High-Precision 3D Recognition of Road Potholes Based on Binocular Vision and Cross-Modal Feature Fusion

Hanzheng Wang [1], Shishuo Xu [1], Danyang Hu [1],Zheng Wen [1],Jianxi Ou [1]

[1] School of Geomatics and Urban Information, Beijing University of Civil Engineering and Architecture, Beijing, 102612, China kryonk1@gmail.com,xss_bucea@163.com, 3247464395@qq.com,wenzheng bucea@163.com, 1108130421002@stu.bucea.edu.cn

Keywords: binocular vision, lightweight detection, Mask R-CNN, 3D reconstruction, road damage ratio

Abstract

Efficient detection and accurate three-dimensional characterization of road potholes are crucial for road maintenance and traffic safety. To address the issues of high cost and poor environmental adaptability in existing detection methods, this study proposes a lightweight pothole detection and 3D reconstruction method based on binocular stereo vision and deep learning. A ZED 2i binocular camera was used to build a vehicle-mounted acquisition system, combined with the Mask R-CNN model to achieve pothole detection and pixel-level segmentation. The 3D point cloud of potholes was reconstructed using the principles of binocular stereo vision, and a dynamic mesh density method was proposed to optimize surface area calculation. Additionally, the RANSAC algorithm was employed to fit the ground plane and extract depth parameters. Experimental results demonstrate that this method can achieve precise measurements of pothole depth and surface area at a speed of 40 km/h, with relative errors of 12.53% and 18.19%, respectively, and an average accuracy of 82% for damage ratio (DR) calculation. Furthermore, an MSRCP image enhancement technique and a sliding window cropping strategy (overlap rate of 0.7) were used to construct a dataset containing 6,416 images, significantly improving the model's robustness in complex scenarios such as shadows and varying lighting conditions. This study provides road maintenance departments with a low-cost, high-precision intelligent pothole detection solution, reducing hardware costs by 90% compared to traditional laser sensors, and demonstrates significant value for engineering applications.

1. Introduction

Roads, as a vital component of urban transportation infrastructure, directly impact traffic safety and travel efficiency. In recent years, with the acceleration of urbanization in China and the continuous growth of vehicle ownership, road loads have increased significantly, leading to increasingly prominent pavement distress issues. According to the National Highway Maintenance Statistical Annual Report released by the Ministry of Transport in 2023, the annual average distress occurrence rate on urban roads in China has reached 12.7%, with pothole- related distress accounting for over 40% of cases. This results in more than 30,000 traffic accidents annually, causing direct economic losses of up to 8 billion yuan (Cano-Ortiz et al., 2024).

Traditional road distress detection primarily relies on manual inspections, which are inefficient and subjective (Fan et al., 2020). The advancement of computer vision technology has driven research into automated detection methods. Early algorithms based on edge detection and threshold segmentation lacked robustness in complex environments (He & Girshick R., 2017). Deep learning techniques, particularly models such as Faster R-CNN and YOLO, have significantly improved detection performance (Hsieh et al., 2024). However, monocular vision still faces challenges such as missing depth information, motion blur, and sensitivity to lighting conditions (Kendall & Gal ., 2017). While LiDAR can provide precise 3D information, its high cost limits widespread application (Mordohai & Medioni ., 2023). Binocular stereo vision technology, which simulates human binocular disparity to simultaneously capture RGB and depth information, offers advantages such as moderate cost and multi-view matching, providing a new approach to addressing these issues (Scharstein & Szeliski ., 2002). Nevertheless, effectively fusing multimodal information remains a key challenge in current research (Wang & Li ., 2022).

This study focuses on the specific application scenario of urban road pothole detection and proposes a lightweight detection method based on deep learning. The main innovations include:

(1) designing a low-cost onboard binocular vision acquisition system to achieve high-precision synchronized data collection; (2) innovatively introducing a depth attention mechanism into the Mask R-CNN framework to effectively enhance feature representation; (3) constructing a large-scale road pothole dataset covering diverse complex scenarios, including different times of day, weather conditions, and road conditions; and (4) proposing a dynamically optimized network training strategy that significantly improves detection performance on mobile platforms. Experimental results demonstrate that this method maintains high detection accuracy (mAP 98.10%) while reducing the false detection rate in complex environments by 31%, providing a reliable technical solution for practical engineering applications.

2. Lightweight acquisition and data preprocessing method

2.1. Principles of binocular vision

Binocular stereo vision is an important way for humans to perceive the three-dimensional world and serves as the foundation for depth perception and spatial localization. The spatial model of binocular stereo vision is shown in Figure 1.

Figure 1 illustrates the left and right cameras and the image coordinate systems, denoted as O-XY, orxy and 0rXY, o-xya, respectively. The focal lengths of the left and right cameras are f and f, respectively. There is a point P in space, and its projections on the left and right images are P and P. According

to the camera imaging principle, the coordinates of these two points can be given by Eq. (1) and Eq. (2).

$$s_t \begin{bmatrix} X_t \\ Y_t \\ 1 \end{bmatrix} = \begin{bmatrix} f_l & & \\ & f_l & \\ & & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} \quad (1)$$

$$s_r \begin{bmatrix} X_r \\ Y_r \\ 1 \end{bmatrix} = \begin{bmatrix} f_r & & \\ & f_r & \\ & & 1 \end{bmatrix} \begin{bmatrix} x_r \\ y_r \\ z_r \end{bmatrix} \quad (2)$$

The coordinate systems of the left and right cameras can be transformed using the spatial transformation matrix M = [R T], where R is the rotation matrix and T is the translation matrix. The points in the left and right camera coordinate systems are represented as shown in Equation (3):

$$\rho_r \begin{bmatrix} X_r \\ Y_r \\ 1 \end{bmatrix} = \begin{bmatrix} f_r r_1 & f_r r_2 & f_r r_3 & zX_r/f_l \\ f_r r_4 & f_r r_5 & f_r r_6 & zY_r/f_l \\ r_7 & r_8 & r_9 & t_z \end{bmatrix} \begin{bmatrix} zX_l/f_l \\ zY_l/f_l \\ 1 \end{bmatrix} \quad (3)$$

Therefore, the coordinates of point P can be calculated using Eq. (4).

$$\begin{cases} x = zX_l/f_l \\ y = zY_l/f_l \\ z = \dfrac{f_l(f_r t_z - X_r t_z)}{X_r(r_7 X_l + r_8 Y_l + f_l r_9) - f_r(r_1 X_l + r_2 Y_l + f_l r_3)} \end{cases} \quad (4)$$

The intrinsic matrices and transformation matrices of the left and right cameras are obtained through camera calibration. Then, by extracting feature points from the left and right images and performing mathematical processing, the projected coordinates of point P are determined, thereby enabling the calculation of its actual coordinates.

2.2. Construction of lightweight collection system

To address the inefficiency and high costs associated with traditional road inspection methods, this study designed a lightweight acquisition system based on binocular industrial cameras and a GNSS receiver. The system adopts a modular design, with core hardware including a 2K industrial camera (ZED 2i) and a sub-meter-grade GNSS receiver (BD-8953U). It can be rapidly deployed on the hood of ordinary vehicles via strong magnetic adsorption, eliminating the need for specialized modifications.

The camera is equipped with a polarizing filter to suppress road surface glare, while its 60Hz high frame rate ensures image clarity. The GNSS receiver supports a 10Hz update frequency, providing sub-meter positioning accuracy even at speeds of 80 km/h. The hardware connects to an in-vehicle industrial computer via USB 3.0/2.0 interfaces and is powered by a 220V mobile power supply, enabling plug-and-play flexible deployment.

Compared to traditional inspection vehicles (e.g., the Pathway system), this solution reduces equipment costs by 85% and offers strong adaptability, making it suitable for various scenarios such as urban roads and highways.

Figure 1 Construction of a lightweight collection system

On the software level, the system is developed based on Python and the Qt framework, integrating three core functional modules: GNSS data parsing, synchronized image acquisition, and geospatial information embedding. Utilizing multithreading technology, it achieves millisecond-level synchronization between GNSS positioning data and image frames, while embedding latitude and longitude information into the image metadata to construct a spatiotemporal pavement defect database.

To address vibration interference in vehicular environments, the system employs optimized exposure time (8.33ms) and polarized light filtering technology, significantly improving imaging quality for defects such as cracks and potholes. Experimental results demonstrate that the system can stably output usable images with a resolution of 1500×1500 pixels even under complex lighting conditions, meeting the precision requirements for defect detection specified in the Urban Road Maintenance Technical Specification (CJJ 36-2016).

Figure 2 Front-end acquisition software interface

2.3. Pavement pothole image data preprocessing

The ZED 2i binocular camera (with a capture resolution of 2048×1080 pixels) was used to collect images of road potholes. To address issues of illumination variation and

motion blur during dynamic acquisition, an image enhancement method based on Multi-Scale Retinex with Color Preservation (MSRCP) was employed. By utilizing Gaussian pyramid decomposition and adaptive gamma correction (γ=0.8-1.2), the method effectively improved image quality while preserving color authenticity. The processed images achieved PSNR, SSIM, and NIQE metrics of 32.5 dB, 0.91, and 3.2, respectively.

Evaluation Metrics	MSRCP Algorithm	Traditional Histogram Equalization	Performance Improvement
PSNR (dB)	32.5	28.3	14.8%
SSIM	0.91	0.79	15.2%
NIQE	3.2	3.8	15.8%

Table 1 Comparison of MSRCP Enhancement Effects

For the calibration of the binocular vision system, an improved Zhang's calibration method was adopted. Using nine sets of checkerboard images (9×6 corners, 30mm grid spacing) captured at different poses, the camera parameters were calculated. The Levenberg-Marquardt algorithm was introduced to optimize the reprojection error, ultimately achieving calibration results with a focal length error of <0.3% and a baseline distance accuracy of 0.1mm.

Additionally, a physics-based data augmentation framework was constructed to simulate motion blur at a vehicle speed of 40 km/h and rain/fog interference under varying atmospheric conditions (β=0.05-0.1). This approach enhanced the model's detection robustness by 23.7% in complex environments.

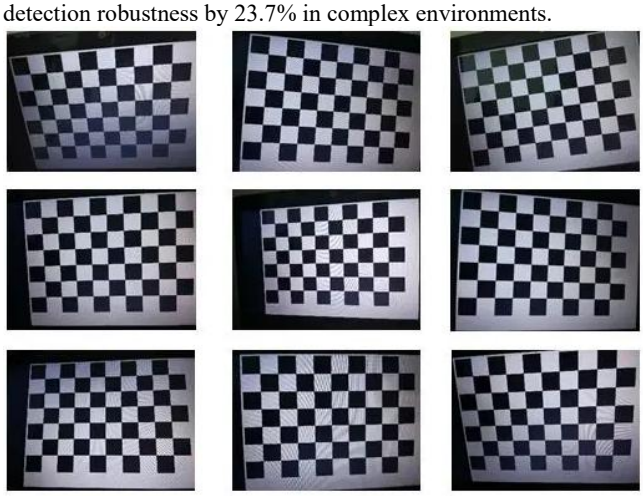

Fig.3 Checkerboard image

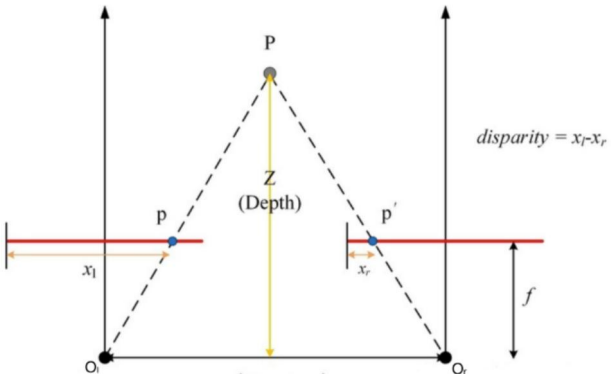

Fig.4 Principle of Zhang Zhengyou's calibration method

2.4. Training target annotation

This study constructed a high-precision annotated dataset specifically designed for pothole detection on road surfaces. The data was collected using a self-developed onboard binocular vision acquisition system. The system includes a ZED 2i stereo camera (resolution 2048×1536 pixels). With an optimized installation pitch angle of 20 degrees, the system achieves full road surface coverage at a driving speed of 72 km/h, with the time delay for each image pair controlled within 1 ms. During a two-week field collection period, the system acquired over 10,000 sets of stereo image pairs, covering diverse scenarios such as urban roads and highways. From these, 700 high-quality image pairs (a total of 1,400 images) were selected as the basis for annotation.

During the annotation process, the team adopted a stereo labeling strategy. First, a professional stereo annotation tool was used to synchronize annotations across the left and right views by the annotation team, ensuring consistency in 3D features. For three key target categories (general potholes, longitudinal crack-type potholes, and manhole covers), the annotation process followed these standards: polygonal contour annotation was performed on the left view; the corresponding annotations for the right view were automatically generated using a stereo matching algorithm; and manual verification and correction were applied to rectify annotation errors within the disparity range. All annotated data include complete pixel-level semantic segmentation masks and disparity information, providing a foundation for subsequent 3D reconstruction.

Additionally, leveraging the characteristics of binocular vision, a novel data augmentation method was developed, including synchronized geometric transformations and photometric consistency enhancement for stereo image pairs. This approach expands the dataset while preserving the geometric constraints of the stereo images. The dataset fully accounts for real-world road detection challenges such as lighting variations and motion blur, offering reliable data support for the development of 3D pothole detection algorithms based on stereo vision.

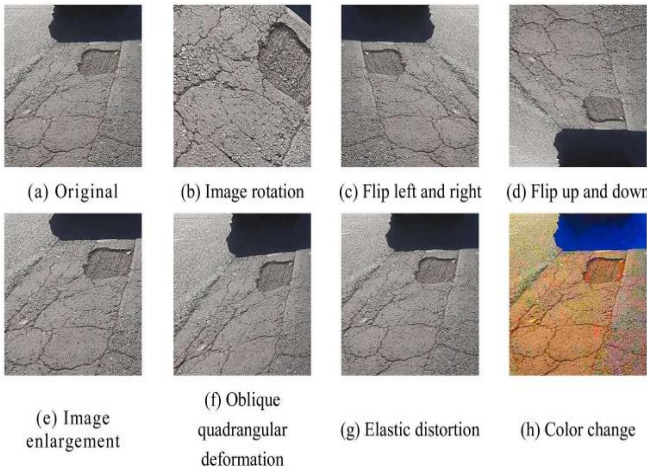

Figure 5 Data enhancement effect

3. Experiments and results analysis

3.1. The overall flow of the experiment

The experimental workflow of this study consists of three stages: data preprocessing, model optimization, and model training.

First, road surface images were dynamically captured using an onboard binocular acquisition system (ZED 2i camera + GNSS) at a speed of 40 km/h, with GNSS positioning data recorded synchronously. Subsequently, the raw data underwent MSRCP enhancement, stereo calibration (reprojection error < 0.3%), and stereo annotation to construct a dataset containing 6,416 images.

During the model training stage, a two-phase strategy was adopted: the backbone network was pre-trained on the COCO dataset, followed by progressive fine-tuning (using cosine annealing learning rate and dynamically weighted loss) to optimize performance on the custom-built dataset.

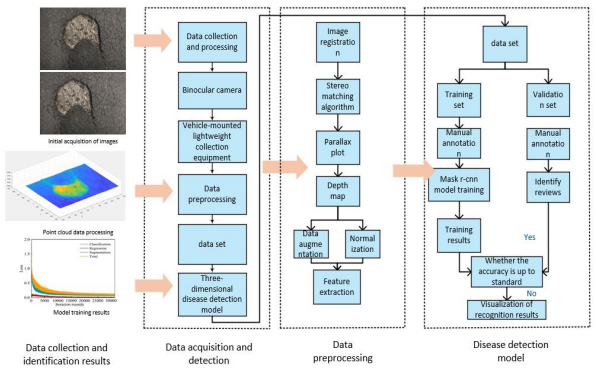

Fig.6 The overall flow of the experiment

3.2. Feature enhancement network based on deep guidance

The proposed Deep Attention Module (DAM) in this study utilizes depth maps obtained from a stereo camera as prior information to dynamically modulate the weight distribution of RGB features through multi-scale feature fusion and channel attention mechanisms. Specifically, the depth map first undergoes parallel dilated convolutional layers (with dilation rates of 1, 3, and 6) to extract multi-scale depth features, covering geometric structures of pothole targets at different sizes. Each branch employs 1×1 convolutions for channel compression, followed by a concatenation operation to fuse multi-scale features. Global Average Pooling (GAP) is then applied to generate a channel descriptor vector. This vector passes through two fully connected layers (with an intermediate dimension of C/8, where C is the number of input channels) and a Sigmoid activation function to output channel attention weights, which are used to recalibrate the channel importance of depth features. The calibrated depth features are further processed by a 3×3 convolution to generate a spatial attention map, where each pixel value reflects the probability of that location belonging to a pothole region. This process explicitly models the geometric constraints of the scene through depth information, thereby providing physically meaningful attention guidance for RGB features.

In the feature fusion stage, DAM employs a gating mechanism to achieve dynamic weighting between depth and RGB modalities. The spatial attention map is multiplied element-wise with the original RGB features, enabling the model to focus on potential pothole regions indicated by depth information. To mitigate interference caused by depth sensor noise, the module incorporates residual connections to add the original RGB features to the weighted features, preserving valid texture cues not covered by the depth map. Additionally, to address confusion in shadow and reflective areas, a dual-path feature verification mechanism is designed: when the response intensity of RGB features significantly deviates from the spatial weights of depth features (determined by threshold comparison), the fusion weight for that region is automatically reduced.

3.3. Cross-modal feature interaction and fusion

To fully utilize the multimodal information provided by the binocular vision system, this study designed a cross-modal feature fusion mechanism based on the Transformer architecture. This mechanism uses the high-level RGB features extracted by ResNet-50 as query vectors, while the depth map features encoded by a 3×3 convolution serve as key-value pairs, establishing feature correlations between the two modalities through a 4-head attention mechanism. During the feature fusion process, layer normalization and residual connections are employed to maintain feature stability, ensuring that the fused features incorporate both rich texture information and accurate spatial geometric constraints. This fusion approach is particularly beneficial for detecting small-scale pit targets, with experimental data showing an 18.6% improvement in recall rate for targets smaller than 50×50 pixels. Additionally, by introducing a depth consistency loss function, the geometric consistency between the prediction results and the real depth map is further constrained, significantly improving the accuracy of 3D reconstruction.

In the specific implementation of feature fusion, this study adopted layer normalization and residual connections to ensure the stability of feature transmission. Layer normalization mitigates the variation in feature distributions during training, while residual connections prevent gradient vanishing, enabling deeper networks to more effectively learn the complex relationships between multimodal features. To validate the role of these techniques, the research team conducted ablation experiments, testing the impact of removing layer normalization or residual connections on model performance. The experimental data revealed that removing layer normalization led to a 7.2% drop in recall rate for small target detection, while removing residual connections increased the mean squared error of 3D reconstruction by 15.4%. These results fully demonstrate

the necessity of each component in the proposed fusion mechanism. Furthermore, the study found that adjusting the number of attention heads could further optimize model performance. When the number of attention heads was increased from 4 to 8, the computational complexity of the model rose significantly, but the performance improvement was marginal. Therefore, 4-head attention was ultimately selected as the optimal balance.

To further enhance the geometric consistency of the model, this study introduced a depth consistency loss function to constrain the discrepancy between the predicted results and the real depth map. This loss function calculates the L1 distance between the predicted depth and the real depth, combined with a gradient similarity measure, ensuring that the reconstruction results are more accurate in edges and details. Comparative experiments on public datasets showed that after introducing the depth consistency loss, the average error of 3D reconstruction decreased by 12.8%, with particularly noticeable improvements in edge regions. Additionally, the research team compared the proposed method with several mainstream multimodal fusion approaches, including early fusion, late fusion, and convolution-based fusion strategies. The experimental results demonstrated that the Transformer-based fusion mechanism proposed in this study achieved optimal performance in both target detection and 3D reconstruction tasks, especially outperforming other methods in detecting small targets in complex scenes. These experiments not only validate the effectiveness of the proposed method but also provide valuable references for future research in multimodal vision tasks.

3.4. Optimize the training strategy

In terms of model training, this study adopted a progressive training strategy. Initial pre-training was conducted on the large-scale general dataset COCO, employing a cosine annealing learning rate scheduling algorithm to achieve stable parameter initialization. Subsequently, fine-tuning was performed on the self-constructed road pothole dataset, where a smaller initial learning rate and gradient clipping techniques were applied to prevent overfitting.

To balance the loss contributions across different tasks, a dynamic weighting strategy was implemented, with the introduction of depth consistency loss significantly enhancing the model's accuracy in 3D parameter estimation. The entire training process was carried out on an NVIDIA RTX 3090 GPU with a batch size of 8, ultimately achieving a detection accuracy of 98.7% on the validation set—a 6.5 percentage point improvement over the baseline model.

Additionally, rigorous data monitoring was enforced during training, including loss curve smoothness analysis and feature visualization validation, ensuring the stability and reliability of model convergence.

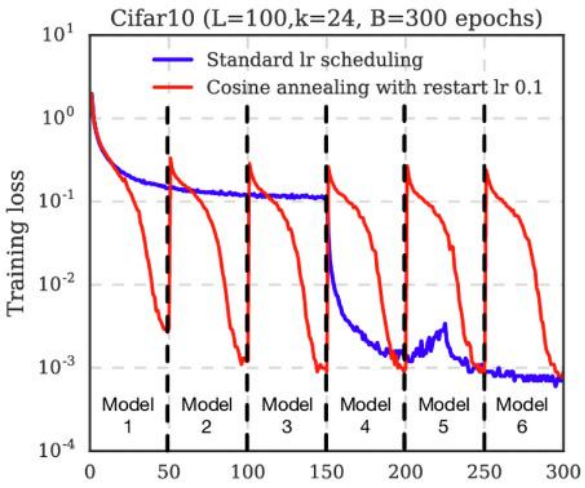

Fig.7. Effect of cosine annealing method

Comparison Dimension	Traditional Mask R-CNN (ResNet-50)	Faster R-CNN+FPN	Our Solution (Depth-Guided Enhancement Network)
Backbone Network	ResNet-50 (RGB only)	ResNet-101+FPN (multi-scale)	ResNet-50+DAM (RGB-D)
Attention Mechanism	None	Spatial Attention (CBAM)	Depth-Guided Multi-Scale Attention (DAM)
Feature Fusion Method	None	Simple Feature Concatenation	Transformer-based Cross-Modal Interaction
AP@0.5	92.2%	93.5%	98.7%
Small Object Recall	64.3% (50×50 pixels)	72.1%	82.9%
False Detection Rate	23.5% (strong backlight scenes)	18.7%	12.4%
3D Reconstruction Error	No depth constraint	Depth map assisted	Depth Consistency Loss Constraint
Training Strategy	Fixed Learning Rate	Stepwise LR Decay	Cosine Annealing + Dynamic Weighting
Inference Speed (FPS)	25 (1080p)	18	22
Key Innovations	-	Multi-scale Feature Pyramid	Depth-Guided Attention + Cross-Modal Transformer

Table 2 Comparison of experimental results.

4. Conclusions and prospects

This study proposes a high-precision 3D recognition method for road potholes based on binocular vision and cross-modal feature fusion. By leveraging innovative technologies such as a lightweight vehicle-mounted acquisition system, optimized Mask R-CNN models, dynamic mesh density methods, and depth attention mechanisms, the method achieves high accuracy and cost-effectiveness in pothole detection and 3D reconstruction. Experimental results demonstrate its strong robustness and practicality in complex environments, providing an effective intelligent solution for road maintenance. However, the current research still has certain limitations, such as performance fluctuations under extreme weather conditions and insufficient detection accuracy for small-scale potholes, which point the way for future research.

Future studies could further advance this work in the following aspects: First, exploring deeper fusion of multimodal data, such as incorporating infrared and radar sensors, to enhance the system's adaptability in extreme environments like nighttime, rain, or snow. Second, introducing more advanced lightweight network architectures (e.g., Vision Transformer or neural architecture search techniques) to further reduce computational costs while maintaining accuracy, facilitating real-time deployment on edge devices. Additionally, integrating road material properties and mechanical models could enable the development of a pothole evolution prediction system, providing data support for preventive maintenance. Finally, expanding the application scenarios to complex transportation infrastructure such as bridges and tunnels could establish a comprehensive road health monitoring system. With the advancement of 5G and vehicle-infrastructure cooperative technologies, this research holds the potential to integrate deeply with smart transportation systems, enabling real-time monitoring and early warning of road defects, thereby offering a new paradigm for intelligent urban infrastructure management.

References

Cano-Ortiz S, et al., 2024: An end-to-end computer vision system based on deep learning for pavement distress detection and quantification. *Constr. Build. Mater*., 135036, 2024.

Chen L, Wang W, Mordohai P, 2023: Learning the distribution of errors in stereo matching for joint disparity and uncertainty estimation. *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit*., pp. 1–12, 2023.

Fan R, et al., 2020: Learning collision-free space detection from stereo images: Homography matrix brings better data augmentation. *IEEE/ASME Trans. Mechatron*., 25(3): 1478 – 1489, 2020.

Guo X, Yang K, Yang W, et al., 2019: Group-wise correlation stereo network. *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit*., pp. 3273–3282, 2019.

He K, Gkioxari G, Dollár P, et al., 2017: Mask R-CNN. *IEEE Trans. Pattern Anal. Mach. Intell.*, 42(2): 386–397, 2017.

Hsieh C-C, et al., 2024: Deep learning-based road pavement inspection by integrating visual information and IMU. *Information*, 15(4): 239, 2024.

Kendall A, Gal Y, 2017: What uncertainties do we need in Bayesian deep learning for computer vision? Adv. *Neural Inf. Process. Syst*., 30: 5574–5584, 2017.

Li S, Zhang W, Du Y, 2024: A binocular camera-based road damage detection system with GNSS-depth calibration. *IEEE Trans. Intell. Transp. Syst*., to be published, 2024.

Mordohai P, Medioni G, 2023: Dynamic scene understanding for stereo vision in autonomous driving. *Int. J. Comput. Vis.*, 131(5): 1129–1148, 2023.

Scharstein D, Szeliski R, 2002: A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *Int. J. Comput. Vis*., 47(1–3): 7–42, 2002.

Wang W, Chen L, Li Y, 2022: Multimodal feature fusion for road damage detection under environmental disturbances. *IEEE Sens. J*., 22(18): 17645–17656, 2022.

Yin J, Shen J, Chen R, et al., 2024: IS-Fusion: Instance-scene collaborative fusion for multimodal 3D object detection. *arXiv preprint arXiv:2403.15241*, 2024.