

# A Dataset and Evaluation Methodology for Depth Estimation on 4D Light Fields

Katrin Honauer<sup>1</sup>(✉), Ole Johannsen<sup>2</sup>,  
Daniel Kondermann<sup>1</sup>, and Bastian Goldluecke<sup>2</sup>

<sup>1</sup> HCI, Heidelberg University, Heidelberg, Germany  
{katrin.honauer,daniel.kondermann}@iwr.uni-heidelberg.de

<sup>2</sup> University of Konstanz, Konstanz, Germany  
{ole.johannsen,bastian.goldluecke}@uni-konstanz.de

**Abstract.** In computer vision communities such as stereo, optical flow, or visual tracking, commonly accepted and widely used benchmarks have enabled objective comparison and boosted scientific progress.

In the emergent light field community, a comparable benchmark and evaluation methodology is still missing. The performance of newly proposed methods is often demonstrated qualitatively on a handful of images, making quantitative comparison and targeted progress very difficult. To overcome these difficulties, we propose a novel light field benchmark. We provide 24 carefully designed synthetic, densely sampled 4D light fields with highly accurate disparity ground truth. We thoroughly evaluate four state-of-the-art light field algorithms and one multi-view stereo algorithm using existing and novel error measures.

This consolidated state-of-the-art may serve as a baseline to stimulate and guide further scientific progress. We publish the benchmark website <http://www.lightfield-analysis.net>, an evaluation toolkit, and our rendering setup to encourage submissions of both algorithms and further datasets.

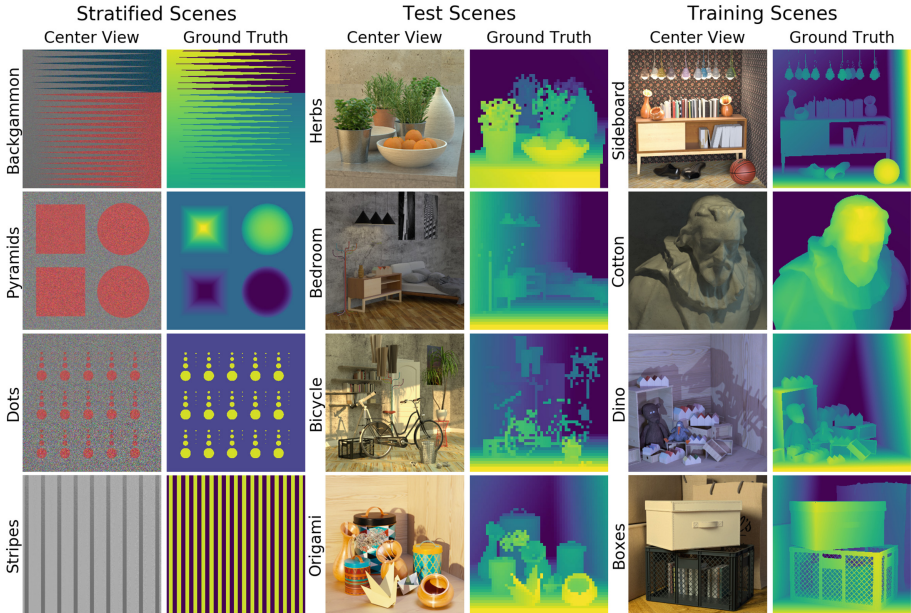
## 1 Introduction

Over the last decade, light field analysis has grown from a niche topic to an established part of the computer vision community. While in its most general form, the light field captures the radiance distribution for every ray passing through every point in space-time [1], one usually simplifies this to a 4D parametrization, where one essentially considers a dense collection of pinhole views with parallel optical axes, sampled on a rectangular grid in a 2D plane. The key difference to the classical multi-view scenario is the dense and regular sampling, which allows to develop novel and highly accurate methods for depth reconstruction [2–6], which can correctly take occlusions into account to recover fine details [7].

---

K. Honauer and O. Johannsen contributed equally.

**Electronic supplementary material** The online version of this chapter (doi:[10.1007/978-3-319-54187-7\\_2](https://doi.org/10.1007/978-3-319-54187-7_2)) contains supplementary material, which is available to authorized users.



**Fig. 1.** We present a new light field benchmark consisting of four stratified, four test, and four training scenes. The stratified scenes are designed to pose specific, isolated challenges with spatially increasing difficulty. To warrant a deep and comprehensive understanding of algorithm performance, we quantify a variety of characteristics such as foreground fattening, texture sensitivity, and robustness to noise. For the stratified and training scenes, we provide high resolution ground truth disparity maps, normal maps and 3D depth point clouds. The same information is provided for twelve additional scenes (see Fig. 2).

In more mature vision communities such as the stereo or tracking community, standard benchmarks of sufficient variety and difficulty have proven their fundamental importance for targeted development and objective judgment of the overall progress in the respective field. Detailed evaluations and comparisons of the precise strengths and weaknesses of different methods are guiding research and thus stimulating progress. However, such a common benchmark is currently lacking in the light field community. For this reason, recent papers often resort to showing qualitative results on real-world datasets to showcase their improved results [2, 6, 7], but performance is very difficult to judge without ground truth. In those cases where a numeric evaluation is performed, the specific ground truth data set and/or quality metrics often vary wildly between papers, again making objective comparison hard [2–7]. Moreover, parameters might be fine-tuned towards a certain quality metric, e.g. more smoothing in general improves the mean squared error at the expense of per-pixel accuracy. Finally, there is currently no benchmarking website which offers the opportunity of a common gathering point for datasets and online performance comparison.

**Contributions.** The light field benchmark we present in this paper is designed to remedy the aforementioned shortcomings. In this first iteration, we focus solely on the problem of depth estimation for Lambertian scenes, although we provide some scenes with specular reflections to offer more of a challenge. Our main contributions can be summarized as follows:

- We introduce a new synthetic dataset with 24 carefully designed scenes, which overcomes technical shortcomings of previous datasets.
- We propose novel error measures and evaluation modalities enabling comprehensive and detailed characterizations of algorithm results.
- We present an initial performance analysis of four state-of-the-art light field algorithms and one multi-view stereo algorithm.
- We publish a benchmarking website and an evaluation toolkit to provide researchers with the necessary tools to facilitate algorithm evaluation.

We consider this benchmark as a first step towards a joint effort of the light field community to develop a commonly accepted benchmark suite. All researchers in the field are kindly invited to contribute existing and future algorithms, datasets, and evaluation measures.

## 2 Related Work

**Existing Light Field Datasets.** The available datasets can be grouped into synthetic light fields, real world light fields captured with a plenoptic camera (usually a Lytro Illum) and real world scenes captured with a camera array or gantry. We are aware of multiple smaller and larger collections, in particular the Stanford Light Field Archive [8], the Synthetic Light Field Archive [9], a collection of Lytro images [10], the 3D High-Resolution Disney Dataset [11], and the New Light Field Image Dataset [12]. All these datasets have in common that no ground truth data is available, making them hard to use for precise benchmarking.

To our knowledge, the only collection of light fields which comes with ground truth depth and an open benchmark is the HCI Light Field Benchmark by Wanner et al. [13]. They provide synthetic as well as real world 4D light fields including ground truth. In the past, this benchmark stimulated the growth of multiple light field algorithms, but it now reaches a point where we think it can no longer satisfy the needs of the light field community. This is due to three major drawbacks. First, their ground truth gives around 130 distinct depth labels, yielding a maximum evaluation accuracy which is already surpassed by state-of-the-art algorithms. Second, the ground truth data contains errors in the form of wrong pixels, as well as inaccuracies at occlusion boundaries, which are a key part of depth accuracy evaluation. Third, due to the way the light fields were rendered, a systematic noise pattern is present that is the same for all views.

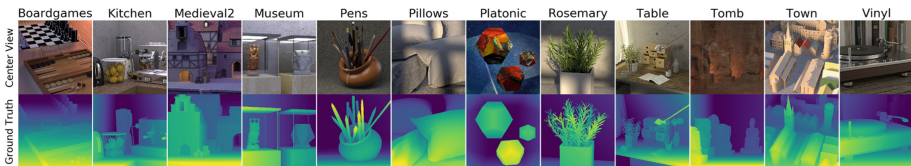
**Insights from Other Popular Benchmarks.** In more mature communities such as stereo, optical flow, and visual tracking, benchmarks play a fundamental

role in boosting scientific progress: they help consolidate existing research [14] and guide the community towards open challenges (e.g. [15] for large motion optical flow, [16] for automotive stereo). Building upon the experience from these successful benchmarks, we identify three key insights for the design of our benchmark. From the Visual Object Tracking Community<sup>1</sup>, we conclude that scientific progress thrives if benchmarks are seen as a joint effort by and service for the community [17, 18]. We therefore encourage researchers to not only contribute algorithms but also datasets and evaluation methods. Second, there is no single best algorithm: algorithms have different strengths and weaknesses and may be used for applications with very different requirements. We therefore use multi-dimensional performance evaluation with carefully designed metrics to reflect this diversity [19–21]. Third, as methods improve, benchmarks may no longer be suitable to differentiate algorithm performance. They may even hinder scientific progress by unintentionally stimulating overfitting. We hence designed our benchmark for a limited lifespan, focusing on those challenges where current algorithms struggle with. Together with the community, similar to [19, 22], we plan to regularly maintain the dataset and add new scenes when necessary.

### 3 Considerations on Benchmark Design

Light field and multi-view stereo algorithms find more and more applications in real world challenges such as the movie set reconstruction and industrial optical inspection. Often, medical or automotive technologies are even safety-relevant. Designing a useful benchmark requires addressing the following four aspects:

(1) *Benchmark Purpose.* Test datasets should be compact to minimize dataset creation cost, maximize information gain, and reduce benchmarking efforts. Researchers across different fields of computer vision agree that a systematic, considerate compilation of imagery is necessary to allow for specific, meaningful algorithm evaluation [19, 23–26]. Unintended biases can occur in dataset creation, causing e.g. an overemphasis on smooth surfaces in the scenes [23]. Using top-down approaches such as requirements engineering [25] or bottom-up methods such as HAZOP studies [23] can alleviate this risk. As shown below, state of the art algorithms often struggle with geometry and texture challenges. Hence,



**Fig. 2.** We provide 12 additional scenes with ground truth. They are not part of the official benchmark but can be used for algorithm development and evaluation.

<sup>1</sup> <http://www.votchallenge.net/>.



we focus on five challenges, namely occlusion boundaries, fine structures, low texture, smooth surfaces and camera noise.

(2) *Scene Design.* Simple scenes focusing on a single challenge allow to decouple the performance analysis of each individual challenge [24]. Thus, we introduce four light fields with controlled parameters for a fixed challenge combination: Backgammon, Dots, Pyramids and Stripes (see Fig. 1). We call these scenes *stratified* since their goal is to create quantifiable challenges which can be used to re-weight performance metrics based on real-world, non-stratified data. To gradually increase each challenge, the scenes exhibit spatially increasing difficulties. This allows both for immediate visual inspection as well as quantitative comparison between algorithms. Yet, complex real-world scenes contain all the challenges in potentially statistically significant spatial combinations. We therefore create additional, photorealistically rendered scenes (see Fig. 1). This suppresses the chance of overfitting parameters to a certain challenge and helps to obtain an intuition on real-world performance.

(3) *Dataset Acquisition.* To date, no measurement technology exists to record real light fields with sufficiently accurate ground truth. Using computer vision algorithms to create ground truth for computer vision algorithms defeats the purpose of benchmarking. Recent research shows promising results that rendering can be a valid approach [15, 27, 28]. We therefore use rendered scenes, building on the advantages of near-perfect ground truth accuracy and the option to systematically vary scene parameters.

(4) *Benchmarking Methodology.* We adopt the approach of Scharstein et al. [19] and divide our dataset into test, training, and additional scenes (see Figs. 1 and 2). To be listed on the public benchmark table, we ask participants to submit their algorithm results and runtimes on the twelve scenes as depicted on Fig. 1. Participants may use the input data and disparity ranges as provided on the website. We further provide an evaluation toolkit which contains: (i) file IO methods for Matlab and Python (ii) a submission validation script (iii) evaluation code to compute and visualize the metrics on the stratified and training scenes. All metric scores and visualizations will be computed on our server and displayed publicly on the benchmark table. The ground truth of the training and stratified scenes may be used to optimize parameter settings. We do not publish ground truth for the four photorealistic test scenes. As in [19] algorithm results of the training scenes will be available for download in full resolution. The twelve additional scenes with full ground truth are not part of the benchmark. They are shared with the community to support further algorithm development. We refer to <http://www.lightfield-analysis.net> for technical submission details.

## 4 Description of Dataset and Metrics

In this section, we present the scenes and corresponding error metrics resulting from our theoretical considerations on scene content and performance analysis.

#### 4.1 Technical Dataset Details

The scenes were created with Blender [29] using the internal renderer for the stratified scenes and the Cycles renderer for the photorealistic scenes. We built the light field setup in a way such that all cameras are shifted towards a common focus plane while keeping the optical axes parallel. Thus, zero disparity does not correspond to infinite depth. Most scene content lies within a range of  $-1.5$  px and  $1.5$  px, though disparities on some scenes are up to 3 px.

For each scene, we provide 8 bit light fields ( $9 \times 9 \times 512 \times 512 \times 3$ ), camera parameters, and disparity ranges. For the stratified and training scenes we further provide evaluation masks and 16 bit ground truth disparity maps in two resolutions ( $512 \times 512$  px and  $5120 \times 5120$  px). We use the high resolution ground truth to accurately evaluate algorithm results at fine structures and depth discontinuities. The textures of the stratified scenes are generated from Gaussian noise to minimize potential unwanted interference of texture irregularities with the actual challenges in the scene. A detailed technical description of the data generation process and the source code of the Blender add-on can be found at <http://www.lightfield-analysis.net>.

#### 4.2 General Evaluation Measures

Algorithms often have different strengths and weaknesses, such as overall accuracy or sensitivity to fine structures, which may be prioritized very differently depending on the application. In the spirit of [21], we quantify a variety of characteristics to warrant a deep and comprehensive understanding of individual algorithm performance. We provide the commonly used MSE \* 100 and BadPix(0.07) metrics as well as Bumpiness and scene specific adaptations of these metrics. The adaptations are introduced together with the respective scenes. The general MSE, BadPix, and Bumpiness metrics are defined as follows:

Given an estimated disparity map  $d$ , the ground truth disparity map  $gt$  and an evaluation mask  $M$ , MSE is quantified as

$$\text{MSE}_{\mathcal{M}} = \frac{\sum_{x \in \mathcal{M}} (d(x) - gt(x))^2}{|\mathcal{M}|} * 100 \quad (1)$$

and BadPix is quantified as

$$\text{BadPix}_{\mathcal{M}}(t) = \frac{|\{x \in \mathcal{M} : |d(x) - gt(x)| > t\}|}{|\mathcal{M}|}. \quad (2)$$

To measure algorithm performance at smooth planar and curved surfaces we further define  $f = d - gt$  to quantify Bumpiness as

$$\text{Bumpiness} = \frac{\sum_{x \in \mathcal{M}} \min(0.05, \|H_f(x)\|_F)}{|\mathcal{M}|} * 100. \quad (3)$$

Hence, the bumpiness metric solely focuses on the smoothness of an estimation but does not assess misorientation or offset. These properties are covered by other metrics.

### 4.3 Scene Descriptions with Corresponding Evaluation Measures

**Backgammon.** This scene (see Fig. 1) is designed to assess the interplay of fine structures, occlusion boundaries and disparity differences. It consists of two parallel, slanted background planes and one foreground plane which is inversely slanted. The foreground plane is jagged to create increasingly thin foreground structures and increasingly fine background slits. On Backgammon, we quantify Foreground Fattening which is defined at occlusion boundaries on a mask  $M$  that only includes background pixels as

$$\text{FG\_Fattening} = \frac{|\{x \in \mathcal{M}: d(x) > h\}|}{|\mathcal{M}|}, \quad (4)$$

where  $h = (BG + FG)/2$ . Thus, Foreground Fattening calculates the fraction of pixels that are closer to the foreground than to the background and should have been estimated as background. Similarly, Foreground Thinning is defined on a mask  $M$  that only includes foreground pixels as

$$\text{FG\_Thinning} = \frac{|\{x \in \mathcal{M}: d(x) < h\}|}{|\mathcal{M}|}, \quad (5)$$

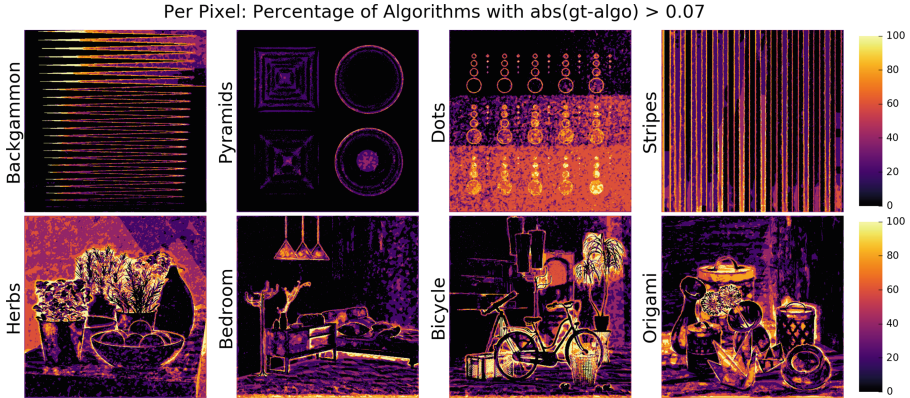
i.e. Foreground Thinning calculates the fraction of pixels that are closer to the background than to the foreground.

**Pyramids.** With this scene, we assess algorithm performance on convex versus concave as well as rounded versus planar geometry. The upper hemisphere and pyramid stick out of the middle plane whereas the lower hemisphere and pyramid are embedded in the plane. We quantify surface reconstruction quality by computing Bumpiness as defined in Eq. 3 on masks for the fronto-parallel plane and the slanted surfaces of the objects respectively.

**Dots.** This scene is designed to assess the effect of camera noise on the reconstruction of objects of varying size. The image features 15 identical grid cells. Each cell consists of 9 increasingly smaller coplanar circles. To approximate thermal and shot noise, we add Gaussian noise with variances growing linearly between 0.0 and 0.2 in row-major order. We quantify robustness against noise by computing the MSE on the background plane. We further quantify sensitivity to small geometries by computing the percentage of detected dots. A dot counts as detected if the majority of its local disparity estimates is distinguishable from the background by being in a BadPix range of 0.4 px to the ground truth dot.

**Stripes.** This scene is used to assess the influence of texture and contrast at occlusion boundaries. It consists of a fronto-parallel background plane and 16 coplanar stripes. The amount of texture on the background plane is gradually increasing from left to right. Likewise, the vertical stripes are increasingly textured from the bottom to the top of the image. The stripes feature alternating intensities with dark, high contrast stripes and bright, low contrast stripes.

To quantify performance, we define three types of image regions and compute BadPix(0.07) on each region individually: First, we use the no-occlusion areas on



**Fig. 3.** The heatmaps illustrate local scene difficulty. Per pixel, they show the percentage of algorithms with a disparity error  $> 0.07$  px. Algorithms struggle particularly with fine structures, noise, and occlusion areas.

the stripes and on the background for low texture evaluation. Second, we use the dark stripes and their occlusion areas to quantify performance at high contrast occlusion boundaries. Similarly, we use bright stripes and their occlusion areas to quantify performance at low contrast occlusion boundaries.

**Photorealistic Scenes.** We designed the photorealistic scenes to allow for performance evaluation on fine structures, complex occlusion areas, slanted planar surfaces, and continuous non-planar surfaces. The scenes contain various combinations of these challenges and allow to obtain an intuition of algorithm performance on real world scenes. For quantitative performance analysis, we use masks for different challenge regions. Apart from the overall MSE and BadPix(0.07) scores, we compute the BadPix(0.07) score at occlusion areas. We further quantify smoothness at planar and non-planar continuous surfaces by computing Bumpiness scores on the respective image areas. Furthermore, we compute Thinning (0.15) and Fattening ( $-0.15$ ) at fine structures by computing adjusted BadPix scores as follows:

$$\text{Thinning}_{\mathcal{M}}(t) = \frac{|\{x \in \mathcal{M}: gt(x) - d(x) > t\}|}{|\mathcal{M}|} \quad (6)$$

where  $\mathcal{M}$  is a mask for fine structure pixels and

$$\text{Fattening}_{\mathcal{M}}(t) = \frac{|\{x \in \mathcal{M}: gt(x) - d(x) < t\}|}{|\mathcal{M}|} \quad (7)$$

where  $\mathcal{M}$  is a mask for pixels surrounding fine structures.

## 5 Experimental Validation of Dataset and Metrics

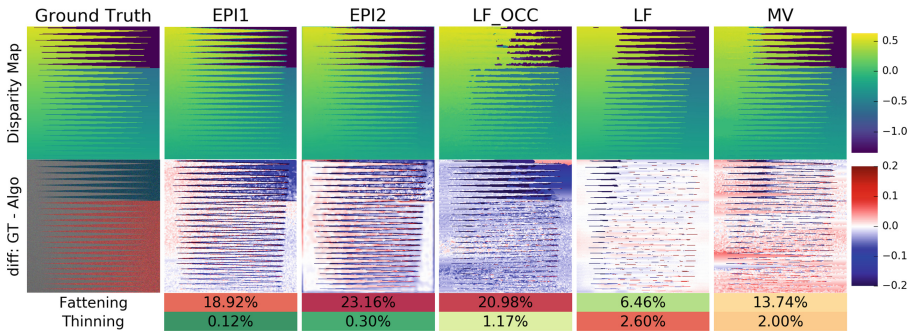
### 5.1 Evaluation of Scene Content

In order to verify our reasoning on challenging scene characteristics, we analyze local scene difficulty as shown in Fig. 3. Challenging regions on the heatmaps (bright) correlate with our intended challenges as described in Sect. 4. On the stratified scenes, the fine gaps on Backgammon, low texture areas on Stripes and noisy regions on Dots feature low algorithm performance. On the photorealistic scenes, complex occlusions on Herbs, fine structures on Bedroom, and fine structure grids on Bicycle represent the most challenging image regions. By contrast, the well-textured fronto-parallel surface of Pyramids, the noise-free area on Dots as well as smooth and high-texture regions on the photorealistic scenes feature good algorithm performance.

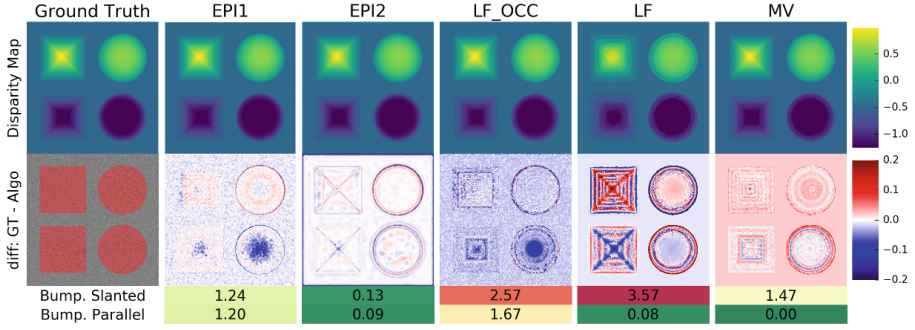
### 5.2 Evaluation of Performance Measures

In this section, we examine whether our metrics appropriately quantify algorithm performance on the stratified and photorealistic scenes. We show algorithm results of one multi-view algorithm (MV) and four light field algorithms (LF, LF\_OCC, EPI2, EPI1). In order to keep the focus on the metrics, we treat the algorithms as black boxes until Sect. 6. For additional results we refer to the supplemental material.

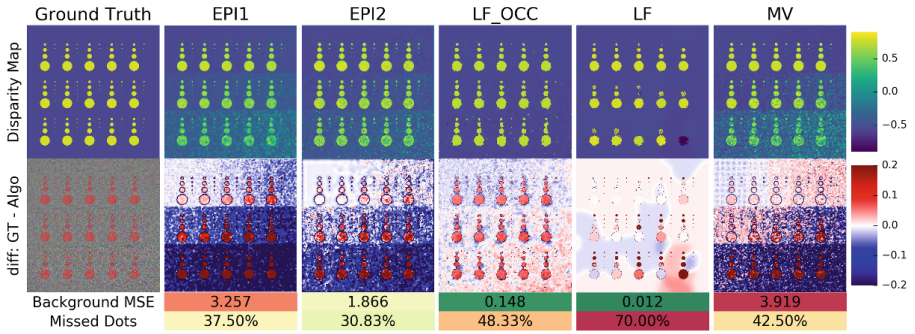
**Backgammon: Fine Structures, Thin Gaps and Occlusions.** The algorithm results on Fig. 4 show that algorithms do indeed struggle at gradually finer peaks and especially at thin gaps of the Backgammon scene. The depicted fattening and thinning scores quantify the respective algorithm performance appropriately. More fattening occurs at occlusion areas on the top part of the image,



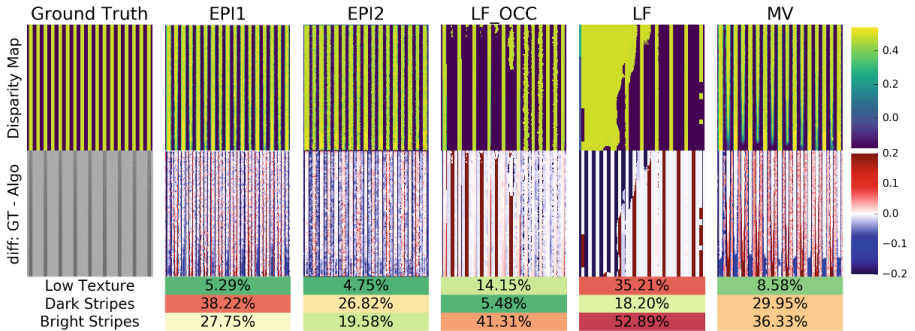
**Fig. 4.** All algorithms struggle with reconstructing the background depth of the narrow gaps on the left side of the image. LF\_OCC and EPI2 have the strongest fattening, for LF\_OCC it is concentrated between the upper bars and for EPI2 it is uniformly distributed around each bar.



**Fig. 5.** The bumpiness scores correctly reflect the observation that LF produces very smooth estimates on the fronto-parallel plane but heavily staircased estimates on the slanted object surfaces. On both types of surfaces LF\_OCC results are bumpy and EPI2 results are smooth.



**Fig. 6.** The performance of most algorithms degrades by increasing levels of noise. Robustness via strong regularization is traded for low sensitivity on the smaller dots (LF) and vice versa (EPI1).



**Fig. 7.** Algorithms struggle with the increasingly low texture towards the bottom of the image. As reflected by our metrics, LF\_OCC and LF handle dark, high contrast stripes much better than bright, low contrast stripes.



where disparity distances are large (see LF\_OCC and MV). In this area, background pixels which are visible from the center view are occluded in many other views. For very thin gaps, an epipolar line belonging to a background point might then be occluded at both ends.

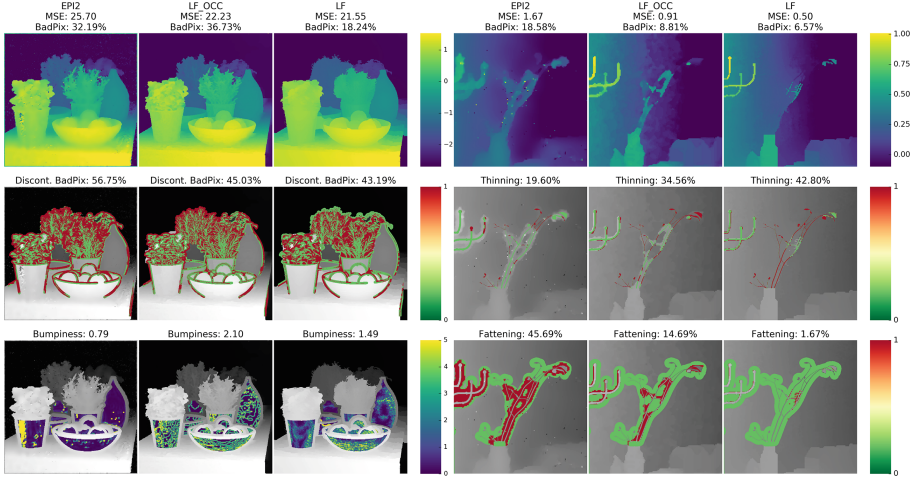
**Pyramids: Slanted and Convex vs. Concave Surfaces.** Algorithms face various difficulties on the Pyramids scene (see Fig. 5), such as systematic offset on the middle plane, bumpy surfaces and inaccurate object boundaries. The continuous disparity ranges of the slanted surfaces are particularly challenging for algorithms which estimate discrete disparity labels such as LF. The bias on the middle planes is also caused by a limited number of disparity steps where no step matches the disparity of the plane. The depicted bumpiness scores for slanted surfaces correctly identify smooth and staircased disparity maps.

**Dots: Noise and Tiny Objects.** Results on Fig. 6 show that algorithms struggle either with reconstructing small dots or with reconstructing smooth and accurate background planes. LF\_OCC and LF robustly yield accurate results on the background, whereas EPI1 and MV are strongly affected by artifacts due to noise. In contrast, LF applies strong regularization, causing poor scores for the number of reconstructed dots; EPI1 and EPI2 perform better. These effects show that the complementary metrics of this scene nicely challenge the algorithms to find a good trade-off between regularization and fine structure sensitivity.

**Stripes: Texture and Contrast at Occlusions.** Algorithms struggle with correctly computing disparities at the low contrast boundaries of the bright stripes and on the low texture regions towards the bottom of the image (see Fig. 7). Our metrics quantify that algorithms such as LF, which use image gradients as priors for their occlusion handling, almost completely miss the low contrast stripes. While EPI2 shows decent performance on both types of occlusion boundaries, LF\_OCC performs almost an order of magnitude better on high contrast stripes as compared to low contrast stripes.

**Photorealistic Scenes.** Figure 8 shows three sample algorithm results for the Herbs scene and a cutout of the Bedroom scene together with region specific challenge evaluations. MSE scores on the Herbs scene are rather similar and relatively high for all three algorithms. On this scene, high errors at the scene background and on the thyme structures reduce the expressiveness of the MSE metric. With our evaluation methods, we specifically quantify performance at smooth surfaces. The bumpiness metric is useful to show that EPI2 features smooth results, whereas the locally smooth but stepped results of LF or the noisy results of LF\_OCC are not suitable in case accurate surface normals are needed per application requirements.

On the Bedroom cutout, MSE scores are much lower. Since fine structures only make up 2.8% of the total cutout, performance on these image regions is poorly reflected by MSE or BadPix scores. Hence quantifying thinning and fattening at fine structures gives additional, more specific characteristics of algorithm performance. LF may have the lowest MSE but it misses most of the fine



**Fig. 8.** Our region specific evaluation on Herbs reveals that EPI2 features the smoothest surfaces but the poorest discontinuities, whereas MSE scores for all three algorithms are close to each other. On the Bedroom cutout we quantify that LF features high fine structure thinning and low fattening whereas EPI2 and LF\_OCC miss fewer structures but tend to fattening.

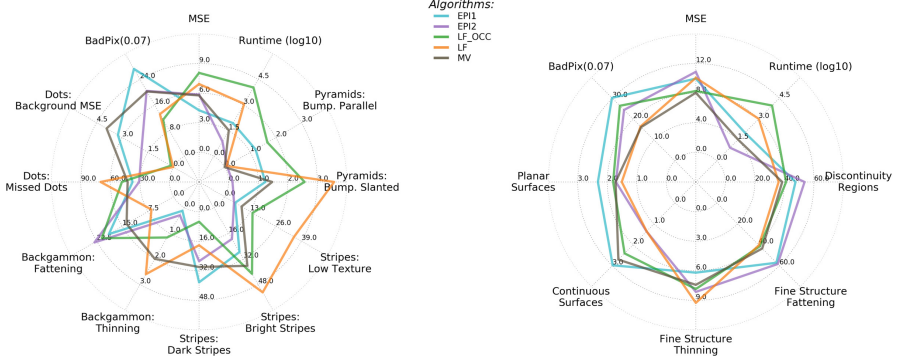
structures which is correctly represented by our thinning scores. By contrast, EPI2 and LF\_OCC have better thinning scores but show very strong fattening.

## 6 Baseline Evaluation of Existing Light Field Algorithms

**Experimental Setup.** We evaluate four state-of-the art light field algorithms and one multi-view stereo approach. The algorithms were selected based on demonstrated state-of-the-art performance and source code availability.

LF [5] poses depth estimation as a multilabel problem which is later refined by locally fitting a quadratic function. For subpixel accurate shifting, the phase-shift theorem is used. LF\_OCC [7] also poses depth estimation as a multi-label problem. As an occlusion cost, boundary orientation in the center view is compared against boundary orientation of so-called scene cam patches, which are constructed from all observed pixels for a single scene point. EPI1 [6] builds a dictionary consisting of atoms of fixed known disparity. By solving a sparse coding problem, the dictionary is employed to recover disparity for each epipolar plane image patch. EPI2 [30] employs the structure tensor to estimate the orientation of patches on the epipolar plane image. A weighted variational regularization is performed to obtain a smooth result. Finally, MV is a lab code implementation of a multi view stereo approach.

**Multidimensional Algorithm Characterization.** In Sect. 5, we used black-box representations of the algorithm results to show that our scenes and metrics



**Fig. 9.** The radar charts summarize all scores of the proposed metrics on the stratified (left) and photorealistic (right) test scenes. Lower scores in the center represent better performance. Neither stratified nor photorealistic scenes can be perfectly solved with a single best algorithm outperforming all others.

are capable of quantifying specific strengths and weaknesses of given algorithms. Here, we demonstrate how our scenes and metrics can be used to obtain an in-depth understanding of algorithm performance. In particular, we show how algorithms can be compared given a range of various scores instead of a single MSE score. Figure 9 summarizes all scores computed on the five algorithms, eight scenes and all associated metrics. Each radar axis represents one metric with zero in the center representing perfect performance.

Neither stratified nor photorealistic scenes can be perfectly solved with a single best algorithm outperforming all others. The radar charts illustrate that every algorithm has different strengths and weaknesses. Thus, if application data mostly contains only a subset of the challenges, optimal choice of algorithm can differ considerably. As algorithm rankings on the MSE and BadPix axes differ from rankings on other performance characteristics, our metrics indeed quantify specific properties which cannot be inferred by simply computing the MSE. For example, the multi-view stereo approach MV scores best in MSE and BadPix over all photorealistic scenes, but in no other dimension.

Furthermore, performance differences and changes in relative rankings per metric are much higher on the stratified scenes than on the photorealistic scenes. Our stratified scenes are very focused on measuring a specific algorithm characteristic, with difficulty levels ranging from feasible to almost impossible. Algorithm performance deteriorates at different levels, allowing quantification of even small differences in top performing algorithms.

**Insights on Specific Algorithm Performance.** The algorithms EPI1 and EPI2 are similar in that they both work on epipolar images. On the radar charts they perform similarly well on most scores of the stratified scenes, but relatively poor in the photorealistic scenes. Our metrics quantify that EPI2 outperforms all other algorithms on reconstructing smooth surfaces in the stratified scenes.

However, on the photorealistic scenes, EPI2 does not feature good scores on planar and continuous surface reconstruction. We speculate that EPI2 is very good on specific challenges but not very robust when different challenges are combined in more complex scenes.

By contrast, LF features solid performance on the photorealistic scenes, but very poor performance on many metrics of the stratified scenes. The strong regularization of LF seems to be good for scoring well on the photorealistic scenes due to the spatial distribution of the contained challenges.

LF\_OCC is the only algorithm explicitly handling occlusions. Indeed, it demonstrates good performance at discontinuities and fine structure in particular on the photorealistic scenes, as well as at the high contrast stripes on the stratified scene. LF\_OCC performance is much lower on the low contrast stripes since it uses image gradients for occlusion handling.

Our dataset reveals several directions for future research: based on the results shown in Figs. 3 and 9, we conclude that occlusion areas, fine structures, the reconstruction of slanted surfaces, and low texture are still unsolved challenges for light field algorithms. Additionally, while most algorithms perform well on some characteristics, there is no algorithm with solid performance on all characteristics simultaneously.

## 7 Conclusion

We presented and carefully justified a novel light field benchmark consisting of 4 stratified and 20 photorealistic light field scenes, a solid evaluation procedure, and a baseline evaluation to seed a public benchmark.

We thoroughly evaluated four state-of-the-art light field algorithms and one multi-view stereo algorithm using our proposed evaluation approach. Thereby, we showed that our dataset highlights open challenges for depth reconstruction algorithms. Moreover, the careful design of our dataset allowed for a structured, quantitative and specific performance analysis of the algorithms at hand. Our evaluation approach facilitated sophisticated and detailed comparisons between the strengths and weaknesses of different algorithms. The presented scenes and evaluation methods are available at <http://www.lightfield-analysis.net>. We encourage researchers to contribute not only algorithms but also datasets and evaluation methods to this benchmark.

In this paper we focused on the geometrical aspects of depth estimation from light fields. In future work we plan to extend the benchmark to include more non-Lambertian materials.

**Acknowledgment.** This work was supported by the ERC Starting Grant “Light Field Imaging and Analysis” (LIA 336978, FP7-2014), the Heidelberg Collaboratory for Image Processing (Institutional Strategy ZUK49, Measure 6.4) and the AIT Vienna, Austria.

## References

1. Levoy, M.: Light fields and computational imaging. *Computer* **39**, 46–55 (2006)
2. Tao, M., Hadap, S., Malik, J., Ramamoorthi, R.: Depth from combining defocus and correspondence using light-field cameras. In: *Proceedings of the International Conference on Computer Vision* (2013)
3. Wanner, S., Goldluecke, B.: Variational light field analysis for disparity estimation and super-resolution. *IEEE Trans. Pattern Anal. Mach. Intell.* **36**, 606–619 (2014)
4. Heber, S., Pock, T.: Shape from light field meets robust PCA. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) *ECCV 2014. LNCS*, vol. 8694, pp. 751–767. Springer, Cham (2014). doi:[10.1007/978-3-319-10599-4\\_48](https://doi.org/10.1007/978-3-319-10599-4_48)
5. Jeon, H., Park, J., Choe, G., Park, J., Bok, Y., Tai, Y., Kweon, I.: Accurate depth map estimation from a lenslet light field camera. In: *Proceedings of the International Conference on Computer Vision and Pattern Recognition* (2015)
6. Johannsen, O., Sulc, A., Goldluecke, B.: What sparse light field coding reveals about scene structure. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3262–3270 (2016)
7. Wang, T., Efros, A., Ramamoorthi, R.: Occlusion-aware depth estimation using light-field cameras. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3487–3495 (2015)
8. Wilburn, B., Joshi, N., Vaish, V., Talvala, E.V., Antunez, E., Barth, A., Adams, A., Horowitz, M., Levoy, M.: High performance imaging using large camera arrays. *ACM Trans. Graph. (TOG)* **24**, 765–776 (2005). ACM. <http://lightfield.stanford.edu/>
9. Marwah, K., Wetzstein, G., Bando, Y., Raskar, R.: Compressive light field photography using overcomplete dictionaries and optimized projections. *ACM Trans. Graph. (Proc. SIGGRAPH)* **32**, 1–11 (2013). <http://web.media.mit.edu/~gordonw/SyntheticLightFields/index.php>
10. Mousnier, A., Vural, E., Guillemot, C.: Partial light field tomographic reconstruction from a fixed-camera focal stack. *arXiv preprint arXiv:1503.01903* (2015). <https://www.irisa.fr/temics/demos/lightField/index.html>
11. Kim, C., Zimmer, H., Pritch, Y., Sorkine-Hornung, A., Gross, M.H.: Scene reconstruction from high spatio-angular resolution light fields. *ACM Trans. Graph.* **32**, 73:1–73:12 (2013). <https://www.disneyresearch.com/project/lightfields/>
12. Rerabek, M., Ebrahimi, T.: New light field image dataset. In: *8th International Conference on Quality of Multimedia Experience (QoMEX)*. Number EPFL-CONF-218363 (2016)
13. Wanner, S., Meister, S., Goldluecke, B.: Datasets and benchmarks for densely sampled 4D light fields. In: *Vision, Modelling and Visualization (VMV)* (2013)
14. Scharstein, D., Szeliski, R.: A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *Int. J. Comput. Vis.* **47**, 7–42 (2002)
15. Butler, D.J., Wulff, J., Stanley, G.B., Black, M.J.: A naturalistic open source movie for optical flow evaluation. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) *ECCV 2012. LNCS*, vol. 7577, pp. 611–625. Springer, Heidelberg (2012). doi:[10.1007/978-3-642-33783-3\\_44](https://doi.org/10.1007/978-3-642-33783-3_44)
16. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? The KITTI vision benchmark suite. In: *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, pp. 3354–3361 (2012)
17. Kristan, M., Matas, J., Leonardis, A., Felsberg, M., Cehovin, L., Fernandez, G., Vojir, T., Hager, G., Nebehay, G., Pflugfelder, R.: The visual object tracking VOT2015 challenge results. In: *Proceedings of the ICCV*, pp. 1–23 (2015)

18. Kristan, M., Pflugfelder, R., Leonardis, A., Matas, J., Porikli, F., Čehovin, L., Nebehay, G., Fernandez, G., Vojir, T.: The VOT2013 challenge: overview and additional results (2014)
19. Scharstein, D., Hirschmüller, H., Kitajima, Y., Krathwohl, G., Nešić, N., Wang, X., Westling, P.: High-resolution stereo datasets with subpixel-accurate ground truth. In: Jiang, X., Hornegger, J., Koch, R. (eds.) GCPR 2014. LNCS, vol. 8753, pp. 31–42. Springer, Cham (2014). doi:[10.1007/978-3-319-11752-2\\_3](https://doi.org/10.1007/978-3-319-11752-2_3)
20. Kristan, M., Matas, J., Leonardis, A., Vojir, T., Pflugfelder, R., Fernandez, G., Nebehay, G., Porikli, F., Čehovin, L.: A novel performance evaluation methodology for single-target trackers (2015)
21. Honauer, K., Maier-Hein, L., Kondermann, D.: The HCI stereo metrics: geometry-aware performance analysis of stereo algorithms. In: Proceedings of the ICCV, pp. 2120–2128 (2015)
22. Menze, M., Geiger, A.: Object scene flow for autonomous vehicles. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition (2015)
23. Zendel, O., Murschitz, M., Humenberger, M., Herzner, W.: CV-HAZOP: introducing test data validation for computer vision. In: Proceedings of the ICCV (2015)
24. Haeusler, R., Kondermann, D.: Synthesizing real world stereo challenges. In: Weickert, J., Hein, M., Schiele, B. (eds.) GCPR 2013. LNCS, vol. 8142, pp. 164–173. Springer, Heidelberg (2013). doi:[10.1007/978-3-642-40602-7\\_17](https://doi.org/10.1007/978-3-642-40602-7_17)
25. Kondermann, D., Nair, R., Honauer, K., Krispin, K., Andrulis, J., Brock, A., Güssefeld, B., Rahimimoghaddam, M., Hofmann, S., Brenner, C., Jähne, B.: The HCI benchmark suite: stereo and flow ground truth with uncertainties for urban autonomous driving. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition Workshops (2016)
26. Perazzi, F., Pont-Tuset, J., McWilliams, B., Gool, L.V., Gross, M., Sorkine-Hornung, A.: A benchmark dataset and evaluation methodology for video object segmentation. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition (2016)
27. Meister, S., Kondermann, D.: Real versus realistically rendered scenes for optical flow evaluation. In: 14th ITG Conference on Electronic Media Technology, pp. 1–6 (2011)
28. Dosovitskiy, A., Fischer, P., Ilg, E., Hausser, P., Hazirbas, C., Golkov, V., van der Smagt, P., Cremers, D., Brox, T.: FlowNet: learning optical flow with convolutional networks, pp. 2758–2766 (2015)
29. Blender Online Community: Blender - a 3D modelling and rendering package (2016)
30. Wanner, S., Goldluecke, B.: Reconstructing reflective and transparent surfaces from epipolar plane images. In: Weickert, J., Hein, M., Schiele, B. (eds.) GCPR 2013. LNCS, vol. 8142, pp. 1–10. Springer, Heidelberg (2013). doi:[10.1007/978-3-642-40602-7\\_1](https://doi.org/10.1007/978-3-642-40602-7_1)



Computer Vision – ACCV 2016

13th Asian Conference on Computer Vision, Taipei,  
Taiwan, November 20–24, 2016, Revised Selected  
Papers, Part III

Lai, S.-H.; Lepetit, V.; Nishino, K.; Sato, Y. (Eds.)

2017, XIII, 490 p. 196 illus., Softcover

ISBN: 978-3-319-54186-0