Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

MOBI Depth: Monocular Depth Plane Decomposition Based Sparse-to-Dense Binocular Depth Estimation

Qibin He
Macao Polytechnic University
Yanming Chai
Macao Polytechnic University, Macao SAR
Haochen Li
Macao Polytechnic University, Macao SAR
Yanxiao Li
Macao Polytechnic University, Macao SAR
Junhao Wei
Macao Polytechnic University, Macao SAR
Dexing Yao
Macao Polytechnic University, Macao SAR
Yifu Zhao
Macao Polytechnic University, Macao SAR
Ka-Hou Chan
Macao Polytechnic University, Macao SAR
Sio-Kei Im
Macao Polytechnic University
Yapeng Wang
yapengwang@mpu.edu.mo

Macao Polytechnic University, Macao SAR

Article

Keywords:

Posted Date: June 2nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9593902/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

MOBI Depth: Monocular Depth Plane Decomposition Based Sparse-to-Dense Binocular Depth Estimation

Qibin He[1], Yanming Chai[1], Haochen Li[1], Yanxiao Li[1], Junhao Wei[1], Dexing Yao[1], Yifu Zhao[1], Ka-Hou Chan[1], Sio-Kei Im[2], and Yapeng Wang[1,*]

[1]Macao Polytechnic University, Faculty of applied science, Macao SAR, 999078, China
[2]Macao Polytechnic University, Macao SAR, 999078, China
[*]Corresponding author:yapengwang@mpu.edu.mo

ABSTRACT

Binocular cameras have been the go-to choice for large-scale robotics applications due to their cost-effectiveness and versatility. Recent advancements in deep learning have significantly enhanced binocular depth estimation, achieving performance comparable to Lidar under certain conditions. However, many binocular depth models rely on self-supervised learning, as high-quality labeled datasets are scarce, raising concerns about their reliability and confidence in unseen scenes. To address these challenges and adapt to different situations, we propose an innovative and explainable depth estimation method. This method starts with monocular depth estimation and extracts matched feature points from binocular images as anchors. Using a sliding window and monocular depth maps, it creates and refines a provisional plane, then calculates the real plane formula from feature point disparity to estimate depth. Experiments confirm its advantages in producing accurate dense depth maps across diverse scenes.

Introduction

Since the first visual rangefinder was invented, binocular depth estimation played an important role in engineering, surveying, astronomy, and many other areas. In the last several decades, the application of digital images and computer vision technology has broken the limitation of single target range measuring and creating the depth map for the whole images, which became the most important foundation of future applications in 3D reconstruction, augmented reality, UAVs, autonomous driving and robotics[1–3]. For more than twenty years, traditional computer vision algorithms have taken the major part of depth estimation. Methods like semi-global matching[4] for stereo vision usually regard depth estimation tasks as an optimization problem. By minimizing the energy function, they can reasonably estimate the depth for each point. However, these methods often exhibit poor performance when images lack sufficient texture and structural information, which significantly limits the effectiveness of binocular depth estimation The revolution of deep learning heavily impacted the field of binocular depth estimation[5]. From

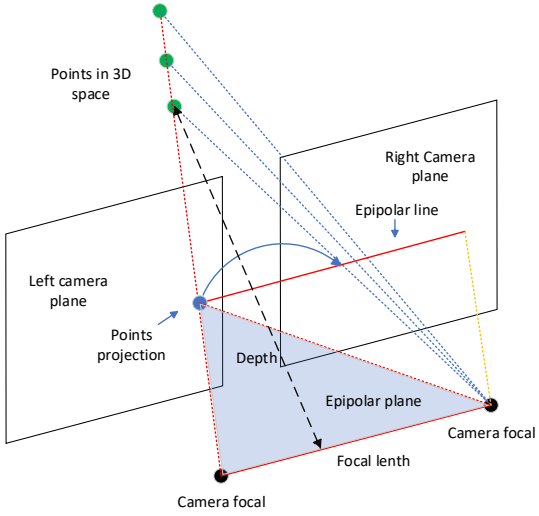


Figure 1. Binocular camera model

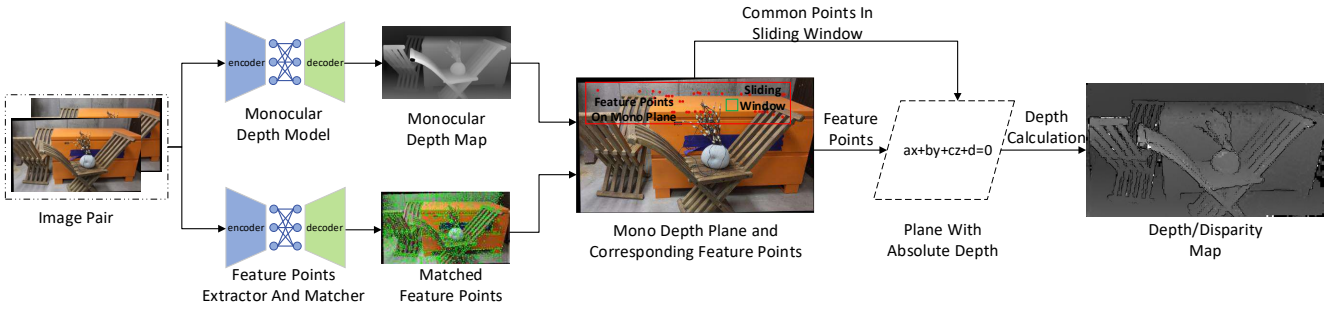


Figure 2. Overview of MOBI Depth. We first extract matched anchor points from input images and perform monocular depth estimation to create false depth maps. Then, we use a sliding window to find corresponding anchor points on the false plane. Finally, we use real planes made of anchor points to calculate true depth values.

convolutional neural network[6,7] to vision transformers[8], various deep learning architectures have been proposed in binocular depth estimation. Those trainable depth estimation models can create smooth, dense depth maps with high accuracy in most areas. Although the accuracy of deep learning methods has reached a superb level, their results cannot be fully trusted[9]. Unlike traditional methods, depth values are estimated rather than matched to calculate, which means the performances of those models depend on the quality and the scale of datasets. In real-world scenarios, the performance of depth estimation models relies more heavily on fine-tuning than other deep learning models for different tasks due to the lack of binocular camera data[10,11]. Regardless of improvements to model architecture, this limitation persists. Thus, we propose MOBI depth, a novel multi-step method that converts reliable sparse matching to the dense depth map by monocular depth. Our contributions to this field are as follows:

- A epipolar line constraint method to extract and match feature points in binocular cameras.
- A sliding-window-based method for representing irregular scenes using monocular depth decomposition of images into different planes
- An explainable method calculates absolute depth by plane formula to create the dense depth map from sparse depth values.

1 Related work

Depth estimation on monocular cameras and binocular cameras is different but related. Compared with monocular cameras, binocular cameras provide a parallel observation of the scene, which inspires our method to adjust monocular depth values with binocular cameras.

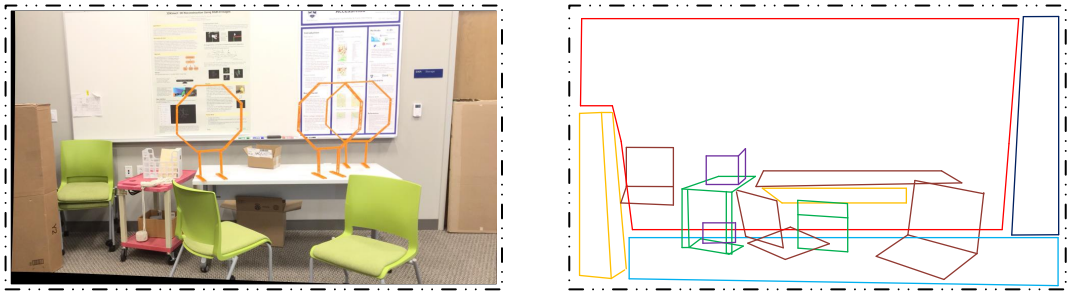
1.1 Binocular depth estimation

Traditional binocular matching-based depth estimation methods typically rely on minimizing an energy function. However, these approaches frequently fail in challenging scenarios, such as textureless regions, occlusions, and specular reflections. Although various refinements have been proposed, their performance remains unsatisfactory in robotics and other applications that demand highly accurate 3D perception. With the advent of deep learning, trainable models have significantly advanced accuracy and robustness in binocular depth estimation. In early efforts, researchers treated this problem as a supervised learning task, much like other computer vision applications[12]. Yet, the scarcity of high-quality depth annotations constrained both the size and fidelity of the available datasets, adversely affecting performance. In response, numerous recent methods employ self-supervision[13,14], wherein the network is guided by cost volumes constructed from stereo image pairs. This training strategy, however, has drawn criticism that current models may exclusively learn features pertinent to stereo matching, disregarding vital geometric cues. Consequently, geometric inconsistency remains a principal challenge in binocular depth estimation models[15].

1.2 Monocular depth estimation

For a long time, monocular depth estimation has been considered an ill problem[16,17] because multiple depth maps may have the same monocular image. However, humans can make rough judgments about near and far with one eye on their experiences and knowledge of geometry[18]. Deep learning models simulate these processes and make this problem easier[19,20].

While a single image makes depth estimation difficult, it also reduces the limitations of supervising the model. Monocular depth estimation models can be trained from binocular image pairs[21], video series[22], 3D movies[23], etc. Compared with binocular depth estimation, monocular depth estimation, though it cannot generate absolute depths, adapts to most scenes



(a) Image from one camera (b) Planes in the images

Figure 3. Image plane decomposition. Points on the same plane share the same plane formula.

because of the sufficient data from multiple sources[24,25]. Thus, we believe finding the right scale for monocular depth values can create high-quality absolute depth maps.

2 Problem Analysis

In the field of computer graphics, 3D models are usually represented by many small triangle planes. Those small planes can be determined by very few parameters, and this characteristic inspires us to use sparse matching to create dense depth maps. Figure. 3 shows the scene represented by planes. For any point on a plane, the coordinates should obey the following plane formula:

$$ax + by + cz + d = 0 \tag{1}$$

where a, b, c, d are parameters of the plane and x, y, z are coordinates of the point. Plane parameters can be determined by at least 3 points on the plane, but one plane contains far more than 3 points. According to Figure. 1, with matching relations, the depth equation is as follows:

$$z = \frac{f_m \times b_m}{d} \tag{2}$$

where z is the depth of the point, f_m is the focal length, b_m is the baseline length, d is the disparity. Once we represent the scene with planes and determine those planes' parameters, the dense depth map is created.

3 Methodology

In this section, we illustrate details of MoBi Depth, which contains three steps: Anchor point extraction, false plane expansion, and depth calculation on a real plane. In this paper, the monocular depth map is created using the Depth Anything model, and feature points are extracted by Super Point and matched by Super Glue.

3.1 Epipolar Line Constraint Sparse Matching

For the input RGB images, our method first extracts and matches their feature points using Super Point and Super Glue. We then filter these matches based on the polar line constraints of the binocular camera. For any matched pair of points $((x1, y1), (x2, y2))$, if $|y1 - y2| > \varepsilon$, we regard it as the wrong match; otherwise, we use its disparity $|x1 - x2|$ to calculate real depth value by equation 1.

Then, we perform monocular depth estimation on these images to get a monocular depth map. After this step, we can have two kinds of 3D points: 1. Common points that only have monocular depth values and pixel coordinates in the image. 2. Feature points with both monocular depth values and absolute depth values. However, we notice feature points always exist on the edges of images. Thus, we use a max pooling layer with kernel size k to process the monocular depth map to eliminate the uncertainty.

Figure. 4 shows we extract anchor points and transfer them into 3D points. Those matched feature points represent the surface structure of the image and have very high accuracy for depth values. We regard them as anchor points for further processes.

3.2 Plane Decomposition and Extension

Though we extract some anchor points with real depth, we still need to calculate depth values for other points. In this situation, the monocular depth becomes the bridge that connects the common points and the anchor points. Traditional algorithms such as SGBM treat points within a small block as having the same disparity; similarly, we consider the points within a small block as

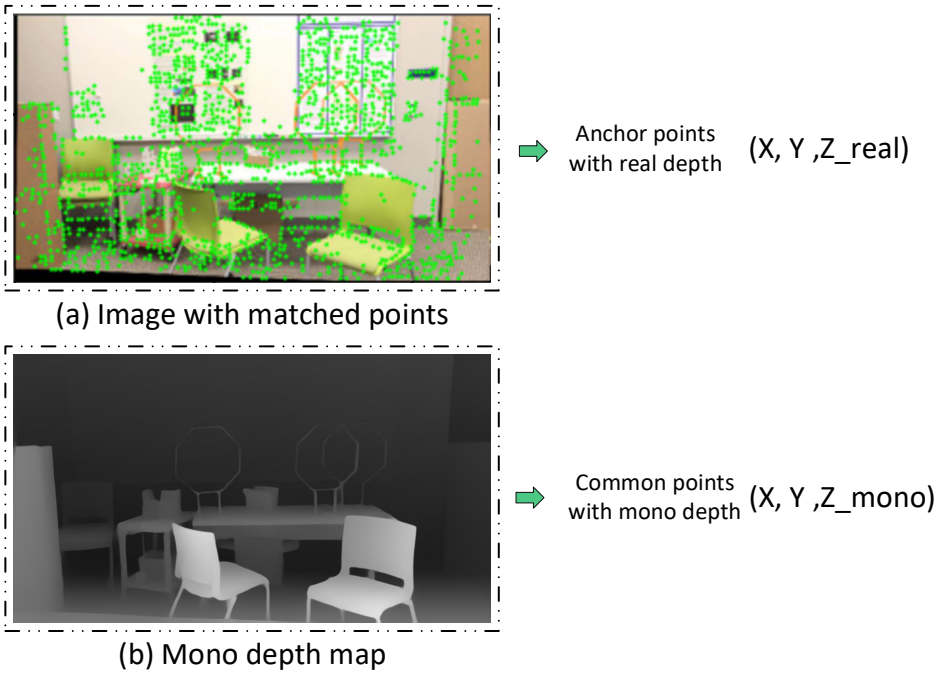


Figure 4. Anchor points extractions. X and Y are pixel coordinates. Matched feature points are sparse but can have both real and mono-depth values. Common points only have mono-depth values, but they are dense.

being on the same plane. We use a sliding window to split the whole image into small blocks. For all points inside the small block with monocular depth values, we use the RANSAC(Random Sample Consensus) method to find the possible monocular depth plane parameters and divide points into inner and outer points of the plane.

However, anchor points may not be inside this small block, but they could exist on the extended plane made of points from the block. To find the relation between this plane and anchor points, we define the distance between points and plane. For any anchor point (x_0, y_0, z_0) where x_0, y_0 are pixel coordinates, z_0 is monocular depth and plane $a_m x + b_m y + c_m z + d_m = 0$, their distance d is :

$$d = \frac{|a_m x_0 + b_m y_0 + c_m z_0 + d_m|}{\sqrt{a_m^2 + b_m^2 + c_m^2}} \tag{3}$$

We then calculate this distance for all anchor points, and when the distance is less than a pre-set threshold τ, we consider the point to be on the monocular plane. Figure. 5 shows one plane in a small block and anchor points close enough to it. Though the depth is not true, this false plane still shows good extending ability. It's very possible for these points to be on a real plane.

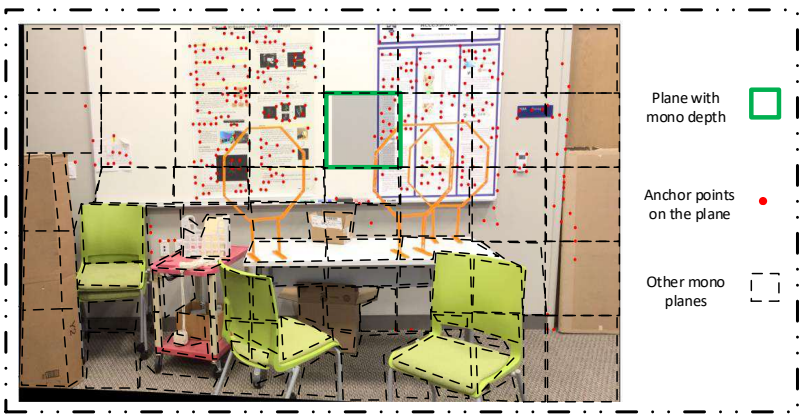


Figure 5. The plane decomposition of an image by sliding window method and corresponding anchor points.

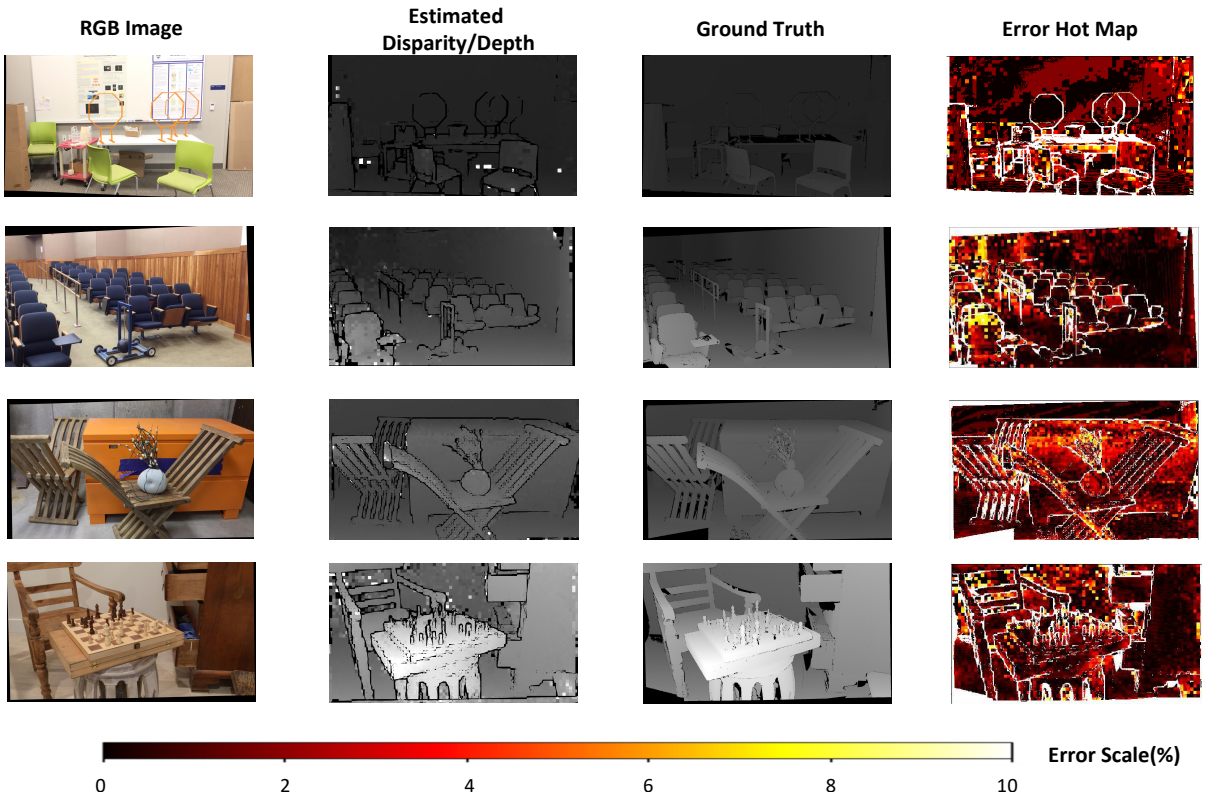


Figure 6. Our qualitative results in the Middlebury dataset. The first column is the first input image, while the estimated depth/disparity map is in the middle. Error hot maps with absolute real errors are shown on the right. For error hot maps, more red colors indicate smaller errors.

3.3 Dense Absolute Depth Calculation on Planes

After the false plane extending step, we determine the false plane and anchor points on it. Though we do not know real depth values for common points in the small block, there is a plane with real depth hidden behind those common points and anchor points, and they share the same plane equation with real depth. We cannot make sure that all corresponding anchor points are on the same plane, but most of them should be. Once again, we use the RANSAC method to find plane parameters among those anchor points by their real depth values. This plane is the most possible plane that both anchor points and common points should be on. We can calculate the real depth value z for every point in the small block and on the plane by the following equation:

$$z = \frac{-a_r x - b_r y - d_r}{c_r} \tag{4}$$

where x, y are pixel coordinates and a_r, b_r, c_r, d_r are plane parameters from anchor points with real depth values.

4 Experiments

In this section, we evaluate the performance of the proposed method on Middlebury and KITTI datasets to test the performance in indoor and outdoor environments.

4.1 Experiment Settings

Table 1 shows the parameter settings of our experiments. τ is the on-plane threshold. When the distance is less than it, we regard the point as on the plane. ε is the same line threshold; when the disparity of the axis is smaller than it, we treat matched points as in the same polar line. s is the size of the sliding window to extract the plane in a small block. This parameter influences the smoothness of the depth map. When the size is too small, the depth map becomes smoother but loses some accuracy. Conversely, accuracy improves when the size is too large, but the density decreases. Iteration times and sampling ratio are RANSAC parameters. The higher these values, the more accurate the plane equations will be, increasing the computation time. Two feature point extraction and matching thresholds affect the quality and number of anchor points.

Table 1. Parameter settings

τ (on plane threshold)	ε (same line threshold)	s (sliding window size)	Iteration times (RANSAC)
1.0	5	30×30	200
SuperPoint confidence	SuperGlue confidence	k (max pooling kernel size)	Sampling Ratio (RANSAC)
0.0015	0.01	5×5	0.1

Table 2. Performances of different methods in Middlebury Dataset

Methods	Bad1.0(%) (all)	Bad2.0(%) (all)	Bad4.0(%) (all)	Avgerr (all)
MC-CNN-acrt[12]	17.1	**8.08**	4.91	3.82
RDNet[27]	26.5	11.3	5.57	3.34
PCVNet[28]	19.5	8.19	**3.71**	**1.53**
RAFT-Stereo[6]	15.1	9.37	6.42	2.71
SGBM[4]	36.0	23.8	18.5	15.9
Ours	**14.8**	8.6	3.9	2.82

4.2 Results in Middlebury Dataset

Middlebury dataset[26] is an indoor dataset consisting of hundreds of stereo images with true depth values. In this section, we compare the performance of the approach with other popular depth estimation methods. Table 2 shows the comparison results of different methods. Our method is comparable in most sections compared with other trainable pure deep learning methods. In bad1.0, our method has the best performance. We believe this is because our process has solid geometry support from matched feature points, constraining the error to a minimal range.

We also noticed that in Bad 2.0 and Bad4.0, the performances were not the best. We guess this is also due to the unique depth calculation method from the plane equation because, for most points in the scene, there is a distance from matched points, and the broadcasting process may cause some minor errors.

4.3 Results in KITTI Dataset

KITTI dataset[29,30] is an outdoor dataset of thousands of stereo images with half auto-labeled depth values. We compare the results with several methods to test the performance in the outdoor environment.

Table 3 shows the results of different methods. Compared with other approaches, our method does not have an advantage.

Table 3. Performances of different methods in KITTI Dataset

Methods	D1(%) (bg)	D1(%) (fg)	D1(%) (all)	Density(%)	Runtime (s)
MC-CNN-acrt[12]	2.89	8.88	3.89	100	67
PCVNet[28]	1.68	3.19	**1.93**	100	**0.05**
PGNet[31]	1.64	3.60	1.96	100	0.7
SGNet[32]	**1.63**	3.76	1.99	100	0.6
SGBM[4]	8.92	20.59	10.86	90.41	1.1
Ours	1.83	**2.62**	2.11	96.2	2.9

We believe this is because, in an outdoor environment, the scale of the scene is much larger than in indoor scenes, which causes more cost for information broadcasting from anchor points. However, the performance is more balanced in the background and foreground, and anchor points once again avoid unacceptable errors. Those results show that our method can adapt to outdoor scenes but performs better in indoor situations.

5 Limitation

From our experiment results, we notice that while the proposed approach can be adapted to different scenes without fine-tuning, it may not be possible to classify thin objects into any plane, which reduces the density of the depth map slightly. Besides, the disparity map has some discontinuous edges between small blocks due to the sliding window method. Anchor points broadcasting mechanism, though, rigidly constrains large errors and can cause small errors. This limits the applications for our method with high accuracy requirements.

6 Conclusions

In this paper, we present MobiDepth, a novel multi-step binocular depth estimation system. Our method can adapt to different situations and create high-quality dense depth maps. Unlike pure deep learning methods, our approach does not need fine-tuning to suit unfamiliar scenes and has excellent interpretability and generalization ability. We can track the origins of every pixel with a depth value from matched anchor points. Our approach almost completely outperforms traditional methods such as SGBM and is comparable to advanced deep learning methods. We hope our methods can provide a new idea for binocular depth estimation in unfamiliar scenes and replace previous self-supervised methods like cost-volume for binocular depth estimation models.

Funding

This work is supported by the grant from FDCT-MOST (0018/2025/AMJ) and Macao Polytechnic University (RP/FCA-01/2025).

References

1. Dong, X., Garratt, M. A., Anavatti, S. G. & Abbass, H. A. Towards real-time monocular depth estimation for robotics: A survey. *IEEE Trans. Intell. Transp. Syst.* **23**, 16940–16961 (2022).
2. Wang, Y. *et al.* Pseudo-LiDAR from visual depth estimation: Bridging the gap in 3D object detection for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8445–8453 (IEEE, 2019).
3. Oskiper, T., Sizintsev, M., Branzoi, V., Samarasekera, S. & Kumar, R. Augmented reality binoculars. *IEEE Trans. Vis. Comput. Graph.* **21**, 611–623 (2015).
4. Hirschmüller, H. Accurate and efficient stereo processing by semi-global matching and mutual information. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 2, 807–814 (IEEE, 2005).
5. Poggi, M., Tosi, F., Batsos, K., Mordohai, P. & Mattoccia, S. On the synergies between machine learning and binocular stereo for depth estimation from images: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**, 5314–5334 (2022).
6. Lipson, L., Teed, Z. & Deng, J. RAFT-stereo: Multilevel recurrent field transforms for stereo matching. In *2021 International Conference on 3D Vision (3DV)* (IEEE, 2021).
7. Khamis, S. *et al.* StereoNet: Guided hierarchical refinement for real-time edge-aware depth prediction. In *Computer Vision – ECCV 2018*, 596–613 (Springer, 2018).
8. Li, Z. *et al.* Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 6197–6206 (IEEE, 2021).
9. Scorzato, L. Reliability and interpretability in science and deep learning. *Minds Mach.* **34**, 1–31 (2024).
10. Xu, S. *et al.* Bifocal-binocular visual SLAM system for repetitive large-scale environments. *IEEE Trans. Instrum. Meas.* **71**, 1–15 (2022).
11. Luo, X., Han, B., Luo, Q. & Zhong, X. Research on SLAM system based on binocular vision and IMU information. In *Advances in Intelligent Systems and Computing*, 557–567 (Springer, 2021).
12. Zbontar, J. & LeCun, Y. Stereo matching by training a convolutional neural network to compare image patches. *J. Mach. Learn. Res.* **17**, 1–32 (2016).

13. Chang, J.-R. & Chen, Y.-S. Pyramid stereo matching network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5410–5418 (IEEE, 2018).

14. Zhang, F., Prisacariu, V., Yang, R. & Torr, P. H. S. GA-net: Guided aggregation net for end-to-end stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 185–194 (IEEE, 2019).

15. Laga, H., Jospin, L. V., Boussaid, F. & Bennamoun, M. A survey on deep learning techniques for stereo-based depth estimation. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**, 1738–1764 (2022).

16. Li, Q. *et al.* Deep learning based monocular depth prediction: Datasets, methods and applications. *arxiv* (2020).

17. Arampatzakis, V., Pavlidis, G., Mitianoudis, N. & Papamarkos, N. Monocular depth estimation: A thorough review. *IEEE Trans. Pattern Anal. Mach. Intell.* **46**, 2396–2414 (2024).

18. Miao, J. *et al.* A survey on monocular re-localization: From the perspective of scene map representation. *arXiv* (2023).

19. Eigen, D., Puhrsch, C. & Fergus, R. Depth map prediction from a single image using a multi-scale deep network. In *arXiv* (2014).

20. Eigen, D. & Fergus, R. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2650–2658 (IEEE, 2015).

21. Zhou, Z. & Dong, Q. Two-in-one depth: Bridging the gap between monocular and binocular self-supervised depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 9411–9421 (IEEE, 2023).

22. Chen, Y., Schmid, C. & Sminchisescu, C. Self-supervised learning with geometric constraints in monocular video: Connecting flow, depth, and camera. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 7063–7072 (IEEE, 2019).

23. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K. & Koltun, V. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**, 1623–1637 (2022).

24. Wang, Y., Liang, Y., Xu, H., Jiao, S. & Yu, H. SQLdepth: Generalizable self-supervised fine-structured monocular depth estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, 5713–5721 (2024).

25. Yang, L. *et al.* Depth Anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10371–10381 (IEEE, 2024).

26. Scharstein, D. *et al.* High-resolution stereo datasets with subpixel-accurate ground truth. In *Pattern Recognition*, 31–42 (Springer, 2014).

27. Yang, X., Feng, Z., Zhao, Y., Zhang, G. & He, L. Edge supervision and multi-scale cost volume for stereo matching. *Image Vis. Comput.* **117**, 104336 (2022).

28. Zeng, J., Yao, C., Yu, L., Wu, Y. & Jia, Y. Parameterized cost volume for stereo matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 18347–18357 (2023).

29. Menze, M., Heipke, C. & Geiger, A. Joint 3D estimation of vehicles and scene flow. In *ISPRS Workshop on Image Sequence Analysis (ISA)* (2015).

30. Menze, M. & Geiger, A. Object scene flow for autonomous vehicles. *ISPRS J. Photogramm. Remote. Sens.* (2018).

31. Chen, S., Xiang, Z., Qiao, C., Chen, Y. & Bai, T. PGNet: Panoptic parsing guided deep stereo matching. *Neurocomputing* **463**, 609–622 (2021).

32. Chen, S., Xiang, Z., Qiao, C., Chen, Y. & Bai, T. SGNet: Semantics guided deep stereo matching. In *Proceedings of the Asian Conference on Computer Vision (ACCV)* (2020).

Author contributions statement

Qibin He conceived and designed the study, performed the experiments, analyzed the data, and drafted the manuscript. Yanming Chai, Haochen Li, Yanxiao Li, Junhao Wei, Dexing Yao, Yifu Zhao, Ka-Hou Chan, Sio-Kei Im, and Yapeng Wang provided technical support, assisted with data collection and validation, and reviewed or revised the manuscript.

Additional information

The authors declare no conflicts of interest

Data Availability

All data generated or analysed during this study are included in this manuscript.