SeLFVi: Self-supervised Light-Field Video Reconstruction from Stereo Video

Prasan Shedligeri[1†] Florian Schiffers[2] Sushobhan Ghosh[2]
Oliver Cossairt[2] Kaushik Mitra[1]
[1]IIT Madras, India [2]Northwestern University, USA
[†]ee16d409@ee.iitm.ac.in

Abstract

Light-field imaging is appealing to the mobile devices market because of its capability for intuitive post-capture processing. Acquiring light field (LF) data with high angular, spatial and temporal resolution poses significant challenges, especially with space constraints preventing bulky optics. At the same time, stereo video capture, now available on many consumer devices, can be interpreted as a sparse LF-capture. We explore the application of small baseline stereo videos for reconstructing high fidelity LF videos.

We propose a self-supervised learning-based algorithm for LF video reconstruction from stereo video. The self-supervised LF video reconstruction is guided via the geometric information from the individual stereo pairs and the temporal information from the video sequence. LF estimation is further regularized by a low-rank constraint based on layered LF displays. The proposed self-supervised algorithm facilitates advantages such as post-training fine-tuning on test sequences and variable angular view interpolation and extrapolation. Quantitatively the reconstructed LF videos show higher fidelity than previously proposed unsupervised approaches.We demonstrate our results via LF videos generated from publicly available stereo videos acquired from commercially available stereoscopic cameras. Finally, we demonstrate that our reconstructed LF videos allow applications such as post-capture focus control and region-of-interest (RoI) based focus tracking for videos.

1. Introduction

Photography and videography has become ubiquitous in our modern lives due to the availability of simple-to-use imaging hardware. With steadily increasing image quality, consumers long for intuitive and simple processing for post-capture finetuning of their images. LF imaging has emerged as a promising imaging technique to overcome the limitations of conventional photography such as post-capture focus control, novel view synthesis, and post-capture depth-of-field control. With video acquisition surging in popularity, LF video capture could enable simple post-capture focus control for videos acquired on consumer devices. However, acquiring LF video data at useful frame-rates remains challenging. For example, commercial LF cameras such as Lytro acquire LF videos at only 3 frames per second (fps) [45]. This is mainly because of the trade-off between angular, spatial, and temporal resolution. Modern cameras easily capture videos at 720p resolution at a reasonable frame-rate of 30 fps. Ignoring the challenges of complex LF sensor, capturing a LF video at reasonable angular resolution of 7×7 requires a staggering $\sim 50\times$ more bandwidth. This is equivalent to capturing a 50MP video at 30 fps, something that is currently unimaginable for consumer devices.

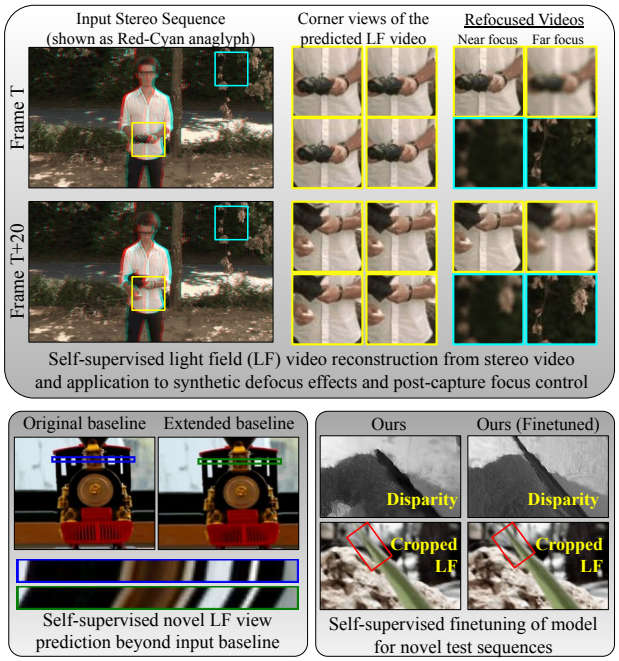

Figure 1: We propose a self-supervised algorithm for LF video reconstruction from a stereo video, enabling applications such as post-capture focus control for videos. Our proposed algorithm allows for post-training fine-tuning on test sequences and variable angular view interpolation as well as extrapolation.

While computational photography is poised to solve some of these problems in the upcoming decade via jointly optimized hardware-software solutions [16, 44, 50, 43], a practical solution is yet to be found. Numerous approaches have been proposed to overcome this challenge of high resolution LF imaging using hardware, commonly available today. Table 1 provides a concise review of such existing methods. We particularly note the recent work attempting to reconstruct LF *images* from sparsely sampled angular views [19, 3]. Considering the current limitations on available LF-hardware, we consider a simple case of sparse samples: the stereo image pair. In this paper, we tackle the task of reconstructing LF video from a sequence of stereo frames and propose a self-supervised learning-based algorithm as our solution.

The LF reconstruction in our self-supervised algorithm is guided via the geometric and temporal information embedded in a stereo video sequence. A recurrent neural network first takes the stereo frames at the current time-step and outputs a low-rank representation for LF frames based on layered LF displays [51]. The full 4D LF frame is then obtained from this representation via a deterministic linear operation. To enforce the LF epipolar consistency, we impose a disparity-based geometric consistency constraint on the generated LF frames. To ensure temporal consistency of the generated LF frames, we enforce an optical flow-based constraint [21]. Two different recurrent neural networks are learned to estimate the disparity maps and optical flow from the input stereo video. All three networks are trained via self-supervised cost functions during training.

A significant advantage of our approach is that it is self-supervised, and hence does not require hard-to-acquire ground-truth data for supervision while training. Our algorithm is able to estimate the full 4D LF with variable number of angular views from the input stereo views. We also show that our algorithm allows us to extend the baseline of the input views and generate novel views outside the original stereo baseline. Finally, our algorithm can be fine-tuned (see Sec. 4.4 and Fig. 1 and 8) on specific video sequences as it does not require ground truth data for supervision. Such self-supervised fine-tuning is especially useful when the test sequences do not follow the same distribution as the training sequences.

We show that our proposed algorithm outperforms the state-of-the-art disparity-based LF reconstruction algorithms. Our algorithm also performs on par with unsupervised LF reconstruction approaches, e.g. X-fields [3] with 4 corner-views of the LF as its input [3]. Overall, our contributions are:

- A self-supervised learning-based algorithm for LF video reconstruction from stereo video.
- Effective use of layered display based low-rank regularization for self-supervised LF video prediction.

Method	Self-Supervision	Stereo-View	Video
LF synthesis [19, 52, 47, 6, 12]	✗	✗	✗
View synthesis [19, 29, 7]	✗	✗	✗
View Synthesis [30, 23, 56]	✔	✗	✗
LF Video [14, 45, 33]	✗	✗	✔
Bino-LF [58]	✔	✔	✗
X-fields [3]	✔	✗	✔
Ours	✔	✔	✔

Table 1: A concise, categorized overview of the related work.

- Facilitate post-training fine-tuning on test sequences and variable angular view prediction for both view interpolation and extrapolation.
- We show LF video recontruction results on publicly available stereo videos captured in the wild.

2. Related Work

LF super-resolution The past decade saw the rise of commercial LF cameras but quickly faded out of popularity due to the inherent angular and spatial resolution trade-off. Exploiting the correlations in the angular and spatial dimensions, several algorithms have been proposed to overcome this trade-off in LF imaging. Some of these approaches involve modified hardware setups such as coded masks on the aperture [16, 44, 50, 43] and near the sensor [13, 27, 14, 43, 42]. However, the complex optical hardware setups hinder small form factors necessary for consumer devices. Hence, other approaches that use conventional cameras have been proposed such as focal-stack [43] and high-resolution LF reconstruction from sparse measurements [19, 52, 47, 6, 12, 3]. Alternative approaches for a 3D scene such as Multi-Plane Image (MPI) [60, 29, 7] and Neural Radiance Fields (NeRF) [30, 23, 56] have also shown how to generate high-quality LFs. With the evolution of machine learning-based methods to estimate disparity from image semantics in a single image, synthesizing LF images from single images has also been popular [22, 36, 40].

LF video reconstruction While the spatial and angular dimensions of LF have received much attention, commercial LF cameras also suffer from low temporal resolution. A hybrid hardware setup with a commercial LF camera and a DSLR to enable capturing of LF videos at 30 fps was proposed in [45]. A single sensor-based compressive imaging approach that requires a mask near the sensor was proposed in [14]. While these require complex hardware setups, an unpublished manuscript proposes to utilize a single monocular camera for 5D LF video reconstruction [2]. Algorithms such as [45, 14, 2] are learning-based approaches that require supervised training data. As collecting large-scale ground-truth LF videos for training is challenging, [3] proposes X-Fields, a self-supervised approach eliminating the need for supervised training datasets. X-Fields interpolates novel views in both angular and temporal directions. However, the X-Field results in the paper [3] use 4-views and

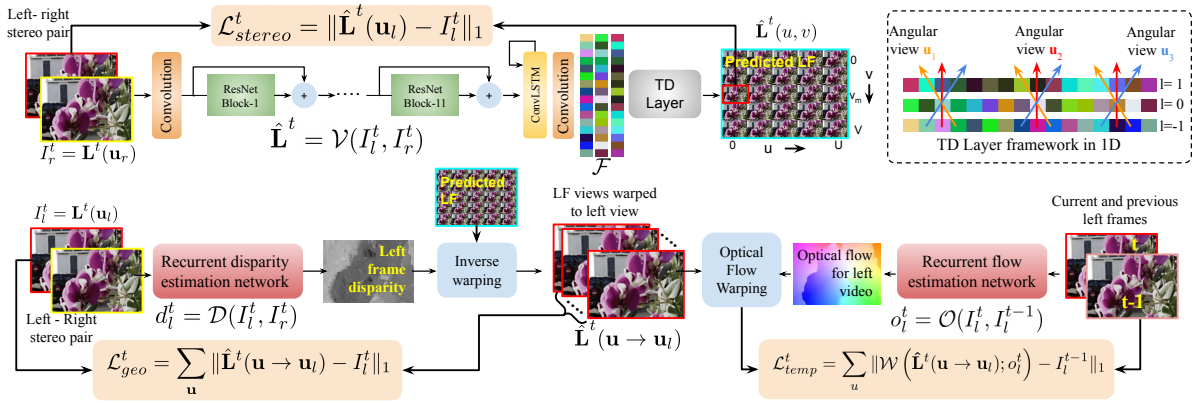

Figure 2: Overall flow of the proposed self-supervised algorithm for LF video reconstruction from stereo video. The LF frames are generated from the input stereo pair via an intermediate low-rank tensor-display (TD) based representation. The self-supervised learning of LF reconstruction is guided via self-supervised cost functions involving stereo pair, disparity maps and optical flow maps.

our experiments in this paper demonstrate that reconstruction quality deteriorates significantly for this method when only two stereo views are available (see Fig. 4, Table 2). We propose a self-supervised algorithm capable of LF reconstruction from only a pair of stereo frames. The distinguishing factor of our work is the rank-constraint on the LF to enforce correlations between horizontal and vertical disparity. This constraint enables high-quality LF reconstruction even when only 1D disparity information is available (e.g., 2 stereo views).

Layered LF displays and neural networks Previously, layered LF display representations have been used in conjunction with neural networks. [26] built an end-to-end pipeline from a coded aperture scene acquisition to displaying the scene on a layered LF display. Similar work in [37, 20] aims at capturing a focal stack and then learning to display the scene onto the LF display. Although layered display representations have been used in conjunction with neural networks, to the best of our knowledge, we are the first ones to use it as a regularizer for self-supervised LF reconstruction.

3. Self-supervised LF Video Reconstruction

In this section, we introduce our self-supervised algorithm for LF video reconstruction from an input stereo video. The input stereo video is assumed to be captured using a pair of rectified, synchronized and identical stereo cameras. A deep recurrent neural network first takes as input an individual stereo pair at the current time-step. It outputs an intermediate low-rank LF representation based on layered LF displays [51] (or Tensor Displays (TD)). A differentiable TD layer then takes this representation as input and generates the corresponding LF frame at the current time-step. Three different self-supervised cost functions based on photometric, geometric, and temporal constraints guide the self-supervised learning for LF reconstruction. The geometric and temporal constraints are imposed by disparity and optical flow maps, respectively. These are obtained via two separate self-supervised recurrent neural networks similar to [8, 28]. Self-supervision of the full 4D LF prediction is explained in Sec. 3.1. Obtaining the LF frame from the intermediate representation is a deterministic linear operation as elaborated in Sec. 3.3.

3.1. Stereo LF estimation

In our proposed algorithm, to obtain the LF video sequence, we estimate the full 4D LF frame for each input pair of stereo frames. Let the required full 4D LF video sequence be denoted by $\mathbf{L}^t(\mathbf{u})$, where $\mathbf{u} = (u, v)$ denotes the 2D coordinates of the LF sub-aperture image (SAI). Here, we assume the input left-right frames, I^t_l and I^t_r as sparse samples of $\mathbf{L}^t(\mathbf{u})$ at SAI co-ordinates $\mathbf{u}_l = (0, v_m)$ and $\mathbf{u}_r = (U, v_m)$, as shown in Fig. 2. Specifically, we have $\mathbf{L}^t(\mathbf{u}_l) = I^t_l$ and $\mathbf{L}^t(\mathbf{u}_r) = I^t_r$. To predict the LF, we use a deep residual neural network [26], $\mathcal{V}$, coupled with a recurrent architecture based on convolutional long short-term memory (ConvLSTM) [34] as shown in Fig. 2. The network $\mathcal{V}$ takes as input the stereo frames (I^t_l, I^t_r) and outputs a low-rank approximation, $\mathcal{F}$, of the desired LF $\hat{\mathbf{L}}^t$. A parameter-free TD layer [26], added after $\mathcal{V}$, takes the representation $\mathcal{F}$ as input and outputs the estimated 4D LF frame $\hat{\mathbf{L}}^t$. We further elaborate on this TD layer in Sec. 3.3 and for now, we assume that $\mathcal{V}$ finally outputs the LF frame $\hat{\mathbf{L}}^t$ from the input frames (I^t_l, I^t_r). As we do not have ground truth LF $\mathbf{L}^t$, we supervise the training of $\mathcal{V}$ by three different self-supervised cost functions based on photometric, geometric and temporal consistency constraints.

Photometric consistency We define the photometric consistency cost as

$$\mathcal{L}^t_{stereo} = \|\hat{\mathbf{L}}^t(\mathbf{u}_l) - I^t_l\|_1 + \|\hat{\mathbf{L}}^t(\mathbf{u}_r) - I^t_r\|_1 \,, \quad (1)$$

which ensures the consistency of $\hat{\mathbf{L}}^t$ with respect to the two known measurements, I_l^t, I_r^t, of $\mathbf{L}^t$.

Geometric consistency The geometric consistency cost enforces $\hat{\mathbf{L}}^t$ to follow the same underlying scene geometry as that of the captured stereo pair. To enforce such a constraint, we first estimate dense disparity maps from the individual input stereo frames via a recurrent neural network $\mathcal{D}$. The network architecture $\mathcal{D}$ is inspired from FlowNet[4] and is augmented with a ConvLSTM network after the encoder network. The disparity maps d_l^t and d_r^t are estimated as,

$$d_l^t = \mathcal{D}\left(I_l^t, I_r^t\right) \quad d_r^t = \mathcal{D}\left(I_r^t, I_l^t\right) \ . \tag{2}$$

As no ground-truth disparity maps are available for supervision, we self-supervise disparity prediction via a photo-consistency based loss [8, 9, 59, 55, 10],

$$\mathcal{L}_{disp}^t = \|\mathcal{W}\left(I_l^t; d_l^t\right) - I_r^t\|_1 + \|\mathcal{W}\left(I_r^t; d_r^t\right) - I_l^t\|_1 \ . \tag{3}$$

Here, $\mathcal{W}$ denotes the bilinear inverse warping operator [17] that takes as input a displacement map and remaps the images. To impose the geometric consitency on $\hat{\mathbf{L}}^t$, we take a SAI $\hat{\mathbf{L}}^t(\mathbf{u})$ at $\mathbf{u}$ and approximate the LF views at $\mathbf{u}_l$ and $\mathbf{u}_r$ via disparity based warping as seen in Fig. 2. But, we already know the ground-truth intensity frame at SAI co-ordinates $\mathbf{u}_l$ and $\mathbf{u}_r$ which are the input stereo frames I_l^t, I_r^t respectively. The error between the approximated and the known input stereo views acts as the supervisory signal for LF estimation. In essence, we warp $\hat{\mathbf{L}}^t(\mathbf{u})$ to the SAIs at $\mathbf{u}_l$ and $\mathbf{u}_r$ to obtain $\hat{\mathbf{L}}^t(\mathbf{u} \to \mathbf{u}_l)$ and $\hat{\mathbf{L}}^t(\mathbf{u} \to \mathbf{u}_r)$ respectively. This can be expressed as,

$$\hat{\mathbf{L}}^t(\mathbf{u} \to \mathbf{u}_r) = \mathcal{W}\left(\hat{\mathbf{L}}^t\left(\mathbf{u}\right); \left(\mathbf{u} - \mathbf{u}_r\right) d_r^t\right) \tag{4}$$

$$\hat{\mathbf{L}}^t(\mathbf{u} \to \mathbf{u}_l) = \mathcal{W}\left(\hat{\mathbf{L}}^t\left(\mathbf{u}\right); \left(\mathbf{u} - \mathbf{u}_l\right) d_l^t\right) \tag{5}$$

The geometric consistency error between the approximated stereo pairs (from the estimated LF) and the known input stereo pairs is then defined as,

$$\mathcal{L}_{geo}^t = \sum_{\mathbf{u}} \sum_{k \in \{l,r\}} \|\hat{\mathbf{L}}^t\left(\mathbf{u} \to \mathbf{u}_k\right) - I_k^t\|_1. \tag{6}$$

Temporal consistency The sequence of estimated LF frames $\hat{\mathbf{L}}^t$ form a video sequence when they are temporally consistent. Here, we use the optical flow estimated from the input sequence of stereo frames to enforce temporal consistency between successive predicted LF frames. With solely the stereo frames as input, it is only possible to estimate optical flow at SAIs $\mathbf{u}_l$ and $\mathbf{u}_r$. We employ a recurrent neural network $\mathcal{O}$ to estimate the optical flows $o_l^t, o_r^t \in R^{h \times w \times 2}$ for the left and right temporal sequences, respectively. The input left-right pairs are input to $\mathcal{O}$ and the optical flow is obtained as

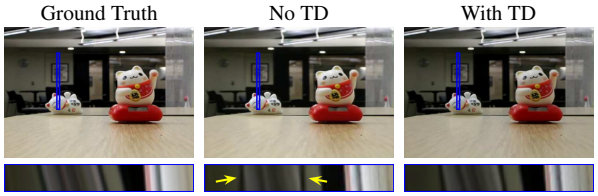

Figure 3: The figure shows epipolar plane image (EPI) for vertical views for a small region of the image. It can be seen that the intermediate representation $\mathcal{F}$ assists in better recovery of the LF frame than direct regression.

$$o_l^t = \mathcal{O}\left(I_l^t, I_l^{t-1}\right) \quad o_r^t = \mathcal{O}\left(I_r^t, I_r^{t-1}\right) \ . \tag{7}$$

Since the ground truth optical flow is unavailable, we choose to learn the optical flow with a self-supervised learning algorithm [28, 32, 49, 48, 18]. We define the photo-consistency based self-supervised cost function [28, 32, 49, 48, 18] for training optical flow network $\mathcal{O}$ as,

$$\mathcal{L}_{flow}^t = \sum\nolimits_{k \in \{l,r\}} \|\mathcal{W}\left(I_k^t; o_k^t\right) - I_k^{t-1}\|_1 \tag{8}$$

where we use $k = l, r$ to sum over both left and right images. To enforce temporal consistency, we utilize the images $\hat{\mathbf{L}}^t\left(\mathbf{u} \to \mathbf{u}_l\right)$ and $\hat{\mathbf{L}}^t\left(\mathbf{u} \to \mathbf{u}_r\right)$ which represent the LF SAIs warped to the stereo SAI co-ordinates $\mathbf{u}_l$ and $\mathbf{u}_r$. With the estimated optical flows o_l^t and o_r^t, $\hat{\mathbf{L}}^t\left(\mathbf{u} \to \mathbf{u}_l\right)$ and $\hat{\mathbf{L}}^t\left(\mathbf{u} \to \mathbf{u}_r\right)$ are warped to approximate the images at the SAIs $\mathbf{u}_l$ and $\mathbf{u}_r$ at the timeframe $t-1$. The SAIs $\mathbf{u}_l$ and $\mathbf{u}_r$ at the timeframe $t-1$ are given by I_l^{t-1} and I_r^{t-1} respectively. The corresponding temporal error is defined as,

$$\mathcal{L}_{temp}^t = \sum_{\mathbf{u}} \sum_{k \in \{l,r\}} \|\mathcal{W}\left(\hat{\mathbf{L}}^t\left(\mathbf{u} \to \mathbf{u}_k; o_k^t\right)\right) - I_k^{t-1}\|_1 \tag{9}$$

where minimizing the error enforces temporal consistency between successive frames.

3.2. Overall loss

We finally add total-variation (TV)-based smoothness constraint on the predicted disparity maps, optical flow and the LF frames. We define the TV smoothness loss as,

$$TV\left(I\right) = \|\nabla_x I\|_1 + \|\nabla_y I\|_1 \ , \tag{10}$$

where ∇_x and ∇_y are the x and y-gradient operators respectively. We define the overall smoothness loss as,

$$\mathcal{L}_{TV}^t = TV\left(\hat{\mathbf{L}}^t\right) + \sum\nolimits_{k \in \{l,r\}} TV\left(d_k^t\right) + TV\left(o_k^t\right). \tag{11}$$

Including all the cost functions, the overall cost function used to optimize the neural networks is defined as,

$$\mathcal{L} = \sum_{t=1}^{T} \lambda_1 \mathcal{L}^t_{disp} + \lambda_2 \mathcal{L}^t_{flow} + \lambda_3 \mathcal{L}^t_{stereo} + \lambda_4 \mathcal{L}^t_{geo} + \lambda_5 \mathcal{L}^t_{temp} + \lambda_6 \mathcal{L}^t_{TV} \ , \quad (12)$$

where T is the total number of frames in the video sequence.

3.3. Low-rank regularization

As elaborated in Sec. 3.1, the LF reconstruction network $\mathcal{V}$ learns to estimate a low-rank representation $\mathcal{F}$ of the desired LF frame. Let's consider the direct estimation of the full 4D LF frame $\hat{\mathbf{L}}^t$ of angular resolution $U \times V$. In this case, $\mathcal{V}$ outputs $U \times V \times 3$ independent channels representing $U \times V$ RGB frames. Such a network design ignores the grid-like structure inherent to a 4D LF frame. Effective utilization of such a structure can lead to a better overall performance of the algorithm. We choose to impose the grid-like structure of the 4D LF frames via the tensor-display [51] based low-rank representation. In Fig. 3 we show that imposing such a low-rank regularizer indeed helps in better recovery of the LF frame. The network $\mathcal{V}$ outputs an intermediate low-rank representation $\mathcal{F} = [\mathbf{f}_{-L/2}, \ldots, \mathbf{f}_0, \ldots, \mathbf{f}_{L/2}]$, where $\mathbf{f}_k = [f^1_k, f^2_k, \ldots, f^M_k]^T$, $f^m_k \in [0,1]^{h \times w \times 3}$ consists of LM RGB channels, where L and M represent the number of layers and the rank, respectively. A linear, parameter-free layer $TD(\cdot)$ takes as input the representation $\mathcal{F}$ and outputs the corresponding LF frame. An intuitive picture of the TD layer is shown in Fig. 2. The operation of $TD(\cdot)$ can be mathematically described as [51],

$$L(x,y,u,v) = TD(\mathcal{F}) = \sum_{m=1}^{M} \prod_{l=-L/2}^{L/2} f^l_m(x + lu, y + lv) \quad (13)$$

where $L(x,y,u,v)$ represents the 4D LF rays, where (x,y) and (u,v) represent the spatial and angular dimensions respectively.

3.4. Implementation details

We employ three different recurrent neural networks, $\mathcal{D}$, $\mathcal{O}$ and $\mathcal{V}$ for predicting disparity maps, optical flow and the 4D LF frames. The LF prediction network $\mathcal{V}$ consists of 1 convolutional layer followed by 11 ResNet blocks [15]. A ConvLSTM [34] layer follows the 11 residual blocks, whose output is then used to predict the intermediate LF representation $\mathcal{F}$. A final convolutional layer outputs the intermediate representation $\mathcal{F}$ with $L = 3$ layers and rank of $M = 12$, *i.e.* a total of 36 RGB channels. We augment the FlowNet [4] architecture with a ConvLSTM [34] after the encoder to form our disparity and flow estimation networks, $\mathcal{D}$ and $\mathcal{O}$. The output of $\mathcal{D}$ and $\mathcal{O}$ consist of 1 and 2 channels respectively. Please refer the supplementary material for the detailed architecture of the neural networks $\mathcal{D}$ and $\mathcal{O}$.

For training our proposed algorithm, we first obtain a LF *image* dataset from [19]. Assuming a static scene, we generate stereo videos by simulating random 6-DoF camera motion through resampling the 4D LF data [25, 35]. The dataset contains a total of 125 LF images, and we generate ten videos of five frames each from each LF image. The camera motion for ten videos is randomly sampled from a pool of 40 simulated camera motions. Hence, in total, we have 1250 stereo video sequences, each with five frames and a spatial resolution of 375×540. More details of the stereo video generation from a given 4D LF image is given in the supplementary material. While training, we obtain a stereo video of 4 frames and randomly crop a patch of size 128×128 from both left and right image pairs. We further augment the data by shifting the focal plane of the stereo images between $[-5, 5]$ pixels. The network is trained in Pytorch [31] using AdamW optimizer [24] for 200 epochs, with an initial learning rate of 0.0001. The learning rate is decreased by $1.1\times$ when the validation loss plateaus for more than 10 epochs. We empirically choose the hyperparameters as $\lambda_1 = 1$, $\lambda_2 = 1$, $\lambda_3 = 0.1$, $\lambda_4 = 1$, $\lambda_5 = 0.1$ and $\lambda_6 = 0.01$ in Eq. (12).

4. Experiments

To validate our proposed algorithm, we perform various experiments on a variety of datasets. For quantitative comparison against the ground truth, we use the *Raytrix* dataset comprising of ground truth LF videos acquired using an industrial LF camera [11]. However, this dataset has only three video sequences with limited scene diversity and a limited angular resolution of 5×5 views. Moreover, it has a maximum disparity of < 2 pixels between adjacent views at a spatial resolution of 1080×1920. To further validate on challenging video sequences from the *Hybrid* video data from [45]. Furthermore, to include more diversity in the scenes, we simulate videos from 15 LF images in the test set of [19] and call this dataset *ViewSynth*. While testing, we obtain the stereo sequences from these datasets and provide them as an input to the network $\mathcal{V}$ and generate the LF sequences. During inference, note that we don't need to estimate the disparity and optical flow maps from $\mathcal{D}$ and $\mathcal{O}$.

4.1. LF video reconstruction

We compare the accuracy of our proposed algorithm with self-supervised [3] and disparity based methods [58, 46, 54, 38, 1, 5]. For each LF test video, we extract a stereo pair from each frame of the sequence. We consider the two extreme SAIs of the central row of the 4D LF frame as the stereo input to our algorithm to estimate the corresponding 4D LF video. We compare our proposed self-supervised algorithm with X-fields [3], also an unsuper-

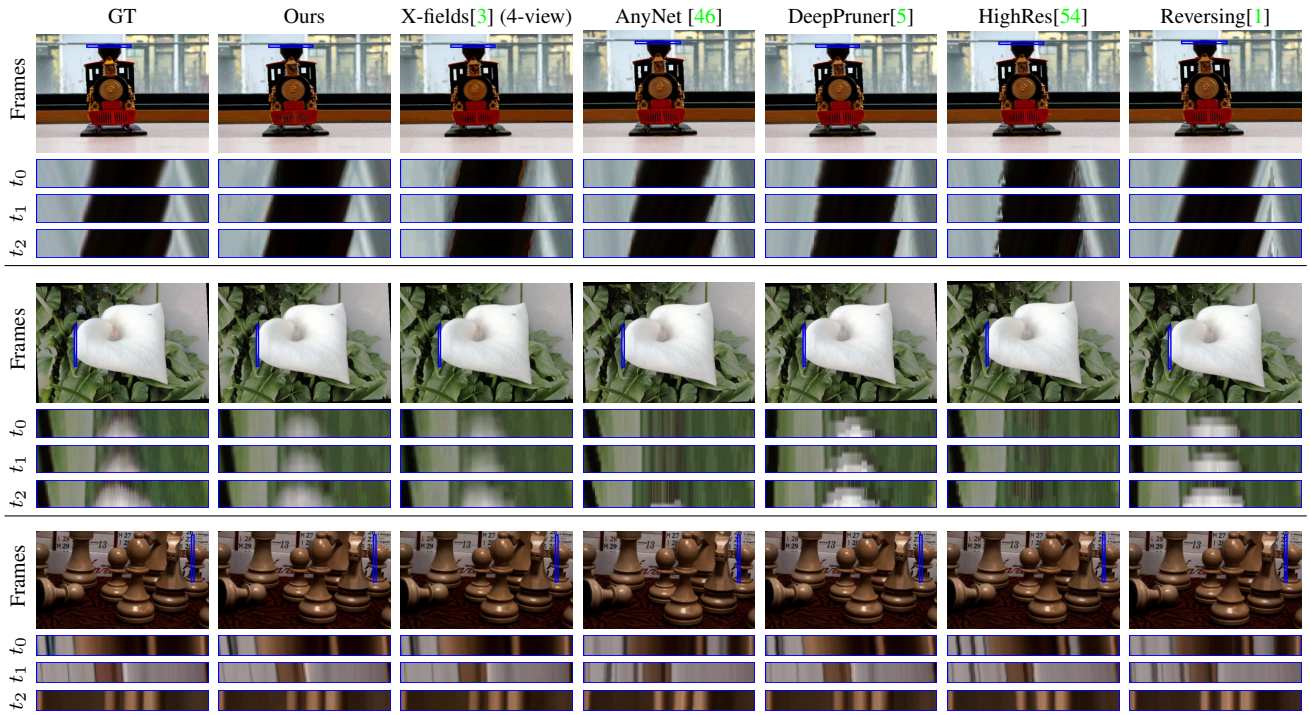

Figure 4: Qualitatively, our algorithm out-performs disparity-based LF prediction techniques. Our proposed algorithm also performs on par with unsupervised LF prediction technique that requires 4 corner views as input.

Datasets	Hybrid		ViewSynth		Raytrix		Average	
	PSNR	LPIPS	PSNR	LPIPS	PSNR	LPIPS	PSNR	LPIPS
AnyNet [46]	27.59	0.070	14.88	0.181	16.35	0.251	19.61	0.167
DeepPruner [5]	30.49	0.068	21.35	0.094	30.98	0.064	27.61	0.075
HighRes [54]	30.79	0.069	25.57	**0.063**	32.59	0.057	29.65	**0.063**
HITNet [38]	30.78	0.070	25.51	**0.078**	32.71	**0.061**	29.67	0.069
Reversing [1]	29.58	**0.061**	13.51	0.262	14.97	0.247	19.35	0.188
X-fields [3] 2-view	25.53	0.089	24.58	0.099	31.24	0.089	27.12	0.092
X-fields [3] 4-view	**31.66**	0.076	**28.21**	0.091	**32.66**	0.095	**30.84**	0.087
Ours	**34.21**	**0.054**	**30.10**	0.122	**35.57**	**0.045**	**33.29**	**0.071**

Table 2: A quantitative comparison of our algorithm against existing algorithms on various datasets. We show that our method outperforms existing methods for self-supervised LF video synthesis. Note that the first five methods require warping. **Blue** and **green** represent the first and second best algorithm in each column.

Model	[46]	[5]	[38]	[1]	[54]	[3] 2-view	[3] 4-view	Ours
Error($\times 10^{-2}$)	2.58	2.50	2.49	2.54	2.43	3.10	**1.73**	**1.58**

Table 3: Mean absolute error (lower is better) obtained after warping successive predicted LF frames via optical flow computed from ground truth LF frames. Our proposed algorithm shows better temporal consistency than other algorithms.

vised algorithm. Since X-fields aim at interpolating views, it fails to generate the full 4D LF from only the stereo views (X-fields (2-view)) as input (Fig. 4). For completeness in our comparisons, we include results for LF generation with the four corner views as input (X-fields (4-view)). We also compare with disparity-based unsupervised LF estimation approach [58], which reconstructs LF via disparity-based warping. Without access to the implementation of [58], we first estimate the disparity from learning-based methods and warp the input views to the LF. We use several state-of-the-art supervised (AnyNet [46], HighRes [54], DeepPruner [5], HITNet [38]) and unupservised (Reversing [1]) stereo disparity estimation algorithms for comparison.

For quantitative comparison, we used two metrics: peak-signal-to-noise ratio (PSNR) (higher is better) and learned perceptual similarity (LPIPS) [57] (lower is better). Table 2 details the quantitative comparisons of various algorithms against all 3 datasets: Raytrix, Hybrid, and ViewSynth. When compared to algorithms that use only 2-views as input, our algorithm outperforms in terms of PSNR. Other algorithms have a slightly better LPIPS [57] metric as their output is just a warped input image and hence tend to be much sharper than the ones generated from our algorithm. However, we can see the real distinction when we compare the images qualitatively in Fig. 4 and especially take into account the EPI for the LF views. Algorithms dependent on disparity-based warping suffer from artifacts arising from incorrect disparity estimation, as seen in Fig. 4. Our proposed algorithm performs consistently better in predicting LF frames as can be seen from both Table 2 and Fig. 4.

Temporal consistency Our proposed algorithm aims to reconstruct LF *video* sequences where temporal consistency is a crucial factor. To establish the temporal consistency, we first predict optical flow for individual ground-truth LF frames [39]. Then the mean absolute error is computed after

Model	TD	$\mathcal{L}_{geo}$	$\mathcal{L}_{temp}$	$\mathcal{L}_{stereo}$	PSNR
$V1$	✔	✔	✔	✗	32.20
$V2$	✔	✔	✗	✔	31.98
$V3$	✔	✗	✔	✔	19.20
$V4$	✗	✗	✔	✔	6.04
$V5$	✗	✔	✔	✔	30.50
Ours	✔	✔	✔	✔	32.39

Table 4: Ablation study of the proposed model with various loss terms from Eq. (12)

Metric	Rank of $\mathcal{F}$ (Layers=3)					$V5$
	1	3	6	9	12	–
PSNR	31.43	32.21	31.87	31.92	32.39	30.50
Time	0.103	0.167	0.248	0.319	0.381	0.108

Table 5: Quantitative comparison of the efficacy of the proposed layered-display regularizer. $V5$, as shown in Table 4, refers to the model where the LF frame is directly output from $\mathcal{V}$ instead of through the intermediate representation $\mathcal{F}$.

warping successive predicted frames via the pseudo-ground truth optical flow. As can be seen in Table 3 our algorithm shows much better temporal consistency.

4.2. Ablation Study

Effect of various loss terms In Table 4, we quantitatively compare our proposed model with its variants based on the loss terms in Eq. (12). The loss terms, $\mathcal{L}_{stereo}$ and $\mathcal{L}_{temp}$ do not have a significant effect on the model performance, but are still important to ensure the photometric and temporal consistencies. Enforcing the epipolar geometric consistency via $\mathcal{L}_{geo}$ is crucial for our task as we observe a significant performance drop in $V3$. However, between $V3$ and $V4$ we observe that the structure imposed by TD layer helps in obtaining reasonable accuracy even in the absence of $\mathcal{L}_{geo}$ term. When using the $\mathcal{L}_{geo}$ constraint, the performance of both without and with TD model, $V5$ and $Ours$ respectively, is enhanced. For $V5$, we modify $\mathcal{V}$ to output 49 RGB frames corresponding to each view of the 7×7 LF frame. Between $V5$ and our proposed model, we observe a PSNR gain of ~1.9dB due to the low-rank intermediate representation. Please refer supplementary material for qualitative comparisons of the various model variants.

Efficacy of layered-display regularizer We study the effect of varying rank configurations ($M = [1, 3, 6, 9, 12]$) for the low-rank representation $\mathcal{F}$, with number of layers fixed to $L = 3$ [51, 26, 37]. The quantitative comparison is shown in Table 5 for 7×7 angular resolution LF output. While the PSNR improves with increasing rank, we also observe a corresponding increase in time complexity. Hence, we use a rank of $M = 12$ for the representation $\mathcal{F}$ in all our experiments, unless stated otherwise. As seen from Table 5, direct regression of LF frame provides the computational advantage but underperforms in terms of PSNR of the output LF. We also see from Fig. 5 that the intermediate representation helps obtain sharper LF reconstructions. More qualitative comparisons can be seen in the supplementary material.

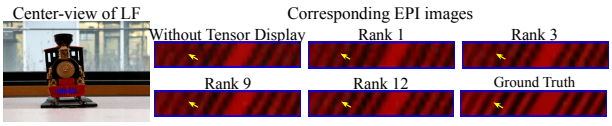

Figure 5: Predicted LF frames are sharper when using higher rank than at lower ranks or not using the low-rank representation at all.

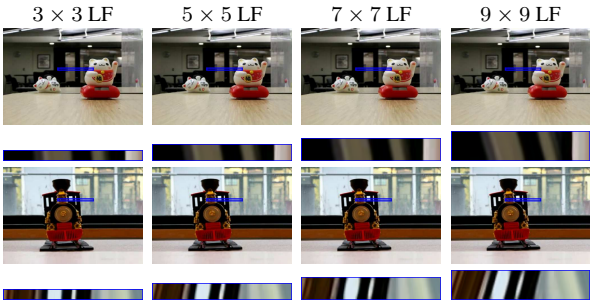

Figure 6: We show the LF sequence predicted at angular resolutions of $3, 5, 7, 9$ and also provide the PSNR / LPIPS metrics where ground truth is available.

4.3. Variable Angular View Prediction

Commercial LF cameras such as Lytro capture 14×14 angular resolution images, where only the central 8×8 views are usable due to vignetting. Unlike supervised techniques, our proposed self-supervised algorithm is not restricted to predict the angular views at the ground-truth angular co-ordinates provided by the LF. In Fig. 6, we demonstrate reconstruction of LF frames with variable angular resolutions such as 3×3, 5×5, 7×7, and 9×9, with our proposed technique. Note that our algorithm allows us to generate frames with higher angular resolution (9×9) than that of the ground-truth frames from Lytro (8×8).

Next, in Fig 7, we demonstrate our algorithm's capability for extrapolating the angular views to new views outside of the input baseline. For extrapolating the views beyond the input baseline, we employ a simple trick: the input stereo views are now assumed to correspond to adjacent horizontal views of the predicted LF frame. We show qualitative results in Fig. 7 where the EPI of the *extended* images show increased slopes, indicating increased disparity between adjacent views compared to original frames.

4.4. Fine-tuning on test sequences

The training procedure for our algorithm is to minimize the overall cost function in Eq. (12), while jointly estimating the LF video, disparity, and optical flow maps from the input stereo video. However, due to domain mismatch, the network can fail to reconstruct reasonable sequences during inference. For such cases, our proposed algorithm allows for *fine-tuning* the neural network on single test sequences. During fine-tuning, the overall cost function in Eq. (12) is minimized with AdamW optimizer for 500 iterations. As can be seen from Fig. 8, fine-tuning consistently improves

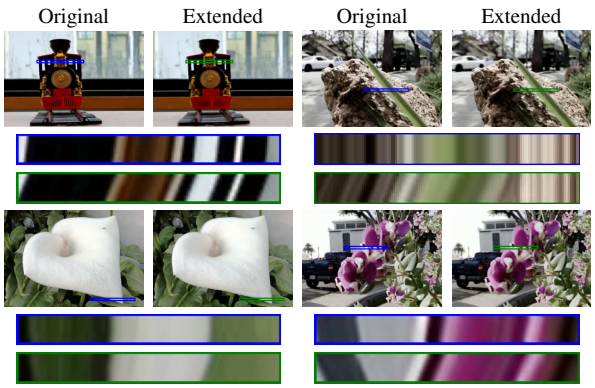

Figure 7: Our proposed self-supervised algorithm can be used to predict novel views beyond the baseline of the input image pair.

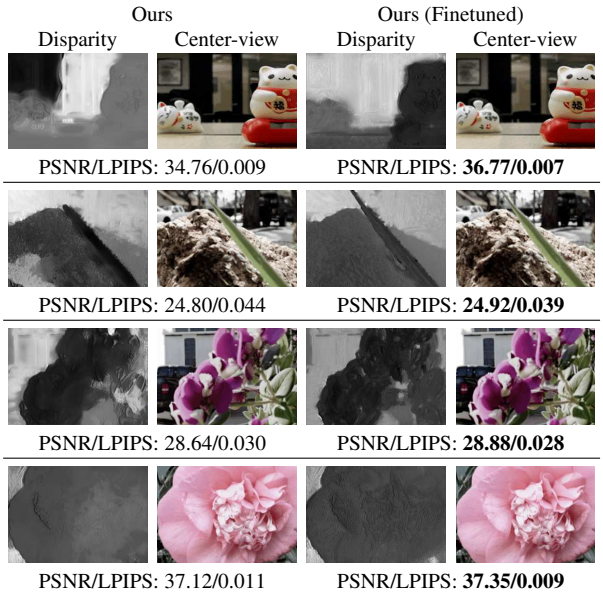

Figure 8: We show the results of finetuning the trained networks on novel test sequences. The first two columns show cases where the network does not perform well initially but we observe significant improvement with finetuning.

the accuracy of the predicted LF sequence while also producing significant qualitative improvements in the reconstructed disparity maps.

4.5. Application to video refocusing

In Fig. 1 we show post-capture focus control on video sequence from [41] acquired using a commercial stereoscopic camera. As the video is acquired with a large baseline (6 cm) stereoscopic camera, we synthetically reduce the baseline by downsampling the spatial resolution to 270×480 from 1080×1920. Each stereo frame is rectified using [53] and the corresponding LF video is generated using our algorithm. The reconstructed LF frames are then used to demonstrate post-capture focus control as seen in Fig. 1. In Fig. 9, we show another instance of post-capture focus control. We extract a stereo video sequence consisting of 8 frames from the LF video dataset in [45]. The proposed algorithm

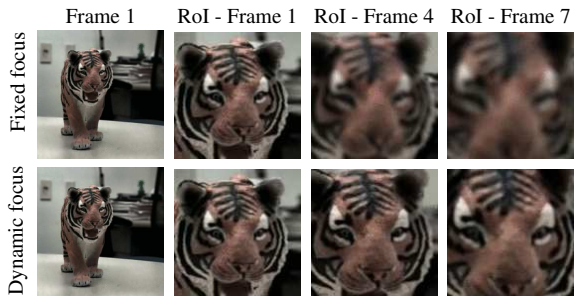

Figure 9: The focal-plane of the video is dynamically adjusted on the toy using the reconstructed LF video.

is used to generate the LF video from the stereo video. The focal plane is fixed in the original video, due to which the toy gets increasingly blurred. However, with our predicted LF video sequence, we can dynamically change the focal plane to be fixed on the toy. More results can be seen in the accompanying supplementary video.

4.6. Discussion

Our proposed algorithm can recover perceptually appealing light-field videos from only a stereo video sequence. With only a stereo video input, there is limited knowledge about the objects being disoccluded in the vertical direction. However, occlusions do not pose a huge challenge because we use a relatively small baseline. The proposed algorithm implicitly learns to inpaint the disoccluded regions. One of the ways to handle occlusions would be to exploit long-range temporal correlations in the input video. Another option would be to use a small corpus of training data for supervised training to handle occlusions.

5. Conclusion

We propose a self-supervised algorithm for light-field video reconstruction from a stereo video. A layered light field display-based low-rank representation is used as a regularizer for guiding the self-supervised reconstruction of light-field frames. The algorithm is applicable for widespread consumer use because we require only a stereo video as input. The proposed self-supervised algorithm confers advantages over supervised learning, such as post-training fine-tuning on test sequences. Other advantages include variable angular view synthesis both between and beyond the input baseline. The reconstructed light-field videos also enable post-capture focus control applications for video sequences.

Acknowledgements

This work was supported in part by NSF CAREER (IIS-1453192, IIS-1652633), RBCDSAI-IITM Research Travel Scholarship, and Qualcomm Innovation Fellowship (QIF) 2021. Prasan is also grateful for the insightful and motivating discussions he had with Subeesh Vasu, Sreyas Mohan, and Matta Gopi Raju.

References

[1] Filippo Aleotti, Fabio Tosi, Li Zhang, Matteo Poggi, and Stefano Mattoccia. Reversing the cycle: self-supervised deep stereo through enhanced monocular distillation. In *European Conference on Computer Vision*, pages 614–632. Springer, 2020.

[2] Kyuho Bae, Andre Ivan, Hajime Nagahara, and In Kyu Park. 5d light field synthesis from a monocular video. *arXiv preprint arXiv:1912.10687*, 2019.

[3] Mojtaba Bemana, Karol Myszkowski, Hans-Peter Seidel, and Tobias Ritschel. X-fields: implicit neural view-, light- and time-image interpolation. *ACM Transactions on Graphics (TOG)*, 39(6):1–15, 2020.

[4] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2758–2766, 2015.

[5] Shivam Duggal, Shenlong Wang, Wei-Chiu Ma, Rui Hu, and Raquel Urtasun. Deeppruner: Learning efficient stereo matching via differentiable patchmatch. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4384–4393, 2019.

[6] Reuben A Farrugia and Christine Guillemot. Light field super-resolution using a low-rank prior and deep convolutional neural networks. *IEEE transactions on pattern analysis and machine intelligence*, 42(5):1162–1175, 2019.

[7] John Flynn, Michael Broxton, Paul Debevec, Matthew DuVall, Graham Fyffe, Ryan Overbeck, Noah Snavely, and Richard Tucker. Deepview: View synthesis with learned gradient descent. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2367–2376, 2019.

[8] Clément Godard, Oisin Mac Aodha, and Gabriel J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, 2017.

[9] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 270–279, 2017.

[10] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3828–3838, 2019.

[11] L Guillo, X Jiang, G Lafruit, and C Guillemot. Light field video dataset captured by a r8 raytrix camera. *Tech. rep., ISO/IEC JTC1/SC29/WG1 & WG11*, 2018.

[12] Mantang Guo, Hao Zhu, Guoqing Zhou, and Qing Wang. Dense light field reconstruction from sparse sampling using residual network. In *Asian Conference on Computer Vision*, pages 50–65. Springer, 2018.

[13] Mayank Gupta, Arjun Jauhari, Kuldeep Kulkarni, Suren Jayasuriya, Alyosha Molnar, and Pavan Turaga. Compressive light field reconstructions using deep learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 11–20, 2017.

[14] Saghi Hajisharif, Ehsan Miandji, Christine Guillemot, and Jonas Unger. Single sensor compressive light field video camera. In *Computer Graphics Forum*, volume 39, pages 463–474. Wiley Online Library, 2020.

[15] 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*, pages 770–778, 2016.

[16] Yasutaka Inagaki, Yuto Kobayashi, Keita Takahashi, Toshiaki Fujii, and Hajime Nagahara. Learning to capture light fields through a coded aperture camera. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 418–434, 2018.

[17] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. *arXiv preprint arXiv:1506.02025*, 2015.

[18] J Yu Jason, Adam W Harley, and Konstantinos G Derpanis. Back to basics: Unsupervised learning of optical flow via brightness constancy and motion smoothness. In *European Conference on Computer Vision*, pages 3–10. Springer, 2016.

[19] Nima Khademi Kalantari, Ting-Chun Wang, and Ravi Ramamoorthi. Learning-based view synthesis for light field cameras. *ACM Transactions on Graphics (TOG)*, 35(6):1–10, 2016.

[20] Yuto Kobayashi, Keita Takahashi, and Toshiaki Fujii. From focal stacks to tensor display: A method for light field visualization without multi-view images. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2007–2011. IEEE, 2017.

[21] Wei-Sheng Lai, Jia-Bin Huang, Oliver Wang, Eli Shechtman, Ersin Yumer, and Ming-Hsuan Yang. Learning blind video temporal consistency. In *Proceedings of the European conference on computer vision (ECCV)*, pages 170–185, 2018.

[22] Qinbo Li and Nima Khademi Kalantari. Synthesizing light field from a single image with variable mpi and two network fusion. *ACM Transactions on Graphics (TOG)*, 39(6):1–10, 2020.

[23] Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. Neural sparse voxel fields, 2021.

[24] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[25] Jonathan Samuel Lumentut, Tae Hyun Kim, Ravi Ramamoorthi, and In Kyu Park. Fast and full-resolution light field deblurring using a deep neural network. *arXiv preprint arXiv:1904.00352*, 2019.

[26] Keita Maruyama, Yasutaka Inagaki, Keita Takahashi, Toshiaki Fujii, and Hajime Nagahara. A 3-d display pipeline from coded-aperture camera to tensor light-field display through cnn. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 1064–1068. IEEE, 2019.

[27] Kshitij Marwah, Gordon Wetzstein, Yosuke Bando, and Ramesh Raskar. Compressive light field photography using overcomplete dictionaries and optimized projections. *ACM Transactions on Graphics (TOG)*, 32(4):1–12, 2013.

[28] Simon Meister, Junhwa Hur, and Stefan Roth. Unflow: Unsupervised learning of optical flow with a bidirectional census loss. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[29] Ben Mildenhall, Pratul P Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics (TOG)*, 38(4):1–14, 2019.

[30] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020.

[31] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[32] Zhe Ren, Junchi Yan, Bingbing Ni, Bin Liu, Xiaokang Yang, and Hongyuan Zha. Unsupervised deep learning for optical flow estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

[33] Kohei Sakai, Keita Takahashi, Toshiaki Fujii, and Hajime Nagahara. Acquiring dynamic light fields through coded aperture camera. In *ECCV (19)*, pages 368–385, 2020.

[34] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 802–810, 2015.

[35] Pratul P Srinivasan, Ren Ng, and Ravi Ramamoorthi. Light field blind motion deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3958–3966, 2017.

[36] Pratul P Srinivasan, Tongzhou Wang, Ashwin Sreelal, Ravi Ramamoorthi, and Ren Ng. Learning to synthesize a 4d rgbd light field from a single image. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2243–2251, 2017.

[37] Keita Takahashi, Yuto Kobayashi, and Toshiaki Fujii. From focal stack to tensor light-field display. *IEEE Transactions on Image Processing*, 27(9):4571–4584, 2018.

[38] Vladimir Tankovich, Christian Häne, Sean Fanello, Yinda Zhang, Shahram Izadi, and Sofien Bouaziz. Hitnet: Hierarchical iterative tile refinement network for real-time stereo matching. *arXiv preprint arXiv:2007.12140*, 2020.

[39] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European Conference on Computer Vision*, pages 402–419. Springer, 2020.

[40] Richard Tucker and Noah Snavely. Single-view view synthesis with multiplane images. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[41] Matthieu Urvoy, Marcus Barkowsky, Romain Cousseau, Yao Koudota, Vincent Ricorde, Patrick Le Callet, Jesus Gutierrez, and Narciso Garcia. Nama3ds1-cospad1: Subjective video quality assessment database on coding conditions introducing freely available high quality 3d stereoscopic sequences. In *2012 Fourth International Workshop on Quality of Multimedia Experience*, pages 109–114. IEEE, 2012.

[42] Anil Kumar Vadathya, Saikiran Cholleti, Gautham Ramajayam, Vijayalakshmi Kanchana, and Kaushik Mitra. Learning light field reconstruction from a single coded image. In *2017 4th IAPR Asian Conference on Pattern Recognition (ACPR)*, pages 328–333. IEEE, 2017.

[43] Anil Kumar Vadathya, Sharath Girish, and Kaushik Mitra. A unified learning-based framework for light field reconstruction from coded projections. *IEEE Transactions on Computational Imaging*, 6:304–316, 2019.

[44] Ashok Veeraraghavan, Ramesh Raskar, Amit Agrawal, Ankit Mohan, and Jack Tumblin. Dappled photography: Mask enhanced cameras for heterodyned light fields and coded aperture refocusing. *ACM Trans. Graph.*, 26(3):69, 2007.

[45] Ting-Chun Wang, Jun-Yan Zhu, Nima Khademi Kalantari, Alexei A Efros, and Ravi Ramamoorthi. Light field video capture using a learning-based hybrid imaging system. *ACM Transactions on Graphics (TOG)*, 36(4):1–13, 2017.

[46] Yan Wang, Zihang Lai, Gao Huang, Brian H Wang, Laurens Van Der Maaten, Mark Campbell, and Kilian Q Weinberger. Anytime stereo image depth estimation on mobile devices. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 5893–5900. IEEE, 2019.

[47] Yunlong Wang, Fei Liu, Zilei Wang, Guangqi Hou, Zhenan Sun, and Tieniu Tan. End-to-end view synthesis for light field imaging with pseudo 4dcnn. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 333–348, 2018.

[48] Yang Wang, Peng Wang, Zhenheng Yang, Chenxu Luo, Yi Yang, and Wei Xu. Unos: Unified unsupervised optical-flow and stereo-depth estimation by watching videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8071–8081, 2019.

[49] Yang Wang, Yi Yang, Zhenheng Yang, Liang Zhao, Peng Wang, and Wei Xu. Occlusion aware unsupervised learning of optical flow. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4884–4893, 2018.

[50] Yu-Ping Wang, Li-Chun Wang, De-Hui Kong, and Bao-Cai Yin. High-resolution light field capture with coded aperture. *IEEE Transactions on Image Processing*, 24(12):5609–5618, 2015.

[51] G. Wetzstein, D. Lanman, M. Hirsch, and R. Raskar. Tensor Displays: Compressive Light Field Synthesis using Multilayer Displays with Directional Backlighting. *ACM Trans. Graph. (Proc. SIGGRAPH)*, 31(4):1–11, 2012.

[52] Gaochang Wu, Mandan Zhao, Liangyong Wang, Qionghai Dai, Tianyou Chai, and Yebin Liu. Light field reconstruction using deep convolutional network on epi. In *Proceed-*

ings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 6319–6327, 2017.

[53] Ruichao Xiao, Wenxiu Sun, Jiahao Pang, Qiong Yan, and Jimmy Ren. Dsr: Direct self-rectification for uncalibrated dual-lens cameras. *3DV*, 2018.

[54] Gengshan Yang, Joshua Manela, Michael Happold, and Deva Ramanan. Hierarchical deep stereo matching on high-resolution images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5515–5524, 2019.

[55] Zhichao Yin and Jianping Shi. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1983–1992, 2018.

[56] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. Nerf++: Analyzing and improving neural radiance fields, 2020.

[57] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[58] Zhoutong Zhang, Yebin Liu, and Qionghai Dai. Light field from micro-baseline image pair. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3800–3809, 2015.

[59] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1851–1858, 2017.

[60] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. In *SIGGRAPH*, 2018.