Accurate Depth and Normal Maps from Occlusion-Aware Focal Stack Symmetry

Michael Strecke, Anna Alperovich, and Bastian Goldluecke
University of Konstanz
firstname.lastname@uni-konstanz.de

Abstract

We introduce a novel approach to jointly estimate consistent depth and normal maps from 4D light fields, with two main contributions. First, we build a cost volume from focal stack symmetry. However, in contrast to previous approaches, we introduce partial focal stacks in order to be able to robustly deal with occlusions. This idea already yields significanly better disparity maps. Second, even recent sublabel-accurate methods for multi-label optimization recover only a piecewise flat disparity map from the cost volume, with normals pointing mostly towards the image plane. This renders normal maps recovered from these approaches unsuitable for potential subsequent applications. We therefore propose regularization with a novel prior linking depth to normals, and imposing smoothness of the resulting normal field. We then jointly optimize over depth and normals to achieve estimates for both which surpass previous work in accuracy on a recent benchmark.

1. Introduction

In light field imaging, robust depth estimation is the limiting factor for a variety of useful applications, such as super-resolution [30], image-based rendering [22], or light field editing [13]. More sophisticated models for *e.g.* intrinsic light field decomposition [1] or reflectance estimation [29] often even require accurate surface normals, which are much more difficult to achieve as the available cues about them are subtle [29].

Current algorithms, *e.g.* [16, 18, 28, 5, 14] and many more cited in the references, work exceedingly well for estimating depth from light field images. However, methods are usually not designed with normal estimation in mind. Thus, depth estimates from algorithms based on cost volumes, even when optimized with sublabel accuracy [19], are often piecewise flat and thus fail at predicting accurate normal maps. Frequently, their accuracy is also naturally limited around occlusion boundaries [18, 15, 30]. The aim of this work is to contribute towards a remedy for these drawbacks.

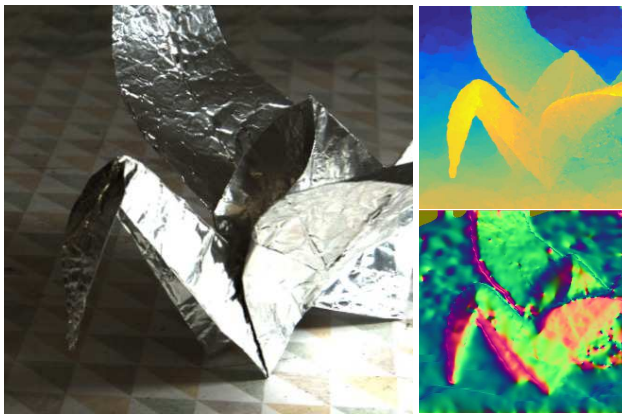

Figure 1. We present a novel idea to compute disparity cost volumes which is based on the concept of occlusion-aware focal stack symmetry. Using the proposed framework, we can optimize jointly for depth and normals to reconstruct challenging real-world scenes captured with a Lytro Illum plenoptic camera.

Contributions. In this work, we make two main contributions. First, we introduce a novel way to handle occlusions when constructing cost volumes based on the idea of focal stack symmetry [18]. This novel data term achieves substantially more accurate results than the previous method when a global optimum is computed with sublabel relaxation [19]. Second, we propose post-processing using joint regularization of depth and normals, in order to achieve a smooth normal map which is consistent with the depth estimate. For this, we employ ideas from Graber *et al*. [7] to linearly couple depth and normals, and employ the relaxation in Zeisl *et al*. [31] to deal with the non-convexity of the unit length constraint on the normal map. The resulting sub-problems on depth and normal regularization can be efficiently solved with the primal-dual algorithm in [4]. Our results substantially outperform all previous work that has been evaluated on the recent benchmark for disparity estimation on light fields [12] with respect to accuracy of disparity and normal maps and several other metrics.

Figure 2. A light field is defined on a 4D volume parametrized by image coordinates (x, y) and view point coordinates (s, t). Epipolar images (EPIs) are the slices in the sx- or yt-planes depicted to the right and below the center view. By integrating the 4D volume along different orientations in the epipolar planes (blue and green), one obtains views with different focus planes, see section 3.

2. Background and related work

Methods for disparity estimation from light fields can roughly be classified according to the underlying representation. For this reason, and to fix notation for the upcoming sections, we will briefly review the most common parametrizations of a light field while discussing related methods.

Two-plane representation and subaperture views. In this work, we consider 4D light fields, which capture the radiance of rays passing through an image plane Ω and focal plane Π. By fixing view point coordinates (s, t), one obtains a 2D *subaperture image* in (x, y)-coordinates as if captured by an ideal pinhole camera. Matching subaperture images means doing just multiview stereo, for which there is a multitude of existing methods. Interesting variants specific to the light field setting construct novel matching scores based on the availability of a dense set of views [10, 8]. A useful transposed representation considers the projections of scene points at a certain distance to the image plane into all subaperture views. The resulting view on the light field is called an S-CAM [5] or angular patch [28], and it can be statistically analyzed to obtain disparity cost volumes robust to occlusion [28] and specular reflections [5].

Epipolar plane images. Analyzing horizontal and vertical slices through the 4D radiance volume, so-called epipolar plane images (EPIs), see figure 2, has been pioneered in [2]. Subsequently, the ideas have been adapted to disparity estimation in various ways. Proposed methods include leveraging of the structure tensor [30] or special filters [26] to estimate the slope of epipolar lines, iterative epipolar line extraction for better occlusion handling [6], fine-to-coarse approaches focusing on correct object boundaries [16], building patch dictionaries with fixed disparities [15], or training a convolutional neural network for orientation analysis [9]. EPI lines are employed in [14] for distortion correction, but they subsequently construct a cost volume from subaperture view matching.

Focal stack. It is well known that shape can be estimated accurately from a stack of images focused at different distances to the camera [20]. Furthermore, a 4D light field can be easily transformed into such a focal stack, see figure 2, to apply these ideas to estimate depth. Lin *et al*. [18] exploit the fact that slices through the focal stack are symmetric around the true disparity, see figure 3. Correct occlusion handling is critical in this approach, and we improve upon this in section 3 as a main contribution of our work. Authors of [24, 28] combine stereo and focus cues to arrive at better results. In particular, Tao *et al*. [24] propose confidence measures to automatically weight the respective contributions to the cost volumes, which we can also apply as an additional step to increase resilience to noise, see section 5.

Regularization and optimization. Regardless of how the disparity cost volume is constructed, more sophisticated methods typically perform optimization of a functional weighting the cost with a regularization term. Key differences lie in the type of regularization and how a minimizer of the cost function is found. Popular optimization methods include discrete methods like graph cuts [17] or semi-global matching [11], or continuous methods based on the lifting idea [21], which was recently extended to achieve sublabel accuracy [19]. If one wants to obtain the exact global minimum of cost and regularization term, the class of regularizers is severely restricted. To become more general and also speed up the optimization, coarse-to-fine approaches are common [7], extending the possible class of regularizers to sophisticated ones like *e.g*. total generalized variation [3]. A recent idea was to construct an efficient minimal surface regularizer [7] using a linear map between a reparametrization of depth and the normals scaled by the local area element. We embrace this idea, but extend it to a coupled optimization of depth and normal map, in order to achieve better smoothing of the latter. This will be shown in section 4.

3. Occlusion-aware focal stack symmetry

For accurate normal estimation, good depth estimates are crucial. There are several algorithms exploiting depth cues available from multiple views and focus information that can be obtained from light field data. Among the most robust algorithms with respect to noise is Lin *et al.*'s [18] cost volume based on focal stack symmetry. This algorithm is based on the observation that for planar scenes parallel to the image plane, focal shifts in either direction from the ground truth disparity result in the same color values and that thus there is a symmetry in the focal stack around the ground truth disparity d. We will briefly review the foundations and then slightly generalize the symmetry property.

Focal stack symmetry [18]. For refocusing of the light field, one integrates a sheared version of the radiance volume L over the subaperture views (u, v) weighted with an aperture filter σ,

$$\varphi_{\boldsymbol{p}}(\alpha) = \int_{\Pi} \sigma(\boldsymbol{v})\, L(\boldsymbol{p} + \alpha \boldsymbol{v}, \boldsymbol{v})\, \mathrm{d}\boldsymbol{v}, \tag{1}$$

where $\boldsymbol{p} = (x, y)$ denotes a point in the image plane Ω, and $\boldsymbol{v} = (s, t)$ the focal point of the respective subaperture view. Without loss of generality, we assume that the center (or reference) view of the light field has coordinates $\boldsymbol{v} = (0, 0)$. To further simplify formulas, we omit σ in the following (one may assume it is subsumed into the measure $\mathrm{d}\boldsymbol{v}$). Finally, α denotes the disparity of the synthetic focal plane.

Lin *et al.* [18] observed that under relatively mild conditions, the focal stack is symmetric around the true disparity d, *i.e.* $\varphi_{\boldsymbol{p}}(d + \delta) = \varphi_{\boldsymbol{p}}(d - \delta)$ for any $\delta \in \mathbb{R}$. The conditions are that the scene is Lambertian, as well as locally constant disparity. In practice, it is sufficient for the disparity to be slowly varying on surfaces. In their work, they leverage this observation to define a focus cost as

$$s^{\varphi}_{\boldsymbol{p}}(\alpha) = \int_0^{\delta_{\max}} \rho(\varphi_{\boldsymbol{p}}(\alpha + \delta) - \varphi_{\boldsymbol{p}}(\alpha - \delta))\, \mathrm{d}\delta, \tag{2}$$

which is small if the stack is more symmetric around α. Above, $\rho(v) = 1 - e^{-|v|_2/(2\sigma^2)}$ is a robust distance function.

The main problem of this approach is that it does not show the desired behaviour near occlusion boundaries. Because pixels on the occluder smear into the background when refocusing to the background, one can observe that the focal stack is actually more symmetric around the occluder's ground truth disparity instead of the desired background disparity, see figure 3. Of course, Lin *et al.* [18] already observed this and proposed handling the problem by choosing an alternative cost for occluded pixels detected by an estimated occlusion map. We propose an alternative approach, which does not require error-prone estimation of an occlusion map, and only uses light field data instead.

Occlusion-aware focal stack symmetry. In order to tackle problems around occlusions, we use occlusion-free partial focal stacks. We do not refocus the light field using all subaperture views, but create four separate stacks using only the views right of, left of, above and below the reference view. The assumption is that the baseline is small enough so that if occlusion is present it occurs only in one direction of view point shift.

We will see that depending on the occlusion edge orientation, there will be symmetry around the background disparity between the top and bottom or the left and right focal stacks. To see this, we prove the following observation, which will lead to the definition of our modified focus cost volume. Essentially, it refines the focal stack symmetry property defined on the complete stack to symmetry along arbitrary directions of view point shift.

Proposition. *Let d be the true disparity value of the point $\boldsymbol{p}$ in the image plane of the reference view. Let $\boldsymbol{e}$ be a unit view point shift. Then for all $\delta \in \mathbb{R}$,*

$$\begin{aligned} &\varphi^{-}_{\boldsymbol{e},\boldsymbol{p}}(d + \delta) = \varphi^{+}_{\boldsymbol{e},\boldsymbol{p}}(d - \delta), \\ \text{where} \quad &\varphi^{-}_{\boldsymbol{e},\boldsymbol{p}}(\alpha) = \int_{-\infty}^{0} L(\boldsymbol{p} + \alpha s \boldsymbol{e}, s\boldsymbol{e})\, \mathrm{d}s \\ &\varphi^{+}_{\boldsymbol{e},\boldsymbol{p}}(\alpha) = \int_{0}^{\infty} L(\boldsymbol{p} + \alpha s \boldsymbol{e}, s\boldsymbol{e})\, \mathrm{d}s \end{aligned} \tag{3}$$

are partial focal stacks integrated only in direction $\boldsymbol{e}$.

Proof. We assume that the scene is (locally) parallel to the image plane and Lambertian with ground truth disparity d. We thus get for any view point $\boldsymbol{v}$

$$L(\boldsymbol{p} + (d \pm \delta)\boldsymbol{v}, \boldsymbol{v}) = L(\boldsymbol{p} \pm \delta \boldsymbol{v}, \boldsymbol{v}_c), \tag{4}$$

since view $\boldsymbol{v}$ is the same as the reference view $\boldsymbol{v}_c = (0, 0)$ shifted by d. The integrals from (3) thus take the form

$$\begin{aligned} \varphi^{-}_{\boldsymbol{e},\boldsymbol{p}}(d + \delta) &= \int_{-\infty}^{0} L(\boldsymbol{p} + \delta s \boldsymbol{e}, \boldsymbol{v}_c)\, \mathrm{d}s, \\ \varphi^{+}_{\boldsymbol{e},\boldsymbol{p}}(d - \delta) &= \int_{0}^{\infty} L(\boldsymbol{p} - \delta s \boldsymbol{e}, \boldsymbol{v}_c)\, \mathrm{d}s. \end{aligned} \tag{5}$$

Since $\int_{-\infty}^{0} f(x)\, \mathrm{d}x = \int_{0}^{\infty} f(-x)\, \mathrm{d}x$ for any real-valued function f, we get

$$\begin{aligned} \varphi^{-}_{\boldsymbol{e},\boldsymbol{p}}(d + \delta) &= \int_{-\infty}^{0} L(\boldsymbol{p} + \delta s \boldsymbol{e}, \boldsymbol{v}_c)\, \mathrm{d}s \\ &= \int_{0}^{\infty} L(\boldsymbol{p} + \delta s(-\boldsymbol{e}), \boldsymbol{v}_c)\, \mathrm{d}s \\ &= \int_{0}^{\infty} L(\boldsymbol{p} - \delta s \boldsymbol{e}, \boldsymbol{v}_c)\, \mathrm{d}s \\ &= \varphi^{+}_{\boldsymbol{e},\boldsymbol{p}}(d - \delta). \end{aligned} \tag{6}$$

This completes the proof. □

(a) slice through focal stack φ of Lin et al. [18]

(b) slice through our partial focal stack φ^+

(c) slice through our partial focal stack φ^-

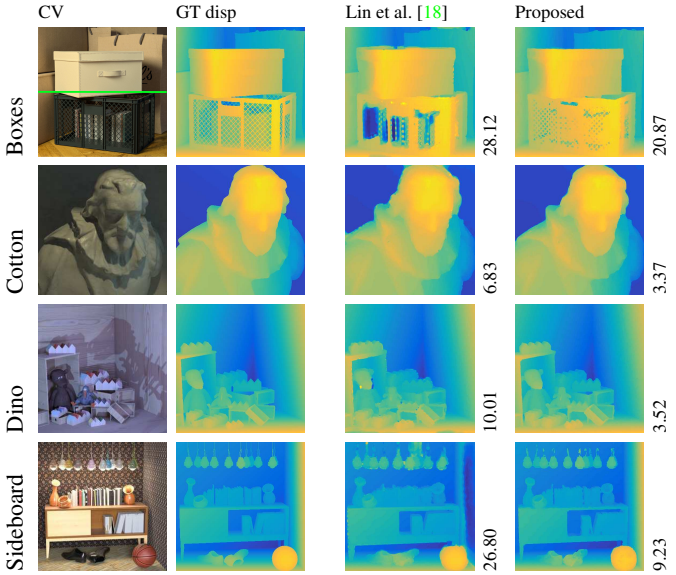

Figure 3. *Left:* comparison of Lin *et al.*'s [18] focal stack φ (a) with our versions φ^+ (b) and φ^- (c) for the green scanline of the light field *boxes* to the right. One can clearly see that our focal stacks do provide sharper edges near occlusion boundaries while still being pairwise symmetric around the true disparity d. *Right:* comparison of disparity maps obtained from Lin *et al.*'s [18] focal stack cost and our proposed cost volume. The numbers show the percentage of Pixels that deviate more than 0.07 from the ground truth.

Taking into account the proposition, we modify the cost function (2),

$$s^{\varphi}_{\boldsymbol{p}}(\alpha) = \int_0^{\delta_{\max}} \min\big(\rho(\varphi^-_{(1,0),\boldsymbol{p}}(\alpha+\delta) - \varphi^+_{(1,0),\boldsymbol{p}}(\alpha-\delta)), \rho(\varphi^-_{(0,1),\boldsymbol{p}}(\alpha+\delta) - \varphi^+_{(0,1),\boldsymbol{p}}(\alpha-\delta))\big)\,\mathrm{d}\delta \quad (7)$$

where ρ is the same robust distance function as defined below equation (2).

Note that we create four partial focal stacks corresponding to a crosshair of views around the center view. In future work, we plan to exploit symmetry in other directions to make the method more rotation-invariant. Assuming occlusion occurs only in one direction, i.e. occluders are not too thin, it is always guaranteed that focal stack regions unaffected by occlusion are compared to each other, and lead to zero (or at least very low) cost if α is the correct disparity. In our experiments, we set $\sigma = 1$ for cost computation and $\delta_{\max}$ to one-fifth of the disparity range.

4. Joint depth and normal map optimization

The result from finding a globally optimal solution for the cost with total variation prior is essentially locally flat, even if one uses sublabel relaxation [19], see figure 6. The resulting normal field is not useful for interesting applications such as intrinsic decomposition of the light field. Unfortunately, only priors of a restricted form are allowed if one wants to achieve the global optimum.

Thus, we propose to post-process the result by minimizing a second functional. The key requirements are that we still want to be faithful to the original data term, and at the same time obtain a piecewise smooth normal field. We achieve this by optimizing over depth and normals simultaneously, and linearly coupling them using the ideas in [7], which we describe first.

Relation between depth and normals. In [7], it was shown that if depth is reparametrized in a new variable $\zeta := \frac{1}{2}z^2$, the linear operator N given by

$$N(\zeta) = \begin{bmatrix} -\frac{\zeta_x}{f} \\ -\frac{\zeta_y}{f} \\ \frac{\hat{x}\zeta_x}{f} + \frac{\hat{y}\zeta_y}{f} \frac{2\zeta}{f^2} \end{bmatrix} \quad (8)$$

maps a depth map ζ to the map of corresponding normals scaled with the local area element of the parametrized surface. Above, f is the focal length, i.e. distance between Ω and Π, and $(\hat{x}, \hat{y})$ the homogenous coordinate of the pixel where the normal is computed, in particular, N is spatially varying. ζ_x and ζ_y denote partial derivatives.

The authors of [7] leveraged this map to introduce a minimal surface regularizer by encouraging small $\|N\zeta\|$. However, we want to impose smoothness of the field of *unit length* normals. It thus becomes necessary to introduce an unknown point-wise scaling factor α to relate $N\zeta$ and $\boldsymbol{n}$, which will converge to the area element.

The final prior on normal maps. Finally, we do not only want the normals to be correctly related to depth, but also to be piecewise smooth. Thus, the functional for depth reparametrized in ζ and unit length normals $\boldsymbol{n}$ we optimize

is

$$E(\zeta, \boldsymbol{n}) = \min_{\alpha>0} \int_\Omega \rho(\zeta, x) + \lambda \left\| N\zeta - \alpha \boldsymbol{n} \right\|_2 \mathrm{d}x + R(\boldsymbol{n}). \tag{9}$$

Above, $\rho(\zeta, x)$ is the reparametrized cost function, and $R(\boldsymbol{n})$ a convex regularizer of the normal map. To obtain a state-of-the-art framework, we extend the total generalized variation [3, 23] to vector-valued functions $\boldsymbol{n} : \Omega \to \mathbb{R}^m$ by defining

$$R(\boldsymbol{n}) = \sup_{\boldsymbol{w} \in C_c^1(\Omega, \mathbb{R}^{n\times m})} \int_\Omega \alpha \left\| \boldsymbol{w} - D\boldsymbol{n} \right\| + \gamma g \left\| D\boldsymbol{w} \right\|_F \mathrm{d}x. \tag{10}$$

The constants $\alpha, \gamma > 0$ defining amount of smoothing are user-provided, while $g := \exp(-c \left\| \nabla I \right\|)$ is a point-wise weight adapting the regularizer to image edges in the reference view I. Intuitively, we encourage $D\boldsymbol{n}$ to be close to a matrix-valued function $\boldsymbol{w}$ which has itself a sparse derivative, so $\boldsymbol{n}$ is encouraged to be piecewise affine.

Optimization. The functional E in (9) is overall non-convex, due to the multiplicative coupling of α and $\boldsymbol{n}$, and the non-convexity of ρ. We therefore follow an iterative approach and optimize for ζ and $\boldsymbol{n}$ in turn, initializing ζ_0 with the solution from sublabel relaxation [19] of (7) and $\boldsymbol{n}_0 = N\zeta_0$. Note that we could just as well embed (9) in a coarse-to-fine framework similar to the implementation [7] to make it computationally more efficient, but decided to evaluate the likely more accurate initialization from global optimization in this work. We now show to perform the optimization for the individual variables. Note that we will provide source code for the complete framework after our work has been published, so we will omit most of the technical details and just give a quick tour.

Optimization for depth. We remove from (9) the terms which do not depend on ζ, replace the norms by their second convex conjugates, and linearize ρ around the current depth estimate ζ_0. This way, we find that we have to solve the saddle point problem

$$\min_{\zeta, \alpha>0} \max_{\left\|\boldsymbol{p}\right\|_2 \le \lambda, |\xi| \le 1} \Big\{ (\boldsymbol{p}, N\zeta - \alpha\boldsymbol{n}) + (\xi, \rho|_{\zeta_0} + (\zeta - \zeta_0)\partial_\zeta \rho|_{\zeta_0}) \Big\}. \tag{11}$$

The solver we employ is the pre-conditioned primal-dual algorithm in [4]. Note that the functional (11) intuitively makes sense: it tries to maintain small residual cost ρ from focal stack symmetry, while at the same time adjusting the surface ζ so that $N\zeta$ becomes closer to the current estimate $\boldsymbol{n}$ for the smooth normal field scaled by α. As $N\zeta$ is the area-weighted Gauss map of the surface, α will converge to the local area element.

Optimization for the normal map. This time, we remove from (9) the terms which do not depend on $\boldsymbol{n}$. As α should be at the optimum equal to $\left\|N\zeta\right\|$, which is now known explicitly, we set $\boldsymbol{w} := N\zeta / \left\|N\zeta\right\|$ and end up with the L^1 denoising problem

$$\min_{\left\|\boldsymbol{n}\right\|=1} \int_\Omega \lambda \left\|N\zeta\right\| \left\|\boldsymbol{w} - \boldsymbol{n}\right\| \mathrm{d}x + R(\boldsymbol{n}). \tag{12}$$

The difficulty is the constraint $\left\|\boldsymbol{n}\right\| = 1$, which makes the problem non-convex. We therefore adopt the relaxation ideas in [31], which solves for the coefficients of the normals in a local parametrization of tangent space around the current solution, thus effectively linearizing the constraint. For details, we refer to [31]. Note that we use a different regularizer, image-driven TGV instead of vectorial total variation, which requires more variables [23]. Regardless, we obtain a sequence of saddle point problems with iteratively updated linearization points, which we again solve with [4].

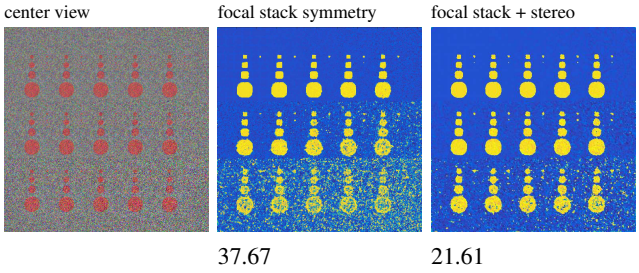

Figure 4. Averaging the proposed focal stack cost volume with a stereo correspondence cost volume using confidence scores from [25] as weights, we can significantly increase the resilience against noise. Numbers show *BadPix(0.07)*.

5. Results

We evaluate our algorithm on the recent benchmark [12] tailored to light field disparity map estimation. The given ground truth disparity is sufficiently accurate to also compute ground truth normal maps using the operator N from the previous section without visible discretization artifacts except at discontinuities.

Benchmark performance. Our submission results with several performance metrics evaluated can be observed on the benchmark web page[1] under the acronym `OFSY_330/DNR`. Note that all our parameters were tuned on the four training scenes to achieve an optimum *BadPix(0.07)* score, *i.e.* the percentage of pixels where disparity deviates by less than 0.07 pixels from the ground truth. The *stratified* scenes were not taken into account for parameter tuning as they are too artificial, but have of course been evaluated together with the *test* scenes. In accordance with the benchmark requirements, parameters are exactly the same for all scenes.

At the time of submission we rank first in *BadPix(0.07)*, with a solid first place on all test and training datasets. We

[1] http://www.lightfield-analysis.net

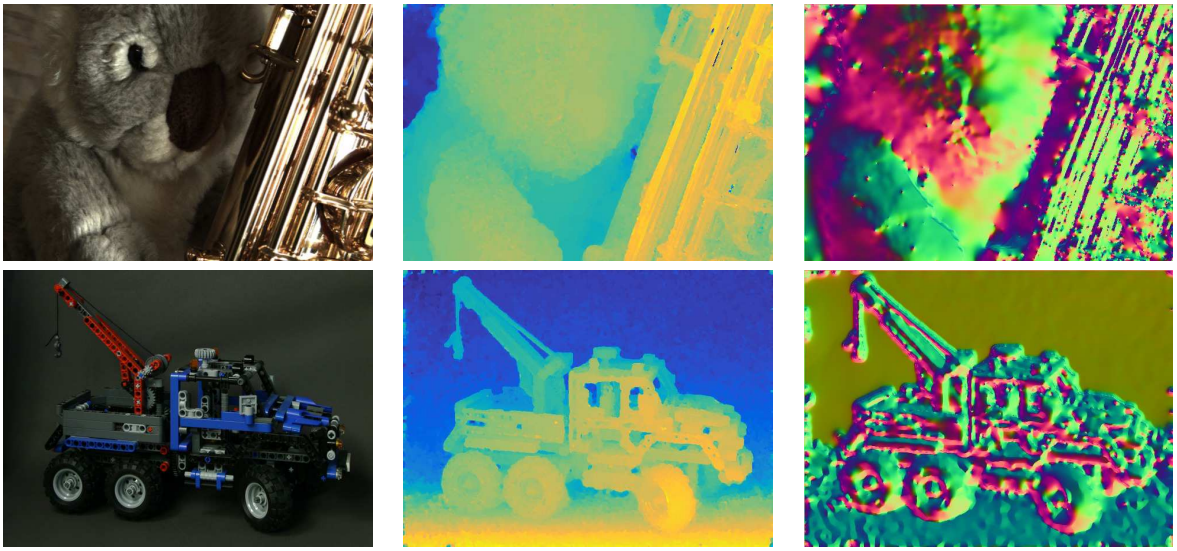

Figure 5. One further result on a challenging light field captured with the Lytro Illum plenoptic camera (top). The *truck* lightfield (bottom) is from the Stanford Light Field Archive [27] and was captured with a gantry. This last example demonstrates that our method is also able to handle larger disparity ranges.

also rank first place in the *Bumpiness* score for planes and continuous surfaces. This demonstrates the success of our proposed joint depth and normal prior to achieve smooth geometry. Finally, we rank first place on the *Discontinuities* score, which demonstrates the superior performance of the proposed occlusion-aware focal stack symmetry at occlusion boundaries. An overview of the results and a comparison for *BadPix(0.07)* on the training scenes can be observed in figure 7, for details and more evaluation criteria, we refer to the benchmark web page.

The single outlier is the performance on the stratified light field *dots*, which can be considered as a failure case of our method. This light field exhibits a substantial amount of noise in the lower right regions, see figure 4, and our method does not produce satisfactory results. A way to remedy this problem is to proceed like [28], and mix the focal stack cost volume with a stereo correspondence cost volume to increase resilience against noise. In contrast to their approach, we computed the mixture weights using the confidence measures in Tao *et al*. [25], and were able to drastically increase the performance on *dots* this way. For the benchmark evaluation, however, we decided to submit the pure version of our method with focal stack symmetry cost only. However, on real-world Lytro data we evaluate later on, combining costs turns out to be very beneficial.

Comparison to focal stack symmetry. In addition to the detailed benchmark above, we also compare our occlusion-aware cost volume to original focal stack symmetry [18], which has not yet been submitted to the benchmark. For this, we compute cost volumes using both [18] as well as the method proposed in section 3, and compute a globally optimal disparity map for both methods using sublabel relaxation, with the smoothness parameter tuned individually to achieve an optimal result. Results can be observed in figure 3. We perform significantly better in terms of error metrics, easily visible from the behaviour at occlusion boundaries, in particular regions which exhibit occlusions with very fine detail.

Please note that these images do not show the results obtained by Lin *et al*.'s final algorithm in [18], but only those achieved by simply using the focal stack symmetry cost function (2), since we aim at a pure comparison of the two cost volumes.

Normal map accuracy. In another run of experiments, we verify that our joint scheme for depth and normal map regularization is an improvement over either individual regularization of just the depth map with the minimal surface regularizer in [7], as well as just regularization of the normal map using [31]. Results can be observed in figure 6. Normal map regularization is non-convex and fails to converge towards the correct solution starting with the piecewise fronto-parallel intitialization from sublabel relaxation [19]. The better result is achieved by smoothing the depth map directly, but when imposing the same amount of smoothing as in our framework, it performs worse smoothing on planar surfaces and fails to preserve details.

Real-world results. To test the real-world performance of our proposed method, we evaluate on light fields captured with the Lytro Illum plenoptic camera. One can clearly see that our method performs relatively well even on non-Lambertian objects like the origami crane in figure 1 or the saxophone in figure 5 (top row). Figure 5 (bottom row) shows results on the lego truck from the Stanford light field archive for comparison and proof that our method works even with a relatively large baseline.

Figure 6. Comparison of normal maps obtained with different methods. Numbers show mean angular error in degrees. The result obtained from sublabel relaxation is overall still fronto-parallel despite the large number of 330 labels. In consequence, the non-convex normal smoothing with [31] fails to converge to a useful solution, as the initialization is too far from the optimum. Smoothing the depth map using [7] yields visually similar results than our method, but we achieve lower errors and smoother surfaces while still preserving details like the eyes of the teddy in *dino*. For visualization, normals $\boldsymbol{n}$ have been transformed to RGB space via $(r, g, b) = \frac{1}{2}(\boldsymbol{n} + [1\ 1\ 1]^T)$.

6. Conclusion

We have presented occlusion aware focal-stack symmetry as a way to compute disparity cost volumes. The key assumptions are Lambertian scenes and slowly varying disparity within surfaces. Experiments show that our proposed data term is to this date the most accurate light field depth estimation approach on the recent benchmark [12]. It performs particularly well on occlusion boundaries and in terms of overall correctness of the disparity estimate. As a small drawback, we get a slightly reduced noise resiliency, as we operate only on a crosshair of views around the reference view as opposed to the full light field. On very noisy scenes, we can however improve the situation by confidence-based integration of stereo correspondence costs into the data term, as suggested in previous literature on disparity estimation with focus cues [28, 25].

With additional post-processing using joint depth and normal map regularization, we can further increase accuracy slightly, but in particular obtain accurate and smooth normal fields which preserve small details in the scene. We again outperform previous methods on the benchmark datasets. Further experiments on real-world scenes show that we can deal with significant amounts of specularity, and obtain depth and normal estimates suitable for challenging applications like intrinsic light field decomposition [1].

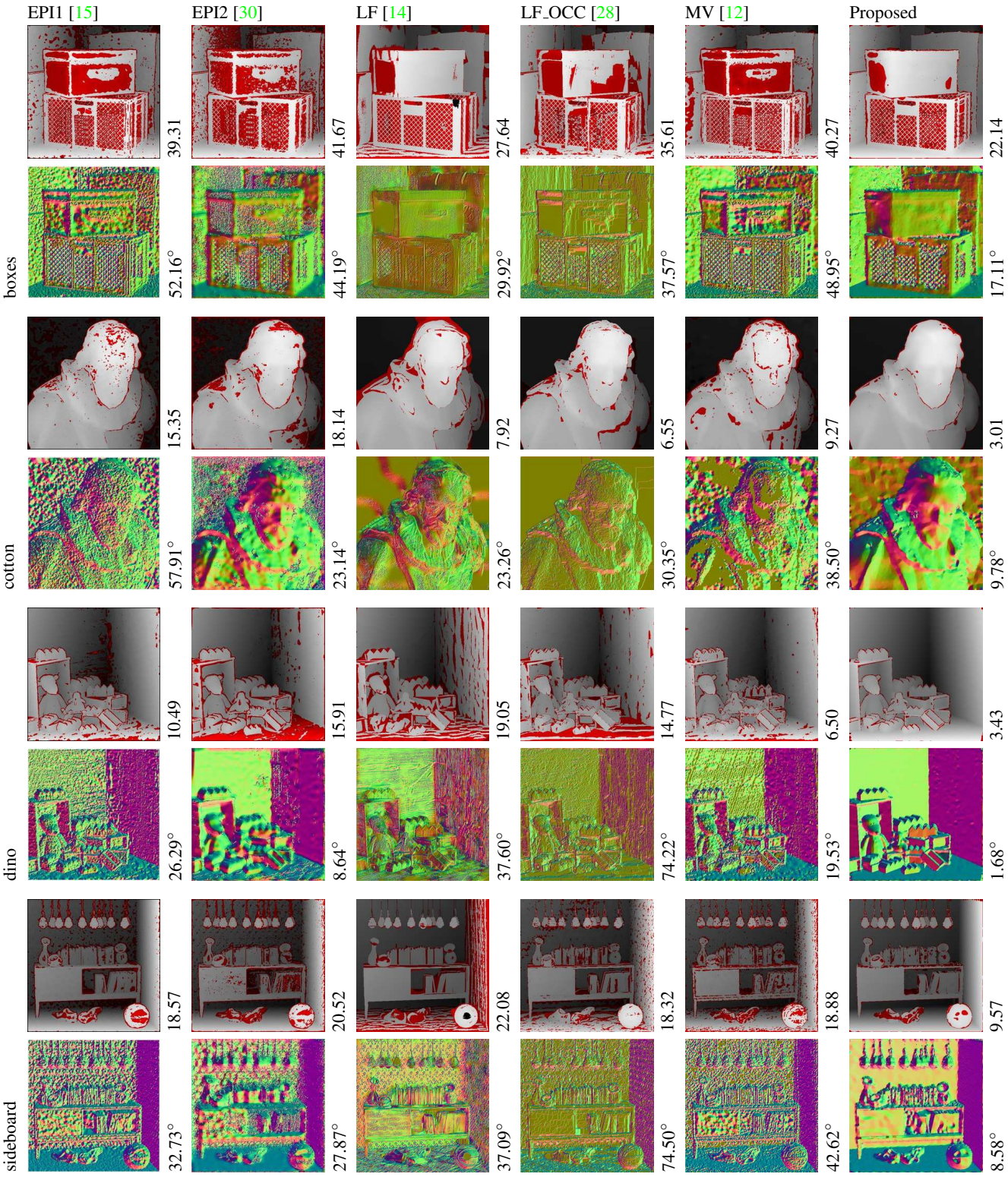

Figure 7. Our disparity and normal maps for the training datasets compared to the results of the other methods listed on the benchmark at the time of submission. For all datasets, we achieve the lowest error for both, measured in percentage of pixels with a disparity error larger than 0.07 pixels (marked red in the disparity map), and mean angular error in degrees for the normal map, respectively. For the full benchmark evaluation with several other accuracy metrics, see http://lightfield-analysis.net, where our method is listed as OFSY_330/DNR.

References

[1] A. Alperovich and B. Goldluecke. A variational model for intrinsic light field decomposition. In *Asian Conf. on Computer Vision*, 2016. 1, 7

[2] R. Bolles, H. Baker, and D. Marimont. Epipolar-plane image analysis: An approach to determining structure from motion. *International Journal of Computer Vision*, 1(1):7–55, 1987. 2

[3] K. Bredies, K. Kunisch, and T. Pock. Total generalized variation. *SIAM Journal on Imaging Sciences*, 3(3):492–526, 2010. 2, 5

[4] A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *J. Math. Imaging Vis.*, 40(1):120–145, 2011. 1, 5

[5] C. Chen, H. Lin, Z. Yu, S.-B. Kang, and Y. J. Light field stereo matching using bilateral statistics of surface cameras. In *Proc. International Conference on Computer Vision and Pattern Recognition*, 2014. 1, 2

[6] 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*, 97(1):51–85, 2005. 2

[7] G. Graber, J. Balzer, S. Soatto, and T. Pock. Efficient minimal-surface regularization of perspective depth maps in variational stereo. In *Proc. International Conference on Computer Vision and Pattern Recognition*, 2015. 1, 2, 4, 5, 6, 7

[8] S. Heber and T. Pock. Shape from light field meets robust PCA. In *Proc. European Conference on Computer Vision*, 2014. 2

[9] S. Heber and T. Pock. Convolutional networks for shape from light field. In *Proc. International Conference on Computer Vision and Pattern Recognition*, 2016. 2

[10] S. Heber, R. Ranftl, and T. Pock. Variational shape from light field. In *Int. Conf. on Energy Minimization Methods for Computer Vision and Pattern Recognition*, pages 66–79, 2013. 2

[11] H. Hirschmuller. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2):328–341, 2008. 2

[12] K. Honauer, O. Johannsen, D. Kondermann, and B. Goldluecke. A dataset and evaluation methodology for depth estimation on 4d light fields. In *Asian Conf. on Computer Vision*, 2016. 1, 5, 7, 8

[13] A. Jarabo, B. Masia, A. Bousseau, F. Pellacini, and D. Gutierrez. How do people edit light fields? *ACM Transactions on Graphics (Proc. SIGGRAPH)*, 33(4), 2014. 1

[14] H. Jeon, J. Park, G. Choe, J. Park, Y. Bok, Y. Tai, and I. Kweon. Accurate depth map estimation from a lenslet light field camera. In *Proc. International Conference on Computer Vision and Pattern Recognition*, 2015. 1, 2, 8

[15] O. Johannsen, A. Sulc, and B. Goldluecke. What sparse light field coding reveals about scene structure. In *Proc. International Conference on Computer Vision and Pattern Recognition*, 2016. 1, 2, 8

[16] C. Kim, H. Zimmer, Y. Pritch, A. Sorkine-Hornung, and M. Gross. Scene reconstruction from high spatio-angular resolution light fields. *ACM Transactions on Graphics (Proc. SIGGRAPH)*, 32(4), 2013. 1, 2

[17] V. Kolmogorov and R. Zabih. Multi-camera Scene Reconstruction via Graph Cuts. In *Proc. European Conference on Computer Vision*, pages 82–96, 2002. 2

[18] H. Lin, C. Chen, S.-B. Kang, and J. Yu. Depth recovery from light field using focal stack symmetry. In *Proc. International Conference on Computer Vision*, 2015. 1, 2, 3, 4, 6

[19] T. Moellenhoff, E. Laude, M. Moeller, J. Lellmann, and D. Cremers. Sublabel-accurate relaxation of nonconvex energies. In *Proc. International Conference on Computer Vision and Pattern Recognition*, 2016. 1, 2, 4, 5, 6, 7

[20] S. Nayar and Y. Nakagawa. Shape from Focus. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(8):824–831, 1994. 2

[21] T. Pock, D. Cremers, H. Bischof, and A. Chambolle. Global Solutions of Variational Models with Convex Regularization. *SIAM Journal on Imaging Sciences*, 2010. 2

[22] S. Pujades, B. Goldluecke, and F. Devernay. Bayesian view synthesis and image-based rendering principles. In *Proc. International Conference on Computer Vision and Pattern Recognition*, 2014. 1

[23] R. Ranftl, S. Gehrig, T. Pock, and H. Bischof. Pushing the limits of stereo using variational stereo estimation. In *IEEE Intelligent Vehicles Symposium*, 2012. 5

[24] M. Tao, S. Hadap, J. Malik, and R. Ramamoorthi. Depth from combining defocus and correspondence using light-field cameras. In *Proc. International Conference on Computer Vision*, 2013. 2

[25] M. Tao, P. Srinivasan, S. Hadap, S. Rusinkiewicz, J. Malik, and R. Ramamoorthi. Shape estimation from shading, defocus, and correspondence using light-field angular coherence. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2016. 5, 6, 7

[26] I. Tosic and K. Berkner. Light field scale-depth space transform for dense depth estimation. In *Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 441–448, 2014. 2

[27] V. Vaish and A. Adams. The (New) Stanford Light Field Archive. `http://lightfield.stanford.edu`, 2008. 6

[28] T. Wang, A. Efros, and R. Ramamoorthi. Occlusion-aware depth estimation using light-field cameras. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3487–3495, 2015. 1, 2, 6, 7, 8

[29] T. C. Wang, M. Chandraker, A. Efros, and R. Ramamoorthi. SVBRDF-invariant shape and reflectance estimation from light-field cameras. In *Proc. International Conference on Computer Vision and Pattern Recognition*, 2016. 1

[30] S. Wanner and B. Goldluecke. Variational light field analysis for disparity estimation and super-resolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(3):606–619, 2014. 1, 2, 8

[31] B. Zeisl, C. Zach, and M. Pollefeys. Variational regularization and fusion of surface normal maps. In *Proc. International Conference on 3D Vision (3DV)*, 2014. 1, 5, 6, 7