Light Field Depth Estimation

Conventional and Deep Learning Solutions

Yongqi Zhou, YZ, Zhou*

Carnegie Mellon University, yongqiz2@andrew.cmu.edu

This work delves into the field of light field depth estimation, researching both conventional and cutting-edge algorithms for predicting disparity information from a light field image. Traditional methods for depth estimation from light field images suffer from the challenge of high computational costs and the difficulty of handling occlusion, making the use of learning-based algorithms highly attractive. This article aims to explore the potential to utilize learning-based algorithms to balance computational cost and depth estimation accuracy, providing accurate and detailed disparity information from a light field image. My work replicated the paper proposed by Kim et al.**[1]** in 2013, which calculates the disparity map based on EPI and edge confidence at multiple resolutions, and experimentally compared two deep learning-based light field disparity estimation methods proposed in 2023, ranking second in the runtime index and third in the MSE index in the 4D Light Field Benchmark. Other deep learning-based approaches are also discussed, with several potential improvements explored through experimentation. The findings of this study may also have practical implications in the development of more efficient and accurate algorithms for estimating depth information from light field images.

CCS CONCEPTS • Machine learning • Computer graphics • Artificial intelligence

Additional Keywords and Phrases: Light Field, Depth Estimation, Deep Learning, Multi-view Stereo

1 INTRODUCTION

Light field (LF) imaging is an intriguing topic in the area of computer vision. Compared to ordinary images, light field images contain information about the angle of light rays. Multi-angle information of an object is obtained through a single shot taken by an array of sub-cameras. Using LF instead of multiple cameras results in reduced costs and errors associated with camera calibration and alignment, offering adaptability to complex scenes, while lowering hardware costs associated with using multiple cameras. These advantages make LF imaging a more suitable option for various research applications, such as view synthesis and image segmentation.

With the increasing prevalence of plenoptic cameras and simplified LF acquisition methods, LF depth estimation has the potential to become a promising breakthrough in camera-based 3D positioning and reconstruction. The research of LF depth estimation poses a significant challenge due to the need to extract depth information from an LF image that contains dense and regularly sampled views. Conventional approaches to this problem that rely on geometric matching, light field structure, or refocusing techniques are mostly optimization-based, which poses the issue of being computationally complex and lacking flexibility, thus not suitable for real-world applications. Furthermore, obstacles such as occlusion and texture-less regions also adversely impact the LF structure.

In recent years, researchers have been applying deep learning methods to LF depth estimation. In comparison to optimization-based methods, deep learning methods can use pre-trained model parameters to make overall predictions on images. On the one hand, this reduces the computing power consumed by traditional methods, which require a large amount of pixel-wise calculations and matches for each image. On the other hand, through non-linear or semantic feature extraction, deep learning methods can be more robust to noise and occlusion. Although research on the application of deep learning to light field algorithms is still limited, the latest methods have surpassed traditional algorithms on various metrics. Further research on this could have significant implications in fields such as autonomous driving, virtual reality, and robot navigation.

To gain a better understanding of LF depth estimation, this article discusses a series of classical algorithms and prominent deep learning-based methods from recent years. Section 2 provides an overview and classification of optimization-based and learning-based methods. Section 3 presents the implementation of a paper proposed by Kim et al. in 2013[1], which calculates the disparity map based on EPI and edge confidence at multiple resolutions, as well as two newer deep learning methods, OACC-Net[2] and DistgDisp[3]. The datasets, evaluation, benchmarks, and the comparation of former results are further discussed in Section 4. A conclusion of the LF depth estimation problem is provided in the last section, which covers some potential areas of application and improvement points for the future

2 RELATED WORK

Light field images contain multiple viewpoints of a scene, making depth estimation possible with light field cameras. However, the short baselines between the multiple viewpoints can lead to matching errors. Currently, methods for depth estimation in light fields can be broadly categorized into conventional (optimization-based) methods and learning-based methods. Some learning-based methods are developed based on the theoretical foundation of optimization methods. In this section, typical algorithms from each category are discussed as examples.

2.1 Optimization-based Methods

2.1.1Multi-views matching.

Depth estimation based on multi-view matching has evolved from traditional stereo-matching methods for 2D images. Traditional stereo-matching methods require two or more cameras to capture a scene, which requires overcoming camera

shake and human operation. In contrast, a light field camera captures a scene as if from multiple cameras at once and is almost unaffected by camera shake, making it very promising for depth estimation. Taking the Lytro Illum as an example, each microlens captures 225 light rays from the main lens in a 15x15 pixel array behind it. Selecting the same position behind each microlens can generate an image from a specific viewpoint. Traversing the 15x15 pixels behind each microlens can generate 225 different perspective images. These images have different viewing angles or disparities, which can be used to calculate depth using a matching method.

However, due to the short baseline of light field cameras, matching errors are common. Accurate matching pairs are critical for multi-view matching-based depth estimation. Jeon et al.[4] proposed a subpixel multi-view stereo matching algorithm to achieve sub-pixel accuracy matching, which solves the problem of a short baseline to some extent. The core of this algorithm is the use of phase-shifting theory, in which a small spatial displacement in the time domain is a product of the exponentiation of the displacement in the frequency domain of the original signal. To enable matching between sub-view images, the authors designed two different cost functions: Sum of Absolute Differences (SAD) and Sum of Gradient Differences (GRAD). The final matching cost C is a function of pixel position x and disparity layer l, as shown in the following formula:

$$C(x,l) = \alpha C_A(x,l) + (1-\alpha) C_G(x,l)$$

where C_A is defined as:

$$C_A(x,l) = \sum\sum \min\left(\left|I(uc,x) - I\left(u, x + \Delta x(u,l)\right)\right|, \tau\right)$$

It is constructed by comparing the differences between the central view image $I(uc,x)$ and the other views $I(u,x)$, repeatedly moving a small distance around the pixel x in a specific sub-view and subtract it from the central view until all sub-views (i=1…N) have been compared. By using the aforementioned phase-shifting theory, the pixel intensity after displacement can be obtained, and Δx increases linearly with the distance between the viewpoints and the central viewpoint. Furthermore, SGD loss is calculated in both x and y directions, and the weighting of the cost function in both directions is determined by the relative distance between any viewpoint and the central viewpoint. Lastly, the authors established a multi-label optimization model and an iterative optimization model to optimize the depth map.

2.1.2 EPI based method

Unlike the multi-view stereo matching method, the EPI method estimates depth by analyzing the structure of the light field data. The slope of the diagonal line in the EPI image can reflect the depth of the scene. The larger the horizontal displacement in the EPI image, the larger the disparity corresponding to the diagonal line, indicating a smaller depth.

The earliest work of EPI for depth estimation was proposed in 1987 by Bolles et al. for structural depth estimation under a moving background, based on the color consistency principle assumption[5]. However, this approach lacked robustness against occlusion and noise. Subsequent work by Zhang et al.[6] sought to enhance the robustness of EPI-based methods to strong occlusion and noise by measuring the slope of the EPI using a rotating parallelogram operator. The operator was integrated into the two-dimensional EPI, measuring the partial distance between two parts of the window. Wanner et al. [7] estimated the local direction of lines using the structure tensor in EPI's spatial domain, and then introduced a smooth optimization to construct global depth.

One representative algorithm is the large scene reconstruction method proposed by Kim et al in 2013[1]. The slope m of a line segment associated with a scene point at distance z can be expressed as:

$$m = 1/d = z/fb,$$

where d is the image space disparity between a pair of images captured at adjacent positions or the displacement between two adjacent horizontal lines in an EPI. f is the camera focal length in pixels, and b is the metric distance between each adjacent pair of imaging positions. The authors used a fine-to-coarse approach to estimate depth in the EPI by starting at the highest resolution edges, propagating the information, and gradually reducing the EPI resolution. A more detailed algorithm process and experimental results will be presented in the next section.

EPI-based methods exhibit excellent performance when the depth undergoes continuous changes along a straight line in space. However, when the line is interrupted by occlusion or noise, these methods may produce erroneous predictions. Although some algorithms as Kim's are efficient and robust to noisy measurements and occlusions, the introduced constraints such as piecewise processing can increase the algorithm's complexity.

2.1.3 Defocus-based method

An important feature of a light field camera is that it allows post-capture refocusing, which is based on the light field shear principle. By measuring the "blur" of each pixel at different focal planes, its corresponding depth can be estimated. One representative algorithm is proposed by Tao et al.[8] Tao's algorithm utilizes the entire light field image captured by the sensor for computation. During the refocusing process, it obtains the sensor image at different focal planes and calculates the correspondences between different images, i.e., the matching relationships of images captured from different viewpoints.

Specifically, for different depth stacks, Tao's algorithm extracts defocus cues and correspondence cues separately. Defocus is defined as

$$D_\alpha(x) = \frac{1}{|W|}\sum |\Delta x \, L\alpha(x')|$$

where W represents the window size of the current pixel neighborhood, Δx represents the horizontal Laplacian operator, and $L\alpha(x')$ is the refocused light field image after averaging.

Furthermore, correspondence cue is defined as

$$C_\alpha(x) = \frac{1}{|W|}\sum |\sigma_\alpha(x')|$$

where $\sigma_\alpha(x')$ represents the standard deviation of the intensity of each macro-pixel.

Based on these two cues, maximizing spatial contrast can obtain the depth corresponding to defocus clues, while minimizing angular variance can obtain the depth of correspondence. Finally, global optimization was performed on these two depth maps using Markov Random Field (MRF).

2.2 Learning-based Methods

Deep learning has been widely applied in the field of computer vision. In the problem of light field depth estimation, learning-based methods are mainly based on two approaches. The first approach involves fusing different features using models in order to achieve more accurate optimization. The second approach involves optimizing the computation cost. These two mainstream approaches are respectively referred to as EPI-based and cost volume-based methods, as described in paper [9].

2.2.1 EPI-based:

The most representative EPI-based model is EPINet[10], which is often used as a benchmark for comparison when developing new algorithms. The characteristic of this network architecture is that it stacks the light field data from four different angles and then performs the stacked convolution operation on features from each angle, aiming to capture the relationships between the features from different angles. Each direction of the sub-aperture image corresponds to a network branch that is responsible for encoding and extracting features from images in the corresponding direction. Each branch of the network consists of three fully convolutional modules, where each module includes a $Conv \rightarrow Relu \rightarrow Conv \rightarrow BN \rightarrow Relu$ block. To address the problem of short baseline, a 2*2 convolution kernel is used with a stride of 1. The results obtained from the four directions are concatenated and input into the subsequent convolutional blocks. Finally, it uses the $Conv \rightarrow Relu \rightarrow Conv$ structure to obtain sub-pixel-level estimation precision.

On the basis of EPINet, some researchers have made effective improvements. For instance, Leistner et al. [11] introduced EPI-shift, which enables virtual shifting of LF stacks, allowing for the retention of a small receptive field to be effective in the case of wide-baseline. However, due to the fact that an EPI is merely a two-dimensional horizontal or vertical section of a four-dimensional light field, it is challenging to encode all of the spatial and angular hints through EPIs, leading to a limited utilization of valuable information [3].

2.2.2 Cost Volume-based:

Some algorithms have optimized the computational consumption of depth estimation while obtaining more accurate results by constructing cost volumes and designing elaborate loss functions. In 2020, Tsai et al. proposed LFattNet [12], which utilizes a view selection module to prioritize important views and reduce redundancy in order to effectively utilize all views for more accurate depth estimation.

The proposed architecture is shown in figure 1. Specifically, to extract unary features from each sub-aperture view of the light field image, four basic residual blocks are used, followed by a spatial pyramid pooling (SPP) module to extract context information and generate effective feature maps. These feature maps are concatenated into a 5D cost volume across all sub-aperture views. Before sending the cost volume for disparity regression, an attention-based view selection module is applied to obtain an attention map that specifies the importance of each view. The cost volume is combined with the attention map and sent to the disparity regression module, which produces the disparity map for the center view in the light field image.

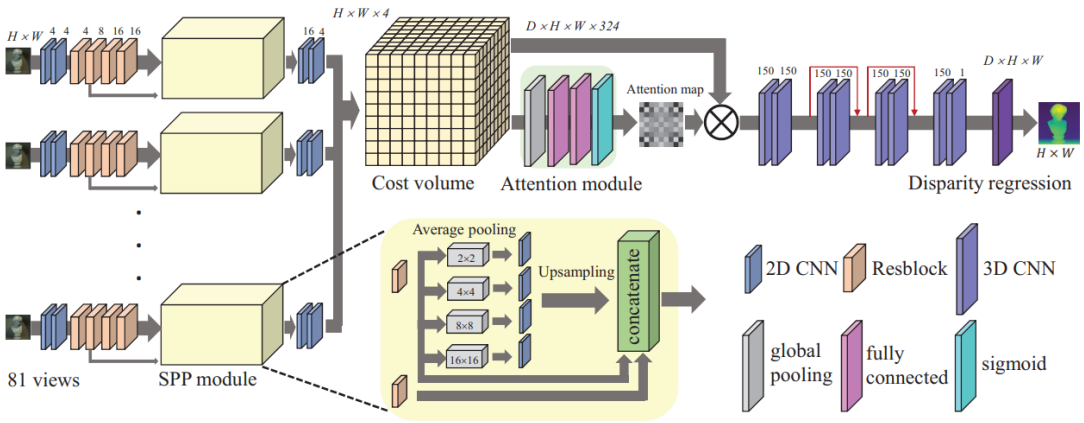

Figure 1: The architecture of LFattNet[12]

3 METHOD AND IMPLEMENTATION

3.1 Non Machine-learning Method

Kim and his team published a paper in 2013 entitled "Scene Reconstruction from High Spatio-Angular Resolution Light Fields." They proposed a method for depth estimation based on EPI that operates on individual rays of light. The core of this method is to first estimate edge depth and assign confidence scores, and then propagate the depth to EPI-pixels that have similar radiance, iteratively calculating all object contour features. Afterward, multiple resolution images are down-sampled to calculate the depth of areas with fewer details, estimating the inner areas from fine to rough, and finally smoothing the depth image using a median filter.

Specifically, the first step of calculating edge confidence is to determine which pixels in the EPI image may have potential for depth estimation.

$$Ce(u,s) = \sum_{u' \in N(u,s)} \|E(u,s) - E(u',s)\|^2,$$

According to the equation above, if there is a large color difference between two pixels, they may correspond to different depths in the light field.

Then, the algorithm computes depth estimates for confident EPI-pixels in the light field image. It performs the computation per scanline in the EPI, assigning a depth estimate to each EPI-pixel using a modified Parzen window estimation.

$$S(u,d) = 1/|R(u,d)| \sum_{r \in R(u,d)} K(r - \bar{r})$$

When a group of brightness values $|R(u,d)|$ are tightly distributed in the color space, they are more likely to represent the same point in the scene, and the density score S(u, d) will be higher. Based on the edges confidence and EPI depth score, a fine-to-coarse iterative approach is used to calculate the entire disparity map.

The results presented in the paper are very promising. However, in our attempts to replicate the experiments, we tested the algorithm using images from the HCI 4D LF Benchmark and the generated results were not as ideal as we had hoped. Furthermore, these results were heavily influenced by the parameters used. The image in Figure 2 shows the performance of the algorithm on Box, Greek, Bedroom, and Bicycle, where the parameters were mainly adjusted based on the effectiveness of generating Box. The top four images are the original images, and the bottom half shows the results of the generated disparity maps. From the results, the algorithm seems to perform well in edge detection but it not accurately estimated depth. It's uncertain if this is due to an error in the code during the reproduction process.

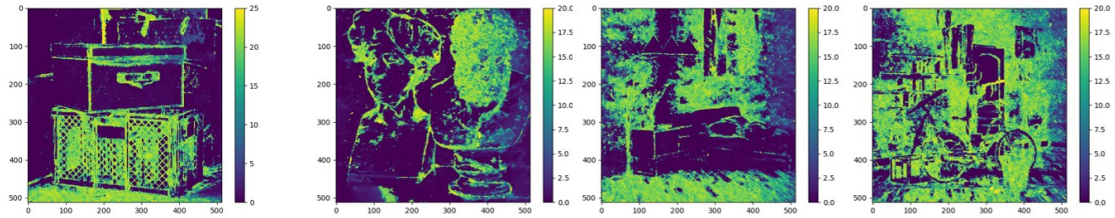

Figure 2 Raw Data and Disparity Map Result of 3.1 Method

3.2 Deep-learning based Method

3.2.1 Disentangling Light Fields

In 2023, Wang and their team proposed the DistgDisp[3] method, which achieved state-of-the-art performance in spatial super-resolution, angular super-resolution, and disparity estimation. They proposed a light field image processing decoupling mechanism at the level of feature extraction, rather than addressing the problem of disparity estimation itself. The mechanism employs different convolutions in spatial, angular, and kernel-line domains to decouple the LF image into different subspaces, enabling more effective acquisition of intrinsic or learned structural features. These features are then utilized to address specific light field image processing problems. The effectiveness of the proposed mechanism is validated through the design of networks for specific problems, including spatial and angular super-resolution and disparity estimation, with the development of the DistgDisp network specifically for the latter.

The LF decoupling mechanism is applied on micro-pixel images (MacPI), which, like sub-aperture image (SAI), are a representation of LF. While MacPI is not human-friendly for visual perception, it enables uniform mixing of spatial and angular information in the LF. This facilitates the use of convolution to extract and merge spatial and angular information.

The backbone architecture designed in the paper is illustrated in the figure below. The model takes MacPI as input and performs feature extraction, cost volume construction, cost aggregation, and disparity regression in sequence. The model includes a spatial residual block (i.e. spatial resolution block) with batch normalization (BN) to model the relationship between each pixel and its spatially neighboring pixels. A SFE module is added to extract spatial contextual information, and a spatial Res block is constructed in the $SFE \rightarrow BN \rightarrow LeakyReLU \rightarrow SFE \rightarrow BN$ structure. The original input features are added to the end of the block for local residual learning.

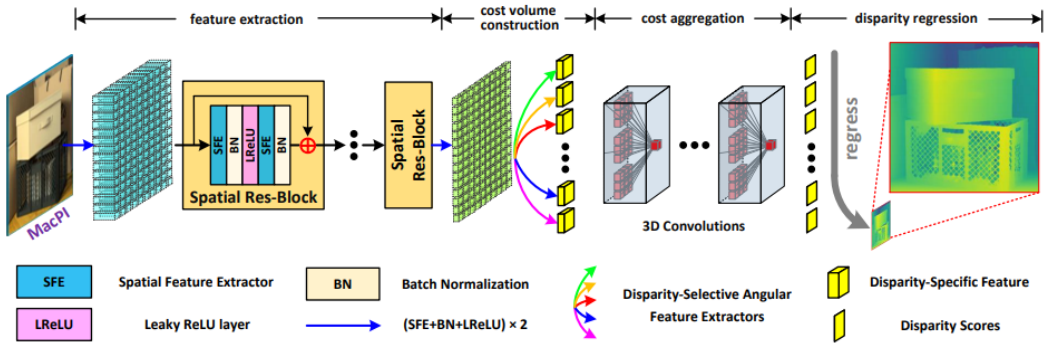

Figure 3: Model architecture of DistgDisp[3]

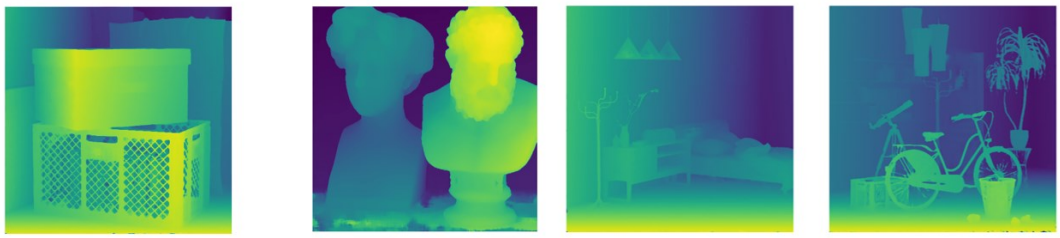

Figure 4: Disparity Map Result of 3.2.1 Method

3.2.2OACC-Net

Another method proposed by the same authors is the OACC-Net[2], which focuses on cost construction to reduce the negative impact of occlusion on disparity estimation. The Occlusion-Aware Cost Constructor module is proposed to achieve occlusion-aware cost construction.

Firstly, as LF images have regular spatial-angular structures, the model performs convolution on the SAI array to construct the cost, with the dilation rate decreasing as the disparity of the pixel increases. Then, a pixel modulation mechanism is designed to dynamically handle occluded areas. With this mechanism, the cost constructor can adjust the cost of each view at each position to achieve occlusion-aware cost construction.

Finally, a parameter-free method is proposed to derive the occlusion mask for each view. Specifically, for the occluded areas, the scene points available in the center view may not be available in the surrounding views, and the pixels in the surrounding views may not have corresponding pixels in the center view. Therefore, a fine-grained occlusion mask can be calculated based on the photometric consistency prior.

The OACC-Net model uses an initial 3x3 convolution for feature extraction, followed by 8 residual blocks for deep feature extraction. The last residual block generates features for cost construction, with shared weights across different views. The OACC module uses an iterative method to optimize the final disparity map, with 8 x 3D convolution layers used for cost aggregation and disparity estimation.

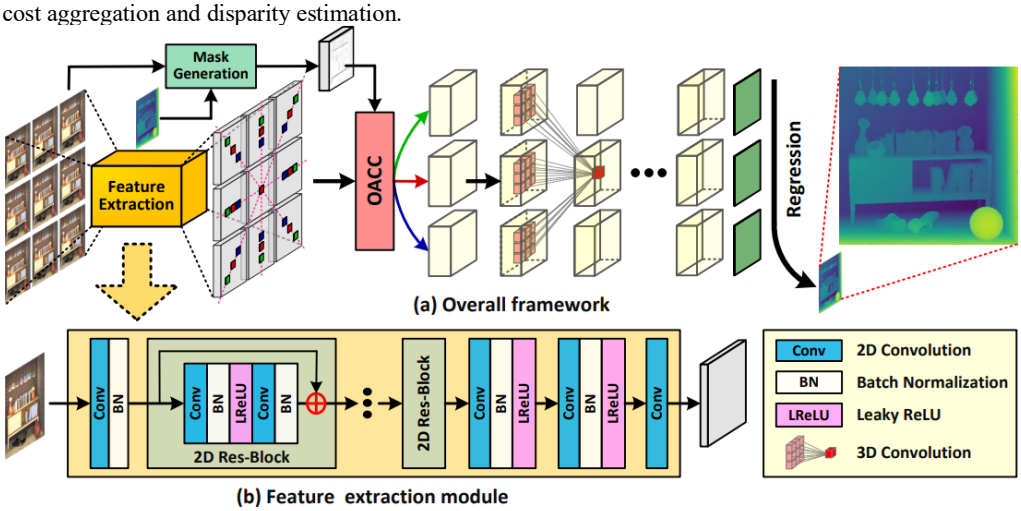

Figure 5: Model Architecture of OACC-Net[2]

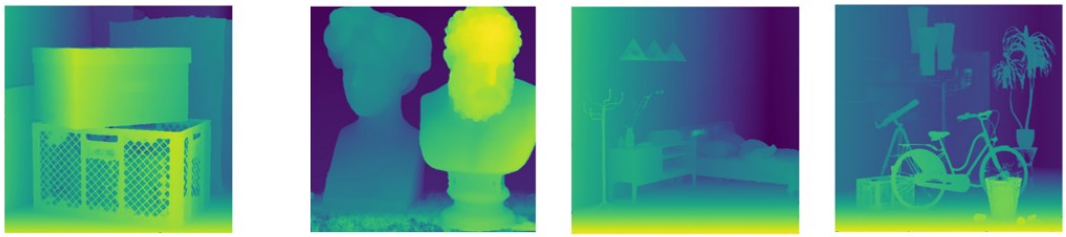

Figure 6: Disparity Map Result of 3.2.2 Method

As comparison, both OACC-Net and DistgDisp use deep learning for disparity estimation, and follow the same four-step process: feature extraction, cost construction, cost aggregation, and disparity regression. The feature aggregation and final disparity regression parts are similar in both methods. The DistgDisp method uses a decoupling mechanism based on MacPI to further extract features and speed up computation, while the OACC-Net model uses occlusion-aware cost construction with mask clustering and pixel modulation based on SAI to better handle occlusion.

4 EVALUATION AND BENCEMARKS

For the problem of depth estimation in light fields, there exists a well-established benchmark evaluation website that provides current algorithm rankings and evaluation tools[13]. Evaluation is conducted on HCI 4D LF dataset, which includes 9x9x512x512x3 light fields as individual PNGs for various scenes. The basic evaluation metrics are shown in the following Table 1, and the current (2024.1) benchmark is shown in figure 7 as ranking by MSE.

Table 1: Light Field Depth/Disparity Estimation Metric

Metric	Description
BadPixel (0.01/0.03/0.07)	The percentage of pixels at the given mask with abs(gt - algo) > 0.01/0.03/0.07
MSE	The mean squared error over all pixels at the given mask, multiplied with 100.
Q25	The 25th percentile of the disparity errors: The maximum absolute disparity error of the best 25% of pixels for each algorithm.
Runtime	The runtime in seconds as reported by the authors.

	General											
Algorithm	BadPix(0.01) Description		BadPix(0.03) Description		BadPix(0.07) Description		MSE Description ▲		Q25 Description		Runtime Description	
ESMNet	21.302	20	8.292	31	3.523	28	0.828	1	0.194	21	5.149	56
LFRNN	22.344	23	5.599	20	3.017	19	0.854	2	0.273	34	0.000	1
OACC-Net	22.041	22	6.173	23	3.135	23	0.862	3	0.198	22	0.034	3
SubFocal-L	10.041	2	3.660	1	2.408	2	0.878	4	0.114	4	54.979	93
SubFocal	13.992	7	4.322	9	2.412	3	0.888	5	0.109	3	5.857	63
CasLFNet	10.922	3	4.177	7	2.463	7	0.894	6	0.106	2	9.679	73
AttMLFNet	15.809	8	5.007	14	2.812	13	0.903	7	0.126	6	4.619	52
mulfnet	12.604	4	4.899	13	2.892	14	0.911	8	0.125	5	1.000	31
Query-EPI	18.115	14	3.896	4	2.283	1	0.920	9	0.155	14	7.319	70

Figure 7: LF Benchmark Ranking by MSE

As the running time recorded on the benchmark is self-reported, we conducted actual measurements of the execution time for the three algorithms during the experiment to verify the results. The running times on Nvidia RTX 3070 Ti + i9-12900 are shown in Table 2. These results do not match those reported in the benchmark, possibly due to differences in

the operating platform or timing measurement methods. However, it can be observed that the DistgDisp method is faster than OACC-Net, and both deep learning methods are significantly faster than the non-deep learning method.

Table 2: Running Time of Three Algorithms on Different Test Data

Algorithm/ Runtime(s)	Box	Bicycle	Bedroom	Greek
DistgDisp	27.064	26.588	27.002	27.032
OACC-Net	50.282	50.679	50.559	51.613
Non ML	408.555	/	/	/

5 CONCLUSION

This paper explored both conventional optimization-based methods and cutting-edge deep learning approaches for extracting depth information from LF images. The findings demonstrate the limitations of traditional methods, particularly their high computational cost and sensitivity to occlusions. Conversely, deep learning methods like OACC-Net and DistgDisp showcased promising results in terms of accuracy and efficiency. However, there are still challenges to address and exciting future directions to explore.

***Computational Cost*:** While deep learning methods achieve high accuracy, their computational cost can be significant. Future research should focus on developing more efficient feature abstraction methods and exploring techniques to fuse redundant information across various views in the light field. This could involve dimensionality reduction techniques or lightweight network architectures.

Occlusion Handling: Occlusions in the scene can lead to inaccurate depth estimation. Integration with semantic understanding modules that can identify objects and their potential occlusions can be beneficial. Additionally, generative models hold promise for filling in occluded regions by learning to synthesize plausible content based on surrounding information.

Future Directions and Applications: Developing real-time capable ML models for LF depth estimation would unlock applications in robotics, autonomous navigation, and interactive 3D scene manipulation. Lightweight network architectures and hardware acceleration techniques are crucial for achieving this goal.

In conclusion, ML-based LF depth estimation is a rapidly evolving field with vast potential. By addressing computational challenges, improving occlusion handling, and exploring new learning paradigms, this technology can unlock a future rich with innovative applications across various domains.

REFERENCES

[1] C. Kim, H. Zimmer, Y. Pritch, A. Sorkine-Hornung, and M. Gross, "Scene reconstruction from high spatio-angular resolution light fields," *ACM Trans. Graph.*, vol. 32, no. 4, pp. 1–12, Jul. 2013, doi: 10.1145/2461912.2461926.

[2] Y. Wang, L. Wang, Z. Liang, J. Yang, W. An, and Y. Guo, "Occlusion-Aware Cost Constructor for Light Field Depth Estimation," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA: IEEE, Jun. 2022, pp. 19777–19786. doi: 10.1109/CVPR52688.2022.01919.

[3] Y. Wang *et al.*, "Disentangling Light Fields for Super-Resolution and Disparity Estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 425–443, Jan. 2023, doi: 10.1109/TPAMI.2022.3152488.

[4] H.-G. Jeon *et al.*, "Accurate depth map estimation from a lenslet light field camera," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2015, pp. 1547–1555. doi: 10.1109/CVPR.2015.7298762.

[5] R. C. Bolles, H. H. Baker, and D. H. Marimont, “Epipolar-plane image analysis: An approach to determining structure from motion,” *Int J Comput Vision*, vol. 1, no. 1, pp. 7–55, Mar. 1987, doi: 10.1007/BF00128525.
[6] S. Zhang, H. Sheng, C. Li, J. Zhang, and Z. Xiong, “Robust depth estimation for light field via spinning parallelogram operator,” *Computer Vision and Image Understanding*, vol. 145, pp. 148–159, Apr. 2016, doi: 10.1016/j.cviu.2015.12.007.
[7] S. Wanner and B. Goldluecke, “Variational Light Field Analysis for Disparity Estimation and Super-Resolution,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 3, pp. 606–619, Mar. 2014, doi: 10.1109/TPAMI.2013.147.
[8] M. W. Tao, S. Hadap, J. Malik, and R. Ramamoorthi, “Depth from Combining Defocus and Correspondence Using Light-Field Cameras,” in *2013 IEEE International Conference on Computer Vision*, Dec. 2013, pp. 673–680. doi: 10.1109/ICCV.2013.89.
[9] H. Sheng *et al.*, “LFNAT 2023 Challenge on Light Field Depth Estimation: Methods and Results,” in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Jun. 2023, pp. 3473–3485. doi: 10.1109/CVPRW59228.2023.00350.
[10] C. Shin, H.-G. Jeon, Y. Yoon, I. S. Kweon, and S. J. Kim, “EPINET: A Fully-Convolutional Neural Network Using Epipolar Geometry for Depth from Light Field Images,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 4748–4757. doi: 10.1109/CVPR.2018.00499.
[11] T. Leistner, H. Schilling, R. Mackowiak, S. Gumhold, and C. Rother, “Learning to Think Outside the Box: Wide-Baseline Light Field Depth Estimation with EPI-Shift,” in *2019 International Conference on 3D Vision (3DV)*, Québec City, QC, Canada: IEEE, Sep. 2019, pp. 249–257. doi: 10.1109/3DV.2019.00036.
[12] Y.-J. Tsai, Y.-L. Liu, M. Ouhyoung, and Y.-Y. Chuang, “Attention-Based View Selection Networks for Light-Field Disparity Estimation,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, Art. no. 07, Apr. 2020, doi: 10.1609/aaai.v34i07.6888.
[13] “4D Light Field Benchmark.” Accessed: Jan. 23, 2024. [Online]. Available: https://lightfield-analysis.uni-konstanz.de/benchmark/table?column-type=metrics&image=median