Single-Shot Metric Depth from Focused Plenoptic Cameras

Blanca Lasheras-Hernandez[†§], Klaus H. Strobl[†], Sergio Izquierdo[§],
Tim Bodenmüller[†], Rudolph Triebel[†‡], and Javier Civera[§]

***Abstract*— Metric depth estimation from visual sensors is crucial for robots to perceive, navigate, and interact with their environment. Traditional range imaging setups, such as stereo or structured light cameras, face hassles including calibration, occlusions, and hardware demands, with accuracy limited by the baseline between cameras. Single- and multi-view monocular depth offers a more compact alternative, but is constrained by the unobservability of the metric scale. Light field imaging provides a promising solution for estimating metric depth by using a unique lens configuration through a single device. However, its application to single-view dense metric depth is under-addressed mainly due to the technology's high cost, the lack of public benchmarks, and proprietary geometrical models and software.**

Our work explores the potential of focused plenoptic cameras for dense metric depth. We propose a novel pipeline that predicts metric depth from a single plenoptic camera shot by first generating a sparse metric point cloud using a neural network, which is then used to scale and align a dense relative depth map regressed by a foundation depth model, resulting in a dense metric depth. To validate it, we curated the Light Field & Stereo Image Dataset[1] (LFS) of real-world light field images with stereo depth labels, filling a current gap in existing resources. Experimental results show that our pipeline produces accurate metric depth predictions, laying a solid groundwork for future research in this field.[2]

I. Introduction

Within the last decades, computer vision has become a relevant domain as numerous applications demand visual perception. For example, it allows robots to navigate, understand, and interact with the environment. Particularly, accurate depth estimation becomes critical for applications where safety and reliability are paramount, such as autonomous driving [1] and robotic manipulation [2].

From a single pinhole camera, we can estimate the 3D reconstruction of an unknown scene using monocular structure-from-motion (SfM), but this is subject to scale ambiguity [3]. In learning-based single-view approaches, scale is also unobservable for self-supervised models [4], as their losses are based on the same multi-view geometric constraints. For supervised ones, the scale accuracy depends on the training and test data [5]. Multi-sensor setups, such as visual-inertial, stereo, or multi-camera, do observe the scale. However, in addition to their higher degree of hardware complexity, they require a precise alignment and calibration, and their accuracy is limited by the baseline between the cameras, which is conditioned by the applications. In the case of visual-inertial systems, their state may not be observable for certain motions.

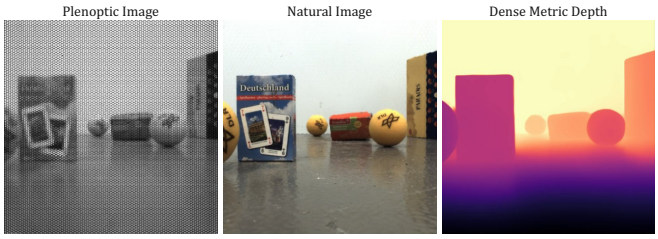

Fig. 1: *Left:* Plenoptic image from a light field camera, displaying the microlens pattern (see detail in Fig. 2). *Center:* Corresponding natural image, synthesized from the central viewpoint of the plenoptic camera. *Right:* Our single-shot metric depth map, at the true scale of the scene.

Light field cameras represent a promising alternative for metric depth estimation. Their inner configuration involves a microlens array (MLA) placed between the main lens and the sensor. This allows the acquisition not only of the intensity but also the direction of light rays that interact with the sensor, by acquiring range-dependent sub-aperture images arranged in a grid pattern [6]. This unique internal structure allows the capture of multiple views and ranges of local areas with a single device and a single shot, overcoming the limitations of traditional monocular and multi-sensor setups. Additional advantages include their higher light capture (as they produce focused images even with large lens apertures), the elimination of disocclusion issues (since the scene is captured in 3D using a single main lens), and the absence of moving parts (unlike optical refocusing systems). These features have already found applications in robotics, such as on-orbit servicing and robotic exploration [7], [8].

In this paper, we explore the potential of a focused plenoptic camera for single-view metric depth estimation (see Fig. 1). Specifically, we present an end-to-end pipeline that leverages light field images to resolve the scale ambiguity of a learning-based monocular depth regressor, hence producing dense metric depth maps. We believe that our approach, which leverages both, priors from a foundation model with global receptive fields and geometric cues between multiple microlenses, is bound to outperform geometric triangulation methods from local correspondences in plenoptic images. To benchmark our approach, we curated a novel real-world Light Field & Stereo Image Dataset (LFS) that we release with this paper. Our results show that we outperform related baselines, specifically the manufacturer's software for depth,

† Institute of Robotics and Mechatronics, German Aerospace Center (DLR) <name>.<surname(s)>@dlr.de
§ I3A, Universidad de Zaragoza {izquierdo,jcivera}@unizar.es
‡ Karlsruhe Institute of Technology <name>.<surname>@kit.edu
[1]Dataset available at https://zenodo.org/records/14224205.
[2]Work partially supported by the DLR Impulse Project *SaiNSOR*.

 DOI: 10.1109/ICRA55743.2025.11128276

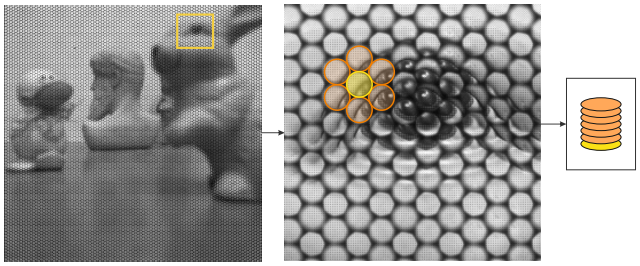

Fig. 2: *Flower stack* illustration. Each *flower stack* is constructed by piling a central microlens and another six in the ring surrounding it. Each stacked microlens is debayered into a 3-channel RGB image.

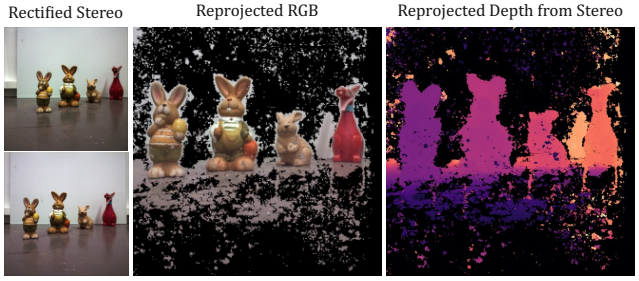

Fig. 3: Visual results produced through the stereo processing pipeline to obtain suitable ground-truth depth values from stereo images, reprojected onto the plenoptic camera's calibrated reference frame.

sparse and presumably based only on geometry, and the state-of-the-art foundation model Depth Anything.

II. Related Work

Depth estimation is a traditional topic in computer vision. **Multi-view approaches** have been extensively studied, offer improvements by capturing data from multiple vantage points rather than just two, as in stereo vision [9]. In practice, most often the sum of squared differences is used to evaluate differences across images, facing the challenge of balancing geometric accuracy with robustness to perspective changes of local image patterns [10], [11]. Techniques such as Kalman filtering enhance depth estimation by processing sequential observations, while aggregation methods like sliding windows or global optimization become necessary when fewer images are available [12]. Structure-from-motion batch pipelines like COLMAP [13] further refine depth estimation by selecting views and ensuring geometric consistency between multiple depth maps, allowing for better handling of occlusions [14], [15], [16], [17], [18].

Single-view depth is geometrically ill-posed [19]. Various computational methods, inspired by human visual perception, have been developed to tackle this challenge using visual cues such as perspective and and visual appearance through lighting and occlusion [20], [21]. Deep neural networks have shown impressive results at predicting per-pixel depth or disparity by learning image patterns from databases with and without supervision [22], [4], [23], [24], [25], [5], [26].

Depth from light field cameras. Capturing light field images in a single exposure combines the simplicity of single-shot setups with the depth information coming from local calibrated multiple views, thereby potentially adding scale information to single-view methods. They have evolved from the plenoptic camera 1.0, which focused the main lens at the microlens array (MLA) plane, allowing for post-capture processing but at a lower resolution, to the plenoptic camera 2.0, which focuses the MLA on the main lens's image plane, enhancing image quality by balancing spatial and angular resolution. For both devices, however, a recurring challenge is the scarcity of available data. This is due to the limited availability of light field cameras and the difficulties in setting up ground-truth measuring systems for supervised learning. While camera arrays offer an effective answer, they are costly and impractical for widespread use. Additionally, plenoptic cameras are hindered by their reliance on proprietary software, high costs, and limited datasets [27], [28], [29].

Some works have addressed the depth estimation task using classical methods developed for earlier light field capturing devices [30], [31], [32], [33]. However, there are no available depth estimation methods for the plenoptic camera 2.0, primarily due to the lack of a publicly available geometric model and the scarcity of available real-world data, which impedes the development of learning-based approaches. This shortage stems from the limited availability of light field capturing devices and from the difficulty of obtaining precise ground truth, as it requires an external system and data registration. Moreover, the technology incorporated into these cameras is conditioned by the dependence on the manufacturer's proprietary camera model and software for decoding raw data [29]. In addition, both their low-volume production and their manufacturing complexity contribute to their elevated cost, preventing widespread adoption [28].

Existing light field datasets, such as the Stanford Light Field Archives [34], [35], [36], provide valuable resources but often lack alignment with modern plenoptic 2.0 cameras, limiting their usability. Although other datasets with similar setups exist [37], [38], their utility is constrained by insufficient data volume, different data formats and hardware version. Hence, the lack of real-world datasets for plenoptic 2.0 cameras has led most research to rely on synthetic datasets, which, while valuable, lack the realism needed for robust deployment [39], [40], [41], [42], [43], [44].

Depth completion methods generate dense depth maps from sparse representations, e.g. in the context of LIDAR data [45]. These methods typically refine and densify sparse depth maps using unguided techniques, such as interpolation or hand-crafted features, within the same modality [46], [47], [48]. However, image-guided methods have proven more successful, especially with noisy depth data [49], [50], [51]. In the case of a light field camera, sparse depth maps are generated from a limited number of microlenses in the sensor. Integrating foundation models as an analogous image-guided method offers a promising approach for depth completion in our domain [52].

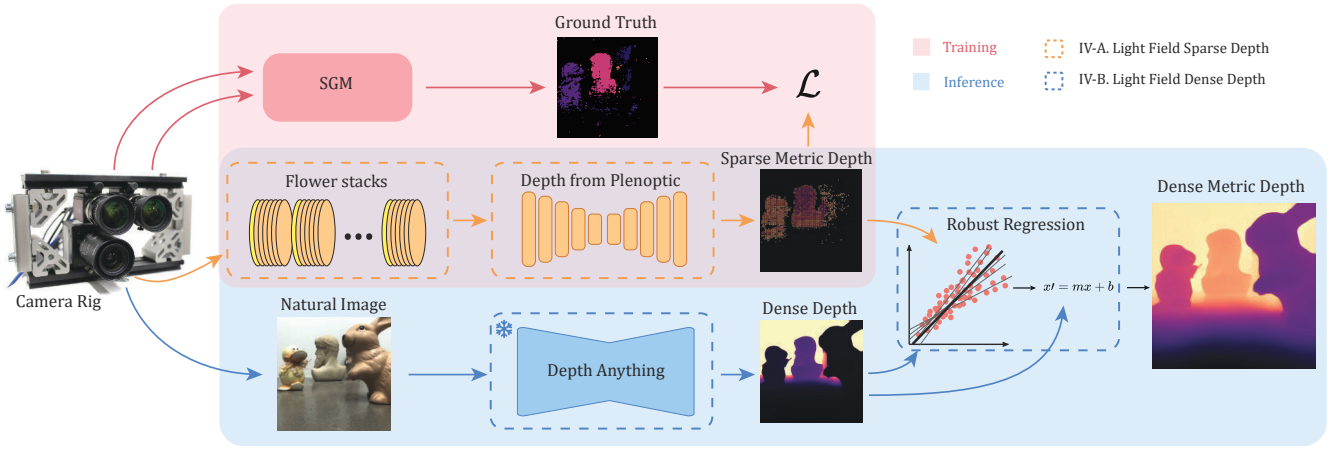

Fig. 4: Overview. The Image Processing Toolkit pre-processes captured data, which is then used to train the *Microlens Depth Network* for sparse metric depth estimation. These are used to refine the inference by Depth Anything, producing a dense metric depth map. Filtered stereo depth serves as the ground truth, following densification and scale alignment.

III. The Light Field & Stereo (LFS) Dataset

Dataset specifications. Our LFS dataset comprises images captured with a plenoptic camera and a stereo. It includes 59 captures from various static scenes in controlled laboratory conditions. Each set contains a *plenoptic image*, a *plenoptic virtual depth image*,[3] and a *natural image*, as well as the *depth from stereo*. Plenoptic images display a pattern of 8,837 microlenses on the sensor, while natural images represent the reconstructed image from the camera's central point of view. Both plenoptic and natural images are produced using the manufacturer's proprietary software, see an example in Fig. 1. We extract semi-dense metric depth from the calibrated stereo, that will serve as a ground-truth.

Hardware and software. The cameras are mounted on a robust mechanical framework made of aluminium profiles. The setup includes two Allied Vision Mako G-419C cameras [53] and a Raytrix R5 plenoptic camera [54], which is based on the Baumer HXG40c model. The Mako cameras are compact, industrial vision cameras, while the Raytrix R5 is a light field camera with up to 1 MP effective output resolution and 4.2 Megarays light field resolution. All cameras feature a CMOS CMV4000 sensor at 4.2 MP resolution, supporting 25 fps under the GigE vision standard. The three cameras are connected to a notebook through a 2.5G switch and a 2.5G Ethernet adapter. To ensure overlapping fields of view between the stereo pair and the plenoptic camera, they are positioned close together with rigid alignment and matching lenses. Synchronization is achieved by using the plenoptic camera's exposure signal as a master trigger for the stereo. The configuration, control, and data processing utilizes the vendors' SDKs, which rely on GenICam transport layer (GenTL) libraries to communicate with the cameras. This setup is extended with in-house packages for concurrent deployment of processes, frame distribution via shared memory, image post-processing, and efficient storage. Geometric camera calibration is carried out using the method by Zhang, Sturm, and Maybank [55], [56] implemented in the camera calibration toolbox DLR CalLab [57], along with the stepwise plenoptic camera calibration method [58] and the extrinsic calibration method by Strobl and Hirzinger [59]. The RxLive software from Raytrix is also utilized to capture plenoptic images and visualize scene representations, with custom methods developed for pre-processing plenoptic images. Metric stereo depth is estimated by an in-house implementation of the SGM algorithm [60]. Lastly, custom scripts were developed for stereo image reprojection.

Scene configuration. Our dataset images a variety of scenes, to benchmark generalization. It includes objects with diverse shapes and textures placed at multiple distances to avoid overfitting. The setup optimizes camera positioning and focal distances across a range of working distances. Illumination conditions are kept constant.

Plenoptic images. We created an *Image Processing Toolkit* that generates a dataset from plenoptic images by modeling the microlenses pattern for efficient indexing and manipulation. Utilizing a hexagonal grid storage system, it uses a dual coordinate system to accommodate the microlens arrangement with interlaced rows (see Fig. 2). A parametric model calculates the coordinates of each microlens based on a 2D calibration of the microlenses pattern. First, the original plenoptic image undergoes debayering into a full-color image. Other key operations include *cropping*, where images are extracted based on centroid positions, and *stacking*, which arranges cropped microlens images from concentric rings to create a *flower stack* (tensor) for enhanced feature extraction and efficient correspondence search.

Plenoptic depth. The data obtained from RxLive is processed as an alternative ground-truth depth source for benchmarking purposes. This involves decoding the virtual depth

[3]Virtual depths are the measured depths of focused projections within the camera relative to the MLA plane, expressed in multiples of the (secretive) sub-millimetric distance between the MLA and the sensor chip. These virtual images resemble disparity maps in stereo vision, as they are closely tied to image processing steps while still related to actual metric depth.

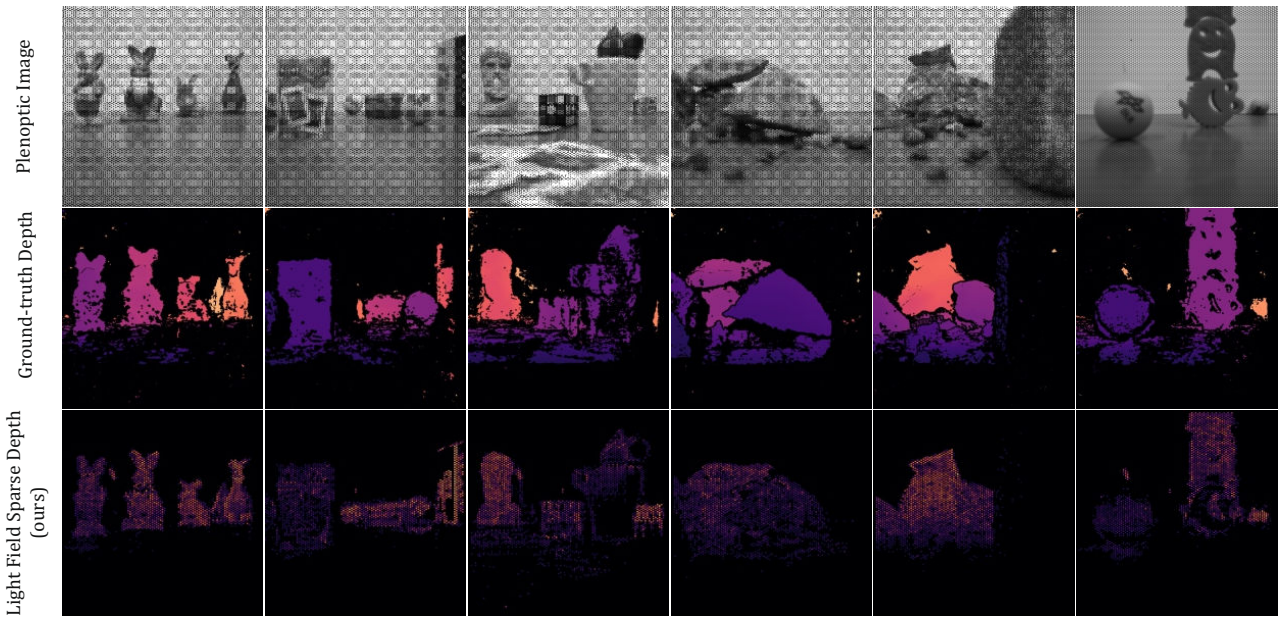

Fig. 5: Qualitative sparse depth results. Top row: Plenoptic images. Center row: Ground-truth depth from stereo. Bottom row: Depth from our *Microlens Depth Network*. Note how our depth values are very close to the ground truth.

values obtained from the manufacturer and transforming them into metric ones using a thin lens camera model [58].

Depth from stereo. The stereo processing pipeline (Fig. 3) involves debayering, undistortion, rectification, and the SGM algorithm (correspondence search, triangulation, and regularization) to produce disparity and metric depth maps. The resulting RGBD image is then re-projected, as a colored pointcloud, onto the calibrated light field camera pose, synthesizing distorted color and metric depth images. While the resulting images resemble natural images, discontinuities occur due to occlusions and untextured regions. To reduce noise in low-texture areas, regularization measures like consistency checks, gradient filters, and hierarchical matching are applied. Finally, depth values are extracted at microlens centroid coordinates of the respective plenoptic image using the *Image Processing Toolkit*.

IV. Single-Shot Depth from Plenoptic Cameras

We present a complete pipeline to obtain dense, metric depth maps from a single light field image, that combines image processing, microlens-scale depth estimation (Section IV-A), and subsequent densification and refinement to generate a dense metric depth map (Section IV-B), see Fig. 4.

A. Sparse Depth from Plenoptic Images

Our model follows an encoder-decoder convolutional architecture. *Flower stacks* are provided to the encoder in the form of 4D tensors X_i of size (N, C, H, W), for which N is the batch size, C the number of channels, and $H \times W$ the height and width of each microlens projection onto the sensor. Our model infers a single metric depth value per stack, corresponding to the prediction at the centroid of the main microlens of each stack. This network, which we denote as *Microlens Depth Network*, is composed of 2D convolutions, batch normalization, activation functions, and fully-connected bottleneck layers. It captures spatial relationships among microlenses, leveraging their redundancy in the *flower stack* to enhance depth estimation robustness, e.g. in regions with occlusions. The estimated depths form a sparse depth map aligned with the central sub-aperture image at a lower resolution due to the limited number of microlenses (8,837). The resulting depth map is inaccurate in regions with weak textures, which we then remove by applying a texture filter.

B. Dense Depth from Plenoptic Cameras

In order to obtain dense per-pixel metric depth values, we fuse these results with the pretrained model *Depth Anything* [26], which generates dense disparity maps from monocular images. *Depth Anything* is trained on extensive datasets, ensuring robustness across a variety of scenarios, and showing state-of-the-art performance in the most recent benchmarks [61]. However, its disparity maps are not in metric scale. We use a robust regression method to align them to the sparse metric depth values predicted by our *Microlens Depth Network*. We use the *Theil-Sen* estimator [62], [63] for this alignment, as it handles outliers in an effective manner. First, dense disparity values are extracted at locations with corresponding sparse metric depths (the latter are converted to disparities for consistency, to achieve a monotonic function that preserves the order), and then aligned using the *Theil-Sen* approach. This non-parametric method calculates the median slope m from all pairs of points (x_i, y_i) and (x_j, y_j) as follows:

$$m = \text{median} \left\{ \frac{y_j - y_i}{x_j - x_i} \mid x_i \neq x_j \right\} \quad . \tag{1}$$

TABLE I: Comparison of our Light Field Sparse and Dense depth against 1) A Random Depth generator, 2) the Sparse Depth provided by the Raytrix manufacturer software, and 3) the foundation model Depth Anything.

	MSE [cm^2] ↓	**RMSE [cm] ↓**	**MARE ↓**	**MSRE ↓**	δ ↑	δ^2 ↑	δ^3 ↑	**BPR ↓**
Random Depth	1551.39	39.27	1403.74	23.45	-	-	-	-
Light Field Sparse Depth (Raytrix) [29]	136.47	10.94	65.43	2.70	**88.62**	**93.04**	**95.43**	**0.3043**
Light Field Sparse Depth (our *Microlens Depth Network*)	**124.68**	**5.55**	**52.75**	**2.63**	84.83	88.40	91.25	0.3949
Depth Anything [26]	129.44	10.96	40.90	5.24	18.30	38.70	65.00	0.8623
Light Field Dense Depth (ours)	**83.21**	**8.50**	**37.30**	**3.94**	**46.40**	**74.60**	**90.00**	**0.4233**

The intercept term b is then determined by taking the median of the individual intercepts

$$b = \text{median}\{y_i - mx_i\} \quad , \tag{2}$$

resulting in the linear model $y = mx+b$. This linear model is then applied to scale and offset the depths predicted by Depth Anything. Finally, disparities are transformed back to metric depth using the intrinsics of the rectified stereo camera.

C. Implementation Details

Architecture. The encoder of the *Microlens Depth Network* comprises five convolutional layers with 2D convolutions, batch normalization, and ReLU activations, processing input tensors of shape $X_i = (N, C, H, W)$, where the batch size is $N = 128$. The *flower stacks* are of size 23×23 pixels with seven RGB images each, resulting in $C = 21$ channels. The bottleneck includes a multilayer perceptron (MLP) with three fully connected layers, compressing the encoded data into a lower-dimensional space to capture high-level features like correspondences and disparities. The decoder, which mirrors the encoder, uses five transposed convolutional layers, with the final layer outputting a single depth channel.

Loss. We use the mean squared error (MSE) between our per-pixel depth predictions $\hat{y}$ and their ground-truth values y

$$MSE = \frac{1}{\sum_{i=1}^{n} M_i} \sum_{i=1}^{n} M_i(\hat{y}_i - y_i)^2 \quad , \tag{3}$$

where M_i is a binary mask indicating whether the ground-truth value at pixel i is available (1) or not (0).

V. Experimental Results

A. Data Setup

As detailed in Section III, the LFS dataset consists of 59 images captured, each containing 8,837 microlenses, resulting in 8,465 *flower stacks*[4] of 7 RGB crops, with a resolution of 23×23 pixels. We split the dataset image-wise in train and test with 49 and 10 images respectively. In Tables I and II we report the standard metrics used by the single-view depth community, find their definitions here [22].

B. Training Details

We train the *Microlens Depth Network* for 125 epochs using batches of 128 *flower stacks*. This requires around 10 hours using a Quadro FV100 Volta GPU. We optimize using Adam [64] with a learning rate of 0.001. For *Depth Anything*, we used 1024×1024 RGB images. *Theil-Sen* regression was applied to align scale and offset using 11,002 valid depth values out of 17,321 sparse data points. The process took 1.78 seconds per image on the same GPU.

[4]The generation of *flower stacks* is only carried out if all microlenses are within the image boundaries. Otherwise, the stack is discarded.

C. Comparison against Baselines

Our *Microlens Depth Network* took 10.54 minutes for a test set of 84,650 *flower stacks*, representing 20% of the LFS dataset (10 images). The test images encompass diverse scenes with elements at varying distances. Table I shows the aggregated results, along with those of related baselines.

We first report the results of a dummy baseline using randomly generated depth values, to properly assess the results of the methods. After that, note how our sparse depth achieves a RMSE of 5.55 cm, almost halving the error over the software provided by the camera manufacturer (Raytrix) with an additional conversion from virtual to metric depths [29], [58]. Finally, observe how our metrics significantly improve the ones of *Depth Anything*, the state of the art on single-view depth estimation, as this last model cannot observe the metric scale of the scene. The accuracy in the prediction of the scale, compared to that of *Depth Anything*, can be clearly observed in the qualitative results of Fig. 6. The source of this scale accuracy traces back to the accurately scaled predictions of our *Microlens Depth Network*, which can also be qualitatively assessed in Fig. 5.

D. Ablation Study

To validate our design choices, we carried out ablative experiments that assessed the importance of each. We summarized them below, and results are shown in Table II.

Flower stacks. In addition to our *flower stacks*, we experimented with single microlens images and double-ringed *flower stacks*. The former lacked multi-view context and thus produced much worse depth estimates, while the latter did not improve the accuracy significantly and had a higher computational cost.

Network's architecture. We tested a fully-convolutional architecture with parallel encoders and a bottleneck, which underperformed compared to our single encoder architecture, which better captured the global context within the *flower stacks*. Integrating an MLP in the bottleneck improved the model's ability to learn complex patterns.

Alternative ground-truth. Comparisons between ground truth depth from the light field camera and the stereo cameras revealed that the stereo depth, despite requiring additional pre-processing from our side, yielded better performance due to a more accurate metric calibration.

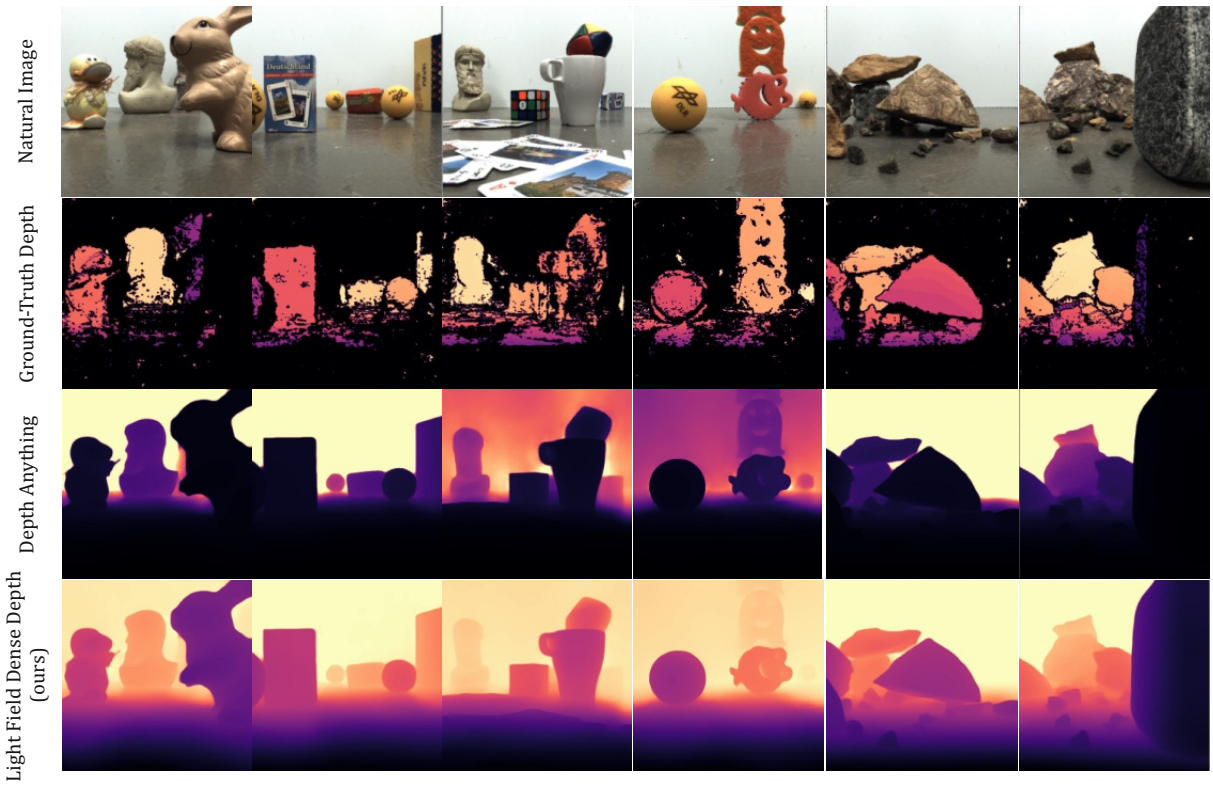

Fig. 6: Qualitative results. From natural images *(first row)*, Depth Anything infers dense depth (yet up-to-scale) *(third row)*. Then, scale alignment with metric values from our Sparse Light Field Depth transforms it into a dense metric depth *(fourth row)*. Note our Light Field Dense depth aligns more closely with the ground truth depth *(second row)* than Depth Anything.

TABLE II: Error metrics for additional methods explored. The first and fourth rows present the proposed methods. Each of these is then ablated and compared quantitatively. The proposed ones outperform the alternatives.

	MSE [cm^2] ↓	**RMSE [cm]** ↓	**MARE** ↓	**MSRE** ↓	δ ↑	δ^2 ↑	δ^3 ↑	**BPR** ↓
Light Field Sparse Depth (ours)	**124.68**	**5.55**	**52.75**	**2.63**	**84.83**	88.40	91.25	0.3949
Parallel encoders	250.18	9.45	104.57	5.13	81.02	**89.74**	**93.03**	0.6872
Weighted Mask	171.57	12.30	80.82	3.64	83.43	88.32	93.01	**0.3802**
Light Field Sparse Depth (ours)	**83.21**	**8.50**	37.30	3.94	**46.40**	**74.60**	**90.00**	**0.4233**
Huber	110.40	9.92	41.30	**2.96**	38.00	73.10	83.60	0.4758
SGD-Huber	110.75	9.94	**27.60**	3.12	43.40	73.60	87.90	0.4631

Metric scaling. We adapted the test-time refinement by Izquierdo and Civera [52], to aggregate our sparse depth with Depth Anything. However, the method struggled with the multi-modal noise distribution of stereo data, which led us to test other regression techniques for robust scale alignment, including the Huber Regressor and the Stochastic Gradient Descent (SGD) with Huber loss. The results showed that Theil-Sen and RANSAC remain most effective for the specific data distributions in this problem.

VI. Conclusion

This work introduces a novel approach for single-shot depth from a plenoptic camera 2.0. We developed a novel *Microlens Depth Network* that successfully infers sparse metric depth from plenoptic images, we then synthesize a natural image from the plenoptic one, that we forward pass through the foundation single-view up-to-scale dense depth model Depth Anything. Finally, we regress the true scale and offset values to correct the dense output from our sparse depth. Our end-to-end pipeline demonstrates the feasibility of generating dense metric depth maps from single shots. Our work makes several key contributions, including the design of a specialized neural network for depth estimation, the creation and release of the Light Field & Stereo Image Dataset (LFS) – a new dataset with plenoptic images and corresponding reprojected metric depth labels – and the development of a comprehensive image pre-processing methodology suited for learning-based applications. These contributions advance the state of the art in light field imaging and single-view depth estimation, establishing a foundation for further research. Future work could focus on refining the developed methodology by exploring more advanced regression techniques, that incorporate object segmentation and independent scale estimation for different image regions, alongside occlusion handling. This would enable more accurate depth scaling, improve metric precision, and address the complexities in the relationships among objects, as opposed to our linear regression approach.

REFERENCES

[1] B. Lasheras-Hernandez, B. Masia, and D. Martin, "DriveRNN : Predicting Drivers' Attention with Deep Recurrent Networks," in *Spanish Computer Graphics Conference (CEIG)*, 2022.

[2] S. Fuchs, S. Haddadin, M. Keller, S. Parusel, A. Kolb, and M. Suppa, "Cooperative bin-picking with time-of-flight camera and impedance controlled DLR lightweight robot III," in *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 4862–4867, IEEE, 2010.

[3] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[4] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth estimation," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3828–3838, 2019.

[5] Z. Li, X. Wang, X. Liu, and J. Jiang, "Binsformer: Revisiting adaptive bins for monocular depth estimation," *IEEE Transactions on Image Processing*, 2024.

[6] M. Feng, S. Z. Gilani, Y. Wang, and A. Mian, "3D face reconstruction from light field images: A model-free approach," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 501–518, 2018.

[7] M. Lingenauber, K. H. Strobl, N. W. Oumer, and S. Kriegel, "Benefits of Plenoptic Cameras for Robot Vision during Close Range On-Orbit Servicing Maneuvers," in *2017 IEEE Aerospace Conference*, pp. 1–18, March 2017.

[8] M. Lingenauber, U. Krutz, F. A. Fröhlich, C. Nissler, and K. H. Strobl, "In-Situ Close-Range Imaging with Plenoptic Cameras," in *2019 IEEE Aerospace Conference*, March 2019.

[9] S. N. Sinha, *Multiview Stereo*, p. 516–522. Boston, MA: Springer US, 2014.

[10] D. Gallup, J.-M. Frahm, P. Mordohai, and M. Pollefeys, "Variable baseline/resolution stereo," in *2008 IEEE conference on computer vision and pattern recognition*, pp. 1–8, IEEE, 2008.

[11] A. Fontan, L. Oliva, J. Civera, and R. Triebel, "Model for multi-view residual covariances based on perspective deformation," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 1960–1967, 2022.

[12] S. B. Kang, R. Szeliski, and J. Chai, "Handling occlusions in dense multi-view stereo," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, vol. 1, pp. I–I, IEEE, 2001.

[13] J. L. Schonberger and J.-M. Frahm, "Structure-from-motion revisited," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4104–4113, 2016.

[14] R. Szeliski, "A multi-view approach to motion and stereo," in *Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149)*, vol. 1, pp. 157–163, IEEE, 1999.

[15] S. B. Kang and R. Szeliski, "Extracting view-dependent depth maps from a collection of images," *International Journal of Computer Vision*, vol. 58, pp. 139–163, 2004.

[16] M. Maitre, Y. Shinagawa, and M. N. Do, "Symmetric multi-view stereo reconstruction from planar camera arrays," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, IEEE, 2008.

[17] G. Zhang, J. Jia, T.-T. Wong, and H. Bao, "Recovering consistent video depth maps via bundle optimization," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, IEEE, 2008.

[18] J. L. Schönberger, E. Zheng, J.-M. Frahm, and M. Pollefeys, "Pixel-wise view selection for unstructured multi-view stereo," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pp. 501–518, Springer, 2016.

[19] B. Chiang and J. Bohg, "Monocular Depth Estimation and Feature Tracking," *Stanford Education*, 2022.

[20] C. Chen, Y. He, H. Mao, L. Zhu, X. Wang, Y. Zhu, Y. Zhu, Y. Shi, C. Wan, and Q. Wan, "A photoelectric spiking neuron for visual depth perception," *Advanced Materials*, vol. 34, no. 20, p. 2201895, 2022.

[21] S. Zhou, T. Zhu, K. Shi, Y. Li, W. Zheng, and J. Yong, "Review of light field technologies," *Visual Computing for Industry, Biomedicine, and Art*, vol. 4, no. 1, p. 29, 2021.

[22] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," *Advances in neural information processing systems*, vol. 27, 2014.

[23] J. M. Facil, B. Ummenhofer, H. Zhou, L. Montesano, T. Brox, and J. Civera, "Cam-convs: Camera-aware multi-scale convolutions for single-view depth," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11826–11835, 2019.

[24] S. F. Bhat, I. Alhashim, and P. Wonka, "Adabins: Depth estimation using adaptive bins," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4009–4018, 2021.

[25] J. Rodríguez-Puigvert, V. M. Batlle, J. Montiel, R. Martinez-Cantin, P. Fua, J. D. Tardós, and J. Civera, "Lightdepth: Single-view depth self-supervision from illumination decline," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21273–21283, 2023.

[26] L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng, and H. Zhao, "Depth Anything: Unleashing the Power of Large-Scale Unlabeled Data," in *CVPR*, 2024.

[27] R. S. Abhilash Sunder Raj, Michael Lowney and G. Wetzstein, "Stanford lytro light field archive." `http://lightfields.stanford.edu/LF2016.html`. [Last access 2024-08-29].

[28] A. Robertson, "VR camera maker Lytro is shutting down, and former employees are going to Google." `https://www.theverge.com/2018/3/27/17166038/lytro-light-field-camera-company-shuts-down-google-hiring`, 2018. [Last access 2023-06-28].

[29] Raytrix, "RxLive 5.0." `https://raytrix.de/experience-the-new-rxlive-5-0-download-try-buy/`. [Last access 2023-12-28].

[30] R. Ng, M. Levoy, M. Brédif, G. Duval, M. Horowitz, and P. Hanrahan, *Light field photography with a hand-held plenoptic camera*. PhD thesis, Stanford university, 2005.

[31] Y. Zhang, P. Yu, W. Yang, Y. Ma, and J. Yu, "Ray space features for plenoptic structure-from-motion," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4631–4639, 2017.

[32] C. Hahne, A. Aggoun, V. Velisavljevic, S. Fiebig, and M. Pesch, "Baseline and triangulation geometry in a standard plenoptic camera," *International Journal of Computer Vision*, vol. 126, pp. 21–35, 2018.

[33] Q. Liu, X. Xie, X. Zhang, Y. Tian, Y. Wang, and X. Xu, "Feature detection of focused plenoptic camera based on central projection stereo focal stack," *Applied Sciences*, vol. 10, no. 21, p. 7632, 2020.

[34] B. Wilburn, *High-performance imaging using arrays of inexpensive cameras*. Stanford University, 2005.

[35] C. Hahne and A. Aggoun, "PlenoptiCam v1. 0: A light-field imaging framework," *IEEE Transactions on Image Processing*, vol. 30, pp. 6757–6771, 2021.

[36] D. G. Dansereau, B. Girod, and G. Wetzstein, "LiFF: Light Field Features in Scale and Depth," in *Computer Vision and Pattern Recognition (CVPR)*, IEEE, June 2019.

[37] A. Mousnier, E. Vural, and C. Guillemot, "Partial light field tomographic reconstruction from a fixed-camera focal stack," *arXiv preprint arXiv:1503.01903*, 2015.

[38] M. Rerabek and T. Ebrahimi, "New light field image dataset," in *8th International Conference on Quality of Multimedia Experience (QoMEX)*, no. CONF in I, 2016.

[39] S. C. G. Laboratory, "The (Old) Stanford Light Fields Archive." `https://www.theverge.com/2018/3/27/17166038/lytro-light-field-camera-company-shuts-down-google-hiring`, 1996. [Last access 2024-07-22].

[40] M. I. of Technology, "Synthetic Light Field Archive." `https://web.media.mit.edu/~gordonw/SyntheticLightFields/index.php`, 2013. [Last access 2023-11-23].

[41] K. Honauer, O. Johannsen, D. Kondermann, and B. Goldluecke, "A dataset and evaluation methodology for depth estimation on 4D light fields," in *Asian Conference on Computer Vision*, Springer, 2016.

[42] P. Paudyal, R. Olsson, M. Sjöström, F. Battisti, and M. Carli, "SMART: A light field image quality dataset," in *Proceedings of the 7th International Conference on Multimedia Systems, MMSys 2016*, pp. 374–379, 2016.

[43] S. Shekhar, S. Beigpour, M. Ziegler, M. Chwesiuk, D. Paleń, K. Myszkowski, J. Keinert, R. Mantiuk, and P. Didyk, "Light-Field Intrinsic Dataset," in *British Machine Vision Conference 2018, BMVC 2018, Northumbria University, Newcastle, UK, September 3-6, 2018*, p. 120, 2018.

[44] V. K. Adhikarla, M. Vinkler, D. Sumin, R. Mantiuk, K. Myszkowski, H.-P. Seidel, and P. Didyk, "Towards a Quality Metric for Dense Light Fields," in *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[45] M. A. U. Khan, D. Nazir, A. Pagani, H. Mokayed, M. Liwicki, D. Stricker, and M. Z. Afzal, "A comprehensive survey of depth completion approaches," *Sensors*, vol. 22, no. 18, p. 6969, 2022.

[46] D. Ferstl, C. Reinbacher, R. Ranftl, M. Rüther, and H. Bischof, "Image guided depth upsampling using anisotropic total generalized variation," in *Proceedings of the IEEE international conference on computer vision*, pp. 993–1000, 2013.

[47] K. Matsuo and Y. Aoki, "Depth image enhancement using local tangent plane approximations," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3574–3583, 2015.

[48] L. Bai, Y. Zhao, M. Elhousni, and X. Huang, "Depthnet: Real-time lidar point cloud depth completion for autonomous vehicles," *IEEE Access*, vol. 8, pp. 227825–227833, 2020.

[49] C. Zhang, Y. Tang, C. Zhao, Q. Sun, Z. Ye, and J. Kurths, "Multitask gans for semantic segmentation and depth completion with cycle consistency," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 12, pp. 5404–5415, 2021.

[50] M. Hu, S. Wang, B. Li, S. Ning, L. Fan, and X. Gong, "Penet: Towards precise and efficient image guided depth completion," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 13656–13662, 2021.

[51] D. Nazir, A. Pagani, M. Liwicki, D. Stricker, and M. Z. Afzal, "Semattnet: Toward attention-based semantic aware guided depth completion," *IEEE Access*, vol. 10, pp. 120781–120791, 2022.

[52] S. Izquierdo and J. Civera, "SfM-TTR: Using Structure from Motion for Test-Time Refinement of Single-View Depth Networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21466–21476, 2023.

[53] A. Vision, "Mako G-419. Small size, powerful performance." `https://www.alliedvision.com/en/camera-selector/detail/mako/g-419/`. [Last access 2023-08-26].

[54] Raytrix, "3D light-field vision technology made in Germany." `https://raytrix.de/products/`. [Last access 2023-08-28].

[55] Z. Zhang, "A Flexible new Technique for Camera Calibration," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, pp. 1330–1334, November 2000.

[56] P. F. Sturm and S. J. Maybank, "On Plane-Based Camera Calibration: A General Algorithm, Singularities, Applications," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (Fort Collins, CO, USA), pp. 432–437, June 1999.

[57] K. H. Strobl, W. Sepp, S. Fuchs, C. Paredes, M. Smíšek, and K. Arbter, "DLR CalDe and DLR CalLab." Institute of Robotics and Mechatronics, German Aerospace Center (DLR). Oberpfaffenhofen, Germany. `http://www.robotic.dlr.de/callab/`, 2005. [Last access 2024-06-11].

[58] K. H. Strobl and M. Lingenauber, "Stepwise calibration of focused plenoptic cameras," *Computer Vision and Image Understanding*, vol. 145, pp. 140–147, 2016.

[59] K. H. Strobl and G. Hirzinger, "Optimal Hand-Eye Calibration," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, (Beijing, China), pp. 4647–4653, October 2006.

[60] H. Hirschmüller, "Stereo Processing by Semi-Global Matching and Mutual Information," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, pp. 328–341, February 2008.

[61] J. Spencer, F. Tosi, M. Poggi, R. S. Arora, C. Russell, S. Hadfield, R. Bowden, G. Zhou, Z. Li, Q. Rao, *et al.*, "The third monocular depth estimation challenge," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1–14, 2024.

[62] H. Theil, "A rank-invariant method of linear and polynomial regression analysis," *Indagationes mathematicae*, vol. 12, no. 85, p. 173, 1950.

[63] P. K. Sen, "Estimates of the Regression Coefficient Based on Kendall's Tau," *Journal of the American Statistical Association*, vol. 63, no. 324, p. 1379–1389, 1968.

[64] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.