to output a multilayer decomposition. (Bottom row) For light field displays, DeepFocus generates dense multiview imagery from a sparse set of RGB-D images.

ABSTRACT

Reproducing accurate retinal defocus blur is important to correctly drive accommodation and address vergence-accommodation conflict in head-mounted displays (HMDs). Numerous *accommodation-supporting* HMDs have been proposed. Three architectures have received particular attention: varifocal, multifocal, and light field displays. These designs all extend depth of focus, but rely on computationally expensive rendering and optimization algorithms to reproduce accurate retinal blur (often limiting content complexity and interactive applications). To date, no unified computational framework has been proposed to support driving these emerging HMDs using commodity content. In this paper, we introduce *DeepFocus*, a generic, end-to-end trainable convolutional neural network designed to efficiently solve the full range of computational tasks for accommodation-supporting HMDs. This network is demonstrated to accurately synthesize defocus blur, focal stacks, multilayer decompositions, and multiview imagery using commonly available RGB-D images. Leveraging recent advances in GPU hardware and best practices for image synthesis networks, DeepFocus enables real-time, near-correct depictions of retinal blur with a broad set of accommodation-supporting HMDs.

SIGGRAPH '18 Talks, August 12-16, 2018, Vancouver, BC, Canada

ACM ISBN 978-1-4503-5820-0/18/08.
https://doi.org/10.1145/3214745.3214769

CCS CONCEPTS

• **Computing methodologies** → *Rendering*; *Neural networks*; *Virtual reality*; *Mixed / augmented reality*;

KEYWORDS

computational displays, deep learning, depth of field, varifocal, multifocal, light fields, accommodation

ACM Reference Format:
Lei Xiao, Anton Kaplanyan, Alexander Fix, Matt Chapman, and Douglas Lanman. 2018. DeepFocus: Learned Image Synthesis for Computational Display. In *Proceedings of SIGGRAPH '18 Talks.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3214745.3214769

1 METHOD

Figure 2 shows the architecture of the DeepFocus network. At a high level, the network is a fully convolutional network augmented with *residual connections*, where a layer is added together with a preceding layer, and *skip connections*, where the next-to-last layer is concatenated with the input layer. The number of layers and the number of filters in each layer are selected to balance quality and inference time for each of our target applications. In the following subsections, we demonstrate that our network can be trained and

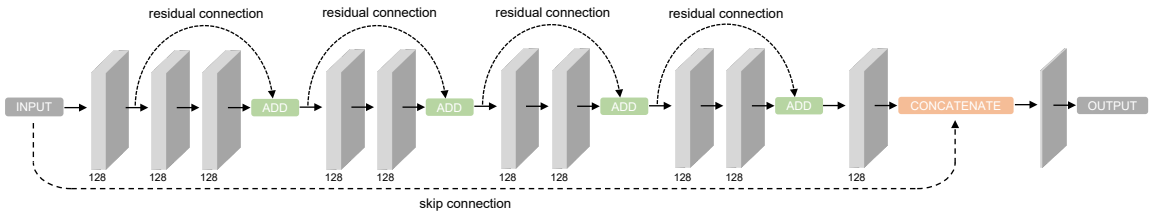

Figure 2: Example DeepFocus network architecture. See the supplementary materials for details of the layer components.

applied to the core rendering and optimization tasks for a broad class of accommodation-supporting HMDs. Most significantly, we show that the DeepFocus network is generalizable: tailoring for each application only involves changing the input and output layers.

1.1 Defocus Blur Rendering from RGB-D

Rendering accurate defocus blur is important for properly driving accommodation in near eye displays. Traditional methods for faithfully rendering defocus blur are either computationally expensive or fail to approximate the blur effects at partially occluded regions.

The DeepFocus network takes as input a single RGB-D (color and depth) image and a circle-of-confusion (CoC) map (pixel-wise blur size on the retina), and generates a high-quality retinal defocus blur image as an output. Example results and comparisons are shown in Figures 1 and 3. On our set of test scenes, DeepFocus significantly outperforms state-of-the-art methods, including Nalbach et al. [2017], and commercial software, such as Unity [2018], both in peak signal-to-noise ratio (PSNR) and in structural similarity (SSIM). More comparisons can be found in the supplementary materials.

1.2 Multilayer Decomposition from RGB-D

Multifocal displays represent 3D scenes with a limited number of display layers. A multilayer decomposition algorithm must solve for the set of displayed images. Akeley et al. [2004] depict each pixel of the RGB-D input on the two nearest layers: a representation that introduces artifacts at occlusion boundaries. Mercier et al. [2017] first render a target focal stack and then solve for the decomposition as an iterative optimization problem. While producing high visual quality, this approach is computationally expensive.

The DeepFocus network *directly* generates the multilayer decomposition from a single RGB-D image, avoiding both focal stack rendering and iterative optimization. It significantly outperforms previous work in both quality and run-time efficiency. Example results are shown in Figure 1. More results and comparisons can be found in the supplementary materials.

1.3 Light Field Rendering from RGB-D

Near-eye light field displays [Lanman and Luebke 2013] require as input a large number of elemental images, each of which is rendered from a distinct viewpoint to form a 4D light field. The high-resolution 2D image shown on the underlying display panel is generated from this light field. The resulting retinal image adapts with the viewer's accommodative state by superimposing multiple projected images to approximate near-correct retinal defocus blur.

Rendering tens or even hundreds of views is computationally expensive. The DeepFocus network efficiently synthesizes dense light

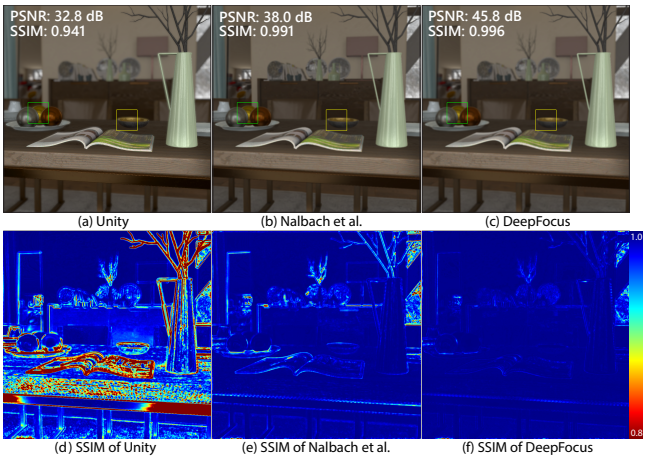

Figure 3: Comparison of defocus blur rendering using RGB-D inputs only. DeepFocus significantly outperforms other methods in both PSNR and SSIM.

fields (e.g., 81 views in our examples) with a single-pass network evaluation, when provided with a sparse set of RGB-D images (e.g., 5 or 9 input views in our examples). As a result, we can synthesize light fields that produce high-quality retinal images (see Figure 1). More details can be found in the supplementary materials.

2 IMPLEMENTATION

The DeepFocus networks are trained using TensorFlow, providing a large set of path-traced scenes as the training data. Network inference is optimized with TensorRT [2018] and evaluated on a Nvidia Titan V. Our networks currently achieve real-time performance on 1024×1024 color images for the applications in Section 1.1 (40.8 fps) and Section 1.2 (37.3 fps), and on 81×512×512 color images for the application in Section 1.3 (30.6 fps).

REFERENCES

Kurt Akeley, Simon J. Watt, Ahna R. Girshick, and Martin S. Banks. 2004. A Stereo Display Prototype with Multiple Focal Distances. *ACM Trans. Graph.* 23, 3 (2004), 804–813.

Douglas Lanman and David Luebke. 2013. Near-Eye Light Field Displays. *ACM Trans. Graph.* 32, 6, Article 220 (2013), 10 pages.

Olivier Mercier, Yusufu Sulai, Kevin Mackenzie, Marina Zannoli, James Hillis, Derek Nowrouzezahrai, and Douglas Lanman. 2017. Fast Gaze-Contingent Optimal Decompositions for Multifocal Displays. *ACM Trans. Graph.* 36, 6 (2017), 237.

O. Nalbach, E. Arabadzhiyska, D. Mehta, H.-P. Seidel, and T. Ritschel. 2017. Deep Shading: Convolutional Neural Networks for Screen Space Shading. *Comput. Graph. Forum* 36, 4 (2017), 65–78.

Nvidia Corporation. 2017–2018. TensorRT. https://developer.nvidia.com/tensorrt. (2017–2018).

Unity Technologies. 2005–2018. Unity Engine. http://unity3d.com. (2005–2018).