Disparity-Guided Light Field Video Synthesis with Temporal Consistency

Chin-Chia Yang, Yi-Chou Chen, Shan-Ling Chen, and Homer H. Chen
National Taiwan University
Graduate Institute of Communication Engineering
{r10942093, r10942061, r10942089, homer}@ntu.edu.tw

Abstract

Light field has great applications in AR/VR. It is particularly useful for resolving the vergence-accommodation conflict (VAC) and creating correct depth cues for AR/VR displays. However, the source data, especially light field video, are not widely available yet. To resolve the scarcity issue, one may resort to data such as stereo image sequences that are commonly available. In this paper, we propose an end-to-end deep learning framework for synthesizing light field sequences from stereo image sequences. Our framework consists of a disparity estimation network, a guided synthesis network, and a refinement network and is able to resolve the flickering issue caused by temporal inconsistency, an artifact that is commonly seen in synthesized light field videos. Our experimental results are quantitatively and qualitatively better than the results of existing light field synthesis algorithms that were originally developed for static light fields.

1. Introduction

Recently, light field display has received considerable attention in both academia and industry. Because the light field simulates light rays perceived by human eyes in a real scene, a realistic and comfortable viewing experience is made possible with light field displays such as a pair of light field glasses. Such displays are free of the vergence-accommodation conflict (VAC) that is the root cause of nausea or dizziness for users of conventional AR displays.

A light field can be represented as a 2D array of sub-images. The height and width of the 2D array are referred to as the angular resolution of the light field in the corresponding dimension, while the size of the sub-images is referred to as the spatial resolution of the light field. In other words, a light field carries the intensity and direction information of all light rays of a scene it represents.

However, due to cost and hardware limitations, high angular and spatial resolution light fields are difficult to obtain. This is especially the case for light field videos. Therefore, various solutions for light field synthesis have been developed to generate dense sub-images from a sparse light field. With advanced neural networks, it is even possible to synthesize a 9×9 light field from a stereo image pair or a single image.

In consideration of accessibility and quality, synthesizing a light field video from a stereo image sequence is a practical choice. Although stereo video data are more costly to obtain than monocular video data, they contain depth information useful for creating high-quality light fields.

An intuitive way to synthesize a light field video from a stereo video is to consider the stereo video input as a sequence of stereo image pairs and convert each stereo image pair to a light field. However, this may create unwanted artifacts, since it does not consider the temporal relationship between the stereo image pairs. Temporal inconsistency may lead to video flickering. This is usually the case for warping-based light field synthesis, which create light fields by using the disparity maps to warp input images. The occlusions and holes commonly seen in warped images can lead to noticeable flicker, especially if the stereo input has a large baseline.

In this paper, we propose a flicker-free light field video synthesis pipeline that takes a stereo sequence as input. This end-to-end learning-based synthesis approach consists of three components: a disparity estimation network, a guided synthesis network, and a refinement network.

First, we generate light fields with the input stereo pairs and disparity maps estimated from the disparity estimation network, similar to how light fields are synthesized using warping-based methods. However, instead of taking the warped light fields as the output, we use them as guidance to train a separate synthesis network, which generates light fields without image warping. The generated light fields are then refined by the refinement network, which is a 3DCNN that utilizes spatial-angular information. To further reduce the flicker that often occurs in the synthesized light field videos, we impose an optical flow loss to ensure the temporal consistency of the synthesis pipeline.

2. Related Work

2.1. Light Field Synthesis

Light field view synthesis or view interpolation, refers to the generation of new sub-images from a sparse light field or even a single image. Over the years, many view synthesis approaches have been developed, with the most common ones being depth or disparity-based methods and multiplane image (MPI)-based methods.

Among the depth or disparity-based light field synthesis methods, Kalantari *et al.* [10] generated novel views from 2×2 light field images by employing a learning-based method that estimates disparity features from the input and uses them to generate novel views. Chao *et al.* [6] further proposed an end-to-end learning-based method that creates a 9 × 9 light field from stereo images by utilizing the left-right consistency of stereo images. Moreover, Srinivasan *et al.* [15] developed a depth-based light field synthesis network for view synthesis from a single RGB image. However, this method heavily depends on the quality of the estimated depth and the color information of the image.

The method proposed by Zhou *et al.* [22] synthesizes a horizontal light field image from a stereo image with a small baseline using the MPI representation, which is a multi-layered image representation with each layer being a 4D RGBA image that represents scenes and objects at different depths. However, to generate novel views, this method usually requires two or more input images and camera parameters, which are more difficult for users to provide. Moreover, the inference of MPI and the rendering of novel views are time-consuming, making it difficult to synthesize light field images.

2.2. Light Field Video Synthesis

Compared to light field image synthesis, light field video synthesis needs to consider temporal consistency between each video frame. Bemana *et al.* [4] suggested the concept of an X-Field, which is a set of 2D images taken across different views, time, or illumination conditions. With the help of neural networks, it is possible to create joint view, time, or light interpolation. Bae *et al.* [3] proposed a learning-based method for synthesizing a light field video from a monocular video using an optical decoder to refine the temporal consistency. Shedligeri *et al.* [14] also proposed a learning-based method for synthesizing light field videos from stereo videos. It utilizes a low-rank light field representation based on layered light field displays [20] as well as a recurrent means to estimate both disparity and optical flow.

3. Method

In this section, we describe the proposed method for light field video synthesis and the associated loss terms.

3.1. Light Field Synthesis

We denote the input left stereo image by I_l and the input right stereo image by I_r. A light field, denoted by F, is an $N \times N \times H \times W \times 3$ tensor, where $N \times N$ denotes the angular resolution and $H \times W$ the spatial resolution of the light field. A sub-view in the light field is denoted by $F(i,j)$, where (i,j), $0 \leq i,j < N$, are the angular coordinates of the sub-view.

In the first step of our framework, disparity maps for guiding light field synthesis are generated from input stereo image pair by the disparity estimation network d. Specifically, the disparity estimation network takes I_l and I_r in order as input and outputs an $H \times W$ left-to-right disparity map $D_{lr} = d(I_l, I_r)$ and a right-to-left disparity map $D_{rl} = d(I_r, I_l)$ of the same dimension. To make the two disparity estimates smoother, we utilize a total variation loss L_v [16], [21] as follows:

$$L_v = \|\nabla D_{lr}\|_1 + \|\nabla D_{rl}\|_1. \tag{1}$$

Once D_{lr} and D_{rl} are obtained, we backward warp I_l and I_r to every sub-view $L(i,j)$. We denote the light field warped from I_l by F_l and the light field warped from I_r by F_r and impose a left-right consistency loss L_c similar to Chao *et al.* [6] as follows:

$$L_c = \|F_l - F_r\|_1 + \|F_l - F_t\|_1 + \|F_r - F_t\|_1 \tag{2}$$

where F_t denotes the ground truth light field. After F_l and F_r are obtained, we apply a distance-weighted blending method [6] to merge F_l and F_r into a guidance light field F_g, which is used in the next step to train the synthesis network and the refinement network.

We train the synthesis network to generate a light field similar to the guidance light field F_g. Our synthesis network, denoted by s, takes I_l and I_r as input and outputs a light field $F_s = s(I_l, I_r)$. Then we apply a guidance loss,

$$L_g = \|F_s - F_g\|_1, \tag{3}$$

to make F_s similar to F_g.

Finally, we feed the light field generated by the synthesis network into a refinement network that utilizes spatial-angular information to handle occlusions and non-Lambertian surfaces. The synthesized light field F_s is reshaped into an $N \times N \times H \times W \times 3$ tensor before it is fed to the refinement network. The refinement network, denoted by r, is a 3DCNN [15] that takes F_s as input and

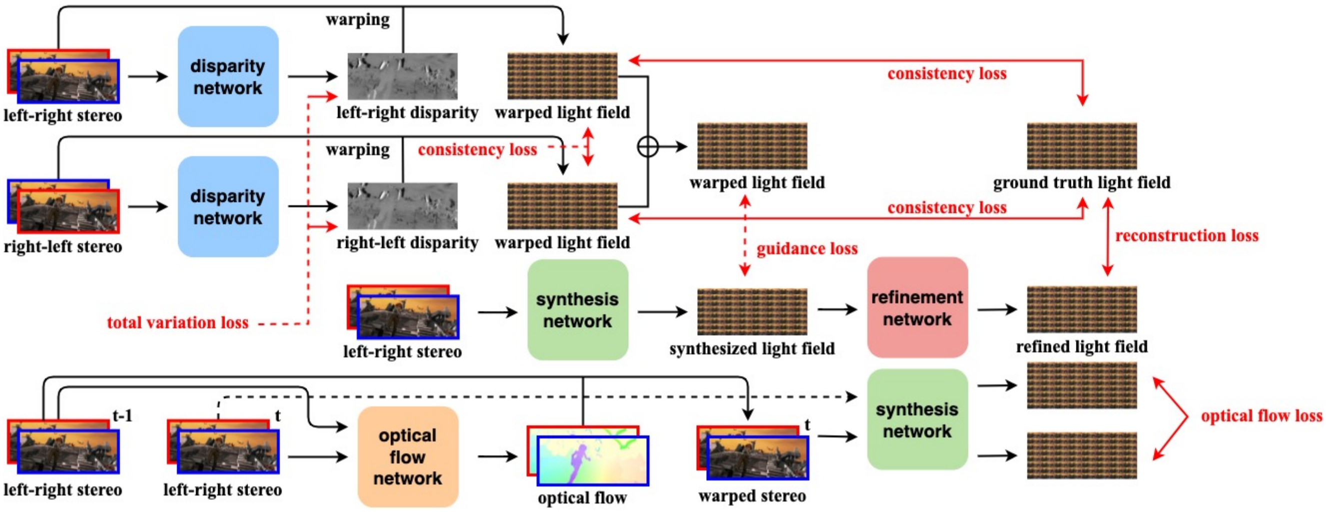

Figure 1. Flow chart of our proposed framework for light field video synthesis, which consists of a disparity network, a synthesis network, and a refinement network. The disparity network generates disparity maps for guiding light field synthesis, and the refinement network takes the synthesized light field as input and outputs the final predicted light field.

outputs a predicted light field F_p by employing the reconstruction loss,

$$L_r = \|F_p - F_g\|_1, \tag{4}$$

to minimize the distance between the predicted light field F_p and the ground truth light field F_g.

3.2. Temporal Consistency

We denote the input stereo pair at time t by I^t, which consists of the left stereo image and the right stereo image. That is, $I^t = (I^t_l, I^t_r)$. We compute the optical flow from time t to time $t-1$ by an optical flow estimation network o [17],

$$o(I^t, I^{t-1}) = f^{t\to t-1}, \tag{5}$$

where $f^{t\to t-1}$ consists of both left and right optical flows estimated from time t to time $t-1$.

We apply a backward warping operation W to warp the stereo image pair at time $t-1$ to the stereo image pair at time t,

$$W(I^{t-1}, f^{t\to t-1}) = I^{t-1\to t}, \tag{6}$$

where $I^{t-1\to t}$ denotes the stereo image pair backward warped from time $t-1$ to time t. An optical flow loss L_f is imposed to ensure that the backward warped stereo image pair and the stereo image pair at time t synthesize a similar light field,

$$L_f = \|s(I^t) - s(I^{t-1\to t})\|_1. \tag{7}$$

The total loss L is a weighted sum of the loss terms in Eqs. (1), (2), (3), (4), and (7). That is,

$$L = w_v \cdot L_v + w_c \cdot L_c + w_g \cdot L_g + w_r \cdot L_r + w_f \cdot L_f \tag{8}$$

where w denotes the weighting of each loss term.The overall flow chart of the proposed method is shown in Figure 1.

3.3. Implementation Details

To achieve efficiency, we implement a lightweight disparity estimation network with an architecture similar to that proposed by Chao *et al.* [6]. The synthesis network is a U-Net [13] with six input channels for input stereo image pairs and $N \times N \times 3$ output channels to generate light fields with $N \times N$ angular resolution. A tanh function is applied in the last layer of the disparity estimation network and the refinement network to normalize the output and stabilize the training procedure. We adopt ELU [7] as the activation function in the disparity estimation network and the refinement network.

Our networks are trained on the HCI 4D Light Field Dataset [9] and the Stereo Ego-Motion Dataset [2]. The HCI dataset, which consists of 24 light fields, is split by 5:1 into training and testing sets for light field synthesis training. The Stereo Ego-Motion Dataset is a real-world stereo video dataset, from which four stereo videos are randomly chosen for temporal consistency training. All networks are trained by using the Adam optimization algorithm [11] with default parameters and with $w_v = 0.001$, $w_c = 1$, $w_g = 1$, $w_r = 1$, and $w_f = 0.005$.

4. Experiments and Results

We perform both quantitative and qualitative comparisons to evaluate the proposed method. For quantitative comparison, we compare the proposed method with Chao *et al.* on two datasets: Hybrid light field video dataset [19] and Raytrix light field video dataset [8]. The former consists of light field videos with 8×8 angular resolution and 352×512 spatial resolution, and the latter consists of light field videos with 5×5 angular resolution and 1080×1920 spatial resolution. For qualitative comparison, we evaluate the proposed method on the testing sequences randomly selected from the Stereo Ego-Motion dataset [2]. We conduct qualitative experiments on this real-world stereo video dataset to evaluate the generalizability of the proposed method.

4.1. Quantitative Results

We adopt peak-signal-to-noise ratio (PSNR) and structural similarity (SSIM) as metrics to evaluate the performance of our method on the task of light field video reconstruction. We first extract stereo sequences from the ground truth light field videos. Then we apply different light field synthesis methods to reconstruct light field videos from the stereo sequences. Finally we compute PSNR and SSIM of the reconstructed light field videos against the ground truth light field videos. Table 1 shows the PSNR and SSIM values of the reconstructed light field videos. A higher PSNR or SSIM indicates a better similarity between the reconstructed light field videos and ground truth light field videos. Our method has better performance in terms of PSNR and is comparable in terms of SSIM.

The temporal consistency of the reconstructed light field videos is evaluated by using a warping loss [14]. We first predict the optical flow between adjacent ground truth light field video frames using an optical flow estimation network [18]. Once the optical flow is obtained, we backward warp adjacent predicted light field video frame using the predicted optical flow and compute the mean absolute error (MAE) of the backward warped frames with respect to the original frames. Table 2 shows the MAE of each method. The lower the MAE, the better the temporal consistency is. Our predicted light field videos have slightly better temporal consistency.

Datasets	Hybrid		Raytrix	
Metrics	PSNR	SSIM	PSNR	SSIM
Chao *et al.*	34.29	**0.964**	35.46	0.961
Ours	**34.62**	0.957	**37.66**	**0.965**

Table 1. Quantitative comparison of light field video reconstruction. The bold numbers indicate the best results.

Datasets	Hybrid	Raytrix
Metric	MAE	MAE
Chao *et al.*	0.0178	0.0208
Ours	**0.0174**	**0.0204**

Table 2. Quantitative comparison of temporal consistency. The bold numbers indicate the best results.

4.2. Qualitative Results

Examples of the refocused views and sub-views generated by our method and the methods of Shedligeri *et al* and Chao *et al* are shown in Figure 2 for qualitative comparison. It can be seen that our refocused views are sharper than the refocused views of Shedligeri *et al.* and Chao *et al.* In addition, we can observe that our method generates distortion-free sub-views, but the sub-views generated by Shedligeri *et al.* and Chao *et al.* have a certain degree of distortion. Furthermore, we observe that the distortion is the root cause of video flickering. A Comparison of the quality of our synthesized light field video with the other methods is provided on YouTube [1]. The qualitative comparison shows that our method performs well for real-world stereo sequences. It also shows the effectiveness of our synthesis network and disparity guidance for light field video synthesis.

4.3. Ablation studies

We conduct ablation studies to evaluate the influence of the loss terms, the blending method, and the refinement network of our light field video synthesis framework. First, we evaluate the effectiveness of total variation loss, left-right consistency loss, and guidance loss by setting, one by one, their weight to zero. Then we evaluate the performance of the distance-weighted alpha blending method with other blending methods. Finally, we evaluate the effectiveness of our refinement network. All ablation studies are tested on the Hybrid dataset [19]. From the results in Table 3, we verify that the loss terms, the blending method, and the refinement network are all effective components of our framework.

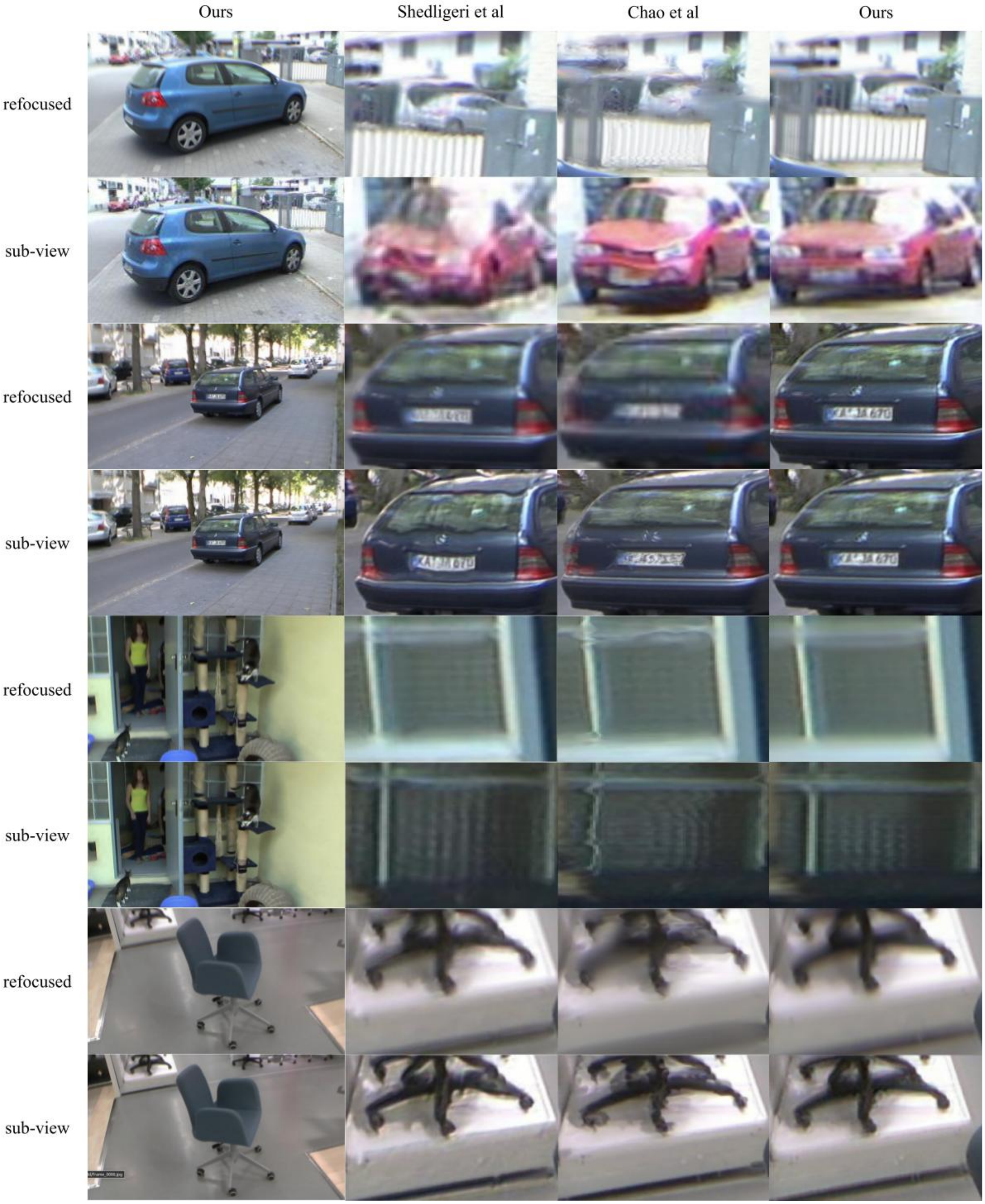

Figure 2. Qualitative comparison of refocused views and sub-views. We compare the visual quality of refocused views and sub-views with Shedligeri et al. and Chao et al. Our refocused views are sharper and our sub-views are distortion-free.

Method	PSNR	SSIM
Ours (w/o total variation loss)	34.59	**0.959**
Ours (w/o consistency loss)	33.62	0.943
Ours (w/o guidance loss)	32.86	0.930
Ours (only left view)	34.58	0.952
Ours (only right view)	30.40	0.910
Ours (average blending)	33.89	0.951
Ours (w/o 3DCNN)	31.04	0.910
Ours (full model)	**34.62**	0.957

Table 3. Ablation studies of our framework. The bold numbers indicate the best performance.

5. Conclusion

We have described a learning-based framework to synthesize light field videos from stereo sequences. Unlike conventional warping-based light field synthesis methods that take warped light fields as the output, the proposed disparity-guided framework takes the warped light field as a guidance to train a synthesis network, which generates light fields without image warping. As a result, our framework is able to avoid distortions caused by inaccurate disparity estimate. Furthermore, we take temporal consistency between video frames into consideration in our light field video synthesis framework and generate light field videos that are sharper and less flickering than those generated by warping-based methods.

References

[1] Comparison of synthesized light field videos. [online] https://www.youtube.com/channel/UCJMbrsFTIoSOzv6mjm3NY5A.

[2] Stereo ego-motion dataset (accessed: Mar. 03, 2022). [online] https://lmb.informatik.uni-freiburg.de/resources/datasets/StereoEgomotion.

[3] K. Bae, A. Ivan, H. Nagahara, and I. K. Park. 5d light field synthesis from a monocular video. *CoRR*, abs/1912.10687, 2019.

[4] M. Bemana, K. Myszkowski, H.-P. Seidel, and T. Ritschel. X-fields: Implicit neural view-, light- and time-image interpolation. *ACM Transactions on Graphics (Proc. SIGGRAPH Asia 2020)*, 39(6), 2020.

[5] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In A. Fitzgibbon et al. (Eds.), editor, *European Conf. on Computer Vision (ECCV)*, Part IV, LNCS 7577, pages 611–625. Springer-Verlag, Oct. 2012.

[6] C.-H. Chao, C.-L. Liu, and H. H. Chen. Robust light field synthesis from stereo images with left-right geometric consistency. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 1844–1848, 2021.

[7] D. Clevert, T. Unterthiner, and S. Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). In Y. Bengio and Y. LeCun, editors, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016.

[8] L. Guillo, X. jiang, G. Lafruit, and C. Guillemot. Light field video dataset captured by a r8 raytrix camera (with disparity maps). 2018.

[9] K. Honauer, O. Johannsen, D. Kondermann, and B. Goldlücke. A dataset and evaluation methodology for depth estimation on 4d light fields. In *ACCV*, 2016.

[10] N. K. Kalantari, T.-C. Wang, and R. Ramamoorthi. Learning-based view synthesis for light field cameras. *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia 2016)*, 35(6), 2016.

[11] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In Y. Bengio and Y. LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[12] N.Mayer, E.Ilg, P.Häusser, P.Fischer, D.Cremers, A.Dosovitskiy, and T.Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. arXiv:1512.02134.

[13] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015.

[14] P. Shedligeri, F. Schiffers, S. Ghosh, O. Cossairt, and K. Mitra. Selfvi: Self-supervised light-field video reconstruction from stereo video. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2471–2481, 2021.

[15] P. P. Srinivasan, T. Wang, A. Sreelal, R. Ramamoorthi, and R. Ng. Learning to synthesize a 4d RGBD light field from a single image. *CoRR*, abs/1708.03292, 2017.

[16] F. Steinbrücker, T. Pock, and D. Cremers. Large displacement optical flow computation without warping. In *2009 IEEE 12th International Conference on Computer Vision*, page 1609–1614, 09 2009.

[17] D. Sun, X. Yang, M. Liu, and J. Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. *CoRR*, abs/1709.02371, 2017.

[18] Z. Teed and J. Deng. RAFT: recurrent all-pairs field transforms for optical flow. *CoRR*, abs/2003.12039, 2020.

[19] T. Wang, J. Zhu, N. K. Kalantari, A. A. Efros, and R. Ramamoorthi. Light field video capture using a learning-based hybrid imaging system. *CoRR*, abs/1705.02997, 2017.

[20] G. Wetzstein, D. Lanman, M. Hirsch, and R. Raskar. Tensor displays: Compressive light field synthesis using multilayer displays with directional backlighting. *ACM Trans. Graph.*, 31(4), jul 2012.

[21] C. Zach, T. Pock, and H. Bischof. A duality based approach for realtime tv-l1 optical flow. In *Pattern Recognition*, page 214–223, 2007.

[22] T. Zhou, R. Tucker, J. Flynn, G. Fyffe, and N. Snavely. Stereo magnification: Learning view synthesis using multiplane images. *CoRR*, abs/1805.09817, 2018.