Research On Depth Estimation and Fast 3D Reconstruction Based on Light Field Images

Yuhang Ma

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 present in people's production and life, and has important applications in industrial manufacturing, cultural relic protection, virtual reality, smart cities, autonomous driving, and other fields. Therefore, research on 3D reconstruction has been enduring and has a wide range of branches. According to the differences in data types, 3D reconstruction techniques can be roughly classified into 3D data based reconstruction and 2D data based reconstruction. The traditional acquisition of 3D data mainly utilizes some special imaging devices to actively emit controllable light beams or electromagnetic waves towards the target object, and obtain scene depth information based on their flight time differences, such as laser scanning, structured light, gratings, etc., which require professional imaging equipment. Although the accuracy is high, its cost is high and its application range is limited. The three-dimensional reconstruction based on two-dimensional data or images can be divided into single image based and multi image based three-dimensional reconstruction. From current research, 3D reconstruction based on a single image is highly challenging due to the lack of depth information. Obtaining depth information from multiple images to achieve 3D reconstruction is currently one of the mainstream research directions.

In summary, the core of 3D reconstruction lies in the acquisition of scene depth information. Scene depth refers to the distance from the target object to the center plane of the camera. Similar to the mechanism by which the human visual system perceives the depth of a target, computers can accurately calculate the depth of the target scene by capturing features such as texture, occlusion, and disparity. Compared to traditional digital cameras, light field cameras can simultaneously record the position and direction information of light, capture complete light field data of the scene, and extend traditional two-dimensional images to four-dimensional. The light field images obtained by the light field camera provide rich and accurate geometric information support for depth estimation, providing favorable conditions for accurate solution of depth estimation. Micro lens array light field imaging is a new type of single lens 3D imaging technology that has attracted much attention in recent years. It has the characteristics of simple structure, recording the position and direction information of light rays in one imaging, and diverse post-processing methods. It can be widely used in fields such as 3D reconstruction and measurement, 3D measurement and recognition, virtual and augmented reality, etc. In recent years, light field depth estimation algorithms have significantly improved the accuracy of scene depth estimation and have received widespread attention from researchers. This article studies the use of micro lens light field cameras to obtain light field images combined with deep learning for scene depth estimation, further achieving fast 3D reconstruction of close range.

Yang et al. (2024) conducted research on large scene adaptive feature extraction using deep learning, demonstrating its effectiveness in handling complex visual environments[1]. In natural language processing, Zheng et al. (2024) performed a comparative study of advanced pre-trained models for named entity recognition, highlighting their superior performance in text analysis tasks[2]. Xu et al. (2025) developed AI-enhanced tools for cross-cultural game design, facilitating online character conceptualization and collaborative sketching, which underscores the potential of AI in creative industries[3]. In the field of computer vision, Lyu et al. (2024)

optimized convolutional neural networks (CNNs) for rapid 3D point cloud object recognition, improving efficiency in real-time applications[4]. In supply chain management, Wang and Liang (2025) applied reinforcement learning methods combining graph neural networks and self-attention mechanisms to optimize supply chain routes, showcasing the effectiveness of AI in logistics[5]. Jin et al. (2025) introduced RankFlow, a multi-role collaborative reranking workflow utilizing large language models, which enhances efficiency in information retrieval and ranking tasks[6]. Additionally, Xie et al. (2025) proposed RTop-K, an ultra-fast row-wise top-K selection method for neural network acceleration on GPUs, addressing computational challenges in deep learning[7].

2. DEPTH ESTIMATION METHOD FOR SCENES

The purpose of depth estimation is to obtain the distance between the target and the camera and output a depth map. 3D reconstruction can be performed based on depth maps. The existing depth estimation methods for light fields can be divided into two categories: optimization based depth estimation methods and learning based depth estimation methods:

The optimization based light field depth estimation method first estimates the initial depth map of the scene in a specific way, and then uses a global optimization framework or local smoothing method to refine the depth map.

The essential feature of light field images is that they contain information from multiple perspectives of the target object. Based on the different ways in which the multi perspective information is represented, existing optimization based light field depth estimation methods can be divided into three types: depth estimation based on multi perspective stereo matching, depth estimation based on refocusing, and depth estimation based on EPI (polar plane) images.

The learning based light field depth estimation method utilizes existing deep learning frameworks to learn models containing scene depth information, and utilizes the powerful performance of computer GPUs to design various networks to achieve depth estimation.

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 requirements of the network for input data, it is necessary to perform dimensionality reduction processing, and the depth relationship of scene space points should still be included after processing. Compared to the other two representation methods of light field data (multi view image and refocused image), the spatial geometric characteristics in the epipolar plane image (EPI) slice more intuitively reflect the depth information of the scene. Only the slope of the corresponding diagonal line in EPI needs to be calculated to obtain the depth information of the scene, and 2D-EPI slice is more convenient as input data for convolutional neural networks. Therefore, most of the existing CNN based light field depth estimation frameworks use EPI Patch as the input of the network, which can be divided into two types of network implementation methods according to the task of the network: light field depth estimation based on classification tasks and light field depth estimation based on regression tasks. Light field depth estimation based on classification tasks divides depth labels into multiple classes according to the depth range of the dataset, and classifies each pixel point. Luo et al. [1] designed two CNN networks to train vertical and horizontal EPI slices, combined with global optimization methods to optimize the output and obtain the final depth map, without implementing an end-to-end network structure. This method transforms the depth estimation problem into a classification problem, which performs well in scenes with small disparity ranges. However, in real scenes with continuous depth, it may result in output discretization and reduced accuracy. Shin et al [2] proposed an EPI network with four directional channels, which enhances the preservation of viewpoint information compared to the selection of two directions. Simultaneously using 2*2 convolution kernels to extract EPI information, but due to the small size of the convolution kernel, it is easily affected by noise. Zhou et al. [3] introduced a scale and direction aware EPI block learning network, but its depth estimation is only based on local information. Tsai et al. [4] proposed a view selection network based on attention mechanism, which selects sub aperture images that have a greater impact on depth maps through attention modules, achieving more accurate estimation of difference values. This type of method implements an end-to-end light field depth prediction network and is based on regression tasks, but only relies on local feature estimation, and the depth map is easily affected by noise.

3. NETWORK STRUCTURE

Based on the depth continuity of real-world scenarios, this paper treats the problem of obtaining light field depth as a regression task, and studies and implements an end-to-end light field depth estimation algorithm based on EPI and focus stack images on the basis of Shin [2] algorithm, and performs 3D reconstruction based on depth maps. Unlike the network used in the Shin [2] algorithm, this design incorporates an attention mechanism module CBAM [5] to enhance the learning of geometric relationships between stacked images, and uses larger convolutional kernels to extract relative global geometric features, effectively facilitating information flow between networks and improving overall network performance. At the same time, data augmentation techniques that fit the internal geometric relationships of light field data are used to support network training. The network structure is shown in Figure 1.

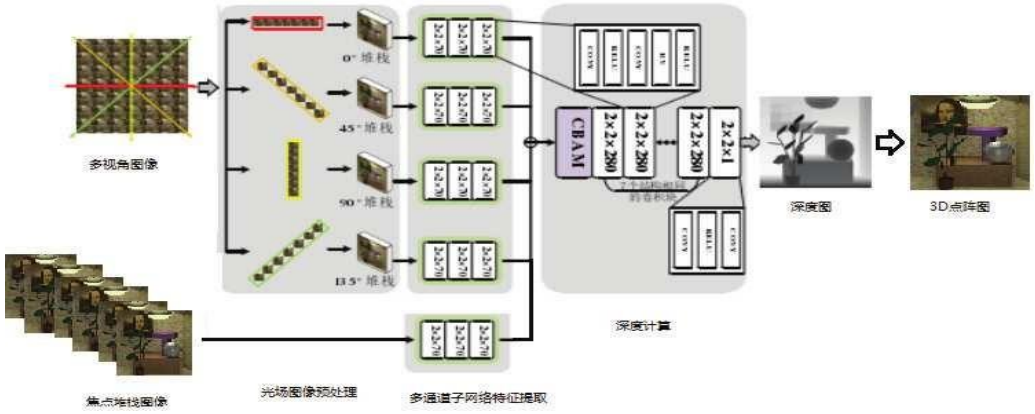


Figure 1: Network Structure Diagram

This article uses stacked images of the four directional angles of the light field image as the input of the network, which ensures the accuracy of the calculation results and reduces the computational cost. Then, through a multi-channel sub network, network encoding and feature extraction are performed on each EPI block separately. Simultaneously add information from the focus stack graph as a channel. Due to the fact that fully convolutional networks can achieve pixel level feature extraction, this module consists of three fully convolutional blocks with identical structures, namely "Conv ReLU Conv BN ReLU", used to measure pixel differences in each local EPI block. To address the issues of short baseline and small disparity changes in light field images, a small convolution kernel with a stride of 1 and a size of 2×2 is used in the convolution block for feature extraction in the EPI block. Finally, there is the deep computing module, which consists of three parts. The first part is a fully convolutional network consisting of seven identical convolutional blocks. Like multi-channel subnetworks, each convolutional block is composed of "Conv ReLU Conv BN ReLU", which is used to learn the relationships between the features conveyed by the attention model. The final part of the network consists of convolutional blocks with a structure of "Conv ReLU Conv", used to output sub-pixel level accuracy disparity values. Finally, 3D reconstruction is performed using depth estimation maps.

4. NETWORK TRAINING AND EVALUATION STANDARDS

The server configuration used for this network training is NVIDIA GeForce GTX 1080 GPU, 16GB RAM, Windows 64 bit operating system, implemented based on TensorFlow architecture. The experimental dataset used is the HCI Old [6] and HCI New [7] light field datasets, both of which were developed by the Heidelberg Image Processing Laboratory (HCI) in Germany. The HCI Old dataset contains 7 synthetic scenes and 6 real scenes, each providing 81 sub aperture images of light fields and real disparity maps. The HCI New dataset has 28 scenes synthesized by Blender software, 24 scenes provide real disparity maps, and 4 scenes do not provide them. The training set for this network was selected from scenes that provide real disparity, with 10 scenes selected from the HCI Old dataset and 20 scenes selected from the HCI New dataset. The remaining scenes with real disparity are used as the test set, while the remaining scenes without real disparity are used as the validation set.

The network takes 23×23 EPI image blocks as input, which are randomly sampled from stacked light field images. Set the block size to 16, the learning rate to 0.1×10^{-4}, and iterate 10000 times per epoch. To improve speed, convolutions are not padded with zeros during training. The loss function of the network model is the Mean Absolute Error (MAE):

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

In the formula, N represents the number of EPI blocks trained, y_{gt} Provide depth labels for corresponding pixels.

Evaluate algorithm performance using mean square error:

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

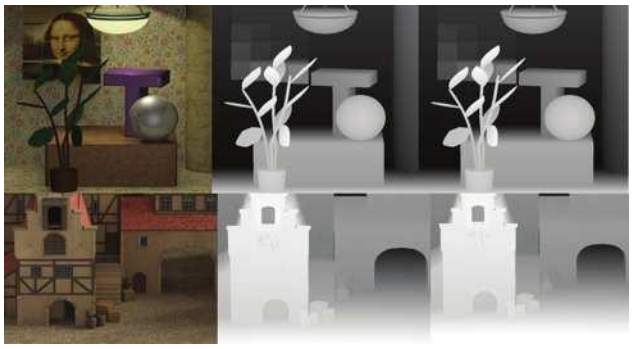
In the above equation, H and W respectively represent the height and width of the image, d_{GT} represents the true depth map, and d represents the depth estimation map.

5. EXPERIMENTAL RESULTS AND ANALYSIS

This article conducts depth prediction and rapid 3D reconstruction on synthetic light field images and real scene light field images, and qualitatively and quantitatively analyzes the experimental results.

5.1 Qualitative analysis

The comparison between the depth estimation results of our network on some HCI Old datasets and the depth of real scenes is shown in Figure 2.



Central perspective image (a) This article's method (b) True depth value

Figure 2: Comparison of Depth Estimation Results between Partial HCI Old Dataset and Real Scene Depth

This method is mainly aimed at close range 3D reconstruction, so the result image is selected as the close range image. From Figure 3, it can be seen that the depth estimation method proposed in this paper performs well on close range images. Due to the adoption of multi-scale input and attention mechanisms, except for the influence of some noise, the basic contour is clear, and the brightness contrast of the depth map is close to the depth value of the real scene.

Central perspective image (a) Depth estimation map

(b) Point cloud diagram front view (c) Side view of point cloud map

Figure 3: Network depth estimation map and 3D reconstruction point cloud map of this article on HCI New dataset

The depth estimation results of Dino and Sideboard on two close-up images of the HCI New dataset are shown in Figure 3 (a). Among them, (b) is the front view of the three-dimensional point cloud map, and (c) is the side view. In the depth estimation map, it can be seen that the relative position edges of objects in the scene are relatively clear. The generated point cloud maps (b) and (c) have strong stereoscopic effects and good texture processing. However, this article lacks consideration for occlusion clues in the network, and there are some noise effects.

5.2 Quantitative analysis

The MSE100 values of the deep estimation evaluation function in the HCI New dataset are shown in Table 1. Simultaneously compare Lou [1] method and Shin [2] method. The underlined values in the table represent the best results among the three methods. It can be seen that for close-up images Dino and Sideboard, our method can achieve good depth estimation results.

Table 1: Comparison of Depth Estimation MSE_{100}

HCI New Dataset	Lou[1]Method	Shin[2]Method	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 computation time per image for three types of networks on two datasets. The underlined part represents the optimal result. It can be seen that the concise Shin [2] method based on deep learning and structure has the best time complexity, far higher than the traditional Lou [1] method. Although this paper adds a multi-scale focusing module and an attention mechanism module, the average time used in this paper's method is very close to that of the Shin [2] method.

Table 2: Comparison of Time consumption for Depth Estimation of Various Networks (Unit: S）

Dataset	Lou[1]Method	Shin[2]Method	Proposed Method
HCI Old Dataset	X	2.55	2.64
HCI New Dataset	287	1.63	1.70

6. CONCLUSION

Light field imaging enables multi-dimensional acquisition of scenes in three-dimensional space with a single shot, bringing new solutions to key computer vision problems such as 3D reconstruction. Especially with the application of deep learning, the processing time and accuracy of depth estimation for light field images have been greatly improved, making fast 3D reconstruction possible. This article attempts to construct a deep learning network for processing light field images to obtain depth estimation maps, and based on this, achieve fast 3D reconstruction of close range scenes. Depth estimation determines the accuracy and speed of 3D reconstruction based on quality and speed. However, there are still some issues at present, such as the small dataset leading to insufficient network training; For mainly Lambertian surfaces, insufficient consideration is given to specular reflection, refraction areas, and occlusion situations; The next step will be to conduct in-depth research on these aspects.

REFERENCES

[1] Yang, Y., Li, I., Sang, N., Liu, L., Tang, X., & Tian, Q. (2024, September). Research on Large Scene Adaptive Feature Extraction Based on Deep Learning. In Proceedings of the 2024 7th International Conference on Computer Information Science and Artificial Intelligence (pp. 678-683).

[2] Zheng, Z., Cang, Y., Yang, W., Tian, Q., & Sun, D. (2024). Named entity recognition: A comparative study of advanced pre-trained model. Journal of Computer Technology and Software, 3(5).

[3] 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.

[4] Lyu, T., Gu, D., Chen, P., Jiang, Y., Zhang, Z., Pang, H., ... & Dong, Y. (2024). Optimized CNNs for Rapid 3D Point Cloud Object Recognition. arXiv preprint arXiv:2412.02855.

[5] 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.

[6] Jin, C., Peng, H., Zhang, A., Chen, N., Zhao, J., Xie, X., ... & Metaxas, D. N. (2025). RankFlow: A Multi-Role Collaborative Reranking Workflow Utilizing Large Language Models. arXiv preprint arXiv:2502.00709.

[7] Xie, X., Luo, Y., Peng, H., & Ding, C. RTop-K: Ultra-Fast Row-Wise Top-K Selection for Neural Network Acceleration on GPUs. In The Thirteenth International Conference on Learning Representations.