Proceedings of CONF-MLA 2025 Symposium: Intelligent Systems and Automation: AI Models, IoT, and Robotic Algorithms
DOI: 10.54254/2755-2721/2025.LD29547

Optimizations and Applications of Binocular Vision Technology Development by Deep Learning

Xinyu Wen

School of Information Science and Technology, Beijing University of Technology, Beijing, China
XinyuWen@emails.bjut.edu.cn

Abstract. Traditional binocular vision technology exhibits poor robustness in weak texture, illumination change, and occluded scenes, along with high computational complexity, making it difficult to meet real-time requirements. The rapid development of deep learning has brought revolutionary breakthroughs to this field. This paper systematically reviews the optimization methods and application progress of deep learning-driven binocular vision technology. Its purpose is to analyze the core foundations of binocular vision and deep learning, clarify the key optimization directions of the technology, and provide references for its practical application and future development. The research starts with the core technical foundations of binocular vision and deep learning, then explores two key dimensions: stereo matching accuracy optimization and inference efficiency optimization. It also summarizes the application of the technology in fields such as industrial manufacturing, autonomous driving, and robotics. The results show that deep learning has realized a paradigm shift in binocular vision from passively following manual rules to actively learning scene geometry, significantly improving its accuracy, robustness, and efficiency. The paper also points out shortcomings such as insufficient discussion on model generalization and quantitative comparison of computational complexity, and prospects for future research directions like self-supervised learning and multi-sensor fusion.

Keywords: Deep Learning, Binocular Vision, Stereoscopic Matching

1. Introduction

As an important branch of computer vision, binocular vision simulates the mechanism of human binocular disparity perception. It uses two cameras to capture images of the same scene from different perspectives, calculates the disparity between corresponding points to obtain depth information, and then achieves the reconstruction of the 3D scene. However, traditional binocular vision methods exhibit poor robustness in weak texture areas, lighting changing areas, and occluded scenes, and also have high computational complexity, making it difficult to meet real-time requirements. In recent years, the rapid development of deep learning has brought revolutionary breakthroughs to binocular vision technology. Through end-to-end deep neural networks, binocular vision technology has achieved full process optimization from feature extraction, cost matching, and disparity calculation, significantly improving the accuracy, robustness, and efficiency of stereo matching.

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2025.LD29547

This review systematically summarizes some methods and applications of deep learning applied to binocular vision technology. Firstly, the core foundations of binocular vision and deep learning are analyzed. Then, key technological breakthroughs are explored in depth from the perspectives of optimizing stereo matching accuracy and inference efficiency. Finally, different fields of application are outlined, and future development trends are discussed. By sorting out the optimization methods of deep learning for binocular vision technology in terms of accuracy, robustness and efficiency, the pain points of traditional binocular vision technology have been solved, and the intelligent upgrade of multiple industries has been promoted. It also provides a reference for subsequent researchers to quickly enter this field, clarify research directions and choose technical solutions.

2. Fundamentals of core technologies in binocular vision and deep learning

2.1. Core process and limitations of traditional binocular vision methods

The core process of binocular vision technology is to first calibrate internal and external parameters through camera calibration, then search for corresponding points in left and right eye images through stereo matching, calculate disparity based on the coordinate difference of corresponding points, and finally convert disparity into depth through camera baseline and focal length.

Traditional binocular vision methods rely on manually designed features such as SIFT and SURF, which have poor robustness in weak textures, lighting changes and occluded areas, and also high computational complexity, making them unsuitable for dynamic scenes. With the development of deep learning, many new technologies have greatly improved the performance of binocular vision technology by automatically learning features and solving various problems in traditional methods, such as matching corresponding point errors and poor real-time performance.

2.2. Key supporting technologies for deep learning

Optimizing binocular vision through deep learning is a collaborative system engineering process. Traditional methods rely on manually designed features, while the core advantage of deep learning lies in its ability to automatically learn features. It focuses on differentiable feature learning, cost volume construction and 3D aggregation, and achieves a paradigm shift from passive adherence to manual rules to active learning of complex scene geometry understanding through an end-to-end training framework and carefully designed loss functions and network architecture. This comprehensively optimizes traditional methods in terms of accuracy, robustness and efficiency.

3. Optimization direction of deep learning driven binocular vision technology

3.1. Precision optimization of the stereo matching network

3.1.1. Feature extraction optimization

In the process of binocular visual feature extraction, traditional methods are constrained by limited receptive fields, insufficient information extraction, and insufficient robustness in complex scenes, resulting in limitations on subsequent matching and reconstruction effects.

In response to the problem of limited receptive field in traditional networks, which makes it difficult for decoders to achieve global modelling within channels, Feng Zongwen introduced a structure-aware self-attention mechanism [1]. This mechanism enhances feature fusion's attention to

DOI: 10.54254/2755-2721/2025.LD29547

edges and textures by expanding the receptive field and performing global attention weighting in the channel dimension, effectively improving the depth restoration detail preservation effect.

To address the issue of insufficient shallow feature extraction within the view, Feng Chengchen proposed a multi-scale residual network based on stereo attention (MSR-SAMnet) [2]. This network captures multi-scale shallow features within the view through dilated convolutions with different dilation rates using multi-scale residual blocks, and synchronously generates binocular disparity attention maps using a stereo disparity attention module to accurately match epipolar features and suppress occlusion mismatches. It was tested on a binocular dataset and compared with other binocular super-resolution methods. It showed better performance and better consistency in binocular reconstruction.

To further enhance the pertinence and effectiveness of binocular image feature extraction, Liu Xiaotian proposed a feature extraction scheme combining an attention mechanism and a dual pooling pyramid [3]. This scheme focuses on key feature areas in left and right eye images through the attention mechanism, and introduces a dual pooling pyramid structure to hierarchically aggregate features of different scales, enhancing the discriminative and integrity of features. The ablation and comparative experiments have verified the improvement of the results compared to the benchmark method, proving the effectiveness of the proposed scheme and providing high-quality and robust feature support for the subsequent stereo matching process.

3.1.2. Optimization of matching cost calculation

In order to solve the problem of mismatched corresponding points in weak texture areas and occluded areas with similar colors, Wu Kaifeng proposed the rice-shaped search arm aggregation method and the arm disparity filling method [4]. This method constructs search arms with different arm lengths based on different pixel points, performs cost aggregation on target pixel points, and then sequentially explores candidate disparity values for the four directional search arms. The optimal disparity value is selected for filling based on the type of error point, effectively improving the accuracy of the algorithm matching. Experimental verification shows that its false matching rate is significantly reduced in multi-weak-texture scenes, and its matching robustness is superior to traditional aggregation methods.

To address the issue of insufficient utilization of cost volume information in high-precision depth estimation scenarios, Zhang Ding proposed a stereo matching network that combines multi-scale cost volume fusion with the attention mechanism [5]. This network integrates different resolution cost volumes layer by layer through a step-by-step fusion strategy, expands the model receptive field to capture multi-scale scene information, and introduces an attention weight generator. Then combined with the information fusion hourglass network to integrate contextual and spatial information and refine the depth estimation results. The network was tested on the KITTI 2015 dataset, and the stereo disparity error rate was only 1.76%, leading in accuracy among similar multi-scale fusion stereo matching algorithms.

Zhao Zhiqi introduced the LightGlue deep learning network in visual technology to address the problems of high computational complexity and high mismatch rate in feature matching [6]. This network models single-view feature association and dual-view feature correspondence through attention mechanism, and combines an adaptive inference mechanism to reduce computational complexity while ensuring matching accuracy. Compared with traditional manual feature methods, this approach can more accurately establish the feature correspondence between binocular images, reduce the initial error of disparity calculation. Testing on low-light datasets, this method can

DOI: 10.54254/2755-2721/2025.LD29547

significantly improve the localization accuracy and enhance the matching robustness of binocular vision in complex scenes.

3.1.3. Parallax optimization and refinement

Traditional disparity optimization suffers from a limited receptive field of the encoder convolution and large semantic differences between high and low layer features caused by the decoder relying solely on simple skip connections. This leads to edge blurring and loss of details. Feng Zongwen et al. proposed the introduction of a dense feature fusion strategy to address the problem of traditional methods being unable to efficiently achieve cross-layer feature interaction and the loss of details caused by semantic gaps between low-level and high-level features [1]. This strategy allows the decoder to closely interact with adjacent feature layers through convolutional fusion blocks, fully utilizing contextual information, modelling local context and global semantics, and dynamically adjusting the fusion ratio of low and high-level features. After conducting experiments on the KITTI dataset and incorporating fusion strategies, the model showed better performance in both error and accuracy metrics.

3.2. Optimization of reasoning efficiency of binocular vision technology

3.2.1. Lightweight network design

Zhao Siqi proposed an improved YOLOv5s-MGEFC algorithm based on the YOLOv5s algorithm to address the issues of high computational complexity and unsuitability to dynamic scenes [7]. This algorithm optimizes the lightweight, mobile adaptability, and detection performance of YOLOv5s. The YOLOv5s-MGEFC algorithm uses the lightweight network MobileNetV3 in the Backbone model, significantly reducing the computational complexity of the model. Also, it introduces GSConv and ConvNeXt modules at specific locations in the neck network, integrates three convolution operations and uses large kernel convolution to expand the receptive field. And it integrates the ECA attention module, which directly learns channel weights after global average pooling, enabling the model to achieve a large range of accuracy improvement while maintaining lightweight.

In order to correctly eliminate redundant information and improve the feature extraction efficiency of the model, Zhu Penghao et al. proposed an improved PD-YOLOv5 personnel detection model based on the YOLOv5 model [8]. This model uses efficient and lightweight PDConv convolution modules and convolutional attention mechanism modules. By comparing with other YOLOv5 optimization algorithm models, it reduces the number of parameters while ensuring high model accuracy.

3.2.2. Reasoning acceleration technology

In response to the problem of reduced predictive performance after model lightweighting, Feng Zongwen et al. introduced a knowledge distillation strategy in the estimation algorithm [1]. It can effectively utilize the high-quality prediction results of the teacher network as the supervision target, alleviate the problem of insufficient representation ability of lightweight networks, and enhance the deep recovery ability of lightweight deep networks. This algorithm has significant advantages and good prediction performance in special areas such as complex textures and edge details.

Based on the hardware characteristics of AI chips, Liu et al. proposed an innovative TradNet lightweight real-time stereo matching algorithm through algorithm and hardware co-design [9]. This

DOI: 10.54254/2755-2721/2025.LD29547

algorithm adopts a multi-stage and unsupervised hybrid paradigm, combining the efficiency of traditional algorithms with the strong feature extraction ability of convolutional neural networks, significantly reducing computational complexity from the source of the algorithm. And the entire network of the algorithm is composed only of operators directly supported by AI chips, greatly reducing the additional overhead in the inference process. This algorithm not only ensures that the algorithm can run at extremely high efficiency on dedicated AI chips, but also achieves a high real-time inference speed of 28FPS while maintaining considerable accuracy, with a power consumption of only 0.88W.

3.2.3. Sparse and dense matching balance

Yu and Gao proposed the Fast-MVSNet framework to address the contradiction between the efficiency of sparse matching and the accuracy of dense matching in binocular vision [10]. This framework first extracts image features through an 8-layer 2D convolutional neural network (CNN), constructs a sparse 3D cost volume, and combines 3D CNN regularization to efficiently predict sparse high-resolution depth maps, solving the problem of high memory overhead of traditional dense cost volumes. Then, reuse 10 layers of small CNNs to learn local pixel depth dependencies and propagate sparse depth maps into dense depth maps. Finally, a multi-view high-dimensional feature optimization depth map extracted by differentiable Gaussian Newton layers is introduced to achieve sub-pixel level refinement. This method has an overall reconstruction error as low as 0.37mm on the DTU, Tanks, and Temples datasets, effectively balancing the efficiency and accuracy of sparse dense matching in binocular vision.

4. Typical application areas of deep learning driven binocular vision

Deep learning improves binocular vision and is a key technology for multi-domain intelligence. In industrial manufacturing, its deep stereo matching network can achieve automatic recognition and positioning of micro-level defects in precision components. In fine scenarios such as electronic component welding inspection and automotive body gap measurement, algorithms optimized by deep learning can overcome metal reflection and weak texture problems, ensuring reliable measurement. In the field of autonomous driving, lightweight stereo matching networks help vehicles obtain dense depth maps in real time. Combined with temporal information, they can accurately estimate the motion status of targets such as vehicles and pedestrians, and improve environmental adaptability in rainy, foggy, and low-light conditions at night. In the field of robotics, by using real-time stereo matching to obtain indoor 3D structures, robots can achieve obstacle avoidance and path planning. Industrial robotic arms combined with binocular vision positioning can achieve more complex workpiece recognition and finer operations.

5. Conclusion

This paper mainly explores the optimization effect and application scenarios of deep learning on binocular vision technology. Deep learning has achieved a paradigm shift in binocular vision from passive adherence to manual rules to active learning of complex scene geometry understanding through an end-to-end trainable framework, surpassing traditional methods in accuracy, robustness, and efficiency.

However, this review did not delve into the generalization problem of models when faced with significant domain differences between training and testing data, nor did it quantitatively compare

DOI: 10.54254/2755-2721/2025.LD29547

the computational complexity of all discussed methods. In addition, the exploration of emerging paradigms such as self-supervised learning is relatively preliminary.

The future research will focus more on enhancing the generalization ability of models through self-supervised and semi-supervised learning, reducing reliance on large-scale annotated data. Integrating binocular vision with other sensing modalities, such as Light Detection and Ranging and inertial navigation units, as well as emerging 3D representation technologies like neural radiation fields, is expected to achieve more robust and rich scene understanding.

References

[1] Zongwen Feng. Research on Self-Supervised Monocular Depth Estimation Algorithm Based on Video Frame Sequence [D/OL]. Shanghai: Donghua University, 2025. DOI: 10.27012/d.cnki.gdhuu.2025.001450.
[2] Chengchen Feng. Research on Stereo Image Super Resolution Based on Deep Learning [D/OL]. Nanjing: Southeast University, 2024. DOI: 10.27014/d.cnki.gdnau.2022.001893.
[3] Xiaotian Liu. Multi-scale binocular stereo matching method based on deep learning [D/OL]. Xian: Xi'an University of Electronic Science and Technology, 2024. DOI: 10.27389/d.cnki.gxadu.2022.000778.
[4] Kaifeng Wu. Research on Stereo Matching Algorithm Based on Binocular Vision [D/OL]. Harbin: Heilongjiang University, 2025. DOI: 10.27123/d.cnki.ghlju.2024.000835.
[5] Ding Zhang. Research on Stereo Matching Algorithm Based on Deep Learning and Binocular Vision [D/OL]. Chongqing: Chongqing University of Technology, 2025. DOI: 10.27753/d.cnki.gcqgx.2025.000075.
[6] Zhiqi Zhao. Research on Visual SLAM Algorithm Based on Deep Learning [D/OL]. Chengdu: University of Electronic Science and Technology, 2025. DOI: 10.27005/d.cnki.gdzku.2025.004513.
[7] Siqi Zhao. Research on Environment Perception and Path Planning Algorithm for Workshop AGV Based on Deep Learning [D/OL]. Xian: Xi'an Technological University, 2025. DOI: 10.27391/d.cnki.gxagu.2025.000076.
[8] Haopeng Zhu, Jun Li, Shiyi Zhang, et al. Dynamic Hazardous Area Intrusion Detection Method Based on Deep Learning [J]. China's Safety Production Science and Technology, 2024, 20(7): 170-178.
[9] Liu Y, Wang W, Xu X, et al. Lightweight real-time stereo matching algorithm for AI chips [J/OL]. Computer Communications, 2023, 199: 210-217. DOI: 10.1016/j.comcom.2022.06.018.
[10] Yu Z, Gao S. Fast-MVSNet: sparse-to-dense multi-view stereo with learned propagation and Gauss-Newton refinement [C/OL]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020: 1946-1955. DOI: 10.1109/CVPR42600.2020.00202.