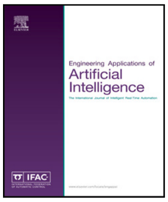
Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai



Research paper

Efficient and robust shoveling control system based on semantic elevation mapping for unmanned loaders

Guangda Chen [a,*], Zhiwen Zhang [a], Lin Cheng [b], Cheng Jin [a], Shunyi Yao [a], Yue Wang [c], Rong Xiong [c], Yingfeng Chen [a,*]

[a] *Fuxi Robotics in NetEase, Hangzhou, 310056, China*
[b] *College of Energy Engineering, Zhejiang University, Hangzhou, 310027, China*
[c] *State Key Laboratory of Industrial Control and Technology, Zhejiang University, Hangzhou, 310027, China*

ARTICLE INFO

Keywords:
Wheel loader
Shoveling control
Semantic segmentation
Elevation mapping
Calibration
Shoveling point selection

ABSTRACT

Improving the automation of wheeled loaders is key to solving labor gaps and boosting safety in construction. This paper proposes an automatic shoveling system for unmanned loaders that, for the first time, balances safety, robustness, efficiency, and energy consumption. The system features automatic calibration of camera and light detection and ranging (LiDAR) using large segmentation models and nonlinear optimization, ensuring stability despite vibrations. A lightweight neural network performs semantic segmentation, and multi-frame point clouds are fused with a confidence algorithm for accurate pile segmentation. The shoveling point selection algorithm integrates semantic and elevation data to prioritize loader and environmental safety. Volume prediction initiates scooping, and a shoveling strategy balances robustness and efficiency. Extensive field tests conducted over two months with two types of loaders in three scenarios, totaling 2090 operations, demonstrate the system's long-term stability, high bucket full rates, efficiency matching manual operations, and an 11% reduction in energy consumption. These results highlight the system's potential to transform automated construction machinery.

1. Introduction

The wheel loader is a pivotal asset in construction and a multitude of industries, owing to its adaptability (Frank et al., 2018). Nonetheless, conventional manual operation encounters several obstacles. Operating a wheel loader demands a high level of skill, with operators needing to frequently alternate between moving forward and backward while controlling the mechanical arm (Zauner et al., 2020). The demanding workload makes it difficult for drivers to sustain high-quality and efficient performance over extended periods (Nezhadali et al., 2016; Frank et al., 2012; Zhang et al., 2021). Consequently, there is a diminishing interest among younger generations to pursue this career, resulting in a labor shortage (Agrawal et al., 2023). In the specialized scenarios such as mining, workers must wait until the dust from explosions has settled to safe levels before they can safely enter the work area, which greatly diminishes productivity (Cardenas et al., 2023). Although remote control technology has been integrated into the construction machinery sector, it remains in its infancy (Lee et al., 2022). Challenges like network latency and the absence of sensory feedback from the field regarding acceleration, vibrations, and depth can affect the accuracy and efficiency of remote operations, potentially posing safety hazards (Dadhich et al., 2018; You et al., 2023). Therefore, enhancing the automation of wheel loaders is essential for alleviating labor shortages, increasing production efficiency, and improving safety.

Automatic shoveling system is a key component of loader automation, directly affecting task success rates, operational efficiency and energy consumption (Filla et al., 2014). Generally, the automatic shoveling process primarily consists of four steps: perceiving information about the material pile, selecting suitable shoveling points, determining the appropriate timing for scooping, and controlling the mechanical arm to complete the loading process (Almqvist, 2009; Hemami and Hassani, 2009). However, automatic shoveling systems are still in the early stages of commercial application (Dadhich et al., 2016; Fernando and Marshall, 2020), and existing technologies face many challenges in perception and interaction with irregular piles, such as unstable perception outcomes, unreasonable shoveling point selection, and inadequate scooping action choices. Thus, optimizing various aspects of

[*] Corresponding authors.

E-mail addresses: cgdsss@mail.ustc.edu.cn (G. Chen), zhangzhiwen01@corp.netease.com (Z. Zhang), 3210105356@zju.edu.cn (L. Cheng), jincheng01@corp.netease.com (C. Jin), yaoshunyi@corp.netease.com (S. Yao), ywang24@zju.edu.cn (Y. Wang), rxiong@zju.edu.cn (R. Xiong), yingfengchen2016@163.com (Y. Chen).

https://doi.org/10.1016/j.engappai.2026.113908
Received 27 April 2025; Received in revised form 5 November 2025; Accepted 17 January 2026

the automatic shoveling process is essential to advance the practical application of unmanned loaders.

Perceiving the information of the material pile is a crucial initial step. Compared to human operators, unmanned systems face significant challenges in real-time acquisition of target pile information and extraction of effective features (Chen et al., 2024). Current mainstream perception schemes are primarily categorized into two classes: site-based fixed sensing and loader-based onboard sensing. In the domain of site-based fixed sensing, Kamari and Ham (2021) proposed an automated approach integrating deep learning with multi-view 3D reconstruction. By employing point cloud semantic segmentation and meshing-based volume calculation, their method achieves accurate volume measurement of material piles at construction sites. Xu et al. (2022, 2024) developed a sliding system based on the integration of a multi-echo scanner and a rangefinder for automated real-time inventory of industrial stockpiles in indoor environments. Coupled with timestamp synchronization and coordinate calibration algorithms, it generates high-precision 3D point clouds of the pile surface, with volume calculated via a slicing integration method. This type of scheme provides comprehensive, stable, and clear observational coverage, yielding more complete perceptual information. However, its drawbacks include the necessity for site modification, leading to high hardware installation and maintenance costs. More critically, transforming the perceived shovel point coordinates from the fixed sensor coordinate system to the robot's native frame can introduce significant calibration and registration errors (Gu et al., 2025). In terms of onboard sensing, Sarata et al. (2008) conducted pioneering research on vehicular measurement systems, developing an automatic loading system for wheel loaders based on a stereo vision system to dynamically detect and update pile geometry. Koyachi and Sarata (2009) focused on perception solutions for unmanned loaders, employing an onboard stereo vision system (dual CCD cameras) to capture real-time images of the pile for generating 3D point cloud models. Such systems offer advantages including high integration, ease of deployment without site modification, and convenient setup. Nonetheless, their perceptual field of view is highly susceptible to changes in the loader's posture, prone to resulting in observational blind spots. Moreover, the intense vibrations and impacts during loading operations make it difficult to maintain long-term stability of the sensor extrinsic calibration parameters, severely compromising perception accuracy. To address the limited sensing range of a single sensor, Chen et al. (2024) experimented with deploying five LiDAR units on top of the loader cabin. While this dense configuration enhances perceptual coverage, it concurrently introduces substantial calibration costs, accumulated errors, and significant computational overhead. It is particularly noteworthy that although LiDAR has become a mainstream choice for environmental perception due to its excellent performance in capturing depth and scale information, it inherently lacks the ability to perceive environmental texture features. Consequently, some studies (Chen et al., 2022a) have incorporated visual sensors to achieve finer material discrimination through image segmentation. However, vision-based methods remain vulnerable to variations in ambient lighting conditions, and their robustness in practical industrial settings still requires further improvement.

Secondly, selecting and calculating the appropriate shoveling point is another challenge for automatic shoveling systems. Research indicates that the shoveling direction should be perpendicular to the convex surface of the material pile (Lindmark and Servin, 2018). If the shoveling direction significantly deviates, the moments on either side of the bucket centerline will become asymmetric, leading to load imbalance on the bucket linkage mechanism and potentially causing overturning (Sarata, 2006). Sarata (2006) and Sarata et al. (2008) employed a columnar model to characterize the material pile and calculated the resistance exerted by the material on the bucket at various points, considering the bucket's trajectory, to determine the direction of the shoveling point that offered the most balanced force distribution. However, this method relies heavily on a single criterion and is contingent upon the accuracy of the complete three-dimensional perception of the material surface. To mitigate the dependence on the precision of surface perception, Magnusson and Almqvist (2011) conducted a quadratic surface fitting on local material surface perception information to estimate local convexity and sideload characteristics. Chen et al. (2024) proposed a shoveling point selection scheme that integrates four indicators: convexity, inclination, slope, and distance traveled, further enriching the selection criteria for shoveling points. Nevertheless, current research primarily focuses on the safety implications of the material pile on the loader's shoveling process, without adequately addressing the potential safety hazards that the loader may pose to the surrounding environment during operation. Materials are often indiscriminately piled in open production environments, necessitating that automated shoveling systems have a more comprehensive understanding of the environment surrounding the material pile.

After identifying the optimal shoveling point, the loader must control the bucket to glide along the ground and approach the material pile. Upon the bucket's penetration into the pile, the automated system must determine the appropriate moment to initiate control over the bucket and boom to execute the scooping process, which constitutes another complex challenge for automatic shoveling systems. Existing studies predominantly employ threshold control strategies based on passive signal triggering. For instance, Chen et al. (2024) utilized changes in boom cylinder pressure as a control signal, while Cao et al. (2023b,a) adopted differences in tire speed and engine torque fluctuations as triggering thresholds, respectively. These thresholds are typically determined empirically through experimentation to initiate the boom movement for loading operations. However, when the loader approaches material piles with varying shapes and properties under different working conditions, the dynamic responses can differ significantly, resulting in poor adaptability of fixed thresholds (Yang et al., 2025a; Yang, 2025; Yang et al., 2025b). Excessively low threshold settings may cause premature triggering, reducing bucket fill factor; whereas overly high thresholds can lead to delayed activation, not only increasing energy consumption (Yao et al., 2023) but also potentially causing tire slippage or even equipment damage (Fernando et al., 2018; Wu, 2003). More fundamentally, such passive triggering mechanisms rely on lagging system response signals and fail to base decisions on the real-time volume of material being crowded during the loading process. For piles with irregular shapes, such as low-lying piles, it is challenging to ensure a stable bucket fill rate. To address this, Sarata (2006) and Magnusson and Almqvist (2011) proposed active triggering approaches based on perceived material volume, dynamically initiating the scooping action by evaluating the displaced material volume in real time. However, current methods still exhibit two major limitations: first, they do not adequately account for the inherent motion delay characteristics of hydraulic systems (Dadhich et al., 2019); second, they rely solely on instantaneous material volume for decision-making without incorporating predictive mechanisms to estimate future states, consequently restricting real-time control accuracy.

Upon determining the optimal timing for initiating the scooping action, the selection of an appropriate scooping strategy remains a complex challenge. Filla et al. (2014) systematically categorized four primary strategies employed in scooping operations. Among these, two have gained prominent adoption in both academic and industrial contexts: the "Just in & out" strategy and the "Stairway" strategy. The "Just in & out" strategy, as implemented by Chen et al. (2024) and further validated in studies such as Chen et al. (2022b, 2023), utilizes a single, continuous motion sequence. This approach is characterized by its operational simplicity and high efficiency, making it particularly suitable for handling easily scoopable materials. However, Li et al. (2021) demonstrated that when encountering dense or highly compacted piles, this method is susceptible to bucket stalling, which significantly reduces scooping efficiency and may lead to operational failure. In contrast, the "Stairway" strategy, extensively investigated by Cao et al. (2023a,b), employs a multi-phase cutting process that progressively engages the pile surface to reduce digging resistance. This method proves more effective in challenging material handling conditions, albeit at the cost of increased cycle time and higher energy consumption, as quantified in earlier work by Filla et al. (2005). Current research efforts often concentrate on optimizing a single strategy in isolation. Consequently, as highlighted by Aoshima et al. (2023) and further supported by Eriksson et al. (2024a), autonomous shoveling systems struggle to maintain a robust balance between operational reliability, efficiency, and energy consumption across varied material properties and dynamic working environments. Furthermore, prevailing technologies predominantly prioritize achieving a high bucket fill rate, while neglecting precise volumetric control of the scooped material. The capability to accurately regulate the loaded volume would not only optimize material handling time and prevent overflow during transport but also enable critical applications requiring precise mixture ratios of multiple material types.

To address the challenges faced during various stages of automatic shoveling, this paper proposes a low-cost, onboard sensor-based automatic shoveling system. To ensure the long-term stability of sensor fusion under the severe vibration conditions of construction machinery, a self-supervised adaptive automatic calibration method for cameras and LiDAR is employed, based on the large-scale neural network model SAM (Kirillov et al., 2023). By integrating sparse historical information from multiple frames of fewer LiDAR sensors, real-time elevation map (Miki et al., 2022) reconstruction of large-scale piles is achieved. And a real-time semantic segmentation model suitable for the working environment of loaders is trained. This fusion of semantic and elevation information enables robust and precise pile surface segmentation under various environmental interferences. An algorithm for selecting shoveling points is proposed, considering both loader and environmental safety. Additionally, unlike passive triggering methods based on hydraulic pressure changes, scooping actions are triggered by material volume, using a data-driven approach to predict the future state of the scooping process and account for hydraulic delay characteristics. An adaptive scooping algorithm is designed to select appropriate strategies based on actual working conditions and adjust in real-time based on perceived volume feedback. The proposed system has been extensively validated through 2090 large-scale real shoveling tasks on various fuel and electric loaders, demonstrating long-term stability and significant advantages. To the best of the authors' knowledge, this is the first method capable of simultaneously considering long-term robustness, efficiency, and energy consumption in actual open production environments, highlighting its importance for the research and application of automatic construction machinery. A demonstration video can be found at https://youtu.be/uHHbI35hjsY. In summary, the contributions of this paper include:

- Proposing the first complete automatic shoveling system suitable for open production scenarios, addressing efficiency, energy consumption, safety, and robustness simultaneously.
- Introducing a self-supervised camera–laser automatic calibration method based on the large model SAM and a probability fusion method for material surface perception based on semantics and elevation confidence, achieving precise and robust segmentation of the material surface.
- Designing a shoveling point selection algorithm that considers semantic information from the semantic elevation map, ensuring both loader and environmental safety.
- Predicting the future state of the scooping process using a data-driven approach to enable timely triggering of scooping actions despite hydraulic delays. An adaptive scooping algorithm based on shoveling volume feedback further enhances efficiency and reduces energy consumption.
- Demonstrating the system's long-term stability and leadership through extensive testing in multiple real mixing plant scenarios with various types of loaders, achieving efficiency comparable to skilled human operators while reducing average energy consumption by 11%.

The main contents are as follows. Section 2 outlines the components of our automatic shoveling system, and then Sections 3–6 explain in detail the principles and methods of calibration, pile perception, shoveling point decision, and shoveling control strategy, respectively. Section 7 presents the implementation of the experiments, data collection, and results analysis. Section 8 addresses the identified limitations and suggests avenues for future research. A concluding summary of the study is provided in Section 9.

2. System overview

The overall framework of the proposed automatic shoveling system is illustrated in Fig. 1, which is primarily composed of three parts: the establishment of the semantic elevation map, the selection of shoveling points, and the execution of shoveling actions. Specifically, A precise calibration technique based on nonlinear optimization has been developed for the kinematics of the arm, integrated with Inertial Measurement Unit (IMU) sensors to achieve real-time perception and precise control of the bucket's attitude. Specifically, two types of LiDAR sensors are employed: a 3D LiDAR, which is integrated with an Inertial Measurement Unit (IMU) using Simultaneous Localization and Mapping (SLAM) technology (Shan et al., 2020) to achieve real-time localization of the loader and provide long-range panoramic perception; and a solid-state LiDAR dedicated to short-range forward perception. The perception data from both sensors are fused to construct a comprehensive environmental model. By integrating multi-source, multi-frame point cloud data from these LiDAR sensors, an elevation map-based environmental perception model is constructed. Additionally, a novel semantic segmentation model based on a single monocular camera and deep neural networks has been utilized to achieve semantic cognition in complex operating scenarios of the loader. After calibration of the LiDAR and camera, semantic information can be combined with the elevation map, and an overall semantic elevation map is formed through a confidence-based fusion algorithm, enabling precise segmentation of the pile surface. Subsequently, each position and angle on the pile surface line are comprehensively evaluated based on the pile elevation map to determine suitability for shoveling. Finally, the vehicle approaches the pile based on the location of the shoveling point, during which the loader predicted the shoveling volume in real-time to determine the moment for executing the scooping action. Once the predicted scooping volume reaches the desired volume, the loader autonomously selects an appropriate scooping strategy based on the material type and vehicle operating conditions, with the option for closed-loop feedback if necessary.

3. Sensor configuration and calibration

As shown in Fig. 2, to provide the automatic shoveling system with real-time and necessary joint state of the loader and environmental information, a variety of sensors have been installed on different parts of the loader. Three IMU sensors are placed on the cab, boom, and bucket, respectively, to obtain real-time attitude information of the loader body and arm. The steering encoder is mounted at the pivot points of the front and rear chassis to capture steering angle information. Additionally, environmental perception equipment has been mounted on the top of the cab, including a wide-angle monocular camera and a solid-state laser sensor that form the forward perception module, as well as a 16-line LiDAR for perception and localization. The solid-state laser effectively compensates for the sparsity of forward close-range material surface perception data associated with the 16-line LiDAR. This setup achieves a balance between performance and cost efficiency.

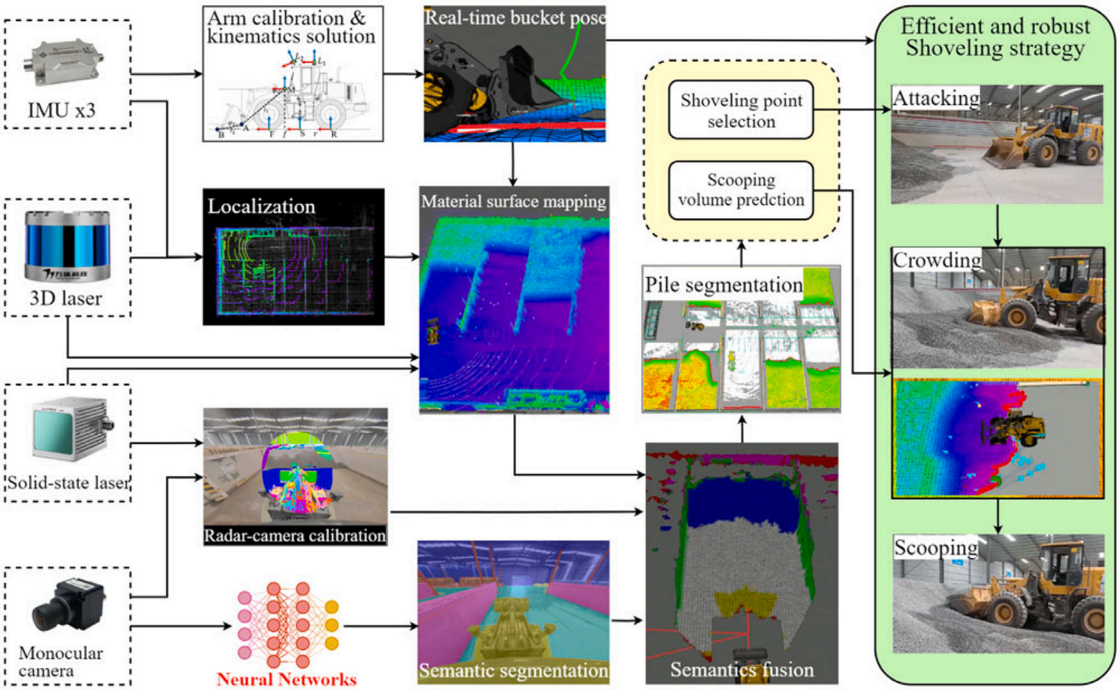


Fig. 1. Framework of the proposed automatic shoveling system for unmanned loaders.

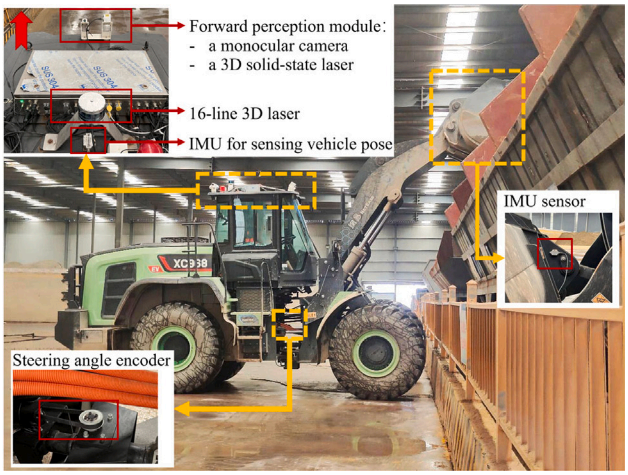


Fig. 2. The key sensor configurations involved in the automatic shoveling system of unmanned loaders.

3.1. Kinematic calibration of the arm

The main components of the loader arm can be simplified to rigid links, with joints represented as hinges (as illustrated in Fig. 3). The coordinate system M is defined at the pivot point of the arm, with the x-axis aligned horizontally towards the front of the vehicle. Point A represents the rotational point of the bucket, and $B(x_b, z_b)$ denotes the tip of the bucket teeth. Thus, the angles (φ_1, φ_2) between the boom and bucket and the loader arm coordinate system M are determined as follows:

$$\begin{cases} \varphi_1 = \theta_2 - \theta_1 + \alpha_1 \\ \varphi_2 = \theta_3 - \theta_1 + \alpha_2 \end{cases} \tag{1}$$

Here, $(\theta_1, \theta_2, \theta_3)$ correspond to the measurements obtained from the IMU sensors, and the parameters α_1 and α_2 denote the offset angles between the IMU sensors and the arm linkages. Since the IMU sensors are rigidly attached to the arm assembly, α_1 and α_2 are constants. Consequently, the kinematic equations of the loader arm can be expressed as:

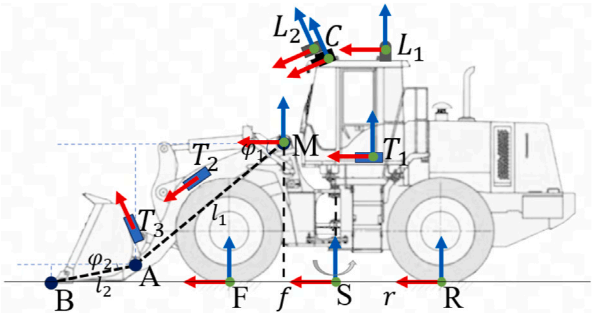


Fig. 3. The kinematic model of the loader arm, where M is the coordinate system of the loader arm.

$$\begin{cases} x_b = l_1 \cos\varphi_1 + l_2 \cos\varphi_2 \\ z_b = l_1 \sin\varphi_1 + l_2 \sin\varphi_2 \end{cases} \tag{2}$$

Concurrently, an accurate calibration method is proposed by minimizing the error between the measured displacement of the bucket endpoint and the calculated displacement to accurately determine the model parameters $(l_1, l_2, \alpha_1, \alpha_2)$. During the calibration phase, the endpoint of the bucket linkage is controlled to various positions (x_i, z_i), where $i \in [1, n]$, and subsequently measure the positional changes $(\Delta x_{ij}, \Delta z_{ij})$ at these points. By integrating the IMU sensor readings at two distinct positions, $\vartheta_i = (\theta_1^i, \theta_2^i, \theta_3^i)$ and $\vartheta_j = (\theta_1^j, \theta_2^j, \theta_3^j)$, two residual terms can be constructed:

$$\begin{cases} R_{ij}^x = \|\Delta x_{ij} - (x_b(\vartheta_i) - x_b(\vartheta_j))\|^2 \\ R_{ij}^z = \|\Delta z_{ij} - (z_b(\vartheta_i) - z_b(\vartheta_j))\|^2 \end{cases} \tag{3}$$

where $(x_b(\vartheta_i), z_b(\vartheta_i))$ and $(x_b(\vartheta_j), z_b(\vartheta_j))$ are the positions calculated based on the arm model (Eqs. (1) and (2)) and are functions of $(l_1, l_2, \alpha_1, \alpha_2)$. The optimization problem is formulated as follows, and

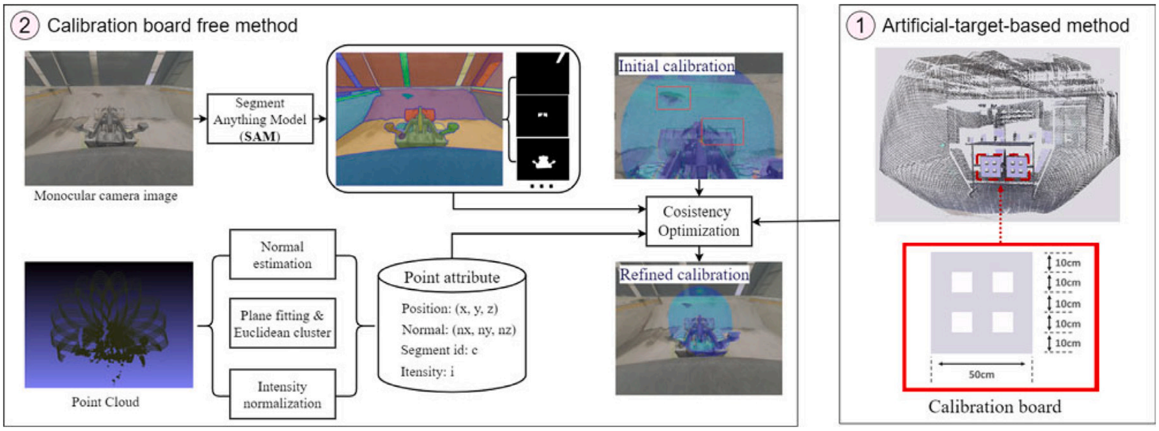


Fig. 4. Framework of the proposed camera and LiDAR extrinsic calibration method. Right: initial calibration with custom board. Left: automatic calibration during operation with SAM and nonlinear optimization.

stochastic gradient descent is employed to solve this problem.

$$\begin{aligned} \min_{l_1,l_2,\alpha_1,\alpha_2} \quad & \sum_{i=1}^{n-1}(R^x_{i,i+1}+R^z_{i,i+1}) \\ \text{s.t.} \quad & l_1>0, l_2>0 \\ & -\pi<\alpha_1,\alpha_2<\pi \end{aligned} \tag{4}$$

Note that the calibration method only requires the measurement of the bucket teeth's displacement, without the need to measure the absolute coordinates at each position. Since the objective function in Eq. (4) is convex, the optimization process is insensitive to initial values. Thus, the initial parameters can be arbitrarily chosen within the ranges $l_1>0, l_2>0, -\pi<\alpha_1,\alpha_2<\pi$. The optimization methods above are implemented using the Ceres Solver library (Agarwal et al., 2023) with the Levenberg–Marquardt algorithm. The solver is considered to have converged to a stable point when the infinity norm of the gradient vector at the current parameter point decreases to below or equal to 1e−10. The larger the dataset, the more residual terms can be constructed, and the more accurate the results will be. Typically, to solve for the model parameters, at least four sets of residual terms are needed, i.e., $n \geq 3$. Once the parameters $(l_1, l_2, \alpha_1, \alpha_2)$ are determined, the real-time estimation of the bucket's pose can be achieved by combining the loader's kinematic model with the inclination readings from the IMU sensors.

3.2. Calibration of the perception module

The forward perception module of the automatic shoveling system integrates a monocular camera and LiDAR. To fuse their data for a high-precision semantic elevation map, spatial alignment between the two sensors via extrinsic calibration is essential. This process determines the transformation matrix between the camera and LiDAR coordinate systems. Since intense loader vibrations can cause calibration shifts, and frequent manual recalibration is impractical in production, an autonomous calibration technique is critically needed.

As illustrated in Fig. 4 and Algorithm 1, the calibration method consists of two steps. The first step employs a method based on artificial calibration boards (*ll*.1–4), which is typically conducted prior to the installation of the forward perception module onto the vehicle. As shown in the right image of Fig. 4, each calibration board features four square holes that can be easily detected by LiDAR and image corner extraction algorithms. By leveraging the detected 2D–3D correspondences, high-precision calibration results can be achieved according to the method mentioned in Liu (2023). The second step leverages the environmental automatic segmentation capabilities of the large-scale neural network model SAM and, based on the initial values obtained from the first step, employs nonlinear derivative-free optimization to achieve automatic extrinsic calibration (*ll*.6–27). This step does not require specific calibration boards or manual intervention and allows for the calibration of parameters between the camera and LiDAR at any time during vehicle operation. The specific method is described as follows:

Firstly, for each frame of image information, the SAM model is used to generate masks for the entire image. Each mask is a binary matrix of the same size as the image, where the value $M_i(u,v) \in \{0,1\}$ indicates whether the pixel (u,v) belongs to the ith mask. All pixels within a mask can be denoted as $P_i = \{p|M_i(p_u,p_v)=1\}$.

Subsequently, the point cloud data is analyzed to obtain the positional information, normal vectors, reflectivity, and segment class of each point p. The 3D point cloud data is then projected onto the 2D image plane:

$$\lambda \begin{pmatrix} p_u \\ p_v \\ 1 \end{pmatrix} = KT \begin{pmatrix} p_x \\ p_y \\ p_z \\ 1 \end{pmatrix} \tag{5}$$

where K represents the intrinsic parameters of the camera, and T denotes the transformation matrix from the camera coordinate system to the LiDAR coordinate system. The consistency score ϵ of the ith mask is defined as follows:

$$\epsilon_i = (\omega_R F_R(P_i) + \omega_N F_N(P_i) + \omega_S F_S(P_i)) f(\|P_i\|) \tag{6}$$

Here, $F_R(\cdot)$, $F_N(\cdot)$, and $F_S(\cdot)$ represent the consistency scores for reflectivity, normal vectors, and segmentation class, respectively. The reflectivity consistency score is calculated from the variance of the reflectivity values across all points (*l*. 20), the normal vector consistency score is derived from the mean of the pairwise inner products of all normal vectors (*l*. 21), and the segmentation class consistency score is based on the weighted sum of the number of points within each category (*l*. 22). ω_R, ω_N, and ω_S are their respective weights, and $f(\cdot)$ is an adjustment function that accounts for the size of the point set, designed to mitigate consistency loss that may arise from an overabundance of points.

The final optimization goal can be expressed as the weighted sum of the consistency scores of all masks:

$$\rho = \sum_i \omega_i \epsilon_i, \quad \omega_i = \frac{\|P_i\|}{\sum_j \|P_j\|} \tag{7}$$

The optimization problem aimed at maximizing the score ρ, with the extrinsic parameter T as the variable, is solved using the Nelder–Mead algorithm (Luo et al., 2024) , with convergence triggered when the standard deviation of simplex function values drops below 1e−10 and the maximum simplex edge length falls below 1e−8. Given the highly nonlinear and discontinuous nature of the problem, the initial estimate

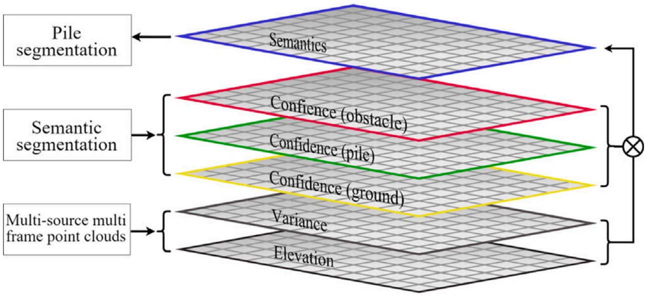


Fig. 5. The semantic elevation map of pile perception contains six layers of information. The bottom two layers represent the estimated height and estimation error obtained from the fusion of multiple frames of LiDAR data. The middle three layers represent the probability confidence of piles, ground, and obstacles obtained from the semantic segmentation model. The top layer represents the category information jointly determined by the fusion of semantic information and elevation information.

from the first-step calibration using a calibration board is used as the starting point for the optimization.

4. Pile perception

Accurate perception of the pile surface is essential for selecting shoveling points and performing scooping actions. To better represent the environment, an elevation map is used to store both category and 3D information per cell, across six layers (see Fig. 5). Multi-frame LiDAR data are first fused to build the elevation map, providing an estimated height and variance per cell. A semantic segmentation model then classifies pixels and projects the results into 3D space, yielding confidence values for each category, namely pile confidence, ground confidence, and obstacle confidence, per cell. These confidences are fused to determine the final semantic classification, as shown in the Semantics layer of Fig. 5. Finally, all cells classified as pile are extracted to generate the pile elevation map.

4.1. Material surface mapping

The elevation grid map uses sampled points to represent the environment and supports multi-frame and multi-sensor fusion. Each cell in this 2.5D map stores height information at its location, offering richer environmental representation than 2D maps with lower computational and memory costs than 3D voxel grids. To build a real-time accurate elevation map, point cloud data are first preprocessed to remove points occluded by the vehicle. The loader's occupied area is abstracted as two rectangles rotating around the steering axis, with positions calculated from the steering encoder's real-time readings. Considering the influence of arm posture changes on the occupied area, the length of the front part is determined using the tip position from Eq. (2).

The elevation map is updated by fusing multiple LiDAR scans. Limited field of view in a single frame (Fig. 6(a)) makes multi-frame fusion with localization essential for a consistent global map. Each grid height is updated via the Kalman filter formula:

$$h^{+} = \frac{\sigma_p^2 h^{-} + \sigma_m^{2-} p_z}{\sigma_m^{2-} + \sigma_p^2}, \quad \sigma_m^{2+} = \frac{\sigma_m^{2-} \sigma_p^2}{\sigma_m^{2-} + \sigma_p^2} \tag{8}$$

Here, h represents the estimated height of the cell, p_z is the height measurement of the point, σ_m^2 is the estimated variance of the cell, which is initially assigned a large value to reflect a high level of uncertainty. σ_p^2 is the variance estimated by the sensor noise model, which is set as $\sigma_p^2 = \alpha_d d^2$, where α_d is a tunable parameter and d is the distance from the point to the sensor. The estimates before updating are denoted with a superscript '−', while the updated data is denoted with a superscript '+'.

Algorithm 1 Automatic Camera–LiDAR Extrinsic Calibration

```
Require: Image $I$, Point cloud $P$, Initial extrinsic $T_0$, SAM model
Ensure: Optimized extrinsic parameters $T^*$
// Step 1: Initial calibration (pre-deployment)
Acquire calibration board data $(I_{board}, P_{board})$
Detect 2D corners in $I_{board}$ and 3D corners in $P_{board}$
Compute $T_0$ using (Liu, 2023)'s method

// Step 2: Online auto-calibration
for each new frame $(I, P)$ do
    // 2.1 Image segmentation via SAM
    $\mathcal{M} \leftarrow \mathrm{SAM}(I)$                          ▷ $\mathcal{M} = \{M_1, \ldots, M_k\}$
    // 2.2 Project all points to image plane
    $P_{\mathrm{proj}} \leftarrow \mathrm{ProjectPoints}(P, T_0, K)$     ▷ Project to 2D coordinates
    // 2.3 Process each mask and compute consistency
    for each mask $M_i \in \mathcal{M}$ do
        // Select points projected inside mask $M_i$
        $P_i \leftarrow \{p \in P \mid \mathrm{proj}(p) \in M_i\}$
        if $|P_i| <$ threshold then
            continue                       ▷ Skip masks with insufficient points
        end if
        // 2.4 Multi-modal consistency evaluation
        $r_i \leftarrow \mathrm{Var}(\{p.r \mid p \in P_i\})$             ▷ Reflectivity consistency
        $n_i \leftarrow \mathrm{Mean}(\{\langle p.n, p'.n\rangle \mid p, p' \in P_i\})$   ▷ Normal consistency
        $s_i \leftarrow \mathrm{Entropy}(\{\mathrm{Class}(p) \mid p \in P_i\})$       ▷ Segmentation consistency
        $\epsilon_i \leftarrow (\omega_R r_i + \omega_N n_i + \omega_S s_i) \cdot f(|P_i|)$
    end for
    // 2.5 Nonlinear derivative-free optimization
    $T^* \leftarrow \arg\max_T \sum_i \frac{|P_i|}{\sum_j |P_j|} \epsilon_i(T)$
    Solve using the Nelder–Mead algorithm (Luo et al., 2024) with $T_0$ as initial guess
end for
```

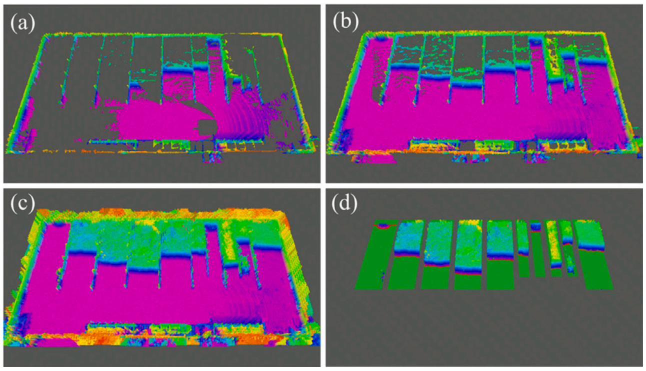


Fig. 6. Process of the material surface mapping. (a) shows the range that a single-frame LiDAR sensor can perceive. (b) displays the perception range after multi-frame data fusion. (c) shows the result of the elevation map after hole filling. (d) presents the final established elevation map.

To clear dynamic obstacles, the system checks if a ray passes through a cell by comparing the ray height with the cell's estimated height. If a point's height p_i^z is below the lower confidence bound $h_i - \sigma_i$ of the elevation map at that point, the cell is marked as penetrated and the obstacle is cleared. In scenarios with large map boundaries and high piles, LiDAR occlusion can cause holes in the elevation map (Fig. 6(b)). Assuming pile continuity, each hole point $p_i(x_i, y_i)$ is assigned the height of its nearest neighbor (as illustrated in Fig. 6(c)):

$$h[x_i, y_i] = h[\arg\min((x - x_i)^2 + (y - y_i)^2)] \tag{9}$$

where $h[x_i, y_i]$ is the height value of the point (x_i, y_i). Each hole-filling operation is independent and can be efficiently parallelized on the GPU.

4.2. Semantic segmentation

To endow loaders with real-time environmental perception capabilities, semantic segmentation models need to balance high performance and lightweight characteristics on edge computing devices. Traditional Convolutional Neural Networks (CNNs) have somewhat limited capabilities in integrating global information, but their structure is particularly suitable for parallel computation, making them advantageous for deployment on ARM platforms. On the other hand, Vision Transformers (ViT) are more conducive to integrating global information and achieving higher accuracy, but they are not suitable for edge computing. To achieve faster inference speeds while ensuring accuracy, we employ the TokenPyramid Vision Transformer (TopFormer (Zhang et al., 2022)) architecture for semantic segmentation, which combines the high performance of CNNs with the high accuracy of transformer structures, demonstrating excellent performance on edge computing devices.

The TopFormer network architecture consists of the Token Pyramid Module, Semantics Extractor, Semantics Injection Module, and Segmentation Head. The Token Pyramid Module, built upon the CNN framework, processes high-resolution images to quickly generate a local feature pyramid. Capitalizing on the lightweight nature and efficiency of CNNs in encoding local image details, the Token Pyramid Module employs stacked lightweight MobileNetV2 (Sandler et al., 2018) blocks and rapid downsampling strategies for fast image processing. The Semantics Extractor module, based on ViT, takes the Token Pyramid as input to generate scale-aware semantics. The transformer-based structure enables explicit modeling of global interactions among pixels, obtaining richer semantics and a larger receptive field. Directly applying global self-attention to high-resolution tokens incurs extremely high computational costs due to the quadratic complexity associated with the number of tokens. By applying an average pooling layer to the tokens produced by the Token Pyramid, the number can be significantly reduced, such as to $1/(64 \times 64)$ of the input size, thus effectively alleviating the high computational complexity while fully utilizing the global information integration capability of ViT. The scale-aware semantics generated by the Semantics Extractor are injected into the corresponding scale tokens through the Semantics Injection Module to enhance representation capability. Finally, the Segmentation Head executes the segmentation task using the enhanced token pyramid.

To establish a dataset for engineering machinery scenarios, the pre-trained open-vocabulary model Grounded-SAM (Ren et al., 2024) is utilized to generate 2D semantic segmentation labels. This model enables the acquisition of 2D labels that closely match the semantics of the given class names, facilitating the rapid construction of a real dataset. Specifically, Grounding DINO (Liu et al., 2025) is provided with the class names that may appear in the production scene (such as piles, ground, pedestrians, loaders, etc.) to generate detection bounding boxes along with corresponding logits and phrases. This information is then passed to SAM to generate precise segmentation binary masks, with the masks and logits jointly determining the label for each pixel. This process achieves automated dataset construction and annotation, ultimately training a semantic segmentation model that is both lightweight and high-accuracy, well-suited for engineering machinery scenarios.

4.3. Semantics–elevation fusion

As shown in Fig. 7, segmenting the pile surface using only height information can lead to significant errors, due to inaccuracies in manually set ground height thresholds and height perception. Fig. 7(a) shows an irregular thin material pile during shoveling, and Fig. 7(c) demonstrates the result of height-based segmentation. Because of uneven ground perception and difficulty in determining accurate height ranges, the elevation map fails to include the thin pile in front of the loader. In comparison, the semantic segmentation model accurately identifies this area (Fig. 7(b)), enabling complete surface reconstruction including the thin pile (Fig. 7(d)). Thus, integrating semantic information is essential for accurate pile surface segmentation.

The semantic segmentation model infers per-point semantic labels in pixel space, which are projected into 3D space via Eq. (5) to obtain the semantic category confidence for each cell $(\hat{p}_1, \hat{p}_2, \hat{p}_3)$, where $\hat{p}_1$, $\hat{p}_2$ and $\hat{p}_3$ represent the probability that the cell is part of the ground, material, and other objects, respectively. We now provide a detailed explanation of how to achieve accurate pile surface segmentation by fusing semantic segmentation confidence with elevation map data.

First, the method for calculating the probability of a location being a material pile or the ground based solely on the elevation and variance layers of the elevation map is considered. The height information of points in the elevation map is modeled as $z \sim \mathcal{N}(h, \sigma_m^2)$. All points in the scene where $z \leq z_0$ are extracted as ground for plane fitting, where z_0 is an empirically determined fixed value. The RANSAC method (Fischler and Bolles, 1981) is employed to filter out outliers and determine the plane equation of the ground $z' = z'(x, y)$ and the fitting variance σ_z^2. Consequently, for each grid in the elevation map, the height category confidence that it is part of the ground is considered p_1, while the height category confidence that it is the pile is $p_2 = 1 - p_1$, where:

$$p_1 = P(z < z' + \sigma_z) = \int_{-\infty}^{z'+\sigma_z} \frac{1}{\sqrt{2\pi}\sigma_m} e^{-\frac{(x-h)^2}{2\sigma_m^2}} \, \mathrm{d}x \tag{10}$$

Then, the semantic information and height information are integrated to assign a category for each cell in the semantic elevation map. If the semantic category is naively adopted as the definitive category, poor environmental lighting may affect the accuracy of the semantic segmentation results, leading to errors in the modeling of the semantic elevation map (as shown in Fig. 8(1)). Here, inspired by soft voting in ensemble learning, the semantic category confidence and the height category confidence are summed, and the category with the highest total confidence is selected. This approach helps to reduce the impact of environmental lighting on the model, enabling accurate modeling of the material surface (as illustrated in Fig. 8(2)). For safety considerations, if the semantic segmentation identifies the highest probability of a certain area being an obstacle category, then that area is directly considered as an obstacle. The final classification result can be expressed as:

$$C[x_i, y_i] = \begin{cases} \text{Obstacle}, & \text{if } \hat{p}_3 > \max(\hat{p}_1, \hat{p}_2), \\ \text{Pile}, & \text{else if } p_1 + \hat{p}_1 < p_2 + \hat{p}_2, \\ \text{Ground}, & \text{otherwise}. \end{cases} \tag{11}$$

By integrating the class information for each cell with its corresponding height data, an accurate semantic elevation map is constructed. From this map, all cells classified as ‘pile’ are extracted to generate a dedicated pile elevation map.

5. Shoveling point selection

The position of the shoveling point refers to the two-dimensional coordinates (x_i, y_i) and the entry angle δ_m at which the bucket first contacts the pile surface during the shoveling process. To ensure the safety and efficiency of the shoveling process, the feasibility of each boundary point is evaluated based on the pile elevation map M, considering the risks of the loader overturning, colliding, or getting damaged, as well as operational efficiency to select the optimal shoveling point. As shown in Algorithm 2, given the loader’s initial position p_0 and the shoveling point (x_i, y_i, δ_m), the loader needs to plan a path $\Gamma_{i,m}$ between these two points, travel along this path to reach the point (x_i, y_i) with an entry angle of δ_m. This process is simulated, and scores for each indicator are calculated, which are then weighted to obtain the overall score for this point. Finally, the shoveling point and orientation with the highest score are selected.

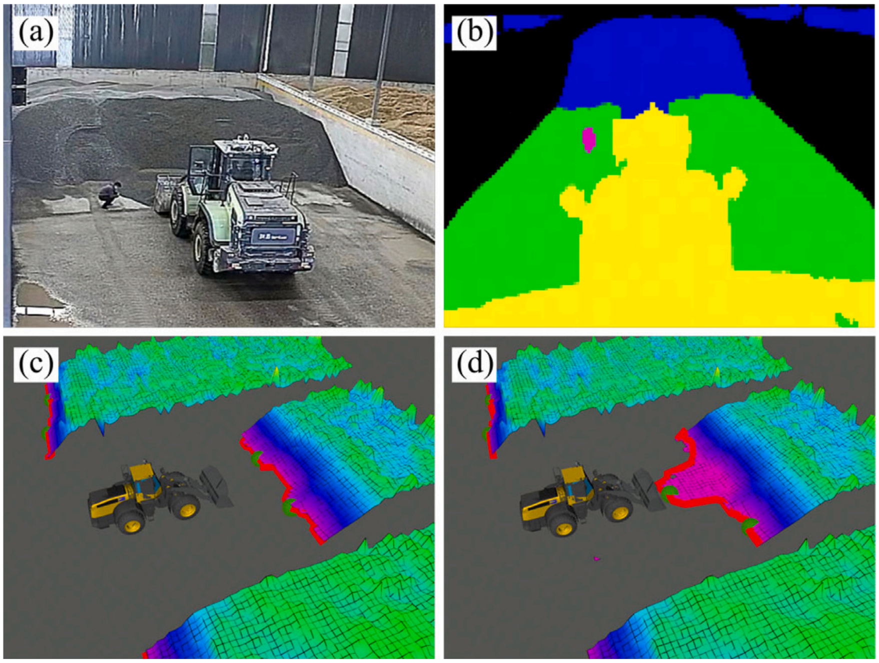


Fig. 7. (a) shows the real production scene of the thin material pile in front of the loader. (b) displays the result of semantic segmentation in pixel space of the loader's front view. (c) represents the result of pile surface segmentation based solely on height thresholds. (d) is the result of pile surface segmentation after fusing height confidence and semantic confidence, indicating that incorporating semantic information is essential for material surface segmentation.

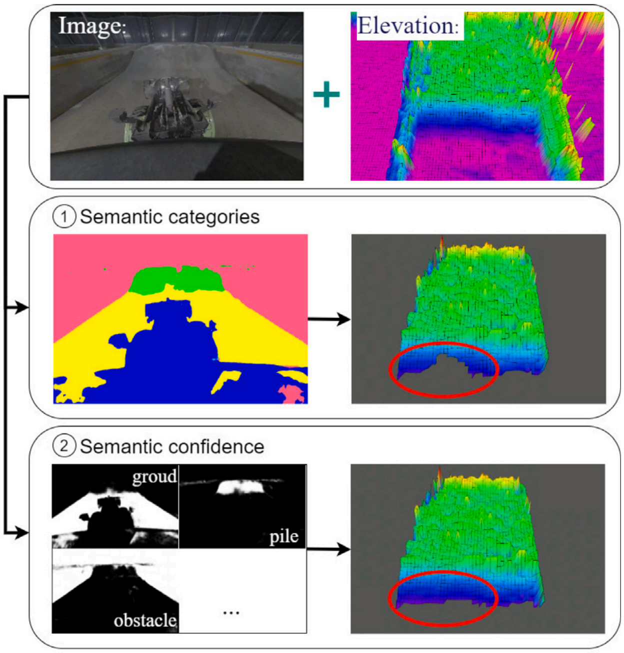


Fig. 8. Comparison of two fusion methods in dim lighting conditions. Method (1) directly uses semantic categories for classification, resulting in noticeable defects in the segmented pile area highlighted in the red box. Method (2) fuses semantic confidence with elevation confidence, significantly mitigating misclassification issues.

Algorithm 2 Shoveling Point Selection

```
Input:
Pile elevation map $M$
Loader initial state $p_0$
Set of candidate shoveling points $P = \{(x_i, y_i)\}$
Set of candidate entry angles $\Delta = \{\delta_m\}$
Safety thresholds: $T_{\text{res}}$, $d_{\text{res}}$, $\kappa_{\max}$
Weight coefficients: $k_t$, $k_s$, $k_c$, $k_e$, $k_a$
Output:
Optimal shoveling point $(x^*, y^*)$ with entry angle $\delta^*$
Initialize:
best_score ← $-\infty$
best_point ← None
best_angle ← None
for each candidate shoveling point $(x_i, y_i) \in P$ do
    for each candidate entry angle $\delta_m \in \Delta$ do
        $\Gamma_{i,m} \leftarrow \text{PlanPath}(p_0, (x_i, y_i), \delta_m)$
        $s_t \leftarrow \text{CalculateTorqueScore}(M, x_i, y_i, \delta_m, T_{\text{res}})$
        $s_s \leftarrow \text{CalculateSafetyScore}(\Gamma_{i,m}, M, d_{\text{res}})$
        $s_c \leftarrow \text{CalculateConcavityScore}(M, x_i, y_i, \delta_m)$
        $s_e \leftarrow \text{CalculateEfficiencyScore}(\Gamma_{i,m}, \Gamma)$
        $s_a \leftarrow \text{CalculateAngleScore}(\Gamma_{i,m}, \kappa_{\max})$
        $F \leftarrow k_t \cdot s_t + k_s \cdot s_s + k_c \cdot s_c + k_e \cdot s_e + k_a \cdot s_a$
        if $F >$ best_score then
            best_score ← $F$
            best_point ← $(x_i, y_i)$
            best_angle ← $\delta_m$
        end if
    end for
end for
return best_point, best_angle
```

5.1. Torque balance score s_t

When the material pile is unevenly shaped, unequal torques act on the bucket ends. An excessive torque difference can damage the steering mechanism or even cause the loader to overturn. Hence, the overturning torque is estimated using a simplified model:

$$T_c = \sum_{i,j \in \Omega} h_{ij} \cdot \lambda_i, \tag{12}$$

where $h_{i,j}$ represents the height of that point in the semantic elevation map, λ_i denotes the lateral distance from that point to the center of mass of the bucket (with left being positive), and Ω indicates the set of points on the pile covered by the bucket. Accordingly, the torque balance score is calculated as:

$$s_t = \begin{cases} 1 - \left(\frac{T_c}{T_{res}}\right)^2, & \text{if } T_c < T_{res}, \\ -\infty, & \text{otherwise,} \end{cases} \tag{13}$$

where T_{res} represents the set safety threshold. If the torque exceeds this threshold, there will be a significant risk of overturning, making that loading point not recommended. As the torque approaches the threshold within a certain proximity, the score drops sharply; conversely, when the torques on both sides are comparatively balanced, the score approaches one.

5.2. Safety distance score s_s

The probability of collision during the shoveling process is generally related to the distance from obstacles. During attacking phase, a certain safety distance from obstacles d_{res} should be maintained. Let the area covered by the loader during the process be Λ, and the area identified by semantic information as containing obstacles be P. The minimum distance d_{min} between the two areas is evaluated. The safety distance score is calculated as:

$$s_s = \begin{cases} 1, & \text{if } d_{min} > d_{res}, \\ \frac{d_{min}}{d_{res}}, & \text{else if } d_{min} > 0, \\ -\infty, & \text{otherwise.} \end{cases} \tag{14}$$

5.3. Concavity score s_c

The concavity of the pile has a significant impact on the forces experienced by the loader. Related studies suggest that to improve operational efficiency, the loader should vertically penetrate the pile and shovel at raised positions, as concave piles introduce greater bucket resistance and adversely affect bucket fill rates. To estimate the concavity at the point, this method employs a three-section bucket model. The future shoveling area is uniformly divided into three sections according to the bucket width, with volumes denoted as V_l, V_m, and V_r for the left, middle, and right sections, respectively. A shoveling point is classified as convex if the volume in the middle section surpasses the maximum volume of the two side sections; otherwise, it is deemed concave. The concavity score is calculated using the formula:

$$s_c = \begin{cases} 1, & \text{if } V_m \geq \max\{V_l, V_r\}, \\ 0, & \text{otherwise.} \end{cases} \tag{15}$$

5.4. Efficiency score s_e

In ensuring safety, loaders generally tend to choose points that are closer in order to enhance operational efficiency. The trajectory length corresponding to the current point is denoted as $l_{i,m}$. For all trajectories $\boldsymbol{\Gamma}$ corresponding to all candidate shoveling points on the same pile, the length of the longest trajectory is denoted as l_{max}, and the length of the shortest trajectory is denoted as l_{min}. The efficiency score is calculated using the formula:

$$s_e = \frac{l_{max} - l_{i,m}}{l_{max} - l_{min}}. \tag{16}$$

5.5. Entry angle score s_a

When shoveling materials, the angle between the front and rear axles of the loader should not be excessively large; otherwise, the lateral pressure on the hinge during entry into the pile increases, raising the risk of structural damage or even tipping over. The angle between the front and rear axles is related to the curvature at the end of the trajectory, denoted as $\kappa_{i,m}$. Therefore, during production operations, the loader should avoid exceeding the permissible curvature κ_{max} while minimizing the trajectory end curvature to ensure safety. The entry angle curvature score is calculated using the formula:

$$s_a = 1 - \frac{|\kappa_{i,m}|}{\kappa_{max}}. \tag{17}$$

Finally, the comprehensive score for the shoveling point is obtained by weighting all indicators:

$$F = k_t \cdot s_t + k_s \cdot s_s + k_c \cdot s_c + k_e \cdot s_e + k_a \cdot s_a, \tag{18}$$

where k_t, k_s, k_c, k_e, k_a corresponds to the weight of the respective score indicator. This calculation method effectively helps to avoid excessive overturning moments, collisions, and other situations that could damage the vehicle and threaten environmental safety. To determine these parameters, we collected elevation maps of material surfaces along with the positions and orientations of both the vehicle's starting point outside the bin and the final selected scooping points during routine operations by expert operators. This process resulted in a human-optimized dataset comprising 482 distinct scenarios. We then employed the Covariance Matrix Adaptation Evolution Strategy (CMA-ES) algorithm to optimize the weighting parameters, with the objective of minimizing the discrepancy between the algorithm-selected scooping points and those chosen by the human operators across all scenarios.

6. Shoveling control strategy

The process of shoveling can essentially be divided into three phases: the attacking phase before contacting the pile, the crowding phase while contacting and continuing to advance, and the scooping phase during which the shoveling action is executed. The goal of shoveling control is to select an appropriate triggering method that enables the automatic loader to transition from the crowding phase to the scooping phase, and to execute the appropriate shoveling strategy during the scooping phase. In this work, the timely triggering of the scooping action is achieved by estimating accurate shoveling volume through perception and predictive algorithms, and an adaptive scooping strategy selection algorithm is proposed based on the two improved scooping strategies.

6.1. Shoveling volume estimation

The volume of material being shoveled refers to the combined volume of material handled during the crowding and scooping phases of a single loading task, which includes the current shoveled volume V_1 and the future shoveling volume V_2 (as shown in Fig. 9).

Due to the inherent delays in signal reception and response of the automation program, as well as the response of the hydraulic actuators to PWM commands, the loader scoops a portion of material, denoted as V_{21}, from the moment the scooping command is issued until the actual initiation of the scooping action, as indicated by the 'Delay' label in Fig. 9. Subsequently, as the loader's arm moves to execute the scooping action, it scoops an additional portion of material, denoted as V_{22}, as shown in the 'Control' segment of Fig. 9. Consequently, the total volume of material that will be scooped in the future is the sum of these two volumes, expressed as $V_2 = V_{21} + V_{22}$.

The current real-time volume of material being shoveled can be calculated based on the pile elevation map:

$$V_1 = \sum_{i \in \Omega} (h_i - h'_i) \cdot \varepsilon^2, \tag{19}$$

where h_i represents the estimated height corresponding to the ith cell in the semantic elevation map, Ω denotes the region of the elevation

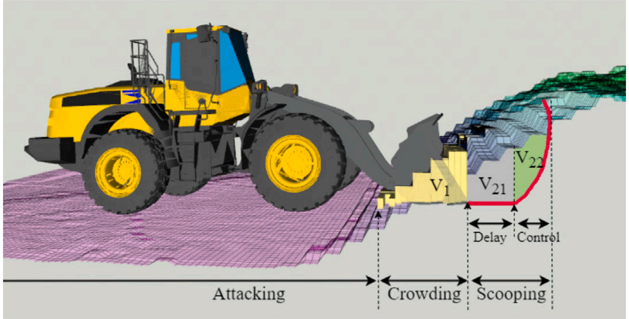


Fig. 9. The shoveling process of a loader consists of three stages: Attacking, Crowding, and Scooping. The total volume of material shoveled includes the current shoveled volume V_1 and the future shoveling volume V_2, which is further divided into the delayed portion V_{21} and the scooped volume V_{22}.

map traversed by the bucket teeth during the shoveling process, h'_i is the height of the bucket teeth when it passes over the ith cell, and ε is the resolution of the semantic elevation map. The calculation of the volume covered by the bucket begins upon contact with the material pile, and this value incrementally increases as the shoveling operation progresses.

The future shoveling volume is equivalent to the amount of material that can still be shoveled by the bucket if it enters the scooping phase at this moment, and can be predicted by estimating the trajectory of the bucket, which essentially involves penetration depth prediction. The speed of the loader during the scooping phase is a time-varying function. Due to the unpredictable resistance encountered during this process, influenced by multiple factors such as material type and pile shape, it is challenging to model the true physical dynamics of the loader's movement during shoveling. Therefore, a data-driven approach is adopted, using Support Vector Regression (SVR) to forecast the penetration depth. This model estimates the nonlinear relationship between several input variables that may affect the future shoveled volume and the penetration depth. The influencing factors mainly include the current speed υ, the instantaneous shoveled volume V_1, the material type M, and pile geometry (in this case, six points are taken along the depth direction with a horizontal spacing of $\Delta l = 0.5$ m on the pile surface in the loader's field of view, denoted as $\mathbf{h} = (h_1, h_2, \ldots, h_6)$). The six-point elevation sampling strategically balances computational efficiency with spatial resolution, capturing critical inflection points in pile topography that dictate force distribution during penetration. Such physically-grounded feature engineering ensures the SVR model effectively predicts the penetration depth in real-time during each shoveling process. The loader's position during shoveling can then be calculated based on the penetration depth, and combined with Eq. (2) to calculate the motion trajectory of the bucket teeth. Finally, using Eq. (19), the future shoveling volume can be calculated. By adding the current shoveled volume to the future shoveling volume, the volume of material that the loader's bucket can obtain in a single operation can be determined. Controlling the loader based on the total volume allows for real-time and precise triggering of scooping actions.

6.2. Scooping operations

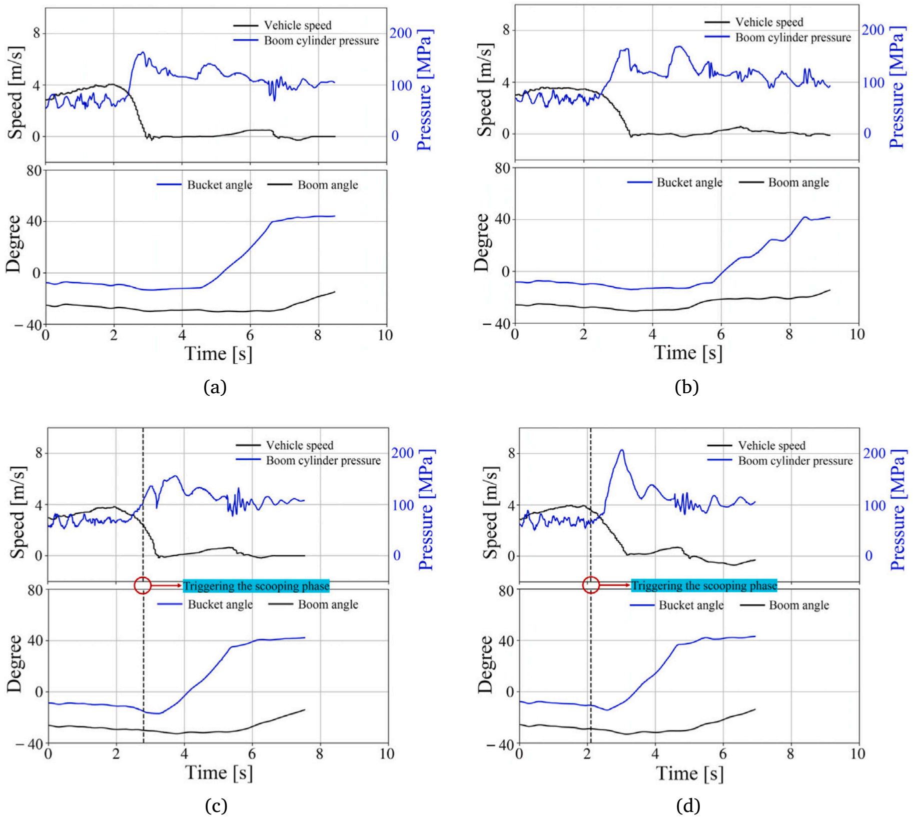
The "Just in & out" strategy involves rotating and retracting the bucket to a certain height, followed by lifting the boom to a specific angle to complete the scooping. In contrast, the "Stairway" strategy entails slightly lifting the boom, then retracting the bucket by a small angle, pausing for a period, and repeating this process multiple times. The performance of the loader using these two strategies during the scooping phase is shown in Figs. 10(a) and 10(b). Retracting the bucket to a designated angle and lifting the boom to a designated angle can be considered as two distinct atomic actions. The "Just in & out" strategy requires the execution of one set each of retracting the bucket and lifting the boom atomic actions, which is suitable for materials with low resistance, thus offering high efficiency and low energy consumption. Conversely, the "Stairway" strategy divides the bucket retraction action into multiple sets of atomic actions executed intermittently to cut through the material, which is suitable for materials with high resistance. As the number of atomic actions increases, the shoveling efficiency decreases, and energy consumption rises.

Improvements have been made to both strategies. For the "Just in & out" strategy, an open-loop predictive approach is employed, executing the scooping action directly when the predicted shoveling volume reaches the target value. This method allows for more precise control of the shoveling volume and faster shoveling compared to traditional perception-triggered methods without shoveling volume prediction (as shown in Figs. 10(c) and 10(d)). For the "Stairway" strategy, a feedback loop between atomic actions is incorporated to assess the real-time scooping volume, allowing the scooping to be completed once the volume reaches the target value, rather than setting a fixed number of bucket retraction atomic actions. This approach ensures the prompt and accurate triggering of the "Just in & out" strategy and minimizes the operational time and energy consumption of the "Stairway" strategy. Additionally, it leverages open-loop prediction and closed-loop feedback to precisely control the shoveling volume.

Algorithm 3 Shoveling Strategy Selection

```
Attacking recommended shoveling point
Calculate the current shoveling volume $V_1$
while $V_1 > 0$, Crowding do
    Calculate the current shoveling volume $V_1$
    Retrieve the material type $\gamma$
    Estimate $V_2$ based on penetration depth prediction
    if $V_1 + V_2 \geqslant V_t$ then
        if $\gamma \in \Upsilon$ then
            Scooping using the "Stairway" strategy
        else
            Scooping using the "Just in & out" strategy
        end if
    else if $V_1 > 0$ and $\upsilon \approx 0$ then
        Scooping using the "Stairway" strategy
    end if
end while
```

Additionally, an adaptive shoveling algorithm capable of flexibly employing the two strategies is proposed, as illustrated in Algorithm 3. Initially, during the attacking phase, the loader controls the bucket to stay close to the ground and navigates along the planned trajectory, advancing towards the chosen loading point. As the bucket impacts and penetrates the material pile, the crowding phase commences, with real-time computation of V_1 and V_2. If $V_1 + V_2$ reaches the target loading volume V_t, the scooping phase is initiated. In this phase, a preliminary scooping strategy is determined based on the material type γ: for difficult-to-load materials such as river sand ($\gamma \in \Upsilon$), the "Stairway" strategy is selected; for easy-to-load materials such as crushed gravel, the "Just in & out" strategy is chosen. If $V_1 + V_2$ is less than the target loading volume but the loader cannot proceed further ($V_1 > 0$ and $\upsilon \approx 0$), it directly enters the scooping phase and adopts the "Stairway" strategy, utilizing closed-loop feedback for stable and precise control. It should be noted that the vehicle stopping determination ($\upsilon \approx 0$) employs a lightweight yet robust mechanism based on confidence thresholds and hysteresis logic. In each control cycle, if the vehicle speed is below 0.15 m/s, the confidence level increases by $(0.1 - \upsilon) \times 100$. When the speed exceeds 0.15 m/s, the confidence level is reset to zero. A stopping state is confirmed once the confidence level reaches 100. After completing the scooping action, the bucket's



(a)

(b)

(c)

(d)

Fig. 10. Comparison of different shoveling strategies. (a) pressure-triggered “Just in & out” strategy, (b) pressure-triggered “Stairway” strategy, (c) perception-triggered “Just in & out” strategy without shoveling volume prediction, (d) perception-triggered “Just in & out” strategy with shoveling volume prediction.

upper edge is leveled, and the material is vibrated to prevent overflow during transportation, before the loader begins to retreat and convey the material. Based on this strategy selection algorithm, the system can autonomously choose the optimal scooping strategy according to the material type and vehicle operation conditions, achieving timely triggering of the scooping phase through open-loop prediction and timely completion of scooping actions through closed-loop feedback, thereby balancing robustness, efficiency, and energy consumption.

Table 1
Parameter list.

Parameter name	Symbol	Value
Reflectivity weight	ω_R	0.20
Normal vectors weight	ω_N	0.30
Segmentation weight	ω_S	0.50
Noise parameter	α_d	0.05
Resolution of elevation map	ε	0.30
Torque balance weight	k_t	0.29
Safety distance weight	k_s	0.23
Concavity weight	k_c	0.13
Efficiency weight	k_e	0.25
Entry angle weight	k_a	0.10

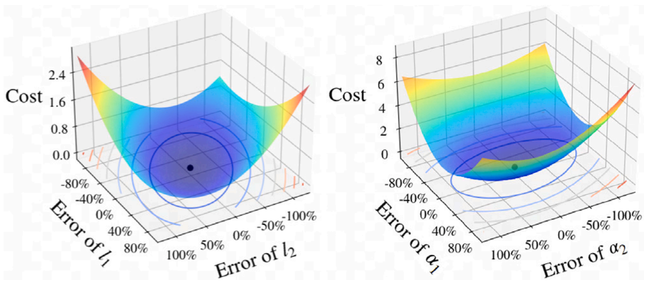


Fig. 11. The surface plot of the calibration cost function as it varies with the errors in the calibration parameters.

The underlying control architecture implements a multi-layered approach for precise material handling operations. For manipulator joint regulation, a PID-based position controller with solenoid valve dead-zone compensation governs the boom and bucket articulation, utilizing pre-calibrated target angles optimized for distinct shoveling phases. The propulsion system employs an acceleration-oriented PID velocity controller, derived from comprehensive throttle-brake acceleration profiling, which maintains constant speed regulation during attacking and crowding phases before transitioning to fixed acceleration control upon scooping initiation. Wheel slip mitigation is achieved through predictive material volume estimation rather than direct slip detection, with real-time motion state prediction algorithms preemptively compensating for potential traction loss through timely bucket actuation that reduces material accumulation resistance ahead of the implement. This integrated approach ensures coordinated electromechanical response while eliminating reliance on explicit slip monitoring subsystems.

7. Experiments and results

To validate the performance and applicability of the proposed system, field production experiments were conducted over two months using two different models of loaders (SDLG L955HE and XCMG 968EV) at three real mixing station scenarios in Shanghai, Hangzhou, and Lanzhou, China (referred to as Mixing Station A, B, and C, respectively). The experimental equipment primarily included a computing platform (NVIDIA Jetson Orin with a 12-core ARM Cortex-A78 CPU, 64 GB RAM,

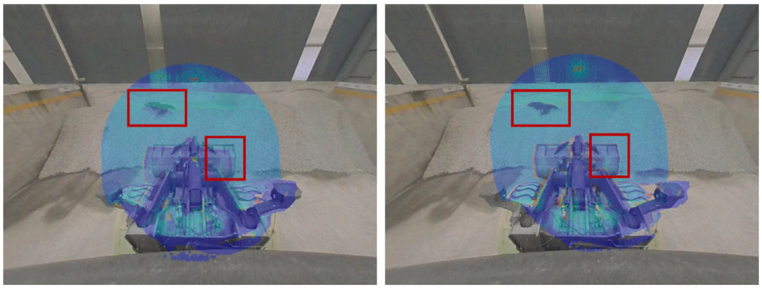


Fig. 12. Camera–laser projection results before and after the automatic calibration.

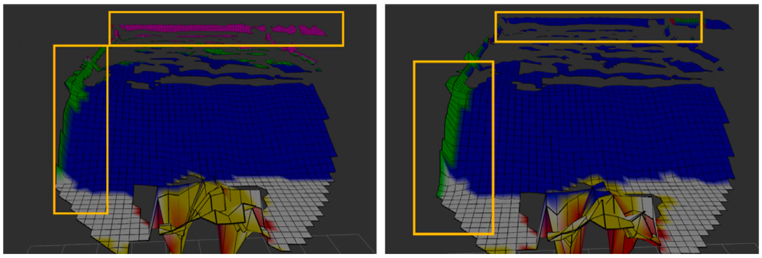


Fig. 13. Fusion results of the semantic elevation maps before and after the automatic calibration.

and NVIDIA Ampere 2048-CUDA GPU), a SENSING monocular camera (model SG3S-ISX031C-GMSL2F, resolution 1920H × 1536V, frame rate 30 fps), a VanJee WLR-720 16-line laser, and a DJI Livox Avia front perception laser. The entire autonomous shoveling system is deployed on an edge computing device (Jetson Orin). The underlying trajectory control, travel velocity control, and joint position control operate at a frequency of 100 Hz. The semantic segmentation model and elevation map update are deployed on the GPU, with the semantic elevation map and shoveling point selection updated at a frequency of 7 Hz. Shoveling action execution and strategy selection are implemented via a state machine with an output frequency of 20 Hz. The parameters used and their corresponding values are shown in Table 1.

7.1. Calibration results

The kinematic calibration data collection for XCMG 968EV loader's arm utilized a low-cost laser rangefinder. Seven sets of end-effector displacement data with corresponding IMU sensor readings were recorded across different arm postures. Fig. 11 demonstrates the variation of calibration cost function surfaces with parameter errors $(l_1, l_2, \alpha_1, \alpha_2)$, revealing a distinct global minimum that facilitates straightforward optimization. Calibration results showed a boom length of 3.13 m, deviating merely 0.02 m from the official 3.11 m specification. Similarly, the bucket length measurement yielded 1.263 m, differing by 0.032 m from the documented 1.295 m. Both discrepancies remained within 1% tolerance thresholds. These experimental outcomes validate the practical precision of the implemented arm kinematic calibration methodology.

An extrinsic calibration experiment was subsequently performed on the forward perception module containing a monocular camera and solid-state laser sensor. As depicted in the left panel of Fig. 12, after a period of operation, there was a deviation in the projection relationship between the camera and the laser. Application of the SAM-based automatic calibration methodology yielded precise extrinsic parameters, evidenced by the corrected projection state marked with a red border in the corresponding right panel. Comparative analysis of semantic elevation map fusion outcomes, shown in Fig. 13, demonstrates performance improvements through automatic calibration. The left panel exhibits erroneous fusion patterns involving construction elements like material piles and walls, while the right panel displays accurately aligned semantic fusion after calibration. These visual comparisons confirm that the automatic calibration algorithm effectively enhances semantic fusion precision through improved calibration accuracy. Notably, the implemented solution eliminates dependency on specialized calibration artifacts or manual adjustments during operation, rendering it particularly advantageous for vibration-prone construction equipment requiring frequent recalibration.

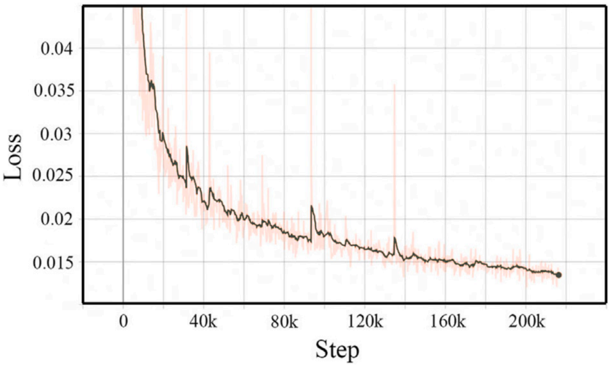


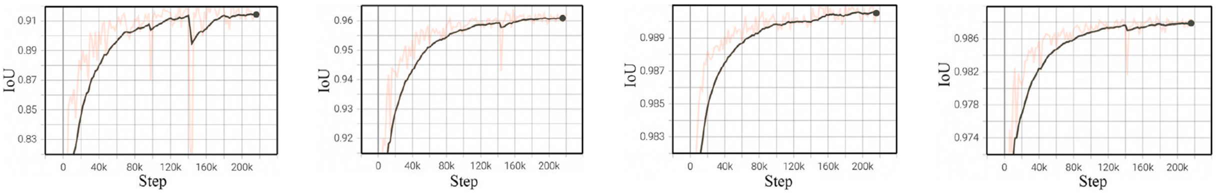
Fig. 14. Loss function variation during the semantic segmentation model training.

7.2. Material surface perception results

In this section, the results of semantic segmentation and semantic-height fusion are introduced, showcasing the effectiveness of the proposed techniques in accurately perceiving and interpreting the material surface within the context of automated loading operations.

7.2.1. Semantic segmentation

A dataset of 3800 images capturing diverse environmental conditions and observational perspectives from typical scenarios at mixing stations was collected. Automatic segmentation of ground surfaces, material piles, loaders, and other obstructions was performed using the Segment Anything Model (SAM). Following manual verification of the automated segmentation outputs and categorical annotation, the



(a) Pile. (b) Ground. (c) Loader. (d) Other.

Fig. 15. Variation of the IoU metrics for different categories during the training of the semantic segmentation model.

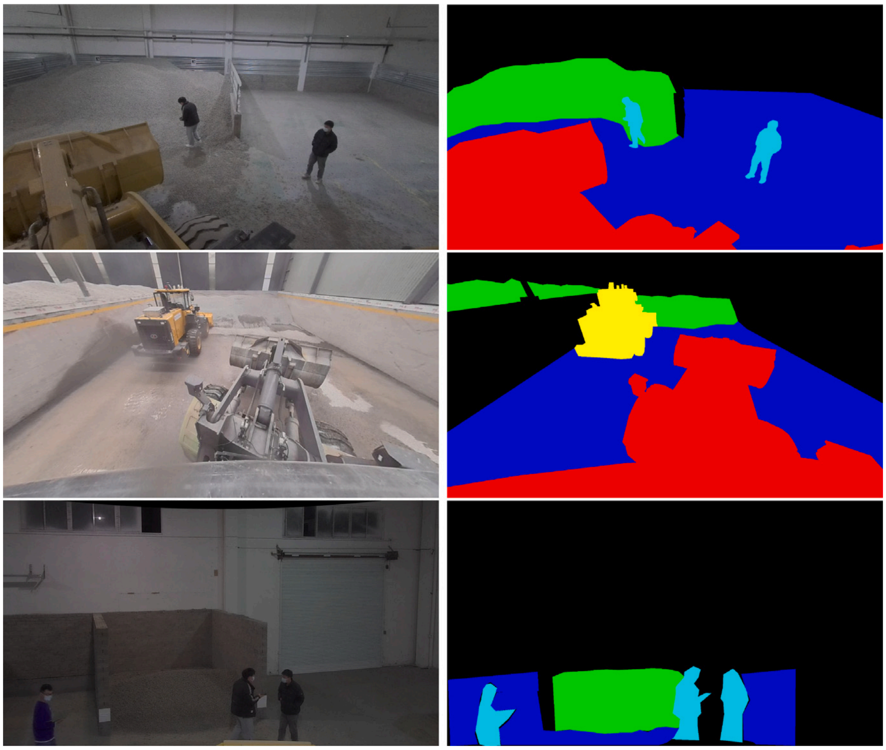


Fig. 16. Results of the semantic segmentation field tests, where different categories are marked in different colors (Red: unmanned loader, Yellow: other loaders, Blue: passable area, Green: material pile, Black: other areas).

dataset was randomly partitioned into 3500 training samples and 300 validation samples. Notably, the SAM framework significantly reduced manual annotation efforts while maintaining cross-frame segmentation consistency. To enhance model robustness and generalization, training images underwent spatial transformations including random scaling, non-uniform cropping, aspect ratio distortion, and horizontal flipping, with final resampling to 480 × 383 px resolution.

The training protocol was implemented on an NVIDIA A30 GPU-accelerated workstation using the Adam optimizer (learning rate = 0.0003, batch size = 16). The model hyperparameters for Semantic Segmentation TopFormer and SAM are shown in Tables 2 and 3, respectively. As depicted in Figs. 14 and 15, model convergence occurred after 200k iterations (3-hour duration), achieving a final training loss of 0.014 and mean intersection-over-union (mIoU) of 97.32%, with per-class IoU metrics surpassing 90%. Experimental validation across varying illumination conditions (optimal, high-intensity, and low-light environments) demonstrated the model's environmental robustness, as illustrated in Fig. 16 showing input–output pairs from test scenarios.

The implemented architecture exhibits real-time performance with 6 ms inference latency on Jetson Orin edge devices, characterized by 1.4M parameters and 0.5G FLOPs computational complexity. Quantitative evaluations confirm the model's capacity to maintain segmentation precision under photometric interference, accurately delineating ground features, material piles, pedestrians, loaders, and environmental obstacles. Comparative analysis reveals computational efficiency advantages over conventional segmentation networks while maintaining competitive accuracy metrics, rendering it particularly suitable for construction machinery applications requiring resource-constrained deployment.

Table 2
Semantic segmentation TopFormer model hyperparameters.

backbone		decode_head	
num_heads	4	num_classes	7
in_channels	[16, 32, 64, 96]	in_channels	[128, 128, 128]
out_channels	[None, 128, 128, 128]	out_channels	128
depths	4	dropout_ratio	0.1
c2t_stride	2	loss_type	CrossEntropyLoss

7.2.2. Perception fusion

To verify the accuracy of the perception fusion, a moment's material surface segmentation result was randomly selected from Mixing Station B. As shown in Fig. 17, the mixing station comprises eight material bins, each containing material piles of varying capacities and shapes. Other loaders are parked in bins 5 and 8, while bins 6 and

Table 3
Semantic segmentation SAM model hyperparameters.

image_encoder		prompt_encoder		mask_decoder	
encoder_depth	32	image_embedding_size	(64, 64)	num_multimask_outputs	3
embed_dim	1280	embed_dim	256	iou_head_depth	3
num_heads	16	mask_in_chans	16	iou_head_hidden_dim	256

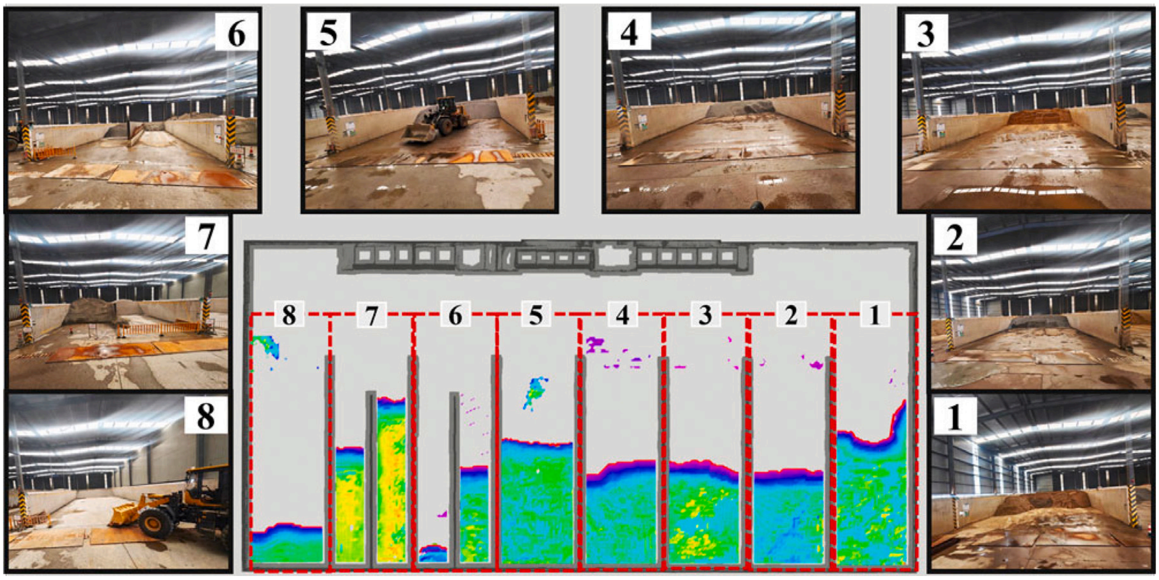


Fig. 17. Perception results of the material surface at the Mixing Station B. Our proposed perception fusion algorithm effectively filtered out interference from various objects, such as parked loaders (e.g., bins 5 and 8) and fences(bins 6 and 7), while also achieving precise perception of material surfaces with different shapes(bins 1–4).

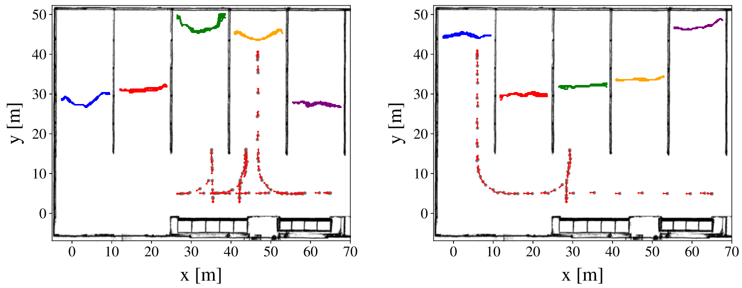


Fig. 18. The perception consistency results of the pile segmentation. Arrows indicate the position and direction of the unmanned loader.

7 are randomly cluttered with obstacles, including fences and other debris. The algorithm was able to detect these anomalies and accurately segment the material surface lines, demonstrating the accuracy and reliability of the method for material surface segmentation based on the fusion of semantic and elevation information. The consistency of the material surface segmentation results under different perception perspectives and distances was also validated. Processes of unmanned loaders performing loading operations in different bins within the mixing station were randomly selected to simulate variations in perception distance and perspective of the sensor towards the material piles. The segmentation results of the material surface lines were recorded as the loader was in different positions. As shown in Fig. 18, the greater the coincidence of the material surface lines in each bin, the more similar the positions of the material surface lines perceived by the loader from different locations, indicating more stable perception results. It can be observed that the contours and positions of the material surface lines perceived by the loader at different locations and moments during the operation in the mixing station are similar. This similarity in the material surface perception results indirectly illustrates the perception accuracy of the proposed algorithm.

7.3. Shoveling points selection

We selected several typical regular and irregular pile surface profiles from three mixing stations and conducted comparative tests with a state-of-the-art shovel point selection algorithm (Chen et al., 2024). As illustrated in Fig. 19, the top 20 scoring shovel points and the optimal shovel point for each surface type are identified and marked with white and red arrows, respectively. For the relatively flat pile in Fig. 19(a), the shovel points are evenly distributed along the surface, and the optimal points recommended by both methods are generally similar. However, for the second irregular surface in Fig. 19(a), Chen's method selects an optimal point closer to the lower region. Comparing the remaining surfaces in Fig. 19(a) and (b), we observe that the primary distinction between the two methods is that our approach tends to recommend safer shovel points. For instance, in Fig. 19(c), our optimal point is farther away from the parked loader and the wall.

To quantitatively evaluate the performance, we performed actual shoveling operations on three surfaces with significant selection differences: the second surface (Scenario a) in Fig. 19(a), the second (Scenario b) in Fig. 19(b), and the third (Scenario c) in Fig. 19(c).

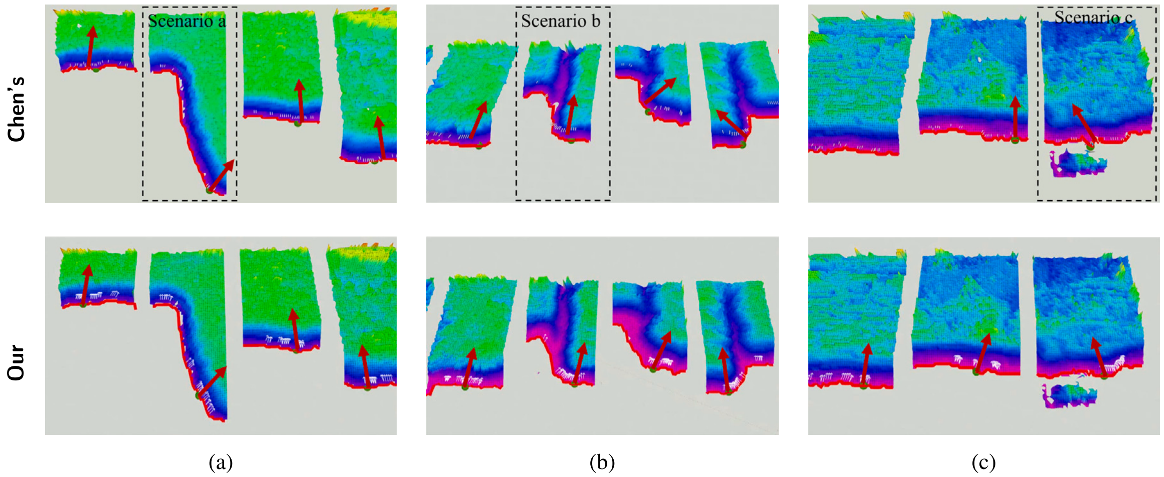


Fig. 19. Performance comparison of shoveling point selection methods across various pile surfaces. Each material surface features 20 candidate shoveling points marked with white arrows, and the most suitable shoveling point is indicated with a red arrow. The three material surface scenarios (Scenario a, b, c) used for the bucket fill rate comparison are labeled.

Table 4
Comparison of bucket fill rates for different shoveling point selection methods.

Method	Scenario a	Scenario b	Scenario c
Chen's approach (Chen et al., 2024)	80.37%	**97.46%**	0%
Our approach	**93.88%**	96.93%	**95.98%**

Using a perception-triggered "Stairway" shoveling strategy uniformly, we conducted five repeated trials for each scenario. As summarized in Table 4, our method achieved a higher bucket fill rate in Scenario a. In Scenario b, both methods yielded comparable bucket fill rates, but the shoveling point of Chen's method required a longer navigation time due to its more distant location. In Scenario c, the shovel point selected by Chen's method posed a collision risk behind the parked loader, making it unsuitable for practical shoveling. The results demonstrate that the scoring metrics and weights incorporated in our method successfully account for a wide range of scenarios encountered in actual production. This enables the loader to select reasonable shoveling points when faced with different material surface shapes, ensuring the loader's own safety while also reducing safety threats to the environment.

7.4. Shoveling control strategy

In this section, the results of shoveling volume prediction and shoveling strategy selection are introduced, followed by a quantitative comparative analysis of different shoveling control strategies.

7.4.1. Shoveling volume estimation

To establish a model between travel distance and various factor variables, 991 instances of real shoveling data were collected over a 60-day period at Mixing Station C. The training and validation set ratio was set to 8:2. The performance of the model trained on the training set is shown in Fig. 20. The left subplot displays the ground truth and predicted values for all test samples, where the solid blue curve represents the actual travel distance and the blue points represent the predictions made by the model. The right subplot shows the error distribution across different material types. The model's absolute error predictions on the test set are within 0.1 m for 75.2% of the points, within 0.15 m for 94.3% of the points, and within 0.2 m for 99% of the points, with a mean squared error of 0.0843. The experimental results indicate that our model can fit the travel distance well within a certain error range.

To test the accuracy of the shoveling volume predictions, field tests were conducted by comparing the predicted bucket path with the actual path of the loader. The distance the loader will continue to travel and the expected bucket teeth trajectory at a certain moment were predicted, and the loader was enabled to enter the scooping phase. Fig. 21 illustrates four intermediate states during the loader's scooping process along with the actual arm postures, with the red line representing the predicted trajectory of the bucket teeth. Stage (a)–(b) indicates the vehicle's travel due to hydraulic delay, stage (b)–(c) represents the process of the loader retracting the bucket, and stage (c)–(d) shows the loader raising the boom and exiting. The highlighted yellow square area in Figure (d) indicates the actual volume of material shoveled by the loader. It is evident that the algorithm can accurately predict the loader's future shoveling process and calculate the volume based on the predictions. Meanwhile, we observed that the time delay between triggering the scooping action (current scoop volume: 1.505 m^3, predicted volume: 1.708 m^3) and the actual movement of the arm (current scoop volume: 2.962 m^3) is approximately 0.6 s, which is caused by computational latency and hydraulic control delays. The final scoop volume reached 3.349 m^3. Therefore, the proportion of V21 (1.457 m^3) in V2 (1.844 m^3) amounts to 79%, indicating a segment that deserves significant attention. The proposed distance-based prediction method achieves a final volume prediction accuracy of 92.6%.

7.4.2. Shoveling control comparison

As shown in Table 5, we conducted a quantitative comparative analysis of the phase-specific performance and final bucket fill rates under different scooping strategies. The distinctions among these strategies lie in both the scooping trigger condition (pressure-based vs. perception-based) and the action policy executed during the scooping phase. Here, "Stairway" refers to an approach where the number of "Stairway" motions is dynamically adjusted based on real-time perceived volume, while "auto-selection" denotes the use of rules from Algorithm 3 to choose between the "Just in & out" and "Stairway" strategies. The last row corresponds to operational data collected from an expert wheel loader operator. Each strategy was tested five times with gravel and five times with sand, and all metrics were averaged to mitigate the impact of single-trial variability. The bucket fill rate was quantified via a perception laser system installed above the mixing station yard, with implementation details referenced from Chen et al. (2025). According to the results, the pressure-triggered "Stairway #3" strategy achieved

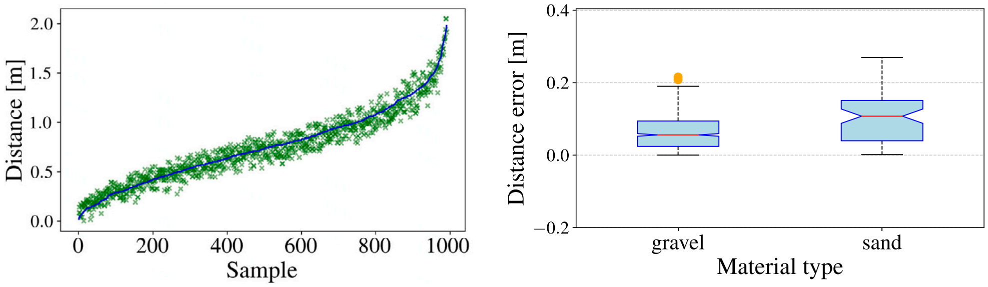


Fig. 20. The performance of the penetration depth prediction on the dataset. (left) The blue curve indicates the true values of the penetration depth, and the green scatter points represent the penetration distances predicted by the model. (right) Error distribution across different material types.

Table 5
The impact of different shoveling strategies on phase-specific metrics and the final bucket fill rate.

Method		Crowding		Scooping		Bucket fill rate
Trigger	Scooping strategy	Time (s)	Energy (kJ)	Time (s)	Energy (kJ)	
Pressure	“Stairway #3”	3.75	57.9	6.34	247.8	98.3%
Perception	“Stairway #3”	0.32	12.7	6.51	242.5	98.8%
Perception	“Stairway”	0.41	11.9	5.78	224.9	97.9%
Perception	“Just in out”	0.36	13.6	4.45	195.4	82.4%
Perception	Auto-selection	0.39	12.8	4.85	213.7	97.8%
Manual	Manual	1.02	30.6	4.89	217.8	98.1%

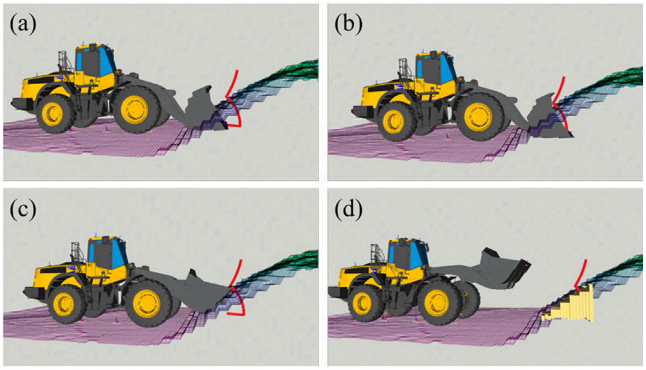


Fig. 21. Comparison between the model-predicted bucket trajectory (red line) and the actual situation, and the yellow areas in the elevation diagram indicate the actual shoveled volume. (a)–(b) represents the process of vehicle movement due to hydraulic delay, (b)–(c) represents the process of the loader retracting the bucket, and (c)–(d) represents the process of the loader raising the boom and exiting.

a relatively high fill rate but consumed the most time and energy. In contrast, our proposed perception-triggered method significantly reduced the triggering time for the scooping phase (from 3.75 s to 0.37 s) and lowered energy consumption (from 57.9 to 12.8 kJ). The choice of scooping strategy also had a considerable impact on the final bucket fill rate. Although “Just in & out” resulted in faster scooping times, it led to a lower bucket fill rate of only 82.4% in high-resistance scenarios such as sand handling. Compared with “Stairway #3”, the adaptive “Stairway” strategy, assisted by perceptual feedback, reduced unnecessary motions with only a marginal decrease in bucket fill rate, thereby shortening the total scooping duration and cutting energy use. Our proposed “auto-selection” method, which integrates three key mechanisms (perception-based triggering, dynamic policy selection, and feedback-controlled adjustment of “Stairway” repetitions) achieved overall performance, demonstrating balanced improvements in bucket fill rate, time efficiency, and energy consumption. Moreover, it is evident that the automated scooping triggering phase outperforms manual operations.

7.5. Large-scale, long-term production

As shown in Table 6, the proposed algorithm combines both strategies and has been stably operational for 8 weeks at Mixing Station B and 5 weeks at Mixing Station A, executing 1603 and 487 loading tasks, respectively, all of which successfully completed the shoveling tasks. Taking Mixing Station B as an example, when scooping gravel, the “Just in & out” strategy was primarily chosen, accounting for 96.6% of all gravel data, indicating that this strategy performs well in most cases. When scooping river sand, the “Stairway” strategy was mainly selected, with about 79.1% of atomic actions being 1 (“Stairway #1”), having similar time consumption but higher energy consumption compared to the “Just in & out” strategy; about 20.9% of atomic actions were 3 (“Stairway #3”), with both time and energy consumption exceeding the “Just in & out” strategy by more than 20% (The “Stairway #2” strategy was employed only on 22 tasks, which was not accounted for in the table). Additionally, 3.4% of gravel were scooped using the “Stairway” strategy, indicating unexpected situations where the bucket was not full but the vehicle speed was zero, necessitating intervention through feedback. Multiple large-scale production experiments have shown that the proposed algorithm can fully consider the possible scenarios in actual production, flexibly select different strategies according to the situation, and truly balance efficiency, energy consumption, and robustness. If the fixed number of “Stairway #3” scooping strategies used in Refs. Cao et al. (2023a,b) is taken as the baseline, the proposed adaptive scooping strategy, after 2090 long-term tests, has reduced the time spent in the scooping phase by 22% and energy consumption by 13%, while maintaining a high bucket fill rate to meet the actual production requirements of the mixing station. As shown in Table 7, we re-evaluated the performance of the automated algorithm based on material type. For sand, the exclusive use of the “Stairway” strategy resulted in longer scooping time and higher energy consumption. In the case of gravel, the fallback “Stairway” strategy was occasionally triggered under special circumstances, such as insufficient forward speed towards the pile. However, such instances accounted for only 3.9% of all gravel scooping actions, relative to the dominant “Just in & out” strategy.

A statistical analysis was conducted on the time consumption and energy consumption of different shoveling strategies in actual production. Specifically, the time consumption refers to the duration from the moment the bucket contacts the material to the completion of the

Table 6
Operational data from stable operations over 8 weeks at Mixing Station B and 5 weeks at Mixing Station C, recording material types processed by different shoveling strategies, average shoveling time, and average energy consumption.

Week	Just in & out			Stairway #1				Stairway #3			
	Gravel	Time (s)	Energy (kJ)	Sand	Gravel	Time (s)	Energy (kJ)	Sand	Gravel	Time (s)	Energy (kJ)
0920–0922	59	5.12	208.86	13	1	5.06	241.58	1	1	7.01	289.8
0923–0929	249	4.97	204.59	44	1	5.06	246.98	8	5	7.12	289.04
1007–1013	285	4.89	201.5	86	2	4.86	234.5	12	8	6.47	252.74
1014–1020	115	4.82	199.02	17	0	4.84	236.55	2	1	6.07	204.1
1021–1027	48	4.58	188.23	1	0	4.29	230.33	0	0	–	–
1104–1110	96	5.12	205.74	69	0	5.01	226.07	40	7	6.44	241.07
1111–1117	109	5.16	207.4	77	4	5.11	242.87	20	5	6.49	222.67
1118–1122	92	4.91	191.18	98	2	4.81	215.66	24	1	6.25	238.85
Sum/Mean/Mean	1053	4.95	201.86	415		4.95	231.86	135		6.48	243.5
Week	Just in & out			Stairway #1				Stairway #3			
	Gravel	Time (s)	Energy (kJ)	Sand	Gravel	Time (s)	Energy (kJ)	Sand	Gravel	Time (s)	Energy (kJ)
1014–1020	64	4.71	202.90	30	1	5.09	266.49	0	0	–	–
1021–1027	53	4.53	191.94	24	2	4.68	222.24	2	3	5.98	255.64
1104–1110	94	4.88	216.63	30	5	5.56	246.58	0	3	6.64	264.04
1111–1117	21	4.96	232.19	24	0	5.15	263.52	0	2	6.15	286.23
1118–1122	90	4.80	223.84	39	0	5.02	260.30	0	0	–	–
Sum/Mean/Mean	322	4.77	212.87	155		5.12	252.55	10		6.21	264.28

Table 7
Shoveling performance by material type. Note that values for “Just in & out” and “Stairway” columns represent occurrence counts.

Material type	Just in & out	Stairway	Time (s)	Energy (kJ)
Big gravel	1221	23	4.96	205.74
Small gravel	154	31	5.04	210.86
Sand	0	661	5.38	241.65

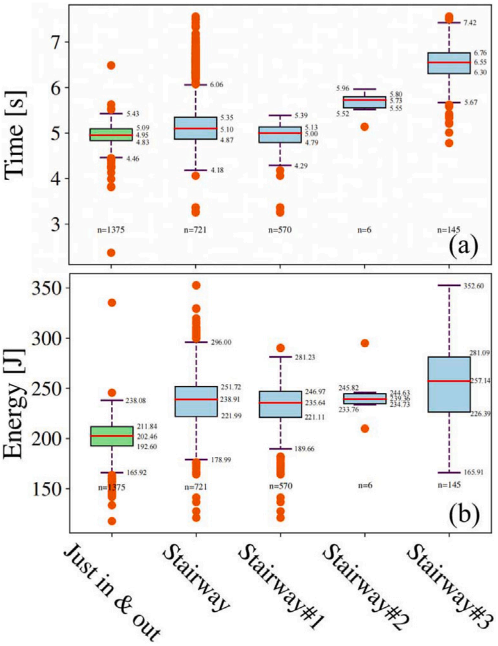


Fig. 22. Boxplot of time and energy consumption for different shoveling strategies; “Stairway#*i*” indicates the “Stairway” strategy with atomic action count *i*.

scooping motion, while the energy consumption is the total energy used by the travel motor and hydraulic motor during the shoveling period. From the boxplot 22(a), it can be observed that the time consumption of the “Stairway” strategy increases with the number of atomic actions. When the number of atomic actions is 1, the time consumption is similar to the “Just in & out” strategy, with a median of approximately 5 s; for each additional atomic action, the median time consumption increases by about 0.7 s, indicating that pauses between multiple atomic actions consume more time. From the boxplot 22(b), it is evident that the “Stairway” strategy consumes significantly more energy. When the number of atomic actions is 1, the median energy consumption is about 25% higher than the “Just in & out” strategy due to the increased difficulty in scooping the material and higher resistance; as the number of atomic actions increases, multiple actions are required, leading to an increasing trend in energy consumption. The experimental results confirm that the “Just in & out” strategy is superior to the “Stairway” strategy in terms of both time and energy consumption. However, due to frequent failures when scooping difficult materials like river sand, it is reasonable to automatically select between the two scooping strategies.

Over a period of three days, data were randomly collected from 20 daily material shoveling operations performed by a human operator using the same XCMG 968EV loader at Mixing Plant B, without prior notice that their operational data would be recorded. These data were compared with the proposed automatic shoveling system, as shown in Fig. 23. The average time for the manual shoveling process was 5.26 s, while the automatic shoveling system’s average time was 5.04 s, slightly lower than the manual shoveling strategy. Additionally, when comparing the average energy consumption of both shoveling processes, it was found that the automatic shoveling strategy (216.26 kJ) was 11% lower than the manual shoveling strategy (244.07 kJ). The proposed autonomous loader shoveling system achieves lower energy consumption than manual operation due to several key factors. Firstly, the system can predict the shoveling volume in real-time and trigger the shoveling action promptly based on the operational conditions, thereby avoiding energy losses caused by redundant movements. Moreover, the system ensures seamless transitions between different operations and selects the most suitable shoveling strategy based on the shoveling volume, material type, and vehicle speed, guaranteeing the completion of the shoveling task with minimal energy expenditure. Additionally, the autonomous loader system can provide accurate and timely outputs of the vehicle’s acceleration and boom angle information, enabling precise control of the loader’s power and transmission systems. The control system still has room for improvement. During the crowding phase, the system uses a pre-set fixed acceleration for advancement, which does not allow for real-time selection of the optimal throttle control based on the actual operating conditions of the vehicle. This limitation may lead to unnecessary energy consumption. Meanwhile, a significant variance in manual performance was also observed, with the minimum time being 4 s and the maximum time being 6 s, and the

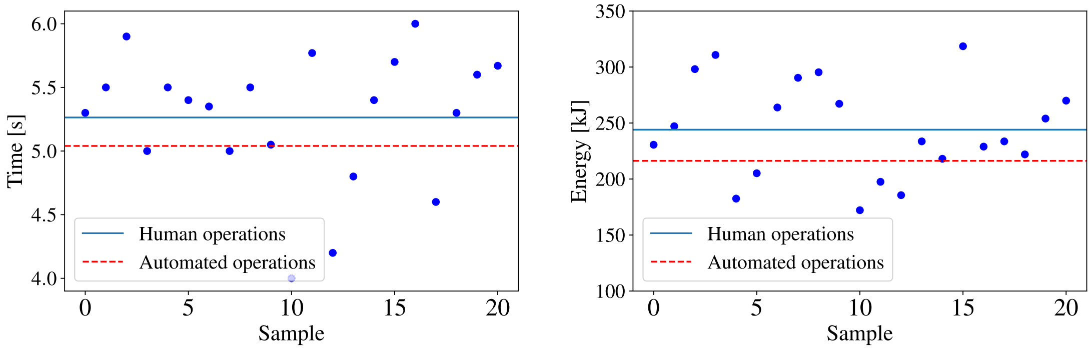


Fig. 23. Randomly collected time (left) and energy consumption (right) of manual daily material shoveling operations, with mean lines exceeding those of the automatic shoveling method.

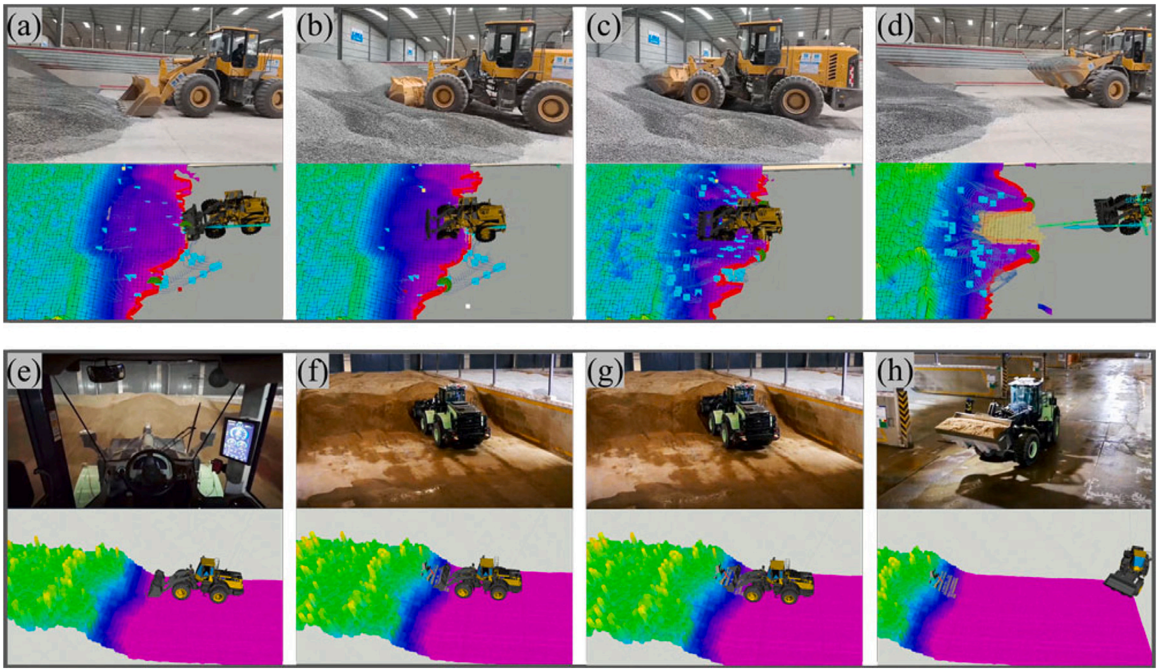


Fig. 24. The actual shoveling process and the changes in the semantic elevation map at Mixing Station A and Mixing Station B. (a)–(d) and (e)–(f) respectively document the entire process of rushing towards and contacting the pile, executing the shoveling action (retracting the bucket, raising the boom), and ultimately exiting. Please refer to the demonstration video at https://youtu.be/uHHbI35hjsY.

lowest energy consumption at 172.17 kJ and the highest at 310.8 kJ. Besides the influence of environmental factors such as the material pile, it is believed that the level of manual shoveling is easily affected by the urgency of mixing plant operations, the duration of continuous work by the operator, and subjective factors. In contrast, unmanned operations demonstrate more stable efficiency and energy consumption levels and can perform continuous work at high intensity for extended periods, which is a significant advantage of the automatic shoveling systems.

Two operation contents were randomly presented at Mixing Station A and Mixing Station B, with the changes in intermediate work states and semantic elevation maps being recorded, as shown in Fig. 24. Panels (a)–(d) depict the scene of the L955HE model unmanned fuel-powered loader at Mixing Station B, where the loading method based on semantic elevation maps and perception-triggered actions can handle piles of low-lying materials. Panels (e)–(h) show the scene of the XCMG 968EV unmanned electric loader at Mixing Station C, illustrating the loader's contact with the material pile, execution of the "Stairway" scooping action, and withdrawal process. The automatic shoveling system is responsible for computing the vehicle's travel acceleration and joint angle control signals at the upper layer and then transmitting these signals to the lower-level control units. For fuel-powered and electric loaders, the primary difference in the lower-level control lies in the control variables: fuel-powered loaders adjust the throttle to increase travel acceleration, while electric loaders control torque. However, the control of arm joint angles is uniformly achieved through the PWM signals of the corresponding hydraulic solenoid valves. This hierarchical control approach enables the system to be compatible with both fuel and electric vehicles, ensuring good versatility. It can be observed that the automatic shoveling system can autonomously select appropriate shoveling points, successfully complete scooping operations under different material types and vehicle models while maintaining a high bucket fill rate, and simultaneously update the semantic elevation map in real-time, ensuring real-time and precise modeling of the surrounding environment to facilitate subsequent operation processes.

7.6. System adaptation

This section details the adaptation process involved in deploying our proposed autonomous shoveling system to a new wheel loader

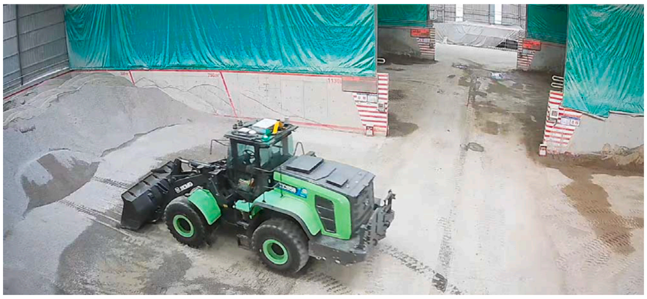


Fig. 25. Field deployment of the autonomous shoveling system on a new XCMG XC958EV wheel loader at a mixing station in Shandong, China.

model and a new mixing plant environment, highlighting the system's transferability. The initial step involves performing precise calibration of key dimensional parameters and sensor intrinsic/extrinsic parameters using the kinematic calibration method for the arm and the camera–LiDAR calibration approach introduced in Section 3. These parameters form the critical foundation for all subsequent modules. Subsequently, baseline vehicle controllers are tuned, including travel velocity control, arm joint position control, and trajectory tracking control. Building upon this foundational layer, adapting the scooping strategy itself requires calibrating only a few key poses, such as the arm and joint angles during the attacking phase (where the bucket must remain grounded) and the target joint angles at the end of the scooping phase.

Two components may require additional data collection: the semantic segmentation model and the shoveling volume prediction model. Should significant environmental differences exist (e.g., a new mixing plant layout), scene-specific data must be collected. The model can then be efficiently fine-tuned using a combination of automated SAM-based annotation and limited manual labeling. Similarly, if material properties differ substantially, the volume prediction model may need retraining with newly collected scooping data. Importantly, inaccuracies in these models rarely lead to critical failures and primarily affect performance metrics such as bucket fill rate, cycle time, and energy consumption. The system incorporates several robustness-oriented design choices: the fusion of semantic and elevation data mitigates the impact of unstable segmentation, while inaccuracies in volume prediction can be compensated for by adjusting operational thresholds (e.g., raising the volume trigger threshold to prioritize bucket fill rate over efficiency). This allows the system to remain operational while continuously improving through data collected during deployment.

A case in point is the successful deployment of the system on a new XCMG XC958EV electric autonomous wheel loader at a mixing station in Shandong, China (as shown in Fig. 25). The calibration and baseline controller tuning were completed within one day. During a one-week trial run, perception and scooping data were collected to refine the models, after which the system achieved stable and efficient long-term operation.

8. Limitations and future work

Although our adaptive selection mechanism between the "Just in & out" and "Stairway" strategies has achieved a balanced performance in bucket fill rate, efficiency, and energy consumption across two mixing stations, the current switching logic still relies on relatively simple rule-based conditions. For instance, the difficulty of scooping is currently determined primarily based on material type alone. This can lead to suboptimal decisions, such as misclassifying a challenging gravel pile as suitable for the "Just in & out" strategy, ultimately resulting in vehicle stoppage and an emergency switch to the "Stairway" strategy. Our observations indicate that scooping difficulty is influenced by multiple factors, including the vehicle's entry speed and material compaction density (Li et al., 2021). In response, we plan to develop a more sophisticated assessment method that integrates multi-modal state and perceptual information during the crowding phase to more accurately evaluate scooping difficulty in real time.

Furthermore, both the "Just in & out" and "Stairway" strategies operate largely in an open-loop manner across most control cycles, with fixed parameters such as throttle value and target joint angles. This rigidity may lead to suboptimal performance given the complex and dynamic interactions between the bucket and the material. Recent studies (Eriksson et al., 2024b) have introduced methods based on world models and reinforcement learning, which adaptively adjust the scooping strategy using real-time perception of vehicle and environmental states. These approaches also leverage historical scooping experiences to learn latent states in a world model, thereby continuously improving future performance.

In future work, we aim to investigate data-driven methods for evaluating scooping difficulty and enabling adaptive adjustment of the scooping strategy. This will include exploring reinforcement learning frameworks capable of closed-loop control and integrating predictive models to enhance overall autonomy and robustness.

9. Conclusions

This paper presents a novel automatic shoveling system that ensures the safety of the loader and its environment while balancing robustness, efficiency, and energy consumption. Such a system can promote technological advancement and intelligent development in the construction machinery industry. Initially, the system achieves automatic calibration of cameras and LiDAR, effectively solving the issue of long-term calibration stability caused by severe vibrations in construction machinery operation scenarios. Secondly, by employing multi-frame, multi-sensor confidence fusion technology for pile surface perception, the system can achieve real-time and accurate segmentation of the pile surface even under environmental light interference and uneven pile surface shapes. Subsequently, a more comprehensive and integrated evaluation index for shoveling point selection is formulated based on the pile elevation map, which fully weighs the safety of the loader, environmental safety, and operational efficiency. Then, by predicting the shoveling volume and calculating the current shoveling volumes, the system enables the timely and accurate triggering of shoveling actions with hydraulic delay characteristics, ensuring accurate control of the shoveling volume and high efficiency of the shoveling operation. Finally, the loader can autonomously select the optimal scooping action between open-loop prediction and closed-loop feedback based on the actual conditions during production operations. The system has undergone two months of field production in three mixing stations, demonstrating stable and outstanding performance throughout the tasks, ensuring high efficiency and low cost of shoveling operations (achieving an average efficiency level comparable to manual operation while reducing energy consumption by 11%).

CRediT authorship contribution statement

Guangda Chen: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Zhiwen Zhang:** Writing – original draft, Visualization, Software, Methodology, Investigation, Data curation. **Lin Cheng:** Writing – review & editing, Writing – original draft, Methodology,S5 PDF Investigation, Conceptualization. **Cheng Jin:** Writing – original draft, Software, Methodology, Data curation. **Shunyi Yao:** Writing – original draft, Software, Methodology. **Yue Wang:** Writing – review & editing, Supervision. **Rong Xiong:** Writing – review & editing, Supervision. **Yingfeng Chen:** Supervision, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors would like to express their gratitude to Qi Sheng for his support in software development, which has ensured the long-term stability of the system. They also appreciate the daily operational support and data collection efforts provided by Guohang Niu in bringing the work into the actual production environment. Additionally, they appreciate the dedication of Yingqing Liao and Xinwei Yang in advancing the commercialization of the project product. Lastly, they would like to thank Changjie Fan for his support in providing project resources.

Data availability

Data will be made available on request.

References

Agarwal, S., Mierle, K., Team, T.C.S., 2023. Ceres solver. URL: https://github.com/ceres-solver/ceres-solver.

Agrawal, T.K., Hanson, R., Sultan, F.A., Johansson, M.I., Andersson, D., Stefansson, G., Katsela, K., Browne, M., 2023. Automating loading and unloading for autonomous transport: Identifying challenges and requirements with a systems approach. In: Alfnes, E., Romsdal, A., Strandhagen, J.O., von Cieminski, G., Romero, D. (Eds.), Advances in Production Management Systems. Production Management Systems for Responsible Manufacturing, Service, and Logistics Futures. Springer Nature Switzerland, Cham, pp. 332–345.

Almqvist, H., 2009. Automatic bucket fill.

Aoshima, K., Fälldin, A., Wadbro, E., Servin, M., 2023. World modeling for autonomous wheel loaders. Automation URL: https://api.semanticscholar.org/CorpusID:262084021.

Cao, B., Liu, X., Chen, W., Li, H., Wang, X., 2023a. Intelligentization of wheel loader shoveling system based on multi-source data acquisition. Autom. Constr. 147, 104733.

Cao, B.-w., Liu, C.-y., Chen, W., Tan, P., Yang, J.-w., 2023b. Shovel-loading cooperative control of loader under typical working conditions. ISA Trans. 142, 702–715.

Cardenas, D., Loncomilla, P., Inostroza, F., Parra-Tsunekawa, I., Ruiz-del Solar, J., 2023. Autonomous detection and loading of ore piles with load–haul–dump machines in room & pillar mines. J. Field Robot. 40 (6), 1424–1443.

Chen, G., Dong, W., Yao, Z., Bi, Q., Li, X., 2025. Estimating bucket fill factor for loaders using point cloud hole repairing. Autom. Constr. 170, 105886.

Chen, Y., Jiang, H., Shi, G., Zheng, T., 2022b. Research on the trajectory and operational performance of wheel loader automatic shoveling. Appl. Sci. 12 (24), http://dx.doi.org/10.3390/app122412919, URL: https://www.mdpi.com/2076-3417/12/24/12919.

Chen, J., Lu, W., Yuan, L., Wu, Y., Xue, F., 2022a. Estimating construction waste truck payload volume using monocular vision. Resour. Conserv. Recycl. 177, 106013.

Chen, Y., Shi, G., Tan, C., Wang, Z., 2023. Machine learning-based shoveling trajectory optimization of wheel loader for fuel consumption reduction. Appl. Sci. 13 (13), http://dx.doi.org/10.3390/app13137659, URL: https://www.mdpi.com/2076-3417/13/13/7659.

Chen, G., Wang, Y., Li, X., Bi, Q., Li, X., 2024. Shovel point optimization for unmanned loader based on pile reconstruction. Comput. Aided Civ. Infrastruct. Eng..

Dadhich, S., Bodin, U., Andersson, U., 2016. Key challenges in automation of earth-moving machines. Autom. Constr. 68, 212–222.

Dadhich, S., Bodin, U., Sandin, F., Andersson, U., 2018. From tele-remote operation to semi-automated wheel-loader. Int. J. Electr. Electron. Eng. Telecommun. 7 (4), 178–182.

Dadhich, S., Sandin, F., Bodin, U., Andersson, U., Martinsson, T., 2019. Field test of neural-network based automatic bucket-filling algorithm for wheel-loaders. Autom. Constr. 97, 1–12. http://dx.doi.org/10.1016/j.autcon.2018.10.013, URL: https://www.sciencedirect.com/science/article/pii/S0926580518305119.

Eriksson, D., Ghabcheloo, R., Geimer, M., 2024a. Automatic loading of unknown material with a wheel loader using reinforcement learning. In: 2024 IEEE International Conference on Robotics and Automation. ICRA, pp. 3646–3652. http://dx.doi.org/10.1109/ICRA57147.2024.10610221.

Eriksson, D., Ghabcheloo, R., Geimer, M., 2024b. Optimizing bucket-filling strategies for wheel loaders inside a dream environment. Autom. Constr. 168, 105804.

Fernando, H., Marshall, J., 2020. What lies beneath: Material classification for autonomous excavators using proprioceptive force sensing and machine learning. Autom. Constr. 119, 103374.

Fernando, H.A., Marshall, J.A., Almqvist, H., Larsson, J., 2018. Towards controlling bucket fill factor in robotic excavation by learning admittance control setpoints. In: Field and Service Robotics: Results of the 11Th International Conference. Springer, pp. 35–48.

Filla, R., Ericsson, A., Palmberg, J.-O., 2005. Dynamic simulation of construction machinery: Towards an operator model. http://dx.doi.org/10.13140/RG.2.1.3915.5680.

Filla, R., Obermayr, M., Frank, B., 2014. A study to compare trajectory generation algorithms for automatic bucket filling in wheel loaders. In: 3rd Commercial Vehicle Technology Symposium. pp. 588–605.

Fischler, M.A., Bolles, R.C., 1981. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Commun. ACM 24 (6), 381–395.

Frank, B., Kleinert, J., Filla, R., 2018. Optimal control of wheel loader actuators in gravel applications. Autom. Constr. 91, 1–14.

Frank, B., Skogh, L., Alaküla, M., 2012. On wheel loader fuel efficiency difference due to operator behaviour distribution. In: 2nd International Commercial Vehicle Technology Symposium. CVT, pp. 1–18.

Gu, Y., Wu, J., Liu, C., 2025. Error analysis and accuracy evaluation method for coordinate measurement in transformed coordinate system. Measurement 242, 115860. http://dx.doi.org/10.1016/j.measurement.2024.115860, URL: https://www.sciencedirect.com/science/article/pii/S0263224124017457.

Hemami, A., Hassani, F., 2009. An overview of autonomous loading of bulk material. In: 2009 26th International Symposium on Automation and Robotics in Construction. ISARC 2009.

Kamari, M., Ham, Y., 2021. Vision-based volumetric measurements via deep learning-based point cloud segmentation for material management in jobsites. Autom. Constr. 121, 103430.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.-Y., et al., 2023. Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4015–4026.

Koyachi, N., Sarata, S., 2009. Unmanned loading operation by autonomous wheel loader. In: 2009 ICCAS-SICE. IEEE, pp. 2221–2225.

Lee, J.S., Ham, Y., Park, H., Kim, J., 2022. Challenges, tasks, and opportunities in teleoperation of excavator toward human-in-the-loop construction automation. Autom. Constr. 135, 104119. http://dx.doi.org/10.1016/j.autcon.2021.104119.

Li, J., Chen, C., Li, Y., Wu, H., Li, X., 2021. Difficulty assessment of shoveling stacked materials based on the fusion of neural network and radar chart information. Autom. Constr. 132, 103966.

Lindmark, D.M., Servin, M., 2018. Computational exploration of robotic rock loading. Robot. Auton. Syst. 106, 117–129. http://dx.doi.org/10.1016/j.robot.2018.04.010, URL: https://www.sciencedirect.com/science/article/pii/S0921889017305511.

Liu, J., 2023. Automatic calibration for camera and solid-state LiDAR (livox). https://github.com/GAfieldCN/automatic-camera-pointcloud-calibration.

Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., et al., 2025. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In: European Conference on Computer Vision. Springer, pp. 38–55.

Luo, Z., Yan, G., Cai, X., Shi, B., 2024. Zero-training lidar-camera extrinsic calibration method using segment anything model. In: 2024 IEEE International Conference on Robotics and Automation. ICRA, IEEE, pp. 14472–14478.

Magnusson, M., Almqvist, H., 2011. Consistent pile-shape quantification for autonomous wheel loaders. In: 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems. IEEE, pp. 4078–4083.

Miki, T., Wellhausen, L., Grandia, R., Jenelten, F., Homberger, T., Hutter, M., 2022. Elevation mapping for locomotion and navigation using gpu. In: 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems. IROS, IEEE, pp. 2273–2280.

Nezhadali, V., Frank, B., Eriksson, L., 2016. Wheel loader operation—Optimal control compared to real drive experience. Control Eng. Pract. 48, 1–9.

Ren, T., Liu, S., Zeng, A., Lin, J., Li, K., Cao, H., Chen, J., Huang, X., Chen, Y., Yan, F., et al., 2024. Grounded sam: Assembling open-world models for diverse visual tasks. arXiv preprint arXiv:2401.14159.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.-C., 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4510–4520.

Sarata, S., 2006. Development of autonomous system for loading operation by wheel loader. pp. 466–471, IS-ARC 2006.

Sarata, S., Kiyachi, N., Sugawara, K., 2008. Measuring and update of shape of pile for loading operation by wheel loader. In: 25th ISARC.

Shan, T., Englot, B., Meyers, D., Wang, W., Ratti, C., Rus, D., 2020. Lio-sam: Tightly-coupled lidar inertial odometry via smoothing and mapping. In: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems. IROS, IEEE, pp. 5135–5142.

Wu, L., 2003. A Study on Automatic Control of Wheel Loaders in Rock/soil Loading. The University of Arizona.

Xu, Z., Lu, X., Xu, E., Xia, L., 2022. A sliding system based on single-pulse scanner and rangefinder for pile inventory. IEEE Geosci. Remote. Sens. Lett. 19, 1–5.

Xu, Z., Peng, Y., Lin, J., Yang, K., Peng, S., 2024. An improved sliding system based on multi-pulse scanner and rangefinder for pile inventory. IEEE Trans. Instrum. Meas..

Yang, C., 2025. Interval riccati equation-based and non-probabilistic dynamic reliability-constrained multi-objective optimal vibration control with multi-source uncertainties. J. Sound Vib. 595, 118742.

Yang, C., Liu, Y., Gao, H., 2025a. Reliability-constrained uncertain spacecraft sliding mode attitude tracking control with interval parameters. IEEE Trans. Aerosp. Electron. Syst..

Yang, C., Zuo, Y., Lu, W., Wang, T., 2025b. Uncertain attitude tracking control for QUAV based on interval LQT with states reliability constraints. Nonlinear Dynam. 1–21.

Yao, J., Edson, C.P., Yu, S., Zhao, G., Sun, Z., Song, X., Stelson, K.A., 2023. Bucket loading trajectory optimization for the automated wheel loader. IEEE Trans. Veh. Technol. 72 (6), 6948–6958.

You, K., Zhou, C., Ding, L., Chen, W., Zhang, R., Xu, J., Wu, Z., Huang, C., 2023. Earthwork digital twin for teleoperation of an automated bulldozer in edge dumping. J. Field Robot. 40 (8), 1945–1963.

Zauner, M., Altenberger, F., Knapp, H., Kozek, M., 2020. Phase independent finding and classification of wheel-loader work-cycles. Autom. Constr. 109, 102962.

Zhang, W., Huang, Z., Luo, G., Chen, T., Wang, X., Liu, W., Yu, G., Shen, C., 2022. Topformer: Token pyramid transformer for mobile semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12083–12093.

Zhang, L., Zhao, J., Long, P., Wang, L., Qian, L., Lu, F., Song, X., Manocha, D., 2021. An autonomous excavator system for material loading tasks. Sci. Robot. 6 (55), eabc3164. http://dx.doi.org/10.1126/scirobotics.abc3164, arXiv:https://www.science.org/doi/pdf/10.1126/scirobotics.abc3164. URL: https://www.science.org/doi/abs/10.1126/scirobotics.abc3164.