Supplementary Material for SeLFVi: Self-supervised Light-Field Video Reconstruction from Stereo Video

Prasan Shedligeri[1], Florian Schiffers[2], Sushobhan Ghosh[2], Oliver Cossairt[2], and Kaushik Mitra[1]

[1]IIT Madras, India [2]Northwestern University, USA

A More Details Regarding Experimental Comparisons Performed

Our proposed self-supervised algorithm is designed to take a stereo video as input and output a light field (LF) video sequence. We evaluate our algorithm against other self-supervised [1] and disparity-based LF generation algorithms [2, 3, 4, 5, 6]. For X-fields [1], we use their publicly available implementation [7] to make our comparisons. As X-fields [1] is mainly designed for interpolation, we make the comparison *X-fields (4-view)* by providing the 4 corner views of the LF as the input. Each novel view of the LF is estimated using the 3 neighboring views with the network capacity multiplier set to 8 (a higher value resulted in GPU memory error). For a fairer comparison, we also run X-fields with only a stereo pair as input. As X-fields only interpolates between the input coordinates, it's not possible to generate the full 4D LF with only stereo pair as input and call it *X-fields (2-view)*. Hence, we employ a trick of duplicating the stereo pair as the corner views of the LF with the baseline in the v axis of the LF set to zero. As can be seen in the accompanying supplementary video, this trick works well for some scenes while failing completely for others. Hence, we mainly compare our algorithm against the *X-fields (4-view)* variant and use the *(2-view)* variant as only a baseline. Further comparisons are shown in the supplementary video, where we observe the 4-view variant fail to generate good LF views in some challenging cases. In the same video, we also observe that X-fields (4-view) performs better than our proposed technique and we discuss this in Sec. E.

For further comparisons, we generate LF frames via warping the input views using stereo disparity [2, 3, 4, 5, 6] as proposed in [8]. We consider various state-of-the-art supervised [2, 3, 4, 6] and unsupervised [5] algorithms for estimating disparity from the input stereo pair. Using this disparity, the input views are then warped to the LF views with the assumption that disparity remains the same in both horizontal and vertical directions. Due to the small baseline of the input views, there are no large holes in the output frames and the small holes due to warping are filled via interpolation.

B Details of Our Proposed Network Architecture

Here, we provide the details of the 3 different network architectures $\mathcal{V}$, $\mathcal{D}$ and $\mathcal{O}$.

Light field prediction network, $\mathcal{V}$ The LF prediction network consists of an input convolutional layer followed by 11 ResNet blocks [9]. The input convolutional layer takes as input a stereo frame (6 channels) and outputs a 64 channel feature map, convolving with a kernel of size 3×3, without any spatial downsampling. This feature map is then input to the ResNet block where the number of channels at the output is kept the same as that of the input (here, 64 channels). Each ResNet block consists of 2 convolutional layers followed by the rectified linear unit (ReLU) [10] activation. In each ResNet block, the first convolutional layer takes the 64 channel feature as input and outputs a 32 channeled feature map. The second convolutional layer takes this intermediate 32 channel feature map as input and outputs again a 64 channel feature map. There is no spatial downsampling or upsampling within the ResNet blocks. The feature map at the output of the 11^{th} ResNet block is then input to a convolutional long short-term memory (ConvLSTM) network [11]. The cell state of this ConvLSTM network is then input to a final convolutional layer which outputs 36 RGB (108) channels corresponding to the $L = 3$ layers and $M = 12$ rank of the low-rank LF representation $\mathcal{F}$. ReLU non-linearity is used at the output of the final convolutional layer to ensure non-negative values in $\mathcal{F}$.

Disparity and optical flow estimation network, $\mathcal{D}$ and $\mathcal{O}$ As shown in Fig. 1, the neural networks $\mathcal{D}$ and $\mathcal{O}$ are derived from the FlowNet [12] network architecture. Although both networks, $\mathcal{D}$ and $\mathcal{O}$, share similar network architecture, the weights are completely independent and are not shared between the two networks. To facilitate temporal consistency in the predicted outputs, a ConvLSTM network is used after the encoder block, following [13]. The major differences between the two networks are in the correlation layer [12] and the final output convolutional layer. The *correlation layer* which computes the cost volume between the two feature maps has 6 parameters [14]: kernel-size, patch-size, stride, padding, dilation, and dilation-patch. The details of these parameters for both the

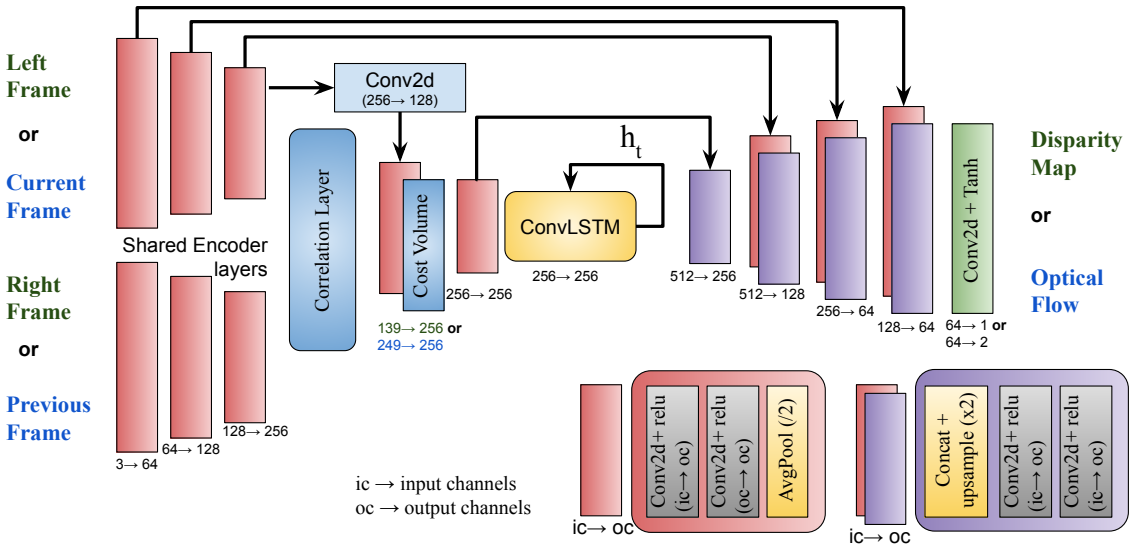

Figure 1: For estimation of the disparity map and optical flow, we modify the FlowNet [12] architecture to include the ConvLSTM network [11] at the encoder. All the layers in the neural network use 2D convolutional layers with kernel size of 3×3.

networks, $\mathcal{D}$, and $\mathcal{O}$, are provided in Table 1. Other network details such as the number of channels per layer are provided in Fig. 1. The final convolutional layer of both the networks uses the tanh non-linearity as shown in Fig. 1 While the disparity estimation network $\mathcal{D}$ outputs a single channel, the flow estimation network $\mathcal{O}$ outputs two channels.

C Generating stereo *video* from a 4D LF *image*

Consider a 4D LF image of the form $L(x, y, u, v)$ where (x, y) are the spatial co-ordinates and (u, v) are the angular co-ordinates. While simulating the video sequence, we assume a model of multiple pinhole cameras located at the co-ordinates (u, v) from which individual views of the LF are captured. Simulating a camera motion through the given light-field is equivalent to resampling the given 4D light-field function and projecting it to the desired camera [15, 16]. We consider the 6-DoF camera motion with translation and rotation defined as $P(t) = [p_x(t), p_y(t), p_z(t)]$ and $R(t) = [\theta_x(t), \theta_y(t), \theta_z(t)]$, respectively. We consider the stereo camera located at the two views $(0, v_m)$ and (U, v_m). For the given 6-DoF translation and rotation $P(t)$ and $R(t)$ respectively, the left view at time t is given by,

$$I_l^t = L(x^j, y^j, p_x^i(t) - x^j p_z(t), v_m + p_y^i(t) - y^j p_z(t)) \tag{1}$$
$$x^j = (x - U/2)\cos\theta_z(t) - y\sin(\theta_z(t)) + U/2 \tag{2}$$
$$y^j = (x - U/2)\sin\theta_z(t) + y\cos(\theta_z(t)) \tag{3}$$
$$p_x^i(t) = p_x(t) + f\theta_x(t) \tag{4}$$
$$p_y^i(t) = p_y(t) + f\theta_y(t) \tag{5}$$

where f is the focal length of the camera. Similarly, the right view of the camera is given by,

$$I_r^t = L(x^j, y^j, U + p_x^i(t) - x^j p_z(t), v_m + p_y^i(t) - y^j p_z(t)) \tag{6}$$

We refer the readers to [16] for a detailed derivation of the above equations.

While we only require stereo videos for training, we require ground-truth LF video in order to quantitatively evaluate the estimated LF videos during testing. For this, we generate full 5D LF videos from a single 4D LF image. The LF video generation process follows that of the stereo video generation. The video generation process described above is repeated across all the views of the LF instead of just 2 extreme views for the stereo video.

D More Qualitative Results

Ablation study In Table 4 of the manuscript, we quantitatively compare 5 different variants of our model. Here, in Fig. 2, we make qualitative comparisons for some of the important variants in Table 4 of the manuscript. The

Network	kernel-size	patch-size	stride	padding	dilation	dilation-patch
$\mathcal{O}$	1×1	11×11	1	0	1	2
$\mathcal{D}$	1×1	1×11	1	0	1	2

Table 1: Values that are used for the different parameters of the correlation layer [14] in the neural networks $\mathcal{O}$ and $\mathcal{D}$.

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

Table 2: Ablation study of the proposed model with various loss terms from Eq. (12)

relevant quantitative comparisons are shown again in Table 2. As we observe from the epipolar plane image (EPI) in Fig. 2 most of the reconstructed frames in $V4$ are zero due to the absence of both the low-rank representation $\mathcal{F}$ and the geometric consistency term $\mathcal{L}_{geo}$. This shows the importance of the intermediate representation $\mathcal{F}$ in the absence of geometric consistency cost, $\mathcal{L}_{geo}$. Comparing $V3$ and Ours, the importance of the epipolar consistency term $\mathcal{L}_{geo}$ is demonstrated. In $V3$, the low-rank representation $\mathcal{F}$ imposes the inherent structure of LF on the output. This ensures that the output frames are reasonably close to the actual LF frames. However, the additional geometric consistency term $\mathcal{L}_{geo}$ in our proposed model provides accurate reconstructions as can be seen from the EPI.

Efficacy of the intermediate low-rank representation In Fig. 3 we qualitatively compare the reconstruction performance in the presence (Ours) and absence ($V5$ in Table 2) of the intermediate low-rank representation $\mathcal{F}$. The training is done with the full loss function as described in Eq. (12) of the manuscript. We observe that the reconstructed LF frames are significantly blurred when we directly predict the LF frame as the output of the network $\mathcal{V}$.

E Discussion on Results

Comparison with X-fields (4-view) [1] On some sequences shown in the supplementary materia, X-fields (4-view) [1] achieves better results for two important reasons. One, X-fields uses both horizontal and vertical disparity information to produce the light-fields. Our technique however has only the horizontal disparity information from the stereo image. Second, it is trained over one particular sequence and is hence expected to perform better. X-fields is certainly a more versatile technique allowing for *interpolation* in time, view and light directions. Our work is a complementary technique to X-fields: we allow for finetuning the reconstruction on a particular sequence, while also utilizing data-driven approaches to improve performance in a way that generalizes well to arbitrary scenes. Our work also provides a technique for *extrapolation*, which X-fields is not designed to handle currently.

Loss in spatial frequency We observe that there's a loss in spatial details for some of the sequences shown in the supplementary material video and in Fig. 4. While we observe blurring in some of our reconstructed sequences, it is not a fundamental limitation of our overall technique. Incorporating detail-preserving losses on top of the low-rank regularizer can preserve high-frequency details. For instance, a low-rank+sparse decomposition model for LF, combined with a perceptual loss, could help recover the high-frequency details. As we see in Fig. 4, the spatial frequency details can be restored to a reasonable accuracy.

References

[1] 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.

[2] 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.

[3] 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.

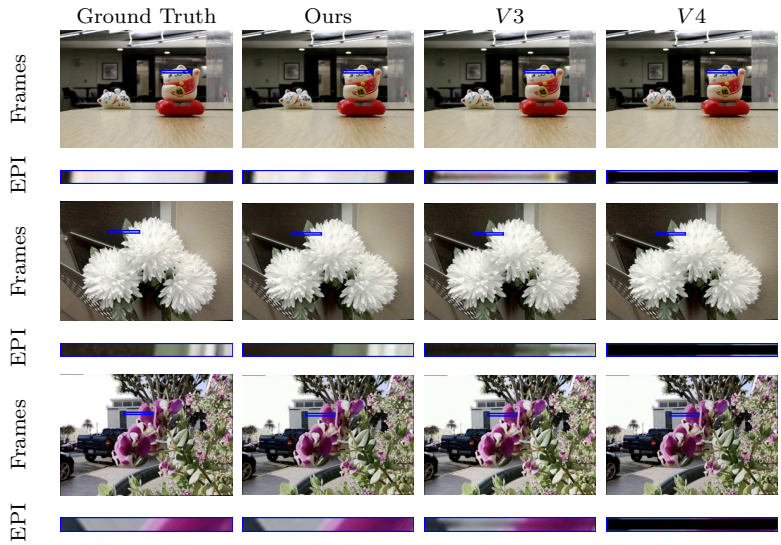

Figure 2: We show the qualitative comparisons for two important configurations of our proposed network architecture, $V3$ and $V4$. The low-rank representation $\mathcal{F}$, inherently imposes the structure of LF in $V3$ producing reasonable reconstructions. However, in the absence of the representation $\mathcal{F}$, most frames predicted by $V4$ are zero. Further, enforcing explicit geometric consistency via $\mathcal{L}_{geo}$ produces significantly better reconstructions as can be seen in the second column.

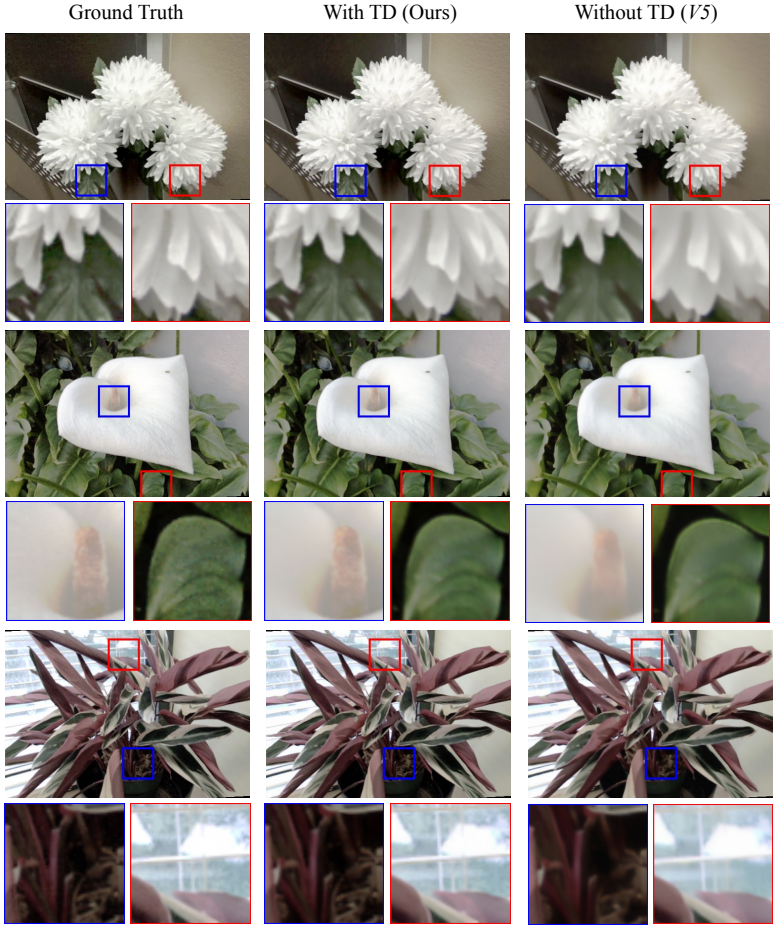

Figure 3: We qualitatively compare the reconstruction performance in the presence (With TD) and absence (Without TD) of the intermediate low-rank representation $\mathcal{F}$. We observe significant blurring in the reconstructed images when not using the low-rank representation.

[4] 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.

[5] 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.

[6] 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.

[7] Source code for x-fields (siggraph asia 2020). https://github.com/m-bemana/xfields. Accessed: 2021-02-23.

[8] 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.

[9] 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.

[10] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *ICML*, 2010.

[11] 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.

[12] 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.

[13] 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.

[14] Pytorch correlation module. https://github.com/ClementPinard/Pytorch-Correlation-extension. Accessed: 2021-01-03.

[15] 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.

[16] J. S. Lumentut, T. H. Kim, R. Ramamoorthi, and I. K. Park. Deep recurrent network for fast and full-resolution light field deblurring. *IEEE Signal Processing Letters*, 26(12):1788–1792, 2019.

[17] 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.

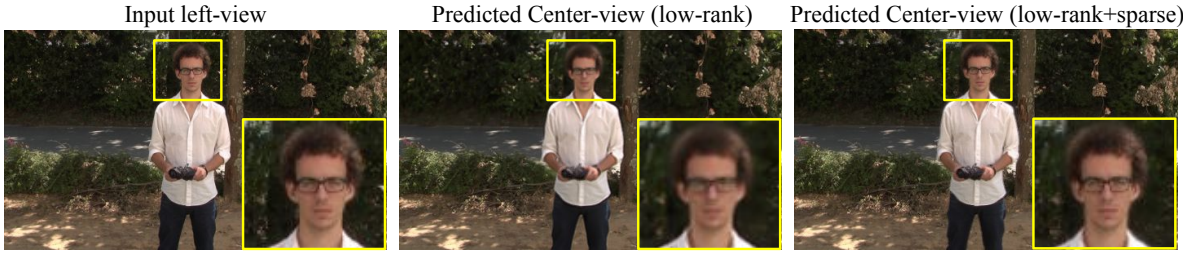

Figure 4: As seen in middle image, we observe loss of spatial details in the reconstructed frames for some video sequences. However, this is not a fundamental limitation of the proposed model. We observe in the right image that the details can be recovered by the use of detail preserving perceptual cost metrics such as LPIPS [17] and low-rank+sparse decomposition model.