Neural EPI-volume Networks for Shape from Light Field

Stefan Heber[1] Wei Yu[1] Thomas Pock[1,2]
Graz University of Technology[1]
Austrian Institute of Technology[2]
{stefan.heber,wei.yu,pock}@icg.tugraz.at

Abstract

This paper presents a novel deep regression network to extract geometric information from Light Field (LF) data. Our network builds upon u-shaped network architectures. Those networks involve two symmetric parts, an encoding and a decoding part. In the first part the network encodes relevant information from the given input into a set of high-level feature maps. In the second part the generated feature maps are then decoded to the desired output. To predict reliable and robust depth information the proposed network examines 3D subsets of the 4D LF called Epipolar Plane Image (EPI) volumes. An important aspect of our network is the use of 3D convolutional layers, that allow to propagate information from two spatial dimensions and one directional dimension of the LF. Compared to previous work this allows for an additional spatial regularization, which reduces depth artifacts and simultaneously maintains clear depth discontinuities. Experimental results show that our approach allows to create high-quality reconstruction results, which outperform current state-of-the-art Shape from Light Field (SfLF) techniques. The main advantage of the proposed approach is the ability to provide those high-quality reconstructions at a low computation time.

1. Introduction

This paper investigates the task of reconstructing the geometry of a scene based on captured Light Field (LF) data. The corresponding problem is referred to as Shape from Light Field (SfLF). A LF represents a densely sampled set of images captured from a regular grid of viewpoints located on a common 2D plane. It should be emphasized that the key difference to the general multi-view stereo setting is the dense and regular sampling of the viewpoints. Thus compared to a traditional 2D image a 4D LF provides two additional dimensions, that can be interpreted as a parametrization of the 2D grid of viewpoints. This multi-view stereo interpretation of the LF data directly shows that a LF provides information about the geometry of the observed scene. The encoded geometrical information in the LF allows to tackle problems that are impossible to solve based on a single 2D image of the scene. Those problems include the geometrical reconstruction of the observed scene itself [32, 9, 13, 30, 11, 17], the generation of images with different focus or aperture settings [16, 25], and the digital viewpoint manipulation [25], to name but a few.

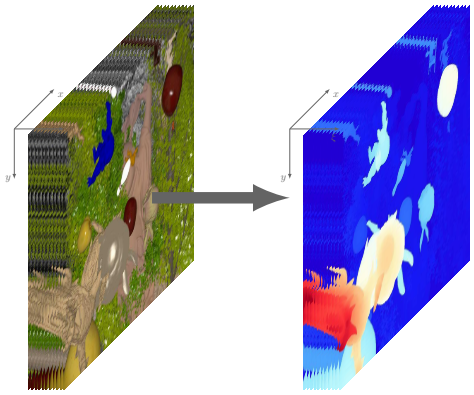

Figure 1. Illustration of the input and output of the proposed network. To the left the figure shows a RGB EPI volume, that is fed to our network, and to the right the corresponding color-coded disparity information is shown.

SfLF is currently a very active area of research. The LF research field has grown from a niche topic to an established part of today's Computer Vision (CV) research. This development occurred not least because there is a growing commercial interest in LF technology. Nowadays LF or plenoptic cameras are used in industrial applications, like for instance automated optical inspection [28], and LF technology is used in consumer cameras to provide features like digital refocusing capabilities [23]. Moreover, there is also an increasing number of companies that are looking for new ways to capture cinematic Virtual Reality (VR) content, where LF imaging might be a perfect solution that allows to take advantage of the freedom of motion offered by devices like Oculus Rift [26] or HTC Vive [15].

In mathematical terms a 4D LF is commonly defined via the so-called two-plane parametrization, where a ray is defined by the intersection points of two parallel planes. Let $\Omega \subseteq \mathbb{R}^2$ and $\Pi \subseteq \mathbb{R}^2$ be two parallel planes ($\Omega \neq \Pi$), then the LF is defined 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 spatial and directional coordinates. Note that Ω corresponds to the traditional image plane and Π is usually referred to as lens or focal plane.

A LF can be visualized in many different ways. Common visualizations are sub-aperture images and EPIs. Both represent 2D slices through the general 4D LF. The representation we work with in this paper is called EPI volume [3], which is equivalent to an orthogonal 3D slice through the LF. In terms of Equation (1) an EPI volume is obtained by holding one directional coordinate constant and varying the remaining coordinates. For instance by choosing a certain directional coordinate η we restrict the 4D LF to the 3D function

$$\Sigma_\eta : \mathbb{R}^3 \to \mathbb{R}, \qquad (x, y, \xi) \mapsto L(x, y, \xi, \eta), \tag{2}$$

that defines the corresponding horizontal EPI volume. Similarly one can also define vertical EPI volumes. Compare Figure 1 for a visualization of an EPI volume. EPI volumes nicely illustrate the linear characteristic of the LF space. By analyzing the orientations of the individual lines in this representation one can infer the depth of the corresponding scene points. This correspondence between depth and orientation has been leveraged in many works on SfLF.

After the huge success of deep learning in a variety of CV applications it was only a matter of time till deep learning principles have found their application also in LF image processing. In [12, 14] it was shown that utilizing learning-based approaches for SfLF has a high potential to reduce depth artifacts due to occlusions, partial visibility, and specular reflections. The advantages of data-driven approaches are not only the capability to learn from data how to handle certain artifacts, but also the facilitation for an efficient implementation of the inference step, which results in a fast computation time.

This work focuses on one main drawback of the deep learning architecture proposed in [14]. We address the lack of regularization in the disparity prediction provided by this method. Due to the fact that the network in [14] is designed to process a single EPI, the input information fed to the network is limited to one spatial dimension. This results in streaking artifacts in the respective direction, which can not be reliably resolved in their proposed framework. To remedy this, we propose to extend the network architecture in [14] to predict disparity based on entire EPI volumes, *i.e.* we allow the network to incorporate information from both spatial dimension. We will show that this simple modification allows to spatially propagate information and avoids the unwanted streaking artifacts.

2. Related Work

SfLF is a fundamental problem in LF image processing. However, despite a substantial amount of progress in this field, 3D scene reconstruction from LFs still struggles with many difficulties, especially in dealing with occlusions, textureless regions, and specular reflections.

There is a wide range of methods for SfLF. The field can be roughly divided into methods based on EPI analysis like *e.g.* [32, 9], and multi-view stereo matching based approaches like *e.g.* [13, 4]. The seminal work of Bolles *et al.* [3] introduced so-called EPI volumes, where they analyze the slopes of lines by line fitting in order to estimate sparse disparity information. In [5] Criminisi *et al.* exploited the high degree of regularity found in the EPIs. They performed a so-called EPI strip rectification, *i.e.* a shearing of the EPI, to estimate lines with smallest color variation and hence dense disparity information. The first order structure tensor was used in [32, 9] to compute the orientation of lines in the EPIs. In [13] the authors proposed a matching-term based on Active Wavefront Sampling (AWS), that is used within a variational multi-view stereo framework. Tao *et al.* [30] suggested to combine correspondence cues with defocus cues to calculate depth. In order to indicate the probability of occlusions Chen *et al.* [4] introduced a bilateral consistency metric on the surface camera. In [11] Heber and Pock defined a new dataterm based on Robust Principal Component Analysis (RPCA), that exploits the redundancy of sub-aperture views. Jeon *et al.* [17] employed the phase-shift theorem to match sub-aperture images.

While for classical stereo reconstruction for color images deep learning approaches are gaining ground [34], SfLF methods still mainly rely on classical variational principles, EPI filtering, and other handcrafted solutions. The main reason for this is the lack of high quality large-scale training data, which is essential to train deep network architectures. Hence only few work has been pursued on utilizing Machine Learning (ML) techniques for LF analysis [31, 18]. To the best of our knowledge there exist only two relevant publications that apply ML techniques to SfLF. Heber and Pock [12] proposed to apply a conventional Convolutional Neural Network (CNN) in a sliding window fashion to predict slope orientation in the EPIs. The main motivation of utilizing ML for SfLF is the fact that phenomena such as occlusions, specular highlights or reflections manifest as certain patterns on the EPI space and can be learned in a data-driven approach. Hence learning-based approaches basically allow to handle cases that are problematic for tra-

ditional non-learning based methods. However, because of the redundancy in overlapping patches, the patch-based method in [12] comes with the drawback of high computational costs. Furthermore, it also relies on an additional refinement step to handle textureless or uniform regions. In a follow up work [14] the authors addressed those drawbacks and proposed a more sophisticated network structure that operates on entire 2D EPIs. The network achieved good reconstruction results in combination with low computational costs. However, one downside of the method is the introduction of streaking artifacts into the final reconstruction result. In this work we tackle this problem by providing the network with additional information from neighboring pixels, *i.e*. extending the input dimension of the network by incorporating the second spatial dimension. Hence we propose to train a network that allows to predict disparity information based on entire EPI volumes. In this work, for the first time, we unify ideas from 2D EPI analysis with spatial matching based approaches by learning 3D filters for disparity estimation based on EPI volumes.

3. Contribution

In this paper we make the following main contributions: We propose a method for SfLF that builds upon the u-shaped architecture proposed in [14]. In particular, we extend the work of [14] by introducing additional spatial regularization in terms of 3D convolutions. By doing so we are able to eliminate the main drawback of the network proposed in [14], which is the tendency to generate visually unpleasing streaking artifacts. Hence, we transfer recent success in predicting disparity information based on EPIs to entire EPI volumes. More specifically, this means that the proposed network sequentially processes 3D subvolumes of the 4D LF instead of 2D EPIs. Compared to [14] this modification allows to propagate information from both spatial dimensions and thus allows to avoid unwanted depth artifacts in the final 4D disparity field. In a fair evaluation we will show that our learning-based method is able to outperform the current state of the art. It not just allows to reduce depth artifacts in the final reconstruction, but it also allows to maintain a low computation time.

4. Methodology

This section describes the methodology of the proposed deep learning approach for disparity prediction based on EPI volumes [3]. The proposed approach consists of three main parts: (i) extending the u-shaped network structure in [14] to perform additional spatial regularization, (ii) preparing a dataset for supervised training, and (iii) training the network using the tensorflow framework [1].

Our data-driven approach is based on CNNs [21], that have been successfully applied to many CV applications. The popularity of CNNs in CV increased drastically after Krizhevsky *et al*. [20] efficiently applied them for large scale image classification. Modern CNN architectures basically alternate between convolutions and Rectified Linear Units (ReLUs) [24]. Note that convolutions only account for short-range dependencies, *i.e*. that they are limited by the size of their kernels. There are different ways to allow long-range dependencies without introducing unfeasible fully connected layers. The simplest solution is to make the network deeper and with that increase the receptive field to the desired size. Another way is to introduce pooling or downsampling layer, but this reduces the output resolution and hence is only one part of the solution. The second part involves so-called unpooling or upsampling layer, that allow to increase the resolution again to a desired output resolution, see *e.g*. [35, 7, 22]. Networks that involve a bottleneck due to downsampling and upsampling operations were first introduced to model auto-encoder. Those networks are basically designed to learn a sparse representation of the given data. Hence they compress the data which also results in a loss of detail when utilized for other prediction tasks. By introducing connections that skip the downsampling parts of the network it is possible to preserve the high frequency information and simultaneously allow for long-range dependencies, as demonstrated in *e.g*. [29, 6, 14]. Those skip-connections are denoted as pinhole connections in the remainder of this paper.

The proposed network is an extension of the network proposed in [14] and is mainly designed to remove depth artifacts occurring in the final 4D disparity field. Due to the fact that the network in [14] only receives input from one spatial dimension, the result is not necessarily consistent w.r.t. the second spatial dimension. This results in depth-artifact in ambiguous regions, which appear as streaking artifact in the sub-aperture images of the final reconstruction. In this work we remedy this problem. We extend the network input by the second spatial dimension. In this way we can exploit 3D convolutional layers to perform an additional spatial regularization.

In what follows we will present details about the suggested network architectures and the training procedure. Due to the fact that our method is not based on natural 2D images, we are not able to exploit existing trained networks in terms of transfer learning. The proposed network is entirely trained from scratch.

Network Architecture. The proposed network is a special version of a Fully Convolutional Network (FCN) [22], that has a network structure similar to an auto-encoder [2] with additional pinhole connections. The overall network architecture is inspired by [14] and it is designed to predict disparity based on RGB EPI volumes. Note that an EPI volume defines a 3D subset of a 4D LF, *i.e*. several net-

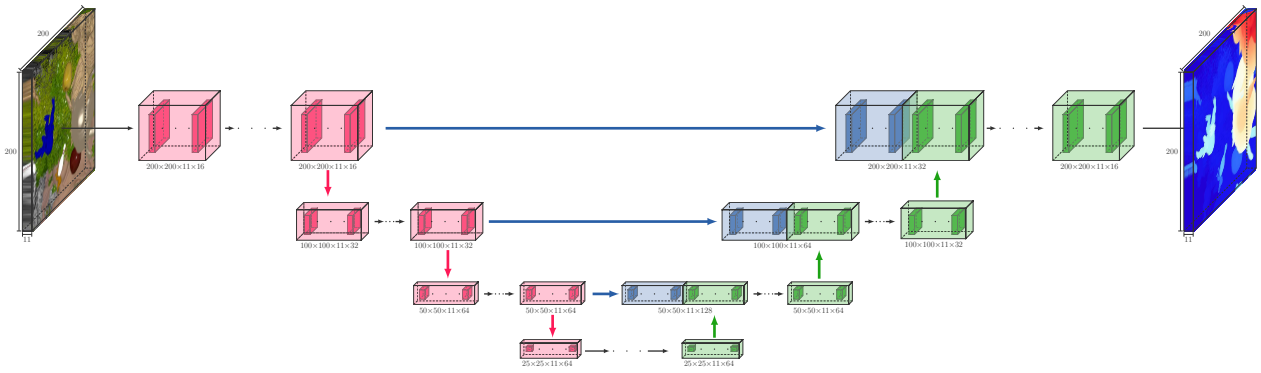

Figure 2. Illustration of the proposed network architecture. The overall network structure builds upon u-shaped networks. Those networks involve two symmetric parts, an encoding and a decoding part. The encoding and decoding parts of the network are highlighted in purple and green, respectively. To preserve high-frequency information the network also uses so-called pinhole connections, marked in blue, that allow to skip the downsampling parts of the network.

work predictions have to be combined to obtain the final 4D disparity field. Figure 2 provides an overview of the entire network structure. Contrary to [14], where the network only operates on a single EPI, the input of the proposed network is an RGB EPI volume and the output is its corresponding disparity volume. Thus the proposed network extends the network in [14] by replacing all 2D operators with their 3D counterparts. The proposed network consists of essentially two symmetric parts, an encoding part (*c.f.* purple part in Figure 2) and a decoding part (*c.f.* green part in Figure 2). Each part is further subdivided into different levels. In the encoding and decoding part those levels are connected via down and up-convolutional layers, respectively. At each down-convolutional layer we reduce the spatial resolution by a factor of two, and at each up-convolutional layer we consequently increase the resolution again by the same factor. In Figure 2 the down and up-convolutional layers are indicated with purple and green arrows, respectively. Each level in the network consists of four convolutional layer each followed by a ReLU non-linearity [24], $\sigma(\mathbf{x}) = \max(\mathbf{0}, \mathbf{x})$. Note that each layer is a four dimensional array of size $h \times w \times d \times c$, where the first three dimensions correspond to the two spatial and one directional dimensions of the LF, and c is the feature or channel dimension. The involved convolutional layers perform 3D convolutions with filter kernels of size $3 \times 3 \times 3$, that correspond to the x, y, and ξ dimension of the LF. Each convolutional layer employs $\max\{64, 16 \cdot 2^{l-1}\}$ filter, where $l \in [4]$[1] denotes the respective level, *i.e.* we start with 16 feature channels in the first level of the encoding part and gradually increase them towards higher levels, till we reach a maximum number of 64 feature channels. Likewise in the decoding part we gradually decrease the number of feature channels at each up-convolutional layer except at the lowest one. At the very end of the network we use a convolutional layer to map to one output channel representing the disparity information. Note that the network does not include any pooling layer, we make use of learned down and up-convolutional layers instead. An important aspect of this network is the use of so-called pinhole connections, that connect the levels from the encoding part with the respective levels in the decoding part of the network. At each pinhole connection (*c.f.* blue arrows in Figure 2) we concatenate the output feature map of the encoding level with the input feature map of the corresponding decoding level. Those pinhole connections allow to preserve high frequency information and thus increase the amount of details in the final reconstruction. Due to the fact that this network belongs to the class of FCNs it allows to process EPI volumes of arbitrary resolutions. The main reason for this is the fact that the involved convolutions are inherently translational invariant. Also note that the network is trained end-to-end and does not make use of pre- and post-processing complications.

[1]Notation: $[n] := \{1, \ldots, n\}$

Dataset. In order to train the proposed network a large amount of training data is needed for supervised training. For this purpose we leverage the synthetic LF dataset proposed in [12]. This dataset provides several interesting features that distinguishes it from other datasets, including highly accurate ground truth depth fields, and a random scene generator, which makes it easy to scale the dataset as required. For our current purpose we generated 900 LFs with a spatial resolution of 640×480 and a directional resolution of 11×11. The entire dataset is split up into a training set of 850 LFs and a test set of 50 LFs.

Data Augmentation. One way to combat overfitting on the training data is called data augmentation [20, 8]. The main idea is to train the model such that it gets invariant to

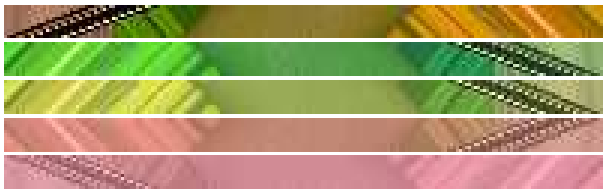

Figure 3. Illustration of the used data augmentation. The figure shows slices through the EPI volume, where the original sample is shown at the top followed by different augmented versions.

certain predefined image deformations. This is done by extending the training set with slightly modified training samples. Although the samples generated via data augmentation are heavily correlated, they allow to increase the robustness of the trained model. We perform a large amount of data augmentation, including hue, saturation, contrast and brightness modifications. To modify the hue and saturation we first convert the RGB images to the HSV color space and add offsets to the hue and saturation channels. The offsets are randomly picked from the interval $[-0.25, 0.25]$ for the hue channel and from $[0.3, 1.0]$ for the saturation channel. After that we convert back to the RGB color space. To augment the contrast, for each channel $\mathbf{x}$ we compute the mean $\bar{\mathbf{x}}$ of the pixel values and calculate the contrast manipulated result as $\bar{\mathbf{x}} + (\mathbf{x} - \bar{\mathbf{x}})s$, where s denotes a contrast factor randomly picked from the interval $[0.1, 1.0]$. The brightness is augmented by adding an offset to each channel, that is randomly picked from $[-0.1, 0.1]$. Besides the changes in pixel values we also flip the x and ξ coordinate axes randomly with a probability of 0.5. Note, when flipping one of the axes we also need to simultaneously flip and negate the disparity values of the corresponding labels. Finally we also add additive Gaussian noise with zeros mean and a standard deviation of 1% of the image dynamic range. Figure 3 provides an illustration of the implemented data augmentation, where slices through augmented EPI volumes are shown.

Network Training. In order to train the proposed network we use the tensorflow framework [1], where we chose Adam [19] to optimize the ℓ_1 loss. Compared to using an ℓ_2 loss this allows to reduce the effect of blurry predictions.

For the training procedure we implemented an input pipeline, that performs the following steps. First we load a random LF from the training set into memory and extract an RGB EPI volume of size $200 \times 200 \times 11 \times 3$ at a random position. This EPI volume constitutes a single training sample, which is added to an input queue, that holds a certain amount of samples. When removing samples from this queue new samples are automatically reloaded. During training we take the first n samples from the queue and triple them using data augmentation. We calculate the gradient of the resulting mini-patch using back-propagation and update the network parameters based on the selected optimization scheme. We use 10 LFs from the test set to monitor overfitting.

Initialization is another important part when training a network. As suggested in [10] we initialize the weights of the network by drawing them from a Gaussian distribution with standard deviation $\sqrt{2/N}$, where N denotes the number of incoming nodes. We train the model for approximately 1000 epochs, where we use a mini-batch size of 48.

5. Experiments

We thoroughly evaluate the proposed model on the following datasets. For the synthetic evaluation we use the LF dataset proposed in [12]. This dataset provides LFs, with a spatial resolution of 640×480 and a directional resolution of 11×11, that are generated using the ray tracing software POV-Ray [27]. The rendered LFs in this dataset are quite challenging because of non-Lambertian surfaces and a great number of objects that are occluding each other. For real world evaluation we use the Stanford Light Field Archive (SLFA) [33]. The LFs from the SLFA are captured with a multi-camera array. They have varying spatial resolutions and a fixed directional resolution of 17×17.

We compare against top performing algorithms for SfLF [32, 30, 11, 17, 12, 14]. Wanner and Goldluecke [32] proposed a method based on EPI analyzes that uses the 2D structure tensor to estimate line orientations. Tao *et al*. [30] suggested to combine correspondence and defocus cues. Heber and Pock [11] proposed a sparse coding method based on RPCA, that shears the 4D LF. Jeon *et al*. [17] utilizes the phase shift theorem to calculate subpixel displacements. In [12] Heber and Pock presented a patch-based deep learning approach to predict depth information for given LF data. Their network takes as input two patches extracted from the vertical and horizontal EPIs and predicts the orientation of the corresponding 2D hyperplane in the domain of the LF. After the pointwise prediction they used an additional 4D higher order regularization step to cope with untextured or uniform regions. They also used a 4D anisotropic diffusion tensor to guide the regularization and a confidence measure to gauge the reliability of the CNN prediction. In a follow up work [14] the approach is extended to predict entire 2D EPIs at once using u-shaped networks. This allows to get rid of the regularization step needed in [12] and thus drastically reduces the overall computation time. However, due to the fact that they process each EPI separately, their approach introduces streaking artifact in the sub-aperture images of the final reconstruction. To overcome this drawback the proposed network predicts disparity information based on entire EPI volumes. This allows to propagate information in both spatial dimensions and thus introduces some kind of spatial regularization. Compared to [14] we will show that this modifi-

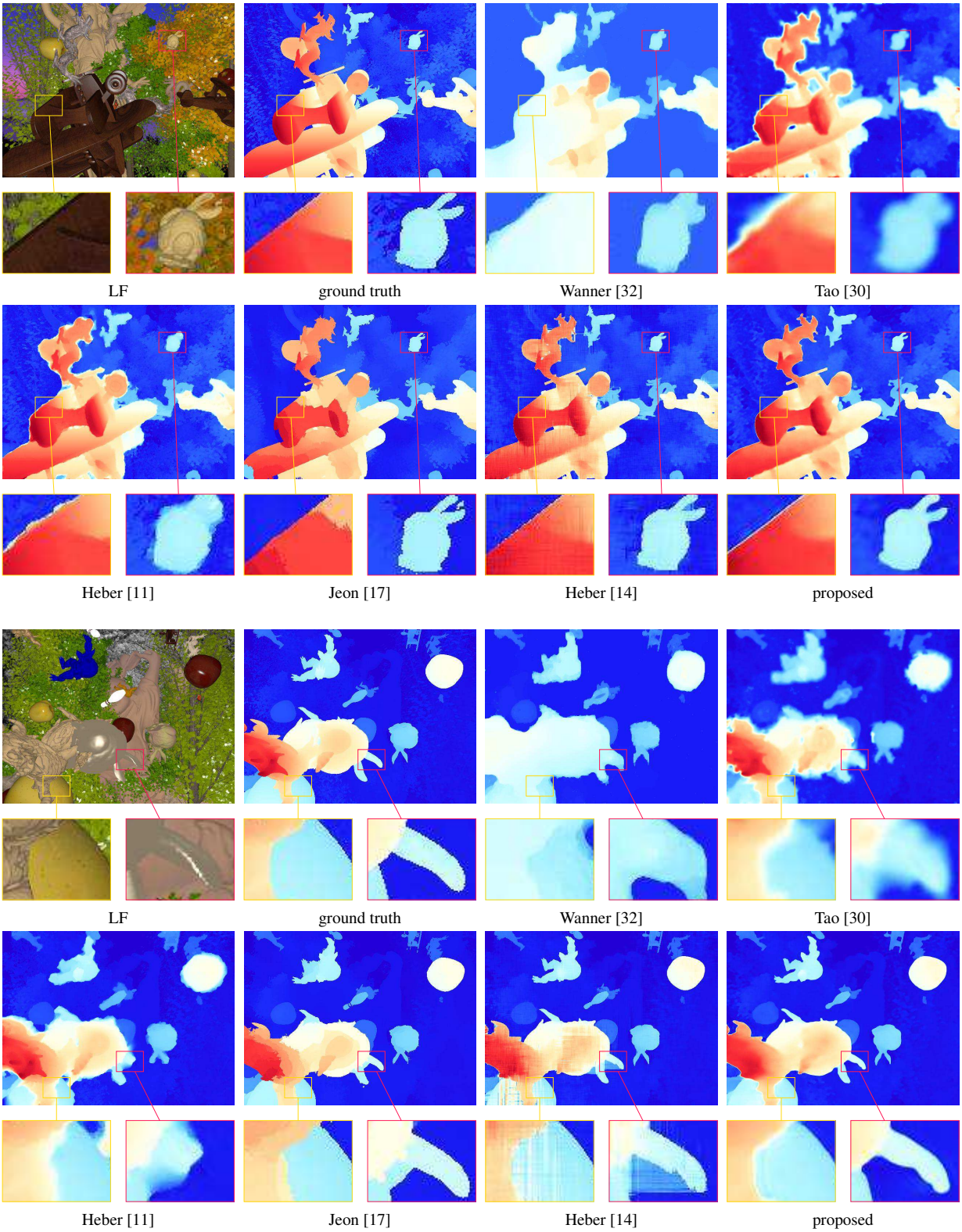

Figure 4. Comparison to state-of-the-art methods on the synthetic POV-Ray dataset. The figure shows the center view of the LF, the color-coded ground truth, the results for five state-of-the-art SfLF methods [32, 30, 11, 17, 14], followed by the result of the proposed method.

	Wanner [32]	Tao [30]	Heber [11]	Jeon [17]	Heber [12] (CNN)	Heber [14]	proposed
RMSE	3.91	2.33	2.50	2.49	1.87	0.80	0.83
MAE	2.94	1.06	0.79	0.75	1.13	0.35	0.34
0.5%	22.00	16.32	8.47	9.64	17.96	7.34	7.28
0.2%	35.22	28.48	13.20	16.46	31.61	14.76	14.42
Time	3min 18s	23min 4s	4min 44s	2h 12min 30s	35s	1.3s	0.8s
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. In each row we indicate with green and yellow the best and second best result, respectively.

cation allows to remove depth artifacts, and simultaneously preserves the advantages of sharp depth discontinuities and a fast computation time.

In this section we mainly focus on a quantitative evaluation based on synthetic data. Besides that we also present some qualitative real world results. Note that for all our experiments we use a horizontal slicing strategy as indicated in Equation (2). Overall our evaluations show that the proposed model is superior to the state of the art.

Synthetic Evaluation. For the synthetic evaluation we use a test set of 50 LFs. The majority of disparity values in this dataset are in the range $[-5, 5]$, with a few exception that exceed this range. Figure 4 provides a visualization of the disparity results of the proposed method and compares them to different state-of-the-art methods. For methods relying on precomputed cost volumes, *i.e*. [32, 30, 17], we set the number of labels to 200 in this experiment. Moreover for those methods the needed disparity range is set based on the ground truth data. When considering Figure 4 we see that the proposed network is able to predict accurate disparity results. Overall the results of the proposed method are on par with those obtained by the method in [14]. However, when considering the closeup views we recognize that the proposed model is able to effectively remove unwanted streaking artifact, that are prevalent in the results predicted by the network proposed in [14]. Also note that the proposed method is barely effected by depth discontinuities.

Table 1 provides quantitative results, that are averaged over the 50 LFs used for testing. Note that for the method proposed in [12] we only compare to the network prediction and exclude the additional refinement step. The table shows the RMSE, the MAE, and the percentage of pixels with a relative disparity error larger than 0.2% and 0.5%. Moreover the table also provides the average computation time for the various methods and an indication if a GPU implementation was used or not. We see that the proposed method provides an excellent performance, and achieves the overall best results. Overall it can improve upon the 2D method proposed in [14]. It provides a low error rate in combination with low computation times. Furthermore we also observe that the proposed method is significantly better compared to non-learning based methods. Especially in terms of computation time we observe a tremendous improvement.

Real World Evaluation. Figure 5 provides a qualitative comparison to state-of-the-art methods based on the SLFA. Note that we had to reduce the directional resolution of the dataset from 17×17 to 11×11 to be able to compute results for the methods by Jeon *et al*. [17] and Tao *et al*. [30] in a reasonable timeframe. Moreover the number of labels for those methods is set to 75. The results show that the proposed method allows to predict excellent disparity fields, although the model was not trained on this specific dataset. Compared to [14] we see that the proposed model removes the streaking artifacts. Especially in regions of depth discontinuities the proposed network is able to reconstruct the scene more accurately than the competing methods. We also indicate the computation time for each method at the bottom right, which shows the main benefit of the learning based approaches. Compared to the fastest non-learning based method, the proposed method is able to reconstruct the scene 100 times faster. Also keep in mind that the proposed method at the same time calculates disparity information for an entire EPI volume and not just for the center view alone.

6. Conclusion

In this paper we have described a new approach for recovering depth information from LF data. We presented an end-to-end system for SfLF that analyzes EPI volumes. Due to stacked convolutional operations the proposed network architecture provides a high efficiency. The suggested network structure extends the u-shaped architecture proposed in [14] to perform additional spatial regularization. By doing so we were able to eliminate depth artifacts present in the results produced with the method in [14].

Our experimental results show that the proposed model is able to predict disparity fields that are significantly better

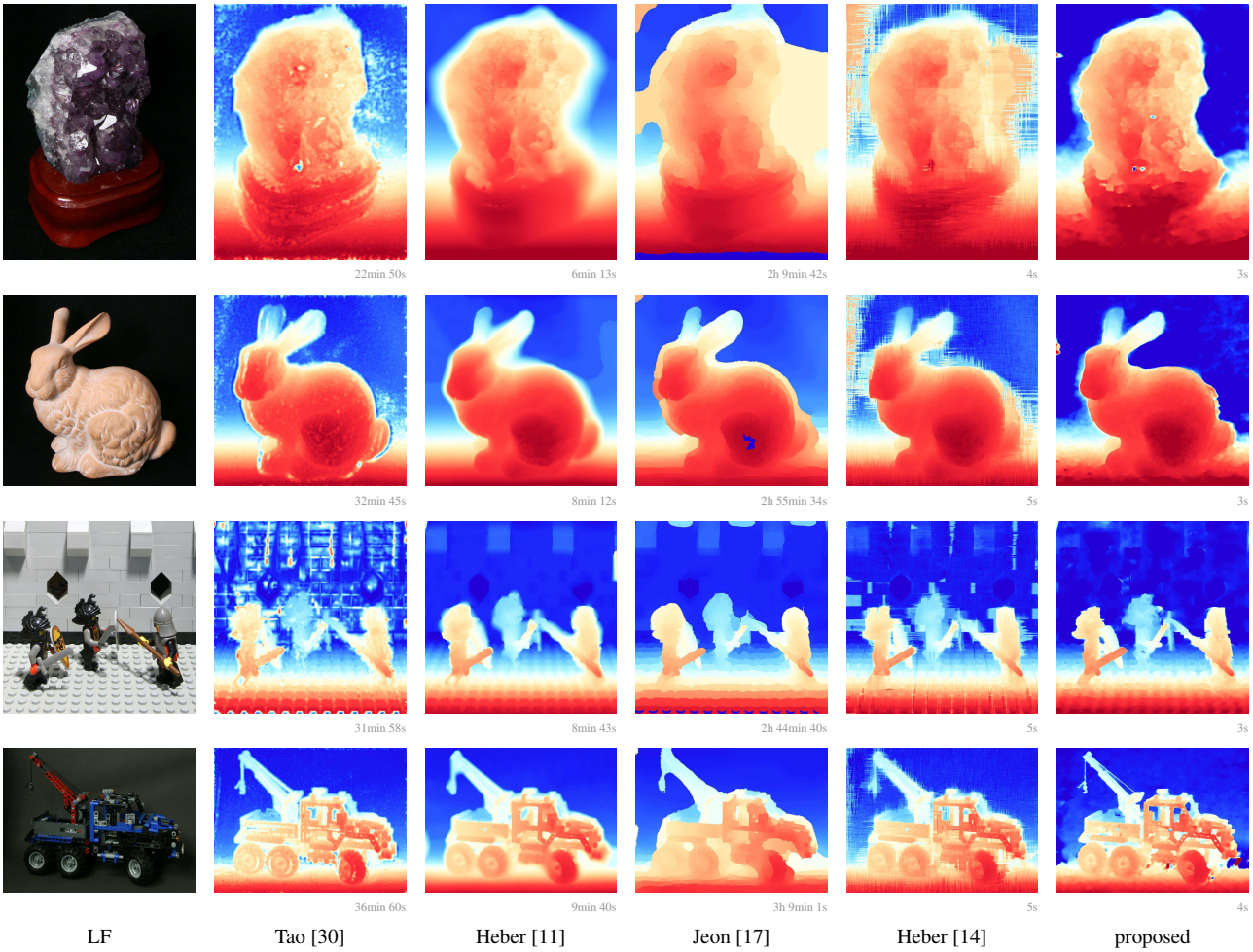

Figure 5. Qualitative comparison for LFs from the SLFA. The figure shows from left to right the center view of the LF, followed by the results for the methods proposed by Tao *et al*. [30], Heber and Pock [11], Jeon *et al*. [17], and Heber *et al*. [14]. The results to the right correspond to the proposed method.

than those produced by competing state-of-the-art methods. The proposed approach combines the two main advantages of the model proposed in [14], namely the low error rate and the low inference time, with an additional spatial regularization that reduces unpleasant depth artifacts.

The presented results suggest that CNNs are well suited for SfLF. More general, applying deep learning to LF image processing tasks is a promising direction of research because the progress remaining to achieve in this area is tremendous. In this paper we focused on SfLF, but the same type of network architecture can also be used for a large variety of applications in LF image processing, including denoising, segmentation, and super-resolution. Testing the proposed network for the above-mentioned applications is left as future work.

Acknowledgment. This work was supported by the Vision+ project *Integrating visual information with independent knowledge*, No. 836630.

References

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[2] M. aurelio Ranzato, C. Poultney, S. Chopra, and Y. L. Cun. Efficient learning of sparse representations with an energy-based model. In B. Schölkopf, J. C. Platt, and T. Hoffman,

editors, *Advances in Neural Information Processing Systems 19*, pages 1137–1144. MIT Press, 2007.

[3] R. C. Bolles, H. H. Baker, and D. H. Marimont. Epipolar-plane image analysis: An approach to determining structure from motion. *International Journal of Computer Vision*, 1(1):7–55, 1987.

[4] C. Chen, H. Lin, Z. Yu, S. Bing Kang, and J. Yu. Light field stereo matching using bilateral statistics of surface cameras. June 2014.

[5] A. Criminisi, S. B. Kang, R. Swaminathan, R. Szeliski, and P. Anandan. Extracting layers and analyzing their specular properties using epipolar-plane-image analysis. *Computer Vision and Image Understanding (CVIU)*, 97:51–85, January 2005.

[6] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. van der Smagt, D. Cremers, and T. Brox. Flownet: Learning optical flow with convolutional networks. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[7] A. Dosovitskiy, J. T. Springenberg, and T. 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.

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

[9] 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.

[10] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *CoRR*, abs/1502.01852, 2015.

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

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

[13] S. Heber, R. Ranftl, and T. Pock. Variational Shape from Light Field. In *International Conference on Energy Minimization Methods in Computer Vision and Pattern Recognition*, 2013.

[14] S. Heber, W. Yu, and T. Pock. U-shaped networks for shape from light field. In *Proc. British Machine Vision Conf.*, 2016.

[15] HTC Vive. `http://www.vive.com`.

[16] A. Isaksen, L. McMillan, and S. J. Gortler. Dynamically reparameterized light fields. In *SIGGRAPH*, pages 297–306, 2000.

[17] 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.

[18] N. K. Kalantari, T.-C. Wang, and R. Ramamoorthi. Learning-based view synthesis for light field cameras. *ACM Trans. Graph.*, 35(6):193:1–193:10, Nov. 2016.

[19] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

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

[21] 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.

[22] 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.

[23] Lytro. `https://www.lytro.com/`.

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

[25] R. Ng. *Digital Light Field Photography*. Phd thesis, Stanford University, 2006.

[26] Oculus. `https://www.oculus.com/`.

[27] Pov-ray. `http://www.povray.org`.

[28] Raytrix. `https://www.raytrix.de/`.

[29] O. Ronneberger, P. Fischer, and T. Brox. *U-Net: Convolutional Networks for Biomedical Image Segmentation*, pages 234–241. Springer International Publishing, Cham, 2015.

[30] M. W. Tao, S. Hadap, J. Malik, and R. Ramamoorthi. Depth from combining defocus and correspondence using light-field cameras. In *International Conference on Computer Vision (ICCV)*, Dec. 2013.

[31] T.-C. Wang, J.-Y. Zhu, E. Hiroaki, M. Chandraker, A. A. Efros, and R. Ramamoorthi. *A 4D Light-Field Dataset and CNN Architectures for Material Recognition*, pages 121–138. Springer International Publishing, Cham, 2016.

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

[33] B. Wilburn, N. Joshi, V. Vaish, E.-V. Talvala, E. Antunez, A. Barth, A. Adams, M. Horowitz, and M. Levoy. High performance imaging using large camera arrays. *ACM Trans. Graph.*, 24(3):765–776, July 2005.

[34] J. Zbontar and Y. LeCun. Computing the stereo matching cost with a convolutional neural network. *CoRR*, abs/1409.4326, 2014.

[35] M. D. Zeiler and R. 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.