Intrinsic Light Field Decomposition and Disparity Estimation with Deep Encoder-Decoder Network

Anna Alperovich
University of Konstanz
Konstanz, Germany
anna.alperovich@uni-konstanz.de

Ole Johannsen
University of Konstanz
Konstanz, Germany
ole.johannsen@uni-konstanz.de

Bastian Goldluecke
University of Konstanz
Konstanz, Germany
bastian.goldluecke@uni-konstanz.de

***Abstract*—We present an encoder-decoder deep neural network that solves non-Lambertian intrinsic light field decomposition, where we recover all three intrinsic components: albedo, shading, and specularity. We learn a sparse set of features from 3D epipolar volumes and use them in separate decoder pathways to reconstruct intrinsic light fields. While being trained on synthetic data generated with Blender, our model still generalizes to real world examples captured with a Lytro Illum plenoptic camera. The proposed method outperforms state-of-the-art approaches for single images and achieves competitive accuracy with recent modeling methods for light fields.**

I. Introduction

Intrinsic images are a well-known computer vision problem that has been extensively studied over decades. The first solution was introduced by Barrow and Tenenbaum [1] in 1978, where the input image is decomposed into reflectance and illumination components. Since then, numerous methods were proposed, based on modeling and data-driven techniques.

Most of the intrinsic image algorithms follow the Lambertian assumption that a scene is composed of objects with only diffuse or body reflection. In this case, the input image is decomposed into the albedo (or reflectance) component which represents surface color/texture and the shading component that represents intensity changes due to geometry and illumination.

In the real world, there are few materials that are purely Lambertian. Most of them exhibit some amount of specularity. Thus, intrinsic images for non-Lambertian objects aims to decompose the input into albedo, shading, and specularity components [2]. Because the specularity component is view dependent it can hardly be estimated from a single image, thus additional information is required.

Light fields appear to be a good source of information that can be used to improve intrinsic image decomposition [3]–[5]. Existing methods rely on modeling of intrinsic components based on their physical properties. Although these approaches give impressive results, they require lots of assumptions and are accurate up-to-the models used.

Contributions. With the growing popularity of data-driven methods, our work focuses on intrinsic light field decomposition and disparity estimation with an encoder-decoder end-to-end deep neural network. These tasks can benefit from each other in a sense that estimated disparity contains information about the geometry of the scene, which, is strongly related to shading and specularity. Moreover, albedo, shading, and specular flow are useful cues for disparity estimation.

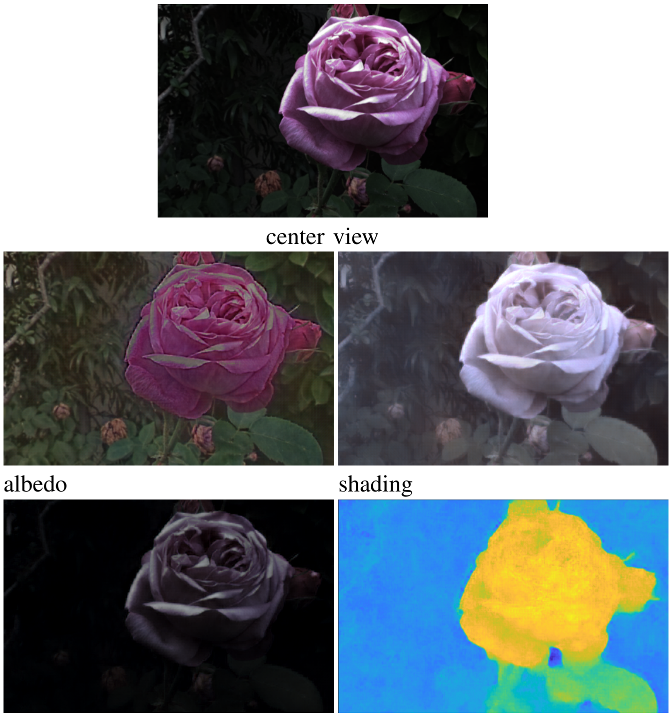

Fig. 1. Our network estimates disparity and decomposes the crosshair-shaped subset of 17 views from the input light field into intrinsic components: albedo, shading, and specularity. Figure illustrates the center view result on real world flowers dataset [6] taken with Lytro Illum plenoptic camera. The light field size is $9\times9\times376\times541\times3$, with the estimated disparity range $[-1.19, 0.57]$.

Similar to Alperovich et al. [7] where the authors only recover disparity, diffuse and specular components, we use horizontal and vertical 3D volumes as an input to our network. Contrary to [7] we perform full intrinsic decomposition, design an architecture that is capable of processing twice larger patches as an input and introduce skip connections from the encoder to corresponding decoder parts that improve the reconstruction quality of decoders. We substitute $3D$ convolutions with a sequence of $2D$ that acting in spatial and angular domains. This design choice decreases the number of parameters in the network and speeds up the training process. As in [7], our network can process light fields where a ground truth is not available.

ISBN 978-90-827970-1-5 © EURASIP 2018

We evaluate our network on synthetic data generated with Blender and real world examples taken with Lytro Illum plenoptic camera. We perform comparisons with a single image CNN-based algorithm and recent methods for light fields.

II. Related work

The problem of **disparity estimation** benefits a lot from the multiple views available in a light field, for an overview of various algorithms we refer to the recent work of Johannsen et al. [8].

Deep learning approaches lead to significant improvement in this task. Heber et al. [9] recover a $4D$ depth field from the light field using CNN followed by convex optimization to refine point-wise predictions from the deep network. Srinivasan et al. [6] synthesize a $4D$ light field from a single image with two neural networks, one estimates the disparity and renders Lambertian light field, and the second one predicts occluding rays and models non-Lambertian effects. Alperovich et al. [7] jointly solve disparity estimation and reflection separation tasks with a fully-convolutional encoder-decoder network.

For an overview of **intrinsic decomposition** algorithms we refer to the work of Bonneel et al. [10] where the authors discuss priors used for modeling intrinsic components.

Among the deep learning approaches Narihira et al. [11] was the first to introduce CNN for recovering relative lightness that was trained on human judgments on relative reflectance [12]. Later they developed a regression CNN-based model that predicts albedo and shading components. Shi et al. [13] introduced a mirror-like, U-shaped architecture that solves non-Lambertian intrinsic decomposition from a single image. Janner et al. [14] developed a self-supervised (RIN) model which predicts reflectance, shape, and lighting conditions given a single image.

III. Underlying model

For the purpose of this paper, we understand a **4D light field** as the radiance function sampled on a space of rays that form regular grid of sub-aperture views. This 4D ray space is parameterized by the two intersection points of each ray $\boldsymbol{r}$ with two different planes. The *image plane* Ω is parameterized in $\boldsymbol{p} = (x, y)$ coordinates, while the *focus plane* Π is parameterized in $\boldsymbol{c} = (s, t)$ coordinates. Both planes are parallel to each other. Thus, the 4D light field is a function $L : \Omega \times \Pi \to \mathbb{R}$ with $(x, y, s, t) \mapsto L(x, y, s, t) = L(\boldsymbol{p}, \boldsymbol{c})$. In practice, it often has several components, i.e. takes values in RGB color space $\mathbb{R}^3$. By convention, the view with focal coordinate $\boldsymbol{c} = \boldsymbol{0}$ is called the center (or reference) view. Most disparity estimation algorithms only compute depth for this specific view.

According to the **dichromatic reflection model** [15], the total radiance I of the reflected light is the sum of two independent parts, the radiance I_d of the reflected light at the surface body and radiance I_s at interface. More precisely,

$$I(\lambda, \boldsymbol{n}, \boldsymbol{l}, \boldsymbol{v}) = I_d(\lambda, \boldsymbol{n}, \boldsymbol{l}, \boldsymbol{v}) + I_s(\lambda, \boldsymbol{n}, \boldsymbol{l}, \boldsymbol{v}), \qquad (1)$$

where λ is the wavelength of the light, $\boldsymbol{n}$ is the surface normal, $\boldsymbol{v}$ is the viewing direction, and $\boldsymbol{l}$ direction to the light source. In particular, it is a complex function of the local surface geometry.

As discussed in detail in [16], we consider the case where the diffuse component is modeled as Lambertian, and can rewrite (1) for a light field L as

$$L(\boldsymbol{v}) = m_d(\boldsymbol{n}, \boldsymbol{l}) c_d(\lambda) + m_s(\boldsymbol{n}, \boldsymbol{l}, \boldsymbol{v}) c_s(\lambda), \qquad (2)$$

where $m_{d,s}$ is a geometric scale factor, $c_{d,s}$ is the spectral power distribution. In short, we can represent a light field as a sum of diffuse and specular H component. By further breaking down diffuse component into albedo A and shading S, we arrive at the intrinsic light fields

$$L(\boldsymbol{r}) = A(\boldsymbol{r}) S(\boldsymbol{r}) + H(\boldsymbol{r}) \qquad (3)$$

model described by Alperovich and Goldluecke [4].

We can see that according to our model, specular and diffuse components behave quite differently in EPIs [17]. While the albedo and shading have constant color along the projections of 3D points given by disparity, specular reflection moves in a non-rigid way and depends on local surface geometry [18], [19]. Based on the definitions and observations above, we can now start to build our network.

IV. Network architecture

Similar to the previous work [7], [13], [14] we use a U-shaped mirror-like architecture [20] to build the network, see Figure 2 for the detailed explanation of the network structure. The key idea is to extract a small number of features, 3.8% of the input patch size, and then upscale them back to the input light field, intrinsic components and disparity.

The input to the network is a pair of $9 \times 96 \times 96 \times 3$ horizontal and vertical $3D$ slices of the light field which have an overlap of 32 pixels. Every patch passes 12 convolution layers where each layer is represented by the residual block, see Figure 3.

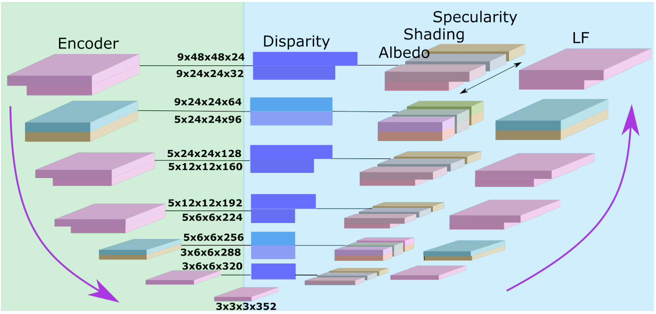

Fig. 2. Encoder-decoder residual network with 12 convolution layers in the encoder part (green rectangle) and corresponding up-sampling layers in the decoder (blue rectangle). The network has four $3D$ decoding pathways: albedo, shading, specularity and light field itself, and one $2D$ pathway for disparity. We illustrate skip connections with lines. The arrow in the last decoding layer for intrinsic components illustrates that albedo, shading and specularity shares their features to better cope with modeling cost (3). Numbers describe the output dimension of a tensor after each layer. Block color corresponds to its kind, for instance for encoder and decoder pathways purple is spatial downscaling, blue is dimension preserving, and orange is angular downscaling.

ISBN 978-90-827970-1-5 © EURASIP 2018

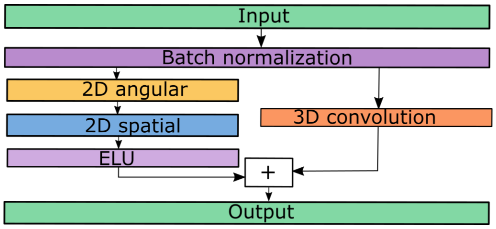

Fig. 3. Example residual block of the network. After batch normalization, we pass the output tensor through two pathways. The right one keeps the input tensor or performs $3D$ convolution if it needs to be resampled. The left one performs angular and spatial $2D$ convolutions followed by ELU layer. The output tensor is the sum of these two pathways.

After batch normalization we duplicate the output tensor and successively apply $2D$ convolution on the EPI followed by $2D$ spatial convolution to one copy. Then we pass the output through the Exponential Linear Units (ELU) layer [21]. Another copy is kept unchanged or resampled to have the same shape as the first copy. The output of the block is the sum of two copies.

Based on strides used, the residual blocks are divided into three groups: spatial downscaling, angular downscaling and dimensions preserving, see Figure 2.

To improve the accuracy of decomposition we add skip connections [13], [14] by copying encoder features to the corresponding outputs of the decoder layers. Then we pass a new tensor through a $1 \times 1 \times 1$ convolution to preserve its original shape. Note that there are no skip connections in the pure autoencoder.

The decoder consists of five pathways. To ensure that intrinsic components follow the model (3), we concatenate their features in the last decoding layer and then up-sample to the output albedo, shading, and specularity.

We use scale invariant loss [11], [22] for albedo A and shading S to reduce the ambiguity caused by the product of those components in the intrinsic decomposition, see model (3). For all other decoders, we use standard MSE loss function.

V. Experiments

Most of the **training data** is generated with the Blender addon provided by [26]. We produced 400 light fields of size $9 \times 9 \times 512 \times 512 \times 3$, which results in a total of $78,400$ training patches. Since our network uses only horizontal and vertical slices, we render high-quality ground truth only for a crosshair-shaped subset of 17 views. For unsupervised training, we use the light field benchmark [26] with only disparity ground truth available and real world light fields [6], [27]–[29] without any ground truth.

We train the network with batch size 10 for $160K$ iterations starting from learning rate 10^{-4} and dropping it every $40K$ iterations till it reaches 10^{-5}. The evaluation of the trained network takes about 11s for $3D$ pathway and 2.6s for $2D$ on a machine with an Intel(R) Core i7-4790 CPU 3.60GHz and an NVIDIA GeForce GTX 1080Ti.

Since there are no algorithms for light fields that jointly solve intrinsic decomposition and disparity estimation, we perform separate **comparisons** for these tasks.

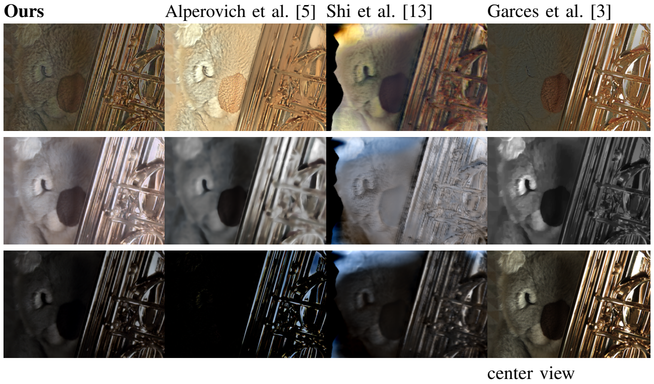

Fig. 4. Intrinsic decomposition for real world $9 \times 9 \times 434 \times 625 \times 3$ light field captured with a Lytro Illum plenoptic camera. Rows from top to bottom: albedo, shading, and specularity. Note, that method by Garces et al. [3] performs only albedo/shading decomposition. We use space for the specular image to illustrate the center view. The single image CNN [13] does not perform decomposition for the background, thus it appears black in the visualization. We conclude that our method copes well with soft shadows and highly specular materials. Also it preserves all the structure compared to over-smoothed results by [5].

For intrinsic decomposition, we select three methods for comparison. The first one is a modeling approach for light fields proposed by Alperovich et al. [5], where the authors model priors according to their physical properties. The second one is proposed by Garces et al. [3], where the authors decompose the input light field into albedo and shading components with extended Retinex theory. The third one is a single image CNN-based method by Shi et al [13], where the authors develop a deep network for non-Lambertian intrinsic decomposition. See Figures 4, 1 for results on the real world data, and Figure 5 for evaluations on the test data.

Note that the method by Garces et al. [3] uses the whole light field as an input, thus we perform comparisons only for the real world data captured with a Lytro Illum plenoptic camera, where we have all 81 views available. Alperovich et al. [5] use horizontal and vertical slices of the light field, and Shi et al. [13] use single image plus object mask, all this information is available in our synthetic data.

For quantitative evaluations we select three error metrics: local mean-squared error (LMSE) [2] computed patch-wise with the size of 40% of the image size, global mean-squared error (GMSE) [5], which is similar to LMSE, but computed for the whole image, and DSSIM index [30] which is defined as $(1 - SSIM)/2$ and measures structure dissimilarity. See Table 8 for numerical evaluations on the center view over 9 test data sets.

For disparity estimation, we select four algorithms for comparison. The recent deep network by Alperovich et al. [7] that jointly decomposes input light field into diffuse and specular components and estimates the disparity. Johannsen et al. [24] employ dictionary learning to recover disparity. Strecke et al. [23] estimate disparity with occlusion-aware focal stack symmetry with additional normals refinement. The last one is the method proposed by Wanner and Goldluecke [25] which is based on orientation of EPI patches. We use the standard MSE metric for the comparisons, see Figure 6 for visual and

ISBN 978-90-827970-1-5 © EURASIP 2018

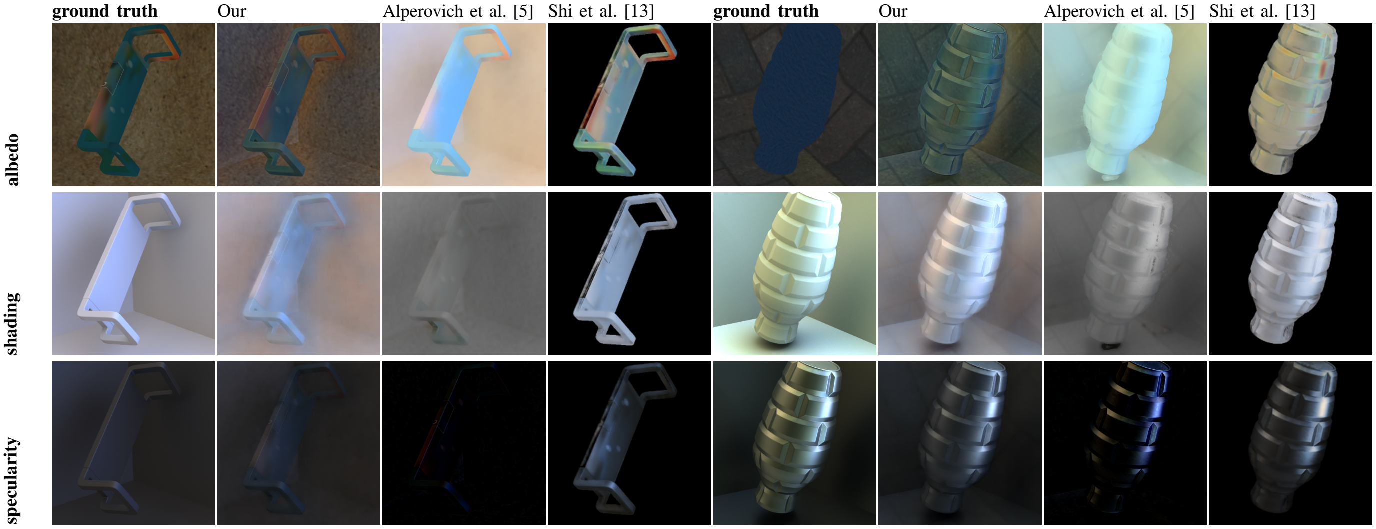

Fig. 5. Comparison on two synthetic data sets generated with Blender. The light field size is $9 \times 9 \times 512 \times 512 \times 3$. We conclude that proposed network outputs more accurate albedo and shading components compared to single image CNN [13]. Compared to the modeling method by [5], the albedo is much sharper and specularity is more intense. The shading component produced by our method still contains some texture compared to shading by [5], also we do not train the network on examples with strong cast shadows, thus shadows are not fully removed from albedo. Due to ambiguity between albedo and shading we observe the difference between scaling of the ground truth and decomposition results.

quantitative evaluations.

We use overlapping patches to compute the decomposition as the weighted average, assuming that pixels that are close to the patch center are more accurate. Thus we exclude border pixels from the final results.

VI. Conclusions

We propose a novel architecture that outperforms recent methods for disparity estimation and intrinsic decomposition. The key idea is to replace $3D$ convolutions with the sequence of $2D$ angular and spatial convolutions, which decreases the number of parameters in the network. The resulting architecture is more computationally efficient and allows larger patch sized compared to [7]. As the result, we are able to train the network with four $3D$ and one $2D$ decoders.

From our experiments, we conclude that with a larger patch size the disparity estimation is much more accurate, especially on the large specular surfaces, see Figure 6.

While ground truth is available only for the synthetic scenes, the proposed architecture still generalizes well to the real world light fields, it successfully removes most of the shading from the albedo component and correctly detects and separates specularity.

Given the difficulty of the tasks, our network achieves superior performance, but there is plenty of work ahead. The estimated shading component contains some structure from the albedo, thus additional post-processing might be needed. Also, our current architecture can deal only with soft shadows, because training scenes are illuminated with environmental maps, that results in soft lighting.

Acknowledgments

This work was supported by the ERC Starting Grant "Light Field Imaging and Analysis" (LIA 336978, FP7-2014).

References

[1] H. G. Barrow and J. M. Tenenbaum, "Recovering intrinsic scene characteristics from images," *Computer Vision Systems, Academic Press*, vol. 23, no. 1, pp. 3–26, 1978.

[2] R. Grosse, M. K. Johnson, E. H. Adelson, and W. T. Freeman, "Ground truth dataset and baseline evaluations for intrinsic image algorithm," in *Proc. ICCV*, 2009.

[3] E. Garces, J. I. Echevarria, W. Zhang, H. Wu, K. Zhou, and D. Gutierrez, "Intrinsic light field images," *Computer Graphics Forum*, 2017.

[4] A. Alperovich and B. Goldluecke, "A variational model for intrinsic light field decomposition," in *Proc. ACCV*, 2016.

[5] A. Alperovich, O. Johannsen, M. Strecke, and B. Goldluecke, "Shadow and specularity priors for intrinsic light field decomposition," in *EMM-CVPR*, 2017.

[6] P. Srinivasan, T. Wang, A. Sreelal, R. Ramamoorthi, and R. Ng, "Learning to synthesize a 4D RGBD light field from a single image," in *Proc. ICCV*, 2017.

[7] A. Alperovich, O. Johannsen, M. Strecke, and B. Goldluecke, "Light field intrinsics with a deep encoder-decoder network," in *Proc. CVPR*, 2018.

[8] O. Johannsen, K. Honauer, B. Goldluecke, A. Alperovich, F. Battisti, Y. Bok, M. Brizzi, M. Carli, G. Choe, M. Diebold, M. Gutsche, H.-G. Jeon, I. S. Kweon, J. Park, H. Schilling, H. Sheng, L. Si, M. Strecke, A. Sulc, Y.-W. Tai, Q. Wang, T.-C. Wang, S. Wanner, Z. Xiong, J. Yu, S. Zhang, and H. Zhu, "A taxonomy and evaluation of dense light field depth estimation algorithms," in *2nd Workshop on Light Fields for Computer Vision at CVPR*, 2017.

[9] S. Heber, W. Yu, and T. Pock, "Neural epi-volume networks for shape from light field," in *Proc. ICCV*, 2017.

[10] N. Bonneel, B. Kovacs, S. Paris, and K. Bala, "Intrinsic decompositions for image editing," *Computer Graphics Forum (Eurographics State of the Art Reports 2017)*, vol. 36, no. 2, 2017.

[11] T. Narihira, M. Maire, and S. X. Yu, "Direct intrinsics: Learning albedo-shading decomposition by convolutional regression," in *Proc. ICCV*, 2015.

[12] S. Bell, K. Bala, and N. Snavely, "Intrinsic images in the wild," *ACM Trans. on Graphics (SIGGRAPH)*, vol. 33, no. 4, pp. 159:1–159:12, 2014.

[13] J. Shi, Y. Dong, H. Su, and S. Yu, "Learning non-lambertian object intrinsics across shapenet categories," in *Proc. CVPR*, 2017.

[14] M. Janner, J. Wu, T. Kulkarni, I. Yildirim, and J. B. Tenenbaum, "Self-Supervised Intrinsic Image Decomposition," in *Proc. NIPS*, 2017.

[15] S. Shafer, "Using color to separate reflection components," *Color Research & Application*, vol. 10, no. 4, pp. 210–218, 1985.

ISBN 978-90-827970-1-5 © EURASIP 2018

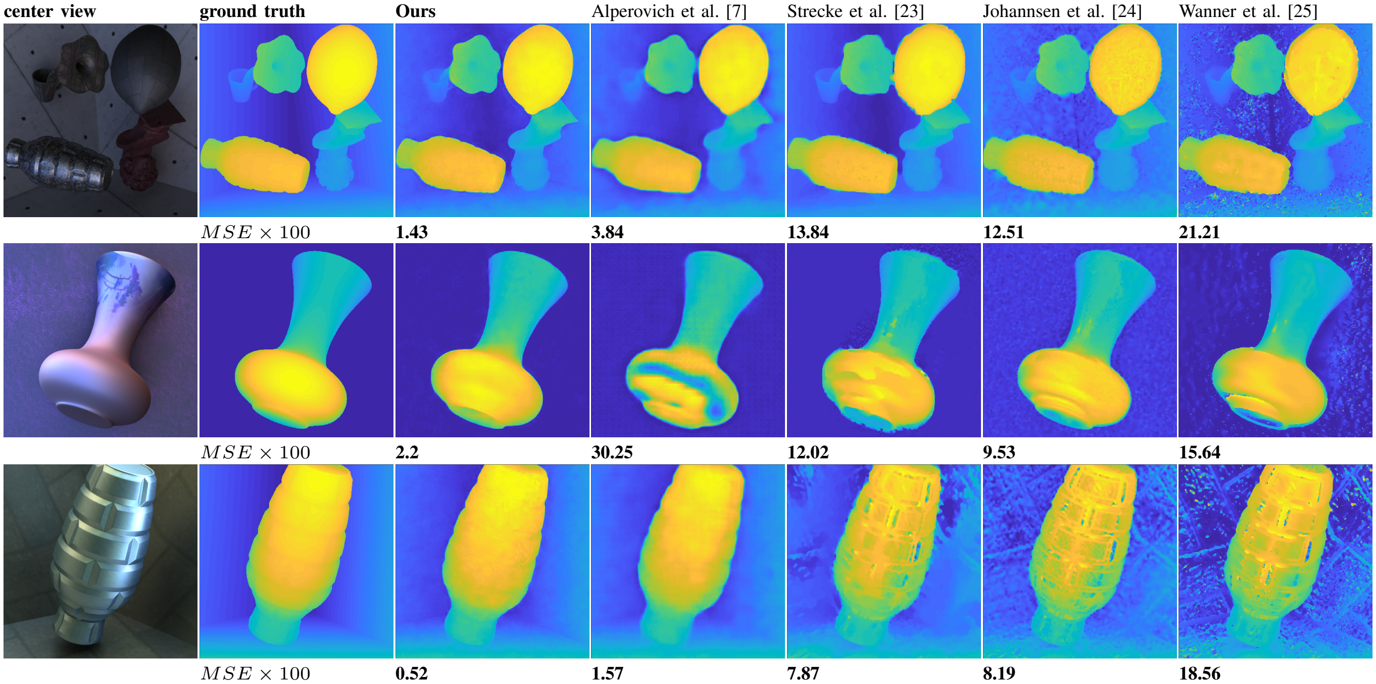

Fig. 6. Ground truth and estimated disparity for three synthetic data sets generated with Blender. The disparity range is $[-2.12, 2.51]$. Compared to [7] our method improves disparity estimation in large specular regions. We explain this behavior by increase of input patch size. Other methods also fail to estimate correct disparity on specular and structureless surfaces. Although we use similar loss function for the disparity, due to the skip connections, the new disparity is sharper than in [7]. Numerical evaluations support qualitative results, MSE error is smaller than for the competing methods.

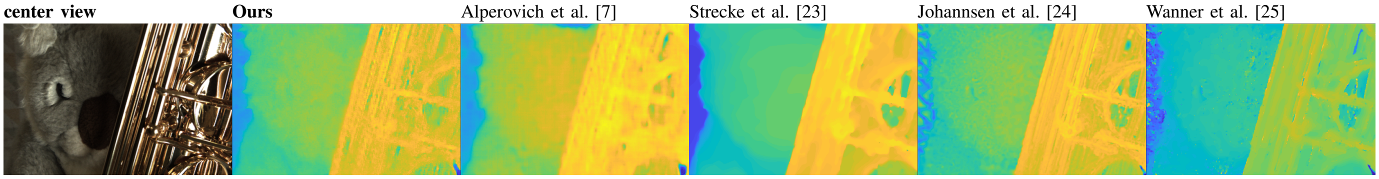

Fig. 7. Disparity estimation for the real world light field from Figure 4. The estimated disparity range is $[-2.44, 1.14]$. This example is particularly difficult for disparity estimation because it has highly specular object. Based on visual comparison we conclude that our method produces similar quality results compared to CNN-based and modeling approaches. Method by Strecke et al. [23] produces visually more accurate results on the saxophone.

	LMSE ×100			GMSE ×100			DSSIM ×100		
	A	S	H	A	S	H	A	S	H
Ours	**0.13**	**0.34**	**0.04**	**0.25**	**0.58**	**0.09**	**9.51**	**5.79**	**4.43**
Alperovich [5]	0.15	0.73	0.49	0.26	1.13	1.0	17.96	14.10	19.72
Shi et al. [13]	0.17	0.9	0.09	0.31	1.47	0.2	13.46	9.74	6.19

Fig. 8. Numerical evaluation of intrinsic decomposition. The numbers represent average LMSE, GMSE, and DSSIM over nine synthetic test data sets. Since the single image CNN [13] does not perform decomposition for the background, we multiply all results and ground truth with object mask before measuring the errors. We conclude that the use of light fields improves intrinsic decomposition, compared to [13]. Our method outperforms modeling approach [5] especially on DSSIM measure.

[16] M. Tao, J. C. Su, T. C. Wang, J. Malik, and R. Ramamoorthi, “Depth estimation and specular removal for glossy surfaces using point and line consistency with light-field cameras,” *IEEE TPAMI*, vol. 38, no. 6, pp. 1155–1169, 2015.

[17] A. Criminisi, S. 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*, vol. 97, no. 1, pp. 51–85, 2005.

[18] R. Swaminathan, S. B. Kang, R. Szeliski, A. Criminisi, and S. K. Nayar, “On the motion and appearance of specularities in image sequences,” in *Proc. ECCV*, vol. I, May 2002, pp. 508–523.

[19] A. Sulc, A. Alperovich, N. Marniok, and B. Goldluecke, “Reflection separation in light fields based on sparse coding and specular flow,” in *VMV*, 2016.

[20] O. Ronneberger, P.Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *MICCAI*, 2015.

[21] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, “Fast and accurate deep network learning by exponential linear units (elus),” *CoRR*, vol. abs/1511.07289, 2015.

[22] D. Eigen, C. Puhrsch, and R. Fergus, “Depth map prediction from a single image using a multi-scale deep network,” in *Proc. NIPS*, 2014.

[23] M. Strecke, A. Alperovich, and B. Goldluecke, “Accurate depth and normal maps from occlusion-aware focal stack symmetry,” in *Proc. CVPR*, 2017.

[24] O. Johannsen, A. Sulc, and B. Goldluecke, “What sparse light field coding reveals about scene structure,” in *Proc. CVPR*, 2016.

[25] S. Wanner and B. Goldluecke, “Globally consistent depth labeling of 4D light fields,” in *Proc. CVPR*, 2012, pp. 41–48.

[26] K. Honauer, O. Johannsen, D. Kondermann, and B. Goldluecke, “A dataset and evaluation methodology for depth estimation on 4D light fields,” in *Proc. ACCV*, 2016.

[27] 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 Transactions on Graphics*, vol. 24, pp. 765–776, July 2005.

[28] V. Vaish and A. Adams, “The (New) Stanford Light Field Archive,” `http://lightfield.stanford.edu`, 2008.

[29] S. Wanner, C. Straehle, and B. Goldluecke, “Globally consistent multi-label assignment on the ray space of 4D light fields,” in *Proc. CVPR*, 2013.

[30] Q. Chen and V. Koltun, “A simple model for intrinsic image decomposition with depth cues,” in *Proc. ICCV*, 2013.