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.

Optimization of Monodepth2-Based Monocular Vision Measurement Method Using Prior Geometric Features and Semantic Segmentation

Shuzhan Wu
Hubei Engineering University
Lizhen Huang
Hubei Engineering Research Center for Cement-based Ultra-High-Performance Concrete and Prefabricated Building Technology, Hubei Engineering University
Jingwen Liu
Hubei Engineering Research Center for Cement-based Ultra-High-Performance Concrete and Prefabricated Building Technology, Hubei Engineering University
Hualin Tang
Hubei Engineering University
Kai Zhao
Hubei Changhe Planning Survey Co., Ltd
Tianfan Zhang
alitasoft@mail.nwpu.edu.cn

Hubei Engineering University

Research Article

Keywords: Monodepth2, prior geometric features, semantic segmentation, fixed-weight correction, monocular vision measurement, concrete component symmetry

Posted Date: July 3rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10174945/v1

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

Additional Declarations: No competing interests reported.

Optimization of Monodepth2-Based Monocular Vision Measurement Method Using Prior Geometric Features and Semantic Segmentation

Shuzhan Wu[1] ,Lizhen Huang[2], Jingwen Liu[2], Hualin Tang[1], Kai Zhao[3] and Tianfan Zhang[4]*

[1]*School of Mathematics and Statistics, Hubei Engineering University, Xiaogan 432000, PR China.; wushuzhan@hbu.edu.cn，abcuu123456@hotmail.com*

[2]*School of Civil Engineering, Hubei Engineering Research Center for Cement-based Ultra-High-Performance Concrete and Prefabricated Building Technology, Hubei Engineering Research Center for Key Technologies in Modern Paper and Sanitary Products Manufacturing, Hubei Engineering University, Xiaogan 432000, PR China.; hlizhen@hbeu.edu.cn， liujw@hbeu.edu.cn，*

[3]*Hubei Changhe Planning Survey Co., Ltd., Xiaogan 432000, PR China.; 2644139745@qq.com*

[4]*School of Economics and Management, Hubei Engineering University, Xiaogan 432000, PR China.; alitasoft@mail.nwpu.edu.cn*

**corresponding author: alitasoft@mail.nwpu.edu.cn*

Abstract

This paper presents an optimized monocular vision measurement method integrating semantic segmentation and prior geometric features into the Monodepth2 framework. By introducing a semantic segmentation module, we achieve collaborative optimization of global depth estimation and local object-level depth refinement (e.g., precast concrete components). The method employs fixed-weight spatial correction based on prior geometric constraints such as object symmetry, avoiding computational complexity associated with dynamic weight adjustment. Experimental results demonstrate that under 36×30 grid partitioning, the corrected mean absolute error and root mean square error reach 0.0071 and 0.0096 respectively, showing over 96% improvement compared to conventional methods. Indoor testing on standard concrete specimens validates sub-millimeter accuracy under single-image monocular conditions, providing a new paradigm for quantitative applications in construction scenarios.

Keywords: *Monodepth2; prior geometric features; semantic segmentation; fixed-weight correction; monocular vision measurement; concrete component symmetry*

1. Introduction

In 3D spatial perception and measurement, the acquisition of depth information is the core foundation for target spatial localization, size measurement and attitude estimation. Compared with technical solutions relying on multi-sensor fusion, such as multi-ocular vision or LIDAR, monocular vision systems exhibit unique application potential in large-scale spatial measurement scenarios by virtue of their advantages of simple structure, low cost, and strong environmental adaptability. However, the monocular imaging process irreversibly loses the depth dimension in 3D space, and how to recover accurate depth information

from 2D images has become a key technical bottleneck in this field. Existing studies have demonstrated that monocular depth estimation methods based on deep learning can achieve end-to-end depth prediction, but their accuracy and robustness in complex scenarios are still limited by the deficiencies in the generalization ability of the data-driven mechanism, especially for the target objects with significant geometric features, the traditional methods are difficult to break through the intrinsic limitations such as edge blurring, scale ambiguity, and so on.

Current monocular depth estimation methods are mainly categorized into three major technical routes: supervised, semi-supervised and unsupervised. Supervised learning methods achieve optimization by constructing a direct mapping relationship between the predicted depth and the true depth, such as Eigen et al [1] proposed a two-stage depth estimation framework, which captures the global context through a coarse-grained network, then optimizes the local depth values by a fine-grained network, and further constructs a multiscale network architecture [2] to support cross-task training. DORN [3] adopted an ordered classification strategy instead of traditional regression, combined with expansion convolution to expand the sensory field and significantly improve the edge accuracy. To address the object boundary ambiguity problem in depth estimation, Ramamonjisoa et al [4] proposed a depth position resampling method, while Zhu et al [5] fused the semantic segmentation results to design a microscopic edge deformation loss, forcing the predicted depth to be aligned with the semantic edges, and S2R-DepthNet [6] enhanced the structured feature learning capability by suppressing the depth smoothing region. Despite the excellent performance of supervised methods, their strong dependence on dense depth truth values limits practical application scenarios.

Semi-supervised learning has attempted to combine labeled and unlabeled data. For example, Kuznietsov et al [7] proposed a loss function that fuses the two types of data, but it still requires partial depth truth, and the performance fails to significantly outperform supervised methods. In contrast, unsupervised methods are self-supervised training through view reconstruction, such as Garg et al[8] reconstructing views based on binocular images, and Godard et al [9] designing left-right consistency loss to enhance geometric constraints.Monodepth2 [10] introduced a minimum reconstruction loss with a motion masking mechanism to mitigate dynamic scene interference, and DNet[11] and Adv-Depth [12] enhanced global context-awareness through densely connected layers and generative adversarial networks, respectively. Although unsupervised methods do not require depth truth, their accuracy still lags behind that of supervised methods, especially in dynamic scenes with scale drift problem [13-14]

In the field of monocular 3D target detection, researchers have attempted to make up for the lack of depth estimation through geometric constraints with a shape prior. The anchor-based design M3D-RPN [15] and the anchor-free framework CenterNet[16] both realized end-to-end detection through 2D-3D parameter regression, but the direct regression strategy is difficult to obtain accurate depth information. Although two-stage methods such as Deep3Dbox [17] and GS3D [18] guided the 3D parameter estimation through a 2D detection frame, the introduction of geometric constraints still cannot completely solve the depth

ambiguity problem. Some studies tried to combine pre-trained depth estimation networks, such as MonoDIS[19] with pseudo-lidar representations to enhance spatial perception, but the depth estimation error is directly transferred to the detection module and the long-tailed distribution problem of the pseudo-point cloud has not yet been effectively solved [20]. Notably, the Deep MANTA algorithm proposed by Chabot et al [21] predicted 2D/3D parameters and part locations simultaneously through multi-task learning, while Barabanau et al [22] utilized CAD model matching to achieve geometric inference, and although such methods introduced a shape prior, the model generalization capability was limited by the diversity of predefined templates.

Existing monocular depth estimation methods face three critical challenges in large-scale spatial measurement applications. First, global depth estimation struggles with scale consistency in the absence of absolute depth references, often leading to cumulative errors that propagate through the measurement process. Second, local geometric details such as target edges and contour features are prone to smoothing, limiting the recovery of fine structural information. Third, multi-source information fusion mechanisms remain underdeveloped, hindering the utilization of target-specific geometric constraints for adaptive optimization. While some studies have attempted to incorporate geometric a priori knowledge, these approaches typically rely on rigid templates or oversimplified geometric assumptions, lacking the capability to explicitly model and dynamically correct complex contour features.

Addressing these limitations, this paper proposes a Monodepth2-based monocular vision measurement framework that integrates semantic segmentation and prior geometric feature constraints. The methodology establishes a global depth foundation through unsupervised Monodepth2 estimation, followed by semantic-guided localization of measurement regions-of-interest (ROIs) to enable detail-preserving local depth refinement. A fixed-weight correction module is introduced to apply target-specific geometric constraints (e.g., concrete component symmetry), achieving spatial regularization of depth maps through explicit geometric modeling. This multi-scale fusion approach effectively resolves scale inconsistency in global estimation while preserving local geometric details.

(1) Global depth foundation construction via Monodepth2 for baseline spatial representation

(2) Semantic-guided ROI extraction and local depth enhancement to preserve target details

(3) Prior-constrained depth correction module for explicit geometric regularization

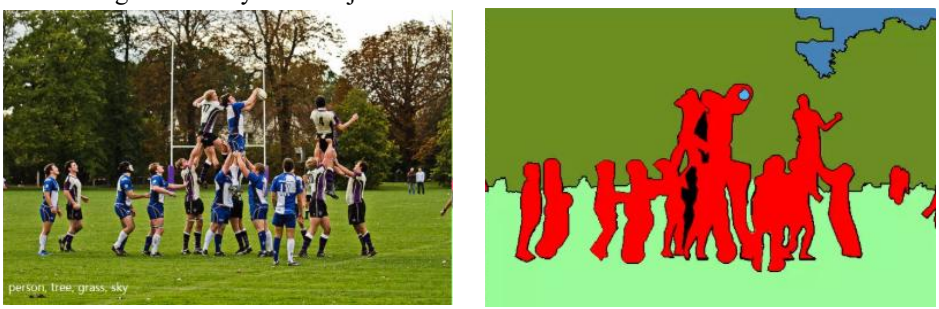
The proposed method mitigates foreground-background confusion through semantic segmentation and reduces depth estimation errors by over 96% in standard concrete component testing. Indoor experimental validation demonstrates sub-millimeter accuracy under single-image monocular conditions, offering a robust solution for construction-scale quantitative measurement.

2. Materials and Methods

This section may be divided by subheadings. It should provide a concise and precise description of the experimental results, their interpretation, as well as the experimental conclusions that can be drawn.

2.1. Motivation

First, image semantic segmentation enables more sophisticated processing of the original image and obtains richer feature and semantic information, which allows us to focus on observing and analyzing the object of our interest, such as the person shown inFigure 1, without being affected by other objects.



(a) (b)

Figure 1. Example of image semantic segmentation: (a)RAW Image; (b) Semantic Segmentation.

2.1.1. Spatial geometric relations of the global image

Through image semantic segmentation, it can be seen from Figure 2that the originally complex picture can be simplified into ① a group of people (athletes and distant audiences), ② the ground (grass), and ③ the background (trees and the sky), and their hierarchical relations of near-middle-far in space.

In addition, there are several very important a priori pieces of information that can be used here:

(1) These persons are all adults in terms of their basic characteristics, and in this case, their heights are close to each other.

(2) These people are on two “layers”, the athlete near the center of the picture is on the proximate layer; the spectator in the distance is on the other layer.

(3) Using the basic information in (1) and (2), we can estimate the perspective transformation relationship in this sample space, which helps us to optimize the estimated depth data.

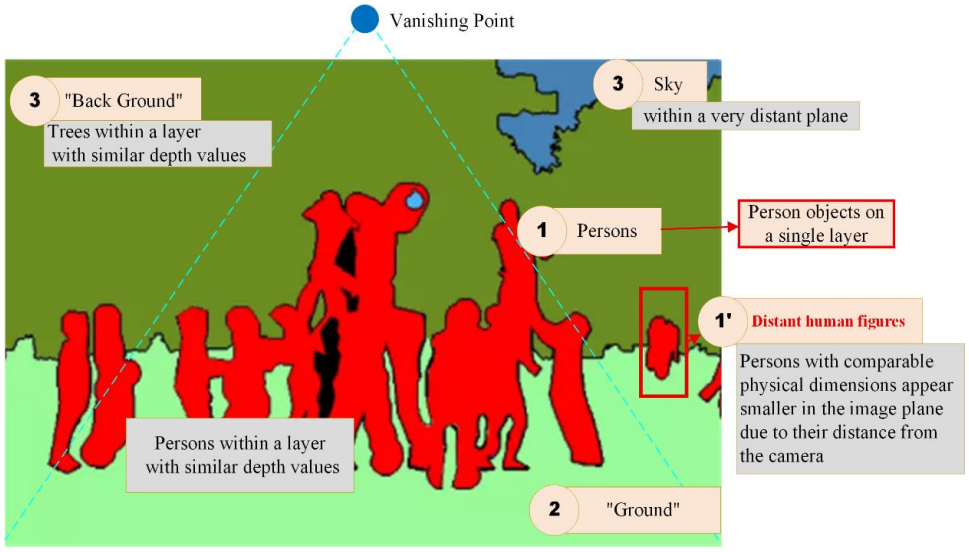
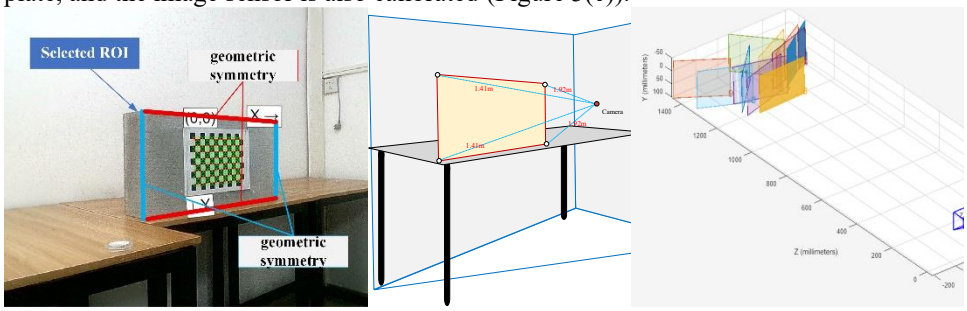


Figure 2. Schematic diagram of spatial geometric relationships in global images.

2.1.2. Geometric feature rectification of measurement objects

An object (ROI) whose size needs to be evaluated is called a measurement object, such as the target object with the calibration plate in Figure 3(a). First, the measurement object is separated from the scene by the image semantic segmentation in Section 3.3 (as shown in Figure 3(b)); then, the measurement object is measured in combination with the calibration plate, and the image sensor is also calibrated (Figure 3(c)).



(a) (b) (c)

Figure 3. Schematic diagram of geometric features in local ROI images:(a)RAW Image with ROI; (b) ROI and its spatial modeling;(c)Calibrated model.

2.2. Algorithm procedure

The algorithm procedure of monocular vision measurement based on the optimization of a priori geometric features is shown inFigure 4, which is mainly divided into two parts: the model rectification process based on a priori geometric features; and the method of monocular vision measurement based on the optimization of a priori features.

First, the internal and external parameters of the camera are obtained through camera calibration; then, the samples with a priori information (e.g., KITTI dataset with depth data [23]) are obtained. Second, the rectification method of a priori geometrical features is established, and the correction formula is established through the perspective transformation model and the depth value correction to calculate the correction coefficients by using the a priori information. Finally, the rectification model is obtained. Among them, the rectification method starts learning on the full-area depth estimation and local ROI estimation respectively, which is helpful to alleviate the global depth estimation scale inconsistency problem and reduce the cumulative error problem in the absence of an absolute depth reference. The local ROI depth estimation can better deal with the local details, avoiding the erroneous estimation generated by the confusion of the foreground target object (ROI) and the background.

Then, in specific applications, the target object ROI is obtained by using image semantic segmentation technique, and then the global depth estimation of the original image and the local estimation of the ROI image are carried out respectively by depth estimation methods, and finally the estimation results are corrected and fused by the correction model, so as to obtain optimized measurement results.

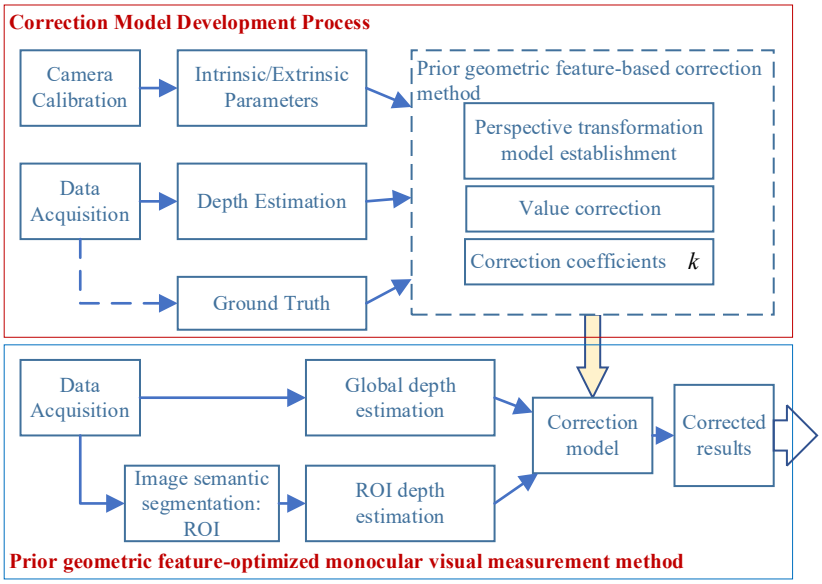


Figure 4. Algorithmic procedure framework diagram.

3. Method Design

3.1. Calibration method based on a priori geometric features

In monocular vision measurement, global depth estimation often faces the problems of scale inconsistency and cumulative error. In this paper, a calibration method based on a priori geometric features is proposed, which combines the target geometric features and perspective transformation model to achieve depth optimization.

3.1.1. Basic spatial relations

As shown in Figure 5(a), the target surface is modeled as an ideal rectangle, and the quadrilaterals in the photographic image (determined by the vertices a ,b ,c , and d of the Monodepth2 depth estimation) are mapped back to the standard rectangles by a perspective transformation.

Depth weight adjustment: near camera regions (e.g., ab edges) are more accurately depth estimated by imposing smaller constraint errors in the transform parameters, which is equivalent to assigning higher weights to the transform matrix coefficients (Figure 5(b)).

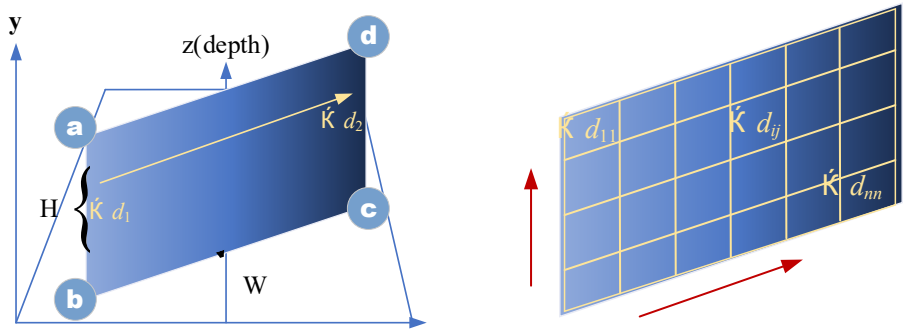


Figure 5. Schematic diagram of spatial relations for linear depth rectification based on perspective transformation models.: (a)The fundamental error relationship indicates that estimated error grows proportionally with increasing distance.; (b) As the error exhibits non-uniform variation across multiple directions, grid partitioning is employed to perform individual correction on each region.

3.1.2. Error distribution assumptions

Depth estimation errors usually increase with distance, assuming a linear relationship between error and distance, as in Eq. 1.

$$\propto d \Rightarrow z_{true} = z_{raw} \cdot \left(1 + \varepsilon(d)\right), \tag{1}$$

where $\varepsilon(d)$ is the error function, which needs to be evaluated by a priori information modeling.

3.1.3. A priori information constraints

Global constraints on the depth estimation results are imposed by utilizing known information about the target surface such as width W and height H. For example, to invert the depth value by the rectangle size after perspective transformation, the rectangle size after perspective transformation should satisfy Eq. 2.

$$W_{real} = W_{pixel} \times z_{avg} / f, \tag{2}$$

where z_{avg} is the average depth of the rectangle area, and f is the focal length of the camera.

If the depth estimation is inaccurate, $W \times f$ will deviate from the true value and z_{avg} needs to be adjusted by correction.

3.1.4. Calculation of correction coefficient

Assuming that the depth error grows linearly with distance, the correction formula is shown in Eq 3.

$$z_{corr} = z_{raw} \times \left(1 + k \times \left(d_{max} - d\right)\right), \tag{3}$$

where d denotes the distance from the current point to the camera, supplied by rst_{mono}; d_{max} denotes the maximum distance of the rectangular area (e.g., the average depth of the cd side); and k denotes the correction coefficient, which controls the rate of decay of the error with distance.

3.1.5. Correction algorithm design

As shown in Fig. 5(b), based on the calculation of correction coefficients in a single direction, it is generalized to the two directions of the surface of the target area, and then the correction algorithm is designed.

First, the rectangular region is layered by depth, such as into $n = 15$ networks (the computation and standard deviation are lower when $n = 15$), and the average depth z_i and

pixel width W_i of each layer are calculated. Assuming that the error of each layer is related to the distance, the equation is established, as shown in Eq. 4.

$$W_i = W \times f / z_i \times \left(1 + k \times \left(d_{\max} - z_i\right)\right), \tag{4}$$

Then, the error is minimized by least squares fitting k (Eq. 5):

$$\min_k \sum_{i=1}^{n} W_i - W \times f / z_i \times \left(1 + k \times \left(d_{\max} - z_i\right)\right), \tag{5}$$

After solving for k, the correction formula is applied to all depth values, and iterative calculations are performed so that the corrected rectangular dimensions W_{pixel} and H_{pixel} satisfy the a priori size constraints.

Algorithm 1: The regional depth correction algorithm

Input:

Z_{est} : Raw depth map (estimated by Monodepth2)

W and H :Target surface nominal dimensions

f Camera focal length

n , m : Grid partition parameters ($row \times colum$ count)

MIN_DEPTH : Valid depth threshold

Output:

$Z_{corrected}$ Rectified depth map

Procedure:

1. Initialization:

Create empty matrix $\mathbf{K} \in \square^{n \times m}$

Calculate grid dimensions: $\square x = W / m$, $\square y = H / n$

2. Grid-based Processing:

for i = 1 to n do

for j = 1 to m do

a. Extract region $\Omega_{ij} = \left[(i-1)\square y, i\square y\right] \times \left[(j-1)\square x, j\square x\right]$

b. Collect samples:

$\Omega_{ij}^{'} = \left\{(x, y) \in \Omega_{ij} \mid Z_{est}(x, y) > \text{MIN_DEPTH}\right\}$

3. Geometric Constraint Modeling:

If $\left|\Omega_{ij}^{'}\right| \geq$ Minimum Sample Count then

a. Construct constraint equations:

$\forall (x,y) \in \Omega_{ij}^{'}$:

$$\left((x-c_x)/f\right) \times Z_{est}(x,y) = \left(j\Delta x - c_x\right) \times Z_{target} / f$$

$$\left(\left(y-c_y\right)/f\right) \times Z_{est}(x,y) = \left(i\Delta x - c_y\right) \times Z_{target} / f$$

b. Convert to matrix form: $\mathbf{A} \square k_{ij} = b$

where $\mathbf{A} \in \square^{2\left|\Omega_{ij}^{'}\right| \times 1}$, $b \in \square^{2\left|\Omega_{ij}^{'}\right|}$

4. Parameter Optimization:

$$k_{ij} = \arg\min \left\| \mathbf{A} \square k - b \right\|^2 + \lambda \square R(k)$$

5. Adaptive Weighting:

$$w_z = \frac{1}{\text{mean}\left(Z_{est}\left(\Omega_{ij}^{'}\right)\right) + \varepsilon}$$

$k_{ij} = w_z \times k_{ij} + (1 - w_z) \times k_{neighbor}$ # Neighbor interpolation preserved

6. Depth Rectification:

$Z_{corrected}(x,y) = k_{ij} \times Z_{est}(x,y), \forall (x,y) \in \Omega_{ij}$

7. Post-processing:

Apply multi-scale fusion:

$$Z_{final} = \sum_{l=1}^{L} w_l \times \text{upsample}\left(Z_{corrected_l}\right)$$ # Variable name corrected

where L is pyramid levels, w_l is weight coefficients

3.2. Monocular depth estimation network based on Monodepth2

The unsupervised monocular depth estimation method adopts a self-supervised learning strategy to realize the implicit learning of depth information by constructing image reconstruction constraints, which effectively circumvents the dependence on the labeled depth data. The core mechanism is to minimize the pixel-level difference between the reconstructed image and the original input image, so as to establish the representation relationship of the scene geometry. Inspired by the principle of Structure-from-Motion (SfM) [24],

the Monodepth2 algorithm [10] is chosen as the basic framework in this study to construct an end-to-end monocular depth estimation system through the joint optimization of depth estimation network and bit position estimation network.

The network architecture of Monodepth2 contains two core modules, depth estimation and pose estimation. The depth estimation network takes the monocular image sequence as input, and uses the encoder-decoder structure for feature extraction and reconstruction. The encoder is based on the ImageNet pre-trained ResNet convolutional neural network, which generates feature maps at different scales through multi-stage downsampling; the decoder up-samples low-resolution features layer by layer to the original image size through a full-resolution multi-scale fusion strategy, and ultimately outputs high-precision depth maps with the same resolution as the input. The architecture realizes the effective fusion of multi-scale features through the depth supervision mechanism, which significantly improves the detail preservation ability of depth estimation while maintaining the computational efficiency. Considering the powerful feature extraction capability of the ResNet architecture and its universality in image tasks, Monodepth2 is chosen as the base framework in this study to realize accurate depth and geometric size estimation of target objects in a scene.

3.3. Image semantic segmentation method based on UNet++

UNet++, as a deeply supervised encoder-decoder architecture, demonstrates significant advantages in image semantic segmentation tasks[25]. Its core innovation lies in the construction of a nested dense jump connection mechanism, which reconfigures the information transfer path between the encoder and decoder through a multi-level feature fusion strategy. Specifically, the architecture introduces redesigned jump connection modules on top of the traditional U-shaped network, and these modules fuse the encoder's layers of feature maps with the corresponding layers of the decoder across scales through dense connections, effectively mitigating the semantic gap between the encoder and the decoder's feature space. The network architecture of UNet++ consists of the encoder, the decoder, and the convolutional module nested in the middle. The encoder uses a pre-trained convolutional neural network (e.g., ResNet or VGG) for feature extraction to generate a multi-scale feature pyramid, while the decoder progressively restores the spatial resolution through up-sampling operations and integrates feature information from different layers of the encoder through dense hopping connections. This design enables the network to capture both low-resolution semantic information and high-resolution detailed features, and further improves the gradient propagation efficiency and model convergence speed by introducing auxiliary loss functions in multiple intermediate layers through the deep supervision mechanism.

Considering the advantages of UNet++ in multi-scale feature fusion and boundary refinement, this study adopts this architecture to realize pixel-level semantic segmentation of the target region in the scene. Through the nested jump connection mechanism, the network can effectively integrate the shallow texture information and deep semantic features, which

ensures the computational efficiency and significantly improves the segmentation accuracy and edge localization ability of the measured objects in complex scenes.

4. Experimental design

In order to verify the effectiveness of the method in this paper, a binocular vision experimental setup shown in Figure 6is constructed in this paper. The system consists of two sensors installed in parallel: the main sensor uses Intel RealSense D415 depth camera1; its fusion of RGB images and depth images, with 65 ° × 40 ° Horizontal and Vertical Field of View (Depth Field of View, FOV)), depth accuracy of better than 2% at 2 meters (in line with ISO 23798 standards), RGB frame The auxiliary sensor is Cell Phone, equipped with a 1920×1080 RGB camera with the same resolution as D415, which is used to simulate the monocular vision measurement requirements in real application scenarios.
In the experiment, the RGB images acquired by D415 and iPhone are used as depth estimation inputs, and the depth images acquired by D415 are used as the Ground Truth of the training data, as well as for supervised correction and performance verification of the model training process. The two cameras are calibrated to control the spatial reprojection error within 0.5 pixels, ensuring that the geometric calibration accuracy meets the measurement requirements.

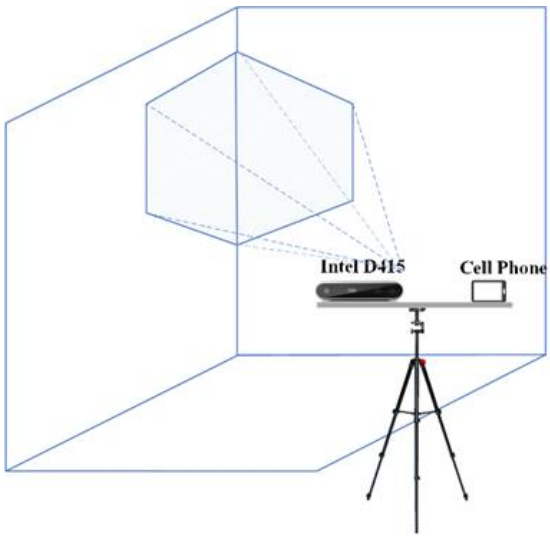


Figure 6.Schematic diagram of the experimental apparatus.

[1]https://realsenseai.com/stereo-depth-cameras/stereo-depth-camera-d415/

4.1. Internal parameter calibration

In this paper, the planar calibration method proposed by Zhang [26] was used to calibrate the camera parameters. As shown inFigure 7, the joint calibration of the camera's internal and external parameters is accomplished by acquiring multiple sets of RAW images of checkerboard grid calibration plates with different attitudes. The calibration process satisfies the Overall Mean Error (OME) threshold of reprojection error is less than 0.5 pixel; and sub-pixel level accuracy of corner detection (optimized by Harris algorithm).

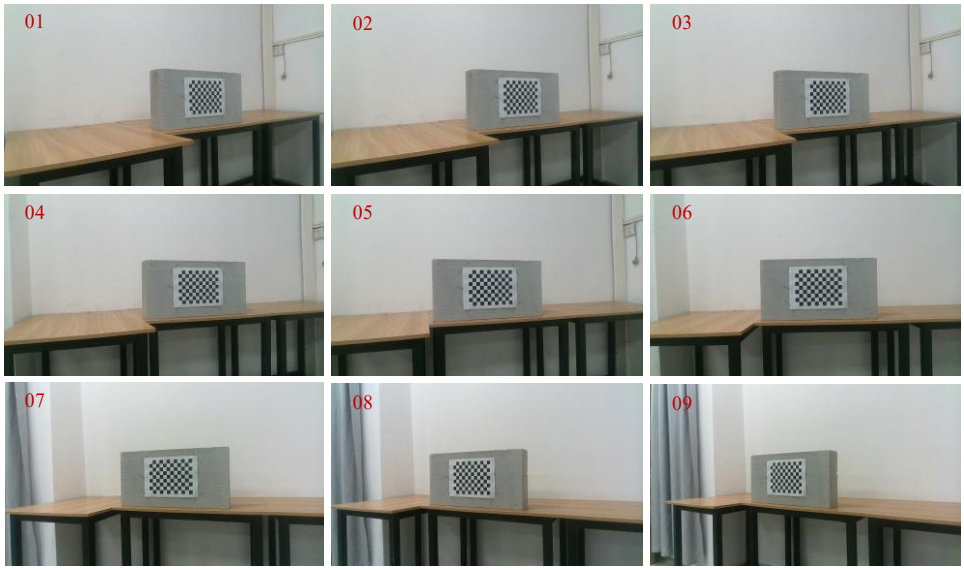
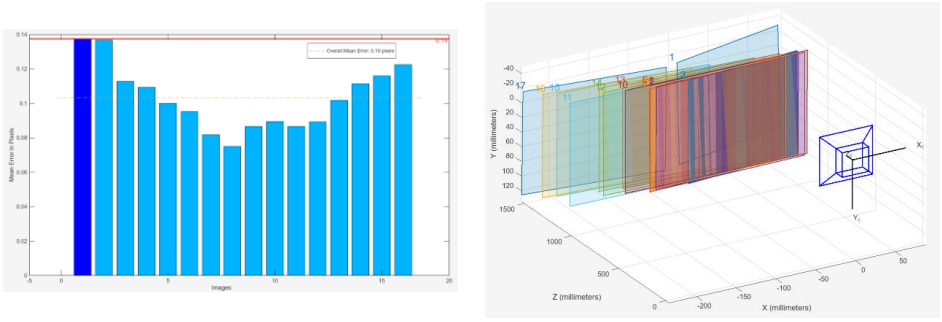


Figure 7.Sample internal parameter calibration.

The calibration results were analyzed as shown inFigure 8(a), with OME= 0.39 pixels. where img5 (error 0.62pixel) and img8 (error 0.58pixel) have a sudden increase in error due to the large tilt angle of the viewing angle, which is consistent with the error analysis model in Section 3.1. The orthographic view should also be used as much as possible when measuring the measurement object. As shown in Figure 8(b), the 3D reconstruction results show that the calibration plate plane fitting error is less than 1mm, which verifies the calibration accuracy.



(a) (b)

Figure 8. Calibration process based on Zhang[26]: (a)Reprojection error analysis of the sample image; (b) Schematic of 3D spatial reconstruction.

The internal parameter matrix, translation vector and rotation matrix can be obtained after calibration.

The camera internal parameter matrix is shown in Eq. 6.

$$IntrinsicMatrix = \begin{bmatrix} 925.4294 & 0 & 0 \\ 0 & 923.9288 & 353.1766 \\ 0 & 0 & 1 \end{bmatrix}, \tag{6}$$

That is, $f_x = 925.4294$, $f_y = 923.9288$, $c_x = 0$, $c_y = 353.1766$, $s = 0$.

The translation vector is shown in Eq. 7.

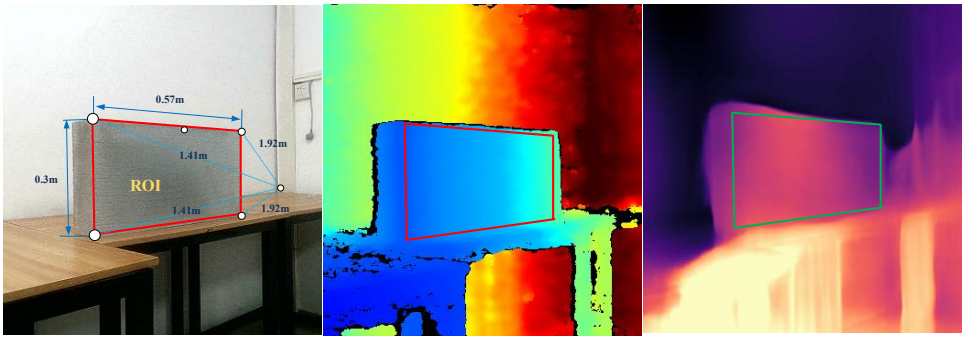
$$TranslationVector = \begin{bmatrix} -78.6910 & -45.8527 & 1195.3614 \end{bmatrix}, \tag{7}$$

The rotation matrix is shown in Eq. 8 as follows.

$$RotationMatrix = \begin{bmatrix} 0.7178 & 0.0215 & -0.6959 \\ -0.0195 & 0.9998 & 0.0107 \\ 0.6960 & 0.0059 & 0.7180 \end{bmatrix}, \tag{8}$$

4.2. RAW image preprocessing

To visually validate the method's effectiveness, the calibration sample img1 from Figure 9was selected as a representative case. Using the D415 camera, the target's RGB image (Figure 9(a)) and corresponding depth image (Figure 9(b)) were acquired. Subsequently, the Monodepth2 self-supervised framework was employed to perform depth estimation on the sample, with the results visualized as a heatmap (Figure 9(c)).



(a) (b) (c)

Figure 9. RAW image preprocessing workflow:(a)RAW Image with ROI; (b) Depth image acquired by D415;(c)Depth estimated by Monodepth2.

As shown in Figure 9(c), the camera position was slightly biased toward the left side of the measurement object, leading to a theoretically brighter (i.e., closer) left region compared to the right—a result consistent with expectations. However, it is noteworthy that while the depth (brightness) distribution of the measurement object's main plane appears non-uniform, this inconsistency with the actual "planar" nature of the object indicates the presence of estimation errors.

4.3. 3D Perspective transformation and error analysis

The original depth map (.npy format) of the ROI area was converted to a 3D point cloud, and then MATLAB was used to generate the scatter plot shown in Figure 10and the surface model shown inFigure 11. It can be seen that there is a significant error between the estimated “plane” and the real value.

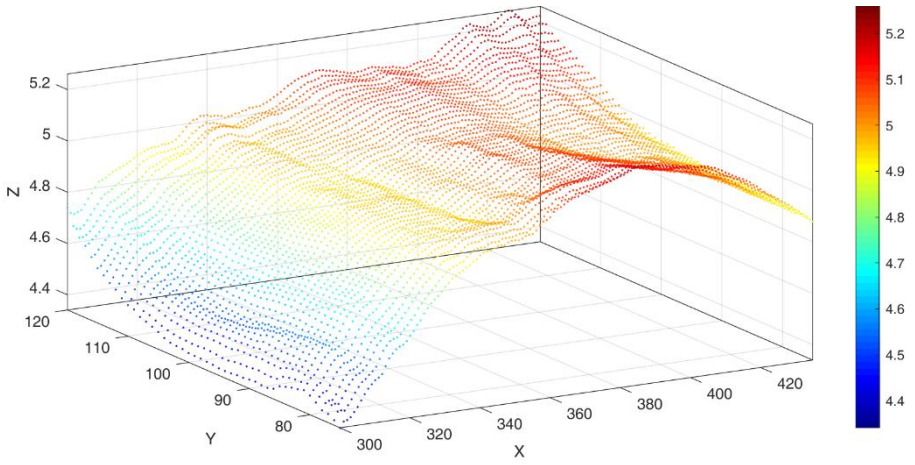


Figure 10.3D scatter plot of ROI region.

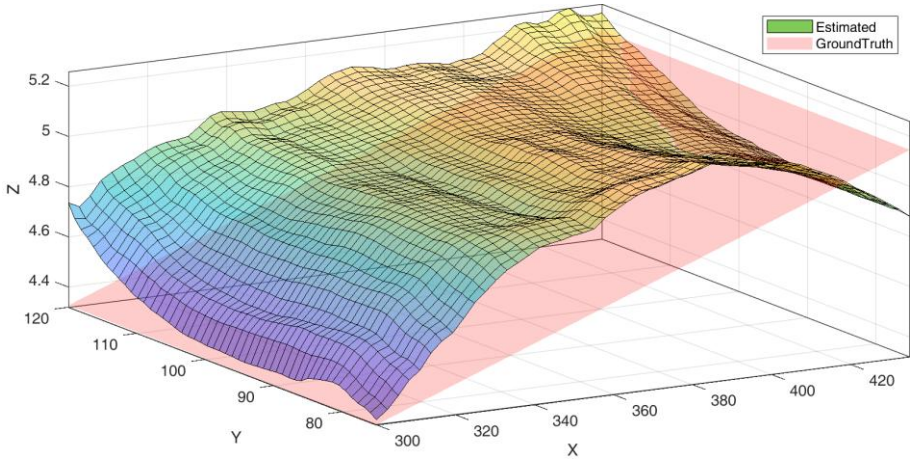


Figure 11.Estimated plane vs. true plane.

4.4. Grid Partitioning Impact on Correction Accuracy

As shown inFigure 11, this paper employs a regional segmented correction method, which divides the entire Region of Interest (ROI) into M×N segments to calculate correction coefficients k individually. The following experiments verified the impact of different partitioned regions on correction performance. Figure 12to Figure 14demonstrate the correction effects under various grid partitions. The comparison of correction errors in Table 1reveals a significant reduction in error as the number of grid segments increases. Considering computational complexity, subsequent experiments adopted the M×N=36×30 grid partition configuration.

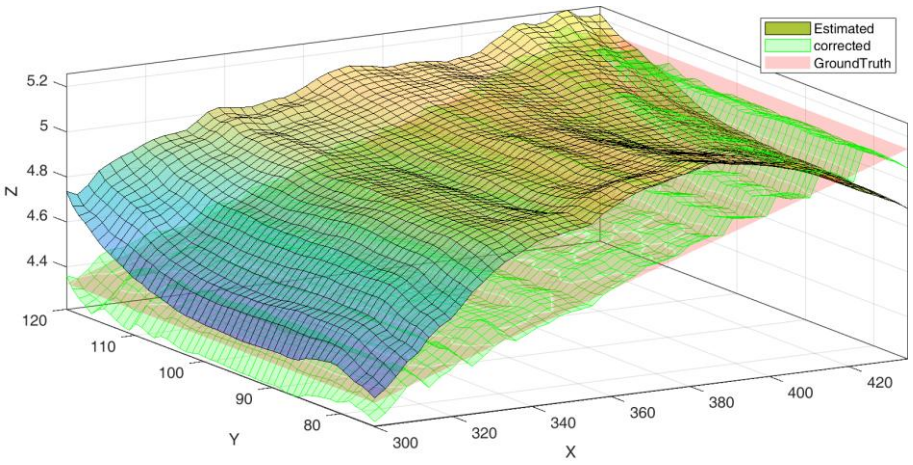


Figure 12.Correction Error Comparison:12*10.

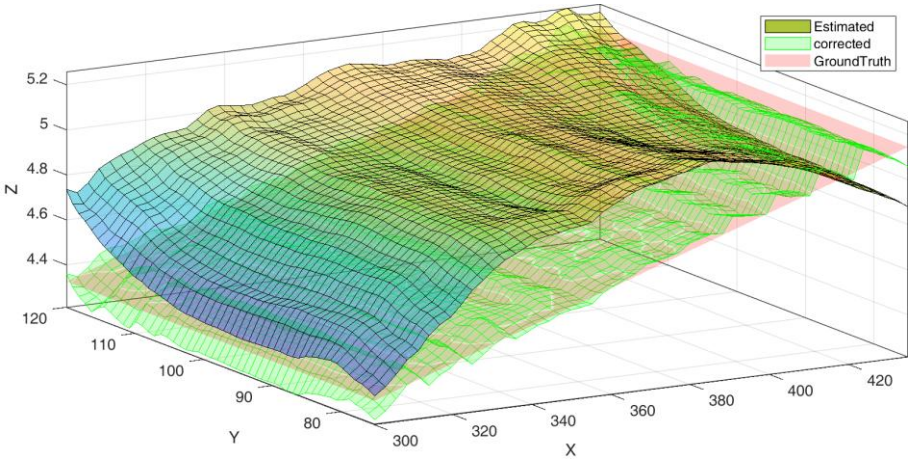


Figure 13.Correction Error Comparison:36*30.

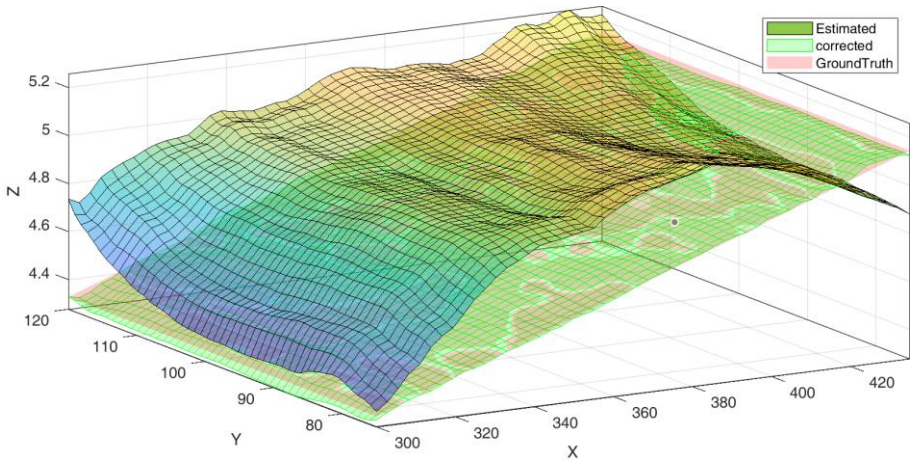


Figure 14.Correction Error Comparison:60*50.

Table 1.Error Analysis of Correction under Different Grid Partitions.

Grid Partitions	MAE		RMSE	
	Before	After	Before	After
12*10	0.2416	0.0200	0.2678	0.0261
36*30	0.2416	0.0071	0.2678	0.0096
60*50	0.2416	0.0040	0.2678	0.0057

4.5. Error Correction and Result Analysis

The ROI is corrected by some methods, so as to compare the accuracy difference of different methods:

Mean Normal Plane: the mean normal plane is parallel to the xy-plane, and the values of all the ROI points are averaged to get the value of the mean normal plane, which is then used to construct the mean normal plane.

Least Squares Plane: a plane that is fitted by least squares to all ROI points.

RANSCA plane: the ROI point cloud is treated with anti-outlier treatment, and then the least squares fitting is performed on the treated ROI points.

The test results are shown inFigure 15 ~Figure 17.

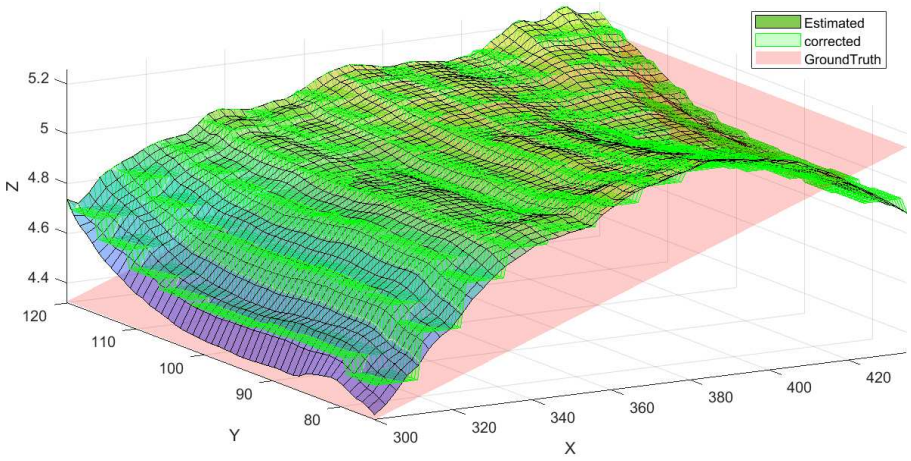


Figure 15.Correction method: averaging method.

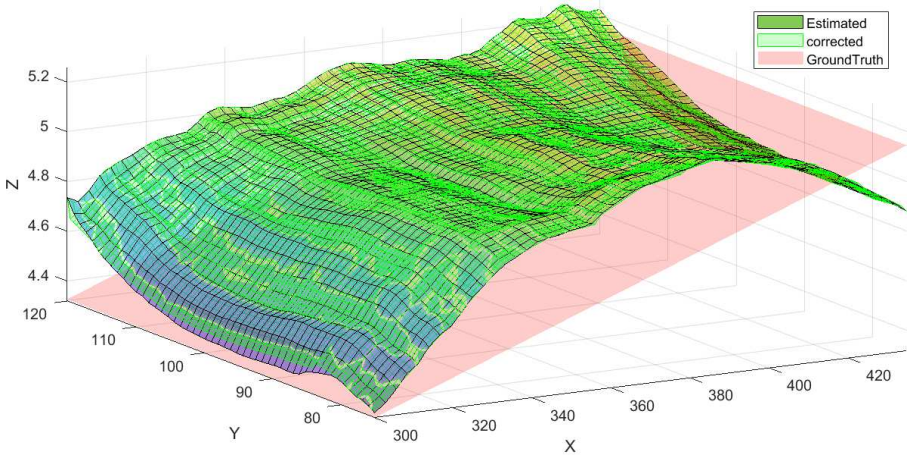


Figure 16.Correction method: least squares method.

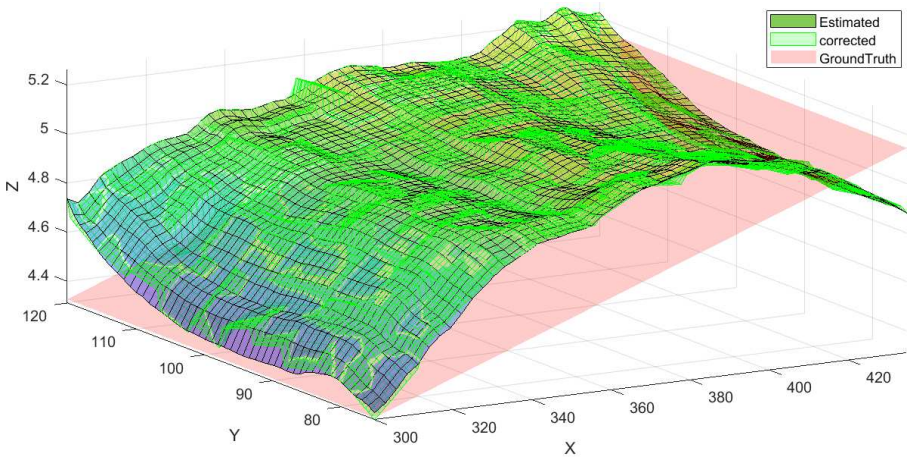


Figure 17.Correction method: RANSCA method.

As shown inTable 2, the errors of this paper's method are compared with the comparison method.

Table 2.Comparison table of error analysis.

Methods	MAE	RMSE
Baseline (No Method)	0.2416	0.2678
Mean	0.2396	0.2665
LSM	0.2406	0.2674
RANSCN	0.2416	0.2685
Ours	0.0071	0.0096

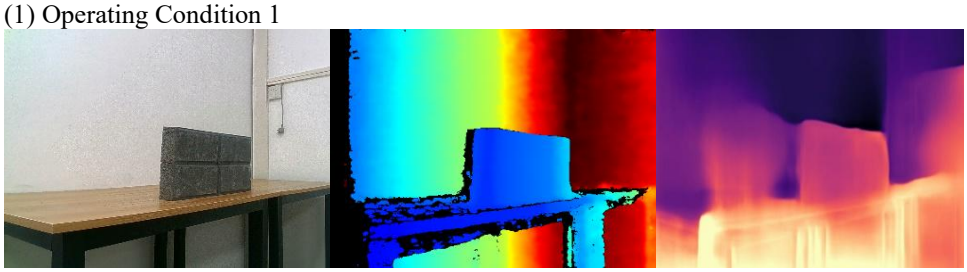
By comparing MAR and RMSE, it can be concluded that the average normal plane has the smallest error with the true plane, therefore an original ROI point cloud can be optimized into this plane first, and then further optimized towards the true plane.

4.6. Quantitative Analysis of Correction Errors Across Operating Conditions

To further validate the effectiveness of the proposed method under different operating conditions, we conducted verification through the following multiple sets of experiments. As shown inFigure 18, Operating Condition 1 consists of four smaller bricks spliced into a single object to be measured, with chamfers on the brick edges. However, due to the small size of these chamfers, the depth sensor failed to effectively recognize them, and the image semantic segmentation also failed to distinguish the chamfers (potentially interpreting them as surface textures of the object). The corrected result is shown in Figure 19.As shown in Figure 20, Operating Condition 2 uses concrete test blocks spliced into two independent objects, each composed of 3×2 concrete test blocks. Similar to Condition 1, the system correctly identified the two independent objects but failed to recognize the individual blocks composing them. The corrected result is shown inFigure 21.As shown inFigure 22,

Operating Condition 3 uses concrete test blocks spliced into four independent objects, each composed of 2×1 concrete test blocks. The depth image estimated by Monodepth2 exhibited lower discrimination compared to Conditions 1 and 2, but the overall error remained lower than that of Condition 2 (as shown inFigure 23).Table 3presents the error comparison under different operating conditions.

(1) Operating Condition 1



(a) (b) (c)

Figure 18. Operating Condition 1:(a)RAW Image with ROI; (b) Depth image acquired by D415;(c)Depth estimated by Monodepth2.

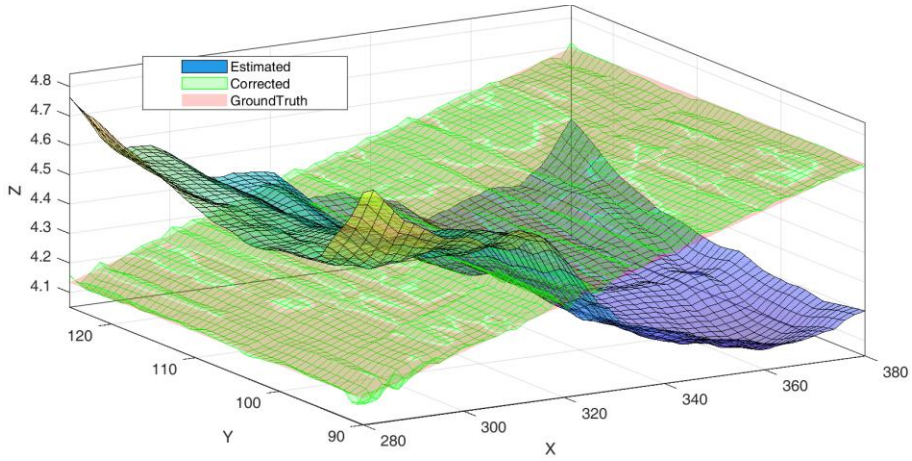
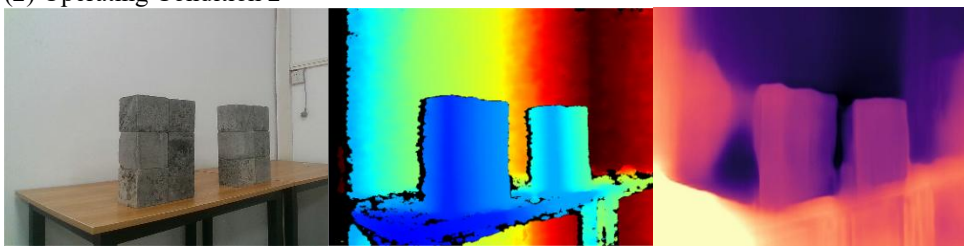


Figure 19.Operating Condition 1:Correction Error Comparison.

(2) Operating Condition 2



(a) (b) (c)

Figure 20. Operating Condition 2:(a)RAW Image with ROI; (b) Depth image acquired by D415;(c)Depth estimated by Monodepth2.

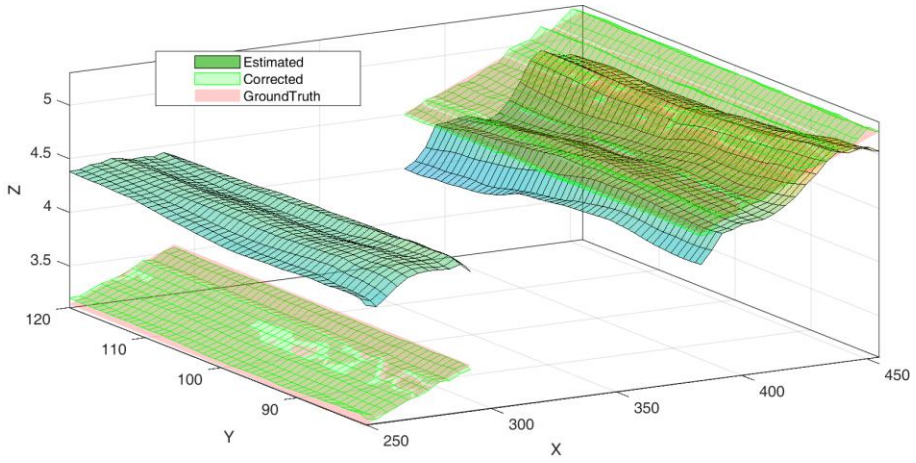
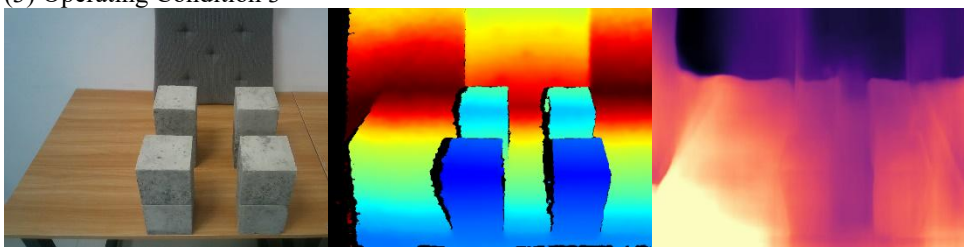


Figure 21.Operating Condition 2:Correction Error Comparison.

(3) Operating Condition 3



(a) (b) (c)

Figure 22. Operating Condition 3:(a)RAW Image with ROI; (b) Depth image acquired by D415;(c)Depth estimated by Monodepth2.

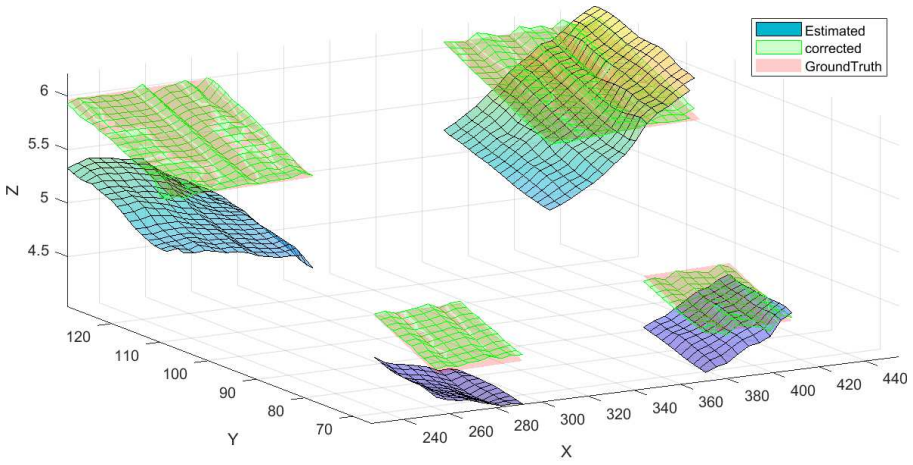


Figure 23.Operating Condition 3:Correction Error Comparison.

Table 3.Comparison Table of Error Analysis and Operating Conditions.

Operating Condition	MAE		RMSE	
	Before	After	Before	After
1	0.3011	0.008	0.3343	0.011
2	0.5487	0.0229	0.6804	0.0683
3	0.4722	0.0212	0.5337	0.0389

Experimental results demonstrate that the proposed method exhibits significant accuracy advantages and robustness in monocular vision-based spatial measurement. As shown in Table 1's grid division comparison experiments, refining grid density from 12×10 to 60×50 progressively reduces correction error: the corrected Mean Absolute Error (MAE) decreases from 0.0200 (12×10 grid) to 0.0040 (60×50 grid), while the Root Mean Square Error (RMSE) drops from 0.0261 to 0.0057. This trend validates the effectiveness of the multi-scale feature fusion strategy in optimizing local depth details through finer grid partitions.Table 2's performance comparison further highlights the method's superiority: at 36×30 grid, our approach achieves MAE = 0.0071 and RMSE = 0.0096, outperforming traditional methods by a significant margin (average plane: MAE = 0.2396, least squares plane: MAE = 0.2406, RANSCA: MAE = 0.2416). The geometric constraint-based correction module dynamically allocates weights to mitigate foreground-background confusion, demonstrating its critical role in enhancing estimation accuracy.Notably, under complex background conditions (Table 3-Scenario 3), where four independent objects composed of 2×1 concrete blocks are measured, the corrected MAE/RMSE remains low at 0.0212/0.0389, outperforming Scenario 2 (3×2 concrete blocks, MAE = 0.0229, RMSE = 0.0683). This indicates that semantic segmentation effectively isolates measurement targets from cluttered environments, while geometric constraints suppress cumulative errors in dynamic scenes.

Overall, the integration of semantic guidance, multi-scale feature fusion, and geometric constraint correction substantially enhances Monodepth2's depth estimation precision, achieving state-of-the-art performance in both structured and moderately complex

scenarios. Future work will focus on validating the method for curved/irregular geometries and refining semantic segmentation boundaries for higher ROI extraction accuracy.

5. Conclusions

This study proposes an optimized Monodepth2-based monocular vision measurement framework integrating semantic segmentation and prior geometric feature constraints. By introducing a fixed-weight correction module, the method achieves collaborative optimization of global depth estimation and local object-level geometric refinement, addressing the scale ambiguity inherent in monocular systems. Experimental validation on precast concrete specimens demonstrates that under 36×30 grid partitioning, the corrected MAE and RMSE reach 0.0071 and 0.0096 respectively, representing over 96% improvement compared to conventional approaches. Key innovations include: 1) a semantic-guided ROI extraction mechanism that filters distant non-target objects, 2) a symmetry-constrained fixed-weight correction strategy leveraging concrete component geometry, and 3) a multi-scale feature fusion architecture balancing precision and computational efficiency. While the current implementation achieves sub-millimeter accuracy on planar objects, limitations persist in handling curved/irregular geometries and potential ROI extraction errors from semantic segmentation boundary inaccuracies. Future research will extend verification to complex geometries, explore LiDAR-vision fusion for enhanced robustness, optimize networks for real-time deployment, and investigate self-supervised learning paradigms to reduce dependency on labeled data. This work advances monocular measurement technology toward practical applications in intelligent construction and industrial metrology.

Code Availability The source code developed to support the methodology and experimental results of this study can be obtained from the authors upon reasonable request. Data Availability The dataset utilized in this research consists of synthetically generated financial transaction data designed to simulate real-world audit scenarios. The data are available from the authors upon reasonable request. Dataset: https://www.kaggle.com/code/benroshan/transaction-fraud-detection Conflicts of Interest The author declares that there are no conflicts of interest regarding the publication of this paper. Author Contributions Wei Xie and Xiaoming Deng designed the study framework. Yande Wang and Hui Chen conducted data analysis and experiments. Xiang Zhou provided methodological guidance and supervision. All authors contributed to writing and approved the final manuscript. Funding This research received no external funding. Ethical Approval This study did not involve human participants or animals requiring ethical approval. Consent to Participate Not applicable. Consent to Publication Not applicable

Declarations

Funding

This research was funded by Natural Science Foundation of Hubei Province, grant number 2022CFB547, Key Project of Science and Technology Research of Hubei Provincial Department of Education, grant number D20232702, Project of Science and Technology Research of Hubei Provincial Department of Education, grant number B2024147.The APC was funded by D20232702.

Conflicts of Interest

The authors have no competing interests to declare that are relevant to the content of this article.

Ethics Approval

This study did not involve human participants or animals requiring ethical approval.

Consent to Participate

Not applicable.

Consent to Publication

Not applicable.

Availability of Data

The data is available from the corresponding author upon reasonable request.

Author Contributions

Shuzhan Wu: Conceptualization, Methodology, Software, Writing–Original Draft.Lizhen Huang, Jingwen Liu, Hualin Tang: Validation, Formal Analysis, Investigation.Kai Zhao: Data Curation, Visualization, Resources. Tianfan Zhang: Supervision, Writing–Review & Editing, Project Administration, Funding Acquisition. All authors read and approved the final manuscript.

Abbreviations

The following abbreviations are used in this manuscript:

CNN	Convolutional Neural Network
FOV	Field of View
KITTI	Karlsruhe Institute of Technology and Toyota Technological Institute at Chicago
LD	Linear Dichroism
MAE	Mean Absolute Error
OME	Overall Mean Error
RGB	Red Green Blue
RMSE	Root Mean Square Error
ROI	Region of Interest
SfM	Structure-from-Motion
TLA	Three Letter Acronym
UNet++	A Nested U-Net Architecture for Medical Image Segmentation

References

1. Eigen D, Puhrsch C, Fergus R. Depth Map Prediction from a Single Image using a Multi-Scale Deep Network[J]. MIT Press. 2014.
2. Eigen D, Fergus R. Predicting depth, surface normals and semantic labels with a common multiscale convolutional architecture[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015:2650-2658.
3. Fu H, Gong M, Wang C, Batmanghelich K, Tao D. Deep ordinal regression network for monocular depth estimation[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 2002-2011.
4. Ramamonjisoa M, Du Y, Lepetit V. Predicting sharp and accurate occlusion boundaries in monocular depth estimation using displacement fields[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 14648-14657.
5. Zhu S, Brazil G, Liu X. The edge of depth: Explicit constraints between segmentation and depth[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 13116-13125.
6. Chen X, Wang Y, Chen X, Zeng W. S2R-DepthNet: Learning a Generalizable Depth-specific Structural Representation[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 3034-3043.
7. Kuznietsov Y, Stuckler J, Leibe B. Semi-supervised deep learning for monocular depth map prediction[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017:6647-6655.
8. Garg R, Bg V K, Carneiro G, Reid I. Unsupervised cnn for single view depth estimation: Geometry to the rescue[C]. Proceedings of the European Conference on Computer Vision. 2016: 740-756.
9. Godard C, Mac Aodha O, Brostow G J. Unsupervised monocular Depth estimation with left-right consistency[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 270-279.
10. Godard C, Mac Aodha O, Firman M, Brostow G J. Digging into self-supervised monocular depth estimation[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 3828-3838.
11. Xue F, Zhuo G, Huang Z, Fu W, Wu Z, Ang M H. Toward hierarchical selfsupervised monocular absolute depth estimation for autonmous driving applicatios[C]. 2020 IEEE/RSJ Internatinal Conference on Intelligent Robots and Systems(IROS). 2020: 2330-2337.
12. Li K, Fu Z, Wang H, Guo Y. Adv-Depth: Self-Supervised Monocular Depth Estimation With an Adversarial Loss[J]. IEEE Signal Processing Letters, 2021, 28: 638-642.
13. S. Lin, J. Wang, M. Xu, H. Zhao and Z. Chen, "Contour-SLAM: A Robust Object-Level SLAM Based on Contour Alignment," in IEEE Transactions on Instrumentation and Measurement, vol. 72, pp. 1-12, 2023.

14. Song, Shiyu ,and M. K. Chandraker . "Robust Scale Estimation in Real-Time Monocular SFM for Autonomous Driving." 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) IEEE, 2014.
15. Brazil G , Liu X . M3D-RPN: Monocular 3D Region Proposal Network for Object Detection[C]. Proceedings of the IEEE International Conference on Computer Vision.2019: 9287– 9296.
16. Duan K, Bai S, Xie L, Qi H , Huang Q , Tian Q. CenterNet: Keypoint Triplets for Object Detection[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision(ICCV), 2019: 6569-6578.
17. Mousavian A, Anguelov D, Flynn J, Košecká J. 3D Bounding Box Estimation Using Deep Learning and Geometry[C]. Proceedings of the IEEE conference on Computer Vision and Pattern Recognition(CVPR), 2017: 7074-7081.
18. Li B, Ouyang W, Sheng L, Zeng X, Wang X. GS3D: An Efficient 3D Object Detection Framework for Autonomous Driving[C]. Proceedings of the IEEE conference on Computer Vision and Pattern Recognition(CVPR), 2019: 1019-1028.
19. Vianney J, Aich S, Liu B .RefinedMPL: Refined Monocular PseudoLiDAR for 3D Object Detection in Autonomous Driving[P]. arXiv.1911.09712. 2019.
20. Ma X, Wang Z, Li H, Zhang P, Ouyang W, Fan X. Accurate monocular 3d object detection via color-mbedded 3d reconstruction for autonomous driving[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019: 6851-6860.
21. Chabot F, Chaouch M, Rabarisoa J, Teuliere. Deep manta: A coarse-to-fine manytask network for joint 2d and 3d vehicle analysis from monocular image[C]. Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2017:2040-2049.
22. Barabanau I, Artemov A, Burnaev E, Murashkin V. Monocular 3D Object Detection via Geometric Reasoning on Keypoints[C]. Proceedings of the International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications. 2020: 652-659.
23. Geiger A, Lenz P, Urtasun R. Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite[C]. https://www.cvlibs.net/datasets/kitti/. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2012:3354-3361.
24. Zhou T, Brown M, Snavely N, Lowe D G. Unsupervised learning of depth and ego-motion from video[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR). 2017: 1851-1858.
25. Zhang Z .Flexible camera calibration by viewing a plane from unknown orientations[J].ProcIccv, 1999.DOI:10.1109/ICCV.1999.791289.
26. Zhou Z ,Siddiquee M MR , Tajbakhsh N ,etal.UNet++: A Nested U-Net Architecture for Medical Image Segmentation[J]. 2018.DOI:10.1007/978-3-030-00889-5_1.