buildings

MDPI

Article

Safety Helmet-Based Scale Recovery for Low-Cost Monocular 3D Reconstruction on Construction Sites

Jianyu Ren [1], Lingling Wang [2], Xuanxuan Liu [2] and Linghong Zeng [1,*]

[1] School of Civil Engineering, Hunan University, Changsha 410082, China; renjianyu66@163.com
[2] Economic and Technology Research Institute, State Grid Hunan Electric Power Company Ltd., Changsha 410000, China; wangll1090196533@126.com (L.W.); echo58613922@outlook.com (X.L.)
* Correspondence: mikema21@hnu.edu.cn

Abstract

Three-dimensional (3D) reconstruction is increasingly being adopted in construction site management. While most existing studies rely on auxiliary equipment such as LiDAR and depth cameras, monocular depth estimation offers broader applicability under typical site conditions, yet it has received limited attention due to the inherent scale ambiguity in monocular vision. To address this limitation, this study proposes a safety helmet-based scale recovery framework that enables low-cost, monocular 3D reconstruction for construction site monitoring. The method utilizes safety helmets as exemplary scale carriers due to their standardized dimensions and frequent appearance in construction scenes. A Standard Template Library (STL) comprising multi-angle safety helmet masks and dimensional information is established and linked to site imagery through template matching. Following three-dimensional scale recovery, multi-frame fusion is applied to optimize the scale factors. Experimental results on multiple real construction videos demonstrate that the proposed method achieves high reconstruction accuracy, with a mean relative error below 10% and outliers not exceeding 5%, across diverse construction environments without site-specific calibration. This work aims to contribute to the practical application of monocular vision in engineering management by leveraging ubiquitous site objects as metrological references.

Keywords: monocular depth estimation; 3D reconstruction; safety helmet; scale recovery; template matching; construction sites

check for updates

Academic Editor: Antonio Caggiano

Received: 16 October 2025
Revised: 16 November 2025
Accepted: 24 November 2025
Published: 26 November 2025

Citation: Ren, J.; Wang, L.; Liu, X.; Zeng, L. Safety Helmet-Based Scale Recovery for Low-Cost Monocular 3D Reconstruction on Construction Sites. *Buildings* **2025**, *15*, 4291. https://doi.org/10.3390/buildings15234291

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Intelligentization represents a prevailing trend in contemporary engineering, and visual perception technology, which enables machines to interpret, analyze, and respond to visual data, has become increasingly essential in modern construction management [1,2]. The integration of computer vision, machine learning, and artificial intelligence has established visual perception as an essential tool for enhancing engineering efficiency, ensuring construction safety, and optimizing resource allocation [3–5]. By analyzing on-site image data, these technologies provide comprehensive insights into complex environments, support proactive risk prediction, and facilitate informed decision-making [6,7]. They have also demonstrated significant potential in improving quality control [8], monitoring construction progress [9], and enhancing on-site safety management [10].

While visual perception technologies have demonstrably transitioned from purely theoretical investigations to validated practical applications, a substantial proportion of existing research within the construction domain predominantly still relies on two-dimensional

(2D) image data for the execution of management and monitoring tasks [11]. Although mature computer vision algorithms are capable of detecting and precisely localizing key entities such as workers, heavy machinery, and construction materials from 2D imagery, the critical absence of absolute depth and metric spatial information frequently culminates in notable geometric discrepancies and inevitably leads to inaccurate quantitative analyses. Consequently, the robust integration of three-dimensional (3D) reconstruction capabilities into vision-based site monitoring frameworks has emerged as a crucial and necessary undertaking toward fully realizing the concept of intelligent, smart construction sites [12]. Accurate 3D reconstruction provides indispensable spatial metrics, including the distribution of construction assets, the precise spatial dimensions of both temporary and permanent structures, and process-level activity and productivity indicators. This foundational data intrinsically supports key management processes, encompassing project progress tracking, stringent safety supervision, and comprehensive productivity analysis [13].

However, construction sites are typically characterized by their inherently large spatial scales, pronounced cost sensitivity, and significantly dynamically changing environments [14]. Consequently, traditional monocular cameras inherently suffer from the scale-depth ambiguity (lack of absolute scale) in 3D reconstruction, meaning that depth maps derived from single images are fundamentally incapable of providing an absolute metric scale for spatial measurements [15–17], as shown in Figure 1. This issue is further compounded in dynamic environments, where standard monocular Structure from Motion (SfM) or Simultaneous Localization and Mapping (SLAM) processes are highly susceptible to scale drift. Scale drift manifests as the cumulative propagation of error in relative depth and translation estimates over time, ultimately leading to a gradual but critical misrepresentation of the absolute scale throughout the reconstructed scene. To convert this relative depth information into absolute metric measurements, a process technically known as scale recovery is mandated, which necessitates the introduction of external or learned metric references to rigorously restore the absolute spatial dimensions. While alternative sensory technologies, such as LiDAR and dedicated depth cameras, can intrinsically resolve the scale issue, their application is frequently restricted due to high capital costs, the necessity for frequent and specialized calibration, and their intrinsic sensitivity to adverse lighting conditions and weather disturbances [18,19].

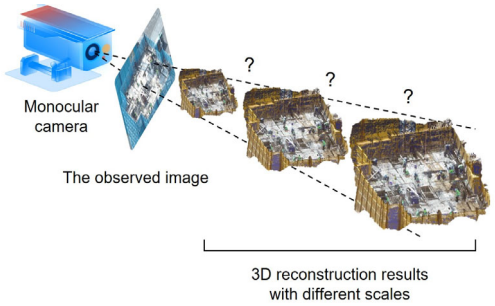


Figure 1. Illustration of scale ambiguity in monocular 3D reconstruction.

To address this problem, researchers have explored a wide range of monocular depth estimation (MDE) and scale recovery methods. Classical approaches such as Structure from Motion (SfM) [20], Shape from Shading (SfS) [21,22], and Markov Random Fields (MRF) [23] estimate relative depth through geometric and photometric constraints. More recently, deep learning-based models [24–26] have achieved impressive results by inferring depth from learned visual cues. Several studies have incorporated scale recovery modules into deep monocular frameworks—for instance, Yin et al. (2017) combined depth estimation with environmental structure priors [27]; Wagstaff and Kelly (2021) introduced a scale

recovery loss based on known camera height [28]; and Shen et al. (2023) predicted per-pixel absolute depth via self-supervised learning [29]. These methods improve depth accuracy but typically require either known geometric references or external sensor data, limiting their suitability for low-cost construction environments. Despite these advancements, most approaches still depend on pre-known geometric parameters or specialized sensors, limiting their practical use in cost-sensitive construction environments.

In the construction industry, various 3D reconstruction frameworks have been applied to enhance automation and efficiency [30]. Sung and Kim (2016) developed a stereo-based terrain reconstruction system using multi-scale descriptors and cascade matching [13]. Golparvar-Fard et al. (2015) integrated SfM and multi-view stereo (MVS) reconstruction with semantic classification to automate asset documentation [31]. More recently, active sensing approaches have demonstrated high-precision 3D perception capabilities. Dual-frequency LiDAR with compressed-sensing imaging has enabled efficient non-scanning depth acquisition [32], and LiDAR-camera fusion robotic platforms have achieved sub-millimeter crack measurement accuracy in structural inspection [33]. However, such systems typically require specialized hardware, precise calibration, and increased deployment cost. In practice, most site monitoring systems still rely on monocular cameras due to their low cost, portability, and ease of installation. This creates a persistent limitation: monocular methods lack inherent metric scale, which restricts their accuracy without additional sensors. To address this challenge, recent studies have explored monocular-based enhancement strategies, such as normalized geometry learning for scale-aware point clouds and lightweight monocular pipelines optimized for deployment efficiency [34,35].

Despite these advances, two critical gaps remain. First, existing monocular methods are rarely adapted to construction environments, where occlusion, motion, and uneven illumination complicate depth estimation. Second, current research often prioritizes algorithmic accuracy over practical deployability, overlooking cost constraints and sensor availability [30]. As a result, a scalable and low-cost solution for metric reconstruction using readily available visual sensors remains lacking.

To bridge these gaps, this study proposes a low-cost monocular 3D reconstruction framework with scale recovery based on standardized scale carriers. The main contributions of this study are summarized as follows:

1. A low-cost and infrastructure-compatible framework for metric 3D reconstruction using ordinary monocular surveillance systems, eliminating the need for LiDAR or depth cameras.

2. A construction-specific scale carrier dataset derived from standardized safety helmets, enabling the creation of a reusable scale template library applicable across diverse environments.

3. A robust scale optimization algorithm that employs multi-frame fusion and temporal regularization to enhance scale stability and accuracy in dynamic site conditions.

These advances establish a new paradigm for cost-sensitive site digitalization, wherein commonly available objects are repurposed as intrinsic calibration targets for achieving metric-level reconstruction accuracy.

2. Methods

The proposed framework integrates advanced computer vision techniques, such as object detection, instance segmentation, and depth estimation. As illustrated in Figure 2, the overall architecture comprises five primary processing stages: detection and segmentation, template matching, scale recovery, scale optimization, and 3D reconstruction.

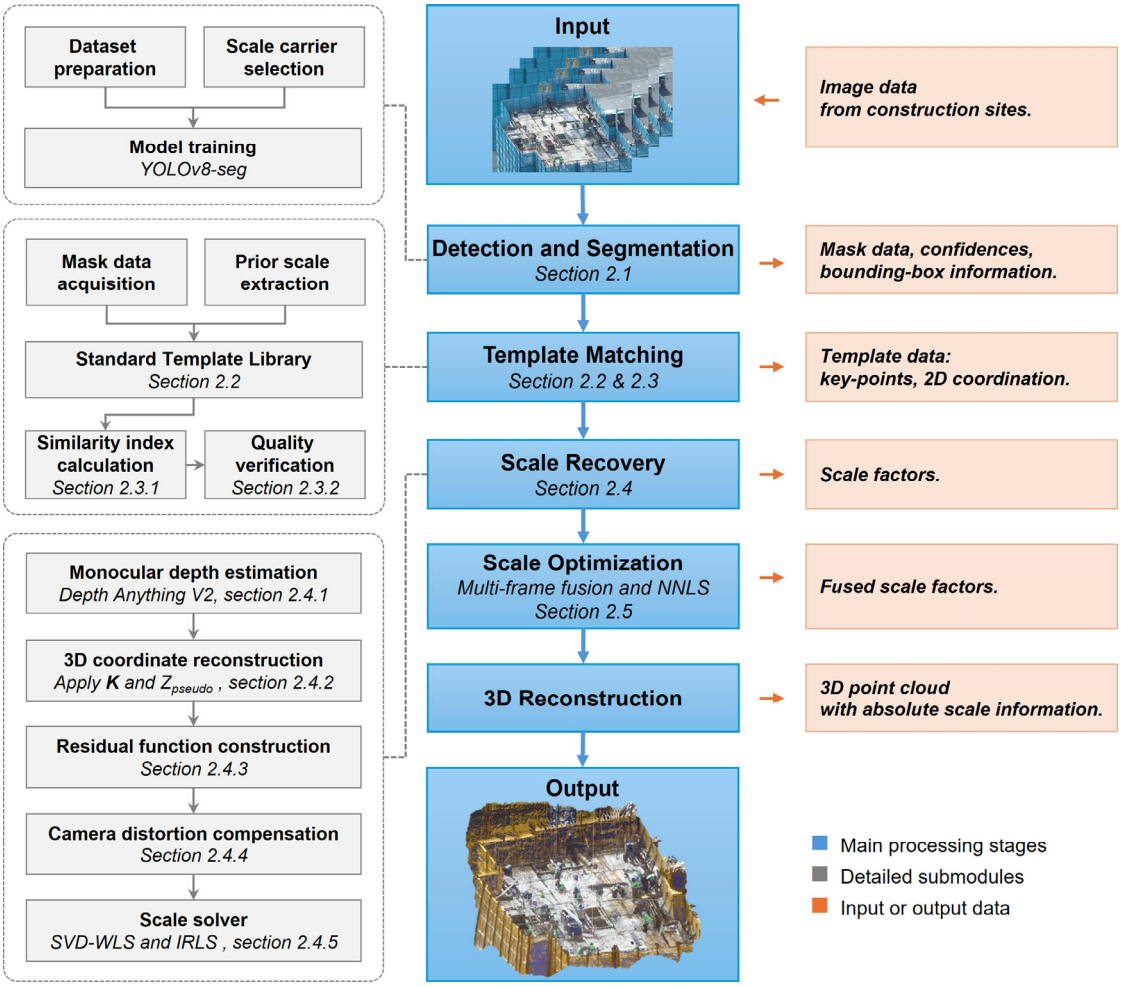


Figure 2. Overview of the proposed monocular 3D reconstruction framework with integrated scale recovery and optimization.

2.1. Detection and Segmentation of Scale Carriers

The foundational step of this framework is the selection of an appropriate scale carrier. An ideal carrier must satisfy four essential criteria: (1) high prevalence on-site (e.g., bricks, scaffold joints), (2) standardized dimensions (e.g., rebar diameters), (3) distinct geometric features for detectability, and (4) adequate appearance asymmetry to avoid ambiguity in geometric matching (thereby excluding highly symmetrical objects like spheres or pipes).

The core technical task, therefore, is the precise detection and instance segmentation of the scale carrier, which serves as the input for both STL creation and subsequent scale recovery. For this task, the YOLOv8 Instance Segmentation model (YOLOv8-seg) was employed. This model was chosen as it represents a state-of-the-art balance of high accuracy and computational efficiency, making it well-suited for complex construction site imagery.

To ensure robust performance on the target objects, the YOLOv8-seg model was fine-tuned on our custom dataset (detailed in Section 3.1). The overall detection and segmentation workflow, including dataset preparation, scale carrier selection, network architecture, training configuration, and primary outputs, is illustrated in Figure 3.

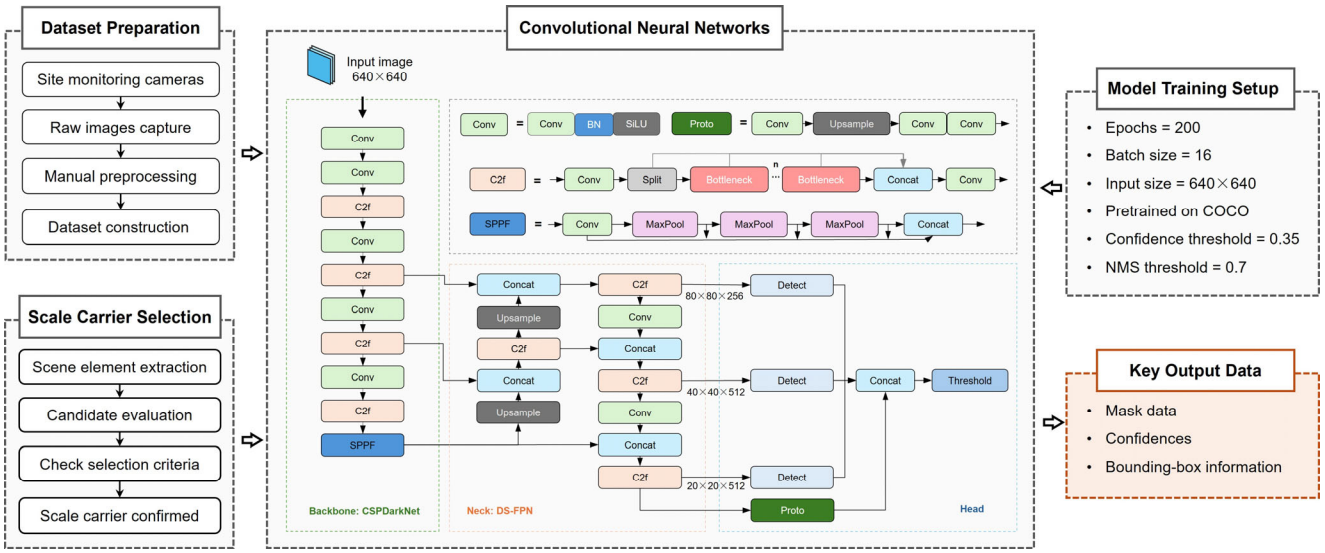


Figure 3. Workflow of the detection and segmentation stage.

2.2. Development of the Standard Template Library (STL)

The STL comprises image data of scale carriers captured in a controlled laboratory environment using a monocular camera at multiple viewing angles. Corresponding mask data are algorithmically generated from these images via computer vision techniques.

Variations among individual instances of the same category of objects on construction sites can be substantial. To counteract these factors, this study employs mask images of library objects to construct the template library and facilitate subsequent template matching processes.

The input data for the scale recovery model is derived from the actual distances between selected key-point pairs on the standard template and their corresponding coordinates in the pixel coordinate system. An equation system is constructed based on multiple key-point pairs from the same object type within the same scene. At least 3 key-point pairs from different dimensions are required to form a solvable equation.

Object detection and algorithms detect multiple instances of objects j within the i-th frame. For each object, the detection confidence $C_{det,j}$ is returned, and the template matching confidence is $C_{match,j}$. Predefined key-points in the template are mapped to corresponding image coordinates $(x_{i,j,1}, y_{i,j,1})$ and $(x_{i,j,2}, y_{i,j,2})$. The confidence for each key-point $C_{i,j}$ is determined by the product of the object detection confidence $C_{det,j}$ and the template matching confidence $C_{match,j}$, as shown in Equation (1).

$$C_{i,j} = C_{\mathrm{det},j} \times C_{match,j} \tag{1}$$

Only key-points with confidence $C_{i,j} \geq Cth$ are selected for subsequent scale factor estimation, where Cth is a preset confidence threshold.

2.3. Template Matching

2.3.1. Similarity Index Calculation

Input site imagery undergoes detection and segmentation of scale carriers. Segmented instances are matched against standardized templates in the library. The similarity between a detected object and each template in the STL is evaluated using the Intersection over

Union (IoU). Let I_n and I_m represent the mask pixel area of the target image and the m-th template, respectively. The $\mathrm{IoU_{tm}}$ is then computed as shown in Equation (2):

$$\mathrm{IoU_{tm}} = \frac{|I_n \cap I_m|}{|I_n \cup I_m|} \tag{2}$$

To improve robustness while maintaining computational efficiency, an additional normalized centroid-distance term is incorporated into the matching score. Let $c(I_n)$ and $c(I_m)$ denote the centroids of the detected mask and the m-th template, respectively, and $D_{\max}$ be the image diagonal length. The centroid distance is normalized as:

$$d_{norm} = \frac{\|c(I_n) - c(I_m)\|}{D_{\max}} \tag{3}$$

The final similarity index (SI) is computed as:

$$SI = \alpha \cdot \mathrm{IoU_{tm}} + \beta \cdot (1 - d_{norm}) \tag{4}$$

where α and β are weighting coefficients ($\alpha = 0.7, \beta = 0.3$ in our experiments). The template yielding the highest SI is selected as the optimal match. This matching method is primarily designed for static camera setups, where the range of viewpoint variation is limited, thereby reducing the likelihood of spurious matches caused by similar contour overlaps.

2.3.2. Matching Quality Verification

Construction sites present dynamically changing and visually cluttered environments, where object occlusion, reflection, and segmentation uncertainty may compromise the reliability of template matching. To enhance robustness and ensure consistent metric accuracy, a quality verification procedure is incorporated as a mandatory pre-step before the scale recovery stage.

This verification module evaluates both the confidence of the YOLOv8 detection results and the SI consistency between detected masks and their matched templates.

Specifically:

1. Detection confidence threshold: Detections with a confidence score below 0.80 are discarded.

2. Template SI threshold: For each remaining instance, SI must exceed 0.85 to be considered valid.

These thresholds were determined empirically through a series of controlled experiments conducted on real construction-site videos with varying occlusion levels. The selected parameters provide an optimal trade-off between data retention and robustness. Instances meeting both conditions are then forwarded to the scale recovery stage.

2.4. Scale Recovery Model

Monocular depth estimation inherently introduces non-uniform scaling and deformation along the three spatial axes, namely the horizontal (X), vertical (Y), and depth (Z), when reconstructing 3D scenes from single-view images. To correct these distortions and restore the true metric dimensions, a scale recovery model is established. The overall process consists of the following steps.

2.4.1. Depth Estimation Process

Assuming the use of a pinhole camera model, image coordinates can be transformed into 3D spatial coordinates through the camera intrinsic matrix $\boldsymbol{K}$, given by Equation (5).

$$\boldsymbol{K} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 0 \end{bmatrix} \tag{5}$$

where f_x and f_y are the focal lengths in the horizontal and vertical directions, and (c_x, c_y) represents the principal point of the image. The depth map is computed where each pixel's depth Z_{pseudo} corresponds to the distance between the camera and the pixel (i, j), given by Equation (6).

$$Z_{pseudo,i,j,k} = f_{depth}(I) \tag{6}$$

where I is the input image, and f_{depth} represents the depth estimation model (Depth Anything V2 [36]).

2.4.2. Three-Dimensional Coordinate Calculation

Based on the pixel coordinates and the monocular pseudo-depth, the 3D coordinates of points can be recovered. Let ($x_{i,j,k}$, $y_{i,j,k}$) (k = 1, 2) denote the pixel coordinates of a key-point pair on object j in frame i. The corresponding normalized image coordinates are computed as:

$$\widetilde{x}_{i,j,k} = \frac{x_{i,j,k} - c_x}{f_x}, \widetilde{y}_{i,j,k} = \frac{y_{i,j,k} - c_y}{f_y} \tag{7}$$

where f_x, f_y and c_x, c_y are camera intrinsic parameters. With the estimated pseudo-depth $Z_{i,j,k}$, the reconstructed 3D point is expressed as:

$$X_{i,j,k} = s_x \widetilde{x}_{i,j,k} Z_{pseudo,i,j,k}, Y_{i,j,k} = s_y \widetilde{y}_{i,j,k} Z_{pseudo,i,j,k}, Z_{i,j,k} = s_z Z_{pseudo,i,j,k} \tag{8}$$

where s_x, s_y, and s_z are scale factors along the horizontal, vertical, and depth axes, respectively.

2.4.3. Residual Function Construction

Given pairs of points ($x_{i,j,k}$, $y_{i,j,k}$) (k = 1, 2), the real-world distance L_{real} between them is known. Equation (9) shows the Euclidean distance between the 3D coordinates of these key-points.

$$D_{i,j} = \sqrt{(X_{i,j,1} - X_{i,j,2})^2 + (Y_{i,j,1} - Y_{i,j,2})^2 + (Z_{i,j,1} - Z_{i,j,2})^2} \tag{9}$$

In order to simplify the optimization problem, the residual function in Equation (10) has been modified.

$$R_{i,j} = \left(L_{real,j}^2 - D_{i,j}^2\right) \times C_{i,j} \tag{10}$$

where $R_{i,j}$ represents the residual for objects j in the i-th frame. Then, incorporating the simplified expression, the total residual R for all key-point pairs that meet the confidence requirement is expressed as Equation (11).

$$R = \sum_i \sum_j R_{i,j}^2 \tag{11}$$

Subsequently, by constructing a linear least squares problem using the total residual function, the scale factors S_x, S_y, and S_z can be solved by minimizing the total residual R, as expressed in Equation (12).

$$\min_{S_x,S_y,S_z} R = \sum_i \sum_j \left(r_{i,j} - a_{i,j}S_x - b_{i,j}S_y - c_{i,j}S_z\right)^2$$
$$\begin{cases} r_{i,j} = L_{real,j}^2 \times C_{i,j} \\ a_{i,j} = \left[(x_{i,j,1} - c_x) \times Z_{pseudo,i,j,1} - (x_{i,j,2} - c_x) \times Z_{pseudo,i,j,2}\right]^2 \times C_{i,j} \\ b_{i,j} = \left[(y_{i,j,1} - c_y) \times Z_{pseudo,i,j,1} - (y_{i,j,2} - c_y) \times Z_{pseudo,i,j,2}\right]^2 \times C_{i,j} \\ c_{i,j} = \left(Z_{pseudo,i,j,1} - Z_{pseudo,i,j,2}\right)^2 \times C_{i,j} \\ S_x = \left(\frac{f_x}{s_x}\right)^2, S_y = \left(\frac{f_y}{s_y}\right)^2, S_z = (s_z)^2 \end{cases} \quad (12)$$

Expressing (11) in matrix form, it can be represented as Equation (13).

$$\boldsymbol{R} = \|\boldsymbol{r}_{N\times 1} - \boldsymbol{A}_{N\times 3} \times \boldsymbol{S}_{3\times 1}\|^2 \quad (13)$$

where $\boldsymbol{r} = [r_{1,1}, r_{1,2}, \ldots, r_{i,j}]^{\mathrm{T}}$ is an $N \times 1$ vector containing all $r_{i,j}$, $\boldsymbol{A}$ is an $N \times 3$ matrix with columns $a_{i,j}$, $b_{i,j}$, and $c_{i,j}$, and $\boldsymbol{S}$ is a 3×1 vector, $S = \left[S_x, S_y, S_z\right]^T$. Here, N denotes the total number of valid point pairs across all frames. This formulation ensures that each geometric constraint contributes one independent row in $\boldsymbol{A}$, allowing the scale vector $\boldsymbol{S}$ to be estimated by solving the over-determined system.

Finally, the linear system is solved using a numerically stable SVD-based weighted least-squares solver to avoid conditioning issues and ensure robust and physically consistent scale estimates.

2.4.4. Camera Distortion Robustness

In construction site deployments, commercial Closed-Circuit Television (CCTV) cameras may exhibit mild radial and tangential distortion. Although the pinhole model used in Section 2.4.1 assumes ideal projection, real pixel coordinates can be modeled by the Brown-Conrady formulation, as shown in Equation (14).

$$\begin{cases} x_{dist} = x(1 + k_1r^2 + k_2r^4) + 2p_1xy + p_2(r^2 + 2x^2) \\ y_{dist} = y(1 + k_1r^2 + k_2r^4) + 2p_2xy + p_1(r^2 + 2y^2) \\ r^2 = x^2 + y^2 \end{cases} \quad (14)$$

where k_1 and k_2 are radial distortion coefficients and p_1, p_2 are tangential distortion coefficients. These equations describe the mapping between undistorted image coordinates (x, y) and distorted coordinates (x_{dist}, y_{dist}).

To evaluate the influence of distortion on scale estimation, synthetic radial distortion perturbations $k_1 = \pm 0.0005, \pm 0.001$ were applied to representative image coordinates, corresponding to typical CCTV lenses. When distortion parameters are available, images can be rectified prior to depth inference using a standard calibration workflow:

1. Capture checkerboard images from multiple viewpoints.
2. Estimate intrinsic parameters and distortion coefficients (K, k_1, k_2, p_1, p_2).
3. Rectify input images (OpenCV undistort).
4. Perform monocular depth estimation.
5. Recover metric scale $\left(S_x, S_y, S_z\right)$.

The compatibility with standard calibration tools further supports the practical deployability of the method.

2.4.5. Numerically Stable Scale Solver

As discussed in Section 2.4.3, the scale factors S_x, S_y, and S_z are estimated by minimizing the total residual function derived from key-point distance consistency constraints. The resulting least-squares problem can be expressed in matrix form. Although the corresponding normal equation can be formed as $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}S = \boldsymbol{A}^{\mathrm{T}}r$, directly solving this expression is known to suffer from numerical instability, particularly when $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$ exhibits ill-conditioning due to noise in depth estimation or pixel localization.

To ensure numerical robustness and physically consistent solutions, the scale vector $S = [S_x, S_y, S_z]^T$ is computed using a singular-value decomposition (SVD)-based weighted least-squares solver. The SVD factorization of matrix $\boldsymbol{A}$ is expressed as:

$$\boldsymbol{A} = \boldsymbol{U}\Sigma\boldsymbol{V}^T \tag{15}$$

where $\boldsymbol{U}$ and $\boldsymbol{V}$ are orthogonal matrices, and Σ is a diagonal matrix containing the singular values of $\boldsymbol{A}$. The scale vector is then obtained by solving:

$$S = \boldsymbol{V}\Sigma^{-1}\boldsymbol{U}^T r \tag{16}$$

which provides a stable closed-form solution without explicitly forming the normal equations. To reduce the influence of outlier key-point pairs arising from occlusions, imperfect detections, or depth inference noise, confidence weights are further incorporated into the residual formulation. The weighted least-squares objective is expressed as:

$$\underset{S}{\operatorname{argmin}} \sum_{i,j} w_{i,j} \cdot r_{i,j}^2 \tag{17}$$

where $w_{i,j}$ represents the confidence weight associated with the corresponding point pair. In addition, a Huber iterative reweighted least-squares (IRLS) step is applied to further mitigate the effect of large residuals while preserving sensitivity to inliers. The Huber loss function is expressed as:

$$\rho(r) = \begin{cases} \frac{1}{2}r^2, \left|r\right| \leq \delta \\ \delta\left(\left|r\right| - \frac{1}{2}\delta\right), \left|r\right| > \delta \end{cases} \tag{18}$$

where δ is the transition threshold between quadratic and linear penalization. This formulation retains the efficiency of least-squares methods while introducing robustness against measurement uncertainty.

The proposed SVD-based weighted IRLS method ensures that the estimated scale factors S_x, S_y, and S_z remain non-negative, numerically stable, and physically meaningful.

2.5. Scale Optimization Based on Multi-Frame Fusion

To ensure stable and consistent scale estimation in dynamic construction environments, the proposed framework employs a multi-frame fusion strategy for scale factor optimization. Instead of relying on a single-frame estimate, temporal information across consecutive frames is integrated to suppress random fluctuations and compensate for occasional detection errors.

Denoting the estimated scale factor at frame t as $S_t = [S_x^t, S_y^t, S_z^t]^T$, a direct frame-wise estimation may be sensitive to detection noise and local occlusion. To enhance robustness, a temporal regularization term is introduced to enforce smoothness over time, as shown in Equation (19).

$$S_t = \min\left(\|A_t S_t - r_t\|_2^2 + \lambda_1\|S_t - S_{t-1}\|_2^2 + \lambda_2\|S_t - S_0\|_2^2\right) \tag{19}$$

where A_t and r_t represent the scale constraint matrix and reference vector for the current frame, S_{t-1} is the previous fused estimate, S_0 is the initialization prior (typically 1,1,1), and λ_1, λ_2 are temporal and regularization weights. The optimization is solved efficiently using non-negative least squares (NNLS) to ensure physical validity of the scale ($S_t \geq 0$).

After obtaining frame-wise estimates, multi-frame fusion is applied within a sliding temporal window of length W. For each frame t, the fused scale $\hat{S}_t$ is computed as a weighted average, as shown in Equation (20).

$$\hat{S}_t = \frac{\sum_{i=t-W+1}^{t} w_i S_i}{\sum_{i=t-W+1}^{t} w_i} \tag{20}$$

where the weights w_i are inversely proportional to the local residual variance σ_i^2, as shown in Equation (21).

$$w_i = \frac{1}{\sigma_i^2 + \varepsilon} \tag{21}$$

This adaptive weighting scheme down-weights unreliable frames affected by occlusion, motion blur, or detection instability. To automatically determine convergence, an uncertainty-based criterion is defined, as shown in Equation (22).

$$\mathrm{Conv(t)} = \begin{cases} \text{True, if } \mathrm{std}(\hat{S}_{t-W+1:t}) < \tau \\ \text{False, otherwise} \end{cases} \tag{22}$$

where τ is a predefined stability threshold (typically 1–2% of the mean scale). Once the convergence condition holds for consecutive frames, the scale is considered stable and used for subsequent 3D reconstruction.

Compared with single-frame estimation, the proposed multi-frame fusion strategy (1) mitigates noise amplification caused by monocular depth uncertainty, (2) provides temporal consistency across dynamic scenes, and (3) enables automatic confidence assessment of the reconstructed scale. The resulting fused scale values exhibit improved stability and accuracy.

3. Case Study

This section presents an experimental validation of the proposed safety helmet-based scale recovery framework for low-cost monocular 3D reconstruction.

3.1. Dataset Preparation and Scale Carrier Selection

This study utilizes two specialized datasets for methodology validation. The lab-dataset comprises 315 multi-angle images of scale carriers acquired under controlled laboratory conditions using a monocular camera (Hangzhou Hikvision Digital Technology Co., Ltd., Hangzhou, China), generating 315 precise annotations. Complementarily, the site-dataset incorporates 2412 construction site images with 5736 annotations, processed through keyframe extraction and spatial normalization to establish task-specific inputs.

Safety helmets are strategically selected as the exemplar scale carrier based on three critical attributes: standardized dimensions enforced by industrial specifications (ISO 3873-1977), near-ubiquitous presence (over 95% worker occurrence frequency in site observations), and geometric stability enabling robust feature extraction.

To validate this selection, multiple potential scale carriers commonly present on construction sites, such as safety helmets, traffic cones, barrels, and toolboxes, were comparatively analyzed to assess their suitability for reliable scale recovery. Among these

candidates, safety helmets demonstrated the most stable performance. They are dimensionally standardized across manufacturers, exhibit distinctive geometric features, and remain consistently visible in site videos due to mandatory worker usage. These characteristics ensure high detection confidence and accurate extraction of their physical dimensions.

In contrast, comparative trials revealed that alternative objects were substantially less reliable. Toolboxes displayed inconsistent bounding-box scaling due to size variability, while geometrically similar items such as barrels and pipes frequently caused template-assignment ambiguity. In these cases, the incorrect matching rate reached up to 80%, and the key-point extraction error rate up to 68%, particularly under variable illumination or partial occlusion. These results confirm that safety helmets provide an optimal balance between accessibility, geometric distinctiveness, and metric reliability for monocular scale recovery.

Utilizing the specific training parameters detailed in Section 2.1, the YOLOv8 model was trained on the custom dataset for detecting and segmenting scale carriers. As shown in Figure 4, the model achieved optimal performance at 200 epochs, with box-level results of Loss(B) = 0.843, mAP@0.5(B) = 0.834, and mAP@[0.5:0.95] = 0.625, and segmentation-level results of Loss(M) = 0.831, mAP@0.5(M) = 0.786, and mAP@[0.5:0.95] = 0.552. In addition, the segmentation IoU evaluated on the held-out test set reached 0.749, indicating strong overall segmentation performance. Representative samples from both datasets are visualized in Figure 5.

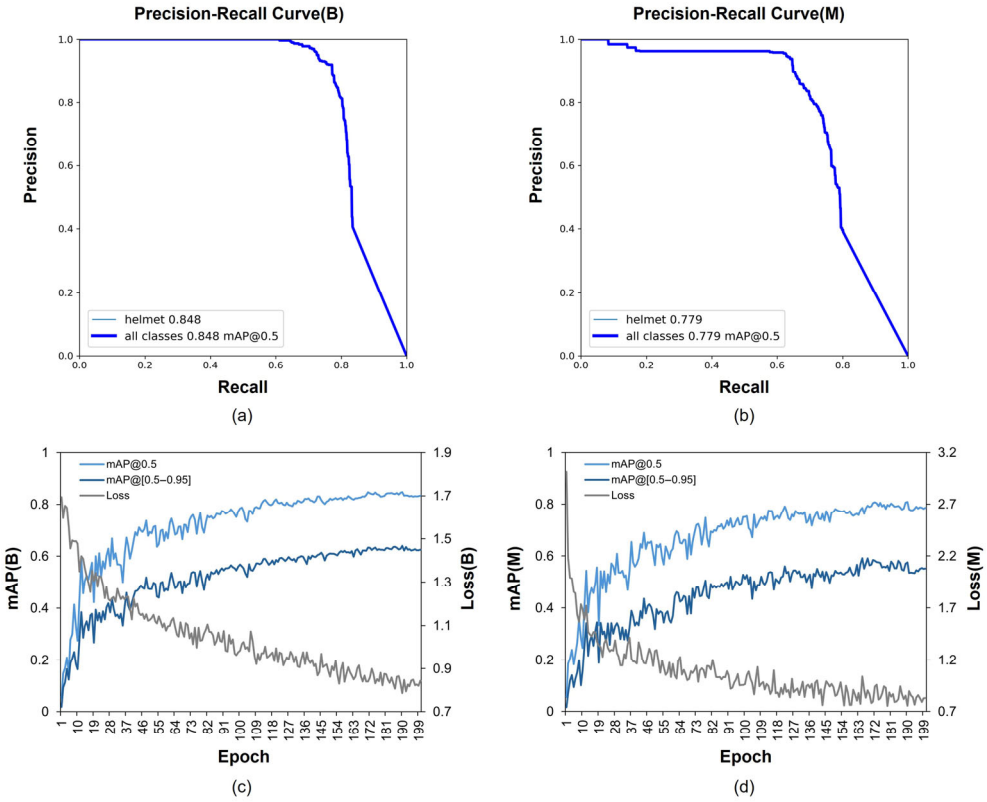


Figure 4. Training performance of the YOLOv8-seg model on the custom dataset. (**a**) Precision–Recall curve (box). (**b**) Precision–Recall curve (mask). (**c**) Loss and mAP evolution for bounding-box predictions. (**d**) Loss and mAP evolution for mask predictions.

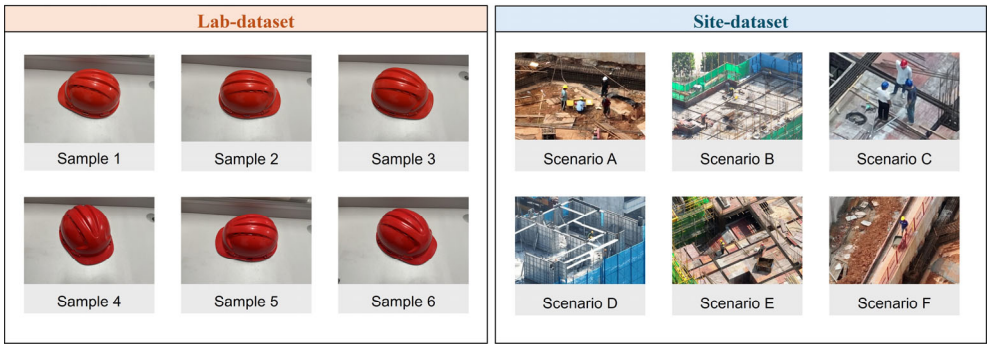


Figure 5. Representative samples from the lab and site datasets used in this study.

Alternative approaches, such as YOLOv8-Pose, can directly detect predefined key-points without the need for instance segmentation and template matching. However, for safety helmets, the low intra-class texture variation and smooth surface characteristics limit the spatial generalization capability of such key-point detectors, particularly under varying lighting and occlusion conditions. In contrast, our template-based matching approach, which operates on precise mask shapes obtained from instance segmentation, demonstrates greater robustness to pose variation, partial occlusion, and deformation. Empirical comparisons in our experiments confirmed that the mask-based matching consistently achieved higher key-point localization accuracy and stability than direct key-point detection in complex construction site environments.

3.2. STL of Safety Helmet as Exemplar Scale Carrier

The STL for safety helmets integrates multi-angle mask images with associated prior dimensional data. This entails mask refinement and prior knowledge extraction from standardized templates. Utilizing the trained YOLOv8 model, 97 lab-dataset images underwent helmet detection and segmentation at over 95% confidence, yielding optimal masks for subsequent processing. To address reflection artifacts and segmentation imperfections (e.g., incomplete helmet boundaries), a dedicated image processing pipeline was implemented (Figure 6). To simulate realistic viewing variations in fixed-surveillance scenarios, each base mask was rotated from 0° to 45° in 3° increments and horizontally mirrored. Additionally, three other helmet models with distinct shapes were included to enhance template diversity. As a result, a total of 12,416 (i.e., $97 \times 16 \times 2 \times 4$) standardized mask templates were generated.

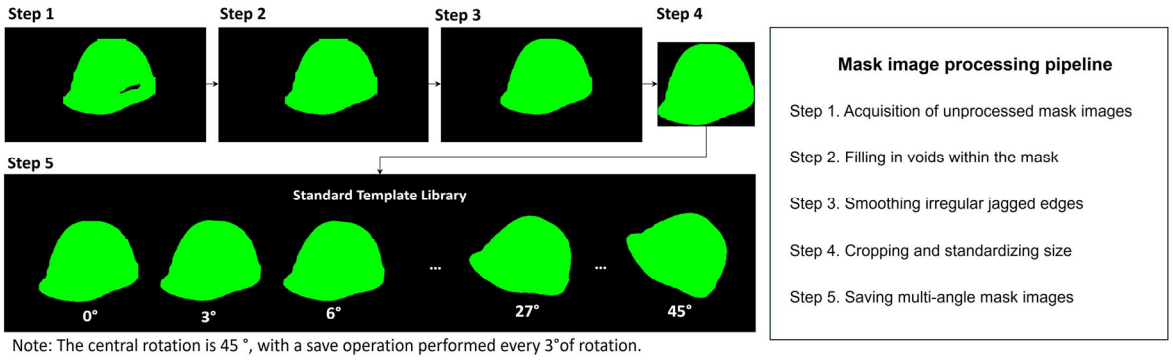


Figure 6. Image processing pipeline for generating refined safety helmet masks.

To ensure labeling accuracy despite the smooth surface of safety helmets, a dual-stage annotation process was adopted. First, key-points were manually labeled on high-resolution lab images using geometrically distinctive features (e.g., brim edge, apex, back

rim). Moreover, a 3D model of the helmet was used to validate these key-points, providing corrective ratios and ensuring label consistency.

Prior scale extraction commenced with defining nine key-points ("K1"–"K9", Figure 7a) on helmet surfaces. Key-point pairs (e.g., "KP69" for "K6"–"K9" distance) were formulated by connecting geometrically significant points. Real-world distances between key-point pairs were measured, with inaccessible dimensions (e.g., "KP78", "KP49") derived via geometric correction using directly measurable pairs ("KP34", "KP45"). Validation employed a safety helmet 3D model (Figure 7b) to compute correction ratios (Tables 1 and 2). Comparative analysis of three key-point pairs confirmed sub-1% cumulative error against ground truth measurements.

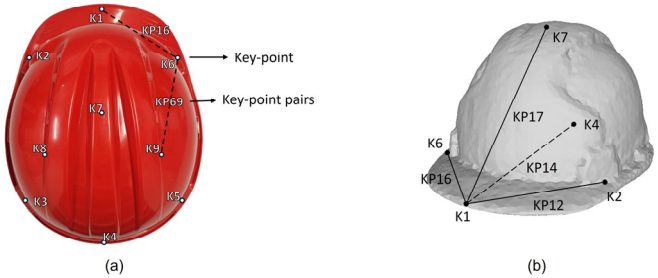


Figure 7. (**a**) Selection and annotation of key-points; (**b**) 3D model of the safety helmet.

Table 1. The modified ratio of the helmet with 3D model-aided measurement.

Attributes	True Value (cm)	Simulated Value (cm)	Modified Ratio	Mean Modified Ratio
Length	27.0	0.98	27.49	
Width	22.0	0.81	27.03	26.72
Height	14.5	0.57	25.64	

Table 2. The errors of key-point pairs with 3D model-aided measurement.

KP	True Value (cm)	Simulated Value (cm)	Calculated Value (cm)	Error (%)
KP12	11.5	0.43	11.54	0.36
KP16	11.5	0.43	11.57	0.61
KP17	21.6	0.81	21.72	0.57

Note: Calculated value = Simulated value × Mean modified ratio. The simulated value is sourced from 3D model-aided measurement, and the mean modified ratio is sourced from Table 1.

3.3. Illustrative Example

A 60 s construction site surveillance video was analyzed to demonstrate the refined reconstruction pipeline, with the 6 s frame exemplifying scale ambiguity resolution. Template matching identified optimal correspondences between detected safety helmets and standardized templates (Figure 8). Worker A and B achieved comparable accuracy (Pose 1: IoU_{tm} = 0.9155, d_{norm} = 0.9534, SI = 0.9269; Pose 2: IoU_{tm} = 0.9247; d_{norm} = 0.9244; SI = 0.9246), confirming high-fidelity correspondence as defined in Equation (4). The mean SI was 0.9267 ± 0.019 (95% CI: [0.9225, 0.9308]), indicating minimal inter-frame variation and stable template matching performance throughout the sequence.

Scale recovery modeling leveraged matched key-points (Table 3) and monocular depth estimation Z_{pseudo}. With optical center at (632, 472), residual minimization (Equations (7)–(14)) yielded dimension-specific scale factors: f_x = 1623.50, f_y = 1521.15, and s_z = 3.50.

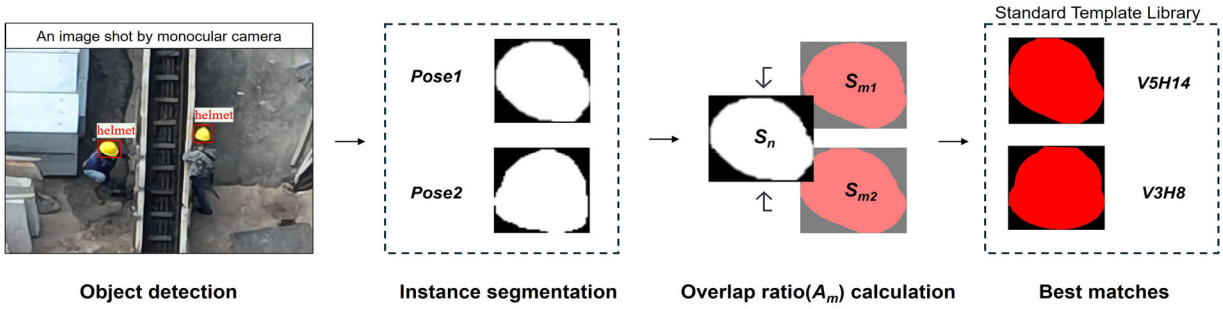


Figure 8. Visualization of the template matching process for scale carriers.

Table 3. The pixel coordinates of key-points and pseudo-depth values.

Pose	Key-Point	U (Pixel)	V (Pixel)
1	K1	371	468
	K4	282	414
2	K3	845	355
	K5	910	371
	K7	887	309
	K4	881	358

Based on the Brown-Conrady distortion model, synthetic radial distortion perturbations corresponding to typical CCTV lenses were applied to evaluate the effect on scale recovery. The re-estimated scale factors are summarized in Table 4, showing deviations within ±1.2%, which demonstrates the robustness of the proposed method to lens distortion.

Table 4. Sensitivity of scale recovery to radial distortion.

k_1	ΔS_x (%)	ΔS_y (%)	ΔS_z (%)
0.0000	0	0	0
−0.0005	+0.3	+0.4	+0.4
−0.001	+0.8	+1.1	+0.9
+0.0005	−0.4	−0.5	−0.6
+0.001	−1.0	−1.2	−1.1

Multi-frame optimization applied the fusion mechanism to consecutive frames. Parameter variations decayed exponentially (Figure 9), converging to stable values: f_x = 1530.71, f_y = 1470.85, s_z = 3.01.

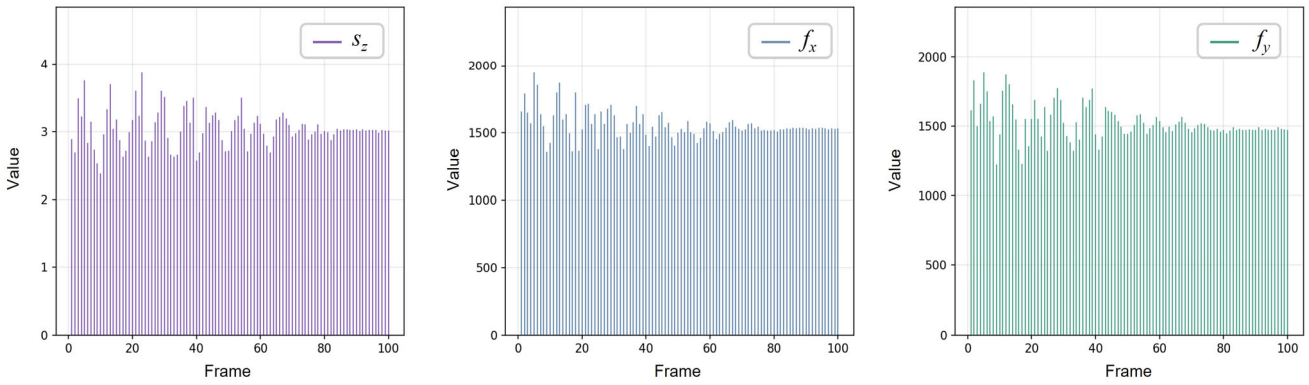


Figure 9. Optimization of scale factors via multi-frame fusion.

3.4. Results Analysis

3.4.1. Uncertainty and Optimization Analysis

Frame-wise scale factors (S_x, S_y, S_z) are estimated independently for each valid video frame. Due to initialization noise and scene variation, early frames can exhibit transient fluctuations before converging to a stable regime. To obtain statistically reliable uncertainty measures, we identify and exclude the initial burn-in period and compute statistics over the subsequent stable frames.

Let $\{S_x^t, S_y^t, S_z^t\}_{t=t_0}^N$ denote the scale estimates from the stable segment starting at frame t_0. For each axis, the mean and standard deviation are reported as:

$$\overline{S} = \frac{1}{N - t_0 + 1} \sum_{t=t_0}^{N} S^t, \sigma = \sqrt{\frac{1}{N - t_0} \sum_{t=t_0}^{N} (S^t - \overline{S})^2} \tag{23}$$

We therefore report mean and standard deviation across frames, together with the 95% confidence interval:

$$CI_{95\%} = \overline{S} \pm 1.96 \frac{\sigma}{\sqrt{N - t_0 + 1}} \tag{24}$$

Table 5 illustrates a representative per-frame scale trajectory. Early frames exhibit transient variation and converge after approximately $t_0 = 70$ frames. Therefore, uncertainty is computed from frames t_0 to N (stable segment).

Table 5. Frame-wise uncertainty of scale recovery (stable frames only).

Scale Factors	Mean	SD	95%CI
f_x	1527.66	12.88	[1523.05,1532.27]
f_y	1475.70	15.37	[1470.20,1481.20]
s_z	3.00	0.05	[2.98,3.02]

After convergence, the recovered scale factors exhibit low variance, with standard deviations (SD) below 1.1% for both focal lengths and 1.7% for depth scaling. The 95% confidence intervals are narrow (e.g., s_z = 3.00 $\pm$ 0.02), demonstrating high temporal stability and reliable scale estimation.

Figure 10 illustrates measurements and relative errors during the scale recovery and optimization process for three distinct areas of varying sizes within the same construction site. The results demonstrate that the model achieves a minimum accuracy of 90.42%. The scale recovery and optimization procedure is divided into five distinct phases: Status 1 and Status 2 represent the initiation and completion of scale recovery, respectively; Status 3 and Status 4 present intermediate data during scale optimization; Status 5 indicates the conclusion of scale optimization.

3.4.2. Multi-Object Error Evaluation

As shown in Figure 11, the measurement results are normalized, with the ground truth for each metric set to 1.0. Based on the overall statistical analysis, the outlier rate of the measurement results is 2.89%, and the total error relative to the true values is less than 10%.

To further examine the influence of object scale on reconstruction accuracy, the relationship between normalized pixel ratio and relative reconstruction error was analyzed. Each data point in the analysis corresponds to one reconstruction result. The normalized pixel ratio was defined as the proportion of an object's width in pixels relative to the total image width, thereby eliminating the influence of image resolution and enabling consistent cross-scene comparison. As shown in Figure 12, a slight decreasing trend is observed, indi-

cating that smaller normalized pixel ratios correspond to marginally higher reconstruction errors. The fitted regression line $y = -0.217x + 0.094$ exhibits a coefficient of determination $R^2 = 0.08$, suggesting that object size has a weak correlation with reconstruction accuracy. This finding demonstrates that the proposed framework maintains stable performance across a wide range of object scales when resolution differences are properly normalized.

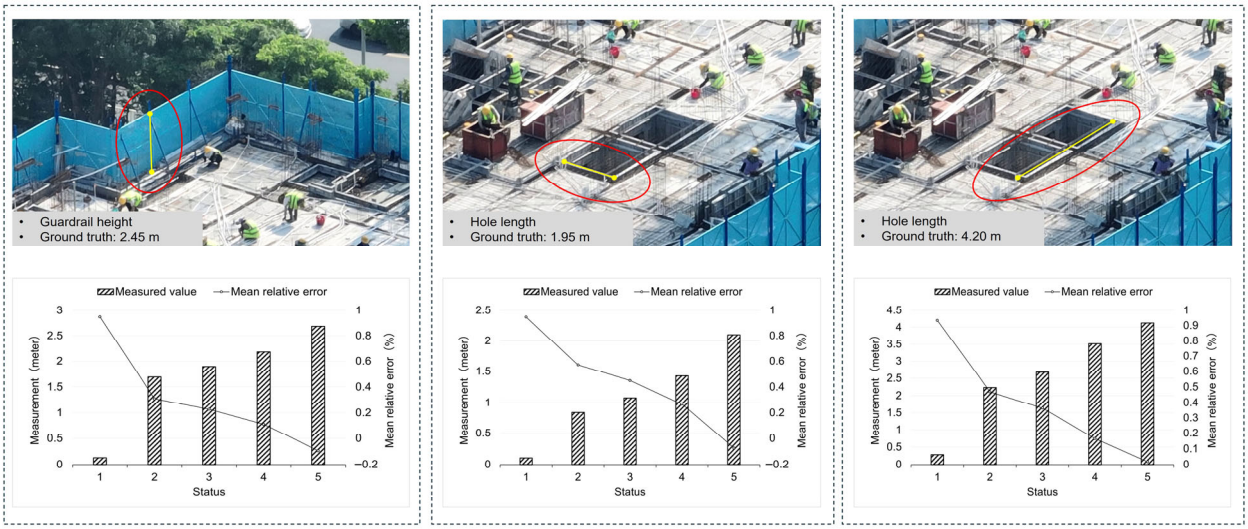


Figure 10. Five status stages of scale recovery and optimization across multiple scenarios.

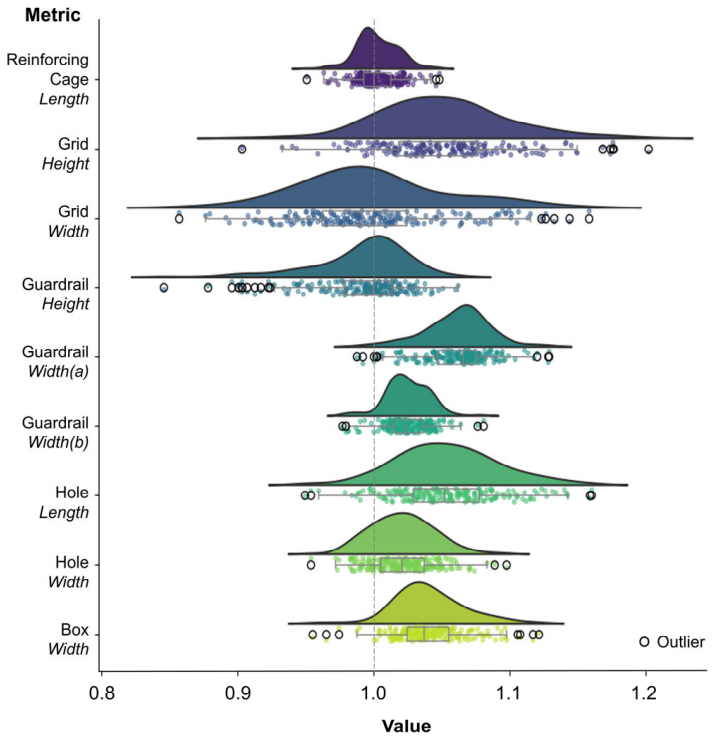


Figure 11. Normalized measurements and error distribution of multi-object dimensions.

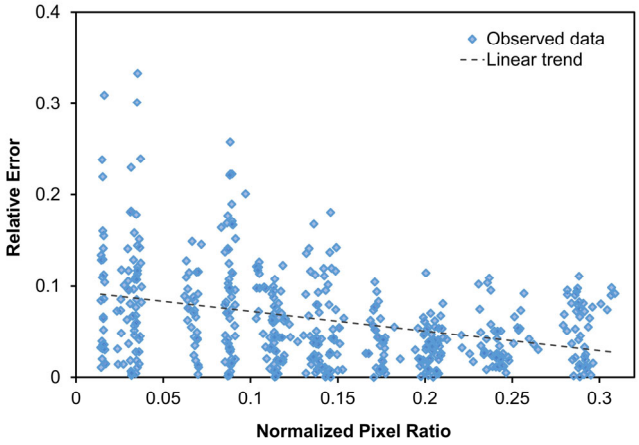


Figure 12. Statistical correlation between normalized object scale and reconstruction accuracy.

In addition, the resolution of input images has a noticeable impact on the performance of small object detection and scale recovery. Specifically, for safety helmets, which serve as the primary scale carriers in this study, a minimum pixel area is required to ensure reliable detection, segmentation, and key-point extraction. Empirical observations suggest that helmets should appear at approximately 120 × 120 pixels or larger in 1280 × 720 resolution images to maintain stable template matching performance. Very low-resolution frames may lead to decreased detection confidence and matching quality, introducing instability in scale estimation. Therefore, resolution requirements should be considered when applying this method in practical settings.

4. Discussion

4.1. Performance Evaluation

To evaluate the reconstruction accuracy of the proposed scale recovery and optimization framework, comparative experiments were conducted against two representative monocular depth estimation models, namely Metric3D and ZoeDepth.

Ground truth dimensions were obtained using manual laser rangefinder measurements (DL331040B, Deli Group Co., Ltd., Ningbo, China) for selected reference objects in each test scene. Multiple measurements were taken per target to reduce random errors, and the mean value was adopted as the ground-truth reference. The measurement uncertainty of the laser rangefinder ensures adequate accuracy for evaluating scale recovery at the tested construction site scales. The mean relative error (MRE), standard deviation (SD), and 95% confidence intervals (CI) for each method are summarized in Table 6. Moreover, all tests returned $p < 0.01$ and Cohen's d values well above 1.4, confirming statistically significant and practical improvements for the proposed method.

Table 6. MRE, SD, and 95%CI statistics of each model across different scenarios.

Scenario	Proposed			Metric3D			ZoeDepth		
	MRE (%)	SD (%)	CI (%)	MRE (%)	SD (%)	CI (%)	MRE (%)	SD (%)	CI (%)
A	8.98	4.21	[6.37, 11.59]	26.76	5.48	[23.37, 30.16]	−54.25	1.98	[−55.48, −53.02]
B	6.33	6.26	[2.45, 10.22]	−25.35	3.31	[−27.40, −23.30]	−51.46	2.15	[−52.79, −50.13]
C	−3.05	2.22	[−4.43, −1.67]	13.65	1.72	[12.58, 14.71]	−55.72	2.58	[−57.32, −54.12]
D	5.32	4.66	[2.43, 8.22]	−14.02	4.30	[−16.68, −11.35]	−55.63	4.18	[−58.22, −53.04]
E	−4.14	7.42	[−8.74, 0.46]	−22.90	11.41	[−29.98, −15.83]	−45.16	3.02	[−47.03, −43.28]
F	−7.27	3.70	[−9.56, −4.98]	−28.83	4.96	[−31.90, −25.76]	−53.12	0.76	[−53.59, −52.64]

In addition, the computational efficiency of the proposed framework was evaluated. All experiments were conducted using PyTorch 2.3.1 with an NVIDIA RTX 3090 GPU. The average runtime from detection and segmentation to the completion of scale recovery

was approximately 1.46 s per frame. Considering that the duration of the video stream directly affects the time required for scale recovery, this stage was excluded from the frame-wise speed calculation. The current implementation therefore operates in an offline post-processing mode, designed for methodological validation and accuracy assessment.

4.2. Cost-Effectiveness Analysis

The framework attains substantial cost savings by utilizing only a monocular camera, a laptop, and regulatory-mandated safety helmets. Critically, monocular surveillance systems dominate existing site infrastructure (deployed in over 90% of Chinese construction sites according to the Research Report on the Digital Development of China's Construction Industry [37]), enabling direct implementation without hardware investment. As quantified in Table 7, monocular solutions reduce sensor costs by over 50% compared to depth-specialized alternatives while maintaining reconstruction accuracy. This hardware minimalism positions our method as an economically viable solution for widespread adoption.

Table 7. Comprehensive analysis of the vision device (2025).

Device and Specifications	Cost (¥)	Complexity of Data Processing	Site Adaptability
Monocular camera (Hikvision Dome, 5MP) *	400	Medium	High
Monocular camera (Hikvision Fixed, 3MP)	270	Medium	High
Stereo Camera (Gemini Pro, 1080p)	1190	Medium	Medium
Depth Camera (Astra Pro Plus)	960	Low	Medium
Depth Camera (Intel D435i, 720p)	2450	Medium	Medium
LiDAR (Intel L515)	5000	High	Low

Note: * Primary experimental device.

4.3. Applicability Analysis

To further evaluate the robustness and field applicability of the proposed scale recovery and optimization framework beyond the controlled experimental scenarios presented in previous sections, video data captured under varying construction site conditions (Figure 13) were analyzed. The applicability assessment was based on the reconstruction accuracy achieved after scale recovery and optimization, as summarized in Table 8.

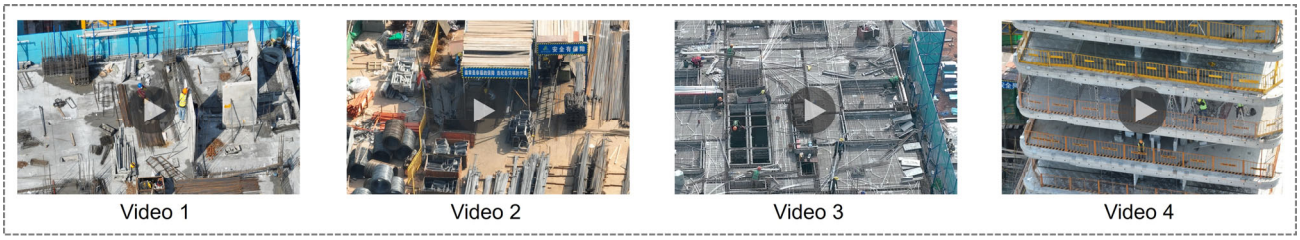


Figure 13. Sample frames from the test videos used in the applicability analysis.

Table 8. Parameters and reconstruction accuracy of different video data.

Video	Camera Setup	Available Scale Carrier	Spatial Openness	Lighting Conditions	Accuracy (%)
1	Fixed, 1080p	Safety helmets, guardrails, rebars	High	Bright	92.08
2	Fixed, 1080p	Safety helmets, rebars	Medium	Bright	86.26
3	Fixed, 720p	Safety helmets, guardrails	High	Dim	67.39
4	Fixed, 1080p	Safety helmets, guardrails	Low	Normal	78.75

Overall, the proposed method delivered optimal performance when applied to high-resolution, stable video data recorded under bright lighting conditions with an adequate number of appropriately sized and distinctive scale carriers. Under such circumstances, the reconstruction accuracy exceeded 90%. Conversely, performance deteriorated significantly in scenarios characterized by dim lighting, insufficient or poorly sized scale carriers, or low camera resolution. For example, in Video 3 (720p resolution and dim lighting), the reconstruction accuracy dropped to 67.39%, highlighting the sensitivity of the method to these factors.

Based on these findings, several conditions were identified as critical for achieving reliable reconstruction results:

1. Presence of standardized, geometrically distinctive scale carriers in the scene (e.g., safety helmets, prefabricated components, scaffold joints) in sufficient quantity and distribution.

2. Video resolution of at least 1280 × 720 pixels, ensuring that scale carriers occupy enough pixels for stable detection, segmentation, and key-point matching.

3. Stable camera installation with minimal vibration and adequate coverage of key work zones.

4. Lighting conditions that avoid prolonged underexposure, overexposure, or frequent extreme fluctuations that could affect detection stability.

To support deployment decision-making, a pre-assessment procedure was developed (Figure 14). This procedure enables site managers to evaluate the feasibility of applying the proposed method before field implementation. The decision flow considers three primary factors: image resolution, the number of effective scale carriers, and site lighting conditions. By filtering unsuitable scenarios in advance, this process helps avoid ineffective deployments and ensures that the method is applied under conditions conducive to achieving high reconstruction accuracy.

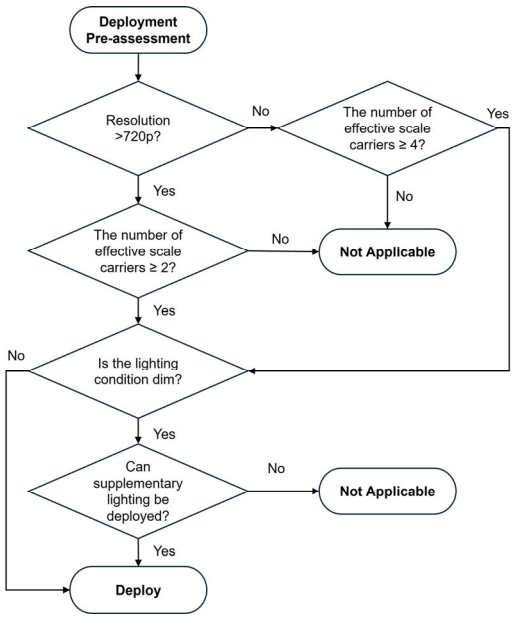


Figure 14. Pre-deployment decision flowchart for monocular 3D reconstruction.

4.4. Engineering Applications

In construction site safety management, scene 3D reconstruction with integrated scale recovery capabilities offers numerous practical applications. By utilizing the 3D coordinates of construction workers and hazardous elements (such as materials, machinery, or specific areas) on the construction site, the shortest true distance between them can be measured.

The "Regulations on the Safety Management of High-Risk Sub-Projects" [38] outlines numerous hazardous sub-projects, including soil excavation, support, and dewatering in foundation pits exceeding 3 m in depth. Therefore, by measuring the shortest distance between the fence and the foundation pit, the pit depth can be determined, allowing an assessment of the project's risk level. The process for measuring safety hazards and evaluating the risk level of sub-projects is shown in Figure 15. The shortest distance between two objects in the image can be calculated based on the Ramer-Douglas-Peucker algorithm [39], as demonstrated in Algorithm 1.

Algorithm 1. Compute Minimum Distance from Point to Polygon

Input: (1) Point Q (x, y).
(2) A polygon defined by $\{P_1 (x_1, y_1), P_2 (x_2, y_2), \ldots, P_n (x_n, y_n)\}$.
Output: Minimum distance d_{min} between Q and the polygon.
1. Initialize: $d_{min} \leftarrow +\infty$
2. For each edge (P_i, P_{i+1}) of the polygon:
2.1 Let P_i (x_1, y_1) and P_{i+1} (x_2, y_2)
2.2 Compute the vector $\boldsymbol{v} \leftarrow (x_2 - x_1, y_2 - y_1)$
2.3 Compute the vector $\boldsymbol{w} \leftarrow (x - x_1, y - y_1)$
2.4 Compute the projection scalar: $t \leftarrow (\boldsymbol{w} \times \boldsymbol{v}) / |\boldsymbol{v}|^2$
2.5 If t < 0: The nearest point is P_i. Compute distance $d \leftarrow ||w||$
2.6 Else if t > 1: The nearest point is P_{i+1}. Compute distance $d \leftarrow ||(x - x_2, y - y_2)||$
2.7 Else: The projection point lies on the edge. Compute the projection point: $R(x_r, y_r) \leftarrow [x_1 + t \times (x_2 - x_1), y_1 + t \times (y_2 - y_1)]$, Compute distance $d \leftarrow ||(x - x_r, y - y_r)||$
2.8 Update minimum distance: $d_{min} \leftarrow min\ (d_{min}, d)$
3. End For
4. Return d_{min}

4.5. Limitations and Future Research

Despite the demonstrated effectiveness of the proposed method, some practical limitations persist. The performance relies on accurate detection and segmentation, sufficient video duration for stable optimization, and adequate image resolution, where helmets typically require approximately 120 × 120 pixels in 1280 × 720 images to ensure robust template matching. In addition, the current framework is best suited for non-symmetrical, dimension-standardized objects with distinct geometric features; geometrically similar or smooth-surfaced objects may lead to pose ambiguity or unstable feature correspondence, thereby reducing scale estimation robustness.

Looking forward, there is currently no universal scale recovery approach applicable to all construction site scenarios. With continued technological development, future research may incorporate advanced AI and large-scale datasets for absolute scale calibration, refine monocular reconstruction networks, and enhance template design to better support symmetrical objects through auxiliary cues such as texture, color patterns, fiducial markers, or contextual scene priors. In addition, future work will systematically examine the influence of color and contrast variations on segmentation robustness to ensure consistent performance across diverse site conditions. Furthermore, extending the framework to nighttime and infrared imaging will be pursued to enable reliable 24 h monitoring under diverse illumination conditions.

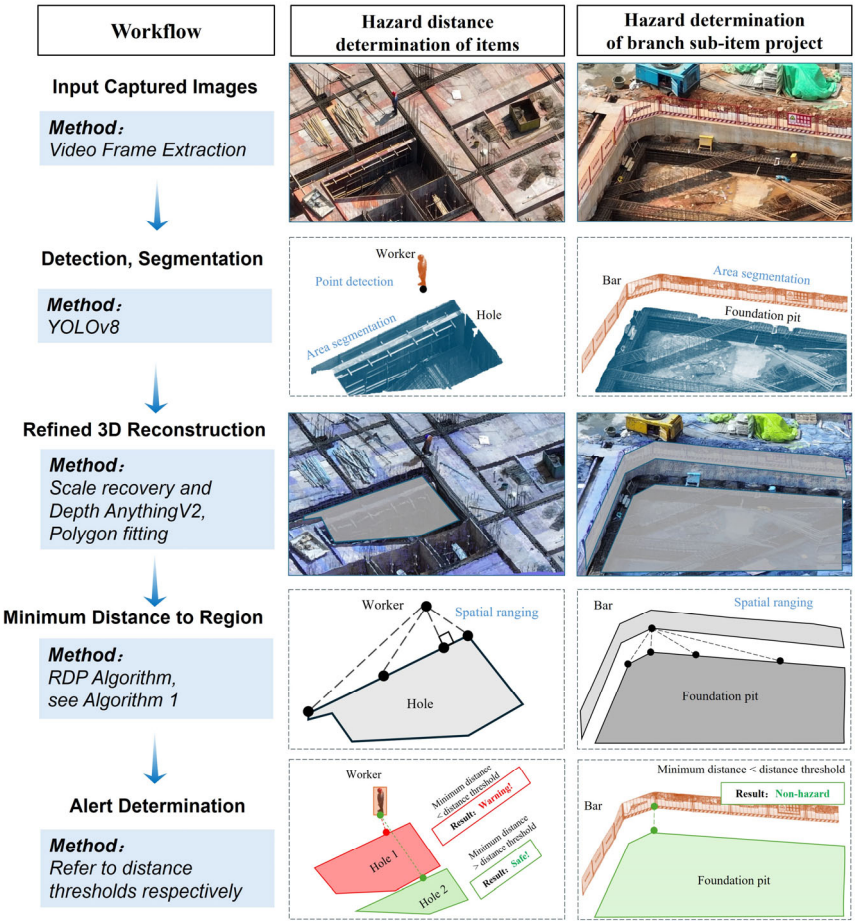


Figure 15. Application of ranging-based safety management in construction environments.

5. Conclusions

The integration of visual perception into construction site management is pivotal for advancing refined management practices. Leveraging visual perception as a tool, combined with scale recovery and optimization techniques, a refined 3D reconstruction of the construction site has been successfully achieved, offering valuable technical support for effective site management.

This research has been designed to accommodate the prevailing conditions of the majority of construction sites, eschewing reliance on high-cost equipment such as LiDAR or depth cameras. Instead, the focus has been directed towards monocular depth estimation-based 3D reconstruction, presenting a cost-effective solution for the attainment of refined management. Innovatively, a multi-dimensional scale recovery model derives scale factors from three spatial axes, while the STL, constructed from standardized safety helmets commonly used on construction sites, serves as a universal metric reference without the need for site-specific calibration. Furthermore, multi-frame data fusion from video streams is utilized to process the scale factor. Stability and accuracy checks are incorporated to continuously optimize the scale factor, further enhancing the overall performance of the reconstructed 3D scene. Experimental results show that the proposed model achieves an accuracy of over 90%, with good stability and outlier values not exceeding 5%.

In conclusion, this research on monocular 3D reconstruction-scale recovery provides a viable solution for refined management on construction site monitoring, capitalizing on the advantage of low cost. This approach is not only of theoretical significance but also highly practical, enabling more refined management of personnel safety, material storage, progress control, and other key aspects of construction site operations, thus improving

management efficiency. It is also anticipated that future research will address the current limitations of the model, thereby providing more comprehensive and precise solutions for the refined management of construction sites.

Author Contributions: Conceptualization, J.R. and L.Z.; methodology, J.R., L.W., X.L. and L.Z.; software, J.R., L.W., X.L. and L.Z.; validation, J.R., L.W., X.L. and L.Z.; resources, L.Z.; data curation, J.R., L.W., X.L. and L.Z.; writing—original draft preparation, J.R.; writing—review and editing, J.R. and L.Z.; visualization, J.R. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by (1) the National Key Research and Development Program of China (Grant No. 2023YFC3806800), (2) the National Natural Science Foundation of China (Grant No. 72201095), and (3) the National Natural Science Foundation of Hunan Province (Grant No. 2023JJ40189).

Data Availability Statement: The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest: Author Lingling Wang and Xuanxuan Liu were employed by the company State Grid Hunan Electric Power Company Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Shi, W.; Alawieh, M.B.; Li, X.; Yu, H. Algorithm and Hardware Implementation for Visual Perception System in Autonomous Vehicle: A Survey. *Integration* **2017**, *59*, 148–156. [CrossRef]
2. Yang, J.; Wang, C.; Jiang, B.; Song, H.; Meng, Q. Visual Perception Enabled Industry Intelligence: State of the Art, Challenges and Prospects. *IEEE Trans. Ind. Inform.* **2021**, *17*, 2204–2219. [CrossRef]
3. Li, J.; Liu, Y.; Wang, L.; Sun, Y. A Vision-Based End Pose Estimation Method for Excavator Manipulator. *Multimed. Tools Appl.* **2024**, *83*, 68723–68741. [CrossRef]
4. Luo, X.; Li, H.; Yang, X.; Yu, Y.; Cao, D. Capturing and Understanding Workers' Activities in Far-Field Surveillance Videos with Deep Action Recognition and Bayesian Nonparametric Learning. *Comput.-Aided Civ. Infrastruct. Eng.* **2019**, *34*, 333–351. [CrossRef]
5. Fang, Q.; Li, H.; Luo, X.; Ding, L.; Luo, H.; Li, C. Computer Vision Aided Inspection on Falling Prevention Measures for Steeplejacks in an Aerial Environment. *Autom. Constr.* **2018**, *93*, 148–164. [CrossRef]
6. Fang, W.; Love, P.E.D.; Ding, L.; Xu, S.; Kong, T.; Li, H. Computer Vision and Deep Learning to Manage Safety in Construction: Matching Images of Unsafe Behavior and Semantic Rules. *IEEE Trans. Eng. Manag.* **2023**, *70*, 4120–4132. [CrossRef]
7. Yilmaz, D.I.; Artan, D. An Occupational Safety Risk Management System for Coastal Construction Projects. *IEEE Trans. Eng. Manag.* **2024**, *71*, 14387–14400. [CrossRef]
8. Kim, M.-K.; Cheng, J.C.P.; Sohn, H.; Chang, C.-C. A Framework for Dimensional and Surface Quality Assessment of Precast Concrete Elements Using BIM and 3D Laser Scanning. *Autom. Constr.* **2015**, *49*, 225–238. [CrossRef]
9. Deng, H.; Hong, H.; Luo, D.; Deng, Y.; Su, C. Automatic Indoor Construction Process Monitoring for Tiles Based on BIM and Computer Vision. *J. Constr. Eng. Manag.* **2020**, *146*, 04019095. [CrossRef]
10. Fang, W.; Ding, L.; Love, P.E.D.; Luo, H.; Li, H.; Peña-Mora, F.; Zhong, B.; Zhou, C. Computer Vision Applications in Construction Safety Assurance. *Autom. Constr.* **2020**, *110*, 103013. [CrossRef]
11. Jiang, Z.; Messner, J.I. Computer Vision Application in Construction and Asset Management Phases: A Literature Review. *J. Inf. Technol. Constr.* **2023**, *28*, 176–199. [CrossRef]
12. Lo, Y.; Zhang, C.; Ye, Z.; Cui, C. Monitoring Road Base Course Construction Progress by Photogrammetry-Based 3D Reconstruction. *Int. J. Constr. Manag.* **2023**, *23*, 2087–2101. [CrossRef]
13. Sung, C.; Kim, P.Y. 3D Terrain Reconstruction of Construction Sites Using a Stereo Camera. *Autom. Constr.* **2016**, *64*, 65–77. [CrossRef]
14. Zhu, Z.; Park, M.-W.; Koch, C.; Soltani, M.; Hammad, A.; Davari, K. Predicting Movements of Onsite Workers and Mobile Equipment for Enhancing Construction Site Safety. *Autom. Constr.* **2016**, *68*, 95–101. [CrossRef]

15. Scharstein, D.; Szeliski, R. A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms. *Int. J. Comput. Vis.* **2002**, *47*, 7–42. [CrossRef]
16. Liu, F.; Shen, C.; Lin, G.; Reid, I. Learning Depth from Single Monocular Images Using Deep Convolutional Neural Fields. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 2024–2039. [CrossRef]
17. Wöhler, C.; d'Angelo, P.; Krüger, L.; Kuhl, A.; Groß, H.-M. Monocular 3D Scene Reconstruction at Absolute Scale. *ISPRS J. Photogramm. Remote Sens.* **2009**, *64*, 529–540. [CrossRef]
18. Saxena, A.; Chung, S.H.; Ng, A.Y. 3-D Depth Reconstruction from a Single Still Image. *Int. J. Comput. Vis.* **2008**, *76*, 53–69. [CrossRef]
19. Wang, Z.; Shen, M.; Chen, Q. Eliminating Scale Ambiguity of Unsupervised Monocular Visual Odometry. *Neural Process. Lett.* **2023**, *55*, 9743–9764. [CrossRef]
20. Forsyth, D.A.; Ponce, J. *Computer Vision: A Modern Approach*; Prentice Hall Professional Technical Reference: Upper Saddle River, NJ, USA, 2002; ISBN 978-0-13-085198-7.
21. Zhang, R.; Tsai, P.-S.; Cryer, J.E.; Shah, M. Shape-from-Shading: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **1999**, *21*, 690–706. [CrossRef]
22. Maki, A.; Watanabe, M.; Wiles, C. Geotensity: Combining Motion and Lighting for 3D Surface Reconstruction. *Int. J. Comput. Vis.* **2002**, *48*, 75–90. [CrossRef]
23. Kasetkasem, T.; Varshney, P.K. An Image Change Detection Algorithm Based on Markov Random Field Models. *IEEE Trans. Geosci. Remote Sens.* **2002**, *40*, 1815–1823. [CrossRef]
24. Saxena, A.; Chung, S.; Ng, A. Learning Depth from Single Monocular Images. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 2005; Volume 18.
25. Eigen, D.; Fergus, R. Predicting Depth, Surface Normals and Semantic Labels With a Common Multi-Scale Convolutional Architecture. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 2650–2658.
26. Godard, C.; Mac Aodha, O.; Brostow, G.J. Unsupervised Monocular Depth Estimation With Left-Right Consistency. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 270–279.
27. Yin, X.; Wang, X.; Du, X.; Chen, Q. Scale Recovery for Monocular Visual Odometry Using Depth Estimated With Deep Convolutional Neural Fields. In Proceedings of the 2017 IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 5871–5879.
28. Wagstaff, B.; Kelly, J. Self-Supervised Scale Recovery for Monocular Depth and Egomotion Estimation. In Proceedings of the 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Prague, Czech Republic, 27 September–1 October 2021; pp. 2620–2627.
29. Shen, J.; Yan, W.; Qin, S.; Zheng, X. A Self-Supervised Monocular Depth Estimation Model with Scale Recovery and Transfer Learning for Construction Scene Analysis. *Comput.-Aided Civ. Infrastruct. Eng.* **2023**, *38*, 1142–1161. [CrossRef]
30. Balasubramanian, S.; Shukla, V.; Islam, N.; Manghat, S. Construction Industry 4.0 and Sustainability: An Enabling Framework. *IEEE Trans. Eng. Manag.* **2024**, *71*, 1–19. [CrossRef]
31. Golparvar-Fard, M.; Balali, V.; de la Garza, J.M. Segmentation and Recognition of Highway Assets Using Image-Based 3D Point Clouds and Semantic Texton Forests. *J. Comput. Civ. Eng.* **2015**, *29*, 04014023. [CrossRef]
32. Li, X.; Hu, Y.; Jie, Y.; Zhao, C.; Zhang, Z. Dual-Frequency Lidar for Compressed Sensing 3D Imaging Based on All-Phase Fast Fourier Transform. *J. Opt. Photonics Res.* **2024**, *1*, 74–81. [CrossRef]
33. Hu, K.; Chen, Z.; Kang, H.; Tang, Y. 3D Vision Technologies for a Self-Developed Structural External Crack Damage Recognition Robot. *Autom. Constr.* **2024**, *159*, 105262. [CrossRef]
34. Xu, B.; Liu, C. A 3D Reconstruction Method for Buildings Based on Monocular Vision. *Comput.-Aided Civ. Infrastruct. Eng.* **2022**, *37*, 354–369. [CrossRef]
35. Ma, S.; Fang, Q.; Zhou, H.; Yin, Y.; Ye, F. Fast 3D Site Reconstruction Using Multichannel Dynamic and Static Object Separation. *Autom. Constr.* **2024**, *168*, 105807. [CrossRef]
36. Yang, L.; Kang, B.; Huang, Z.; Zhao, Z.; Xu, X.; Feng, J.; Zhao, H. Depth Anything V2. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 21875–21911.
37. Information Center of the Ministry of Housing and Urban-Rural Development of China. *Research Report on the Digital Development of China's Construction Industry*; Information Center of the Ministry of Housing and Urban-Rural Development of China: Beijing, China, 2022.

38. Ministry of Housing and Urban-Rural Development of the People's Republic of China. *Regulations on the Safety Management of High-Risk Sub-Projects*; Ministry of Housing and Urban-Rural Development of the People's Republic of China: Beijing, China, 2018.
39. Radeta, M.; Rodrigues, C.; Silva, F. Lost in the Deep? Performance Evaluation of Dead Reckoning Techniques in Underwater Environments. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **2023**, *7*, 73. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.