# An Object-Based Refocusing Scheme for Light Field Video Content

Nusrat Mehajabin, Yixiao Wang, Hamid Reza Tohidypour, Panos Nasiopoulos, Mahsa Pourazad
Electrical and Computer Engineering
University of British Columbia
Vancouver, BC, Canada
e-mails:{nusratm, yixiaow, htohidyp, panosn, pourazad}@ece.ubc.ca

*Abstract*— **Existing Light Field (LF) refocusing techniques refocus to all the pixels of a certain depth plane. However, in reality, the human eye focuses only on the object of interest while everything else is in depth-wise out of focus. In this paper, we propose a LF refocusing technique that is consistent with human visual perception. To this end, we perform instance segmentation on LF content and bring the whole object of interest in-focus rather than only the parts of the object that are in the same depth plane or all the objects that are on the same depth plane as the object of interest. Experimental results show that the proposed method is more consistent with the human visual perception than existing methods, yielding significantly improved visual quality results.**

***Keywords-light field; refocusing; Human Visual System (HVS); instance segmentation.***

## I. INTRODUCTION

LF technology emerged as an "upgrade" to the way we capture and reproduce visual information [1][2]. It offers even more immersive and holistic imaging experience than 360 video and omnidirectional stereo videos [3]. It allows post-shoot refocusing, perspective shifts, depth of field change, and 3D-like content generation with great precision. In fact, refocusing is the most far-reaching potential application of LF technology and the foundation for more realistic mixed reality applications. As augmented reality (AR)/ virtual reality (VR)/ mixed reality (MR) technologies use position and eye tracking, the experience needs to adjust to user's head and eye movement making it more comfortable and immersive. However, with the current state of LF refocusing, we can only alter the focal plane of a captured LF and refocus on all the objects (pixels) belonging to that focal plane. In other words, we bring all the pixels pertaining to the desired depth in-focus and everything else is out of focus.

LFs are captured either by plenoptic cameras or camera arrays. Plenoptic cameras [1] place an array of microlenses in between primary lens and photosensor to capture angular information. Limited photosensor resolution forces us to choose between spatial or angular resolution. In contrast, camera arrays [2] arrange multiple cameras on a rig to capture a scene. The distances between the lenses of the cameras (centimeters) are far greater than they are in plenoptic cameras (nanometers). This means that the point of views are further apart in camera array than they are in plenoptic camera. For camera array setups, the number of cameras dictates the angular resolution/number of viewpoints. Each camera's photo-sensor captures a single image, resulting in higher spatial resolution LFs. Traditional camera array systems are bulky and require a lot of hardware [4] but through camera miniaturization [5] these problems are being rapidly solved. As camera array content has larger baseline, typical refocusing methods lead to aliasing problem in the out of focus region [6].

A stereo-image refocusing method was introduced by Busam et al. [7]. This method selectively blurs the image based on the estimated depth using the stereo image, to create a refocused effect. However, this method was not extended for more than two cameras. Similar problems were found in the works of Cossairt et al. [8] and Bando et al. [9]; the former creates a refocused image using three images, whereas the latter generates the entire LF from a single image. Recent method by Wang et al. [10] also proposes to generate the out of focus region using depth-based anisotropic filters and in-focus region is produced by reconstruction based super resolution. Up to this point, all LF refocusing research assumed the new focal plane for refocusing will be parallel to the camera. Another recent work by Alain et al. [11] considered the scenario where the viewer tilts their head and the desired refocus plane and viewpoint are no longer parallel to the camera. In summary, all the above-mentioned approaches are designed for refocusing to a new focus plane, thereby, bringing all the pixels that are a certain distance from the camera in-focus and rendering everything else as depth wise blur. However, when we observe a scene and focus on an object, every part of the scene having the same distance from our eyes does not come into focus for us. Instead, we see that object clearly and everything else is in depth wise blurred with respect to that object.

In this paper, we propose a realistic, more immersive, and more visually pleasing object-based LF refocusing technique, which uses a deep learning network to synthesize an appropriate number of new views and another deep learning network for object segmentation.

 ISBN: 978-1-61208-927-0 

(a) (b) (c)

Figure. 1. (a) original view (90th frame), (b) COCO annotator segmented view (90th frame) (c) Deep learning network segmented view (91st frame)

The rest of this paper is organized as follows. In Section II, we present our object segmentation-based enhancement method. Experimental results are presented and discussed in Section III. We conclude our work in Section IV.

## II. PROPOSED METHOD

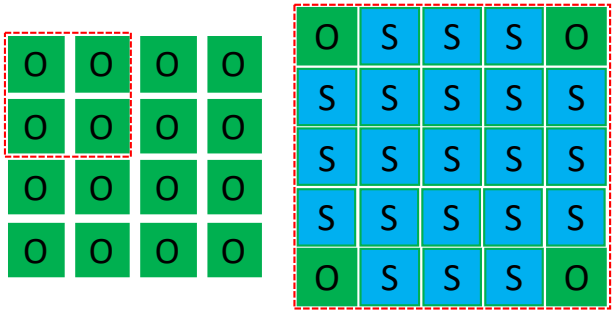
Our proposed refocusing scheme for LF camera array videos uses a deep learning-based view synthesis method and a shift and sum approach combined with a unique object segmentation-based enhancement technique for the in-focus object. Viewers select the object they wish to refocus on, and our method refocuses on that object while depth wise blurring the rest of the scene. The following subsections describe our approach in detail.

### A. Object Segmentation of LF videos

At first, we segment individual objects in the video. We used an object segmentation deep learning network Detectron2 [12] on LF videos [13] and segmented the objects. We used COCO Annotator [14] to annotate 1 frame per second for the LF videos and trained Detectron2 using these frames. Then, we used the trained network to segment all the other frames of the video. Figure 1 (a)-(c) show the original frame, the COCO Annotator annotated frame and the Detectron2 segmented frame correspondingly.

### B. View Synthesis

After object segmentation, to reduce the aliasing artifacts caused by the wide baseline of camera arrays, we synthesize novel views in between existing views using a factor of $n$. Here, $n=1$ is equivalent to one novel view between two adjacent views and 1 view between each pair of horizontal/vertical synthesized views. Therefore, the total number of views is $(nN - n+N)^2$, where $N \times N$ is the arrangement and number of cameras on the camera array. We used a pre-trained fully-convolutional encoder-decoder deep learning network [15] architecture (modeled after the popular convolutional neural network VGG-19 [16] by the Visual Geometry Group) to synthesize the novel views. This network does not require any camera parameter information, and thus it can be generalized to any LF video dataset. Given two input views captured by two horizontally/vertically aligned cameras and the distance between those two cameras, the network can synthesize as many novel views between the input views as desired. Figure 2 shows view synthesis for $n=3$.




(a) (b)

Figure. 2. (a) Original Views, (b) After n=3 view synthesis for top left original 2×2 views

### C. LF Refocusing

During the refocusing process, one camera will be picked as the reference camera. Then, the user selects any object on the reference view to refocus on. We estimate the depth of that object using Depth Estimation Reference Software (DERS) [17]. Next, we need the disparity shift of all the views from the reference view at the determined depth. The disparity shift for original views can be found in [13]. The synthesized views are calculated by linearly interpolating between the disparity shifts of the two adjacent original views. Finally, we apply the generalized shift and sum algorithm [11] using the interpolated and original disparities to produce the refocused video.

### D. Object Segmentation based Enhancement Method

After refocusing, we see that all the objects/pixels of the refocused plane are in-focus rather than the object of interest. We also observe that the refocused plane appears to be a little blurry, especially when it is close to the camera. The reason for this is that the synthesized views are not as accurate as the original views and as a result the average values for the in-focus pixels are not accurate either. To address these problems, we introduce an object segmentation-based enhancement method. First, we segment the object of interest from the reference view using the trained deep learning network. Next, we replace the pixels values of the segmented object at the refocused view with those of the reference view.

## III. EXPERIMENTAL RESULTS

We used the Interdigital LF video dataset [13] to evaluate the proposed approach. The LF videos were captured by a

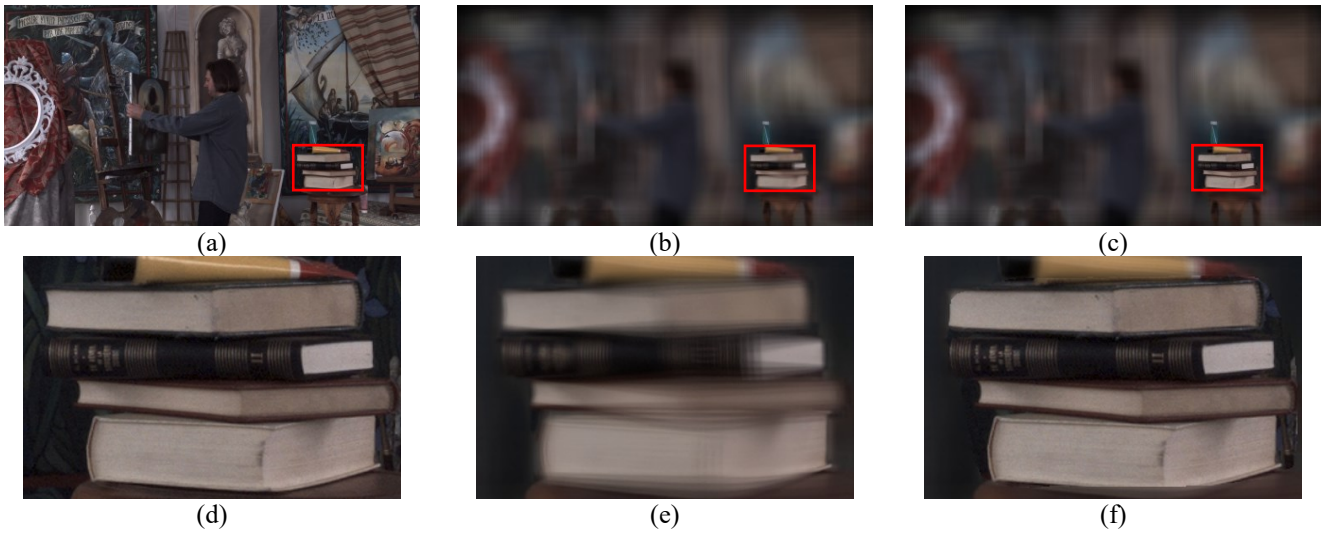


Figure. 3. (a) Original reference view, (b) Refocusing without enhancement, (c) Refocusing with object segmentation-based enhancement, (d), (e) and (f) are the enlargements of refocused objects of (a), (b) and (c), respectively

synchronized 4×4 camera array at 30fps. The cameras are 70*mm* apart with 50°×37° field of view. Each LF video has 2048×1088 spatial resolution in raw 4:2:0 8bit YUV format and is 12.3 seconds long (i.e., 372 frames). We demonstrate the original reference frame, refocused frames without and with our proposed object segmentation- based enhancement in Figure 3. We have experimented with different numbers of synthesized views $n = \{0, 1, 2, 3, 5, 10\}$, at various depths (in meters) from the camera: $z = \{1.9, 2.0, 3.0, 3.2, 4.0\}$.

For demonstration purposes, we show the 75$^{th}$ frame of the "Painter" LF video sequence in Figure 3. We observed that as we increase the value of $n$ for view synthesis, the refocusing results improve, with acceptable results achieved by $n = 3$ (total 169 views with 16 original views). Based on the above observations and the increased computational complexity of generating more views, we decided to fix $n = 3$. Results of both without and with enhancement refocused at a 2m distance are shown in Figure 3 (b) and Figure 3(c), respectively. From Figure 3, it is obvious that object segmentation-based enhancement results are more consistent with human visual perception, more visually pleasing and natural compared to the results where object segmentation has not been used. The object of interest here are the books on the stool and they have been brought in sharp focus with object segmentation-based enhancement, whereas the background experiences depth wise blur. Please refer to the zoomed results in Figure 3(f). When refocused without object segmentation-based enhancement, as in Figure 3 (b), we observe the in-focus region/object is not at sharp focus, but rather looks a little blurry and suffers from some aliasing artifacts. This is more evident in the zoomed in version of Figure 3 (e). This is because we used original and synthesized views for refocusing and the synthesized views are not accurate and as a result the refocusing is not accurate either. Therefore, an enhancement technique is essential. For both with and without enhancement, the background still has some aliasing artifacts. This can be mitigated through more accurate view synthesis or by using blurring filters [10]. Comparing the backgrounds of the original all in-focus Figure 3 (d) and the proposed method in Figure 3 (f), we can notice that the background patterns (floral foliage from the painting behind) in Figure 3 (d) are clearly visible, whereas in Figure 3 (f) those patterns experience depth wise blur as they should in a human visual perception consistent refocused view. We also observe that the refocused region is blurrier if the object/region is closer to the camera. In Figure 4, we present the results of the same frame refocused on the painter at 3.2m distance from the camera. We notice the refocused region is blurry but free of any aliasing artifact. Even then, the object segmentation-based enhancement method brings the painter in sharp focus. The transition from in-focus to not in-focus region for Figure 4 (c) as well as for Figure 3 (f) looks slightly abrupt. For example, we do not see a smooth transition from the painter to background, while we

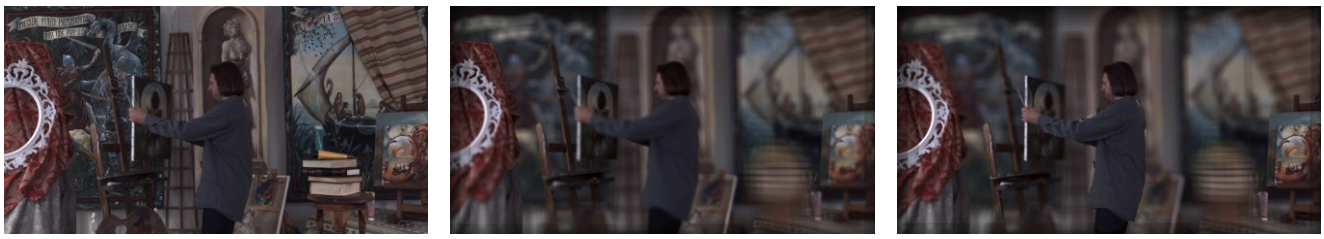


Figure. 4.(a) original reference view, (b) refocused on the painter at 3.2m distance and (c) refocused on the painter at 3.2m distance with enhancement

 ISBN: 978-1-61208-927-0 

would expect a smooth transition from the in-focus books on the table to the not in-focus region. This is a drawback of the proposed method. We leave this as future work, planning to explore the use of depth wise blending techniques for achieving better smoothing transition.

## IV. Conclusion and Future Work

In this work, we presented an efficient and human-like visual perception refocusing scheme for LF camera array content. Our approach uses a deep learning network to synthesize an appropriate number of new views and an object segmentation-based enhancement technique to improve the overall visual quality of the refocused frame. We found that the quality enhancement approach improves the visual quality of the in-focus regions by replacing the blurry pixels of the object of interest with corresponding pixels from the reference view using object segmentation. As a result, our method achieves visually acceptable and natural-looking refocused LF videos. To the best of our knowledge, this is the first method designed to refocus camera array LF content, offering unprecedented immersiveness, consistency with human visual perception and an excellent infrastructure for producing high-quality mixed reality content. Along with this paper we publish the code and an easy-to-use UI application [18] which might be of special interest for the community. Our future work will focus on finetuning the object segmentation process and increasing the size of our training dataset to include as many unique objects as possible.

## References


[1] R. Ng *et al.*, "Light Field Photography with a Hand-held Plenoptic Camera," Doctoral dissertation, Stanford University, 2005.

[2] B. Wilburn *et al.*, "High Performance Imaging Using Large Camera Arrays," ACM SIGGRAPH, pp. 765-776, 2005.

[3] R. S. Overbeck, D. Erickson, D. Evangelakos, M. Pharr, and P. Debevec, "A system for acquiring, processing, and rendering panoramic light field stills for virtual reality," *SIGGRAPH Asia 2018 Tech. Pap.*, vol. 37, no. 6, 2018, doi: 10.1145/3272127.3275031.

[4] G. Wu *et al.*, "Light Field Image Processing: An Overview," *IEEE J. Sel. Top. Signal Process.*, vol. 11, no. 7, pp. 926–954, 2017, doi: 10.1109/JSTSP.2017.2747126.

[5] H. M. Kim, M. S. Kim, G. J. Lee, H. J. Jang, and Y. M. Song, "Miniaturized 3D depth sensing-based smartphone light field camera," *Sensors (Switzerland)*, vol. 20, no. 7, pp. 2129, 2020, doi: 10.3390/s20072129.

[6] C. Huang, J. Chin, H. Chen, Y. Wang, and L. Chen, "Fast Realistic Refocusing for Sparse Light Fields," *IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 1176–1180, 2015.

[7] B. Busam, M. Hog, S. McDonagh, and G. Slabaugh, "SteReFo: Efficient image refocusing with stereo vision," *Proc. - 2019 Int. Conf. Comput. Vis. Work. ICCVW 2019*, pp. 3295–3304, 2019, doi: 10.1109/ICCVW.2019.00411.

[8] O. Cossairt, N. Matsuda, and M. Gupta, "Digital refocusing with incoherent holography," *2014 IEEE Int. Conf. Comput. Photogr. ICCP 2014*, pp. 1-9, 2014, doi: 10.1109/ICCPHOT.2014.6831819.

[9] Y. Bando and T. Nishita, "Towards Digital Refocusing from a Single Photograph," *15th Pacific Conference on Computer Graphics and Applications*, pp. 363-372, 2007

[10] Y. Wang, J. Yang, Y. Guo, C. Xiao, and W. An, "Selective Light Field Refocusing for Camera Arrays Using Bokeh Rendering and Superresolution," *IEEE Signal Process. Lett.*, vol. 26, no. 1, pp. 204–208, 2019, doi: 10.1109/LSP.2018.2885213.

[11] M. Alain, W. Aenchbacher, and A. Smolic, "Interactive Light Field Tilt-Shift Refocus with Generalized Shift-and-Sum," arXiv preprint arXiv:1910.04699 (2019).

[12] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick, "Detectron2," Available from: https://github.com/facebookresearch/detectron2, 2019.

[13] N. Sabater *et al.*, "Dataset and Pipeline for Multi-view Light-Field Video," *IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work.*, vol. 2017-July, pp. 1743–1753, 2017, doi: 10.1109/CVPRW.2017.221.

[14] J. Brooks, "COCO Annotator." Available from: https://github.com/jsbroks/coco-annotator/, 2019.

[15] T. Zhou, R. Tucker, J. Flynn, G. Fyffe, and N. Snavely, "Stereo magnification: Learning view synthesis using multiplane images," *ACM Trans. Graph.*, vol. 37, no. 4, 2018, doi: 10.1145/3197517.3201323.

[16] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc.*, pp. 1–14, 2015.

[17] S. Rogge *et al.*, "MPEG-I Depth Estimation Reference Software," in *2019 International Conference on 3D Immersion (IC3D)*, 2019, pp. 1–6, doi: 10.1109/IC3D48390.2019.8975995.

[18] N. Mehajabin, "Human Visual Perception Consistent Light Field Refocusing." Available from: https://github.com/Nusrat17/Light-Field-Refocusing, Vancouver, 2021.


 ISBN: 978-1-61208-927-0 