ITC 4/54

Information Technology and Control
Vol. 54 / No. 4/ 2025
pp. 1178-1188
DOI 10.5755/j01.itc.54.4.42487

Real-time Binocular 3D Reconstruction for Enhanced Depth Perception and Virtual Reality

Received 2025/08/08 | Accepted after revision 2025/10/05

HOW TO CITE: Huang, Y. (2025). Real-time Binocular 3D Reconstruction for Enhanced Depth Perception and Virtual Reality. *Information Technology and Control*, 54(4), 1178-1188. https://doi.org/10.5755/j01.itc.54.4.42487

Real-time Binocular 3D Reconstruction for Enhanced Depth Perception and Virtual Reality

Yuhan Huang

College of Information Science and Technology, Donghua University, Shanghai, 201600, China

Corresponding author: yuhan@mail.dhu.edu.cn

Motivated by the demand for immersive VR, binocular 3D reconstruction, estimating 3D structure from two camera views is key to enhancing depth perception. In response to existing challenges such as depth ambiguity, calibration inaccuracies, and occlusion handling, this paper introduces a novel real-time binocular 3D reconstruction framework based on the integration of feature-based and learning-based process. Specifically, the proposed approach features a hybrid neural network combining ResNet-50 extractors with attention fusion, and a deep learning model using surface normal priors to refine disparity maps, thereby enhancing depth prediction in occluded regions of the inputs. Experimental results demonstrate significant improvements in the accuracy and efficiency of real-time binocular 3D reconstructions for enhanced depth perception, with the system capable of producing high-quality, detailed models in real time. This work not only addresses some key limitations of current technologies but also significantly advances the potential for more sophisticated and immersive VR applications.

KEYWORDS: Binocular 3D Reconstruction, Disparity Refinement, Depth Perception, VR Depth Perception, Virtual Reality

1. Introduction

Virtual Reality (VR) has advanced rapidly. This progress is driven by the need for more immersive and realistic experiences. Depth perception is central to this. It enables natural interaction in virtual environments. Binocular 3D reconstruction plays a pivotal role in enhancing depth perception, enabling the creation of more accurate and engaging VR worlds [14]. Human vision is inherently stereoscopic, relying on the subtle differences between the images captured by each eye to gauge depth. Drawing inspiration from this biological mechanism, binocular vision systems capture dual images from slightly different perspectives to reconstruct the three-dimensional structure of a scene. This approach not only mirrors human visual processing but also enhances the realism and interactivity of virtual environments. Recent VR systems like Meta Quest 3 and Apple Vision Pro use depth sensing for mixed reality. Their accurate depth improves object occlusion and user interaction [4].

Despite significant progress, state-of-the-art works face challenges: depth ambiguity (uncertainty in depth due to low texture), calibration inaccuracies (misalignment between cameras), occlusion handling, etc. To address these issues, lots of studies have developed various solutions, ranging from traditional feature-based approaches, dense matching techniques to that learning-based frameworks and hybrid systems. Each of these strategies brings unique strengths and limitations to the table.

Feature-based methods focus on extracting and matching distinctive features between stereo images to achieve accurate reconstructions [5, 20]. Chen et al. [1] developed an optimized feature point matching algorithm for binocular vision that boosts the accuracy and efficiency of 3D reconstruction, especially under occlusion and varying lighting conditions. Feature-based approaches offer high accuracy at the cost of computational intensity [6]. Li et al. [7] introduced a binocular line laser system that resolves single-line laser multi-angle positioning issues through the use of a spherical 3D target for precise homologous point pair matching. Wang et al. [17] reviewed vision-based 3D reconstruction methods, comparing active and passive techniques like structured light and laser rangefinders, and discussed their suitability for various applications. In summary, feature-based approaches offer high accuracy at the cost of computational intensity, while the binocular line laser system excels in precision for specific applications but lacks versatility, and active methods, like those using structured light, involve higher setup costs, while passive methods are more dependent on environmental conditions. Recent advancements on optimized algorithms and innovative systems like binocular line laser setups, continue to enhance their performance and applicability across diverse 3D reconstruction tasks.

Dense matching techniques aim to explore pixel-wise correspondence, which is crucial for high-quality 3D reconstruction. These methods often employ algorithms such as Semi-Global Matching (SGM) or Belief Propagation (BP) to achieve their goals [8]. SGM checks multiple paths to reduce errors. BP updates match probabilities iteratively for better results. Zhang et al. [21] proposed a CNN-based dense matching technique that leverages the powerful representation capabilities of deep learning to directly predict disparity maps from images without the need for traditional feature extraction and matching processes. This approach has shown significant improvements in handling textureless areas and providing more accurate depth estimations. Similarly, Li et al. [9] introduced a method that combines local and global optimization [22], aiming to enhance both the accuracy and completeness of the reconstructed models. Ye et al. [19], on the other hand, utilized the Block Matching (BM) algorithm for stereo matching and established a triangulation model. This method is simple to implement and manages to keep errors within 5%, making it highly feasible for practical applications. They offer high-quality pixel-wise correspondence, which is essential for accurate and detailed 3D reconstruction. However, the accuracy of the triangulation model heavily relies on the precision of camera calibration and is susceptible to environmental factors [10]. while they have advanced the binocular 3D reconstruction with their completeness and accuracy, they still face challenges related to noise sensitivity and environmental dependencies, which may significantly affect performance in real-world applications. Therefore, while delivering superior results in controlled environments, their practicality in uncalibrated settings remains a challenge.

As for learning-based approaches, leveraging deep learning to predict disparity maps directly from stereo images, these methods have gained popularity due to their ability to handle complex scenes with little pre-processing [23]. Sun et al. [15] broke through traditional camera calibration methods by using a BP neural network, collecting datasets for training without needing to establish a binocular vision model beforehand. Neural networks provide strong robustness for high-speed camera calibration, though dataset training can be time-consuming. Zhao et al. [24] used epipolar constraints, NCC similarity measures, and disparity gradient compatibility to enhance stability and robustness in sparse feature-based stereo matching [11]. Li et al. [16] employed a Vision Transformer (ViT) for image classification, utilizing self-attention mechanisms to extract regional features, improving speed and accuracy [2]. However, these methods often require extensive and diverse training datasets, leading to high computational costs during training. Additionally, their performance can be sensitive to data quality and environmental conditions, and they may lack interpretability compared to traditional model-based approaches. Thus, while learning-based techniques show great promise in terms of adaptability and performance, they still face challenges related to efficiency, generalization, explainability, and potential ethical concerns such as data privacy and model bias.

Hybrid methods aim to combine two or more of the aforementioned methods to leverage the strengths of each approach while mitigating their weaknesses [10]. For instance, Meng et al. [12] proposed a passive-active hybrid binocular intelligent inspection system based on CNN for high-precision 3D reconstruction. Rabab et al. [13] introduced a new hybrid polynomial Stochastic K-Means Plus (SKMP), minimizing errors during image processing and achieving faster and more robust 3D image reconstruction compared to traditional methods. The incorporation further enhances the system's ability to process and interpret the collected data, resulting in highly detailed and reliable 3D reconstructions. hybrid methods often depend on careful tuning of multiple components and may still be sensitive to environmental factors such as lighting conditions and scene texture. Therefore, while hybrid approaches represent a promising direction for high-precision 3D reconstruction, they require balanced design considerations to ensure practical applicability and scalability.

Each of these approaches presents unique strengths and challenges. For instance, feature-based methods are fast but less accurate in texture-less regions, while dense matching provides high accuracy but at a computational cost. Learning-based methods offer flexibility and adaptability but require extensive training data. Hybrid method usually provides a compromise strategy but it may still face the challenges in real-time applications. This paper proposes a novel method that combines the efficiency of feature-based approaches with the robustness of learning-based models. The three main contributions of this paper are listed as follow.

1 ***Real-time Binocular 3D Reconstruction Framework.*** It is designed to capture and process visual data for enhancing depth perception and 3D Reconstruction. With a focus on the feature extraction, matching, depth calculation, and 3D reconstruction modules, the framework utilizes advanced convolutional layers, regression processing, and a stacked hourglass network to iteratively refine depth estimates. This results in a highly efficient system capable of producing high-quality 3D models in real time, thus significantly advancing the 3D reconstruction.

2 ***Deep Learning Model Incorporating Local Geometric Priors for Occluded Regions.*** It incorporates local geometric priors to refine disparity maps, which aims at enhancing depth prediction accuracy in the occluded regions where conventional techniques often struggle. By considering local geometric information, the work can more accurately predict the depths even in complex scenes, leading to more reliable and detailed 3D reconstructions.

3 ***Feature-Learning Model for Enhanced Depth Accuracy in Textureless Areas.*** It integrates the computational efficiency of feature-based computation with robustness of learning-based processing, specifically addressing challenges encountered by traditional methods in texture-poor regions. By synergizing these strengths, the work achieves improved depth accuracy.

2. 3D Reconstruction Framework

2.1. Problem Statement

The proposed real-time binocular 3D reconstruction framework is a sophisticated system designed to capture and process visual data for the purpose of enhancing depth perception and creating immersive virtual reality environments. The work aims to address depth ambiguity, occlusion, and textureless regions in real-time (operating at 27–30 FPS) binocular 3D reconstruction. Figure 1 illustrates the process of the binocular 3D reconstruction task, involving camera calibration, image collection, stereo correction, feature extraction, feature math, depth calculation, and final 3D reconstruction, where the set of $C_{parameter}$ represents the camera parameters, F_{map} is the feature map matrix, $P_{distance}$ denotes the distance matching set, $D_{distance}$ is the generated depth distance set, and I_{image} is the final constructed image. We focus on the design of latter four modules. Feature extraction and matching modules operate by identifying and extracting feature points from the left and right images captured by the binocular cameras. Advanced feature detection strategy is employed to ensure the selection of robust and distinctive points. Subsequently, an efficient feature matching is utilized to identify corresponding points between the two images, establishing the spatial relationship necessary for depth estimation. With the feature points matched, the depth calculation module takes over to compute the depth information of the scene. This is achieved by leveraging the disparity between the matched points, along with the intrinsic and extrinsic parameters of the cameras. The disparity-to-depth conversion is a critical step that allows for the extraction of accurate depth informa-

Figure 1

Process of the real-time binocular 3D reconstruction. Modules: camera calibration aligns sensors; image collection captures stereo pairs; stereo correction rectifies images; feature extraction and matching find correspondences; depth calculation computes disparity; 3D reconstruction builds the model.

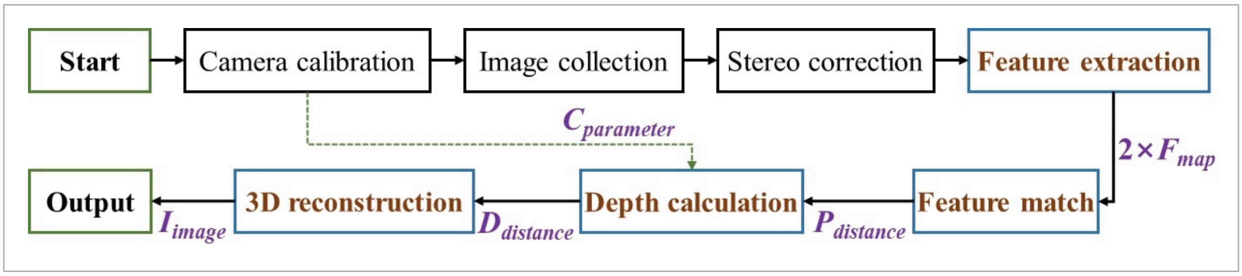
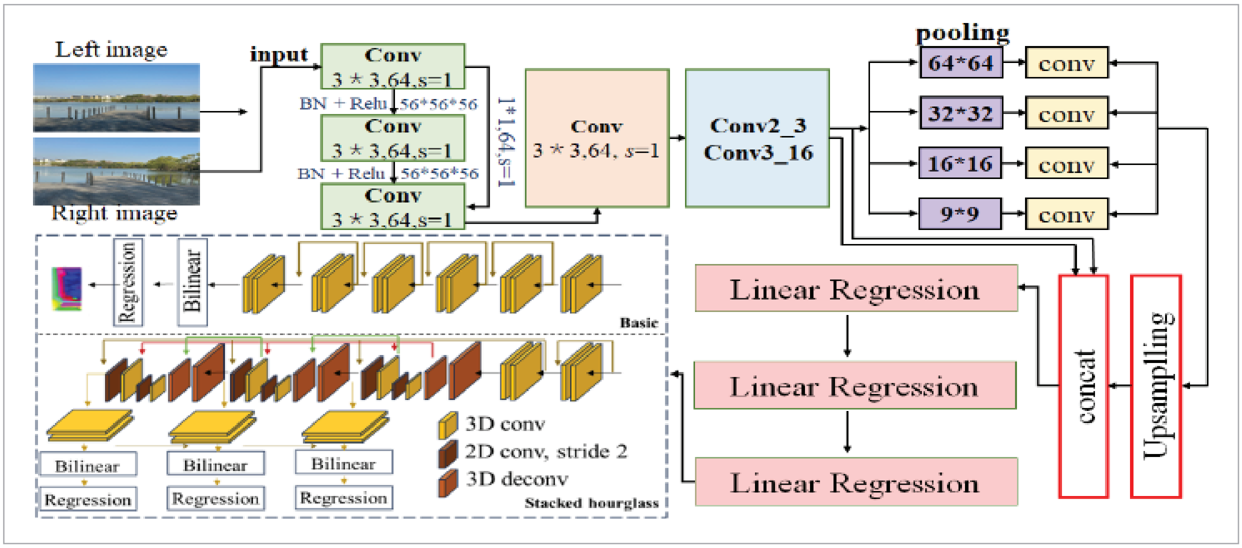


Figure 2

The detailed network structure of the proposed real-time binocular image 3D reconstruction framework.



tion for each feature point in the scene. The final 3D reconstruction module integrates the depth information with the 2D images to construct a 3D environment. The workflow culminates in the output and display of the reconstructed 3D model, providing an enhanced depth perception and a captivating virtual reality experience

2.2. Overview of the Framework

The framework is designed to efficiently handle feature extraction and matching, depth calculation, and 3D reconstruction computation, as depicted in the provided figure 2, which illustrates the network structure, showing convolutional layers, pooling stages, regression blocks, and the stacked hourglass refinement module. The process begins with two input images. First, convolutional layers extract features. Then, pooling downs-samples them. Next, regression layers estimate depth. Outputs are up-sampled and refined via bilinear interpolation. Finally, a stacked hourglass network iteratively improves depth estimates for high-quality 3D output. The two image includes a left image and a right image, which are passed through a series of convolutional layers. The first convolutional layer, Conv1, uses a 3x3 kernel with 64 filters and a stride of 1, followed by batch normalization and a ReLU activation function. This step is repeated multiple times to extract rich features from both images. The extracted feature maps then undergo further processing, including downsampling through pooling layers of varying sizes such as 64x64, 32x32, 16x16, and 9x9, followed by additional convolutional layers to refine the features. These refined feature maps are subsequently used in regression tasks, where multiple linear regression layers estimate depth information. The outputs from these regression layers are concatenated and up-sampled to match the original input image size. To ensure the constructed accuracy, real-time bilinear interpolation is applied to refine the depth map, and further regression steps are performed. The architecture also incorporates a stacked hourglass network, which iteratively refines the depth estimates operation, ensuring that the final 3D reconstruction is both detailed and precise. Through the structured and iterative strategy, the proposed framework can achieve efficient and high-quality real-time binocular 3D reconstruction, producing accurate and high-quality 3D models from the input's images.

2.3. Feature Extraction Module

The module is meticulously constructed around the ResNet architecture, a renowned and robust network that acts as the backbone for the extraction of hierarchical image features and the generation of rich feature maps. ResNet, short for Residual Network, is distinguished by its use of residual connections as a core mechanism to facilitate effective learning [18]. The process begins with the collection of images, which are then subjected to an initial training phase. Following this, the images are passed through a pooling layer, a critical step that leads to the creation of a series of residual structures. These residual structures are then activated using the Rectified Linear Unit (ReLU) function, a popular choice for introducing non-linearities into the model while maintaining computational efficiency. One of the significant advantages of leveraging residual learning is the mitigation of common issues that plague deep neural networks, such as gradient vanishing and network degradation. These issues can hinder the training process and lead to suboptimal performance. However, the improved ResNet design effectively addresses these challenges, ensuring that the network can be trained at greater depths without sacrificing performance. This design choice ensures the extraction of robust and multi-scale features from the input images. The module is capable of capturing both fine-grained details and high-level semantic information, which is crucial for a wide range of applications, from object detection to image segmentation. By incorporating advanced techniques to enhance feature extraction, the module is able to improve the overall quality of the extracted features, leading to better performance and more accurate results. The module is greatly centered around the ResNet architecture and enhanced with residual connections and advanced feature extraction techniques, which represents a significant advancement in the field of computer vision. We use ResNet-50, pretrained on ImageNet. The final classification layer is removed and replaced with regression layers for depth estimation. It is well-equipped to handle the complexities of modern image processing tasks.

2.4. Feature Matching for Disparity Estimation

The feature matching module is designed to employ the PSMNet (Pyramid Stereo Matching Network) framework to progressively estimate the disparity of each pixel in the images captured by a binocular camera system. This module begins with feature extraction from both left and right camera images using a parameter-sharing convolutional network, designed to extract multi-resolution features through downsampling and pyramid structures. To maintain the resolution of feature maps while expanding the receptive field, dilated convolutions are incorporated. These extracted features from both images are then used to construct a Cost Volume, which encapsulates the similarity information across different disparity levels between the left and right feature maps. Subsequently, 3D convolutions are applied to this Cost Volume to further extract and fuse information among the left and right feature maps at various disparity levels, resulting in a refined Cost Volume that effectively captures the relationships and disparities within the stereo pair. The refined Cost Volume is then up-sampled to the original image resolution, ensuring high fidelity in the final disparity map. By identifying the disparity values that minimize matching errors, the module accurately estimates the depth for each pixel, providing a robust foundation for subsequent 3D reconstruction processes. This approach not only enhances the accuracy of disparity estimation but also ensures computational efficiency and effectiveness in handling complex scenes.

2.5. Depth Estimation Module

This module converts the disparity values obtained from the disparity estimation module into depth distances by leveraging the camera parameters and applying a linear transformation layer. Based on the fundamental assumptions of binocular stereo vision—where the intrinsic parameters of the left and right cameras are identical, and the relative motion between the cameras is restricted to translation along the X-axis, the mathematical relationship between disparity and depth is derived. In-camera parameter and that off-camera parameter are depicted using the following two representations.

$$C_L=C_R=C_{parameter}=\begin{bmatrix} f_x & \gamma & u_0 \\ 0 & f_y & v_0 \\ 0 & 0 & 1 \end{bmatrix} \quad (1)$$

$$R_{R\text{-}L}=E\ ,t_{R\text{-}L}=\begin{bmatrix} t_x \\ t_y \\ t_z \end{bmatrix}=\begin{bmatrix} b \\ 0 \\ 0 \end{bmatrix} \quad (2)$$

where C_R and C_L are the right and left in-camera matrix, respectively. f_x and f_y are the focal lengths of the camera in the x and y directions, respectively, measured in pixels. γ is the skew parameter, and v_0 represents the coordinates of the principal point. RR→L represents the rotation matrix from the right camera coordinate system to the left camera coordinate system, and E denotes the Essential Matrix. t_R→L is the translation matrix from the right camera coordinate system to the left camera coordinate system, where $t_x=b$, $t_y=0$, $t_z=0$ indicates that the right camera in the binocular camera setup has a translation distance of b only in the x-direction relative to the left camera, b is the baseline distance. Let the coordinates of POL, POR, PL, PR be:

$$P_{OL}=\begin{bmatrix} X_L \\ Y_L \\ Z_L \end{bmatrix}, \quad P_{OR}=\begin{bmatrix} X_R \\ Y_R \\ Z_R \end{bmatrix} \quad (3)$$

$$P_L=\begin{bmatrix} u_l \\ v_L \\ 1 \end{bmatrix}, \quad P_R=\begin{bmatrix} u_R \\ v_R \\ 1 \end{bmatrix} \quad (4)$$

where POL and POR are the coordinate of point P in the left and right camera coordinate system, PL and PR are the pixel coordinates according to the pinhole imaging model, and that the PL and PR can be expressed in another way with the following representations (Equations (5) - (8)). According to the representative equation, the output of the 3D depth estimation can be obtained with Equations (9) - (10).

$$P_L=\begin{bmatrix} u_L \\ v_L \\ 1 \end{bmatrix}=K_L \frac{1}{Z_L} P_{OL}=K_L \frac{1}{Z_L}\begin{bmatrix} X_L \\ Y_L \\ Z_L \end{bmatrix} \quad (5)$$

$$P_R=\begin{bmatrix} u_R \\ v_R \\ 1 \end{bmatrix}=K_R \frac{1}{Z_R} P_{OR}=K_R \frac{1}{Z_R}\begin{bmatrix} X_R \\ Y_R \\ Z_R \end{bmatrix} \quad (6)$$

$$P_{OL}=R_{R\text{-}L}P_{OR}+t_{R\text{-}L}=EP_{OR}+t_{R\text{-}L}=P_{OR}+\begin{bmatrix} b \\ 0 \\ 0 \end{bmatrix} \quad (7)$$

$$P_L\text{-}P_R=\begin{bmatrix} u_r\text{-}u_R \\ v_L\text{-}v_R \\ 0 \end{bmatrix}=K_{\frac{1}{Z}}\begin{bmatrix} b \\ 0 \\ 0 \end{bmatrix} \tag{8}$$

$$P_{OL}=R_{R\text{-}L}P_{OR}+t_{R\text{-}L}=EP_{OR}+t_{R\text{-}L}=P_{OR}+\begin{bmatrix} b \\ 0 \\ 0 \end{bmatrix} \tag{9}$$

$$P_L\text{-}P_R=\begin{bmatrix} u_r\text{-}u_R \\ v_L\text{-}v_R \\ 0 \end{bmatrix}=K_{\frac{1}{Z}}\begin{bmatrix} b \\ 0 \\ 0 \end{bmatrix},\ d=U_L\text{-}U_r,Z=\frac{bf_x}{d} \tag{10}$$

2.6. 3D Reconstruction Module

The module ues the PSMNet [3] framework for the final stage of constructing a three-dimensional model from stereo images. After obtaining refined disparity maps through previous stages, this module integrates depth information with the original 2D images to generate detailed and accurate 3D reconstructions. Initially, the depth information derived from the disparity and depth estimation modules is mapped back onto the corresponding pixels in the left and right images. This mapping process involves transforming the disparity values into 3D coordinates based on the camera intrinsic and extrinsic parameters. Specifically, for each pixel Pl in the left image, its corresponding 3D coordinate can be calculated using the following transformation:

$$X=\frac{(u\text{-}c_x)*Z}{f},\qquad Y=\frac{\left(v\text{-}c_y\right)*Z}{f},\qquad Z=\frac{B*f}{d} \tag{11}$$

where u and v are the pixel coordinates, cx and cy are the principal point coordinates, f is the focal length, B is the baseline distance, and d is the disparity value. Subsequently, these 3D points are integrated into a coherent 3D model through a series of operations including triangulation and surface reconstruction. PSMNet is chosen for its accuracy in disparity estimation and efficient 3D cost volume processing. No modifications were made. This iterative refinement process helps in smoothing out noise and filling gaps in the reconstructed model, thereby enhancing the overall accuracy and completeness of the 3D structure. Furthermore, the module incorporates additional convolutional layers and regression techniques to further optimize the reconstructed 3D model. These steps ensure that the output is precise and computationally efficient, providing an enhanced depth perception and a reality experience.

3. Experimental Evaluation on Binocular 3D Reconstruction

The experiment involves a dataset of stereo image pairs with diverse scenes and conditions, capturing both texture-rich and - areas, varying lighting, and potential occlusions. Evaluation metrics include RMSE and MAE for depth error, and F1-Score to measure edge accuracy and object boundary preservation in reconstructed models. We conduct comparisons with existing methods, including Zhang et al. [20], Zhang et al. [21], Dong et al. [2] to provide insights into the framework's performance. Analyzing the impact of individual components through ablation studies and visual inspections of reconstructed models would further validate the framework's effectiveness in addressing depth ambiguity, calibration inaccuracies, and occlusion handling. Figures 3, 4, and 5 illustrate the performance of the proposed binocular 3D reconstruction method on three different datasets: ″community,″ ″lecythus,″ and ″guitars.″ In Figure 3, the proposed method demonstrates its capability in reconstructing a complex outdoor scene with multiple buildings and greenery. The feature generation (panel c) successfully captures the essential features of the scene, including architectural details and vegetation. The depth map (panel d) accurately represents the spatial relationships between objects, with clear distinctions between foreground and background elements. The 3D reconstruction result (panel e) shows a well-defined structure of the buildings and pathways, while the constructed result (panel f) further validates the accuracy of the reconstruction by presenting a coherent and visually plausible 3D model. Despite the complexity of the scene, the method maintains high fidelity and detail in the reconstructed output. Figure 4 focuses on a more controlled indoor environment featuring a lecythus (a type of ancient Greek vase). The feature generation (panel c) highlights the intricate patterns and textures on the vase′s surface. The depth map (panel d) provides precise depth information, capturing the curvature and contours of the vase effectively. The 3D reconstruction result (panel e) faithfully reproduces the shape and details of the vase, demonstrating the method′s ability to handle smooth surfaces and subtle variations in depth. The constructed result (panel f) confirms the accuracy of the reconstruction, showing a highly de-

Figure 3

The real-time binocular 3D reconstruction results using the data of “community” with the proposed method.

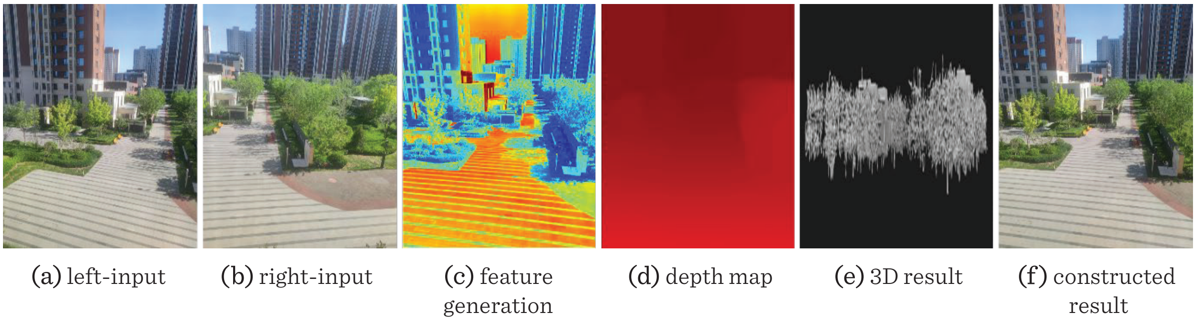

(a) left-input (b) right-input (c) feature generation (d) depth map (e) 3D result (f) constructed result

Figure 4

The real-time binocular 3D reconstruction results with the data of “lecythus” using the proposed method.

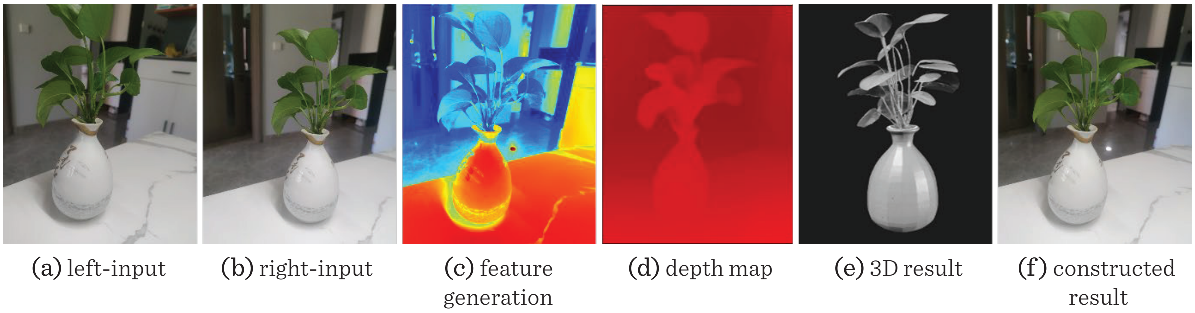

(a) left-input (b) right-input (c) feature generation (d) depth map (e) 3D result (f) constructed result

Figure 5

The real-time binocular 3D reconstruction results with the data of “guitars” using the proposed method.

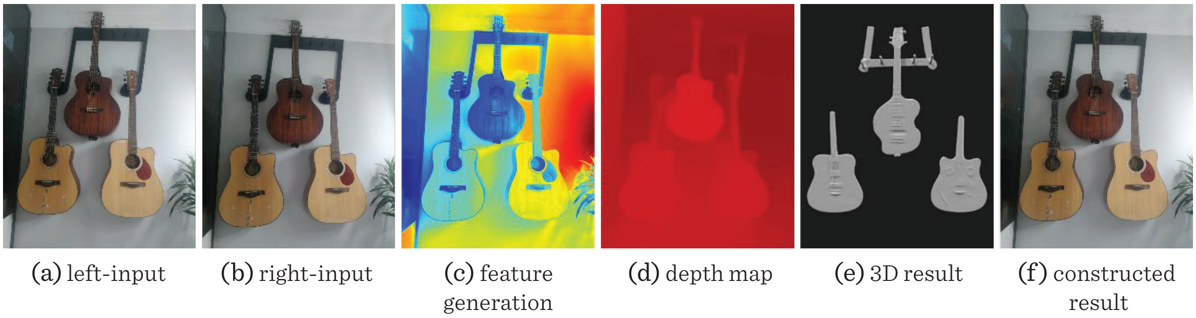

(a) left-input (b) right-input (c) feature generation (d) depth map (e) 3D result (f) constructed result

Table 1

The average results of the MSE, RMSE, F1-Score metrics form the compared methods.

Reconstruction method	MSE	RMSE	F1-Score
Zhang et al. [4]	28.365	62.917	0.846
Zhang et al. [10]	23.476	51.009	0.883
Dong et al. [20]	16.703	42.744	0.920
The proposed method	**9.552**	**31.006**	**0.951**

tailed and realistic 3D model of the lecythus. Figure 5 captures the distinct shapes and colors of each guitar, as well as the fine details such as strings and frets. The depth map (panel d) accurately distinguishes between the individual guitars and their relative positions, providing a comprehensive understanding of the scene's depth. The 3D reconstruction result (panel e) presents a clear and structured arrangement of the guitars, maintaining their proportions and spatial relationships. The constructed result (panel f) offers a complete and accurate 3D representation of the scene, showcasing the method's effectiveness in handling multiple objects with varying shapes and textures. The subjective evaluation across these three representivie diverse datasets indicates that the proposed 3D reconstructed method achieves high-quality binocular 3D reconstruction. It successfully captures detailed features, generates accurate depth maps, and produces faithful 3D models, making it suitable for vairous applications from outdoor scenes to indoor objects with intricate details.

For the objective evaluations, Table 1 illustrates the average results across three metrics. These metrics provide a quantitative assessment of the accuracy and effectiveness of each reconstruction method. The proposed method demonstrates superior performance across all three metrics when compared to existing approaches. Specifically, for the MSE metric, which measures the average squared difference between the estimated values and the actual values, the proposed method achieves a value of 9.552. This is significantly lower than the values obtained from Zhang et al. [20] (28.365), Zhang et al. [21] (23.476), and that work of Dong et al. [2] (16.703). The reduction in MSE indicates that the proposed method provides more accurate estimations with less deviation from the true values. Similarly, for the RMSE metric, which is the square root of the MSE and gives a measure of the magnitude of the error in the same units as the data, the proposed method achieves a value of 31.006. This is considerably lower than the RMSE values reported by Zhang et al. [20] (62.917), Zhang et al. [21] (51.009), and Dong et al. [2] (42.744). A lower RMSE signifies that the proposed 3D reconstruction method has a smaller average error magnitude, further emphasizing its superior performance. Lastly, the F1-Score, which is a harmonic mean of precision and recall and provides a balanced measure of the method's ability to correctly identify positive instances while minimizing false positives and false negatives, shows that the proposed method attains a score of 0.951. This is notably higher than the scores achieved by Zhang et al. [20] (0.846), Zhang et al. [21] (0.883), and Dong et al. [2] (0.920). A higher F1-Score indicates that we not only accurately identified the features but also maintained a high level of precision in its reconstructions. For the comparisons, the proposed method outperforms the compared methods in terms of MSE, RMSE, and F1-Score, demonstrating its robustness and accuracy in binocular 3D reconstruction tasks. The significant improvements in these metrics highlight the effectiveness of the proposed approach and its potential for practical applications in various fields requiring precise 3D reconstruction.

4. Conclusion

This paper presents a novel real-time binocular 3D reconstruction framework that significantly enhances depth perception. By integrating feature-based and learning-based approaches, our method addresses key limitations of current technologies, including depth ambiguity, calibration inaccuracies, and occlusion handling. The proposed framework features a hybrid model designed to improve depth accuracy in textureless areas and employs deep learning models incorporating local geometric priors to refine disparity maps, thereby enhancing depth prediction in occluded regions. Experimental results demonstrate that the proposed system is capable of producing high-quality, detailed 3D models in real time (>25 FPS). Specifically, the method achieves superior performance across multiple evaluation metrics—MSE, RMSE, and F1-Score—compared to existing methods. These improvements are particularly evident in challenging scenarios such as complex outdoor scenes with multiple buildings and greenery, controlled indoor environments featuring intricate objects like ancient Greek vases, and scenes with multiple objects of varying shapes and textures. These results were achieved under real-time constraints on consumer-grade GPUs, showing practical deployment potential.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, author-ship, and/or publication of this article.

Data Sharing Agreement

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

1. Chen, M., Liu, S. Research on 3D Reconstruction of Binocular Vision Based on Feature Point Matching. Sensors, 2023, 23(17), 7372. https://doi.org/10.3390/s23177372
2. Dong, Y., Wu, H., Chen, X., Yang, X., Xi, J. Shape-Aware Speckle Matching Network for Cross-Domain 3D Reconstruction. Neurocomputing, Jun. 7, 2024, 585, 1-12. https://doi.org/10.1016/j.neucom.2024.127617
3. Dai, Z., Geng, Y., Xue, M., Qiang, X. ZHNet: Improved Deep Stereo Matching Network Based on PSMNet. In Proceedings of the 2024 International Conference on Intelligent Perception and Computer Vision (CIPCV), Xiamen, China, 2024, 57-61. https://doi.org/10.1109/CIPCV61763.2024.00020
4. Hu, T., Zhou, F., Tan, H. Design of Single-Camera Mirrored Binocular Vision Sensor with Single-Plane Mirror. IEEE Sensors Journal, 2024, 24(21), 34326-34336.https://doi.org/10.1109/JSEN.2024.3466932
5. Lin, X., Wang, J., Lin, C. Research on 3D Reconstruction in Binocular Stereo Vision Based on Feature Point Matching Method. In Proceedings of the IEEE International Conference on Information Systems and Computer Aided Education (ICISCAE), Dalian, China, 2020, 551-556. https://doi.org/10.1109/ICISCAE51034.2020.9236889
6. Li, H., Wang, S., Zhang, Y., Bai, Z., Wang, H., Li, S., Wen, S. Research on 3D Reconstruction of Binocular Vision Based on Thermal Infrared. Sensors, 2023, 23(17), 7397.https://doi.org/10.3390/s23177372
7. Li, J., Xu, R. A Novel 3D Reconstruction Method with a Binocular-Line Laser System. Robotics and Autonomous Systems, 2024, 145, 103837.
8. Li, H., Yang, R., Wang, L. Enhanced Semi-Global Matching Algorithm for Real-Time 3D Reconstruction. ISPRS Journal of Photogrammetry and Remote Sensing, 2023, 198, 145-156.
9. Li, Z., Ma, X., Wang, J. Combining Local and Global Optimization for Efficient Stereo Matching. Journal of Visual Communication and Image Representation, 2023, 91, 103350.
10. Liu, G., Huang, H., Song, H., Qin, T., Qin, F. Binocular Multi-Line Laser Segment Stereo Matching Based on Projective Geometric Constraints and Scoring Mechanism. Measurement, 2025, 245, 116596. https://doi.org/10.1016/j.measurement.2024.116596
11. Li, X., Kuang, P. 3D-VRVT: 3D Voxel Reconstruction from a Single Image with Vision Transformer. In Proceedings of the International Conference on Culture-Oriented Science & Technology (ICCST), Beijing, China, 2021, 343-348. https://doi.org/10.1109/ICCST53801.2021.00078
12. Meng, Z., Qianqian, S. Active Passive Hybrid Binocular Intelligent Detection System for New Energy Batteries. In Proceedings of the 2023 International Symposium on Computer Science and Intelligent Control (ISCSIC), Nanjing, China, 2023, 70-73. https://doi.org/10.1109/ISCSIC60498.2023.00024
13. Rabab, O., Tahiri, M.A., Bencherqui, A., Amakdouf, H., Jamil, M.O., Jamil, Q.H. Efficient Localization and Reconstruction of 3D Objects Using the New Hybrid Squire Moment. In Proceedings of the 2022 International Conference on Intelligent Systems and Computer Vision (ISCV), Fez, Morocco, 2022, 1-8. https://doi.org/10.1109/ISCV54655.2022.9806086
14. Shuang, Y., Fan, J., Liu, T., Tan, Y. In Situ Calibration of Binocular Vision System for 3-D Measurement of Particle Size During Crystallization. IEEE Transactions on Instrumentation and Measurement, 2024, 73(5), 1-15, no. 5035915. https://doi.org/10.1109/TIM.2024.3470016
15. Sun, J., Ma, Y.Z., Yang, H., Zhu, X.L. Camera Calibration and Its Application of Binocular Stereo Vision Based on Artificial Neural Network. In Proceedings of the International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), Datong, China, 2016, 761-765. https://doi.org/10.1109/CISP-BMEI.2016.7852811
16. Song, Z., Zhu, H., Wu, Q., Wang, X., Li, H., Wang, Q. Accurate 3D Reconstruction from Circular Light Field Using CNN-LSTM. In Proceedings of the 2020

IEEE International Conference on Multimedia and Expo (ICME), London, UK, 2020, 1-6. https://doi.org/10.1109/ICME46284.2020.9102847

17. Wang, Z., Zhang, Y. A Comprehensive Review of Vision-Based 3D Reconstruction Methods. Sensors, 2024, 24(7), 2314. https://doi.org/10.3390/s24072314

18. Xiao, L., Ge, L., Su, J. A Furniture Image Style Classification Model Based on the Fusion of ResNet and Swin Transformer. In Proceedings of the 2024 International Symposium on Digital Home (ISDH), Guilin, China, 2024, 67-72. https://doi.org/10.1109/ISDH64927.2024.00018

19. Ye, Q., Cheng, Y., Zhang, M., Wang, G. Research on Flame Location and Distance Measurement Method Based on Binocular Stereo Vision. In Proceedings of the 2020 Chinese Automation Congress (CAC), Shanghai, China, 2020, 4089-4094. https://doi.org/10.1109/CAC51589.2020.9327595

20. Zhang, L., Wang, Y. Research on 3D Reconstruction Methods Based on Binocular Structured Light. Journal of Physics: Conference Series, 2021, 1744(3), 032002. https://doi.org/10.1088/1742-6596/1744/3/032002

21. Zhang, C., Wu, D., Liu, G. A Convolutional Neural Network Approach for Dense Stereo Matching. Neurocomputing, 2024, 514, 186-196.

22. Zhang, Y., Yang, F., Yuan, H., Zhang, S. 3D Reconstruction of Coal Pile Based on Visual Scanning of Bridge Crane. Measurement, 2024, 242, 116146. https://doi.org/10.1016/j.measurement.2024.116146

23. Zhao, J.M., Wei, L., Yuan, X.A., Yin, X.K., Li, X., Chen, Q.Y. An End-to-End Physics-Informed Neural Network for Defect Identification and 3-D Reconstruction Using Rotating Alternating Current Field Measurement. IEEE Transactions on Industrial Informatics, 2023, 19(7), 8340-8350. https://doi.org/10.1109/TII.2022.3217820

24. Zhao, F., Jiang, Z. A New Algorithm for Three-Dimensional Construction Based on the Robot Binocular Stereo Vision System. In Proceedings of the International Conference on Intelligent Human-Machine Systems and Cybernetics, Nanchang, China, 2012, 302-305. https://doi.org/10.1109/IHMSC.2012.168