U-shaped Networks for Shape from Light Field

Stefan Heber[1]
stefan.heber@icg.tugraz.at
Wei Yu[1]
wei.yu@icg.tugraz.at
Thomas Pock[1,2]
pock@icg.tugraz.at

[1] Graz University of Technology
Graz, Austria
[2] Austrian Institute of Technology
Vienna, Austria

Abstract

This paper presents a novel technique for Shape from Light Field (SfLF), that utilizes deep learning strategies. Our model is based on a fully convolutional network, that involves two symmetric parts, an encoding and a decoding part, leading to a u-shaped network architecture. By leveraging a recently proposed Light Field (LF) dataset, we are able to effectively train our model using supervised training. To process an entire LF we split the LF data into the corresponding Epipolar Plane Image (EPI) representation and predict each EPI separately. This strategy provides good reconstruction results combined with a fast prediction time. In the experimental section we compare our method to the state of the art. The method performs well in terms of depth accuracy, and is able to outperform competing methods in terms of prediction time by a large margin.

1 Introduction

In this paper we investigate the problem of estimating depth information for given Light Field (LF) data. This problem is also referred to as Shape from Light Field (SfLF). A LF [7, 19] is a 4D function that provides in addition to the spatial information, that corresponds to the information of a traditional 2D image, also directional information. The additional directional information includes information about the geometry of the observed scene, and thus gave rise to interesting applications, like for instance digital re-focusing [13, 22], digital viewpoint manipulation [22], or depth estimation [6, 10, 12, 14, 27, 28]. All of these tasks are basically impossible to realize given a single traditional 2D image, that only provides the spatial intensity information.

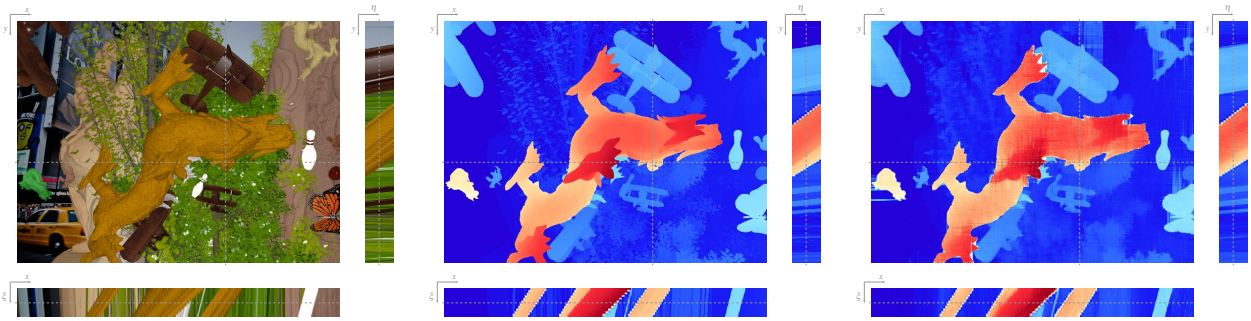
A LF is commonly described using the so-called two-plane parametrization. This type of parametrization defines a ray by the intersection points of two parallel planes. Those planes are referred to as image plane $\Omega \subseteq \mathbb{R}^2$ and lens plane $\Pi \subseteq \mathbb{R}^2$. Thus in mathematical terms the LF is given as

$$L : \Omega \times \Pi \to \mathbb{R}, \qquad (\mathbf{p}, \mathbf{q}) \mapsto L(\mathbf{p}, \mathbf{q}), \tag{1}$$

where $\mathbf{p} = (x, y)^\top \in \Omega$ and $\mathbf{q} = (\xi, \eta)^\top \in \Pi$ represent the spatial and directional coordinates.

© 2016. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.

Pages 37.1-37.12
DOI: https://dx.doi.org/10.5244/C.30.37



(a) LF (b) ground truth (c) proposed

Figure 1: Illustration of LF data. (a) shows a sub-aperture image with vertical and horizontal EPIs. The EPIs correspond to the positions indicated with dashed lines in the sub-aperture image. (b) shows the corresponding color-coded ground truth disparity field. (c) shows the result of the proposed model.

There are different ways of visualizing the 4D LF. In this work we use the so-called Epipolar Plane Image (EPI) representation. In terms of Equation (1) an EPI is obtained by holding one spatial and one directional coordinate constant. For instance by choosing a certain y and a certain η we restrict the 4D LF to the 2D function

$$\Sigma_{y,\eta} : \mathbb{R}^2 \to \mathbb{R}, \qquad (x,\xi) \mapsto L(x,y,\xi,\eta)\,, \tag{2}$$

that defines a horizontal EPI. In a similar way one can also define vertical EPIs. The EPI representation can be considered as a 2D slice through the 4D LF, and it illustrates the linear characteristic of the LF space. See Figure 1(a) for an illustration.

In this work we aim at automatically converting EPIs to corresponding disparity images. Our approach, based on fully Convolutional Neural Networks (CNNs) [20], consists of processing an EPI with a series of convolution operations, that are able to detect line orientations. Knowing the line orientations allows to reconstruct the geometry of the observed scene. The kernels used for the involved convolutions are learned by leveraging a LF dataset that was recently presented in [11]. The proposed data-driven approach has two main advantages compared to prevailing methods: First, it allows to learn necessary heuristics from the training data to cope with artifacts due to, for instance, occlusion and aliasing. Secondly, the convolutions can be implemented efficiently on the GPU allowing for fast prediction times.

2 Related Work

One of the most important research topics in LF image processing is the development of efficient and reliable shape extraction methods. Those methods are the foundation of various applications, like for instance digital refocusing [13, 22], image segmentation [30], or super-resolution [2, 29], to name but a few. The main focus of research regarding Shape from Light Field (SfLF) lies on developing methods to accurately reconstruct the observed scene at depth discontinuities or occlusion boundaries. For this purpose various approaches have been proposed, including specialized multi-view stereo techniques [3, 12] and methods based on an EPI analysis [6, 28]. Wanner and Goldluecke [6, 28] used for example the 2D structure tensor to measure the direction of each position in the vertical and horizontal EPIs. The results are then fused using variational methods by incorporating additional global visibility constraints. In [12] Heber et al. proposed a variational multi-view stereo method based on

a technique called Active Wavefront Sampling (AWS). Tao et al. [27] proposed a fusion method that uses correspondence and defocus cues. Chen et al. [3] introduced a bilateral consistency metric on the surface camera to indicate the probability of occlusions, which was further used for LF stereo matching. Heber and Pock [10] proposed a variational method, that shears the LF by applying a low-rank assumption, where the depth information is provided by the amount of shearing. Jeon et al. [14] proposed an algorithm for SfLF, that utilizes phase shift based subpixel displacements. In [11] Heber and Pock presented a method for SfLF that applies a conventional CNN in a sliding window fashion. Up to this point deep learning techniques were barely used in LF image processing. Utilizing trained models for SfLF is an interesting idea to address certain limitations of previous methods. On the one hand a trained model has the ability to learn how to handle occlusion and aliasing artifacts, and on the other hand a CNN also allows faster computation times.

In this paper we seize ideas presented in [11]. Furthermore this work also builds upon fully convolutional networks [20] and up-convolution-based approaches [4, 20, 32], i.e. the proposed network architecture consists of a contracting and an expanding path, that involve only convolutional layers. The former path compresses the information and simultaneously captures context, and the latter path extracts the information and up-samples it to the original size. The expanding path is more or less symmetric to the contracting path, yielding a u-shaped architecture, that can be trained in an end-to-end scheme.

3 Contribution

The proposed method is inspired by the method of Heber and Pock [11], that uses a conventional CNN in a sliding window fashion to predict depth information. They showed that CNNs have a large capacity to learn from data to predict the orientation of the lines in the EPIs. However, due to the sliding window approach, their method suffers from considerable high computational costs. Compared to [11] we were able to significantly reduce the computation time by predicting complete EPIs at once using u-shaped networks. Besides drastically reducing the prediction time the proposed network architecture also allows to reduce the errors in homogeneous regions, because the proposed model can overcome the patch-nature of the network proposed in [11]. Our experiments demonstrate that the proposed method is able to predict an entire 4D disparity field within a few seconds. Moreover, due to the fact that our network architecture does not include any fully connected layer, our method also allows to process LFs with varying resolutions.

4 Methodology

In this section we describe the methodology of the proposed approach. The success of the proposed CNN depends on leveraging a set of recent improvements, that include up-convolutions [20], no explicit pooling [26], and the Adam optimization method [15]. The section starts with a short introduction to CNNs, followed by the description of the used u-shaped network architecture. At the end of the section we provide details regarding the network training and the leveraged dataset.

Convolutional Neural Networks. In the late 1980s, Yann LeCun et al. [17, 18] introduced a special type of multi-layer Neural Networks (NNs), where weights are shared across layers. By sharing the weights they were able to resemble an important operation in signal

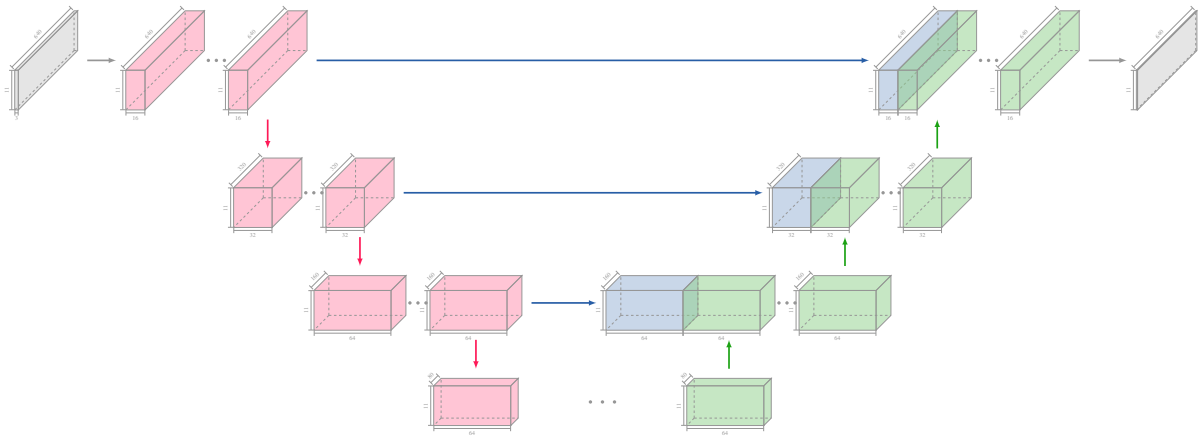


Figure 2: Illustration of the proposed u-shaped network architecture. The encoding and decoding parts of the network are highlighted in purple and green, respectively. The pinhole connections are marked in blue.

processing known as convolution, leading to the CNN architecture. A CNN consists of several layers, where the different layers are connected such that layer l creates the input for layer $l+1$. Layer l can be seen as a multi-channel image of size $H_l \times W_l \times C_l$, where H_l, W_l and C_l denote image height, image width, and number of channels of the l^{th} layer, respectively. The first and last layer are called input and output layers, respectively. Hence their size also corresponds to the input size and to the desired output size. Successive layers are connected via a convolutional mapping with an additional additive bias term, i.e. each channel of the layer $l+1$ is defined as a convolution with a kernel of size $k_h \times k_w \times C_l$ followed by the addition of a constant bias, where k_h and k_w denote the kernel width and height.

Yann LeCun [18] introduced CNNs trained in a supervised manner via back-propagation. Since Krizhevsky et al. [16] utilized CNNs effectively for the task of large-scale image classification the popularity of CNNs or deep learning techniques increased drastically in the computer vision literature. Nowadays CNNs are especially popular in image classification and objection recognition [9, 24]. The entire field of deep learning flourishes with innovations, one after another. However, the exploration of 4D LF data by CNNs is still limited.

Network Architecture. In contrast to methods that use natural images we are not able to exploit existing trained networks, i.e. we opt for designing our network entirely from scratch. However, not relying on pre-trained networks also allows to better adapt the network structure to the problem at hand. The proposed network is a fully convolutional network consisting of a contracting part and an expanding part. The first part acts as an encoder, that spatially compresses the image and thus reduces the input data to an essential feature representation. The bottom part processes the essential features, before the expanding part of the network decodes the simple feature representation to an output disparity image. The encoding and decoding parts of the network are basically symmetric leading to an u-shaped network architecture. An overview of the network structure is depicted in Figure 2, where the encoding and decoding parts of the network are highlighted in purple and green, respectively.

The u-shaped network uses down and up-convolutional layers for the encoding and decoding part, respectively. A down-convolution layer is obtained by increasing the stride of the convolution, i.e. it only computes a subset of all positions. This decreases the spatial resolution of the following layer, and simultaneously increases the spatial support of all subsequent layers. To increase the image resolution again we use so-called up-convolutional lay-

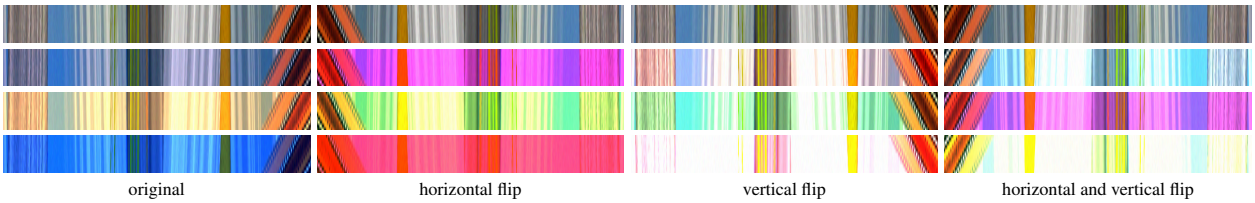


Figure 3: Illustration of the used data augmentation. The figure shows the original sample to the top left. The different columns represent horizontally and vertically flipped samples, and the different rows illustrate random brightness and color changes.

ers. Those layers use fractional strides to increase the resolution. Note that in the proposed u-shaped network we use the down and up-convolutional layers to decrease and increase only the spatial direction of the given EPI.

The basic building block of the overall network is a convolutional layer followed by a Rectified Linear Unit (ReLU) non-linearity [21], $\sigma(\mathbf{x}) = \max(\mathbf{0}, \mathbf{x})$. We combine two convolutional layers to one level. For the convolutional layers within one level, we use padding to compensate for the kernel size. This ensures that the output of one level has the same size as the input. In the first part of the network we use three of those levels, where we use down-convolutional layers after each level to increase the spatial support of subsequent layers. At each down-sampling step we double the number of feature channels, except for the last level. The bottom part of the network consists of another level, that processes the compressed data. The decoding part of the network uses again three levels, but now we utilize up-convolutional layers before each level. Hence we up-convolve the whole coarse feature maps allowing to transfer high-level information to the fine prediction, and finally increase the image resolution back to the original size. All the involved convolutions use kernels of size 3×5, except for the down and up-convolutional layers that use 3×3 kernels. We also use so-called pinhole connections between the encoding and decoding part of the network, i.e. we concatenate the input of each level in the decoding part with the corresponding output feature map from the encoding path. We want to emphasis that the network structure involves only convolutional layers, i.e. we are not using any fully connected layers nor any pooling operations. A main advantage of avoiding fully connected layers is the ability to process EPIs of arbitrary resolutions.

Dataset. In order to train the proposed u-shape network a large amount of labeled training data is needed. Fortunately, we were allowed to use the synthetic dataset proposed in [11]. This dataset was generated using POV-Ray [23] and comes with highly accurate ground truth depth fields. Moreover the dataset also provides a random scene generator that allows to generate the desired amount of LFs. We render 200 LFs with a spatial resolution of 640×480 and a directional resolution of 11×11, out of which 150 are used to generate training data and 50 are used for testing.

Data Augmentation. Data augmentation [5, 16] is a common way to combat overfitting and to improve the generalization of the trained model. It basically allows the model to become invariant to certain predefined image deformations. We perform excessive data augmentation, including brightness changes, color changes, and additive Gaussian noise. We also flip the EPIs horizontally and vertically, where each flipping results in a sign change of the disparity map. Our augmentation procedure results in 8 times the original amount of image pairs. Although they are heavily correlated they allow to increase the robustness of

the trained model. Figure 3 provides some augmentation examples.

Network Training. While NNs learned with back-propagation have been around for several decades [25], only recently the computational power and data has been available to fully exploit this training technique [16]. In order to train the proposed u-shaped network we use the tensorflow framework [1], where we use Adam [15] as the optimization method to minimize the ℓ_1 loss. Out of the 150 rendered LFs used for training we extract 20e3 EPIs. The extracted samples are then increased eightfold using data augmentation. To monitor overfitting we use a test set of 10e3 samples. In deep networks with many convolutional layers a good initialization of the weights is extremely important. Ideally the weights in the network should be initialized such that each feature map has approximately unit variance. This can be achieved by drawing the initial weights of a given node from a Gaussian distribution with standard deviation $\sqrt{2/N}$, where N denotes the number of incoming nodes [8]. After initializing the weights as suggested in [8] we train our model for 400 epochs, where we use a mini-batch size of 2^8 samples.

5 Experiments

We have preformed an extensive analysis of our proposed model. We conducted synthetic and real world experiments. For the synthetic evaluation we used a recently presented LF dataset [11], where all LF scenes within the dataset have a directional resolution of 11×11, and a spatial resolution of 640×480. For the real world evaluation we used a LF captured with a Lytro camera as well as LFs from the Stanford Light Field Archive (SLFA). The used Lytro data provides a spatial resolution of 328×328 and a directional resolution of 7×7. LFs within the SLFA are captured using a multi-camera array [31] and contain 289 views on a 17×17 grid. We trained a u-shaped network based on the description in Section 4, where we use the same model for all the presented experiments. To obtain the final result we predict the horizontal and vertical EPIs and take the pointwise average of the two predictions.

We compare our model against the following state-of-the-art SfLF methods [10, 11, 14, 27, 28]. The method by Wanner and Goldluecke [28] analyzes the EPIs using the 2D structure tensor, before combining the obtained information using a variational framework. Tao et al. [27] proposed a fusion method that uses correspondence and defocus cues. Both local cues are combined to a global depth estimate by using a Markov Random Field (MRF) model. Heber and Pock [10] proposed a variational multi-view stereo model based on low rank minimization. This model includes a matching term based on Robust Principal Component Analysis (RPCA), that can be interpreted as an all vs. all matching term. Jeon et al. [14] proposed an algorithm for SfLF, that utilizes phase shift based subpixel displacements. Besides the use of the phase shift theorem the algorithm is quite straightforward. They first calculate various cost volumes, that are processed using edge-preserving filtering, before extracting a disparity map based on the winner-takes-all strategy. To correct the obtained disparity map in weak textured regions they proceed with a multi-label optimization using graph cuts. At the end they refine the discrete disparity map to a continuous one using an iterative refinement scheme. In [11] Heber and Pock presented the first attempt to predict depth information for given LF data by utilizing deep learning strategies. Their network was trained in a sliding window setup to predict for each imaged scene point the orientation of the corresponding 2D hyperplane in the domain of the LF. This corresponds to estimating the line orientations in the horizontal and vertical EPIs simultaneously. They also use a 4D regularization step to overcome prediction errors in textureless or uniform regions, where they

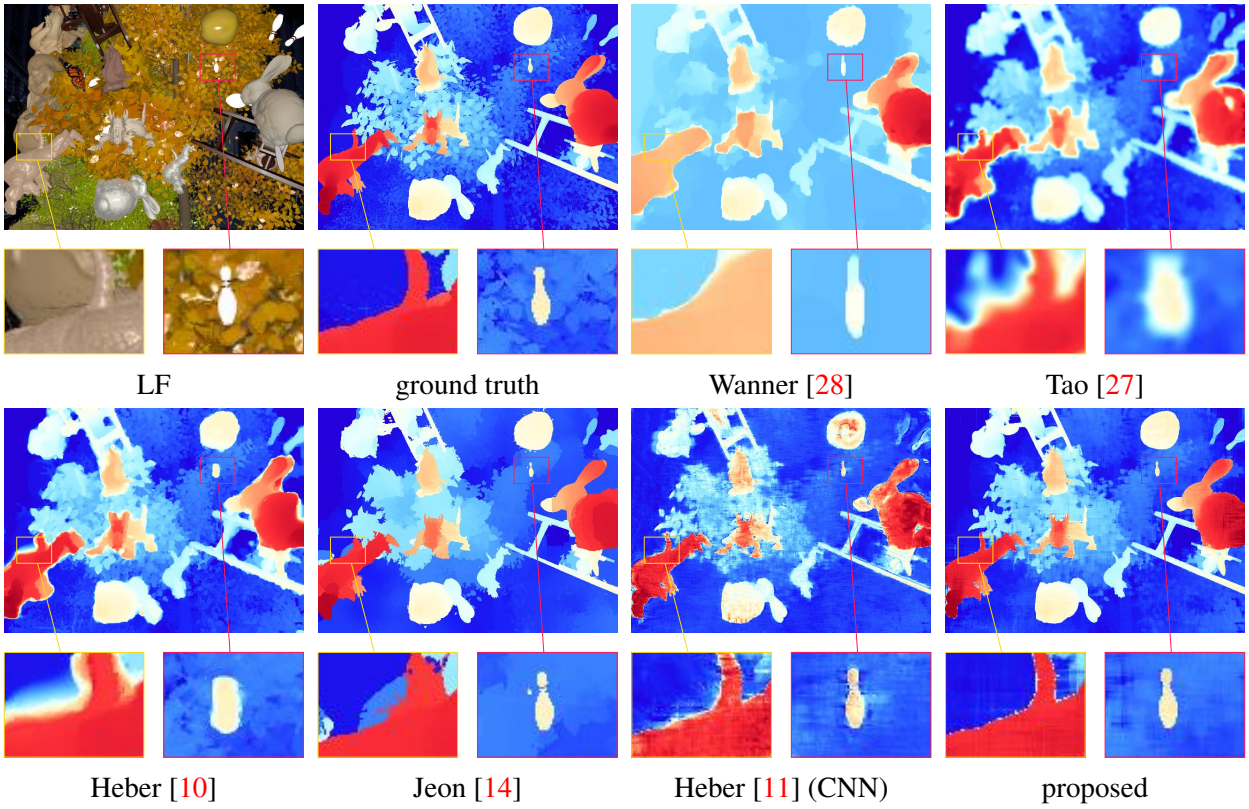


Figure 4: Comparison to state-of-the-art methods on the synthetic POV-Ray dataset. The figure shows the center view of the LF, the color-coded ground truth, the results for five state-of-the-art SfLF methods [10, 11, 14, 27, 28], followed by the result of the proposed method.

use a confidence measure to gauge the reliability of the estimate. This additional regularization step is not used in the following comparison, because such a post processing step can also be applied to the prediction of the proposed model. The method of Heber and Pock [11] works well but has drawbacks due to the sliding window scheme. First, the per-patch nature disallows to account for global output properties, and second, it leads to higher computational costs compared to the proposed approach. In what follows we will first provide some synthetic evaluations before presenting qualitative real world results.

Synthetic Evaluation. We start with the synthetic evaluation. Figure 4 provides a comparison of different state-of-the-art methods. Note that for all methods that rely on precomputed cost volumes [14, 27, 28], the number of labels is set to 200. Moreover we also set the necessary known disparity range for those methods based on the ground truth data. We can see that, despite the complexity of the scene, our model is able to predict accurate disparity results, that are on par with the competing methods. When comparing the results of the proposed model to the predictions obtained by the conventional CNN used in [11], we see that the proposed model provides better results in textureless regions. Also note that the proposed model is barely effected by depth discontinuities.

Quantitative results in terms of RMSE and MAE are presented in Table 1. The table also shows the percentage of pixels with a relative disparity error larger than 0.2% and 0.5%. Besides the various disparity errors the table also provides the computation times for estimating a disparity map for one sub-aperture view of the LF. Moreover we also indication if a GPU implementation was used or not. The presented results represent the average over the 50 LFs used for testing. We observe that the proposed model was able to accurately

	Wanner [28]	Tao [27]	Heber [10]	Jeon [14]	Heber [11] (CNN)	proposed
RMSE	3.91	2.33	2.50	2.49	1.87	0.80
MAE	2.94	1.06	0.79	0.75	1.13	0.35
0.5%	22.00	16.32	8.47	9.64	17.96	7.34
0.2%	35.22	28.48	13.20	16.46	31.61	14.76
Time	3min 18s	23min 4s	4min 44s	2h 12min 30s	35s	2s
GPU	✓	✗	✓	✗	✓	✓

Table 1: Quantitative results for various SfLF methods averaged over 50 synthetic LFs. The table provides the RMSE, MAE, the percentage of pixels with a relative disparity error larger than 0.2% and 0.5%, and the computational time of the method.

learn the characteristics of this dataset. Furthermore, we also see that the proposed method is significantly better than all the competing methods in terms of the computation time. The presented method takes about 15 seconds to compute the disparity field for the entire LF (i.e. 121 views).

Real World Evaluation. We continue with the real world evaluation. Figure 5 provides a qualitative comparison to the methods by Tao et al. [27], Heber and Pock [10], and Jeon et al. [14]. To be able to compute results for the methods by Jeon et al. [14] and Tao et al. [27] in a reasonable time, it was necessary to reduce the directional resolution of the data to 11×11 and the number of labels to 75. The results show that although the proposed model was not trained on this dataset, nor have we performed any fine-tuning for this dataset, it allows to predict depth maps that are on par with the competing methods. However, the results are not perfect because the model produces streaking artifacts in homogeneous background regions. The main benefit of the proposed method is again the computational time of a few seconds. Also keep in mind that we are not using any post-processing, the results shown in the figure are the raw network predictions.

In Figure 6 we also present results for a LF captured with a Lytro camera. Note, that the Lytro data includes a significant amount of noise and outliers, for which the proposed u-shape network was not trained for. Nevertheless, the proposed model is able to predict a reasonable disparity field with clear depth discontinuities.

6 Conclusion

We have presented a novel end-to-end system for SfLF. Our model is based on stacked convolution operations, that result in a high efficiency. The model comprises an encoding and a decoding part. Those parts are symmetric resulting in a u-shaped network architecture. We avoided fully connected layers thus our model allows to process LFs of any resolution. Our results show that the proposed u-shaped network is able to predict disparity fields that are on par with the state of the art while maintaining a low computation time. We believe our proposed approach is an important step towards realtime LF image processing. We also want to emphasize that the results shown in the experimental section are the raw network predictions without any additional post-processing. Investigating suitable methods for post-processing the network output is left for future work.

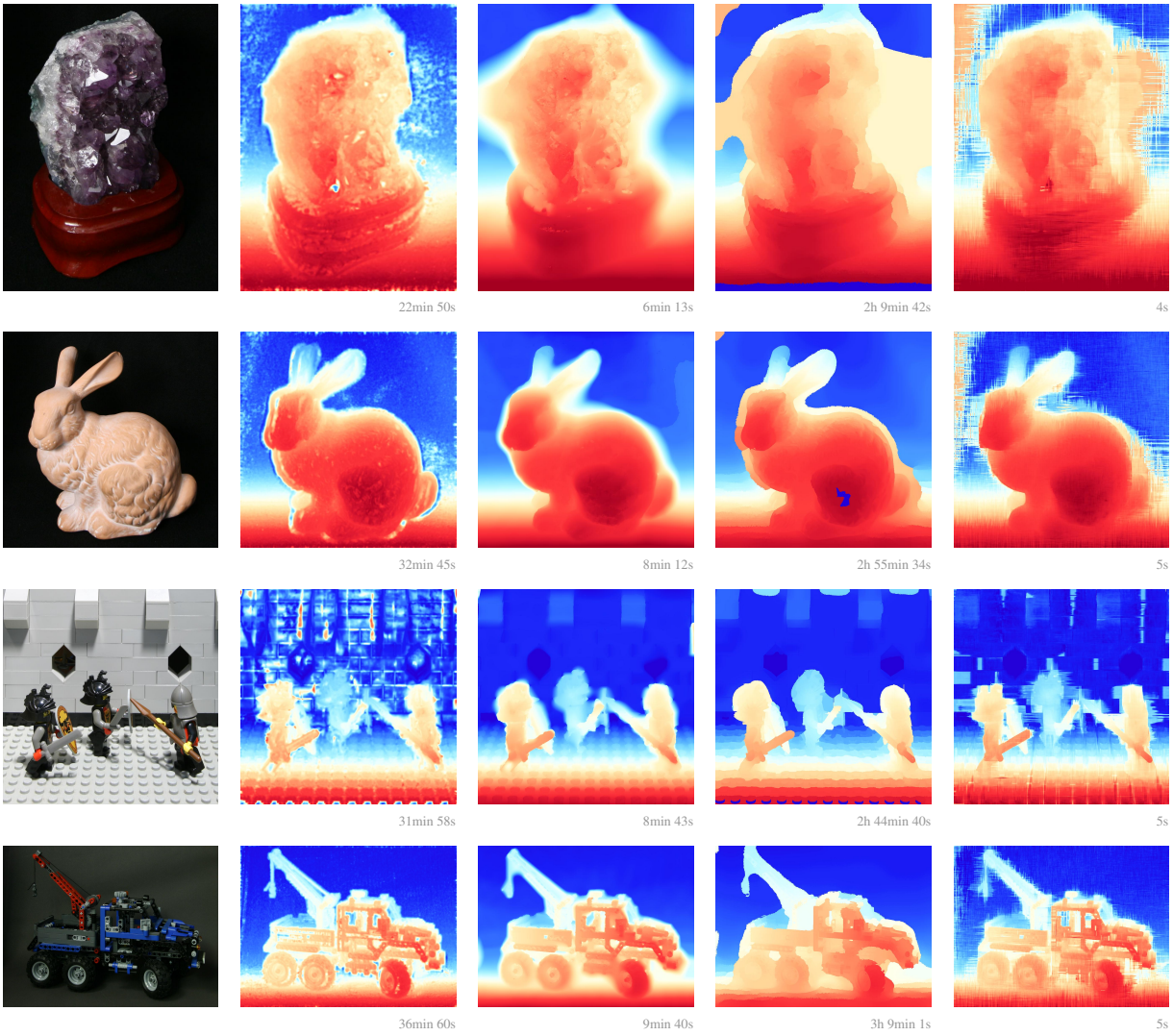


Figure 5: Qualitative comparison for LFs from the SLFA. The figure shows from left to right the center view of the LF, followed by the results for the methods proposed by Tao et al. [27], Heber and Pock [10], and Jeon et al. [14]. The results to the right correspond to the proposed method.

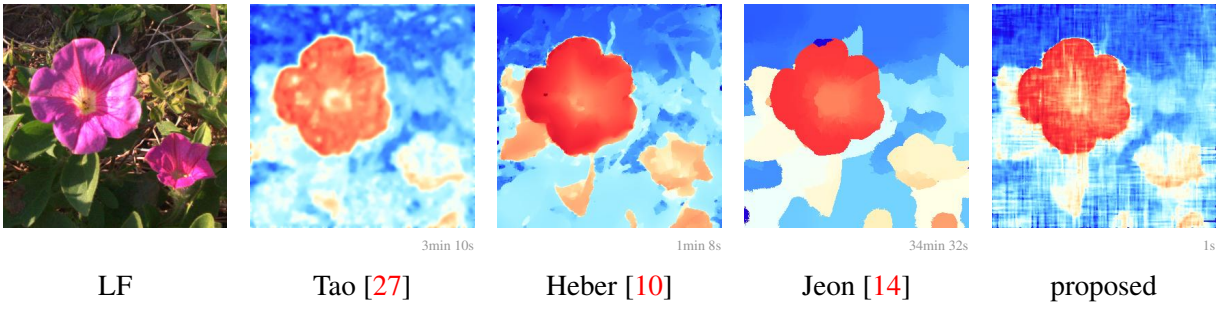


Figure 6: Qualitative comparison for a LF captured with a plenoptic camera. The figure shows from left to right the center view of the LF, followed by the results for the methods proposed by Tao et al. [27], Heber and Pock [10], and Jeon et al. [14]. The result to the right corresponds to the proposed method.

Acknowledgment. This work was supported by the FWF-START project *Bilevel optimization for Computer Vision*, No. Y729 and the Vision+ project *Integrating visual information with independent knowledge*, No. 836630.

References

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL http://tensorflow.org/. Software available from tensorflow.org.

[2] Tom E. Bishop and Paolo Favaro. The light field camera: Extended depth of field, aliasing, and superresolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(5):972–986, 2012. ISSN 0162-8828.

[3] Can Chen, Haiting Lin, Zhan Yu, Sing Bing Kang, and Jingyi Yu. Light field stereo matching using bilateral statistics of surface cameras. June 2014.

[4] Alexey Dosovitskiy, Jost Tobias Springenberg, and Thomas Brox. Learning to generate chairs with convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 1538–1546, 2015.

[5] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In Z. Ghahramani, M. Welling, C. Cortes, N.d. Lawrence, and K.q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2366–2374. Curran Associates, Inc., 2014.

[6] B. Goldluecke and S. Wanner. The variational structure of disparity and regularization of 4d light fields. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[7] Steven J. Gortler, Radek Grzeszczuk, Richard Szeliski, and Michael F. Cohen. The lumigraph. In *SIGGRAPH*, pages 43–54, 1996.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *CoRR*, abs/1502.01852, 2015.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015.

[10] Stefan Heber and Thomas Pock. Shape from light field meets robust PCA. In *Proceedings of the 13th European Conference on Computer Vision*, 2014.

[11] Stefan Heber and Thomas Pock. Convolutional networks for shape from light field. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[12] Stefan Heber, Rene Ranftl, and Thomas Pock. Variational Shape from Light Field. In *International Conference on Energy Minimization Methods in Computer Vision and Pattern Recognition*, 2013.

[13] Aaron Isaksen, Leonard McMillan, and Steven J. Gortler. Dynamically reparameterized light fields. In *SIGGRAPH*, pages 297–306, 2000.

[14] H. G. Jeon, J. Park, G. Choe, J. Park, Y. Bok, Y. W. Tai, and I. S. Kweon. Accurate depth map estimation from a lenslet light field camera. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1547–1555, June 2015.

[15] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

[16] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[17] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Comput.*, 1(4):541–551, December 1989. ISSN 0899-7667.

[18] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, Nov 1998. ISSN 0018-9219.

[19] Marc Levoy and Pat Hanrahan. Light field rendering. In *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '96, pages 31–42, New York, NY, USA, 1996. ACM. ISBN 0-89791-746-4. doi: 10.1145/237170.237199.

[20] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440, June 2015. doi: 10.1109/CVPR.2015.7298965.

[21] Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines. In Johannes Fuernkranz and Thorsten Joachims, editors, *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 807–814. Omnipress, 2010.

[22] Ren Ng. *Digital Light Field Photography*. Phd thesis, Stanford University, 2006.

[23] POV-Ray. Pov-ray. http://www.povray.org.

[24] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, pages 91–99, 2015.

[25] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Neurocomputing: Foundations of research. chapter Learning Representations by Back-propagating Errors, pages 696–699. MIT Press, Cambridge, MA, USA, 1988. ISBN 0-262-01097-6.

[26] J.T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. Striving for simplicity: The all convolutional net. In *ICLR (workshop track)*, 2015.

[27] Michael W. Tao, Sunil Hadap, Jitendra Malik, and Ravi Ramamoorthi. Depth from combining defocus and correspondence using light-field cameras. In *International Conference on Computer Vision (ICCV)*, December 2013.

[28] S. Wanner and B. Goldluecke. Globally consistent depth labeling of 4D lightfields. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[29] S. Wanner and B. Goldluecke. Spatial and angular variational super-resolution of 4d light fields. In *European Conference on Computer Vision (ECCV)*, 2012.

[30] S. Wanner, C. Straehle, and B. Goldluecke. Globally consistent multi-label assignment on the ray space of 4d light fields. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[31] Bennett Wilburn, Neel Joshi, Vaibhav Vaish, Eino-Ville Talvala, Emilio Antunez, Adam Barth, Andrew Adams, Mark Horowitz, and Marc Levoy. High performance imaging using large camera arrays. *ACM Trans. Graph.*, 24(3):765–776, July 2005. ISSN 0730-0301. doi: 10.1145/1073204.1073259.

[32] Matthew D. Zeiler and Rob Fergus. *Computer Vision – ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I*, chapter Visualizing and Understanding Convolutional Networks, pages 818–833. Springer International Publishing, Cham, 2014. ISBN 978-3-319-10590-1.