OPTIMIZATION OF COMPRESSIVE LIGHT FIELD DISPLAY IN DUAL-GUIDED LEARNING

Yangfan Sun [1 4] *Zhu Li* [1] *Li Li* [2] *Shizheng Wang* [3] *Wei Gao* [4]

[1] University of Missouri-Kansas City, USA
[2] University of Science and Technology of China, China
[3] Chinese Academy of Sciences, China
[4] Peng Cheng National Laboratory, China

ABSTRACT

Glass-free compressive light field (CLF) display gains much attention due to their compatibility in holographic-like and three-dimensional (3D) demonstration. Opposite to other analogous devices, CLF display can provide binocular and motion parallaxes by stacking multiple liquid crystal screens without any extra accessories. It is possible to bring the immersive and accommodative experience upon a well-pleasing visual consequence. Conventionally, the excessive processing time impacts its practical value in commercial, along with the severe degradation of display brightness. Therefore, in this paper, we propose a learning-based factorization framework to promote the visual results and expedite the layer decomposition and display adaption. It utilizes the advantage of a dual-guided system and residual learning to implement pixel-wise information extraction and refinement. The experimental results illustrate the outperformance of our proposed method over the conventional iterative factorization. Furthermore, a three-layered CLF prototype has been assembled to verify the practicality of our method.

Index Terms— Compressive light field (CLF) display, dual-guided system, learning-based method, light field (LF), residual learning.

1. INTRODUCTION

Three-dimensional (3D) displays have been gradually emerging in the commercial market due to the development of stereoscopic sensation technologies, such as parallax barriers, volumetric or holographic display, along with visual-reality (VR) technology [1]. However, the drawbacks of these technologies exceedingly influence the immersive experience or limit the practical usages, which can be summarized as followed: 1) Discrepancy between visual accommodation and convergence [2]; 2) Excessive amount of data storage space needs; 3) Additional accessories, such as polarizing lenses or headset mounts. In this case, the free-glass compressive light field (CLF) displays [3, 4, 1] were investigated to handle the aforementioned issues due to their optimization in both optical and algorithm aspects [2, 5, 6, 7, 8].

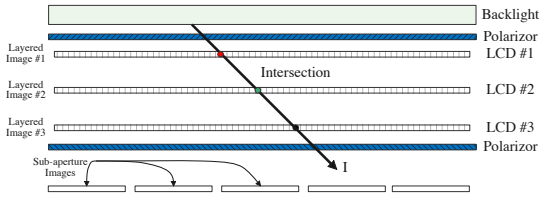

Fig. 1. Physical architecture of polarization-based CLF display.

Generally, CLF display utilizes the non-negative tensor factorization (NTF) algorithms and the multi-layer spatial light modulators (SLMs) that compresses and factorizes the sub-aperture images (SAIs) to each physical display layer in accordance with the corresponding objective depth, which allows reflecting observers' binocular and motion parallaxes. In [3, 4, 1], initial prototypes were assembled in alternative mechanisms: polarization-based and attenuation-based multi-layer architectures. Subsequently, Wang et al. [5, 2] calibrated the initial depth dynamically based on the accurate estimation of the salience map. Zhu et al. [6] proposed a depth-guided parameter initialization method, following the identical principle. It shows that perceptual and subjective quality improvement is the mere consideration in previous research. However, the execution efficiency significantly limits CLF display from being widely applied in low-latency scenarios or sequence-level applications.

Therefore, in this paper, we propose a learning-based factorization framework for CLF display to decompose the content of SAIs to each screen layer in a rapid and precise manner. Essentially, it leverages the advantage of convolution networks to generate precise projected relations between SAIs and layered images. The main contributions are listed as below: First, we initially establish a dual-guided architecture in promoting the visual performance since the interaction of vi-

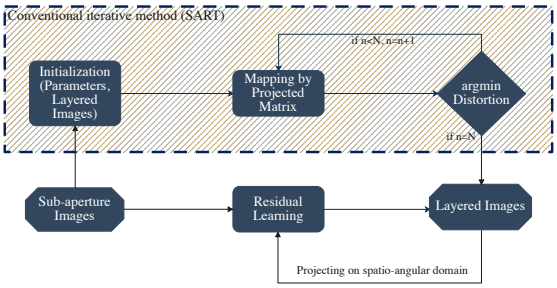

Fig. 2. Processing pipeline of our method over the conventional iterative method (SART)

sualized and refinement loss is to activate the pixel-wise transformation and reconstruction. Second, the residual learning is applied to speed up the convergence of distortion that synthesizes iterative-like results without repeated executions. Moreover, we assemble a high-power-backlight prototype for the visual demonstration and performance verification. In section 2, the technical details of CLF display will be elaborated. Then, the proposed method will be illustrated in section 3. In sections 4 and 5, the experimental results and conclusion will be discussed.

2. COMPRESSIVE LIGHT FIELD DISPLAY

In this paper, we aim at the optimization of polarization-based CLF display. As known, uniform light rays are generated by the screen backlight that goes through diffuse reflection, then, intersected with each display layer. The rotation angles of a light ray across each layer are summed up as the overall shifting. Fig. 1 illustrates the physical structure of polarization-based CLF display, whereas its model function is below,

$$I_{out} = I_{in} \cdot \sin^2(\Phi_a + \Phi_b + \Phi_c), \tag{1}$$

where Φ are the symbol of rotation angles. I_{out} and I_{in} represent the intensity of polarized light emission and entry. The pixels on each layer are fused based on the perspective of observation that located at each intersection from the given viewing directions.

Previously, the simultaneous algebraic reconstruction technique (SART) [9] has been used as a fast solver since it could acquire polarization-based consequences at interactive refresh rates. However, for SAI reconstruction, it does not perform well in convergence as many iterations are demanded [5]. Besides, It cannot suspend at the optimal performance but the pre-set iteration.

3. PROPOSED METHOD

In this section, we will interpret the processes of our proposed method, as shown in Fig. 3. First, we collect SAIs in RGB format to transpose and merge them from $\mathcal{L} \in \mathbb{R}^{U\times V\times S\times T\times C}$ to $\mathbb{R}^{F\times S\times T}$, where U and V are angular sizes, S and T are spatial resolutions, while C denotes the color channel. Note that F is the product of angular and color channels.

The initial training features can be extracted from the processed SAIs, as followed,

$$\mathcal{F}_{initial} = \mathcal{H}_{conv}(R_s(\mathcal{L}(u,v,s,t)) \mid \phi_{conv}), \tag{2}$$

where $\mathcal{L}(u,v,s,t)$ denotes the target images, since $\mathcal{H}_{conv}$ and ϕ_{conv} are the convolutional operation and its according trainable parameters. R_s is an operator of dimensional transposition and merging.

The residual learning [10] has been proven its functionality in many vision tasks. In this network, we involve numerous residual blocks that process residual learning within each block as well, which contain two convolutional layers and a ReLU activated function. In order to comprehensively exploit the mapping correlations from deep feature extraction, avoiding the overloaded computing due to the complex algorithm architecture,

$$\mathcal{F}_{res,i} = \begin{cases} \mathcal{H}_{res,i}(\mathcal{F}_{initial} \mid \phi_{res,i}), i = 0 \\ \mathcal{H}_{res,i}(\mathcal{F}_{res,i-1} \mid \phi_{res,i}), i > 0, \end{cases} \tag{3}$$

where i represents the index of residual blocks. $\phi_{res,i}$ and $\mathcal{H}_{res,i}$ are the i-th trainable parameters and convolutional layer. $\mathcal{F}_{res,i}$ are represented as the i-th features extracted from the corresponding block.

By concatenating the initial and rear residual features, we select the most useful ones through a convolutional layer towards the reconstruction of display layered images $\hat{\mathcal{D}}_l(s,t,c)$, as shown,

$$\hat{\mathcal{D}}_l(s,t,c) = \mathcal{H}_r(R_{con}(\mathcal{F}_{res}, \mathcal{F}_{initial}) \mid \phi_r), \tag{4}$$

where $\mathcal{F}_{res}$ denotes as the rear residual features since R_{con} operates concatenation in the feature domain. ϕ_r and $\mathcal{H}_r$ are the trainable parameters and the corresponding convolution layer. It is noted that the network utilizes the extracted depth information across multi-view features, and further distributes content in each physical layer according to their depth.

As shown in Fig. 3, we establish a dual-guided system that gives a thought to layered images and reconstructed SAIs simultaneously, involving the visualized and refinement loss. The former one is defined as the mean square error (MSE) between the display layered images from the iterative and our proposed method,

$$\ell_v = \frac{1}{N}\sum_{n=0}^{N}(D_n(s,t,c) - \hat{D}_n(s,t,c))^2, \tag{5}$$

where n is the index of display LCD layers. Here, we concentrate and optimize the three-layers CLF monitor ($N = 2$).

Even though, we consider iterative results as the guided images. They might not be the best quality representatives,

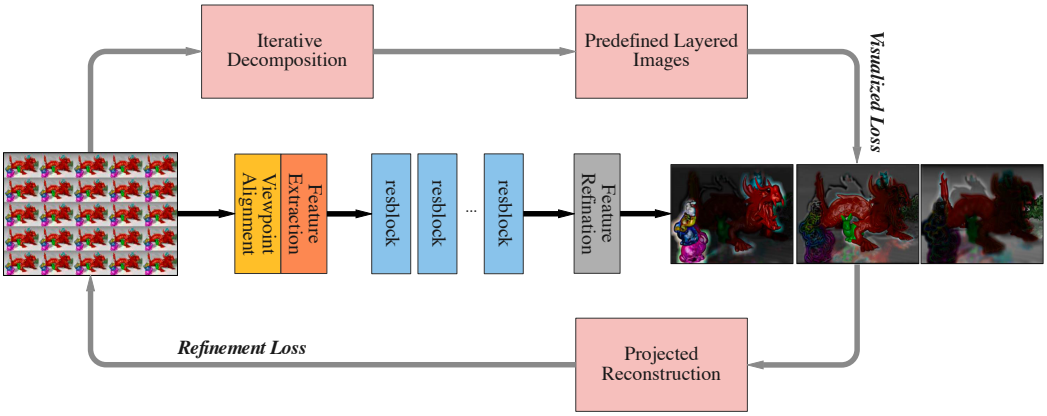

Fig. 3. The architecture of proposed dual-guided system

where the refinement loss is adopted in order to adjust the initial learning-based layered images pixel-wisely,

$$\ell_r = \frac{1}{F}\sum_{f=0}^{F}(\mathcal{L}_f(s,t) - \hat{\mathcal{L}}_f(s,t))^2, \tag{6}$$

where f is the index of reshaped dimension.

Eventually, the visualized loss and refinement loss are weighted and sum as the total loss,

$$\ell_{total} = \ell_v + \lambda * \ell_r. \tag{7}$$

4. EXPERIMENTAL RESULTS

In this paper, we collect five public datasets [12, 13, 14, 11, 15] to ensure the variety of parallax disparities, sampling content and capturing methods. The SART solver achieved by the tomographic LF synthesis tool [1] is used as a baseline. All LF SAIs are initialized to the interval of $(0, 1)$ and participate in off-line training. We crop the training patches into 7x7x64x64. PSNR (Peak Signal to Noise Ratio) and SSIM (Structural SIMilarity) are used as the quantitative metrics. The hyper-parameters are given to 16, 128, and 3x3 as the number of residual blocks, feature channels, and kernel size. The weight of refinement loss λ is set to 0.1. We run our network in an Nvidia RTX 3090 GPU and terminate at 200 epochs since the learning rate is fixed to 0.0001.

4.1. Quantitative and Qualitative Comparisons

To evaluate the effectiveness of the proposed method, we first compare the reconstructed performance over the baseline and previous learning-based method (CNN) [8]. Table 1 illustrates the promotion of reconstructed quality from our method apparently that a similar consequence is obtained in the qualitative comparison, as shown in Fig. 4, where we evaluate the quality of zoom-in patches by our method and the baseline. Ours reconstructed image show superior quality than the baseline, especially in some smooth regions. It is worth noticing that both our method and CNN are using baseline results as supervised data and our method surpasses the baseline in SAI reconstruction due to the advantage of the dual-guided training system. We have it tested in ablation study that the refinement loss is capacity to improve the performance, however, would impact the integrity of layered image whether its weight is too high.

Table 1. Quantitative comparisons of SAI reconstruction for different factorization methods.

	Baseline [1]	CNN [8]	Proposed
EPFL [14]	30.66	30.12	31.20
HCI [12]	30.03	29.40	30.46
Inria [15]	29.11	28.71	29.87
Kalantari [13]	35.04	34.68	35.65
Stan_Gen [11]	34.00	33.51	34.32
Stan_Occ [11]	33.27	32.99	33.79
Average	32.02	31.56	32.55

Furthermore, we evaluate the time consumption of each method, shown in Fig.5. The learning-based solutions are taken their advantage in a low-latent application for granted. They reduce time consumption to approximately one tenth of the baseline. Furthermore, our method shows the superiority of performance compared to the CNN-based method at a comparable decomposing speed.

4.2. Prototype of compressive light field (CLF) display

In order to reveal the practicality of CLF display optimization, we assemble a 15.4-inch three-layered prototype with an external uniform backlight in industrial power that achieves its observable brightness in daylight. The first row of Fig. 6

(a) Inria dataset

(b-d) Patches from the [4,4] view; (e-g) Patches from the [0,0] view

(h-j) Layered images by the proposed method; (k-m) Layered images by the baseline

Fig. 4. Qualitative comparisons. (a) The original image from the Inria dataset. (b-g) The cropped patches for a better comparison are from ground truth, our method, and baseline, respectively (from the left column to right). (h-m) The layered images correspond to the front, the middle, and the rear display screen (from the left column to right).

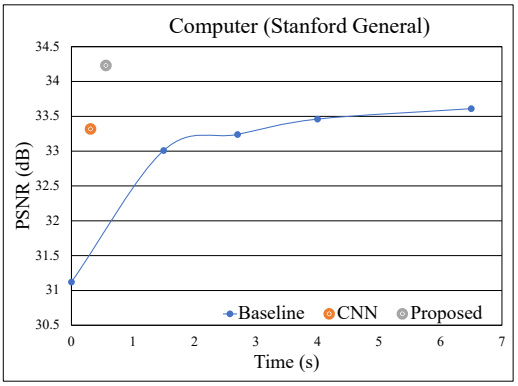

Fig. 5. Changes of reconstructed quality (X-axis) over time (Y-axis) for *Computer* scene (Stanford dataset [11]).

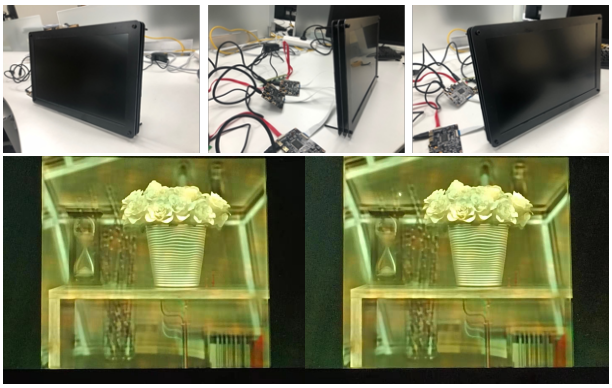

Fig. 6. The prototype of three-layered CLF display. First row shows the appearance of display from various observing angles. Second row demonstrates the display outcomes of our method (left) and baseline (right).

shows the appearance of our prototype with an exoskeleton frame, multi-layered display screens, and external driver chips. In the second row, we demonstrate the phenomenons of display in real-world factorized by our method and baseline, which are shot by a smart cellphone. In actual observation, the motion parallax can be discovered from different observing angles. However, due to the color deviation derived from its multi-layer structure, the background color needs to be calibrated in our future work.

5. CONCLUSION

In this paper, we establish a learning-based factorization framework towards the low-latent high-quality LF reconstruction. The dual-guided system with residual learning is employed to achieve this goal. The experimental results show that it can exceed 1 dB over the previous learning-based method and 10 times faster than the conventional iterative method. Moreover, we assemble a three-layered CLF display prototype to verify the practical demonstrations. It overcomes the issue of low brightness in daylight by rising the power of backlight.

Acknowledgment

This research is accomplished in Peng Cheng National Laboratory and supported by Beijing, Maoming, and National Natural Science Foundation (Grant No: 4194095, 2020028, 61802390).

6. REFERENCES

[1] Gordon Wetzstein, Douglas Lanman, Matthew Hirsch, Wolfgang Heidrich, and Ramesh Raskar, "Compressive light field displays," *IEEE computer graphics and applications*, vol. 32, no. 5, pp. 6–11, 2012.

[2] Shizheng Wang, Wenjuan Liao, Phil Surman, Zhigang Tu, Yuanjin Zheng, and Junsong Yuan, "Salience guided depth calibration for perceptually optimized compressive light field 3d display," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2031–2040.

[3] Douglas Lanman, Gordon Wetzstein, Matthew Hirsch, Wolfgang Heidrich, and Ramesh Raskar, "Polarization fields: dynamic light field display using multi-layer lcds," in *Proceedings of the 2011 SIGGRAPH Asia Conference*, 2011, pp. 1–10.

[4] Gordon Wetzstein, Douglas R Lanman, Matthew Waggener Hirsch, and Ramesh Raskar, "Tensor displays: compressive light field synthesis using multilayer displays with directional backlighting," 2012.

[5] Shizheng Wang, Zhenfeng Zhuang, Phil Surman, Junsong Yuan, Yuanjin Zheng, and Xiao Wei Sun, "Two-layer optimized light field display using depth initialization," in *2015 Visual Communications and Image Processing (VCIP)*. IEEE, 2015, pp. 1–4.

[6] Liming Zhu, Guoqiang Lv, Liye Xv, Zi Wang, and Qibin Feng, "Performance improvement for compressive light field display based on the depth distribution feature," *Optics Express*, vol. 29, no. 14, pp. 22403–22416, 2021.

[7] Keita Takahashi, Yuto Kobayashi, and Toshiaki Fujii, "From focal stack to tensor light-field display," *IEEE Transactions on Image Processing*, vol. 27, no. 9, pp. 4571–4584, 2018.

[8] Keita Maruyama, Keita Takahashi, and Toshiaki Fujii, "Comparison of layer operations and optimization methods for light field display," *IEEE Access*, vol. 8, pp. 38767–38775, 2020.

[9] D Lanman, G Wetzstein, M Hirsch, and R Raskar, "Depth of field analysis for multilayer automultiscopic displays," in *Journal of Physics: Conference Series*. IOP Publishing, 2013, vol. 415, p. 012036.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[11] Abhilash Sunder Raj, Michael Lowney, Raj Shah, and Gordon Wetzstein, "Stanford lytro light field archive," 2016.

[12] Katrin Honauer, Ole Johannsen, Daniel Kondermann, and Bastian Goldluecke, "A dataset and evaluation methodology for depth estimation on 4d light fields," in *Asian Conference on Computer Vision*. Springer, 2016, pp. 19–34.

[13] Nima Khademi Kalantari, Ting-Chun Wang, and Ravi Ramamoorthi, "Learning-based view synthesis for light field cameras," *ACM Transactions on Graphics (TOG)*, vol. 35, no. 6, pp. 1–10, 2016.

[14] Martin Rerabek and Touradj Ebrahimi, "New light field image dataset," in *8th International Conference on Quality of Multimedia Experience (QoMEX)*, 2016, number CONF.

[15] Jinglei Shi, Xiaoran Jiang, and Christine Guillemot, "A framework for learning depth from a flexible subset of dense and sparse light field views," *IEEE Transactions on Image Processing*, vol. 28, no. 12, pp. 5867–5880, 2019.