separations

MDPI

Article

RGB-D Semantic-Guided 3D Geometric Estimation for Intelligent Coal Gangue Separation

Zhiyuan Sun [1], Zhongjun Zhou [2], Tianyang Sun [2], Shiyu Wang [2] and Kaile Xiao [1,*]

[1] College of Applied Science and Technology, Beijing Union University, Beijing 100012, China; 13522080222@163.com
[2] School of Mechanical and Electrical Engineering, China University of Mining and Technology (Beijing), Beijing 100083, China; zz20251215@gmail.com (Z.Z.); sty2550068@yeah.net (T.S.); wsy98658@163.com (S.W.)
* Correspondence: yktkaile@buu.edu.cn

Abstract

Coal and gangue separation is crucial for clean and efficient coal utilization. However, practical sorting environments are highly unstructured, where irregular fragments exhibit random distribution, varying orientations, and occlusion. Conventional 2D vision methods cannot provide the 3D geometric information required for intelligent sorting, while model-based 6D pose estimation is unsuitable for coal gangue fragments due to their irregular morphology and lack of predefined models. This study proposes an RGB-D cascaded fusion framework for partial point cloud-based 3D geometric estimation. The framework follows a detection–segmentation–reconstruction pipeline, where object detection provides semantic priors, zero-shot segmentation generates pixel-level masks, and the masks guide depth extraction and local point cloud reconstruction. Based on the reconstructed point clouds, task-oriented geometric parameters, including centroid position, oriented bounding box dimensions, visible volume, and principal orientation, are estimated for intelligent separation. Experiments on a dedicated RGB-D coal and gangue dataset demonstrate that the proposed method achieves a centroid localization error of 14.22 mm, dimension error of 9.34%, and volume error of 11.3%. Compared with detection-box-based depth reconstruction, the proposed semantic-guided strategy significantly improves geometric estimation accuracy. The proposed framework provides reliable 3D perception for intelligent coal gangue separation and robotic sorting.

Keywords: coal gangue separation; RGB-D perception; 3D geometric estimation; semantic-guided reconstruction

Check for updates

Academic Editors: Yadong Zhang, Guangqing Zhu and Xuan Xu

Received: 10 August 2026
Revised: 4 September 2026
Accepted: 8 September 2026
Published: 15 September 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Gangue, the principal solid waste byproduct of coal mining, accounts for approximately 15–20% of raw coal production. Its massive cumulative stockpiles—exceeding 7 billion tons in China alone—pose severe environmental threats including spontaneous combustion and heavy metal leaching [1]. Different qualities of gangue correspond to differentiated utilization pathways—including power generation, construction materials, and chemical extraction—making efficient separation of coal from gangue a prerequisite for quality-based utilization [2]. Among conventional sorting methods, manual picking is inefficient, dense-medium separation consumes substantial water and generates secondary pollution, and γ/X-ray-based sorting entails high equipment costs, complex maintenance requirements, and regulatory compliance burdens [3]. These limitations underscore an urgent need for non-contact, cost-effective intelligent sorting technologies.

Machine vision, owing to its rich informational content and real-time feedback capability, has emerged as the primary sensing technology for intelligent sorting. Early coal gangue recognition relied predominantly on handcrafted features such as color and texture; however, the complex illumination and severe dust interference characteristic of underground mines limited its robustness. In recent years, deep learning has substantially advanced recognition performance. Pu et al. [4] proposed the CG-YOLO model, which optimizes anchor boxes via K-Means clustering for coal gangue classification and localization on belt conveyors, yet the localization error remained at 9–12.5 mm and only two-dimensional coordinates were provided. Wen et al. [5] integrated Swin Transformer with YOLOv5 to improve detection accuracy under lightweight constraints. He et al. [6] proposed a dual-energy X-ray-based coal gangue identification method that despite optimized feature thresholds was constrained in practical deployment by radiation risks. These methods primarily perform bounding-box-level localization and cannot capture precise target contours, thereby falling short of the fine-grained shape perception required by sorting robots.

To obtain more precise target contours, image segmentation techniques have been introduced into coal gangue recognition. In semantic segmentation, Wang et al. [7] achieved coal gangue segmentation on the top-coal caving face using DeepLabV3+. He et al. [8] proposed the SSNet_CG network based on an improved PSPNet, optimizing the backbone with attention mechanisms and depthwise separable convolutions to achieve rapid segmentation under multi-scale, adherent, and semi-occluded conditions, reaching an mIoU of 95.4%. Yuan et al. [9] embedded feature pyramid and atrous spatial pyramid pooling modules into U-Net, achieving an mIoU of 95.65% and applying it to top-coal caving control. Guo et al. [10] proposed an adaptive low-illumination enhancement approach for real-time segmentation, using CycleGAN to enhance low-light images followed by a lightweight DeepLabV3+ for segmentation, achieving an mIoU of 91.76%. Wang et al. [11] implemented coal gangue semantic segmentation based on SegFormer, reaching an mIoU of 93.03%. Although semantic segmentation enables pixel-level classification of coal gangue regions into two categories (coal and gangue), it cannot distinguish overlapping or connected lumps of the same category, precluding instance-level localization and necessitating instance segmentation methods. In this regard, Cao et al. [12] proposed MFCCM-Mask R-CNN, enhancing feature extraction through multi-channel forward-link confusion convolution and achieving a segmentation accuracy of 97.38%. Lai et al. [13] employed multispectral imaging with an improved Mask R-CNN, utilizing a hybrid attention mechanism to mitigate accuracy degradation under occlusion, albeit at the cost of increased hardware complexity. For real-time applications, Zhu et al. [14] proposed an improved YOLOv8 segmentation model with an mAP of 96.88%, and Bai et al. [15] proposed the DASeg framework, combining object detection with SAM2 for coal lump segmentation, achieving a pixel accuracy of 93.1%. These methods have yielded significant progress in segmentation accuracy; nevertheless, the irregular shapes, blurred boundaries, and high inter-class visual similarity of coal gangue mean that pixel-level annotation requires laborious point-by-point manual tracing along target contours, making large-scale annotated dataset construction prohibitively expensive. Zero-shot segmentation models can generate high-quality masks without annotation, yet they lack semantic discrimination capability and cannot differentiate coal from gangue. More critically, all the aforementioned methods are inherently limited to two-dimensional representations and fail to integrate segmentation outcomes with three-dimensional geometric reconstruction. Automated object separation, however, fundamentally relies on geometric attributes—not merely semantic cues. For instance, robotic sorting systems require precise 3D coordinates and principal orientation estimates to execute robust grasp planning; similarly, size- and mass-based

classification necessitates accurate dimensional and volumetric quantification of individual blocks—capabilities that cannot be reliably inferred from 2D projections alone.

To acquire three-dimensional information, several researchers have explored 3D sensing technologies. By integrating depth geometry, the inherent limitations of RGB-only methods are fundamentally resolved through four key mechanisms: illumination invariance against variable underground lighting, robust background separation via depth discontinuities, topographic feature discrimination for 2D-invisible surface details, and occlusion-robust geometric reasoning for stacked/partially occluded scenes. Dong et al. [16] proposed a dynamic coal volume measurement method based on binocular vision, estimating volume by separating targets through disparity maps. Wang et al. [17] applied 3D laser scanning, deriving the volume relationship of coal gangue under the assumption that void ratio and true volume follow a normal distribution, and analyzed error sources. He et al. [18] utilized line lasers and industrial cameras for belt conveyor coal volume measurement. Wu [19] employed 3D laser scanning for earthwork volume calculation, comparing static and airborne point cloud volume methods. These methods, however, are primarily oriented toward macroscopic volume measurement of coal on conveyor belts rather than geometric feature estimation of individual coal gangue lumps, and thus cannot provide target-level pose and dimensional information for sorting robots.

In the domain of 3D pose estimation, He et al. [20] proposed PVN3D for RGB-D pose estimation via keypoint voting. He et al. [21] proposed FFB6D, constructing a full-flow bidirectional fusion network. Wang et al. [22] proposed DenseFusion for pixel-wise RGB-D feature fusion. Wen et al. [23] further established FoundationPose as a unified foundation model framework. Cai et al. [24] introduced GS-Pose, a cascaded pose estimation architecture, and Lin et al. [25] developed SAM-6D, incorporating SAM into zero-shot 6D pose estimation. Nevertheless, all these methods rely on target CAD models for pose alignment, whereas coal gangue lumps exhibit irregular shapes that defy prior modeling, rendering this paradigm inapplicable.

The preceding review reveals that existing methods face three fundamental challenges in achieving reliable coal gangue separation:

(1) Irregular morphology without prior models. Unlike industrial components with predefined CAD geometries, coal and gangue lumps exhibit highly irregular shapes and varying sizes, rendering model-based 6D pose estimation paradigms inapplicable.

(2) Unstructured and cluttered acquisition scenarios. In practical sorting environments, fragments are randomly distributed with frequent occlusions, contacts, and adherence, posing significant challenges for both 2D segmentation and depth-based reconstruction.

(3) Gap between semantic recognition and geometric reconstruction. At the semantic level, coal and gangue share highly similar appearances with low contrast against the dark background, exhibiting small inter-class differences and large intra-class variations. At the geometric level, directly extracting depth from detection boxes introduces substantial background noise, while existing 3D methods either target macroscopic bulk measurement or require predefined object models.

To address these challenges, this study proposes a 3D geometric estimation method based on RGB-D cascade fusion, with the following main contributions:

(1) Employing the segmentation mask as a “semantic–geometric bridge” to transfer semantic discrimination capability to the depth reconstruction stage, effectively suppressing background and interferent depth noise.

(2) Designing a “detection–segmentation–reconstruction” three-stage cascade architecture that achieves progressive refinement from semantic priors through pixel-level segmentation to 3D geometry.

(3) Defining a 3D geometric description for irregular objects—comprising the 3D centroid, oriented bounding box dimensions, volume, and principal axis direction—that characterizes the spatial state required for sorting without dependence on CAD models.

(4) Constructing a coal gangue RGB-D dataset and validating the effectiveness of the proposed method under simulated working conditions.

2. Materials and Methods

2.1. Overview of the Proposed Framework

A cascaded RGB-D fusion framework is proposed in this study for three-dimensional geometric parameter estimation of coal and gangue fragments. The framework adopts a three-stage detection–segmentation–reconstruction pipeline, where RGB images and depth maps acquired from an RGB-D camera are used as inputs. Considering the irregular morphology, random spatial distribution, and lack of predefined geometric models of coal and gangue lumps in sorting scenarios, the objective of the proposed framework is to obtain the object-level three-dimensional geometric information required for intelligent separation.

Specifically, the proposed framework consists of three sequential stages. First, the object detection module extracts the category information and coarse locations of coal and gangue fragments from RGB images, providing semantic priors for subsequent processing. Second, the zero-shot segmentation module uses the detected bounding boxes as prompts to generate pixel-level instance masks within the target regions, refining object representation from bounding-box-level semantics to pixel-level contours. Finally, the generated instance masks are employed as spatial constraints for depth extraction. Valid depth values belonging to each target object are selected from the depth map, and local three-dimensional point clouds are reconstructed. Based on the reconstructed point clouds, several geometric parameters, including the three-dimensional centroid, oriented bounding box (OBB) dimensions, volume, and principal orientation, are estimated.

Within this cascaded framework, the segmentation mask serves as an intermediate representation connecting semantic perception and geometric reconstruction. On the one hand, the mask preserves semantic information, including object category and accurate boundaries. On the other hand, it defines the valid spatial region for depth processing, constraining three-dimensional reconstruction within the target area and reducing the influence of background regions and neighboring objects on geometric estimation. The overall workflow of the proposed method is illustrated in Figure 1.

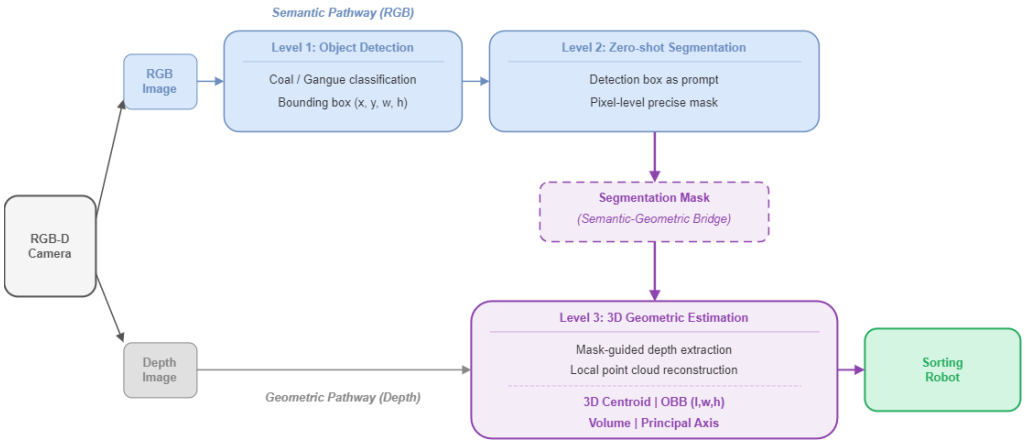


Figure 1. Workflow of the proposed detection–segmentation–reconstruction framework for RGB-D-based geometric estimation.

2.2. Semantic Prior Extraction

The object detection module takes RGB images as input and extracts semantic prior information of coal and gangue lumps, including object categories and two-dimensional locations. Since the detection module is mainly used to provide candidate regions for subsequent instance segmentation rather than to introduce a novel detection algorithm, YOLOv11 is adopted as the single-stage detector to achieve a balance between detection accuracy and inference efficiency.

The YOLOv11 detector is trained on the RGB coal gangue dataset established in this study (Section 2.5), with two categories defined: coal and gangue. During training, data augmentation strategies are applied, including random horizontal flipping, brightness and contrast perturbation, and random scaling and cropping, to improve robustness against illumination variations, object orientation changes, and fragment stacking conditions.

The detector outputs the category label c_i, confidence score s_i, and bounding box coordinates b_i:

$$b_i = (x_{\min}, y_{\min}, x_{\max}, y_{\max}) \tag{1}$$

where i denotes the detected object index. A confidence threshold τ_{det} is applied to remove unreliable predictions, and Non-Maximum Suppression (NMS) is used to eliminate redundant bounding boxes with an IoU threshold of 0.5. The final detection results $\{c_i, b_i, s_i\}_{i=1}^{n}$ are used as prompts for the subsequent zero-shot instance segmentation stage.

2.3. Semantic-Guided Zero-Shot Instance Segmentation

The second stage uses the detection results as spatial prompts to guide a zero-shot segmentation model in generating pixel-level instance masks, thereby refining semantic information from bounding-box-level representation to pixel-level object contours. In this study, Segment Anything Model 2 (SAM2) is adopted as the instance segmentation model. SAM2 is a prompt-driven vision foundation model with strong zero-shot generalization capability, enabling instance mask generation for coal and gangue fragments without additional fine-tuning on the target dataset.

SAM2 mainly consists of an image encoder, a prompt encoder, and a mask decoder. The image encoder based on the hierarchical vision Transformer (Hiera) architecture extracts multi-scale image features. The prompt encoder converts the input bounding box coordinates into spatial prompt embeddings, while the mask decoder integrates image features and prompt information through attention mechanisms to generate candidate masks and corresponding quality scores. The mask with the highest predicted quality score is selected as the final segmentation result.

For each detected object bounding box obtained from the first-stage detection module, the box is used as a prompt input to SAM2. The model performs instance segmentation within the prompted region and outputs a binary mask,

$$M_i \in \{0,1\}^{H \times W} \tag{2}$$

where pixels with a value of 1 represent the target object and pixels with a value of 0 represent the background. Compared with rectangular detection boxes, instance masks follow the actual object boundaries and thus provide refined spatial constraints for subsequent depth extraction and 3D reconstruction.

In complex stacking scenarios, multiple blocks may appear within a single detection box due to contact or adhesion between objects. Therefore, connected component analysis is applied to the generated masks to separate independent regions and assign individual instance labels, and connected regions with areas smaller than a predefined threshold are removed as noise. Morphological smoothing using opening operations (erosion followed

by dilation) is further applied to remove isolated pixels and improve boundary continuity. The refined instance masks $\{M_i\}_{i=1}^{n}$ are then used as spatial constraints for the depth extraction and point cloud reconstruction stage.

2.4. Mask-Guided 3D Reconstruction and Geometric Parameter Estimation

The third stage performs semantic-guided three-dimensional reconstruction and geometric parameter estimation using the instance masks generated in the previous stage. The segmentation masks serve as spatial constraints to extract object-specific depth information, reconstruct local point clouds, and estimate task-oriented geometric parameters. This stage consists of three main components: mask-constrained depth extraction, local point cloud reconstruction, and three-dimensional geometric parameter estimation.

2.4.1. Mask-Constrained Depth Extraction

In coal gangue separation, the detection bounding boxes usually contain not only the target blocks but also background regions, such as conveyor belts, surrounding materials, and neighboring objects. Due to the similar appearance of coal and gangue fragments and the low contrast between objects and the dark industrial background, directly extracting depth information from rectangular detection regions may introduce substantial background noise into subsequent point cloud reconstruction.

To address this issue, the generated instance masks are employed as spatial constraints to isolate object-specific depth information. For each detected object with bounding box b_i and corresponding mask M_i, the masked depth map is defined as

$$D_{\text{masked}}(x, y) = D(x, y) \cdot M_i(x, y) \tag{3}$$

where $D(x, y)$ represents the original depth value at pixel coordinate (x, y), and $M_i(x, y)$ denotes the binary mask. Only depth values within the foreground region are retained, while background pixels are removed.

To evaluate the effectiveness of mask-based depth extraction, two strategies are compared:

(1) Bounding-box-based extraction: All depth pixels within the detected rectangular region are used for reconstruction;

(2) Mask-based extraction: Only depth pixels belonging to the segmented foreground region are retained.

2.4.2. Local 3D Point Cloud Reconstruction

The mask-constrained depth map is converted into a local three-dimensional point cloud using the intrinsic parameters of the RGB-D camera. For each valid pixel (x, y), where $M_i(x, y) = 1$ and $D(x, y) > 0$, the corresponding 3D coordinate (X, Y, Z) is recovered through the pinhole camera model

$$\begin{aligned} Z &= D(x, y), \\ X &= \frac{(x - c_x)Z}{f_x}, \\ Y &= \frac{(y - c_y)Z}{f_y}. \end{aligned} \tag{4}$$

where f_x, f_y, c_x, and c_y denote the intrinsic parameters of the camera. The local point cloud of the i-th object is therefore represented as $P_i = \{p_1, p_2, \ldots, p_N\}$, where $p_j = (X_j, Y_j, Z_j)$ denotes an individual 3D point.

In this way, the reconstructed point cloud mainly contains object surface points, improving the robustness of subsequent geometric estimation.

Although semantic masking effectively suppresses background interference, residual outliers may still occur due to depth measurement uncertainty, surface reflectance variations, and missing depth values in low-reflectivity regions of coal gangue surfaces. Therefore, statistical outlier removal (SOR) filtering is applied to refine the reconstructed point cloud. For each point, the mean distance to its nearest neighbors is calculated, and points beyond the range of $\mu \pm \alpha \cdot \sigma$ are removed.

When necessary, voxel downsampling is further applied to normalize point density while preserving the overall geometric structure. The processed point cloud is denoted as $P_i^{'}$, which is subsequently used for three-dimensional geometric parameter estimation.

2.4.3. Three-Dimensional Geometric Parameter Estimation

Four geometric parameters required for intelligent coal gangue separation are extracted from the filtered local point cloud $P_i^{'}$, including centroid position, oriented bounding box (OBB) dimensions, visible volume, and principal orientation. The geometric state of each object is represented as

$$G_i = (O, l, w, h, V, \theta) \tag{5}$$

(1) Centroid Position

The centroid represents the spatial location of the object in the camera coordinate system and is calculated as the average coordinate of all points,

$$O = \frac{1}{N} \sum_{j=1}^{N} p_j \tag{6}$$

where N denotes the number of points in the reconstructed point cloud.

The depth coordinate z_c indicates the distance between the object and the camera, while x_c and y_c describe the lateral position of the object in the camera coordinate system. These spatial coordinates provide essential localization information for robotic sorting operations.

(2) Oriented Bounding Box Dimensions

To characterize the irregular shape of lumpy coal gangue, principal component analysis (PCA) is performed on the reconstructed point cloud.

The covariance matrix is calculated as

$$\Sigma = \frac{1}{N} \sum_{j=1}^{N} (p_j - O)(p_j - O)^T \tag{7}$$

Eigenvalue decomposition of Σ produces three eigenvalues $\lambda_1 \geq \lambda_2 \geq \lambda_3$ and corresponding eigenvectors v_1, v_2, and v_3, which define the principal axes of the point cloud.

The point cloud is projected onto this PCA coordinate system, and the object dimensions are calculated as

$$l = \max_j (p_j \cdot v_1) - \min_j (p_j \cdot v_1) \tag{8}$$

$$w = \max_j (p_j \cdot v_2) - \min_j (p_j \cdot v_2) \tag{9}$$

$$h = \max_j (p_j \cdot v_3) - \min_j (p_j \cdot v_3) \tag{10}$$

Compared with axis-aligned bounding boxes (AABB), OBB provides a more compact representation for irregular lumps because the estimated dimensions are aligned with the intrinsic geometric orientation of each object.

(3) Visible Volume Estimation

Due to the single-viewpoint RGB-D acquisition, only the upper-surface point cloud of each coal or gangue fragment is captured, while the bottom geometry remains unobservable. To address this incomplete reconstruction, we propose a height integration modeling method based on Random Sample Consensus (RANSAC) baseline plane fitting and bottom-surface truncation closure.

First, the background point cloud outside the target segmentation mask is used to automatically fit the supporting baseline plane of the conveyor belt via the RANSAC algorithm:

$$Ax + By + Cz + D = 0 \tag{11}$$

This approach eliminates the need for manual pre-calibration of the conveyor belt height and effectively compensates for height drift caused by slight belt tilting or mechanical vibration.

Second, to construct a complete 3D volumetric entity, the algorithm assumes that the bottom of each fragment is fully in contact with the conveyor belt surface. The filtered upper-surface point cloud P_i' is orthogonally projected downward along the normal direction onto the fitted baseline plane, thereby achieving bottom-boundary truncation and geometric closure. The entire 3D projected region is then discretized on the XY plane into an orthogonal 2D grid with a resolution of Δ (set to 2.0 mm in this study). For each occupied grid cell (u, v), the maximum height value $h(u, v)$ along the normal direction is extracted, and the volumetric increment of each discrete columnar cell is computed as

$$v(u, v) = [h(u, v) - z_{\text{belt}}(u, v)] \times \Delta^2 \tag{12}$$

By summing the volumetric contributions of all occupied grid cells, the visible 3D reconstructed volume of the fragment is obtained,

$$V = \sum v(u, v) = \Delta^2 \sum [h(u, v) - z_{\text{belt}}(u, v)] \tag{13}$$

where Ω denotes the set of grid cells corresponding to the orthogonal projection of the target onto the conveyor belt plane. By projecting the upper-surface point cloud onto the RANSAC-fitted baseline plane for closure and performing columnar volume integration, this approach avoids the artificial overestimation caused by the Convex Hull algorithm on non-convex surfaces and blind regions, while ensuring the rigor and consistency of irregular fragment volume modeling.

(4) Principal Orientation

The principal orientation describes the dominant spatial direction of the object and provides useful information for robotic grasp planning. In this study, the first principal component v_1, corresponding to the largest eigenvalue λ_1, is defined as the principal axis.

The orientation angle θ is calculated from the projection of v_1 onto the horizontal plane.

However, for approximately isotropic objects, the first principal component becomes unstable because the eigenvalues are close to each other. To quantify this uncertainty, an aspect ratio indicator is introduced:

$$R = \frac{\lambda_1}{\lambda_2} \tag{14}$$

When $R > 1.5$, the difference between the first and second principal directions is sufficiently large, and the estimated orientation is considered reliable. When $R \leq 1.5$, the object exhibits similar dimensions along the two dominant directions, resulting in high uncertainty; therefore, the estimated orientation is marked as low confidence.

2.5. Experimental Setup

All images were captured using an Intel RealSense D455 camera (Intel Corporation, Santa Clara, CA, USA). Model training and inference were conducted on a workstation equipped with an NVIDIA RTX PRO 4000 Blackwell (24GB) GPU (NVIDIA Corporation, Santa Clara, CA, USA) and dual Intel Xeon Gold 6530 CPUs (2 × 32 cores, 64 physical cores, 128 threads total) (Intel Corporation, Santa Clara, CA, USA). The detection branch adopted YOLOv11s (Ultralytics, version 8.4.87), and the segmentation branch adopted SAM 2.1 (Meta AI, model: sam2.1_hiera_base_plus). All models were implemented in Python 3.12.13 with PyTorch 2.10.0+cu128 and CUDA 12.8.

2.5.1. Data Acquisition

Coal and gangue lumps usually exhibit dark surfaces with strong light absorption characteristics, which may degrade depth acquisition performance. Considering that structured-light and ToF-based depth sensors are sensitive to low-reflectivity surfaces, an active stereo RGB-D camera was selected in this study due to its better adaptability to low-texture and dark objects. A comparison of depth acquisition results under different sensing principles is presented in Figure 2.

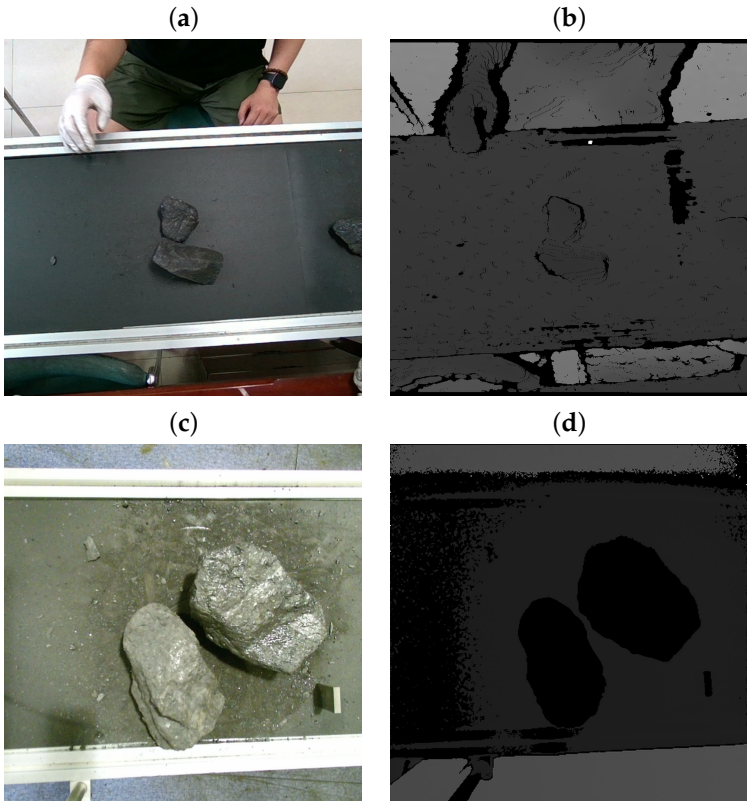


Figure 2. Comparison of RGB-D imaging results obtained from different depth sensing technologies. (**a**) RGB image captured by the active stereo RGB-D camera; (**b**) depth image captured by the active stereo RGB-D camera; (**c**) RGB image captured by the structured-light camera; (**d**) depth image captured by the structured-light camera.

RGB-D cameras consist of independent RGB and depth sensors with different optical centers and viewing geometries. Therefore, depth measurements need to be registered to the RGB coordinate system before 3D reconstruction. As illustrated in Figure 3, the depth frame is transformed into the RGB frame through extrinsic calibration using rotation and translation parameters, ensuring pixel-wise correspondence between RGB images and depth measurements.

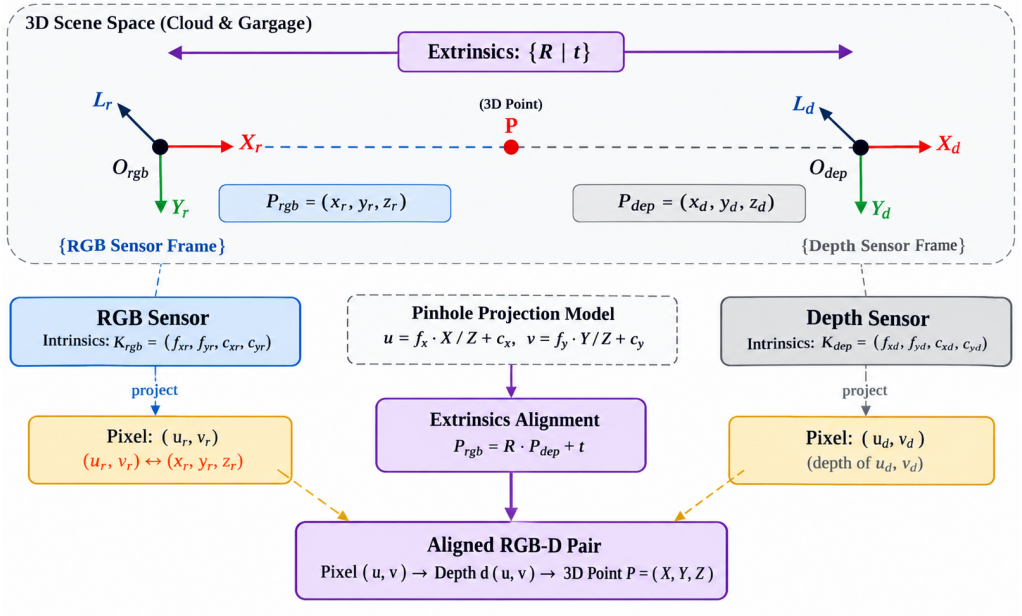


Figure 3. Illustration of RGB-D coordinate transformation and image registration. The depth sensor coordinate system is transformed into the RGB camera frame using extrinsic parameters $[R|t]$, allowing spatial alignment between RGB pixels and depth measurements for 3D reconstruction.

A 3D point P observed by the RGB and depth sensors is represented in their respective coordinate systems as $P_{\mathrm{rgb}} = (x_{\mathrm{r}}, y_{\mathrm{r}}, z_{\mathrm{r}})$ and $P_{\mathrm{dep}} = (x_{\mathrm{d}}, y_{\mathrm{d}}, z_{\mathrm{d}})$. The extrinsic transformation $[R|t]$ is applied to transform the depth camera coordinate system into the RGB camera coordinate system. Due to the different optical centers and viewing geometries, the projected pixel coordinates $(u_{\mathrm{r}}, v_{\mathrm{r}})$ and $(u_{\mathrm{d}}, v_{\mathrm{d}})$ are initially misaligned. After extrinsic calibration and image registration, the depth measurements are spatially aligned with the RGB image, allowing each RGB pixel location to be associated with its corresponding depth value for subsequent 3D reconstruction.

The RGB and depth images were spatially aligned at a resolution of 1280 × 720 pixels. The depth sensor originally provided 16-bit depth measurements. For compact storage, depth values within the 0–3 m operating range were retained and linearly mapped to the 8-bit range [0, 255],

$$D_8 = \mathrm{round}\left(\frac{\mathrm{clip}(D_{\mathrm{raw}}, 0, 3)}{3} \times 255\right) \tag{15}$$

where D_{raw} denotes the original metric depth in meters and D_8 denotes the stored 8-bit depth value. When the stored depth images were used for three-dimensional reconstruction, the metric depth was recovered as

$$D(x, y) = D_8(x, y) \times \frac{3}{255} \tag{16}$$

The mapping range of 0–3 m covers the operating distance of the coal gangue sorting system.

2.5.2. RGB-D Camera Calibration and Intrinsic Parameters

An RGB-D depth camera with a resolution of 1280 × 720 pixels was employed for data acquisition. The intrinsic parameters of the camera were calibrated using the ChArUco calibration method. The ChArUco board combines the advantages of checkerboard corner detection and ArUco marker identification, providing accurate sub-pixel corner localization and robust feature recognition during calibration.

As shown in Figure 4, multiple images of the ChArUco calibration board were captured under different viewing angles, distances, and poses. The camera intrinsic parameters were estimated by jointly optimizing the detected chessboard corners and ArUco marker features. The intrinsic matrix K is expressed as

$$K = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \tag{17}$$

where f_x and f_y represent the focal lengths, and c_x and c_y denote the principal point coordinates.

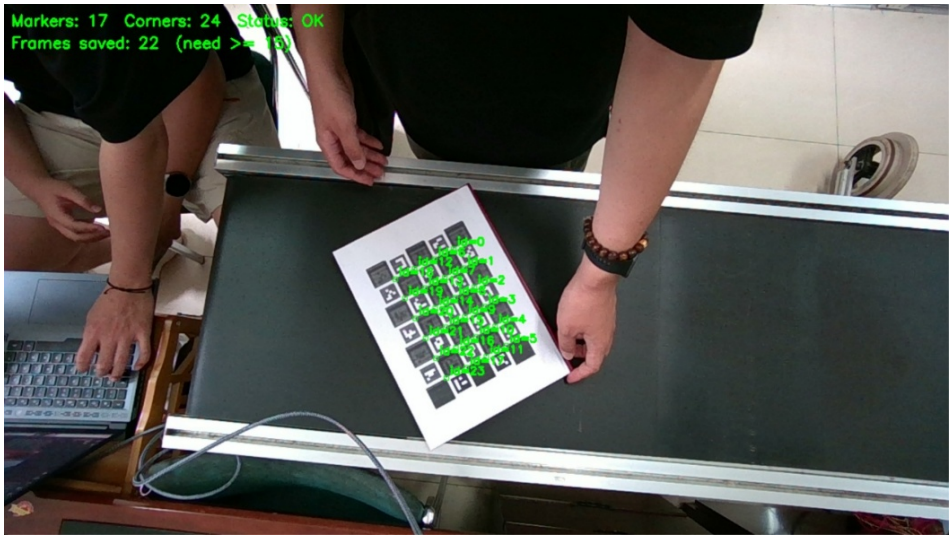


Figure 4. RGB-D camera intrinsic calibration using the ChArUco calibration board.

The obtained intrinsic parameters are listed in Table 1.

Table 1. Intrinsic parameters of the RGB-D camera.

Parameter	Value
f_x	649.284
f_y	647.722
c_x	644.706
c_y	364.784

The distortion coefficients $(k_1, k_2, p_1, p_2, k_3)$ were estimated using the rational distortion model, yielding $[-0.111630,\ 0.314275,\ -0.003405,\ 0.005280,\ -0.396332]$.

The calibration accuracy was evaluated using the reprojection error. The average reprojection error over all calibration images was 0.0471 pixels, with the minimum and maximum errors of 0.0337 pixels and 0.0815 pixels, respectively, indicating accurate camera parameter estimation.

2.5.3. Dataset Construction

To improve dataset diversity, data acquisition considered several factors affecting visual perception in coal gangue sorting scenarios, including illumination variations, object spatial distributions, and coal gangue mixing ratios. Regarding illumination, the dataset includes normal lighting, low-light, and uneven illumination conditions to simulate different visual environments. Regarding object distribution, samples were collected under isolated, stacked, and densely contacted conditions to cover various levels of occlusion and

boundary ambiguity. In addition, different coal gangue mixing ratios were configured to improve the diversity of the dataset. The size range of the fragments in our experiments is approximately 30 mm to 300 mm. A total of 5000 RGB-D image pairs were collected and divided into training, validation, and testing sets with a ratio of 7:2:1.

For object detection, each RGB image was annotated with object categories and bounding boxes. Two categories were considered: coal and gangue. The bounding boxes were defined as the minimum enclosing rectangles of individual objects.

For instance segmentation, a zero-shot segmentation-assisted semi-automatic annotation strategy was adopted. Initial instance masks were generated using the zero-shot segmentation model with detection boxes as prompts, followed by manual inspection and boundary refinement. This strategy improves annotation efficiency while maintaining sufficient mask accuracy for subsequent evaluation and reconstruction.

To obtain the ground truth of 3D geometric parameters, physical measurements were performed. The centroid position $O(x, y, z)$ was determined through spatial measurements, where the depth coordinate was obtained by distance measurement and the horizontal coordinates were determined using a reference grid. The OBB dimensions (l, w, h) were measured using a vernier caliper. The true volume V was measured using the water displacement method and used as the reference value for volume estimation evaluation. The principal orientation θ was obtained by measuring the angle between a reference line marked on the object surface and the horizontal direction.

2.6. Evaluation Metrics

To comprehensively evaluate the proposed framework, four groups of evaluation metrics were considered, including object detection, instance segmentation, 3D geometric estimation, and computational efficiency.

(1) Detection Performance

The object detection performance was evaluated using mean average precision (mAP), precision, and recall,

$$mAP = \frac{1}{C}\sum_{c=1}^{C} AP_c \tag{18}$$

where C denotes the number of object categories and AP_c represents the average precision of category c. The IoU threshold was set to 0.5.

(2) Segmentation Performance

The quality of instance masks was evaluated using mIoU, pixel accuracy (PA), and Dice coefficient,

$$mIoU = \frac{1}{C}\sum_{c=1}^{C} \frac{TP_c}{TP_c + FP_c + FN_c} \tag{19}$$

$$PA = \frac{\sum_i p_{ii}}{\sum_i \sum_j p_{ij}} \tag{20}$$

where TP_c, FP_c, and FN_c denote the numbers of true positive, false positive, and false negative pixels of class c, respectively, and p_{ij} denotes the number of pixels of class i classified as class j. All metrics are computed over all labeled pixels in the test set, excluding unlabeled or ignored regions such as conveyor belt borders and out-of-scene areas.

$$Dice = \frac{2|A \cap B|}{|A| + |B|} \tag{21}$$

The Dice coefficient is complementary to IoU and is more sensitive to segmentation quality for small targets; it is widely used in medical image segmentation. To ensure unambiguous ground-truth labels for overlapping gangue lumps, our annotation protocol

follows a depth-ordered assignment rule: each pixel is assigned to exactly one instance. For overlapping or touching instances, boundary pixels are assigned to the topmost (closest to the camera, i.e., with the smallest depth value) instance based on depth ordering. This rule is consistent with physical visibility in the scene. During evaluation, predicted masks are compared against these depth-ordered ground-truth masks using standard IoU computation.

(3) 3D Geometric Estimation Accuracy

The accuracy of estimated geometric parameters was evaluated using four metrics:

- Position error: The Euclidean distance between the estimated and ground-truth centroid positions.
- Dimension error:

$$E_d = \frac{|d_{\text{est}} - d_{\text{gt}}|}{d_{\text{gt}}} \times 100\% \tag{22}$$

  where d represents l, h, and w, and the final dimension error is calculated as the average error of the three directions.
- Volume error:

$$E_v = \frac{|V_{\text{est}} - V_{\text{gt}}|}{V_{\text{gt}}} \times 100\% \tag{23}$$

- Orientation error: The angular difference between the estimated principal orientation and the reference orientation.

(4) Computational Performance

The computational efficiency was evaluated using the average inference speed (FPS) of the complete RGB-D cascade pipeline. The inference time of each module was also recorded.

3. Results

3.1. Instance Segmentation Results

The segmentation module is an important component of the proposed RGB-D cascade framework, connecting semantic perception with subsequent 3D geometric reconstruction. The detection results provide spatial priors for segmentation, while the generated instance masks are further used to constrain depth extraction. This section evaluates the proposed box-prompted zero-shot segmentation method through quantitative metrics and qualitative visualization.

3.1.1. Experimental Settings

Experiments were conducted on the test split of the dataset described in Section 2.5, which covers different illumination conditions and spatial configurations (single-lump flat placement, multi-lump stacking, and lump contact). Coal and gangue samples were obtained from the Tongxin Mine in Datong, Shanxi Province.

3.1.2. Quantitative Comparison

Two fully supervised instance segmentation methods, YOLO-Seg and Mask R-CNN, were selected as baselines. Evaluation metrics include mean Intersection over Union (mIoU), pixel accuracy (PA), and Dice coefficient. All methods were evaluated on the same test set, with the results shown in Table 2.

Without requiring pixel-level annotation training on coal gangue data, the proposed method achieved an mIoU of 96.69% and a Dice coefficient of 98.32%, outperforming YOLO-Seg and Mask R-CNN by 7.72 and 18.27 percentage points, respectively. All three methods achieved pixel accuracy above 99.6%, with the main difference observed in boundary region

segmentation accuracy. Since the segmentation masks directly serve as spatial constraints for depth extraction in the subsequent stage, the differences in mask boundary accuracy propagate to the point cloud reconstruction phase—higher boundary precision corresponds to fewer background depth values included in the reconstructed point cloud.

Table 2. Performance comparison of different segmentation methods on the coal gangue dataset.

Method	Type	Requires Coal Gangue Training	mIoU (%)	PA (%)	Dice (%)
Ours (box-prompted zero-shot)	Zero-shot	No	96.69	99.94	98.32
YOLO-Seg	Fully supervised	Yes	88.97	99.80	94.16
Mask R-CNN	Fully supervised	Yes	78.42	99.69	87.91

3.1.3. Visual Analysis

Representative scenes were selected for visual comparison of segmentation results across methods (Figure 5).

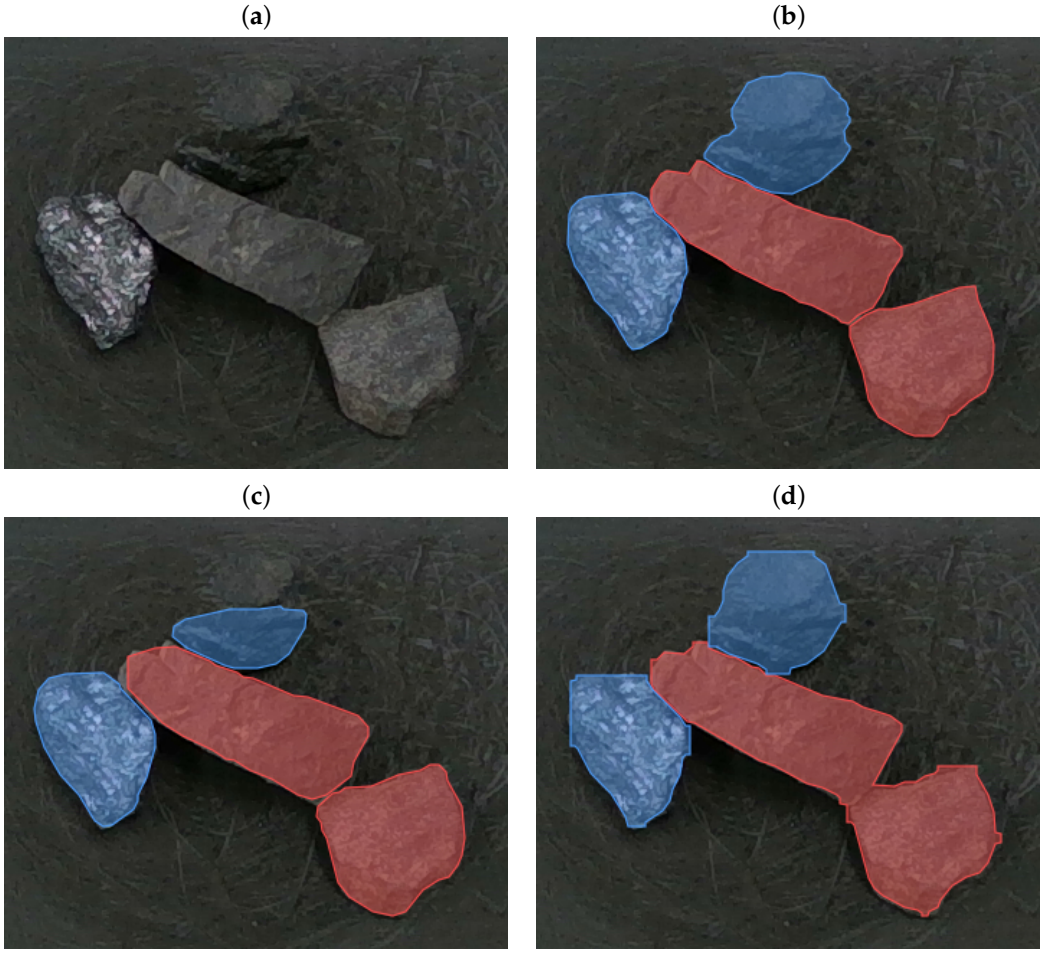


Figure 5. Visualization comparison of instance segmentation results for coal and gangue. Red indicates the segmented gangue regions, and blue indicates the segmented coal regions. (**a**) RGB image; (**b**) proposed method; (**c**) Mask R-CNN; (**d**) YOLO-Seg.

As illustrated in Figure 5, Mask R-CNN suffered from missed detections and incomplete target coverage, producing under-segmented masks that failed to enclose the full target contour in some irregular lump regions. YOLO-Seg achieved detection coverage with no missed targets, but exhibited unsmooth boundary predictions with noticeable deviations from the ground-truth edges, particularly on irregular lump contours; additionally, adjacent touching lumps were incorrectly merged into a single instance and could not be properly separated. In contrast, the proposed method generates instance masks that closely conform to the actual target contours, accurately preserving the irregular boundary characteristics of coal gangue lumps.

3.2. 3D Geometric Parameter Estimation Results

Each object was characterized by an oriented bounding box (OBB) with centroid position O, dimensions (l, w, h), and principal orientation angle θ, together with the volume estimated from the reconstructed point cloud. Since the depth camera captures only the visible surface of the objects, the reconstructed point cloud is regarded as a partial point cloud, and different evaluation strategies were therefore adopted for different geometric parameters.

3.2.1. Centroid Localization Accuracy

The centroid of the reconstructed point cloud was evaluated against the manually measured ground-truth position for each fragment, and the detailed results are summarized in Table 3.

Table 3. Quantitative evaluation of centroid localization accuracy.

Object ID	ΔX (mm)	ΔY (mm)	ΔZ (mm)	Position Error (mm)
0166 Left	0	3	18	18.248
0176 Left	0	3	14	14.318
0348 Bottom	0	2	14	14.1428
400 Right	0	1	6	6.083
501 Top	0	0	2	2.000
543 Right	0	0	3	3.000
657 Right	0	5	29	29.428
735 Left	1	5	25	25.515
828 Top	0	2	10	10.198
1081 Right	0	3	19	19.235
Average	0.10	2.40	14.00	14.22

3.2.2. OBB Dimension Estimation Accuracy

Since manual measurement cannot provide the true value of the complete three-dimensional size, the maximum manually measured length and width are used as reference values, and the size accuracy is evaluated using the point cloud OBB estimation results, as listed in Table 4. The length error and width error are defined as follows:

$$\text{Error}_L = \frac{|L_{\text{est}} - L_{\text{gt}}|}{L_{\text{gt}}} \times 100\% \tag{24}$$

$$\text{Error}_W = \frac{|W_{\text{est}} - W_{\text{gt}}|}{W_{\text{gt}}} \times 100\% \tag{25}$$

Table 4. Accuracy evaluation of OBB dimension estimation based on partial point clouds.

Object ID	Category	Measured L (mm)	Estimated L (mm)	Length Error (%)	Measured W (mm)	Estimated W (mm)	Width Error (%)
0166 Left	coal	100.7	102.4	1.69	63.6	76.9	20.91
0176 Left	gangue	116.86	115.9	0.82	70.9	76.2	7.48
0348 Bottom	gangue	164.2	156.6	4.63	75.1	68.2	9.19
400 Right	coal	105.4	91.4	13.28	49.69	52.9	6.46
501 Top	gangue	104.9	106.6	1.62	30.94	32.6	5.37
543 Right	gangue	104.7	100.7	3.82	37.6	37.8	0.53
657 Right	coal	198.45	168.5	15.09	98	105.3	7.45
735 Left	gangue	165	151.9	7.94	118.5	117.2	1.1
828 Top	gangue	208.5	216.3	3.74	31.89	54.2	69.96
1081 Right	gangue	164.4	165	0.36	130.43	123.5	5.31
Average				5.3			13.38

3.2.3. Volume Estimation Accuracy

To evaluate the volume estimation method, ten coal gangue fragments with different sizes and shapes were selected, and the estimated visible volume was compared with the physical volume measured by water displacement, as shown in Table 5.

Table 5. Comparison between estimated visible volume and physical volume measured by water displacement.

Object ID	Water Displacement Volume (mL)	Estimated Visible Volume (mL)	Error (%)
1	183	185.2	1.2
2	102	108.6	6.5
3	153	150.7	−1.5
4	27	28.8	6.8
5	44	51.5	17.1
6	21	29.1	38.5
7	75	85.9	14.6
8	50	60.4	20.7
9	505	514.6	1.9
10	67	63.9	−4.6
Average	—	—	11.3

The results demonstrate that the proposed height integration-based visible volume estimation method achieves good agreement with the water displacement measurements, with a mean relative error (MRE) of 11.3% and a mean absolute error (MAE) of 6.2 cm^3.

To illustrate the height distribution representation used for visible volume estimation, Figure 6 presents the height maps of several representative coal gangue fragments. The color variation indicates the height information of the reconstructed point cloud projected onto the XY plane. The height value of each occupied grid cell is integrated to calculate the local volume contribution, and the accumulated value is used as the estimated visible volume of each fragment.

Furthermore, linear regression analysis between the estimated visible volume and the water displacement volume demonstrated a strong linear correlation, with an R^2 value of 0.9988. This result indicates that although the complete object geometry cannot be fully recovered from partial point clouds, the height integration-based visible volume can reliably characterize the relative size variation in coal gangue blocks and serve as an effective geometric descriptor for intelligent sorting applications.

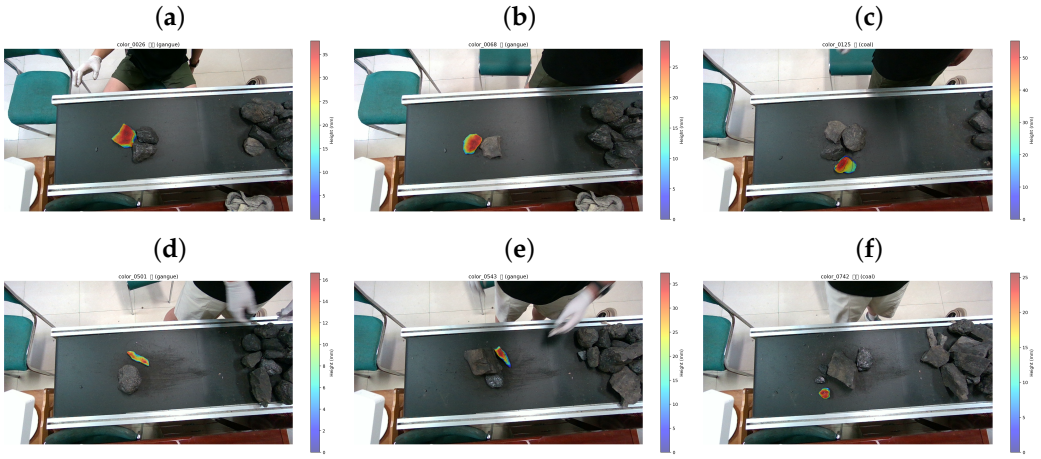


Figure 6. Depth heatmaps projected onto RGB images. (**a**–**f**) correspond to samples 0026_left_top, 0068_left, 0125_bottom, 0501_top, 0543_right, and 0742_left_bottom, respectively. The estimated depth of each target is projected onto the RGB image, with warmer colors indicating greater depth values.

For the target application of robotic sorting, the typical fragment sizes range from 30 mm to 300 mm. The key accuracy requirements are:

(1) Centroid localization: The average position error of 14.22 mm is well within the gripping tolerance of industrial sorting robots (typically ±20–30 mm), providing sufficient spatial accuracy for robotic picking or targeted removal.

(2) Dimension estimation: The average dimension error of 9.34% (length: 5.3%, width: 13.38%) is acceptable for size-based classification, as industrial sorting systems typically use size bins of 25–50 mm intervals.

(3) Volume estimation: The average volume error of 11.3% (MAE = 6.2 cm^3) provides sufficient resolution for mass estimation and quality assessment of individual fragments.

3.2.4. Principal Orientation Estimation Accuracy

As described in Section 2.4.3, the first principal component obtained by PCA is used as the dominant orientation of each fragment, whose reliability is judged by the aspect ratio indicator $R = \lambda_1/\lambda_2$. Figure 7 illustrates the projected principal directions of reconstructed point clouds on the horizontal plane.

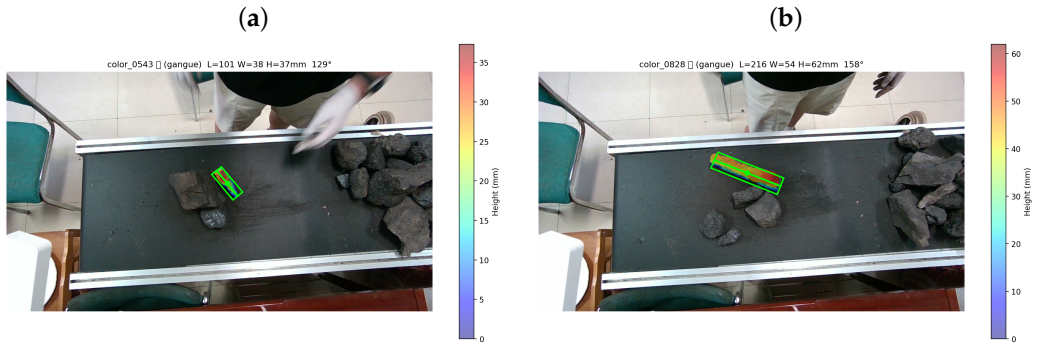


Figure 7. Estimated principal axes projected onto RGB images. (**a**) Sample 0543_right; (**b**) Sample 0826_top. The green line indicates the estimated principal orientation (major axis) of each fragment, projected from the reconstructed point cloud onto the RGB image.

The projected results show stable principal axes for elongated fragments and more scattered directions for approximately isotropic ones. Since quantitative orientation references are difficult to obtain for irregular lumps, the principal orientation is evaluated qualitatively: the reliability of each estimated orientation is indicated by the aspect ratio

indicator R, where elongated fragments ($R > 1.5$) provide reliable principal axes while approximately isotropic fragments ($R \leq 1.5$) yield low-confidence orientations.

3.3. Comparative Ablation Study of Semantic-Guided Geometric Estimation

To evaluate the contribution of semantic constraints to three-dimensional geometric estimation, an ablation study was conducted by comparing three different perception strategies:

(1) 2D object detection

The detection model was used only for semantic classification and localization. Since no depth-based geometric reconstruction was performed, this method could not provide three-dimensional geometric parameters.

(2) Detection box-guided depth reconstruction

The detected bounding boxes were directly used to extract depth information from the corresponding image regions. The extracted point clouds were then used for centroid, OBB dimension, and volume estimation.

(3) Proposed cascaded framework

The proposed method first obtains object-level segmentation masks through the cascaded detection–segmentation strategy and then uses the masks to constrain depth point extraction and three-dimensional geometric estimation.

The quantitative comparison results are presented in Table 6.

Table 6. Comprehensive comparison of multiple methods across multiple metrics. ✓ indicates the capability is provided; × indicates it is not; N/A indicates not applicable.

Method	Semantic Classification	3D Geometry	Coal Gangue Training Required	Position Error (mm)	Dimension Error (%)	Volume Error (%)
2D object detection	✓	×	Yes	N/A	N/A	N/A
Detection box + depth image	✓	✓	Yes	69.66	77.15	536
Proposed cascaded framework	✓	✓	Partial	14.22	9.34	11.3

The 2D detection method provides semantic information but cannot satisfy the requirements of three-dimensional object characterization, while direct depth extraction from detection boxes yields three-dimensional parameters with large estimation errors (position error of 69.66 mm, dimension error of 77.15%, and volume error of 536%). In contrast, the proposed cascaded framework achieved the lowest errors among all evaluated methods (position error of 14.22 mm, dimension error of 9.34%, and volume error of 11.3%), demonstrating that introducing pixel-level semantic information before geometric reconstruction significantly improves the accuracy of three-dimensional parameter estimation for irregular coal and gangue fragments.

3.4. Computational Performance

To evaluate the computational efficiency of the proposed RGB-D cascade framework, the average inference time of each module was measured, including object detection, zero-shot segmentation, and point cloud reconstruction with geometric estimation. The computational results are summarized in Table 7.

Table 7. Inference time analysis of different modules.

Module	Average Time (ms)	Percentage (%)
Stage 1: Object detection	47.7	16.9
Stage 2: Zero-shot segmentation	140.3	49.6
Stage 3: Point cloud reconstruction and geometric estimation	94.9	33.5
Total	282.9	100
FPS	3.5	—

The complete cascade pipeline requires an average processing time of 282.9 ms per frame, corresponding to an inference speed of 3.5 FPS. Among all modules, the zero-shot segmentation stage introduces the largest computational cost, accounting for 49.6% of the total inference time. The point cloud reconstruction and geometric estimation stage contributes 33.5%, while the object detection stage accounts for only 16.9%.

An additional ablation test without the segmentation module increased the processing speed to approximately 7 FPS, indicating that segmentation is the primary computational bottleneck of the proposed framework. However, removing the segmentation stage resulted in significant degradation of geometric estimation accuracy, demonstrating the importance of semantic mask constraints for reliable three-dimensional reconstruction. Therefore, the proposed framework achieves a balance between computational efficiency and geometric estimation accuracy.

4. Discussion

4.1. Contribution and Potential Applications for Intelligent Coal and Gangue Separation

Unlike conventional image-based recognition methods that only provide category information, the proposed framework estimates the spatial position, oriented dimensions, volume, and principal orientation of individual lumps, providing the object-level geometric information required for selective separation rather than bulk processing. These parameters can directly support robotic grasping and targeted removal of gangue lumps, serving as a perception module that bridges the gap between recognition and physical separation execution.

The advantages of the proposed method for intelligent coal gangue separation can be summarized in three aspects. First, improved separation selectivity: segmentation-guided reconstruction isolates individual fragments from complex scenes, avoiding the background interference suffered by bounding-box-based approaches. Second, improved resource utilization: accurate dimension and volume estimation allows valuable lumpy coal to be distinguished from gangue, reducing unnecessary removal. Third, enhanced automation capability: the estimated centroid and principal orientation directly support robotic manipulation and grasp planning, providing a potential perception solution for integrating visual sensing with automated coal gangue separation systems.

4.2. Role of Semantic Segmentation in Improving Separation Reliability

Lumps in practical separation environments often exhibit irregular boundaries, random orientations, and spatial overlap, which makes reliable identification and localization of individual lumps a prerequisite for accurate separation. The ablation study (Table 6) explains why pixel-level segmentation matters for geometry: when depth was extracted directly from detection boxes, the box interiors inevitably contained the conveyor surface, neighboring lumps, and dust, which dominated the reconstructed point cloud and inflated the position error to 69.66 mm and the volume error to 536%. Restricting depth extraction to the instance masks excludes these background points before 3D projection, reducing the position error to 14.22 mm and the volume error to 11.3%. More accurate object-level

geometry also reduces the probability of misclassification or incorrect manipulation, which is critical for separation efficiency and product quality.

4.3. Analysis of Three-Dimensional Geometric Estimation Performance

(1) Centroid localization

The average centroid localization error obtained from ten tested lumps was 14.22 mm, which provides sufficient spatial information for lump-level separation operations, especially for robotic picking or targeted removal of gangue lumps. As shown in Table 3, the depth error (ΔZ = 14.00 mm on average) dominates the position error, whereas the lateral errors are small (ΔX = 0.10 mm and ΔY = 2.40 mm). Although the proposed method achieves accurate 3D geometric estimation, the Z-direction error is relatively larger than the X/Y localization error. This phenomenon is mainly attributed to the characteristics of RGB-D sensing, where depth measurements are directly affected by sensor noise, object surface reflectivity, and measurement distance. In coal gangue separation scenarios, the dark and heterogeneous surfaces of fragments may further increase depth uncertainty. Future improvements could incorporate depth calibration, multi-view RGB-D fusion, and depth refinement algorithms to enhance the robustness of spatial measurement.

(2) Dimension estimation

Table 4 shows that the width error (13.38%) is substantially larger than the length error (5.3%). This is because the width direction corresponds to the second principal axis, whose orientation becomes unstable for nearly isotropic fragments, and because the visible extent along the width direction is more affected by occlusion and boundary uncertainty in stacked scenes. Nevertheless, the estimated dimensions support lump size classification and selective separation strategies, providing more comprehensive geometric information than two-dimensional image measurements, especially for lumps with arbitrary orientations.

(3) Volume estimation

Volume is an important factor affecting separation efficiency because coal and gangue lumps with similar appearance may exhibit different physical properties. The error distribution in Table 5 shows that small, low fragments suffer the largest relative errors (up to 38.5%), because the depth uncertainty and the reference plane estimation error are relatively large compared with the object height, while for large fragments the error falls below 7%. Although incomplete surface information introduces uncertainty, the strong linear correlation (R^2 = 0.9988) between estimated and measured volume indicates that the visible volume is a consistent relative size descriptor, providing potential support for volume-based lump characterization and intelligent separation decision-making.

4.4. Limitations and Future Perspectives

Although the proposed framework achieves promising performance in three-dimensional characterization of coal and gangue fragments, several limitations remain.

First, the current system relies on single-view depth acquisition, resulting in partial point clouds rather than complete object reconstruction. The missing information from invisible surfaces may introduce uncertainty in volume estimation and orientation determination. Future studies will investigate multi-view depth fusion strategies, such as multiple camera configurations above conveyor systems or multi-frame reconstruction during object movement, to improve geometric completeness. Specifically, deploying multiple RGB-D cameras from different viewpoints (e.g., top-view and side-view configurations) could capture previously occluded surfaces and reduce depth uncertainty at object boundaries. While this would increase hardware costs, the improvement in geometric completeness may justify the investment for high-precision sorting applications. Alternatively, a single 360° camera or rotating mirror mechanism could provide omnidirectional coverage, though

such configurations require more complex calibration and synchronization. Compared with RGB-D cameras, LiDAR provides higher geometric accuracy and longer sensing range. However, for short-range coal gangue sorting tasks, RGB-D sensors offer advantages in cost, compactness, and simultaneous semantic and depth acquisition. Future studies may explore LiDAR-RGB-D fusion to further improve geometric perception in complex industrial environments.

Second, the principal orientation estimation depends on the geometric characteristics of individual lumps. For elongated objects, the major axis can be reliably determined, whereas approximately isotropic lumps may exhibit larger orientation uncertainty. Future work will explore adaptive confidence evaluation based on geometric features, such as surface curvature and normal consistency, and investigate learning-based methods for orientation estimation with uncertainty prediction.

Third, practical deployment in mining environments requires further engineering optimization. Dust, humidity, vibration, and illumination variations may affect depth acquisition and model robustness. Future work should focus on developing protective hardware structures, online calibration strategies, and environmental-adaptive perception models to improve long-term operational stability.

Fourth, the current framework operates at 3.5 FPS, which limits direct deployment in high-speed industrial sorting lines. Future optimization will focus on lightweight segmentation networks, accelerated inference frameworks, and edge-computing solutions to achieve a better balance between accuracy and computational efficiency. In addition, the processing speed can be significantly improved through hardware-level parallelization. For instance, deploying multiple cameras in series—each connected to a separate processing unit—could enable simultaneous acquisition and analysis of multiple objects on the conveyor belt. This distributed computing architecture would scale throughput linearly with the number of camera-PC pairs, potentially achieving processing rates of 10–20 FPS or higher with 3–6 parallel units. Such a configuration would be particularly suitable for industrial sorting lines requiring 5–10 FPS throughput.

Finally, integrating the proposed perception framework with robotic sorting systems will be an important future direction. These geometric parameters can provide direct inputs for robotic grasp planning, enabling a complete perception–decision–execution loop. The practical effectiveness of the integrated system can be further evaluated using sorting accuracy, grasping success rate, and processing throughput.

5. Conclusions

This study proposed an RGB-D cascaded fusion framework for three-dimensional geometric estimation of coal and gangue fragments in intelligent sorting scenarios. The framework adopts a cascaded detection–segmentation–reconstruction pipeline, in which the detection stage extracts object categories and coarse locations from RGB images, the segmentation stage generates pixel-level instance masks under detection guidance, and the masks serve as spatial constraints for depth extraction and local point cloud reconstruction. The core contribution is a segmentation mask-based semantic–geometric bridge that transfers semantic information progressively from image space to geometric space (bounding box level $\rightarrow$ pixel level $\rightarrow$ spatial domain), allowing the segmentation results to serve not only as object recognition information but also as effective spatial constraints for three-dimensional reconstruction.

Validated on a self-collected RGB-D coal gangue dataset, the proposed method achieved a centroid localization error of 14.22 mm, a dimension estimation error of 9.34%, and a volume estimation error of 11.3%, significantly outperforming direct depth extraction based on detection bounding boxes. Unlike CAD-model-based six-degree-of-freedom

pose estimation, the proposed approach requires no predefined object models and directly extracts geometric characteristics from observed point clouds, making it suitable for irregular objects with complex morphology. Compared with conventional two-dimensional vision-based sorting methods, it provides the additional spatial and geometric information required for intelligent separation and robotic manipulation.

Although the current system still faces limitations in geometric completeness and computational efficiency, it demonstrates the feasibility of RGB-D-based geometric perception for intelligent coal gangue separation. Future work will focus on multi-view depth fusion, adaptive principal orientation estimation, deployment optimization for harsh mining environments, and integration with robotic sorting systems to establish a complete perception–decision–execution pipeline.

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

Funding: This research was funded by the R&D Program of Beijing Municipal Education Commission, grant number KM202311417004 and KM202311417003, and the Academic Research Projects of Beijing Union University, grant number ZK20202214. The APC was funded by the National Science and Technology Major Project of China, grant number 2025ZD1008205.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data is contained within the article.

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

References

1. National Bureau of Statistics of China. *Statistical Communique of China on the 2025 Economic and Social Development*; National Bureau of Statistics of China: Beijing, China, 2026.
2. Snehasree, N.; Nuruddin, M.; Moghal, A.A.B. Critical Appraisal of Coal Gangue and Activated Coal Gangue for Sustainable Engineering Applications. *Appl. Sci.* **2025**, *15*, 9649. [CrossRef]
3. Xu, Z.; Lü, Z.; Wang, W.; Zhang, K.; Lü, H. Machine Vision Recognition Method and Optimization for Intelligent Separation of Coal and Gangue. *J. China Coal Soc.* **2020**, *45*, 2207–2216.
4. Pu, Y.; Apel, D.B.; Szmeret, Y.; Wang, Z.; Zhang, L. Coal and Gangue Separating Robot System Based on Computer Vision. *Sensors* **2021**, *21*, 1349. [CrossRef]
5. Wen, X.; Li, B.; Wang, X.; Li, J.; Wei, D.; Gao, J.; Zhang, J. A Swin Transformer-Functionalized Lightweight YOLOv5s for Real-Time Coal–Gangue Detection. *J. Real-Time Image Process.* **2023**, *20*, 47. [CrossRef]
6. He, L.; Guo, Y.; Zhi, Y.; Wang, S.; Li, D.; Hu, K.; Cheng, G. Accurate Identification Method of Coal and Gangue Based on Geometric Feature Constraints by DE-XRT. *Coal Sci. Technol.* **2024**, *52*, 262–275. [CrossRef]
7. Wang, Z.; Wang, J.; Li, L.; An, B. Study on the Prediction of Gangue Content Rate in Fully Mechanized Caving Face Based on DeepLab v3+. *J. Mine Autom.* **2024**, *50*, 90–96. [CrossRef]
8. He, L.; Wang, X.; Guo, Y.; Wang, S.; Cheng, G. Rapid Detection of Incomplete Coal and Gangue Based on Improved PSPNet. *Measurement* **2022**, *201*, 111646. [CrossRef]
9. Yuan, Y.; Qin, Z.; Xia, Y.; Wu, R.; Li, L.; Li, Y. Coal Gangue Image Segmentation Model and Coal Drawing Control Technology Based on Improved U-Net. *J. China Coal Soc.* **2025**, *50*, 2722–2738. [CrossRef]
10. Guo, Y.; Yang, Y.; Li, D.; Wang, S.; He, L.; Luo, Q. Real-Time Segmentation Method of Coal Gangue Based on Adaptive Low Illumination Image Enhancement. *Coal Sci. Technol.* **2025**, *53*, 1–19. [CrossRef]
11. Wang, Y.; Tao, W.; Chen, K.; Meng, X.; Zhang, Y. Coal Gangue Recognition Technology Based on SegFormer-CG. *J. China Coal Soc.* **2026**, *51*, 2771–2782. [CrossRef]

12. Cao, Z.; Li, Z.; Fang, L.; Li, J.; Yang, H.; Hui, D. Research on Coal and Gangue Segmentation Based on MFCCM-Mask R-CNN. *Energy Sci. Eng.* **2024**, *12*, 2958–2973. [CrossRef]
13. Lai, W.; Hu, F.; Kong, X.; Yan, P.; Bian, K.; Dai, X. The Study of Coal Gangue Segmentation for Location and Shape Predicts Based on Multispectral and Improved Mask R-CNN. *Powder Technol.* **2022**, *407*, 117655. [CrossRef]
14. Zhu, L.; Gu, W.; Liu, C.; Zhang, B.; Liu, W.; Yuan, C. Image Segmentation Algorithm Based on Improved YOLOv8 Model and Its Application in Underground Coal and Gangue Recognition. *PLoS ONE* **2025**, *20*, e0321249. [CrossRef]
15. Bai, H.; Liu, H.; Li, Y.; Liu, C. A Dual-Stage Adaptive Segmentation Framework for Coal Particle Images. *Ind. Mine Autom.* **2025**, *51*, 102–108. [CrossRef]
16. Dong, L.; Song, W.; Fu, L. Dynamic Coal Quantity Measurement Method Based on Binocular Vision. *Coal Sci. Technol.* **2022**, *50*, 196–203. [CrossRef]
17. Wang, W.; Zhang, C. Separating Coal and Gangue Using Three-Dimensional Laser Scanning. *Int. J. Miner. Process.* **2017**, *169*, 79–84. [CrossRef]
18. He, J.; Wang, G.; Liu, J.; Yang, L. Coal Volume Measurement on Belt Conveyor Based on Image Processing. *Acta Metrol. Sin.* **2020**, *41*, 1516–1520.
19. Wu, J. Application of 3D Laser Scanning Technology in Earthwork Survey. *Beijing Surv. Mapp.* **2019**, *33*, 1344–1347. [CrossRef]
20. He, Y.; Sun, W.; Huang, H.; Liu, J.; Fan, H.; Sun, J. PVN3D: A Deep Point-Wise 3D Keypoints Voting Network for 6DoF Pose Estimation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2020; pp. 11632–11641.
21. He, Y.; Huang, H.; Fan, H.; Chen, Q.; Sun, J. FFB6D: A Full Flow Bidirectional Fusion Network for 6D Pose Estimation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2021; pp. 3003–3013.
22. Wang, C.; Xu, D.; Zhu, Y.; Martin-Martin, R.; Lu, C.; Fei-Fei, L.; Savarese, S. DenseFusion: 6D Object Pose Estimation by Iterative Dense Fusion. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2019; pp. 3343–3352.
23. Wen, B.; Yang, W.; Kautz, J.; Birchfield, S. FoundationPose: Unified 6D Pose Estimation and Tracking of Novel Objects. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2024; pp. 17868–17879.
24. Cai, D.; Heikkilä, J.; Rahtu, E. GS-Pose: Generalizable Segmentation-Based 6D Object Pose Estimation with 3D Gaussian Splatting. *arXiv* **2024**, arXiv:2403.10683.
25. Lin, J.; Liu, L.; Lu, D.; Jia, K. SAM-6D: Segment Anything Model Meets Zero-Shot 6D Object Pose Estimation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2024; pp. 27906–27916.

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