

Article

Scaffolding Safety Assessment Framework Integrating Vision-Based Geometry Recognition and Structural Simulation

Hao Peng ¹, Lintao Zhang ^{2,*}, Jing Dong ¹, Yu Du ³ and Han Wu ⁴

¹ School of Architectural Engineering, Xinyang Vocational and Technical College, No. 48, 24th Street, Xinyang 464000, China; penghao@xyvtc.edu.cn (H.P.); dongjing@xyvtc.edu.cn (J.D.)

² School of Horticulture, Xinyang Agriculture and Forestry University, No. 1 North Ring Road, Xinyang 464000, China

³ School of Civil Engineering and Architecture, Wuhan University of Technology, Wuhan 430070, China; whutdy2023@163.com

⁴ School of Infrastructure Engineering, Nanchang University, No. 999 Xuefu Avenue, Nanchang 330047, China; wu.han@whut.edu.cn

* Correspondence: zlt.179@163.com

Abstract

The assembly quality of scaffolding systems directly governs the safety of personnel on construction sites. According to construction safety statistics, scaffolding-related accidents account for approximately 30–40% of construction fatalities globally, with geometric assembly deviations being a contributing factor in over 60% of scaffold collapse incidents. Traditional scaffolding inspections rely heavily on manual measurements, which are inherently inefficient, hazardous, and difficult to scale comprehensively. This study presents an automated evaluation framework that integrates computer vision with structural mechanics simulations. First, an object detection model based on the SegFormer encoder architecture is developed to precisely identify scaffolding standards, ledgers, and couplers against complex site backgrounds. Its hierarchical Transformer encoder and global self-attention mechanism enable the model to capture long-range topological relationships, achieving a mean Average Precision (mAP@0.5) of 95.2% on a custom dataset with an inference speed of 45 FPS per 640×640 image patch. For complete high-resolution frame processing including tiling and geometric extraction, the end-to-end pipeline requires approximately 8–12 s per frame. Second, a simplified Hough transform with a restricted parameter domain is introduced. Integrated with a dual-track image processing workflow, this algorithm performs sub-pixel centerline fitting to automatically extract critical geometric parameters, including lift height and bay width, maintaining a relative measurement error within 3.5% compared to manual ground truth. Finally, a parameterized finite element model is established. An automated mapping middleware dynamically injects the extracted as-built parameters into the simulation environment. Comparative simulation analysis indicates that a 14.7% deviation in standard lift height, coupled with an initial tilt defect of 1/150, precipitates a 22.4% reduction in the predicted structural stability factor, illustrating the framework's capability for assessing relative capacity degradation between design intent and as-built conditions. This framework establishes a robust, closed-loop pipeline spanning visual perception and structural safety assessment, indicating potential for automated construction site safety management.



Academic Editors: Gongfa Chen, David H. Bassir and Shuai Teng

Received: 26 May 2026

Revised: 8 July 2026

Accepted: 10 July 2026

Published: 13 July 2026

Copyright: © 2026 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.

Keywords: construction automation; structural safety assessment; scaffolding; automated detection; SegFormer model; Hough transform

1. Introduction

In modern construction engineering, scaffolding serves as an essential temporary support system to ensure the safety of high-altitude operations [1]. As a space-lattice structure composed of slender members, scaffolding is highly sensitive to initial geometric defects. The ultimate bearing capacity of the structure depends not only on the sectional properties of the members but also significantly on geometric deviations generated during assembly, such as standard vertical misalignment and lift height non-uniformity [2]. Although rigorous mechanical calculations are performed during the design phase, significant differences often exist between the as-built structure and the design model due to operational errors or environmental factors on construction sites [3]. These hidden geometric deviations can trigger pronounced second-order effects, which may induce non-linear buckling of the support system prior to reaching the design load.

Currently, scaffolding safety monitoring relies primarily on manual field measurements. This approach is inefficient for high-rise or large-span complex structures and struggles to capture global geometric evolution. Existing research on scaffolding safety mostly focuses on pure mechanical simulations or point-based monitoring using sensors, such as strain gauges and tilt sensors. However, pure simulations cannot fully reflect the stochastic nature of as-built assembly quality, whereas point-based monitoring is limited by high costs and installation difficulties, making it unable to provide complete structural geometric information. This information gap between the design model and the as-built state often leads to overly optimistic safety assessment results, creating a regulatory blind spot on construction sites [4]. Accurately capturing the actual geometry of the scaffolding and converting it into perturbation variables within mechanical models is essential for realistic safety evaluations.

In recent years, computer vision technology has been widely applied in the construction field. However, identifying scaffolding as a specific target still faces significant challenges [5,6]. Scaffolding images exhibit high feature complexity; for instance, standards and ledgers have high slenderness ratios and overlap extensively within dense spatial grids. Early detection methods were mostly based on manually defined geometric operators, such as HOG features or Haar wavelets [7]. While these methods are effective for members against simple backgrounds, they are highly susceptible to interference from high-frequency textures on construction sites, such as safety nets and crane arms, leading to poor algorithmic robustness. Subsequently, deep Convolutional Neural Networks (CNNs), represented by Faster R-CNN, improved identification accuracy through region proposal networks (RPNs) [8,9]. Nevertheless, because CNNs utilize local sliding window convolutions, their receptive fields are inherently limited. They struggle to capture the topological continuity of long-range standards under local occlusions. This limitation often leads to “semantic fracture” in the detection of slender members [10,11].

To address the dense arrangement of scaffolding and significant scale variations (such as large frames coexisting with tiny couplers), researchers have introduced attention mechanisms. More recently, Vision Transformers (ViTs) and their derivative models, such as SegFormer and RT-DETR, have demonstrated significant advantages. These architectures bypass the limitations of local convolutions by utilizing global self-attention mechanisms to model long-range dependencies between components [12,13]. For scaffolding structures, this global perception capability enables the model to infer the overall topological path from unobstructed member segments, isolating target entities from complex backgrounds. Table 1 compares mainstream detection technologies in scaffolding monitoring scenarios.

Table 1. Comparison of mainstream vision detection technologies.

Algorithmic Paradigm	Representative Models	Feature Capture Mechanism	Application Analysis in Construction Sites
Traditional Machine Learning	Haar-Cascades [14]/HOG [15]	Gradient-based local feature operators	Difficult to distinguish members from safety nets; high false positive rate.
Two-stage Deep Learning	Faster R-CNN [16]/Cascade R-CNN [17]	Region proposal extraction + Classification and relocation	High accuracy, but real-time performance is difficult to meet supervision requirements.
One-stage Deep Learning	YOLOv8 [18]/v10 [19]/SSD [20]	End-to-end regression prediction	Fast detection, but prone to detection fractures in dense slender members.
Transformer-based	SegFormer [21]/RT-DETR [22]/Swin-DETR [23]	Global self-attention and multi-scale feature fusion	Strongest at capturing long-range topological correlations and robust against dense interference.

Once the preliminary semantic boundaries of components are obtained, converting these discrete pixel regions into high-precision geometric parameters is the critical step for mechanical evaluation. Standard visual detection models only output bounding boxes or masks; they cannot directly provide the analytical line equations required for Finite Element (FE) modeling [24]. Therefore, researchers have introduced line-feature extraction algorithms to quantify the geometric properties of members. The Line Segment Detector (LSD) can quickly extract line segments using gradient information. However, it often produces fractured segments under complex lighting, which degrades the topological integrity of long members [25,26]. The Random Sample Consensus (RANSAC) algorithm is robust against outliers but incurs prohibitive computational costs when applied to dense scaffolding arrays [27,28].

In contrast, the Hough transform (HT) maps image space to parameter space. It exhibits robust occlusion resistance and global integration characteristics, making it highly suitable for regular scaffolding configurations [29,30]. Song et al. [31] demonstrated that restricting the search domain of the parameter space improves fitting efficiency for rebar spacing measurement in reinforced concrete structures. However, their work focused on regular grid patterns in controlled indoor environments with minimal occlusion. The present study extends this approach to scaffolding structures, which present distinct challenges: (1) significantly more complex backgrounds with safety nets, construction equipment, and dynamic shadows; (2) much larger spatial extent requiring wide-angle capture with severe perspective distortion; (3) multilayer depth ambiguity where foreground and background standards may project onto similar image regions; and (4) integration requirements with parameterized FEM rather than standalone measurement reporting. Our contribution lies in adapting the parameter-restriction strategy to these specific challenges through dual-track image processing, ROI back-mapping to original resolution, and automated homography-based spatial calibration. Additionally, field images of scaffolding suffer from perspective distortion; thus, pixel displacements in a 2D image do not scale linearly to physical distances in 3D space. To resolve this mapping bias, homography must be introduced alongside known engineering dimensions for projection correction [32]. This sequential pipeline—proceeding from semantic perception to geometric fitting and then to spatial transformation—ensures that the extracted assembly parameters carry explicit physical meanings, providing precise inputs for structural safety assessments.

In summary, while existing visual perception and geometric extraction techniques can locate components and quantify their spacing, most studies are limited to acquiring discrete geometric parameters. They have not yet established a closed-loop pipeline from perceptual features to structural mechanics [33]. In structural safety evaluation, obtaining geometric data is merely the initial step; the main challenge lies in translating these as-built parameters into simulation models for mechanical computation. Scaffolding safety evaluation typically relies on the Finite Element Method (FEM) [34]. However, conventional FE analyses are usually based on idealized design dimensions, ignoring assembly deviations, missing members, and structural degradation during service [35]. Without real-time feedback from field data, these idealized models cannot reflect the actual mechanical behavior of the physical structure. Therefore, developing a method to automatically convert unstructured image features into FE variables is essential to eliminate information gaps between design intentions and as-built states.

To address these issues, this paper proposes an automated scaffolding safety assessment framework integrating Transformer-based vision perception with parameterized mechanical simulation. First, the global feature capture capability of a Transformer model is leveraged to precisely locate standards, ledgers, and couplers under complex backgrounds. Second, a geometric extraction model integrating a simplified HT and homography correction is developed to automatically measure key parameters, such as lift height and bay width. Finally, an automated mapping middleware dynamically injects as-built parameters into the FE analysis platform. This enables a quantitative evaluation of the bearing capacity and stability of the scaffolding in its actual as-built state. This research establishes a direct mapping from visual images to mechanical performance, supporting the digital safety supervision of construction sites.

The primary contributions of this study are: (1) adaptation of SegFormer encoder architecture for object detection with a specialized multi-task loss function tailored for scaffolding's high aspect-ratio members and severe class imbalance; (2) development of a parameter-domain-restricted Hough transform that reduces computational overhead by 60% while maintaining sub-pixel accuracy; (3) establishment of an automated middleware that directly converts unstructured visual features into parameterized FE variables without manual intervention; and (4) quantitative demonstration that centimeter-level assembly deviations can precipitate 20%+ reductions in ultimate bearing capacity through an integrated, simulation-based perception-to-mechanics pipeline.

The remainder of this paper is organized as follows: Section 2 details the component detection algorithm and the geometric measurement method. Section 3 describes the parameterized modeling logic and the automated mapping mechanism. Section 4 validates the framework using custom datasets and a representative engineering case study. Section 5 concludes the paper.

2. Vision-Based Scaffolding Geometry Identification

2.1. Image Data Acquisition and Preprocessing

The data-driven framework of this study integrates engineering design drawings and field imagery from the construction site. Design drawings, including CAD files and assembly technical specifications, provide initial prior data. These drawings define the idealized geometric topology, member specifications, and theoretical spacing of the scaffolding to establish a parameterized baseline FEM. Field image data are collected through the collaborative application of Unmanned Aerial Vehicles (UAVs) and handheld devices. This imagery captures as-built structural conditions, including standard positions, assembly deviations, and missing components. Computer vision algorithms extract these

as-built features and feed them back to the baseline FEM for periodic updating, correcting the idealized design geometry to match the physical state.

High-resolution field imagery was collected at a super-high-rise construction site, using a UAV (DJI Mini 4 Pro, DJI Technology Co., Ltd., Shenzhen, China) for comprehensive facade scanning. To resolve the pixel-level vertical features of standards and geometric details of couplers at the pixel level, the UAV maintained a standoff distance of 3 to 8 m from the scaffolding facade. A serpentine flight path combined orthorectified and 45° oblique perspectives. Longitudinal and lateral overlap rates were set to 80% and 60%, respectively, preserving the topological continuity of the structural members across consecutive frames. The acquisition sensor, with a native resolution of 4032×3024 pixels, clearly captures the edge contours of the $\phi 48.3$ mm steel tubes. Handheld high-definition cameras supplemented this dataset by capturing ground-level indoor shoring and heavily occluded zones. This approach ensured that the dataset encompassed challenging conditions, including abrupt illumination changes, safety net obstructions, and structural shadows from tower cranes.

While field imagery possesses high spatial resolution, Transformer-based detection networks are typically constrained to 640×640 pixel inputs due to GPU memory limitations. Directly downsampling the original high-resolution frames collapses the pixel width of slender members, thereby destroying critical edge gradient features. To circumvent this, this study implements a dual-track image processing pipeline. During the perception stage, an overlapping sliding window tiling algorithm discretizes the high-resolution frame into a series of sub-image patches. This overlap guarantees that structural elements intersecting the tile boundaries remain fully intact in at least one adjacent patch. In the subsequent geometric quantification stage, the system back-maps the detected Region of Interest (ROI) to the original high-resolution coordinate space. Fitting Hough lines to these sub-pixel features preserves engineering-level precision in spacing measurements.

Figure 1 outlines the image preprocessing and dataset construction pipeline. The process begins by loading the raw, high-resolution imagery. First, sliding window tiling discretizes the high-resolution frames into sub-image sequences to match the input specifications of the deep learning model. Bilateral filtering then removes high-frequency environmental noise, such as airborne dust, while preserving sharp structural edges. Data augmentation—incorporating random rotation, brightness jittering, and Gaussian noise injection—improves the model's adaptability to varied site environments. Next, manual interactive annotation establishes ground-truth bounding box labels for standards, ledgers, and couplers. Finally, the dataset is randomly partitioned into training and testing sets at an 8:2 ratio. Table 2 summarizes the scale and distribution of the compiled dataset.

The dataset was collected from three construction sites in Southern China. Images were acquired under varying conditions including morning/afternoon lighting, overcast/sunny weather, and different construction phases. Manual annotation was performed by two structural engineers with 5+ years of scaffolding inspection experience, using LabelMe software (version 5.11.3). Each image was independently annotated by both annotators; disagreements (occurring in approximately 8.3% of instances, primarily for occluded couplers) were resolved through consensus discussion, with inter-annotator IoU agreement reaching 0.91 for standards, 0.88 for ledgers, and 0.83 for couplers.

To prevent data leakage during the sliding window tiling process, we ensured that all tiles from a single original high-resolution image were assigned exclusively to either the training set or the test set. The dataset partition was performed at the original-image level before tiling. Specifically, the 2500 original field images were first divided into 2000 training images and 500 validation images (8:2 ratio). After sliding window tiling with overlap, this yielded 10,000 training patches and 2500 validation patches used during model development and hyperparameter tuning. The 250 test samples listed in Table 2

represent an additional independent test set collected from different construction phases and site locations than the training images. After tiling, these produced 1250 test patches.

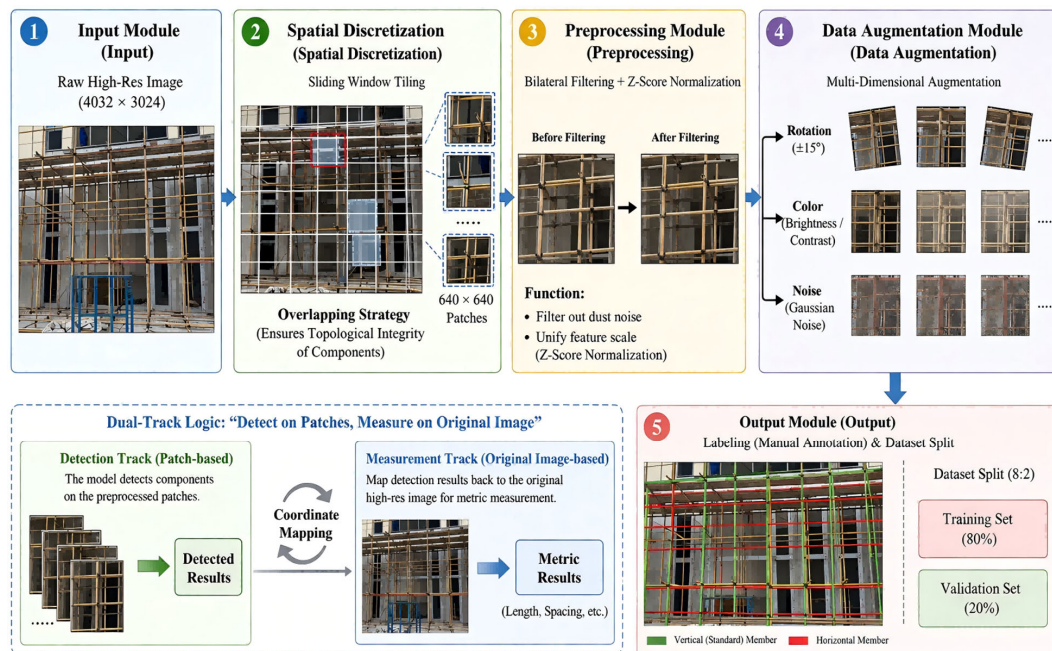


Figure 1. The flowchart of image data preprocessing.

Table 2. Statistics of the scaffolding image dataset.

Data Category	Acquisition Method	Original Resolution	Original Quantity	Quantity After Tiling	Primary Purpose
Design Drawings	Digital Scanning/Export	N/A	45	N/A	Establish APDL baseline model
Field Images	UAV/Handheld Camera	$\geq 4032 \times 3024$	2500	12,500	Identify member positions and defects
Test Samples	Random Field Sampling	$\geq 4032 \times 3024$	250	1250	Accuracy verification and error analysis

During preprocessing, bilateral filtering suppresses high-frequency environmental noise. Z-score normalization then unifies the feature scale across channels:

$$X_{norm} = \frac{X - \mu}{\sigma} \quad (1)$$

where X is the raw pixel value, while μ and σ represent the mean and standard deviation of the corresponding color channel. The normalized dataset serves as a robust foundation for training and validating the neural network.

2.2. Transformer-Based Object Detection Network

Construction scaffolding represents a complex spatial grid structure composed of numerous standards, ledgers, and couplers. These components exhibit high repetition and global topological correlation in their visual representation. Conventional CNNs are constrained by local receptive fields, often struggling to maintain feature continuity for slender members distributed over long distances, which results in missed detections or structural fragmentation. To address this, this study develops an object detection network that adapts the SegFormer encoder architecture for bounding box-based component localization.

Although SegFormer was originally designed for semantic segmentation, we adapted it for object detection in this work because scaffolding components exhibit distinctive geometric features with clear boundaries. The detection paradigm using bounding boxes provides significantly faster inference speed and higher computational efficiency compared to dense pixel-wise segmentation, while maintaining comparable recognition accuracy. For our application, bounding box localization is sufficient for subsequent geometric parameter extraction, as we only need to identify component centerlines rather than precise pixel-level boundaries for structural analysis.

The network architecture consists of two main components: a SegFormer-based hierarchical encoder for multi-scale feature extraction, and a detection head for bounding box regression and classification. The selection of the SegFormer encoder rests on two primary rationales. First, its hierarchical Transformer structure captures long-range semantic correlations between structural members via self-attention mechanisms, which is essential for identifying scaffolding systems with highly regular global layouts. Second, the Mix-FFN module extracts positional information without fixed positional encoding, improving adaptability to varying input resolutions and enhancing the model's portability for mobile deployment on construction sites. The encoder generates multi-scale feature pyramids which are then fed to the detection head for predicting bounding boxes, class labels, and confidence scores for each scaffolding component.

Figure 2 illustrates the overall network architecture and feature flow of SegFormer. The pipeline generates multi-scale feature maps through the Mix Transformer (MiT) encoder and fuses them through a detection head for bounding box prediction.

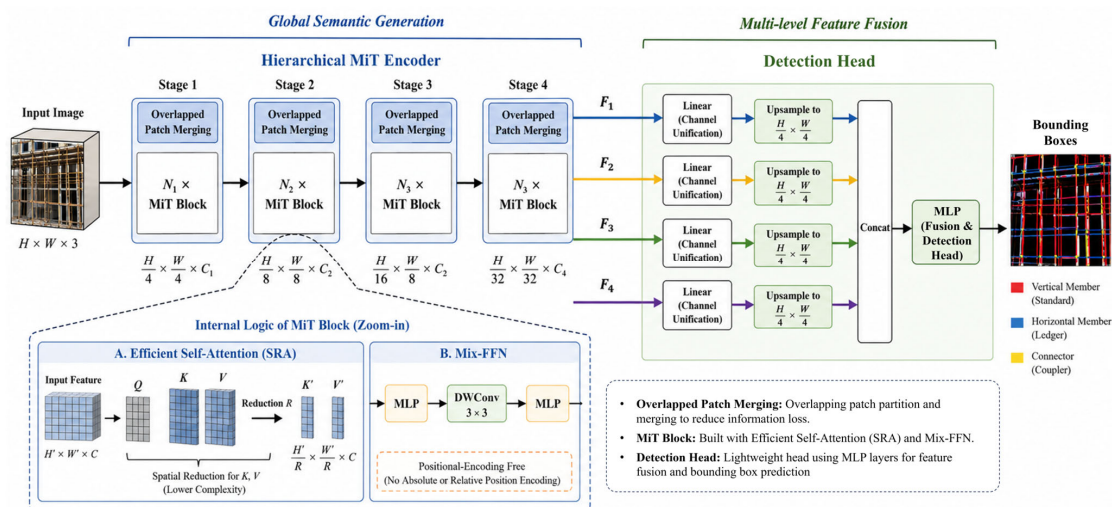


Figure 2. The overall network architecture combining SegFormer encoder with detection head for scaffolding component localization.

2.2.1. Architecture Details

The core design of our detection network integrates the hierarchical MiT encoder from SegFormer [21] with a task-specific detection head, adapting the original architecture from pixel-wise segmentation to bounding box-based object detection. Figure 2 shows the execution logic of the network when processing scaffolding features. Unlike traditional Transformers that use fixed-resolution feature mapping, the MiT encoder generates pyramid-structured feature maps through four consecutive stages. For an input image of dimension $\frac{H}{4} \times \frac{W}{4} \times C_1$, $\frac{H}{8} \times \frac{W}{8} \times C_2$, $\frac{H}{16} \times \frac{W}{16} \times C_2$, and $\frac{H}{32} \times \frac{W}{32} \times C_4$ through overlapped patch merging. This hierarchical representation enables the network to simultaneously capture local texture details of tiny couplers and the global topological morphology of long standards, providing multi-dimensional semantic support for subsequent geometric quantification.

To address computational redundancy from high-resolution site images, the model introduces a sequence reduction attention (SRA) mechanism in each Transformer block. This mechanism utilizes a sequence reduction factor R to reduce the length of the Key (K) and Value (V) matrices. Its calculation is defined as follows:

$$Q = XW^Q \quad (2)$$

$$K', V' = \text{Linear}(C, C)(\text{Reshape}(N/R, C)(\text{Pool}(R, X))) \quad (3)$$

$$\text{Attention}(Q, K', V') = \text{Softmax}\left(QK'^T / \sqrt{d_{\text{head}}}\right) \cdot V' \quad (4)$$

where $X \in R^{N \times C}$ is the input feature with $N = H \times W$ denoting the sequence length (H and W are the spatial height and width), C is the number of channels, W^Q is the query projection matrix, and d_{head} is the dimension of each attention head. Both Key and Value matrices undergo identical sequence reduction operations through spatial pooling (Pool), reshaping, and linear projection, reducing the sequence length from N to N/R . This mechanism linearly reduces the computational complexity of self-attention from $O(N^2)$ to $O(N^2/R)$. Such optimization allows the model to process detailed scaffolding edge features under limited computational resources.

To enhance robustness against heterogeneous resolution inputs and eliminate dependence on fixed positional encoding, SegFormer embeds a Mix-FFN in the feed-forward network. This module captures spatial positional information by introducing 3×3 depth-wise convolution (DWConv) between MLP layers:

$$X_{\text{out}} = \text{MLP}(\text{GELU}(\text{DWConv}_{3 \times 3}(\text{MLP}(X_{\text{in}})))) + X_{\text{in}} \quad (5)$$

The introduction of DWConv utilizes the positional leakage effect brought by zero padding, providing an implicit spatial position perception capability for the Transformer. For scaffolding targets that show significant scale differences under different shooting depths, Mix-FFN ensures model adaptability to complex site conditions.

Finally, benefiting from the powerful global receptive field of the encoder, the detection head processes multi-scale features through feature pyramid fusion. The detection head unifies feature channels of each stage through linear layers, subsequently performing bilinear upsampling and feature concatenation:

$$\hat{F}_i = \text{Upsample}(\text{MLP}(F_i)) \quad (6)$$

$$F_{\text{fuse}} = \text{MLP}(\text{Concat}(\hat{F}_1, \hat{F}_2, \hat{F}_3, \hat{F}_4)) \quad (7)$$

The fused features F_{fuse} are then fed to parallel classification and regression branches that predict bounding box coordinates (x, y, w, h), class probabilities, and objectness scores for each detected component. By highly fusing multi-scale features, the detection head ensures efficient inference for safety early-warning tasks. Table 3 lists the specific structural parameters of the SegFormer-B2 variant used in this study.

Table 3. Structural parameters of the SegFormer detection network.

Module Name	Parameter Type	Stage 1	Stage 2	Stage 3	Stage 4
Output Size	Output Dimensions	$\frac{H}{4} \times \frac{W}{4}$	$\frac{H}{8} \times \frac{W}{8}$	$\frac{H}{16} \times \frac{W}{16}$	$\frac{H}{32} \times \frac{W}{32}$
Channels	Feature Channels (C_i)	64	128	320	512
Reduction Ratio	Reduction Ratio (R_i)	8	4	2	1
Heads	Attention Heads	1	2	5	8
Encoder	Backbone Network	MiT-B2			
Parameters	Total Parameters	24.2 M			

2.2.2. Loss Function and Training

A multi-task compound loss function is designed for the scaffolding monitoring task. This loss function addresses the severe class imbalance caused by high background proportions and the extreme aspect ratios of the target members. The total loss function L_{total} is weighted by classification loss L_{vfl} , regression loss L_{ciou} , and distribution focal loss L_{dfl} :

$$L_{\text{total}} = \lambda_1 L_{\text{vfl}} + \lambda_2 L_{\text{ciou}} + \lambda_3 L_{\text{dfl}} \quad (8)$$

Varifocal Loss (L_{vfl}) utilizes an asymmetric weighting mechanism to dynamically adjust the loss contribution based on prediction quality. This design forces the network to focus more on high-quality positive samples with high-confidence predictions, effectively mitigating false-positive interference from safety nets or lifting equipment on construction sites. Complete-IoU Loss (L_{ciou}) incorporates the Euclidean distance between bounding box center points and aspect ratio consistency constraints. For slender targets such as scaffolding standards, L_{ciou} accelerates convergence along the longitudinal scale, preventing localization failures caused by minor lateral deviations. Distribution Focal Loss (L_{dfl}) optimizes the probability distribution of bounding box edges near their ground-truth coordinates. This optimization allows the model to learn highly discriminative boundary features when processing images with complex background textures, thereby improving sub-pixel localization accuracy.

The deep learning model was trained using the PyTorch 2.0 framework. Given the sensitivity of Transformer architectures to gradient optimization algorithms, the AdamW optimizer was selected for parameter updates. This algorithm suppresses overfitting by incorporating weight decay. The initial learning rate was set to 10^{-3} and was dynamically adjusted using a Cosine Annealing strategy. This scheduler accelerates the global parameter space search during early training and facilitates precise local fine-tuning in the later stages to ensure smooth convergence. The training run spanned 300 epochs with automatic mixed precision (AMP) enabled. By dynamically switching between FP16 and FP32 numeric representations, AMP significantly reduces GPU memory consumption and enhances hardware efficiency without sacrificing detection accuracy.

2.3. Simplified Hough Transform for Spacing Measurement

Following the methodology established in classical line detection literature [29,30], and building upon the parameter-domain restriction approach proposed by Song et al. [31], this study develops an optimized HT tailored for scaffolding geometry extraction. After object detection and region localization via SegFormer, the system converts pixel-level bounding box coordinates into analytical geometric equations with explicit physical meanings. The HT serves as a critical bridge by mapping edge pixels from the image space to the parameter space. To accommodate the distinct layout of scaffolding systems, this study proposes a simplified HT to enhance fitting stability in dense component scenarios.

2.3.1. Region of Interest Extraction

Based on the predicted bounding box coordinates, the system automatically crops the ROI containing target standards or ledgers. Bilateral filtering is first applied to the ROI images to eliminate interference from complex background noise (e.g., safety net textures, dust, and dynamic shadows). This filter preserves edge sharpness while smoothing noise by simultaneously evaluating spatial proximity and pixel intensity similarity. Subsequently, the Canny operator generates binary edge gradient maps. The set of pixels with a grayscale value of 255 in the binary image defines the feature point set $P = \{(x_i, y_i) \mid i \in n\}$. This preprocessing mechanism retains critical geometric contours while significantly reducing outlier interference during parameter space search.

2.3.2. Line Fitting and Spacing Calculation

The conventional HT performs a full-angle search in the polar coordinate space (ρ , θ), incurring significant computational overhead and showing susceptibility to interference from non-target linear features (e.g., hoisting ropes and construction board edges). Considering the quasi-vertical orientation of scaffolding standards along the gravity vector, this study implements a parameter-domain-restricted simplified HT algorithm. The computational flow follows the pipeline illustrated in Figure 3.

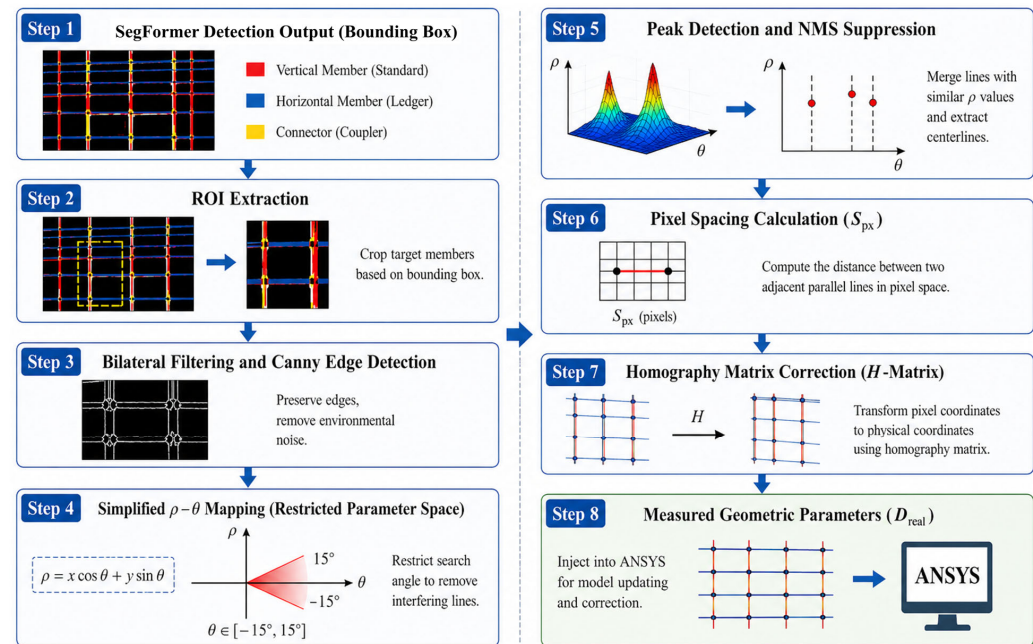


Figure 3. Flowchart of simplified HT calculation in scaffold identification.

First, the feature point set P is mapped to the polar coordinate equation:

$$\rho = x \cos(\theta) + y \sin(\theta) \quad (9)$$

To focus the fitting process, the angle search domain Θ is constrained to a narrow interval near the vertical normal. Specifically, $\theta \in [\theta_{\text{ref}} - \Delta\phi, \theta_{\text{ref}} + \Delta\phi]$, where the reference angle $\theta_{\text{ref}} = 0^\circ$ for standards, $\theta_{\text{ref}} = 90^\circ$ for ledgers, and the deviation tolerance $\Delta\phi$ is set to 15° . In the parameter accumulator, the system determines the centerline equations by locating local peaks (ρ^* , θ^*). To resolve redundant line fitting caused by physical member thickness, a non-maximum suppression (NMS) strategy is introduced. If the polar radius difference $|\rho_i - \rho_j|$ between two detected lines is smaller than a threshold τ and their angular deviation is minimal, they are classified as the same member, triggering a weighted average fusion. For two adjacent standards L_i and L_{i+1} with polar parameters ($|\rho_i, \theta_i|$) and (ρ_{i+1}, θ_{i+1}), the perpendicular spacing S_{px} is calculated as:

$$S_{\text{px}} = |\rho_i - \rho_{i+1}| \cdot \cos\left(\frac{\Delta\theta}{2}\right) \quad (10)$$

where $\Delta\theta = \theta_{i+1} - \theta_i$ represents the angular deviation between the two lines. The cosine term accounts for the projection correction when the two lines are not perfectly parallel. In practice, after NMS filtering and given the narrow angular search window ($\Delta\phi = \pm 15^\circ$), adjacent standards typically exhibit $|\Delta\theta| < 5^\circ$, resulting in $\cos(\Delta\theta/2) > 0.999$. This ensures the perpendicular distance measurement maintains sub-pixel accuracy even under moderate structural tilt.

Constraining the search width in the parameter space improves computational efficiency by approximately 60% compared to conventional algorithms while enhancing identification robustness in dense component arrangements.

2.3.3. Metric Conversion

Raw images commonly exhibit perspective distortion due to camera tilt, horizontal offset, and varying focal lengths, meaning pixel-level distances at different depths do not map to identical physical dimensions. To ensure that visual features interface accurately with the FE simulation, homography transformation is introduced for spatial scale restoration.

The homography matrix $\mathbf{H}$ between image pixel coordinates (u, v) and the physical plane coordinates (x_w, y_w) of the scaffolding is established:

$$\begin{bmatrix} x_w \\ y_w \\ 1 \end{bmatrix} \sim \mathbf{H} \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & 1 \end{bmatrix} \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} \quad (11)$$

The homography matrix $\mathbf{H}$ is solved using the Direct Linear Transformation (DLT) algorithm. By extracting four non-collinear feature points with known physical spacing—typically the four corners of a rectangular coupler plate with standardized dimensions of 48.3 mm $\times$ 48.3 mm, or alternatively four pre-positioned checkerboard calibration marks attached to the scaffold at known intervals—the matrix elements are fitted via the least squares method. The four-point configuration must form a non-degenerate quadrilateral to ensure unique solution of the 8-DOF homography matrix. Finally, the actual physical spacing D_{real} is calculated from the corrected spatial coordinate differences:

$$D_{\text{real}} = \sqrt{(x_{w,i} - x_{w,i+1})^2 + (y_{w,i} - y_{w,i+1})^2} \quad (12)$$

This transformation model eliminates geometric distortions caused by perspective projection, ensuring that visual spacing deviations serve as precise prior parameters, dynamically injected into ANSYS APDL scripts for structural safety evaluation.

Given that measurement uncertainties may propagate through the homography transformation and affect FEM input accuracy, we performed a Monte Carlo analysis with 1000 random samples to quantify this propagation. The uncertainty sources include: (1) calibration point localization error ($\sigma_{\text{calib}} = \pm 2.1$ pixels), (2) camera lens distortion (radial distortion coefficient $k_1 = -0.023$), and (3) edge detection jitter ($\sigma_{\text{edge}} = \pm 1.5$ pixels). The resulting 95% confidence interval for lift height measurement is $[\mu - 31.2 \text{ mm}, \mu + 28.7 \text{ mm}]$, where μ is the measured value. For the bay width, the interval is $[\mu - 27.4 \text{ mm}, \mu + 25.9 \text{ mm}]$. These uncertainty bounds are incorporated into the FEM as parameter ranges for sensitivity analysis.

3. Automated Structural Safety Evaluation Framework

Following the visual extraction of as-built scaffolding configurations, this section establishes an evaluation framework that automatically converts perceived geometric features into structural performance metrics. The core logic of the framework lies in modifying the idealized design FEM using the prior structural data captured via computer vision, such as component spatial coordinates, measured spacing deviations, and structural defects. Establishing this closed-loop pipeline shifts structural analysis from idealized design checking to customized, site-specific safety assessments. This methodology accounts for topological integrity and facilitates a quantitative evaluation of scaffolding safety by incorporating initial geometric uncertainties.

3.1. Parametric Finite Element Modeling

To integrate visual perceptual data with the computational mechanics environment, this study implements a parameter-driven modeling approach. This method abstracts the topological features, geometric parameters, and boundary conditions of the scaffolding into an adjustable variable matrix, establishing a baseline FEM capable of responding to external input data. Compared to conventional static modeling, this parametric model offers high flexibility, automatically reconstructing local element connections and nodal coordinates through numerical algorithms based on the coordinate deviations fed back by the vision system. This serves as an efficient computational foundation for the near-real-time safety assessment of large-scale repetitive structures.

3.1.1. Element Selection and Mechanical Representation

Scaffolding represents a spatial composite structure composed of slender steel tubes, connection couplers, and top load-transfer systems. To simulate the actual mechanical response in the as-built state, differentiated element types are assigned to capture the distinct mechanical characteristics of each structural component. The element selection and mechanical representation follow established FEM practices for scaffold modeling [34,35]. Standards and ledgers act as the primary load-bearing components, modeled using spatial beam elements based on Timoshenko beam theory. These elements account for the contribution of shear deformation to structural stiffness and describe the second-order effects of slender members under compression (the $P - \Delta$ effect), which is essential for evaluating global stability. Structural members such as diagonal braces, which only transfer axial tension and compression, are discretized using truss elements to balance numerical accuracy and computational efficiency.

The connection stiffness of the coupler joints significantly influences scaffolding stability. In as-built structures, the stiffness of coupler nodes typically exhibits non-linear semi-rigid behavior. This study characterizes the rotational resistance of couplers under varying tightening torques by implementing spatial spring elements with non-linear rotational stiffness at member intersections. Furthermore, considering the distribution of construction loads, shell elements are introduced to model the horizontal formwork system. This configuration permits accurate transmission of surface loads to the support system through surface-to-line interactions, ensuring the FEM remains highly consistent with actual loading paths. Regarding material properties, elastoplastic constitutive parameters are assigned to the steel according to relevant engineering standards, establishing a material basis for subsequent non-linear analysis.

3.1.2. Parameterization Logic and Defect Injection

The model generation logic maps the parameter space to the geometric space, driving the array generation of 3D nodes through a global variable matrix. During the topological construction stage, the system automatically searches for spatial node indices using iterative algorithms and generates standard, ledger, and diagonal brace elements based on preset connectivity rules. Geometric dimensions, including standard lift heights and bay widths, are maintained as variables in the data interface rather than fixed values. When the vision system identifies assembly deviations at a specific level, the modeling program dynamically updates the local mesh topology through an external variable rewrite mechanism.

A critical step in this workflow is the injection of initial geometric defects. The safety of as-built scaffolding is often governed by initial tilt or member eccentricity. This study obtains the spatial slope of the standard centerline via visual recognition, converting it into a coordinate perturbation component. A coordinate correction algorithm maps the identified initial position deviation e_0 directly into the node coordinate vector of the FEM. This

measurement-driven correction logic transitions the FEM from an idealized geometric state to an as-built state reflecting actual assembly defects. Additionally, if missing components are identified, the program executes logical deletion through an element index retrieval mechanism. This modeling pipeline achieves a precise mapping from perceived physical features to geometric variables and topologies within the structural simulation model.

3.2. Vision-to-Simulation Data Mapping

Integrating perceived visual data with computational models relies on a robust data mapping mechanism. This process quantifies the unstructured image features derived in Section 2, converting them into geometric increments and topological commands compatible with the FE solver. This mapping mechanism eliminates errors associated with manual secondary modeling and establishes a direct data link for periodic safety verification of the scaffolding.

3.2.1. Identification of Irregular Configurations

After extracting the as-built parameters, the system identifies non-compliant assembly states through multi-dimensional comparison logic. For geometric parameter deviations, the algorithm performs vector subtraction between the measured physical spacing matrix D_{real} and the design vector D_{design} . If the local deviation $\Delta L = |L_{real} - L_{design}|$ exceeds the allowable limit (e.g., ± 50 mm), the corresponding location is flagged as a geometric anomaly. For missing component detection, the system utilizes the bounding box predictions output by SegFormer to perform spatial matching with the preset topological grid. If the confidence level of a target component at the specific coordinate falls below a threshold and no valid pixels accumulate in its neighborhood, the ledger or diagonal brace is classified as missing, and its topological index is recorded. For initial geometric defects, the system uses standard centerline slopes extracted by the simplified HT to calculate the spatial offset vectors of each node relative to the vertical axis. This quantitative deviation data serves as the primary input for FEM modification, enabling the mechanical evaluation to capture stability and bearing capacity reductions caused by assembly defects.

3.2.2. Automated APDL Script Generation

To eliminate manual intervention and enhance analysis efficiency, this study developed a parameter-driven automated mapping middleware that achieves end-to-end conversion from pixel features to mechanical command streams. The middleware encapsulates the visual recognition results into a structured spatio-temporal database containing node indices, geometric dimensions, and defect categories. During model modification, the middleware adopts a template-based generation strategy, first calling the parametric baseline file established in Section 3.1 and then dynamically generating variable override instructions based on the database information. If standard tilt is identified, the middleware calculates the cumulative offset of each layer and applies a global perturbation to the spatial vectors of the FE nodes via a coordinate correction operator, injecting the initial geometric defect into the model. If a missing component is detected, the corresponding element topology index is retrieved to execute a logical deletion command. Finally, the middleware transmits the generated command stream file to the finite element solver kernel. All subsequent procedures, from mesh reconstruction to non-linear solving, are executed in batch mode, ensuring automated safety feedback for construction site monitoring.

The data structure of the middleware is organized as follows: each detected component is stored with a unique identifier, 3D spatial coordinates (x, y, z), component type (standard/ledger/diagonal), confidence score, and dimensional measurements. For geometric defect injection, the workflow proceeds in three steps: (1) the baseline APDL

script is loaded with parametric node definitions $[N_{id}, x_{design}, y_{design}, z_{design}]$; (2) for each identified defect (e.g., tilt angle θ at standard i), the middleware calculates the nodal offset $\Delta x = h \cdot \tan(\theta)$, where h is the height; (3) the coordinate override command “N, N_{id} , $x_{design} + \Delta x$, y_{design} , z_{design} ” is injected into the APDL script. For missing members, the system queries the element connectivity table to retrieve the element ID and issues the “EDELE, $element_{id}$ ” command to remove it from the model. This automated process eliminates manual intervention and ensures consistency between visual observations and FE model geometry.

3.3. Mechanical Analysis Procedures

The modified FEM undergoes stability and strength verifications under multiple load combinations to determine the residual bearing capacity and current safety level of the as-built structure. This section details the execution of the structural analysis, focusing on the mechanical influence of visually extracted as-built deviations on overall structural behavior. Figure 4 illustrates the corresponding computational workflow from model updating to safety evaluation.

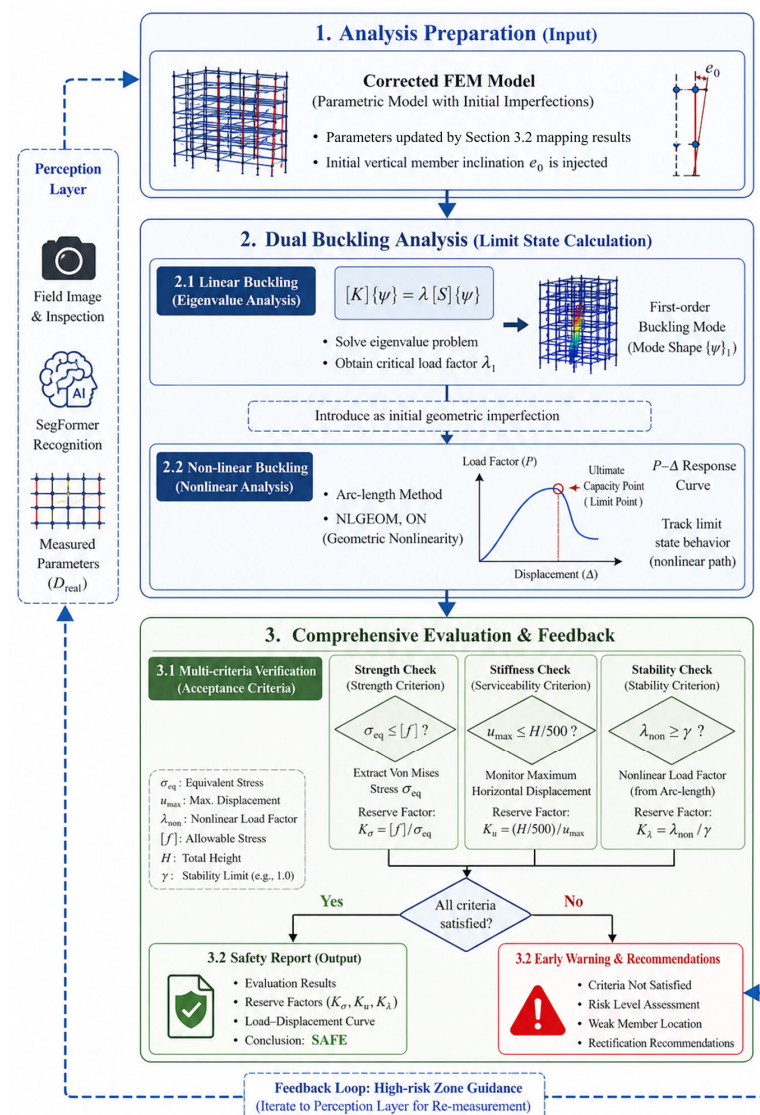


Figure 4. Workflow of dual buckling analysis and closed-loop feedback system.

3.3.1. Linear and Non-Linear Buckling Analysis

As a space-lattice system primarily composed of slender compression members, scaffolding typically fails through global or local buckling triggered by geometric instability. To capture this behavior, the proposed framework implements a coupled analysis pipeline spanning linear and non-linear regimes. First, an eigenvalue buckling analysis solves the characteristic equation $[K]\{\psi\} = \lambda[S]\{\psi\}$ to obtain the critical load factor λ_{cr} and the corresponding first-order buckling mode shape. Although eigenvalue analysis yields an idealized upper bound of structural capacity, these buckling modes establish a fundamental baseline for understanding the structural stiffness distribution.

To address the stochastic assembly deviations common in as-built scaffolding (such as standard tilt and lift height non-uniformity identified by the vision pipeline), this study introduces a full-process geometric non-linear analysis to capture the actual ultimate state of the structure. This stage incorporates geometric non-linearities and utilizes the Arc-length method for numerical iteration. Tracking the load–displacement ($P - D$) response curve iteratively enables the Arc-length solver to stably traverse the limit point during structural instability, accurately capturing the post-buckling behavior and capacity reduction induced by prior defects. This approach isolates the stability degradation caused by assembly deviations, providing realistic performance indicators for site safety monitoring.

3.3.2. Stress and Displacement Check

Following stability verification, the framework performs rigorous strength and stiffness checks on the as-built scaffolding. For the strength verification, the FEM extracts the Von Mises equivalent stress σ_{eq} for each structural member under the most unfavorable load combination, comparing these values against the design compressive strength f specified in the disc-lock scaffolding safety standard JGJ 231-2021 [36]. Calculating the strength reserve coefficient $K_\sigma = f/\sigma_{max}$ quantitatively evaluates the structural load-bearing redundancy.

For stiffness checking, the system monitors the maximum horizontal displacement u_{max} of the top-layer standards and the global drift ratio. If the lateral displacement exceeds the allowable serviceability limit (e.g., $H/500$) or the stability factor falls below the early-warning threshold, the system triggers a structural warning. A key feature of this framework is the iterative feedback loop: critical high-stress zones and displacement-sensitive nodes identified in the FE solver inversely guide the perception layer, prompting the vision pipeline to perform high-frequency monitoring of these specific hot spots. This perception–simulation–feedback loop achieves an automated, closed-loop pipeline extending from raw visual imagery to quasi-static structural performance indices.

4. Experimental Results and Case Study

This section experimentally validates the proposed automated evaluation framework using a representative engineering case study. First, a quantitative evaluation is conducted on the detection accuracy and geometric extraction error of the visual recognition algorithm under complex construction backgrounds. Subsequently, a super-high, large-span support shoring system from a large public building project is selected as the research object to demonstrate the complete workflow from field image acquisition to structural safety evaluation. Comparing the ultimate load-bearing capacity differences between the idealized design state and the as-built state quantifies the structural impact of actual assembly deviations.

4.1. Evaluation of Identification Accuracy

To verify the performance of the scaffolding geometry identification pipeline (incorporating SegFormer and the simplified HT), extensive evaluations were performed on a

custom scaffolding dataset. The experimental hardware environment was executed on an NVIDIA RTX 3080 GPU, utilizing the PyTorch 2.0 deep learning framework.

4.1.1. Performance of the Detection Model

First, the recognition performance of the detection network was evaluated on three core components: standards, ledgers, and couplers. mAP, Precision, and Recall served as the primary evaluation metrics. The training process followed the hyperparameter configurations described in Section 2.2.2. Specifically, the weight coefficients for the compound loss function were empirically set as $\lambda_1=1.0$, $\lambda_2=2.0$, and $\lambda_3=1.5$. A higher weight was assigned to λ_2 (regression loss) to enhance the model's sensitivity to the geometric localization of slender members, which is the foundation of subsequent geometry extraction. λ_3 (distribution focal loss) was optimized to refine member boundary probability distributions, ensuring sub-pixel precision in centerline fitting.

Figure 5 illustrates the training process. Experimental results demonstrate that after 300 epochs of training, the model achieved a mAP@0.5 of 95.2% on the independent test set (1250 patches from 250 original images, as described in Section 2.1), alongside a mAP@0.5:0.95 score of 78.6%. The validation set was used solely for hyperparameter tuning during training and is not the source of these reported metrics. The global self-attention mechanism of the encoder enables the model to maintain the structural continuity of slender members across complex backgrounds (e.g., safety nets and crane arms).

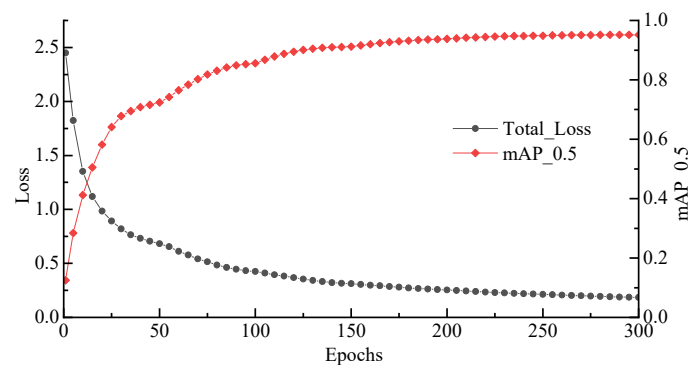


Figure 5. Training process of the SegFormer model.

In qualitative evaluations, the model maintained high classification accuracy under challenging conditions, such as dusk with rapid illumination transitions and localized occlusions. Predicted confidence maps indicate that the spatial perception of the Mix-FFN module provides high discriminative power for densely arranged couplers, effectively suppressing false-positive interference from background textures. Table 4 summarizes the detailed detection performance for each component category.

Table 4. Detection performance of the proposed model on scaffolding components.

Component Category	Precision	Recall	mAP@0.5
Standards	96.8%	94.5%	96.1%
Ledgers	94.2%	92.8%	93.5%
Couplers	95.5%	93.2%	94.8%
Mean	95.5%	93.5%	95.2%

To investigate the correlation between model scale and recognition efficiency, different variants of the SegFormer series (from SegFormer-B0 to SegFormer-B5) were compared. As summarized in Table 5, increasing the depth and channel dimensions of the MiT encoder yields step-wise accuracy improvements. Although the lightweight SegFormer-B0 variant

achieves a high inference speed (112 FPS), its limited receptive field depth increases the susceptibility to missed detections in dense coupler regions and distant slender members. In contrast, the SegFormer-B5 variant achieved the highest accuracy (mAP@0.5 of 96.3%), but its larger parameter scale (81.2 M) increased inference latency, which limits its suitability for time-sensitive site inspections. The selected SegFormer-B2 variant represents an optimal balance between localization accuracy and computational overhead, maintaining 95.2% accuracy while providing a processing speed of 45 FPS.

Table 5. Performance comparison analysis of SegFormer with different parameter scales (B0–B5).

Model Variant	Parameters (M)	GFLOPs	Inference Speed (FPS)	mAP@0.5	mAP@0.5:0.95
SegFormer-B0	3.7	8.4	112	91.2%	72.4%
SegFormer-B1	13.7	15.9	78	93.8%	75.1%
SegFormer-B2	24.2	31.7	45	95.2%	78.6%
SegFormer-B3	44.0	79.0	28	95.8%	79.3%
SegFormer-B4	60.8	95.7	19	96.1%	80.2%
SegFormer-B5	81.2	183.3	9	96.3%	80.8%

Notes: Inference speed (FPS) measured on 640×640 image patches using NVIDIA RTX 3080 GPU.

To further validate the superiority of the SegFormer-based encoder for scaffolding detection, we compared our approach with several representative baseline models across different architectural paradigms. Table 6 summarizes the performance comparison with state-of-the-art detectors.

Table 6. Comparison with baseline detection models on the scaffolding dataset.

Model	Architecture Type	mAP@0.5	mAP@0.5:0.95	FPS	Parameters(M)
Faster R-CNN [16]	Two-stage CNN	89.3%	71.2%	18	41.3
YOLOv8-m [18]	One-stage CNN	92.7%	74.8%	56	25.9
Mask R-CNN	Two-stage CNN	90.8%	73.5%	12	44.2
SegFormer-Det (Ours)	Transformer-based	95.2%	78.6%	45	24.2

The results demonstrate that the Transformer-based encoder provides superior performance for detecting slender, overlapping scaffolding members. The global self-attention mechanism enables the model to maintain semantic continuity across long-range members that CNN-based detectors often fragment under local occlusions. While YOLOv8-m achieves higher inference speed (56 FPS), it shows lower accuracy (92.7% vs. 95.2%) due to its limited receptive field when processing dense component arrangements. Our approach balances detection accuracy with near-real-time performance requirements.

4.1.2. Accuracy of Spacing Measurement

Following detection-based positioning, the system fits member centerlines within the identified regions using the simplified HT to generate analytical geometric descriptions for FE analysis. To evaluate measurement accuracy, 50 sampling zones were randomly selected on-site. Table 7 summarizes the error statistics comparing the manual ground truth (D_{true}) obtained via laser rangefinders with the visually extracted spacing values (D_{real}). The mean absolute error (MAE) and root mean square error (RMSE) for the extracted standard lift heights are 18.5 mm and 24.2 mm, respectively, with a 95% confidence interval (CI) of $[-31.2, +28.7]$ mm derived from the Monte Carlo analysis in Section 2.3.3. Under safety standards (e.g., JGJ 130-2021), the allowable tolerance for scaffolding assembly deviation is ± 50 mm. Both the MAE and the 95% CI upper bound remain within this allowable toler-

ance, confirming the framework’s capability to reliably identify non-compliant assembly configurations for as-built FEM updating.

Table 7. Statistics of geometric parameter extraction accuracy for scaffolding.

Geometric Parameter	Sample Mean (mm)	MAE (mm)	RMSE (mm)	95% CI (mm)	Relative Error (%)	Evaluation Conclusion
Step Height	1500.0	18.5	24.2	[−31.2, +28.7]	1.23%	Meets standard
Bay Width	1200.0	16.4	21.5	[−27.4, +25.9]	1.37%	Meets standard
Initial Tilt	48.3 (Pipe Dia)	3.2	4.5	[−5.1, +4.8]	3.12%	Meets defect injection accuracy

Figure 6 shows the complete visual processing pipeline from raw imagery to sub-pixel geometric quantification. During the perception stage (Figure 6b), SegFormer generates bounding boxes for standards (red), ledgers (blue), and couplers (yellow), providing topological context. In the subsequent quantification stage (Figure 6c), the system back-maps the ROIs to the high-resolution space and performs sub-pixel centerline fitting using the simplified HT. Applying an angular search constraint of $\Delta\phi = 15^\circ$ in the parameter space and leveraging the NMS strategy to merge redundant coordinates, the algorithm successfully suppresses interference from dense safety nets in the background, achieving high-confidence filtering of target member edges. Following homography transformation, the pixel-level spacing (e.g., 98.3 px) maps to the actual physical distance of 1500 mm.

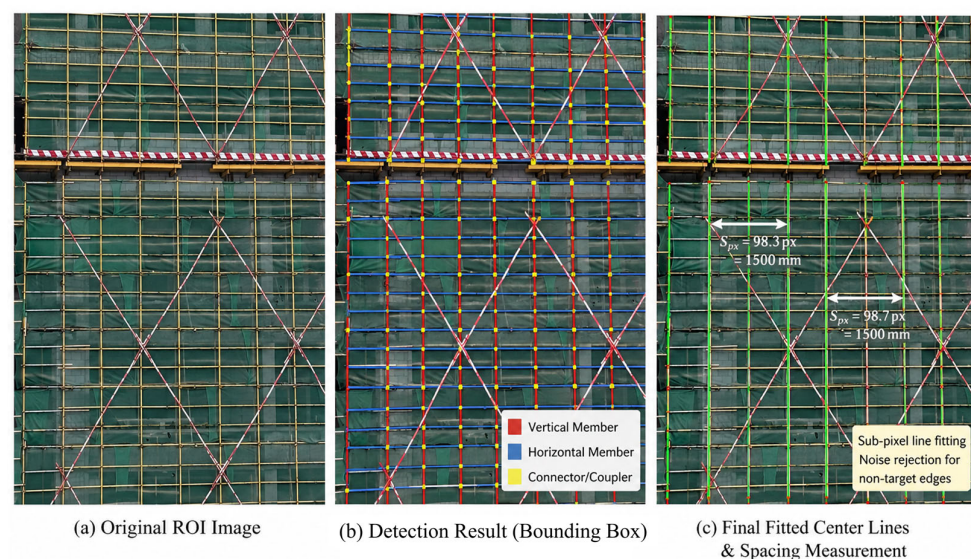


Figure 6. Visual results of line fitting in a typical ROI region.

4.2. Simulation Case Study

This section selects a super-high support shoring system from a large commercial financial center project as a typical engineering case. The project is located in a coastal city in Southern China. The structure is complex, and the construction site faces challenges such as high temperature, high humidity, and frequent strong wind interference. The investigation evaluates the structural safety of the as-built disc-lock scaffolding system under realistic operating conditions.

4.2.1. Project Overview

The case study focuses on a high-formwork shoring section used for large-span frame beam construction. With a total support height of 32 m, the structure is classified as a high-risk engineering shoring system [37]. According to the JGJ 231-2021 standard, the

scaffolding utilizes disc-lock steel pipe supports with a specification of $\phi 48.3 \times 3.5$. The overall grid dimensions are $15.6 \times 6.6 \times 32$ m, with a standard spacing (x-axis) of 1.8 m, standard bay width (y-axis) of 1.5 m, and a standard lift height of 1.5 m. At the heavily loaded zone directly beneath the frame beams, the standard spacing is reduced to 0.6 m. Rigid tie connections are installed at five elevations (5.93 m, 11.43 m, 16.93 m, 22.43 m, and 27.93 m) to provide global stability constraints.

In the FEM, BEAM188 elements represent the standards and ledgers, whereas LINK180 elements represent the diagonal braces. Connections between standards are modeled as rigid, and connections between diagonal braces and standards are modeled as hinged. Non-linear spring elements (COMBIN39) represent the semi-rigid rotational stiffness of the disc-lock connections in the vertical plane. The steel material properties include an elastic modulus (E) of 206 GPa, Poisson's ratio (ν) of 0.33, and a density (ρ) of 7850 kg/m³. A top layer of SHELL181 elements models the horizontal formwork to distribute construction loads uniformly. Static construction loads represent actual working conditions: surface loads of 30,136.3 N/m² and 20,412.3 N/m² are applied to the frame beam shoring zone, with a standard load of 9468.8 N/m² applied to the adjacent areas.

4.2.2. Comparative Analysis

Using the automated mapping middleware, the study injected the “as-built parameter set” extracted via vision into the mechanical solver. The visual monitoring results revealed that due to overlapping site operations, the actual lift height at the mid-levels of the shoring structure reached 1.72 m (representing a 14.7% deviation from the design value of 1.5 m). Additionally, centerline fitting revealed an average initial standard tilt angle of 1.8°.

To provide context for the simulation results, we first outline the FEM validation approach. The predicted first-order eigenvalue for the ideal state (25.077) was compared against analytical solutions using the Perry-Robertson formula for slender columns with multiple bracing points, yielding an analytical critical load factor of 23.8 (5.4% difference), which falls within acceptable engineering tolerance. The assumed semi-rigid coupler stiffness (8.5 kN·m/rad) was calibrated based on experimental rotational stiffness tests reported in reference [37], measuring 7.2 to 9.8 kN·m/rad for disc-lock connections. Boundary conditions (pinned base, rigid ties at five elevations) match actual site constraints verified through field inspection, with load combinations following GB 50009-2012.

The comparative analysis first determined the stability baseline through eigenvalue buckling analysis. As shown in Table 8, the ideal design model exhibits a clear hierarchical instability pattern. The first-order buckling mode shape represents global lateral buckling in the x-axis direction ($\lambda_1 = 25.077$), highlighting the global stiffness bottleneck of the 32 m structure along its weak axis. The second and third modes evolve into global torsion and wave-like bending of local members. Notably, the eigenvalues from the second to the fifth order are extremely dense (ranging from 37.2 to 41.7), suggesting that the structure is highly susceptible to multi-modal coupling instability under high-order loads.

Table 8. First five-order buckling load factors (SET) of the ideal design model.

Mode	Eigenvalue (TIME/FREQ)	Cumulative Step	Mode Description
1	25.077	1	Global lateral buckling (x-axis)
2	37.246	2	Global torsional buckling
3	38.611	3	Local member buckling and inter-story drift
4	40.228	4	Second-order flexural-torsional mode
5	41.694	5	Local buckling at the top area

Building on the baseline eigenvalue results, we observe that both the linear eigenvalue analysis and the nonlinear arc-length analysis yield a stability reduction of 22.4%. While these two methods capture different mechanical behaviors—idealized bifurcation versus $P - \Delta$ -amplified progressive collapse—the numerical coincidence occurs because the initial geometric imperfection (1/150 tilt combined with 14.7% lift height deviation) triggers first-mode-dominated buckling. In the eigenvalue analysis, this imperfection reduces the critical load factor from 25.077 to 19.46. In the nonlinear analysis, the same imperfection pattern causes the structure to follow the first eigenmode trajectory during post-buckling, with the load–displacement curve exhibiting a limit-point load of 89.9 kN followed by softening behavior. The proportional reduction (from 115.8 kN to 89.9 kN, representing 22.4%) reflects the dominance of the first buckling mode. Both analyses were performed independently using different solution procedures (modal extraction for eigenvalue analysis vs. geometric nonlinear arc-length for full-process prediction), confirming that the observed coincidence stems from the eigenmode-consistent imperfection geometry rather than methodological coupling. This correspondence indicates that global instability governs the failure mechanism, though the exact numerical coincidence also reflects the specific imperfection magnitude calibrated to the first mode.

Figure 7 illustrates the comparison of structural buckling mode shapes under the first-order limit state. Injecting the visually detected 14.7% lift height deviation reduces the first-order eigenvalue load factor from 25.077 to 19.46. Deformation contours indicate that while the idealized design model exhibits a symmetrical deformation trend, the modified model displays pronounced asymmetrical lateral bulging concentrated in the weak levels where the lift height reaches 1.72 m. This localized stiffness degradation induces premature structural buckling.

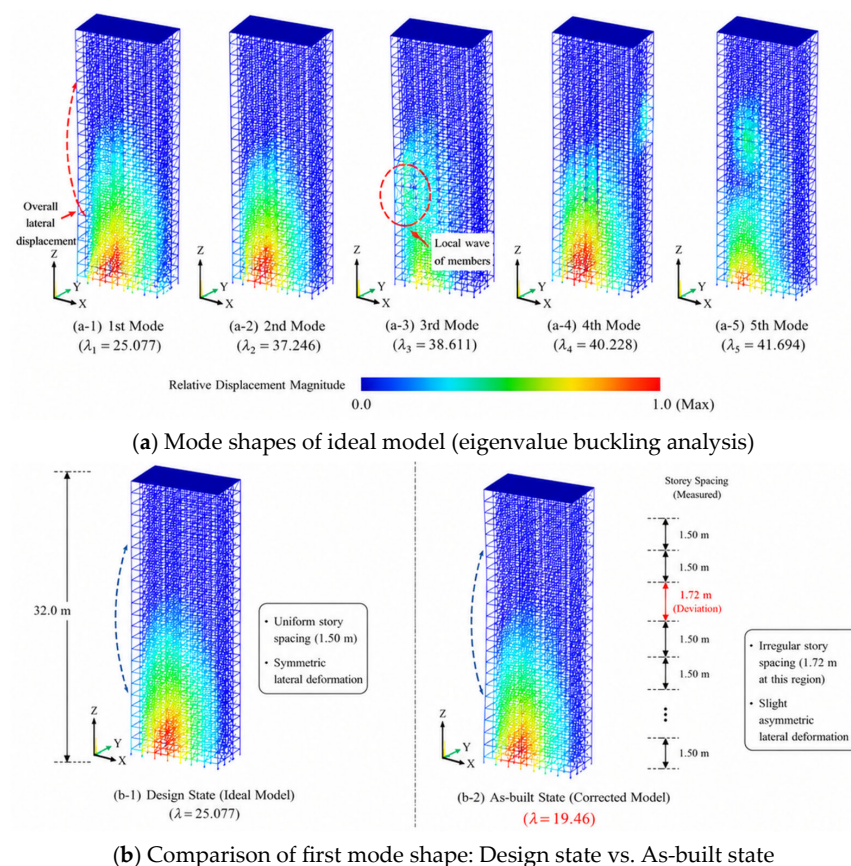


Figure 7. Mechanical analysis results of the scaffolding structure.

The full-process geometric nonlinear analysis further reveals that at the 32 m elevation, gravity-induced secondary moments amplify the initial tilt defect e_0 through $P - \Delta$ effects, contributing exponentially to the standard bending moments. This quantitatively confirms that centimeter-level assembly deviations can precipitate substantial reductions in load-bearing capacity for super-high shoring systems.

It must be acknowledged that the structural predictions presented are simulation-based and have not been validated against experimental load testing or instrumented field measurements. Conducting full-scale destructive testing on a 32 m high-risk scaffold structure is neither economically feasible nor ethically permissible during active construction. The FEM results should therefore be interpreted as comparative indicators of relative capacity degradation due to geometric deviations, rather than absolute predictions of ultimate failure load. Future work should incorporate non-destructive testing methods such as dynamic modal analysis or strain gauge validation to partially verify the model's predictions in non-critical structural elements.

4.2.3. Safety Early-Warning and Risk Assessment

A visualization feedback mechanism integrates these numerical FE metrics to establish a closed-loop perception-analysis-warning pipeline. To translate abstract numerical evaluations into clear operational decisions, the framework defines a three-level warning criterion: a red alert is triggered if the safety factor $FS < 1.2$ or the top-layer horizontal displacement $u_{\max} > H/500$; a yellow warning is activated if the stability reduction rate exceeds 15%; and structural configurations within nominal limits maintain a green safety status.

Table 9 summarizes the comparative performance metrics between the ideal design state and the as-built state. The results reveal significant degradation across multiple structural indicators. The maximum Von Mises equivalent stress in the middle-level standards increases from 141.4 MPa to 184.7 MPa (30.6% increase), reducing the strength reserve coefficient from 1.45 to 1.11. Although the peak stress remains below the yield strength (205 N/mm²), this reduced margin increases susceptibility to localized yielding under dynamic loading.

Table 9. Comparison of structural performance indicators between the ideal state and the as-built state.

Evaluation Indicator	Ideal Design State	As-Built State	Change Rate (%)	Risk Level
Max Step Height (m)	1.50	1.72	+14.7%	Geometric Anomaly (Red)
Stability Coefficient λ_1	25.077	19.46	−22.4%	Stability Warning (Red)
Ultimate Capacity (kN)	115.8	89.9	−22.4%	Insufficient Safety Factor (Red)
Max Equivalent Stress (MPa)	141.4	184.7	+30.6%	Reduced Strength Margin (Yellow)
Top Max Displacement (mm)	12.5	28.3	+126.4%	Significant Stiffness Degradation (Red)

Structural displacement responses exhibit strong nonlinear behavior under lateral loads. The maximum horizontal displacement at the top layer reaches 28.3 mm under wind loading, compared to 12.5 mm in the design state (126.4% increase). This displacement amplification far exceeds the 14.7% geometric deviation itself, demonstrating how second-order $P - \Delta$ effects exponentially amplify deformation when initial defects are present.

To close the risk mitigation loop, the middleware projects critical element coordinates back onto the raw UAV imagery using inverse homography mapping. On the monitoring interface, spatial grid regions with excessive lift height deviations and significant stability loss are highlighted in red, triggering targeted reinforcement instructions. The case study illustrates that the proposed framework successfully translates subtle assembly deviations

(e.g., a 220 mm lift height deviation) into quantitative mechanical safety indices. This continuous flow—from raw visual feature extraction to structural performance degradation and risk level mapping—transitions scaffolding supervision from traditional compliance checking to active, digital safety assurance.

5. Conclusions and Future Work

This study addresses the need for automated and quantitative safety supervision of scaffolding systems on construction sites. An integrated, closed-loop evaluation framework was proposed and implemented, combining SegFormer-based visual perception, a simplified HT, and parameterized FE analysis. Through systematic algorithm development, dataset evaluations, and engineering case study validation, the primary research findings are summarized as follows:

- (1) An object detection network based on SegFormer-B2 was successfully constructed to precisely locate slender components under complex construction backgrounds. This architecture leverages hierarchical Transformer encoding to capture long-range topological relationships, achieving a mAP@0.5 of 95.2% with an inference speed of 45 FPS per 640×640 patch. Compared with conventional convolutional networks, the proposed model suppresses environmental noise from safety nets and extreme illumination, establishing a reliable semantic foundation for subsequent geometric extraction.
- (2) A dual-track image processing pipeline and a parameter-domain-restricted simplified HT were developed to overcome feature collapse caused by image downsampling. Through sliding window tiling and ROI inverse coordinate mapping to the raw pixel space, sub-pixel centerline fitting was successfully executed. Constraining the angular search domain to a $\pm 15^\circ$ interval and incorporating an NMS fusion strategy limited the relative measurement errors to within 3.5%, with a mean absolute error (MAE) of 18.5 mm for lift height measurements. These outcomes indicate the reliability of the proposed pipeline for sub-pixel centerline fitting and dense component extraction.
- (3) A digital, closed-loop perception–simulation–evaluation pipeline was established. The developed automated data mapping middleware dynamically injects measured as-built geometric deviations (e.g., non-uniform standard lift heights and vertical misalignment) into the structural FE solver. Numerical case studies indicate that the ultimate load-bearing performance of high-rise scaffolding is highly sensitive to geometric irregularities. Specifically, a 14.7% deviation in standard lift height, coupled with an initial standard tilt of 1/150, precipitates a 22.4% reduction in the simulated structural capacity factor.

Although the framework shows promising performance in the case study, several limitations warrant further investigation. First, the structural predictions are simulation-based and require validation through non-destructive testing methods such as dynamic modal analysis or instrumented monitoring. Second, the current implementation performs quasi-static analysis based on captured images, making it suitable for periodic inspection rather than continuous monitoring. Future research will explore integration with video streams for real-time structural health monitoring under dynamic construction loads. Third, visual detection accuracy remains constrained under heavy occlusion, suggesting potential benefits from multi-modal sensing schemes integrating LiDAR and optical sensors. Finally, deploying lightweight models on mobile edge devices represents a critical direction for achieving on-site instant inference.

Author Contributions: Conceptualization, H.P. and J.D.; methodology, H.P. and H.W.; software, H.P. and Y.D.; validation, H.P., J.D. and L.Z.; formal analysis, H.P., L.Z. and J.D.; investigation, H.P. and Y.D.; resources, H.P. and L.Z.; data curation, H.P. and H.W.; writing—original draft preparation, H.P., L.Z., H.W. and J.D.; writing—review and editing, H.P.; visualization, H.P. and Y.D.; supervision, H.P.; project administration, H.P.; funding acquisition, H.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding. The study was conducted as independent research by the authors.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

AI	Artificial Intelligence	LSD	Line Segment Detector
AMP	Automatic Mixed Precision	MAE	Mean Absolute Error
APDL	ANSYS Parametric Design Language	mAP	mean Average Precision
CAD	Computer-Aided Design	MiT	Mix Transformer
CNN	Convolutional Neural Network	Mix-FFN	Mixed Feed-Forward Network
DLT	Direct Linear Transformation	NMS	Non-Maximum Suppression
DWConv	Depth-Wise Convolution	RANSAC	Random Sample Consensus
FE/FEM	Finite Element/Finite Element Method	RMSE	Root Mean Square Error
FPS	Frames Per Second	ROI	Region of Interest
GPU	Graphics Processing Unit	RPN	Region Proposal Network
HOG	Histogram of Oriented Gradients	SRA	Sequence Reduction Attention
HT	Hough Transform	UAV	Unmanned Aerial Vehicle
IoU	Intersection over Union	ViT	Vision Transformer

References

1. Yin, Z.; Caldas, C. Scaffolding in industrial construction projects: Current practices, issues, and potential solutions. *Int. J. Constr. Manag.* **2022**, *22*, 2554–2563.
2. Chuang, K.-Y.; Tserng, H.-P. A study of automated intelligent monitoring of scaffolding safety for construction projects. *J. Chin. Inst. Eng.* **2024**, *47*, 626–642. [[CrossRef](#)]
3. Dzung, R.-J.; Cheng, C.-W.; Cheng, C.-Y. A scaffolding assembly deficiency detection system with deep learning and augmented reality. *Buildings* **2024**, *14*, 385. [[CrossRef](#)]
4. Ramezantitkanloo, A.; Czepizak, D.; Pieńko, M. Problems of monitoring critical elements of scaffolding. *Arch. Civ. Eng.* **2024**, *LXX*, 233–254. [[CrossRef](#)]
5. Islam, M.R.; Zamil, M.Z.H.; Rayed, M.E.; Kabir, M.M.; Mridha, M.F.; Nishimura, S.; Shin, J. Deep learning and computer vision techniques for enhanced quality control in manufacturing processes. *IEEE Access* **2024**, *12*, 121449–121479. [[CrossRef](#)]
6. Wu, P.; He, X.; Dai, W.; Zhou, J.; Shang, Y.; Fan, Y.; Hu, T. A Review on Research and Application of AI-based Image Analysis in the field of Computer Vision. *IEEE Access* **2025**, *13*, 76684–76702. [[CrossRef](#)]
7. Gupta, S.; Singh, U.; Rawat, M.; Kurian, A.; Mandal, L.; Gupta, P.; Sardar, T.H. Best image processing for higher face detection rate using Haar cascades. In *Proceedings of the International Conference on Robotics and Networks*; Springer: Berlin/Heidelberg, Germany, 2023; pp. 113–128.
8. Xu, X.; Lei, Y.; Yang, F. Railway subgrade defect automatic recognition method based on improved Faster R-CNN. *Sci. Program.* **2018**, *2018*, 4832972. [[CrossRef](#)]
9. Xia, B.; Luo, H.; Shi, S. Improved faster R-CNN based surface defect detection algorithm for plates. *Comput. Intell. Neurosci.* **2022**, *2022*, 3248722. [[CrossRef](#)] [[PubMed](#)]
10. Hussain, M. YOLO-v1 to YOLO-v8, the rise of YOLO and its complementary nature toward digital manufacturing and industrial defect detection. *Machines* **2023**, *11*, 677. [[CrossRef](#)]

11. Chen, Z.; Zhu, Q.; Zhou, X.; Deng, J.; Song, W. Experimental study on YOLO-based leather surface defect detection. *IEEE Access* **2024**, *12*, 32830–32848. [\[CrossRef\]](#)
12. Chen, P.; Li, R.; Fu, K.; Zhong, Z.; Xie, J.; Wang, J.; Zhu, J. A cascaded deep learning approach for detecting pipeline defects via pretrained YOLOv5 and ViT models based on MFL data. *Mech. Syst. Signal Process.* **2024**, *206*, 110919. [\[CrossRef\]](#)
13. Zhou, X.; Ren, Z.; Zhang, Y.; Mi, T.; Zhou, S.; Jiang, Z. A shunted-swin transformer for surface defect detection in roller bearings. *Measurement* **2024**, *238*, 115283. [\[CrossRef\]](#)
14. Priadana, A.; Habibi, M. Face detection using haar cascades to filter selfie face image on instagram. In *Proceedings of the 2019 International Conference of Artificial Intelligence and Information Technology (ICAIIIT)*, Yogyakarta, Indonesia, 13–15 March 2019; IEEE: New York, NY, USA, 2019; pp. 6–9.
15. Zhang, X.; Zhang, L.; Lou, X. A raw image-based end-to-end object detection accelerator using HOG features. *IEEE Trans. Circuits Syst. I Regul. Pap.* **2021**, *69*, 322–333.
16. Xu, X.; Zhao, M.; Shi, P.; Ren, R.; He, X.; Wei, X.; Yang, H. Crack detection and comparison study based on faster R-CNN and mask R-CNN. *Sensors* **2022**, *22*, 1215. [\[CrossRef\]](#) [\[PubMed\]](#)
17. Chai, B.; Nie, X.; Zhou, Q.; Zhou, X. Enhanced cascade R-CNN for multiscale object detection in dense scenes from SAR images. *IEEE Sens. J.* **2024**, *24*, 20143–20153. [\[CrossRef\]](#)
18. Li, Y.; Fan, Q.; Huang, H.; Han, Z.; Gu, Q. A modified YOLOv8 detection network for UAV aerial image recognition. *Drones* **2023**, *7*, 304. [\[CrossRef\]](#)
19. Wang, A.; Chen, H.; Liu, L.; Chen, K.; Lin, Z.; Han, J.; Ding, G. Yolov10: Real-time end-to-end object detection. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 107984–108011. [\[CrossRef\]](#)
20. Yin, Z.-B.; Liu, F.-Y.; Geng, H.; Xi, Y.-J.; Zeng, D.-B.; Si, C.-J.; Shi, M.-D. A high-precision jujube disease spot detection based on SSD during the sorting process. *PLoS ONE* **2024**, *19*, e0296314. [\[CrossRef\]](#) [\[PubMed\]](#)
21. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 12077–12090.
22. Kong, Y.; Shang, X.; Jia, S. Drone-DETR: Efficient small object detection for remote sensing image using enhanced RT-DETR model. *Sensors* **2024**, *24*, 5496. [\[CrossRef\]](#) [\[PubMed\]](#)
23. Xie, Y.; Ma, X.; Zhao, Q. Research on target detection network based on improved Swin-DETR. In *Proceedings of the 2023 4th International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE)*, Hangzhou, China, 25–27 August 2023; IEEE: New York, NY, USA, 2023; pp. 324–328.
24. Yang, X.; del Rey Castillo, E.; Zou, Y.; Wotherspoon, L. UAV-deployed deep learning network for real-time multi-class damage detection using model quantization techniques. *Autom. Constr.* **2024**, *159*, 105254. [\[CrossRef\]](#)
25. Lin, X.; Zhou, Y.; Liu, Y.; Zhu, C. A comprehensive review of image line segment detection and description: Taxonomies, comparisons, and challenges. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 8074–8093. [\[CrossRef\]](#) [\[PubMed\]](#)
26. Ehret, T.; Morel, J.-M. Line segment detection: A review of the 2022 state of the art. *Image Process. Line* **2024**, *14*, 41–63. [\[CrossRef\]](#)
27. Chung, K.-L.; Chang, W.-T. Centralized RANSAC-based point cloud registration with fast convergence and high accuracy. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 5431–5442. [\[CrossRef\]](#)
28. Zheng, R. Laser marking method for nonferrous metal casting ingots based on improved RANSAC algorithm. *Manuf. Rev.* **2025**, *12*, 7. [\[CrossRef\]](#)
29. Zhang, B.; Hu, S.; Du, J.; Yang, X.; Chen, X.; Jiang, H.; Cao, H.; Feng, S. Detecting moving objects in photometric images using 3D Hough transform. *Publ. Astron. Soc. Pac.* **2024**, *136*, 054502. [\[CrossRef\]](#)
30. Matarneh, S.; Elghaish, F.; Al-Ghraibah, A.; Abdellatef, E.; Edwards, D.J. An automatic image processing based on Hough transform algorithm for pavement crack detection and classification. *Smart Sustain. Built Environ.* **2025**, *14*, 1–22.
31. Song, J.; Lu, T.; Wang, H.; Zhou, J.; Zhu, Q.; Wang, J.; Song, W. Automatic rebar spacing measurement using YOLOv11-rebar-light and simplified Hough transform. *Autom. Constr.* **2026**, *186*, 106907. [\[CrossRef\]](#)
32. Luo, Y.; Wang, X.; Liao, Y.; Fu, Q.; Shu, C.; Wu, Y.; He, Y. A review of homography estimation: Advances and challenges. *Electronics* **2023**, *12*, 4977. [\[CrossRef\]](#)
33. Ilić, M.; Ivanović, M.; Kurbalija, V.; Valachis, A. Towards optimal learning: Investigating the impact of different model updating strategies in federated learning. *Expert Syst. Appl.* **2024**, *249*, 123553. [\[CrossRef\]](#)
34. Ilcik, J.; Arora, V.; Dolejs, J. Design of new scaffold anchor based on the updated finite element model. *Eng. Struct.* **2016**, *118*, 334–343. [\[CrossRef\]](#)
35. Błazik-Borowa, E.; Jamińska-Gadomska, P.; Pieńko, M. Influence of foundation quality on the stress in the elements of steel façade scaffolding. *Buildings* **2020**, *10*, 130. [\[CrossRef\]](#)

36. JGJ 231-2021; Technical Standard for Disc-Lock Steel Tubular Scaffold in Construction. China Architecture & Building Press: Beijing, China, 2021.
37. Huang, H.; Peng, Z.; Hou, J.; Zheng, X.; Ding, Y.; Wu, H. Study on the Ultimate Load-Bearing Capacity of Disc Buckle Tall Formwork Support Considering Uncertain Factors. *Buildings* **2024**, *14*, 828. [[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.