Efficient Depth Estimation and Fast 3D Reconstruction Using Light Field Imagery

Yuhang Zheng, Heng Chen

Nanjing Institute of Engineering, Nanjing, Jiangsu 211167

Abstract: ***Three-dimensional reconstruction is one of the classical problems in computer vision, and its application area is widely used , which has been a hot spot for research in related fields. And the accuracy and speed of 3D reconstruction depends on the estimation of scene depth information. With the development of light field imaging technology, it is more and more convenient to acquire light field images, which contain four-dimensional information and are beneficial to the accurate estimation of scene depth information. The application of deep learning in light field image depth estimation improves the speed and accuracy of light field image depth estimation, and further enables the 3D reconstruction of the scene. In this paper, we study the use of light-field images combined with deep learning for scene depth estimation, and finally realize the near- field fast 3D reconstruction.***

Keywords: Light Field Image; Depth Estimation; Deep Learning; Fast 3D Reconstruction.

1. INTRODUCTION

3D reconstruction technology is widely used in people's production life, and has important applications in industrial manufacturing, cultural relic conservation, virtual reality, smart cities, autonomous driving and other fields. Therefore, the study of 3D reconstruction is enduring and branching is extensive. According to the different data types, 3D reconstruction techniques can be classified into 3D reconstruction and 2D reconstruction. The acquisition of traditional 3D data mainly uses special imaging equipment to actively emit controlled beams or electromagnetic waves to the object. To obtain scene depth information based on its flight time difference, such as laser scanning, structural light, and raster, professional imaging equipment is required, although the accuracy is high, but its cost is high, and the application range is limited. Three-dimensional reconstruction based on two-dimensional data, that is, an image, is divided into three-dimensional reproduction based on a single image and multiple images. From the current research, 3D reconstruction based on a single image is extremely challenging due to its lack of deep information, and obtaining deep information based on multiple images to achieve 3D reproduction is one of the current mainstream research directions. In summary, the core of 3D reconstruction lies in the acquisition of scene depth information. The depth of the scene is the distance between the target object and the central plane of the camera. Similar to human vision system, computer can accurately calculate the depth of the target scene by capturing the texture, occlusion, parallax and other features of the scene. Compared with traditional digital cameras, light field cameras can record the position information and direction information of light at the same time, capture the complete light field data of the scene, and extend traditional 2D images to 4D. The photo field image obtained by the photo field camera provides rich and accurate geometric information support for depth estimation, providing favorable conditions for accurate solving of depth estimations. Microlensed array photo field imaging is a new single lens 3D imaging technology that has received much attention in recent years. It has simple structure, records light position and direction information in one image, and various post-stage data processing methods, and can be widely used in 3D reconstruction and measurement, 3D measurement and recognition, virtual and augmented reality and other fields. In recent years, the light field depth estimation algorithm has significantly improved the accuracy of scene depth estimations and has received widespread attention from researchers. This paper investigates the use of light field images obtained by microlensing light field cameras combined with deep learning to estimate scene depth, further achieving rapid 3D reconstruction of near-field views. Guo and Tao [1] focused on modeling and simulation analysis of robot environmental interaction. In the medical domain, We et al. [2] explored intelligent monitoring of anesthesia depth by leveraging multimodal physiological data. Tang and Zhao [3] applied neural networks to investigate the relationship between aging population distribution and real estate market dynamics. Tu [4] introduced ProtoMind, a modeling-driven approach for NAS and SIP message sequence modeling aimed at smart regression detection. Wang [5] employed Bayesian optimization for adaptive network reconfiguration in urban delivery systems, while Meng et al. [6] conducted research on green warehousing logistics site selection and path planning based on deep learning. Wu [7] addressed fault detection and prediction in models to optimize resource usage in cloud infrastructure. Chen [8] emphasized the importance of efficient and scalable data pipelines as the core of data processing in gig economy platforms. Yuan [9] exploited GPT-4 for multimodal medical data processing in

electronic health record systems. Li and Wang [10] developed a deep learning-enhanced adaptive interface to improve accessibility in e-government platforms. Ren [11] proposed a novel approach for role-oriented dialogue summarization aimed at balancing role contributions, and also developed a feature fusion-based and complex contextual model for smoking detection [12]. Zhou [13] introduced a digital precision distribution strategy for social media content on private domain platforms in the automotive industry using a collaborative filtering model based on user behavior. Lin et al. [14] modeled multivariate degradation data with dynamic covariates under a Bayesian framework. Wang and Liang [15] applied reinforcement learning methods combining graph neural networks and self-attention mechanisms to supply chain route optimization. Zhao et al. [16] developed a CNN-Bi-GRU model for short- and long-term renewable electricity demand forecasting. Liu, Wang, and Liang [17] proposed MiM-UNet, an efficient building image segmentation network integrating state space models. Xu et al. [18] explored AI-enhanced tools for cross-cultural game design to support online character conceptualization and collaborative sketching. Finally, Tian, Wang, and Cui [19] presented an improved U-Net brain tumor image segmentation model based on the GSConv module and ECA attention mechanism. Collectively, these studies demonstrate the growing integration of machine learning, deep learning, and advanced optimization techniques across fields ranging from healthcare and logistics to urban systems and human-computer interaction.

2. SCENE DEPTH ESTIMATION METHOD

The purpose of depth estimation is to obtain the distance between the target and the camera and output a depth chart. Three-dimensional reconstruction is possible based on a depth chart. Existing light field depth estimation methods can be divided into two categories: optimization-based depth estimations and learning-based ones:

An optimized light field depth estimation method first estimates the initial depth diagram of a scene in a specific way, then refines the depth diagram using a global optimization framework or a local smoothing method. The essential characteristic of a light field image is that it has information from multiple perspectives of the target object, depending on the way it represents multiperspective information, Existing optimized light field depth estimation methods can be divided into three types: depth estimations based on multi-perspective stereo matching, depth estimATIONS based on heavy focus, and depth estimATION based on an EPI (polar plane) image.

A learning-based approach to light field depth estimation uses existing deep learning frameworks to learn models that contain depth information for a scene, and uses the powerful performance of computer GPUs to design networks to achieve depth estimations.

Due to the high dimensionality of light field data, four-dimensional light field data cannot be directly applied to existing deep learning frameworks. Therefore, in order to meet the needs of the network for the input data, it is necessary to reduce the dimension and still contain the depth relationship of the scene points. Compared to the other two ways of characterizing light field data (multi-perspective images and refocused images), The spatial geometry in the slice of Epipolar Plane Image (EPI) is more intuitive to reflect the depth of the scene, and 2D-EPI slice is more convenient as the input data of convolutional neural network. Therefore, most of the existing CNN-based light field depth estimation frameworks use light field EPI blocks (EPI-Patch) as input to the network, which can be divided into two types of network implementations according to the network's tasks: light field exploration based on classification tasks and light field estimates based on regression tasks. The light field depth estimation, based on the classification task, divides the depth label into multiple classes based on a depth range of the dataset, classifying each pixel point. Lo et al [1]. A two-channel CNN network is designed to train vertical and horizontal EPI slices, and the output is optimized with the global optimization method to obtain the final depth map, without realizing the end-to-end network structure. This method turns the problem of depth estimation into a classification problem, which works well for scenarios with small visual difference ranges, and may produce problems such as discrete output results and reduced accuracy in real scenarios where depth is continuous. Shin et al [2]. An EPI network with four direction channels is proposed. Compared with two direction channels, this preprocessing method enhances the information of view reservation. At the same time, 2 * 2 convolution kernel is used to extract EPI information, but because the convolution kernel is too small, it is easy to be affected by noise. Zhou et al. [3] A scale and direction-aware EPI block learning network is introduced, but its depth estimation is only based on local information. Tsai et al. [4]. A view selection network based on attention mechanisms is proposed, which achieves a more accurate estimation of difference values by selecting sub-aperture images with a greater impact on the depth chart by the attention module. This class of methods implements an end-to-end network of light field depth prediction and is based on a regression task, but is based only on local feature estimation, and the depth graph is susceptible to noise.

3. NETWORK STRUCTURE

According to the depth continuity of realistic scene, the depth acquisition problem of light field is treated as a regression task [2]. An end-to-end depth estimation algorithm based on EPI and focus stack image is researched and implemented, and 3D reconstruction is performed based on depth map. With Shin [2]. The network used in the algorithm is different, the design in the network to join the CBAM[5]The reinforcement learned geometric relationships between stack images and used larger convolutional nuclei to extract relative global geometric features. This effectively helped the flow of information between networks and improved the overall performance of the network, while data enhancement techniques adapted to the geometric relationships within the light field data were used to support the training of the network. The network structure is shown in Figure 1.

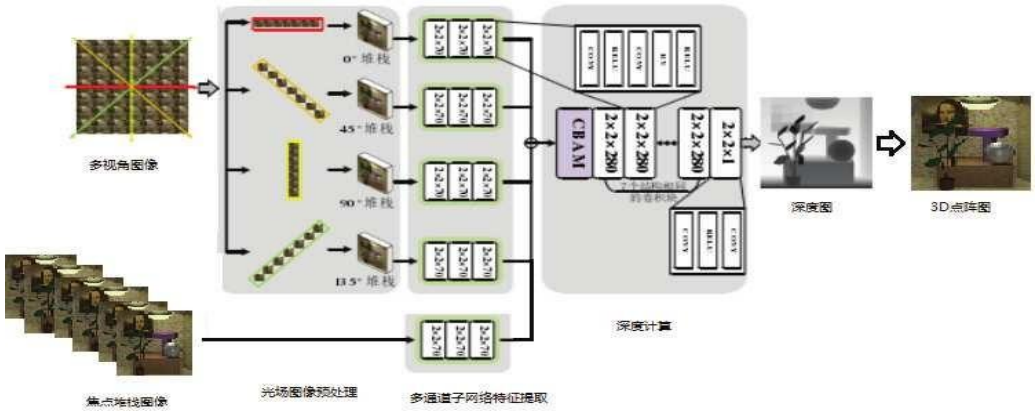


Figure 1: Graph of network structure

Four direction angles $(0°, 45°, 90°, 135°)$ the stack image as input to the network ensures both the accuracy of the calculation results and the reduction of the computing cost. Then through a multi-channel subnetwork, respectively to $0°, 45°, 90°, 135°$, the EPI blocks are used for network coding and feature extraction. Also add information from the focus stack diagram as a channel. Because a full convolutional network is capable of pixel-level feature extraction, the module consists of three full convolution blocks, which are identical in structure and consist of "Conv-ReLU-Conv-BN-ReLU" to measure pixel differences in each local EPI block. At the same time, in order to deal with the problem of short baseline and small parallax change, the step size is 1 and the size is 2 & times in the convolution block. 2 small convolution kernel for feature extraction in EPI blocks. Finally, the deep computing module is composed of three parts, the first part is a full convolutional network composed of 7 convolutional blocks with the same structure. Like the multi-channel subnetwork, each convolutional block is made up of "Conv-ReLU-Conv-BN-ReLU," which is used to learn the relationship between the features passed by the annotation model. The last part of the network consists of convolution blocks with the structure "Conv-ReLU-Conv," which are used to output parallax values with subpixel precision. Finally, a three-dimensional reconstruction is carried out with a depth estimate.

4. STANDARDS FOR NETWORK TRAINING AND EVALUATION

This network training using the server configuration for NVIDIA GeForce GTX 1080 GPU, 16GB RAM, Windows64-bit operating system, based on TensorFlow architecture implementation. The experimental dataset used is HCI Old [6] and HCI New [7] Both of them were introduced by Heidelberg image processing lab (HCI). The HCI Old dataset contains seven synthetic scenes and six real scenes, each providing 81 light field sub-aperture images and real parallax maps. In the HCI New dataset, 28 scenes were synthesized by Blender software, 24 scenes provided true parallax maps, and 4 did not. The training set of the network was selected from scenes that provided real parallax, 10 scenes were selected from the HCI Old dataset and 20 scenes from the HCI New dataset. The remaining scenarios that contain real errors are used as test sets, and the remaining the scenarios which do not contain real values are used as validation sets.

Network thinks 23 & times; The 23-size EPI image block as input is obtained by random sampling of stacked light field images. The block size is set to 16, the learning rate is 0.1 & times;10 -4 , 10000 iterations per epoch。In order to increase speed, the contraction does not replace zero during training. The loss function of the network model is the mean absolute error (MAE):

$$\varepsilon(y, y_{gt}) = \frac{1}{N}\sum_{i=1}^{N}\left|(y, y_{gt})\right| \quad (1)$$

Where N is the number of training EPI blocks , y gt is the depth label of the corresponding pixel .

Using mean error to evaluate algorithm performance:

$$MSE_{100} = \frac{\sum(d_{GT}-d)^2}{H\times W} \times 100 \quad (2)$$

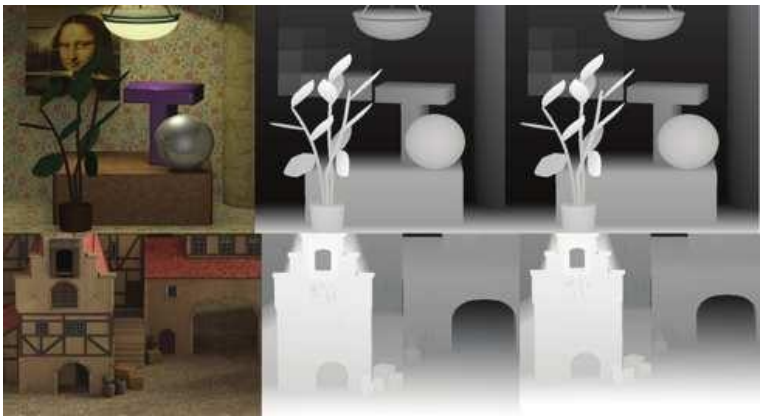
In the formula , H and W represent the height and width of the image respectively , d GT represents the real depth map , and d represents the depth estimation map .

5. EXPERIMENTAL RESULTS AND ANALYSIS

The network conducted in-depth prediction and rapid 3D reconstruction of synthetic and real-world light field images respectively, and conducted qualitative and quantitative analysis of the experimental results.

5.1 Qualitative analysis

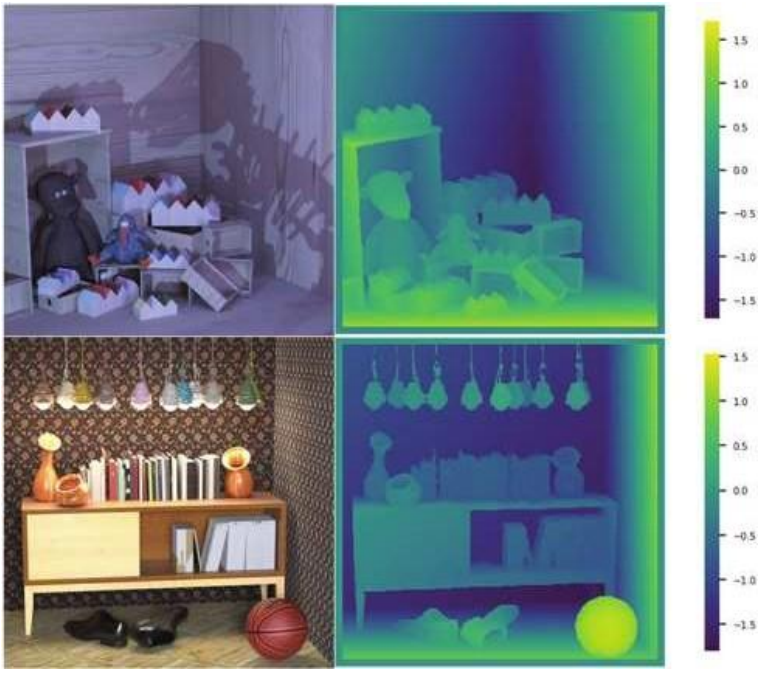
The comparison of the network depth estimation results with the real scene depth on some HCIOld datasets is shown in Figure 2.



Center view image (a) Method of this paper (b) True depth value

Figure 2: Some HCIOld datasets compare depth estimation results with real scene depth

This method is primarily aimed at near-field 3D reconstruction, so the resulting image is selected as a near-field image. As can be seen from Figure 3, this paper's depth estimation method works well with near-field images. The contrast of luminance is close to that of the real scene.



Center View Image (a) Depth Estimation Graph

(b) Point cloud map front view (c) Point cloud map side view

Figure 3: Network depth estimation map and its 3D reconstruction point cloud map on the HCINew dataset

Two close-ups, Dino and Sideboard depth estimates on the HCINew dataset are shown in Figure 3 (a). Wherein (b) is a frontal view of a three-dimensional point cloud map, and (c) is a side view. The depth estimation diagram shows that the relative edges of the objects in the scene are clearer. The generated point cloud images (b) and (c) have strong stereo sense and better texture processing. However, in this article, the network does not consider blocking leads properly and has some noise effects.

5.2 Quantitative analysis

The values of the depth estimation evaluation function MSE100 in the HCINew dataset are shown in Table 4-1. Use of Lou at the same time[1]Methods and Shin[2]Methods are compared. The underlined values in the table are the optimal results of the three methods. It can be seen that for close-range images Dino and Sideboard, the proposed method can achieve better depth estimation results.

Table 1: Comparison of depth estimates MSE 100

HCI New D ata Set	Lou[1] M ethod	Shin[2] M ethod	Proposed Method
Backgammon	4.8507	3.6229	3.6578
Dino	0.8743	1.0881	0.7966
Sideboard	1.0861	1.0615	1.0402

Table 2 shows the average computational time for each image on the two datasets of the three networks. The underlined portion is the optimal result. You can see that deep learning-based and structured concise Shin[2]The method of time complexity is optimal, much higher than the traditional lou[1]Method, although this paper adds the multi-scale focusing module and the annotation mechanism module, but this method and Shin[2]The average time taken by the method is very similar.

Table 2: Comparison of Various Network Depth Estimation Time Consuming (s)

D to Set	Lou[] M ethod	Shin[] M ethod	Proposed Method
HCI Old D to Set	X	2.55	2.64
HCI New D to Set	287	1.63	1.70

6. CONCLUSION

Optical field imaging can realize multi-dimensional collection of scenes in three dimensional space with one shot, which brings new solutions to the key problems of computer vision such as three dimensional reconstruction. In particular, the application of deep learning has greatly improved the processing time and accuracy of depth estimation of light field images, which in turn makes rapid 3D reconstruction possible. This paper attempts to construct a deep learning network for optical image processing to obtain depth estimation map, and on this basis to achieve fast 3D reconstruction of close-range scene. Depth estimation is that mass and speed determine the accuracy and speed of 3D reconstruction. However, there are still some problems, such as the small size of the data set, which makes the network inadequately trained; For the mainly Lubberian surface, less consideration was given to mirror reflecting and refractive areas and blocking conditions; The next steps will be an in-depth study of these aspects.

REFERENCES

[1] Guo, Y., & Tao, D. (2025). Modeling and Simulation Analysis of Robot Environmental Interaction. Artificial Intelligence Technology Research, 2(8).
[2] We, X., Lin, S., Pruś, K., Zhu, X., Jia, X., & Du, R. (2025). Towards Intelligent Monitoring of Anesthesia Depth by Leveraging Multimodal Physiological Data. International Journal of Advance in Clinical Science Research, 4, 26–37. Retrieved from https://www.h-tsp.com/index.php/ijacsr/article/view/158
[3] Tang, Y., & Zhao, S. (2025). Research on Relationship Between Aging Population Distribution and Real Estate Market Dynamics based on Neural Networks.
[4] Tu, Tongwei. "ProtoMind: Modeling Driven NAS and SIP Message Sequence Modeling for Smart Regression Detection." (2025).
[5] Wang, J. (2025). Bayesian Optimization for Adaptive Network Reconfiguration in Urban Delivery Systems.
[6] Meng, Q., Wang, J., He, J., & Zhao, S. (2025). Research on Green Warehousing Logistics Site Selection Optimization and Path Planning based on Deep Learning.
[7] Wu, W. (2025). Fault Detection and Prediction in Models: Optimizing Resource Usage in Cloud Infrastructure.
[8] Chen, J. (2025). Efficient and Scalable Data Pipelines: The Core of Data Processing in Gig Economy Platforms.
[9] Yuan, J. (2024). Exploiting gpt-4 for multimodal medical data processing in electronic health record systems. Preprints, December.

[10] LI, X., & Wang, Y. (2024). Deep learning-enhanced adaptive interface for improved accessibility in e-government platforms.
[11] Zheng Ren, "Balancing role contributions: a novel approach for role-oriented dialogue summarization," Proc. SPIE 13259, International Conference on Automation Control, Algorithm, and Intelligent Bionics (ACAIB 2024), 1325920 (4 September 2024); https://doi.org/10.1117/12.3039616
[12] Z. Ren, "A Novel Feature Fusion-Based and Complex Contextual Model for Smoking Detection," 2024 6th International Conference on Communications, Information System and Computer Engineering (CISCE), Guangzhou, China, 2024, pp. 1181-1185, doi: 10.1109/CISCE62493.2024.10653351.
[13] Zhou, Z. (2025, November). Digital precision distribution strategy for social media content on private domain platforms in the automotive industry: a collaborative filtering model based on user behavior. In Proceedings of the 2025 International Conference on Digital Society and Intelligent Computing (pp. 516-521).
[14] Lin, Z., Liu, X., Xiang, Y., & Hong, Y. (2025). Modeling AI End-to-End Autonomous Driving. covariates under a Bayesian framework.
[15] Wang, Y., & Liang, X. (2025). Application of Reinforcement Learning Methods Combining Graph Neural Networks and Self-Attention Mechanisms in Supply Chain Route Optimization. Sensors, 25(3), 955.
[16] Zhao, S., Xu, Z., Zhu, Z., Liang, X., Zhang, Z., & Jiang, R. (2025). Short and Long-Term Renewable Electricity Demand Forecasting Based on CNN-Bi-GRU Model. IECE Transactions on Emerging Topics in Artificial Intelligence, 2(1), 1-15.
[17] Liu, D., Wang, Z., & Liang, A. (2025). MiM-UNet: An efficient building image segmentation network integrating state space models. Alexandria Engineering Journal, 120, 648-656.
[18] Xu, Y., Shan, X., Lin, Y. S., & Wang, J. (2025). AI-Enhanced Tools for Cross-Cultural Game Design: Supporting Online Character Conceptualization and Collaborative Sketching. In International Conference on Human-Computer Interaction (pp. 429-446). Springer, Cham.
[19] Tian, Q., Wang, Z., & Cui, X. (2024). Improved Unet brain tumor image segmentation based on GSConv module and ECA attention mechanism. arXiv preprint arXiv:2409.13626.