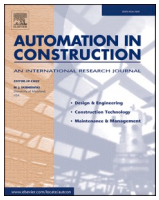
Automation in Construction 184 (2026) 106841

Contents lists available at ScienceDirect

Automation in Construction

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



Training-free open-vocabulary monocular 3D object detection for industrial assets

Masoud Kamali *, Behnam Atazadeh, Abbas Rajabifard, Yiqun Chen

The Centre for Spatial Data Infrastructures and Land Administration, Department of Infrastructure Engineering, The University of Melbourne, Victoria 3010, Australia

ARTICLE INFO

Keywords:
Monocular 3D object detection
Open-vocabulary object detection
Industrial environments
Foundation models

ABSTRACT

3D scene understanding in brownfield industrial environments plays a critical role in operation and maintenance activities. However, the absence of reliable 3D information poses significant challenges for accurate object detection. Existing 3D object detection methods rely on labelled datasets, which are scarce in industrial contexts given the complexity and diversity of assets. In addition, current methods often require depth information and camera parameters for 2D-to-3D transformation. To address these limitations, this paper proposes a training-free open-vocabulary monocular 3D object detection approach that eliminates the reliance on labelled datasets, depth information, and camera parameters. The approach integrates prompt-guided detection, depth and camera parameter estimators, and an adaptive noise filtering method for 3D bounding box prediction. To evaluate the proposed approach, experiments were conducted in an industrial environment in Australia, achieving 3D mIoU of 0.4251 for electrical kits, 0.3205 for pumps, 0.3659 for tanks, 0.2823 for valves, and 0.2217 for wires.

1. Introduction

Operation and maintenance (O&M) activities in industrial environments often impose substantial financial burdens across various systems [1]. In the United States, interoperability issues in the Architecture, Engineering, and Construction (AEC) sector result in annual inefficiencies of approximately US $15.8 billion, with around US $10.6 billion occurring during the O&M phase [2]. In water and wastewater treatment facilities, O&M expenses can account for 50–75% of total lifecycle cost of these assets, rising to 85% in some cases [3]. In process manufacturing industries, recent studies report that maintenance costs may represent 40–70% of production costs [4]. Furthermore, the adoption of advanced maintenance strategies, such as predictive maintenance, has been shown to reduce unplanned downtime by 52.7% and defects by 78.5% compared to reactive maintenance approaches [5].

Industrial environments contain a diverse range of small and large-scale assets. 3D scene understanding in these settings is critical for applications including asset management [6] and autonomous inspection [7]. Although 3D models are created for newly-designed industrial environments, most brownfield sites lack up-to-date spatial information required for effective 3D scene understanding [8]. Data capturing in these environments and identification of critical assets can improve maintenance workflows, such as remote monitoring [9]. However, the manual detection and localisation of assets within the captured data is a time-consuming and resource-intensive process [10]. Therefore, object detection in brownfield industrial sites is essential to facilitate 3D scene understanding and O&M activities.

Object detection in brownfield industrial environments can be performed using LiDAR-based [11], stereo vision-based [12], and monocular techniques [13]. However, the high cost of LiDAR sensors and the complex calibration requirements of stereo cameras limit their adoption in downstream applications [14]. In contrast, monocular 3D object detection (M3OD) aims to identify and localise objects in 3D space from a single image. Accordingly, the affordability and simple configuration of M3OD have received a significant attention in computer vision. Although numerous datasets have been employed to train M3OD models, they remain constrained to a limited range of object categories [15,16]. For instance, Omni3D [17], the largest publicly available dataset for image-based 3D object detection, contains 234 k images with annotations for over 3 million instances, yet it covers only 98 distinct categories. This limitation indicates the need to enhance the generalisation capability of M3OD methods for industrial assets. Besides, preparing training datasets for M3OD tasks in industrial assets is a labour-intensive process due to the complexity and diversity of these assets.

To address these challenges, we propose a training-free M3OD approach that eliminates the reliance on closed-set object detectors, depth information, and camera intrinsic parameters. This approach employs open-vocabulary object detection (OVOD) to enable zero-shot

* Corresponding author.
E-mail address: mkkamali@student.unimelb.edu.au (M. Kamali).

https://doi.org/10.1016/j.autcon.2026.106841
Received 5 September 2025; Received in revised form 8 February 2026; Accepted 12 February 2026
Available online 17 February 2026

object detection in industrial environments. The main contributions of this paper are summarised as follows:

- A training-free open-vocabulary monocular 3D object detection framework is introduced that does not require known camera intrinsic parameters, enabling 3D object detection from a single RGB image where camera parameters are unavailable. Pre-trained camera intrinsic estimators are applied to infer focal length and principal point from single images.
- To address the limitations of existing OVOD approaches in detecting objects with challenging visual appearances, a prompt-guided object detection strategy is proposed that integrates LLM-based asset description extraction with vision language models to improve 2D localisation accuracy in complex industrial scenes.
- Due to the absence of depth information in RGB images, pre-trained monocular depth estimation (MDE) models are utilised to predict pixel-level object distances from the camera without auxiliary inputs.
- An adaptive noise filtering method is introduced to remove noise and outliers based on asset-specific geometric characteristics, improving the accuracy of 3D bounding box prediction across a wide range of industrial assets. Additionally, principal component analysis (PCA) is employed to refine the orientation of 3D bounding boxes.

The remainder of the paper is structured as follows: Section 2 reviews the current state-of-the-art M3OD, MDE, and OVOD methods. Section 3 presents the proposed methodology for the training-free M3OD approach. The quantitative and qualitative results are discussed in Section 4. Section 5 provides an in-depth discussion of the findings and highlights the study's limitations. Finally, Section 6 concludes the paper by summarising the key contributions and outlining potential directions for future research.

2. Literature review

This section reviews recent advances in 3D object detection, M3OD methods for industrial settings, state-of-the-art MDE techniques, open-vocabulary 3D object detection, and training-free open-vocabulary monocular 3D object detection.

2.1. 3D object detection

3D object detection aims to identify and localise objects in three-dimensional space by estimating their spatial position, orientation, and dimensions. Existing approaches can be categorised based on input and modelling architectures into LiDAR-based [18,19], image-based [20,21], and multi-modal detection methods [22,23].

LiDAR-based methods utilise dense or sparse point clouds to derive precise depth and geometric information. They learn geometric representations from point cloud features to predict a 3D bounding box for each object. While these methods achieve high accuracy and robust performance in various environments, they rely on costly sensors and are limited by point sparsity. Image-based methods estimate 3D object geometry directly from visual features in RGB images without relying on explicit depth measurements. These methods are categorised into multi-view and monocular-based approaches. Multi-view methods require images captured from different viewpoints to estimate depth and predict 3D bounding box. Although this approach offers high spatial accuracy, it relies on multiple cameras and large-scale labelled training datasets, making it expensive and labour-intensive in complex industrial settings. In contrast, monocular-based methods infer 3D scene structure from a single RGB image. However, they face challenges with depth ambiguity and limited generalisation across diverse environments. Multi-modal methods integrate information from LiDAR and cameras to combine accurate geometric depth with rich visual features. Since both image and point cloud data need to be collected, these methods require complex sensor configurations and calibration procedures.

2.2. Monocular 3D object detection

M3OD has gained notable attention among image-based 3D object detection approaches due to its cost-effectiveness and straightforward setup [24]. Previous research on M3OD has largely focused on limited domains [25–28]. Besides, publicly available datasets for training learning-based often lack coverage of many objects in real-world scenarios, such as industrial assets. Shen et al. [29] utilised moving objects in construction sites (MOCS) dataset [30] to design a M3OD framework for construction scene analysis. This study integrated a Swin transformer-based cascade network with boundary-patch refinement for precise 2D detection and segmentation, and a self-supervised monocular depth estimator that infers camera intrinsics from videos with unknown parameters. Shen et al. [30] leveraged AIM dataset [31] to develop a 3D object detection framework for construction-site safety that integrates enhanced 2D detection, instance segmentation, and pseudo-LiDAR point cloud generation. Ding et al. [32] created a dataset of 22,500 virtual and real construction images with 3D bounding box annotations and proposed a M3OD framework that regresses bounding box position, size, and orientation to classify worker proximity risk for collision warning. While some datasets for industrial assets have been proposed [33,34], current M3OD approaches still rely heavily on the asset categories in their training datasets. This limitation restricts their ability to generalise to unseen asset types.

2.3. Monocular depth estimation

To address the lack of explicit depth information in monocular settings, MDE methods can be employed to generate depth maps for M3OD tasks. The main challenge of MDE is to infer depth information from a single 2D image, which is difficult due to the inherent loss of 3D spatial information. MDE plays a pivotal role in various applications, including autonomous driving [35] and 3D reconstruction [36]. MDE models are categorised into two main groups: 1) Relative depth estimation and 2) Metric depth estimation.

2.3.1. Relative depth estimation

In relative depth estimation, the depth information for each pixel is defined relative to other pixels within the image. However, the actual scale of the measurements remains unknown. MiDaS [26] introduced novel loss functions to address major incompatibilities across datasets, such as unknown and inconsistent scales and baselines. Marigold [28] proposed an affine-invariant monocular depth estimation method based on Stable Diffusion [29]. In this approach, the image and depth map were encoded into the latent space. The U-Net architecture was fine-tuned by optimising the standard diffusion objective relative to the depth latent code. Image conditioning is achieved by merging the two latent codes before feeding them into the U-Net. Relative depth estimation methods demonstrate strong generalisation to unseen scenarios. However, their inability to recover metric depth constrains their applicability in M3OD tasks that require absolute distance measurements.

2.3.2. Metric depth estimation

Metric depth estimators predict the actual distance of objects from the camera, providing depth information of each pixel with known scale units. ZoeDepth [37] employed MiDaS framework to estimate relative depth and incorporated additional heads for metric depth estimation. ZoeDepth was pre-trained on 12 datasets using relative depth and fine-tuned on two datasets for metric depth estimation. Depth Anything [31] developed a data engine to automatically generate depth annotations for unlabelled images, enabling data scaling to an arbitrary scale. Trained on 1.5 million labelled and 62 million unlabelled images, it used self-training to simultaneously learn from both labelled and unlabelled data. Depth Anything V2 [38] enhanced the Depth Anything model by

Table 1
Comparative summary of monocular 3D object detection methods with respect to input modality, training requirements, and open-vocabulary object detection capability.

Method	Input				Training-Free	Open-Vocabulary Object Detection
	RGB	Depth/PC	Camera Intrinsic	Camera Pose		
MonoDETR	✓	✖	✓	✖	✖	✖
MonoCON [49]	✓	✖	✓	✖	✖	✖
MonoWAD [50]	✓	✖	✓	✖	✖	✖
SRDDP-M3D [51]	✓	✖	✓	✖	✖	✖
AuxDepthNet [52]	✓	✖	✓	✖	✖	✖
OpenSU3D	✓	✓	✓	✓	✓	✓
FM-OV3D	✓	✓	✖	✓	✖	✓
OpenScene [53]	✓	✓	✓	✓	✖	✓
OV-Uni3DETR [54]	✓	✓	✓	✓	✖	✓
OVM3D-Det	✓	✖	✓	✖	✖	✓
OVMono3D-LIFT	✓	✖	✓	✖	✖	✓
OVMono3D-GEO	✓	✖	✓	✖	✓	✓
Ours	✓	✖	✖	✖	✓	✓

replacing labelled real images with synthetic data, scaling up the dataset with pseudo-labelled real images and using a teacher-student model framework for training. Depth Pro [39] utilised a multi-scale vision transformer (ViT) to produce metric depth maps from a single image without requiring camera intrinsics. The network extracts overlapping patches at multiple scales using shared-weight ViT encoders, merges the features, and progressively upsamples them to generate dense depth predictions. UniDepth [40] introduced a camera self-prompting mechanism and a pseudo-spherical 3D output representation to jointly infer camera parameters and metric depth from a single image. The network uses an encoder backbone to extract features, processes them through a camera module that predicts dense angular embeddings, and conditions the depth module via cross-attention. The pseudo-spherical representation disentangles camera and depth components, while a geometric invariance loss enforces consistency across different camera viewpoints. In M3OD scenarios without available depth information for RGB images, pre-trained monocular metric depth estimation models can generate scale-consistent depth maps. The depth information enables absolute distance reasoning and enhances the accuracy of 3D object localisation from a single image.

2.4. Open-vocabulary 3D object detection

OVOD refers to the detection and localisation of both seen and unseen object categories by aligning visual features with textual descriptions. The use of vision language foundation models for OVOD has led to substantial improvements in the detection of unseen objects [40]. Grounding DINO [41] is a transformer-based open-set object detector that unifies object detection and phrase grounding by aligning image regions with arbitrary text queries. This approach utilised a dual-encoder for visual and textual inputs with cross-modal attention to predict text-conditioned bounding boxes, enabling detection of seen and unseen categories. FM-OV3D [42] employed pre-trained foundation models to enhance open-vocabulary 3D detection without requiring manual annotations. For 3D localisation, it used GroundedSAM [43] to generate 2D bounding boxes from paired images, which are then projected and clustered to form 3D boxes. For object detection, it integrated GPT-3, Stable Diffusion, and CLIP by generating textual and visual prompts, and aligning them with 3D point cloud features from the detector. INHA [44] introduced an image-guided novel class discovery and hierarchical feature space alignment framework for open-vocabulary 3D object detection. For feature alignment, it aligned 3D features with vision language model features at the instance, category, and scene levels using a permutation-invariant scene feature extraction module and cross-modal contrastive losses. This study integrated both text and image guidance while leveraging 3D geometry to detect novel objects in point cloud scenes. Although these methods achieve open-vocabulary 3D object detection using foundation models, they remain dependent on point cloud data as the primary input for 3D object detection. OpenSU3D [45] introduced an open-world 3D scene understanding framework by integrating multi-modal information from foundation models. This study extracted 2D semantic features and object labels using GroundedSAM and GPT-4 V, then projected these features into 3D space through depth and camera parameters. This method enables open-vocabulary 3D scene understanding by integrating multi-modal features from foundation models. Nevertheless, it depends on RGB-D images and known camera parameters for 3D projection and feature alignment. OV-3DET [46] proposed an open-vocabulary 3D point cloud detection framework that removed the need for 3D annotations. It trained a 3D detector using pseudo-labels from 2D pre-trained detectors and aligns text, image, and point cloud features through de-biased triplet cross-modal contrastive learning. This study also relies on paired RGB images and point cloud data during training to learn robust cross-modal representations.

2.5. Training-free open-vocabulary monocular 3D object detection

To address the limitations of RGB-D images and point cloud data, several recent studies have focused on training-free open-vocabulary monocular 3D object detection. OVM3D-Det [47] proposed an open-vocabulary monocular 3D object detection framework that operates without any 3D supervision and relies solely on RGB images. It leveraged open-vocabulary 2D detectors and depth estimation to generate pseudo-LiDAR representations, enabling 3D reasoning and bounding box prediction. While the framework introduced a robust pipeline for M3OD, it assumes a consistent ground plane for orientation estimation and employs uniform size priors for bounding box validation, which cannot be applied to industrial assets. For example, assets such as electrical kits and valves are often mounted at varying elevations and orientations, including on walls or pipes rather than on a common ground plane. Furthermore, although objects such as cars, pedestrians, and sofas have relatively uniform dimensions, industrial assets, including tanks, pumps, and valves vary widely in scale and geometry. These variations in industrial settings make it less feasible to rely on standard dimension priors to assess the validity of the generated bounding boxes. OVMono3D [48] introduced two frameworks for open-vocabulary monocular 3D object detection. OVMono3D-LIFT is a learning-based method that trained a class-agnostic neural network to directly lift 2D detections into 3D space. Besides, OVMono3D-GEO is a training-free pipeline that infers 3D bounding boxes from 2D detections via geometric unprojection, leveraging pre-trained modules such as Depth Pro for depth estimation, SAM for segmentation, and Grounding DINO for open-vocabulary 2D detection. This method employed a two-stage outlier removal process. First, the point cloud has been down-sampled using random sampling. Then, DBSCAN clustering was applied iteratively with adaptive parameters to separate the main object cluster

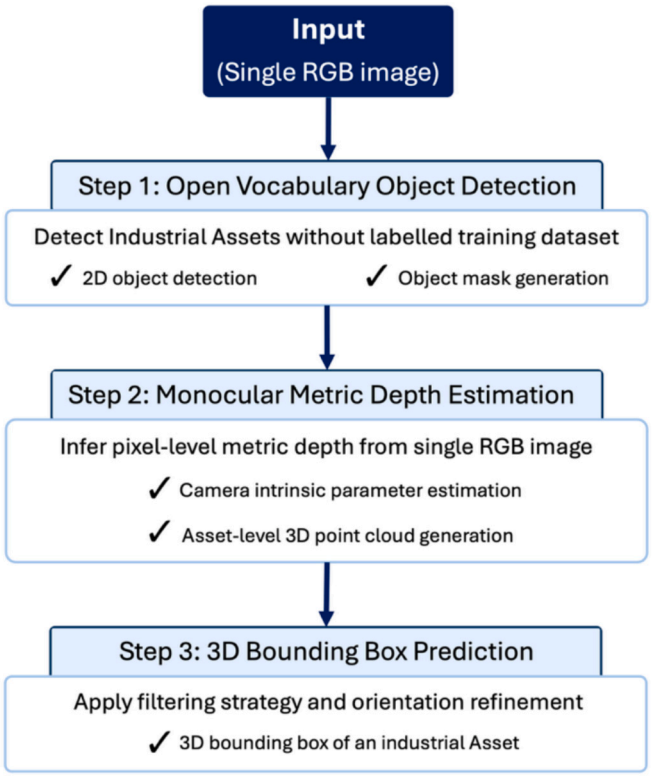


Fig. 1. High-level workflow of proposed training-free open-vocabulary monocular 3D object detection pipeline.

from noise. Although OVMono3D-GEO is the most comparable method to our approach, it assumes that camera intrinsic parameters are available for M3OD task. Moreover, its exclusive reliance on Grounding DINO as the vision–language foundation model for OVOD constrains detection accuracy, particularly for complex industrial assets. Table 1 provides a structured comparison of different 3D object detection methods across key aspects, including input modality, training requirements, and OVOD capability. The comparison indicates that most existing M3OD methods have been developed within supervised and closed-set settings, which restricts their ability to generalise to unseen object categories. Recent open-vocabulary approaches have extended category generalisation to 3D perception. However, they mostly rely on auxiliary depth or point cloud inputs and assume the availability of camera intrinsic or pose information. In addition, existing training-free M3OD methods typically rely on known camera parameters and cannot be generalised to industrial assets, thereby reducing their robustness in complex industrial environments. In this paper, we propose a M3OD method to address the limitations in 2D detection accuracy, reliance on known camera parameters, and limited generalisation to industrial environments.

3. Methodology

This study proposes a training-free open-vocabulary monocular 3D object detection pipeline for industrial assets using a single RGB image. The approach is divided into three distinct processes: (1) OVOD, (2) monocular metric depth estimation, and (3) 3D bounding box prediction. As summarised in Fig. 1, the pipeline begins with OVOD, in which industrial assets are detected and localised in 2D. Camera intrinsic parameters are then predicted to estimate metric depth information from a single RGB image. Finally, a filtering strategy is applied to asset-level point clouds for noise and outlier removal and to improve 3D bounding box prediction.

Fig. 2 illustrates the proposed methodology for training-free open-vocabulary monocular 3D object detection. In the first stage, multi-modal large language models are employed to generate structured descriptions and detect assets in the image. For each detected asset, a 2D bounding box is created. Mask generation is then applied to isolate the asset from neighbouring objects. In the second stage, monocular metric depth estimation is applied to obtain pixel-level depth values and generate 3D point clouds of each asset. In the final stage, adaptive noise filtering and orientation adjustment are applied to predict 3D bounding boxes with accurate geometry and alignment to the object's geometry.

3.1. Open-vocabulary object detection

In the initial phase of the methodology, GPT-4o [55], a multi-modal large language model, is utilised to perform asset detection within industrial scene images. The model generates structured textual descriptions, including asset types, quantities, approximate spatial positions within the images, relative sizes, and descriptions of their appearance and surroundings. This approach enables precise and consistent extraction of asset-related data directly from visual inputs without requiring labelled training datasets. A general prompt is used to identify all visible assets in the image without providing specific guidance. The geometry and visual features of the assets are not predefined in the prompt instructions to enhance the method's generalisability across diverse industrial environments and asset categories. In the next step, Grounding DINO is employed to generate precise 2D bounding

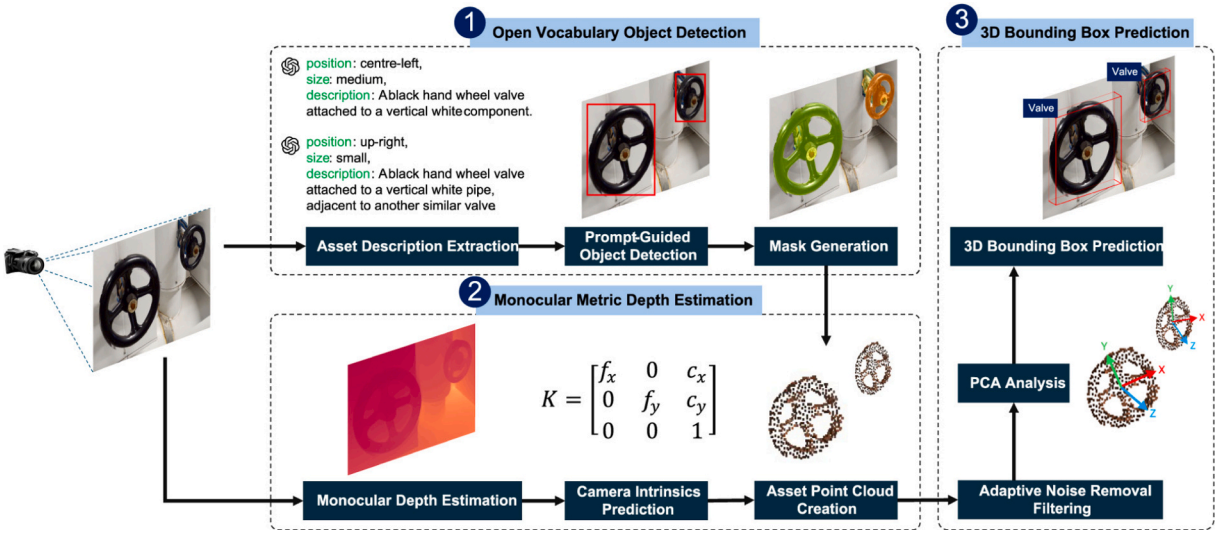


Fig. 2. Methodological workflow of proposed training-free open-vocabulary monocular 3D object detection approach for industrial assets.

boxes for industrial assets. The model uses the textual descriptions from the previous stage as prompts to detect object locations in the images. Finally, SAM is applied to generate pixel-level segmentation masks for each detected asset. Using the generated 2D bounding boxes from previous stage, SAM segments the corresponding regions in the original images to extract each individual asset. This enables precise 3D bounding box prediction by filtering out surrounding elements and focusing on the geometry of the target asset.

3.2. Monocular metric depth estimation

To obtain the depth information required for accurate 3D bounding box prediction, MDE is applied to each single image. Due to the lack of labelled depth information for training learning-based models, a pre-trained metric depth estimator is used to predict pixel-level depth values. The segmented asset regions are used as input to extract the depth of relevant assets. Additionally, camera intrinsic parameters are required to transform coordinates from image coordinate system to 3D camera coordinate system. The transformation is defined as:

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = z \begin{bmatrix} \frac{u - c_x}{f_x} \\ \frac{v - c_y}{f_y} \\ 1 \end{bmatrix} \tag{1}$$

where (u, v) are the 2D pixel coordinates and z represents the depth value at the pixel location (u, v). The 3D coordinates (X, Y, Z) are in the camera's coordinate system. The parameters c_x and c_y are the principal point offsets of the camera, while f_x and f_y correspond to the focal lengths in the horizontal and vertical directions, respectively. A pre-trained camera calibration model is employed to predict the camera intrinsic parameters. This enables 3D point cloud generation from masked images and depth information.

3.3. 3D bounding box prediction

Accurate 3D bounding box prediction depends on clean and reliable point cloud of asset. Noise preservation or original data removal may distort object dimensions and orientations. Errors in depth estimation and occlusions in complex environments frequently introduce noise. Moreover, the diverse geometries and sizes of industrial assets limit the effectiveness of conventional denoising methods such as DBSCAN, voxel grid downsampling, and radius-based filtering, as their fixed parameters often prevent generalisation across asset categories. For example, small components such as valves risk losing geometrical details under aggressive filtering, while large assets such as tanks often retain noise when thresholds are too lenient. To address this limitation, an adaptive noise filtering method is introduced. This approach analyses the geometric characteristics of each asset, including its size and point density, to adjust filtering parameters and preserve geometric fidelity during noise removal.

Voxel grid downsampling reduces the density of a 3D point cloud by dividing the space into uniform voxel grids. Within each voxel, all points are replaced with a single point as centroid, preserving the object's overall geometry while reducing computational cost (see Algorithm A). Each point $\mathbf{p}_i = (x_i, y_i, z_i)$ is assigned to a voxel indexed by $\mathbf{v}_i = \lfloor \frac{\mathbf{p}_i}{v} \rfloor = \left(\lfloor \frac{x_i}{v} \rfloor, \lfloor \frac{y_i}{v} \rfloor, \lfloor \frac{z_i}{v} \rfloor \right)$, where $\lfloor \bullet \rfloor$ denotes the element-wise floor operation. All points sharing the same voxel index $\mathbf{v}_k$ from a subset $\mathbf{P}_k = \{\mathbf{p}_i \mid \mathbf{v}_i = \mathbf{v}_k\}$, and the centroid of each voxel is computed as $\mathbf{c}_k = \frac{1}{|\mathbf{P}_k|} \sum_{\mathbf{p}_i \in \mathbf{P}_k} \mathbf{p}_i$. The final downsampled point cloud is expressed as $\mathbf{P}' = \{\mathbf{c}_k \mid k = 1, 2, \ldots, M\}$, where M is the total number of occupied voxels.

Algorithm A. VoxelGridDownsample

Algorithm A: VoxelGridDownsample

Input: $\mathbf{P} \in \mathbb{R}^{N \times 3}$(input 3D point cloud)
v (voxel size)
Output: $\mathbf{P}'$ (voxelised point cloud)

Initialise an empty hash map $\mathcal{V}$
Initialise $\mathbf{P}' \leftarrow \emptyset$
for each point $\mathbf{p}_i \in \mathbf{P}$ **do**
  Compute voxel index $\mathbf{v}_i = \left\lfloor \frac{\mathbf{p}_i}{v} \right\rfloor$
  Append $\mathbf{p}_i$ to $\mathcal{V}[\mathbf{v}_i]$
end for
for each voxel index $\mathbf{v}_k \in \mathcal{V}$ **do**
  Compute centroid $\mathbf{c}_k \leftarrow \text{mean}(\mathcal{V}[\mathbf{v}_k])$
  Append $\mathbf{c}_k$ to $\mathbf{P}'$
end for
return $\mathbf{P}'$

Distance-based filtering is a statistical outlier removal technique that identifies and removes points by measuring their distance from the centre of cluster. The algorithm first computes the median point of the cluster and calculates the Euclidean distance of each point from this centre. Points exceeding a percentile-based distance threshold are classified as noise and discarded. (see Algorithm B). Formally, let the input point cloud be $\mathbf{P} = \{\mathbf{p}_i = (x_i, y_i, z_i) \mid i = 1, 2, \ldots, N\}$. The cluster centre is defined as the component-wise median of all points, $\mathbf{p}_c = \text{median}(\mathbf{P}) = (\text{median}(x_i), \text{median}(y_i), \text{median}(z_i))$. The Euclidean distance of each point from the cluster centre is then computed as $d_i = \|\mathbf{p}_i - \mathbf{p}_c\|_2$. Let T_p denote the distance value at a specified percentile p of all distances $\{d_i\}$. The adaptive distance threshold is defined as $T = T_p \times \alpha$, where α is a distance multiplier controlling the strictness of filtering. Points satisfying $d_i \leq T$ are retained, and the remaining points are considered as outliers. The filtered point cloud is therefore given by $\mathbf{P}' = \{\mathbf{p}_i \in \mathbf{P} \mid d_i \leq T\}$.

Algorithm B. **DistanceBasedFiltering**

Algorithm B: DistanceBasedFiltering

Input: $\mathbf{P} \in \mathbb{R}^{N \times 3}$(input 3D point cloud)
α (distance multiplier)
p (distance percentile)
Output: $\mathbf{P}'$ (filtered point cloud)

Initialise $\mathbf{P}' \leftarrow \emptyset$
Initialise an empty list $\mathcal{D}$
Compute cluster centre $\mathbf{p}_c \leftarrow \text{median}(\mathbf{P})$
for each point $\mathbf{p}_i \in \mathbf{P}$ **do**
  Compute distance $d_i \leftarrow \| \mathbf{p}_i - \mathbf{p}_c \|_2$
  Append d_i to $\mathcal{D}$
end for
Compute base threshold $T_p \leftarrow \text{percentile}(\mathcal{D}, p)$
Compute distance threshold $T \leftarrow T_p \times \alpha$
for each point $\mathbf{p}_i \in \mathbf{P}$ with corresponding distance d_i **do**
  if $d_i \leq T$ **then**
    Append $\mathbf{p}_i$ to $\mathbf{P}'$
  end if
end for
return $\mathbf{P}'$

Radius-based filtering removes isolated or weakly connected points by assessing their local neighbourhood density. Each point is retained only if it has a sufficient number of neighbours within a defined radius. This method eliminates scattered noise while preserving the core geometry of the object (see Algorithm C). For each point $\mathbf{p}_i \in \mathbf{P}$, the local neighbourhood is defined as $\mathcal{N}(\mathbf{p}_i, r) = \left\{ \mathbf{p}_j \in \mathbf{P} \;\middle|\; \left\| \mathbf{p}_i - \mathbf{p}_j \right\|_2 \leq r \right\}$, where r denotes the search radius and $\| \cdot \|_2$represents the Euclidean distance. The number of neighbouring points within this radius is given by $n_i = |\mathcal{N}(\mathbf{p}_i, r)|$. A point $\mathbf{p}_i$ is retained if $n_i \geq n_{\min}$, where $n_{\min}$ is a predefined threshold specifying the minimum required neighbours. The filtered point cloud is therefore expressed as $\mathbf{P}' = \{\mathbf{p}_i \in \mathbf{P} \mid n_i \geq n_{min}\}$.

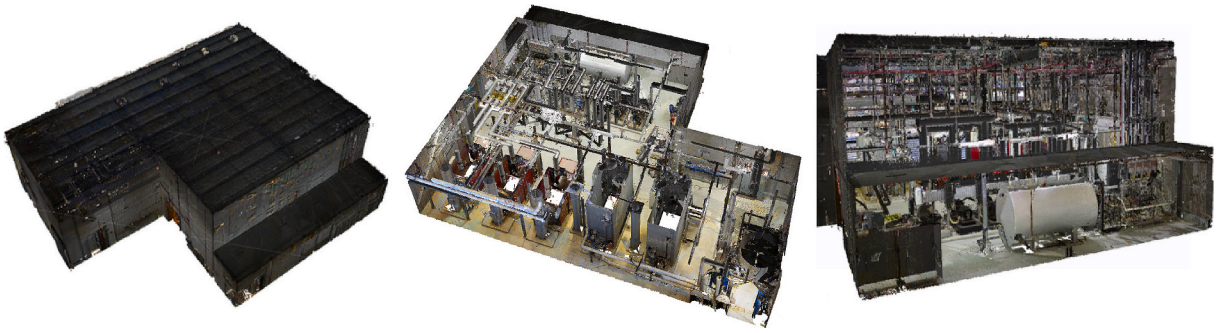


Fig. 3. Point cloud of industrial plant room.

Table 2
Distribution of captured images across different scene conditions.

Asset	Total Images	Scene Condition		
		Single-Asset	Multi-Asset	Occluded
E Kit	70	15	30	25
Pump	60	15	20	25
Tank	60	20	15	25
Valve	70	20	30	20
Wire	60	20	25	15
Total	320	90	120	110

Algorithm C. RadiusBasedFiltering

Algorithm C: RadiusBasedFiltering

Input: $\mathbf{P} \in \mathbb{R}^{N\times 3}$(input 3D point cloud)
r (search radius)
$n_{\min}$ (minimum number of neighbours)
Output: $\mathbf{P}'$ (radius-filtered point cloud)
Build a spatial index (KD-tree) over $\mathbf{P}$
Initialise $\mathbf{P}' \leftarrow \emptyset$
for each point $\mathbf{p}_i \in \mathbf{P}$ **do**
  Query neighbourhood $\mathcal{N}(\mathbf{p}_i, r)$
  Compute neighbour count $n_i \leftarrow |\mathcal{N}(\mathbf{p}_i, r)|$
  if $n_i \geq n_{\min}$ **then**
    Append $\mathbf{p}_i$ to $\mathbf{P}'$
  end if
end for
return $\mathbf{P}'$

Connectivity-based filtering uses DBSCAN to detect consistent clusters in the point cloud. It preserves the largest connected component and removes small or disconnected regions. An additional proximity validation to the cluster centre ensures the spatial integrity of the final refined result (see Algorithm D). For each point $\mathbf{p}_i \in \mathbf{P}$, DBSCAN partitions the point cloud into clusters based on two parameters: the maximum neighbourhood radius ε and the minimum number of points required to form a dense region $n_{\min}$. Points with at least $n_{\min}$ neighbours within distance ε are assigned to the same cluster, while points not satisfying this criterion are considered as noise. Let the resulting cluster labels be $\mathbf{L} = \{l_i \mid i = 1, 2, \ldots, N\}$, where $l_i \in \{1, 2, \ldots, C\}$ and C denotes the total number of detected clusters. The largest connected component is identified as $l_{max} = \arg\max_l |\{i \mid l_i = l\}|$, and the corresponding cluster points are given by $\mathbf{P}_{main} = \{\mathbf{p}_i \in \mathbf{P} \mid l_i = l_{max}\}$. The geometric centre of the main cluster is then computed as $\mathbf{p}_c = \text{mean}(\mathbf{P}_{main})$. To remove points located far from the cluster centre, the Euclidean distance of each point to $\mathbf{p}_c$ is calculated, and the percentile-based distance threshold T is applied. Points satisfying $d_i \leq T$ and belonging to valid clusters are retained in the final refined point cloud $\mathbf{P}' = \{\mathbf{p}_i \in \mathbf{P} \mid l_i \neq -1, d_i \leq T\}$.

Algorithm D. ConnectivityBasedFiltering

Algorithm D: ConnectivityBasedFiltering

Input: $\mathbf{P} \in \mathbb{R}^{N\times 3}$(input 3D point cloud)
ε (neighbourhood radius)
$n_{\min}$ (minimum number of neighbours)
α (distance multiplier)
Output: $\mathbf{P}'$ (refined point cloud)
Initialise an empty list $\mathcal{D}$
Initialise $\mathbf{P}' \leftarrow \emptyset$
Apply DBSCAN to $\mathbf{P}$ with parameters $(\varepsilon, n_{\min})$ to obtain labels $\mathbf{L}$
Identify largest cluster $l_{max} \leftarrow \arg\max_l |\{i \mid l_i = l\}|$
Extract main cluster $\mathbf{P}_{main} \leftarrow \{\mathbf{p}_i \in \mathbf{P} \mid l_i = l_{max}\}$
Compute cluster centre $\mathbf{p}_c \leftarrow \text{mean}(\mathbf{P}_{main})$
for each point $\mathbf{p}_i \in \mathbf{P}_{main}$ **do**
  Compute distance $d_i \leftarrow \| \mathbf{p}_i - \mathbf{p}_c \|_2$
  Append d_i to $\mathcal{D}$
end for
Compute distance threshold $T \leftarrow$ percentile $(\mathcal{D}, 95) \times \alpha$
for each point $\mathbf{p}_i \in \mathbf{P}$ with label l_i **do**
  if $l_i \neq -1$ **and** $\| \mathbf{p}_i - \mathbf{p}_c \|_2 \leq T$ **then**
    Append $\mathbf{p}_i$ to $\mathbf{P}'$
  end if
end for
return $\mathbf{P}'$

The adaptive noise filtering method dynamically adjusts filtering

Table 3
3D mIoU results for open-vocabulary monocular 3D object detection methods based on monocular metric depth estimators and camera intrinsic parameter estimators.

Method	Open-Vocabulary Object Detection	Monocular Metric Depth Estimator	Camera Intrinsic Parameter Estimator	Asset				
				E Kit	**Pump**	**Tank**	**Valve**	**Wire**
Our Method	GPT-4o (Asset Description)	UniDepthV2	UniDepthV2	**0.4251**	0.2509	0.2521	0.2516	**0.2217**
			PerspectiveField	0.3528	0.2365	0.2751	0.2401	0.2088
	Grounding DINO (Prompt-guided object detection)	Depth Anything V2	GeoCalib	0.3919	0.3025	0.2523	**0.2823**	0.2165
			PerspectiveField	0.2622	0.2344	**0.3659**	0.2018	0.1399
			GeoCalib	0.2995	**0.3205**	0.3385	0.2314	0.1461
OVMono3D-LIFT	–	–	–	0.3201	0.1736	0.2501	0.219	0.0341
OVMono3D-GEO	Grounding DINO	Depth Pro	Depth Pro	0.411	0.0752	0.1453	0.2556	0.0668
OVM3D-Det	Grounded-SAM	UniDepthV2	UniDepthV2	0.194	0.0711	0.1192	0.1828	0.0332

Table 4
Depth estimation and 3D bounding box orientation accuracy for different combinations of pre-trained MDE models and camera intrinsic parameter estimators.

Monocular Metric Depth Estimator	Camera Intrinsic Parameter Estimator	RMSE (m)	MAE (m)	$\bar{e}_{axis}$ (°)
UniDepthV2	UniDepthV2	1.7417	1.1011	10.29
	PerspectiveField	1.8658	1.3373	38.35
	GeoCalib	1.7884	1.1709	19.29
Depth Anything V2	PerspectiveField	1.9773	1.1306	39.71
	GeoCalib	1.8234	1.1785	20.61

parameters according to the size of the detected object to preserve its geometric fidelity while removing noise (see Algorithm E). The algorithm begins by calculating the spatial extent of the asset through its bounding dimensions. The spatial extent is computed as $\mathbf{b} = \max(\mathbf{P}) - \min(\mathbf{P})$, where $\mathbf{P}$ denotes the input point cloud. The object size is then defined by the Euclidean norm of its bounding vector, $s = \|\mathbf{b}\|_2$. Based on this measure, the point cloud is categorised into one of three groups: small (<0.3 m), medium (0.3–2.0 m), or large (>2.0 m). For small objects, voxel grid downsampling with fine resolution is applied to reduce redundancy, followed by distance-based filtering with conservative thresholds to retain fine geometrical details. For medium-sized objects, a multi-stage filtering strategy is employed, consisting of voxel downsampling, distance-based filtering with balanced thresholds, and radius-based filtering to remove sparse outliers. Connectivity-based filtering using DBSCAN is then applied to isolate the dominant cluster and discard weakly connected or scattered points. For large objects, more aggressive parameter settings are adopted. To reduce widespread noise in large objects, the algorithm applies coarser voxelisation and stricter distance thresholds. This is followed by radius-based filtering and density-driven clustering to preserve the geometry of the object and discard disconnected regions.

Table 7
Parameter sensitivity analysis reporting the percentage change in mIoU (ΔmIoU) **resulting from systematic variations** of key filtering parameters. Default baseline values are **distance multiplier = 1.2**, **distance percentile = 85%**, **voxel size = 0.01 m**, **DBSCAN ε = 0.025 m**, and **DBSCAN minimum samples = 5**.

Parameter	Change (%)	ΔmIoU (%)				
		E Kit	Valve	Pump	Tank	Wire
Distance Multiplier	−20	−7.89	−5.14	−8.04	−1.46	−10.01
	−10	−2.94	−0.42	−0.90	−6.24	−8.45
	10	−11.63	−3.89	−2.24	−16.76	−4.24
	20	−1.18	−6.65	−2.24	0.3	−3.07
DBSCAN ε	−20	−4.70	−11.05	−10.97	−2.13	−3.49
	−10	−16.15	−2.33	−0.87	−12.65	−0.72
	10	−9.52	−7.43	−7.43	−3.65	−5.51
	20	−3.04	−11.48	−13.48	−0.25	0.0
DBSCAN min points	−20	−5.29	−2.49	−1.09	0.365	0.0
	20	−0.21	−0.78	−0.22	−0.98	0.0

Table 5
Mean removal percentage based on asset size and noise filtering methods.

Object Size	Filtering Method	Asset					
		E Kit	Pump	Tank	Valve	Wire	Overall
Small (<0.3 m)	DBSCAN	–	–	–	0.48	–	0.48
	Voxel Downsampling + Statistical Outlier Removal	98.7	–	–	86.77	–	92.74
	Adaptive	99.73	98.94	99.89	87.26	–	96.45
Medium (0.3 m < 2 m)	DBSCAN	0.14	0.43	1.18	0.98	0.04	0.55
	Voxel Downsampling + Statistical Outlier Removal	90.02	84.6	63.41	83.79	89.76	82.32
	Adaptive	97.28	96.69	93.11	95.34	96.97	95.88
Large (> 2 m)	DBSCAN	0.02	0.82	1.34	3.43	0.75	1.27
	Voxel Downsampling + Statistical Outlier Removal	87.26	62.6	43.73	82.54	77.37	70.7
	Adaptive	95.94	–	–	95.74	97.74	96.47

Table 6
3D mIoU results for monocular 3D object detection based on asset type and noise filtering methods.

Asset	Filtering Method	Mean Removal (%)	Mean Volume (m^3)	Mean Points	mIoU
E Kit	DBSCAN	0.13	0.63	195,093	0.3350
	Voxel Downsampling + Statistical Outlier Removal	90.04	0.75	13,336	0.3262
	Adaptive	97.49	0.47	3991	**0.4251**
Valve	DBSCAN	1.02	0.07	19,088	0.1979
	Voxel Downsampling + Statistical Outlier Removal	84.51	0.42	1086	0.2171
	Adaptive	92.07	0.02	484	**0.2823**
Pump	DBSCAN	0.47	1.02	76,595	0.1804
	Voxel Downsampling + Statistical Outlier Removal	81.13	1.16	10,414	0.2043
	Adaptive	96.93	0.45	2177	**0.3205**
Tank	DBSCAN	1.2	3.65	120,685	**0.4086**
	Voxel Downsampling + Statistical Outlier Removal	60.95	3.71	22,801	0.2022
	Adaptive	93.95	0.84	3595	0.3659
Wire	DBSCAN	0.31	2	48,987	0.1094
	Voxel Downsampling + Statistical Outlier Removal	85	1.97	6309	0.1332
	Adaptive	97.03	0.31	1293	**0.2217**

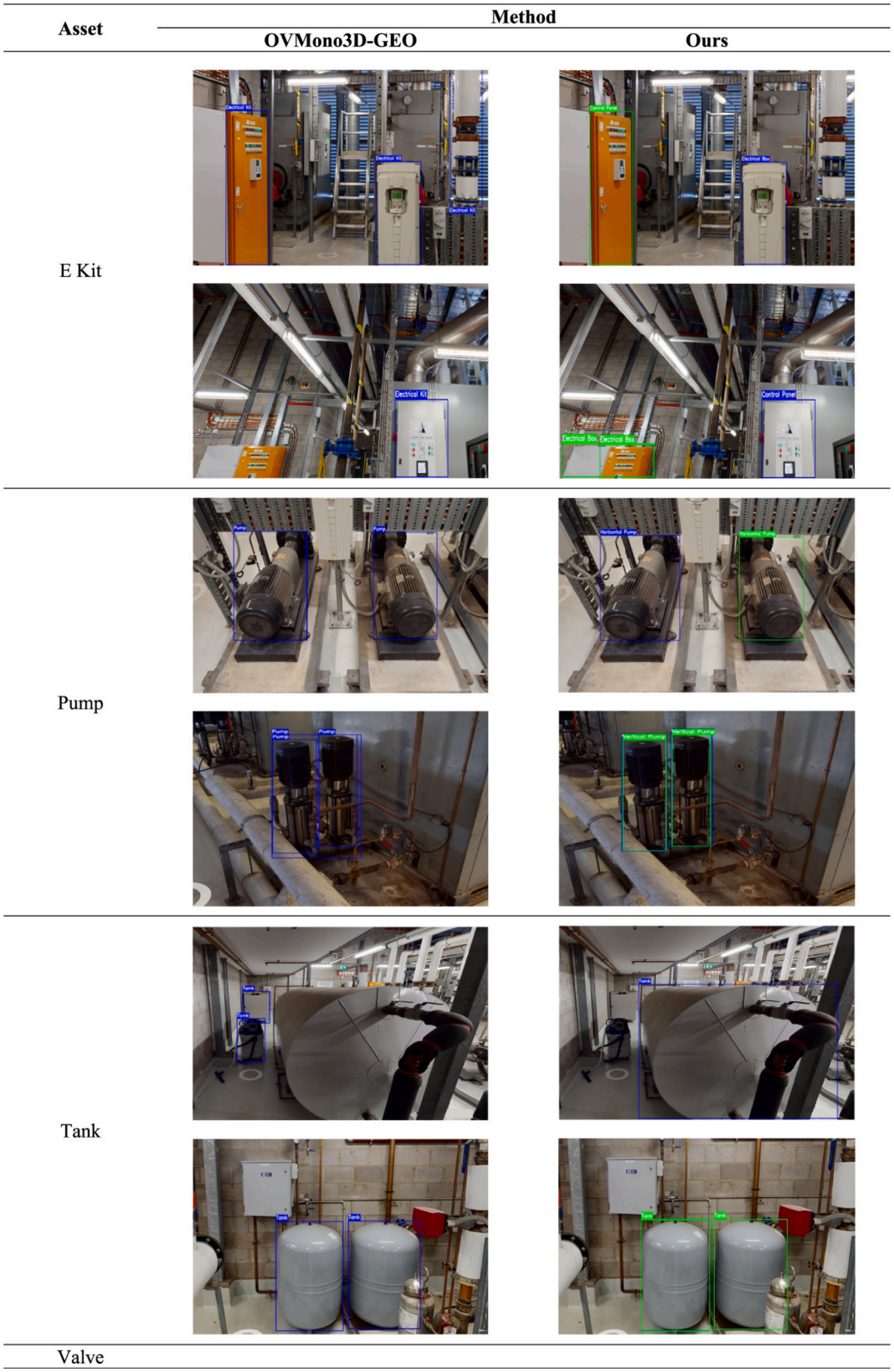


Fig. 4. Comparison of open-vocabulary object detection results for industrial assets.

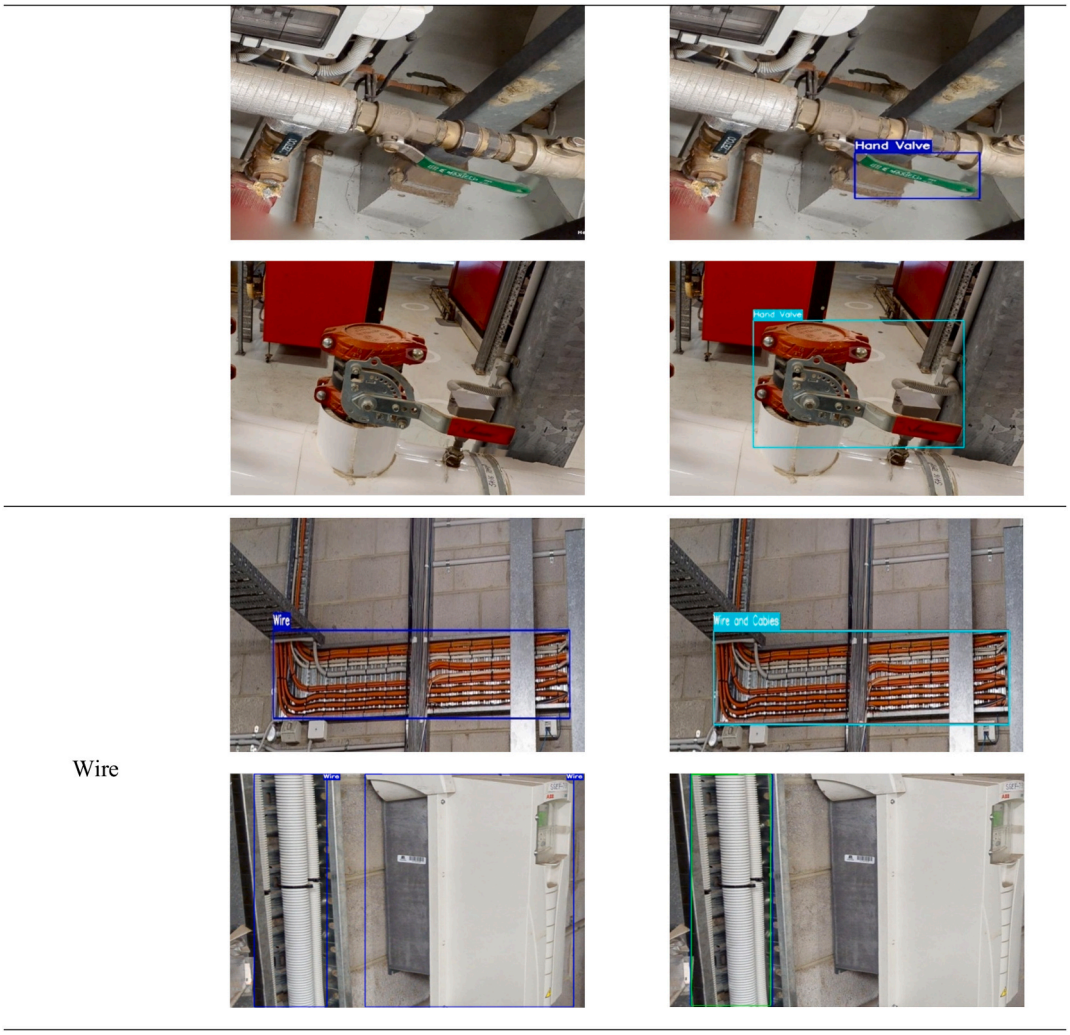


Fig. 4. (*continued*).

Algorithm E. AdaptiveNoiseFiltering algorithm

Algorithm E: AdaptiveNoiseFiltering algorithm

Input: $\mathbf{P} \in \mathbb{R}^{N\times 3}$(input 3D point cloud)
$v_{\text{small}}, v_{\text{medium}}, v_{\text{large}}$ (voxelisation parameters)
p, α (distance-based filtering parameters)
$r_{\text{medium}}, r_{\text{large}}, n_{\text{min}}$ (radius-based filtering parameters)
$\varepsilon_{\text{medium}}, \varepsilon_{\text{large}}, \alpha_{\text{db}}$ (connectivity-based filtering parameters)
Output: $\mathbf{P}'$ (final cleaned point cloud)

Initialise $\mathbf{P}' \leftarrow \emptyset$
$\mathbf{b} \leftarrow \max(\mathbf{P}) - \min(\mathbf{P})$
$s \leftarrow \|\mathbf{b}\|_2$
if $s < 0.3$m **then**
  $\mathbf{P}_v \leftarrow \text{VoxelGridDownsample}(\mathbf{P}, v_{\text{small}})$
  $\mathbf{P}' \leftarrow \text{DistanceBasedFiltering}(\mathbf{P}_v, p+10, \alpha+0.5)$
else if $s < 2.0$m **then**
  $\mathbf{P}_v \leftarrow \text{VoxelGridDownsample}(\mathbf{P}, v_{\text{medium}})$
  $\mathbf{P}_d \leftarrow \text{DistanceBasedFiltering}(\mathbf{P}_v, p-10, \alpha-0.4)$
  $\mathbf{P}_r \leftarrow \text{RadiusBasedFiltering}(\mathbf{P}_d, r_{\text{medium}}, n_{\text{min}})$
  $\mathbf{P}' \leftarrow \text{ConnectivityBasedFiltering}(\mathbf{P}_r, \varepsilon_{\text{medium}}, n_{\text{min}}, \alpha_{\text{db}})$
else
  $\mathbf{P}_v \leftarrow \text{VoxelGridDownsample}(\mathbf{P}, v_{large})$
  $\mathbf{P}_d \leftarrow \text{DistanceBasedFiltering}(\mathbf{P}_v, p-15, \alpha-0.5)$
  $\mathbf{P}_r \leftarrow \text{RadiusBasedFiltering}(\mathbf{P}_d, r_{large}, n_{\text{min}})$
  $\mathbf{P}' \leftarrow \text{ConnectivityBasedFiltering}(\mathbf{P}_r, \varepsilon_{large}, n_{\text{min}}, \alpha_{\text{db}})$
end if
return $\mathbf{P}'$

After applying the adaptive noise filtering method, noise is removed from each generated point cloud, and the dimensions of the 3D bounding boxes are adjusted accordingly. However, the orientation of the bounding boxes remains unresolved, as the point distribution does not directly reveal the object's directional alignment. Without orientation refinement, the bounding box may misrepresent the true pose of the object, leading to inaccurate 3D localisation. For example, a valve placed at an angle on a tilted pipe may be incorrectly represented as axis-aligned if the bounding box is not aligned with the actual direction of the object. To address this limitation, PCA is employed to determine the dominant geometric direction of the object by analysing the covariance of point distributions. The eigenvector corresponding to the largest eigenvalue represents the direction of maximum variance and is assigned as the orientation of the bounding box.

4. Experiments

In this section, the proposed training-free open-vocabulary monocular 3D object detection framework is evaluated. The experimental setup and data collection process are described, followed by quantitative and qualitative results across different industrial assets and scene conditions.

4.1. Data collection

To assess the effectiveness of the proposed method, a complex plant room containing a range of industrial assets within a building in Melbourne was selected. A total of 320 RGB images (4032 × 3024 pixels)

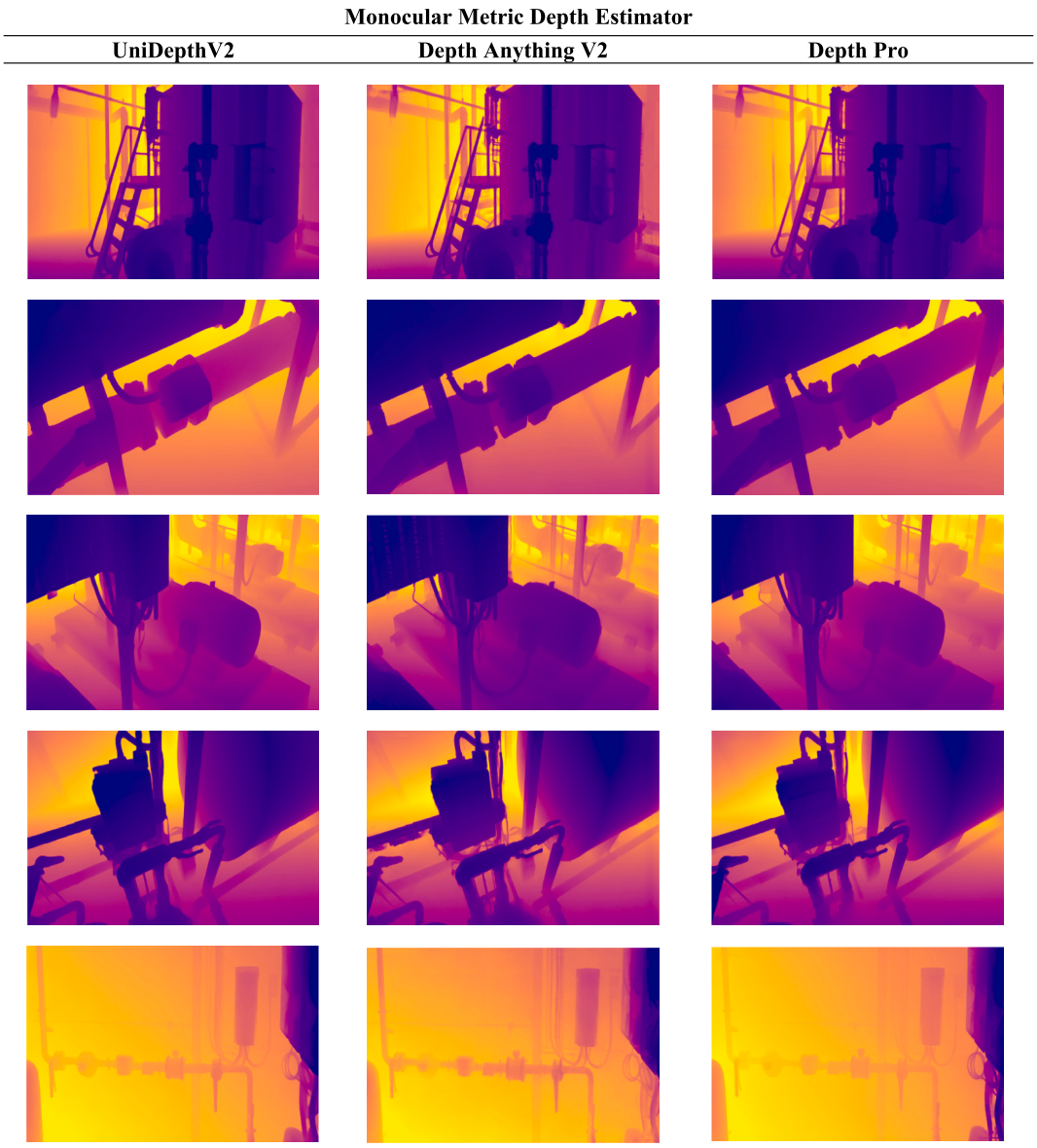


Fig. 5. Metric depth estimation results generated by pre-trained monocular depth estimators: UniDepthV2, Depth Anything V2, and Depth Pro.

were collected for open-vocabulary monocular 3D object detection. In addition, point cloud data of the entire plant room were captured using a Leica BLK360 laser scanner to provide ground truth data for assessing the accuracy of 3D bounding box predictions (Fig. 3). OVOD and MDE were performed on two NVIDIA L40 GPUs with 48 GB of VRAM, 16 vCPUs, and 241 GB of system memory. Besides, mask generation task was conducted on an NVIDIA L40s-16c GPU with 24 GB of VRAM, 16 vCPUs, and 118 GB of system memory. Ubuntu 20.04 was employed as the operating system for this study.

To evaluate generalisation across different scenarios, the images were captured to represent variations in asset type and scene complexity, including single-asset, multi-asset, and occluded scenes. Table 2 provides the distribution of test images across these scene conditions. In this context, single-asset scenes represent isolated objects with clear visibility, multi-asset scenes contain multiple objects within one frame, and occluded scenes include partially visible assets. The number of images for each asset was selected to provide adequate coverage across various scene configurations. Images were captured from different angles and distances to enhance diversity and reflect realistic variations in asset geometry and visual characteristics.

4.2. Results

Once assets are detected through OVOD, 2D bounding boxes are generated for each identified asset. These 2D bounding boxes are then transformed into 3D space using depth information and camera intrinsic parameters. UniDepthV2 and Depth Anything V2 are employed as MDE models to generate per-pixel depth maps and evaluate their impact on 3D bounding box prediction. As camera intrinsic parameters were not available for the images, two pre-trained models, GeoCalib [56] and PerspectiveField [57], were employed to estimate the focal length and principal point for each image. Based on the literature (Section 2), the proposed approach is compared with two state-of-the-art methods, OVMono3D-GEO and OVMono3D-LIFT, to evaluate generalisation capability and detection accuracy in open-vocabulary monocular 3D object detection for industrial assets. To address the absence of camera parameters for OVMono3D-GEO and ensure a fair comparison, Depth Pro is employed for both depth estimation and camera intrinsic parameter prediction.

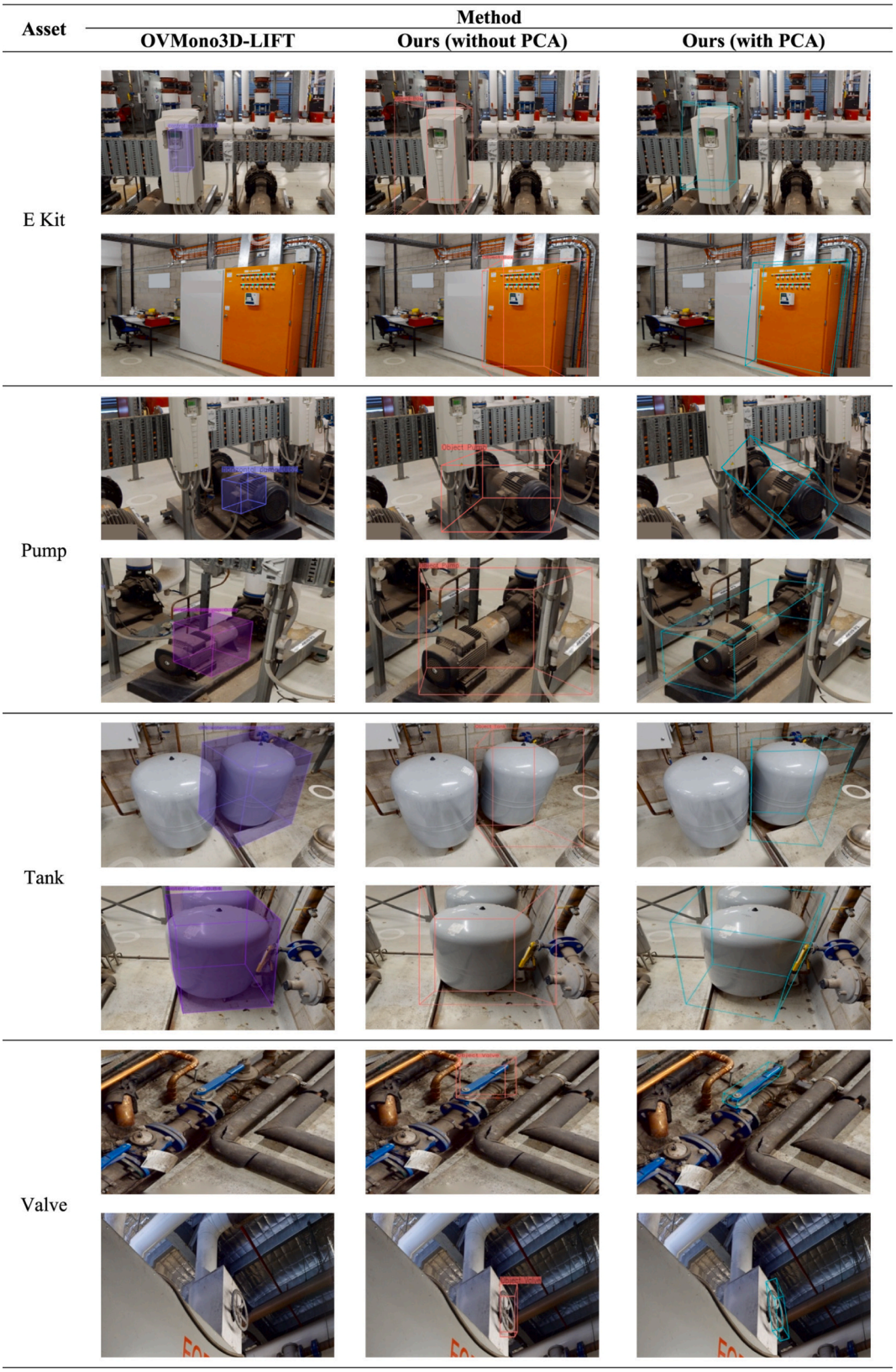


Fig. 6. Monocular 3D object detection results for industrial assets, comparing OVMono3D-LIFT with proposed method using PCA and without PCA.

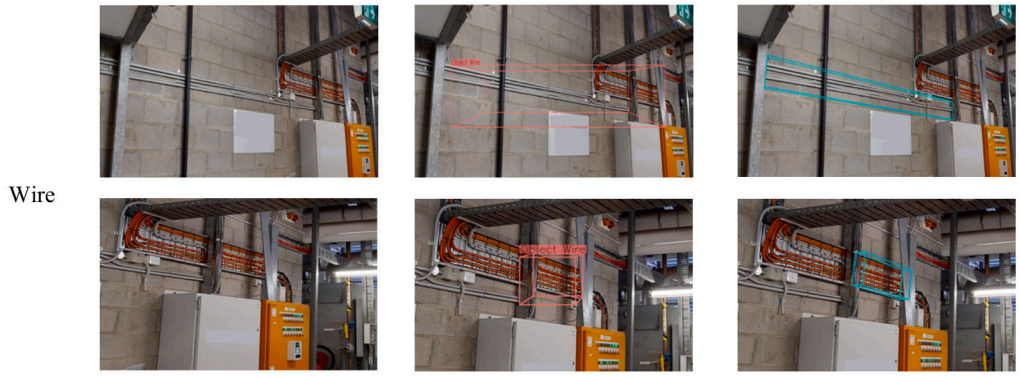


Fig. 6. (*continued*).

Fig. 7. Cross-scene open-vocabulary monocular 3D object detection results.

Table 8
Comparative analysis of computational efficiency among training-free M3OD methods.

Method	Average inference time (ms/image)	Processing speed (FPS)
OVMono3D-LIFT	142	7
OVMono3D-GEO	612	1.6
Ours	560	1.8

4.2.1. Quantitative results

Table 3 presents the 3D mIoU results for open-vocabulary monocular 3D object detection across five industrial asset types: electrical kit, pump, tank, valve, and wire. The evaluation includes different combinations of OVOD, MDE, and camera intrinsic parameter prediction models. The results indicate that our method achieves the best performance across all asset types, with mIoU scores of 0.4251 for kit, 0.3205 for pump, 0.3659 for tank, 0.2823 for valve, and 0.2217 for wire. In contrast, OVMono3D-LIFT and OVMono3D-GEO achieve lower mIoU values, with OVMono3D-GEO exhibiting notably weak performance for assets such as pumps and wires. This limitation is due to OVMono3D-GEO's reliance on Grounding DINO for OVOD, which fails to detect several industrial asset types. Similarly, OVM3D-Det demonstrates consistently lower performance across all asset categories. The integration of asset description extraction with a prompt-guided detection strategy in our method enhances 2D detection accuracy, resulting in higher 3D mIoU scores. Within the proposed framework, the highest mIoU scores for electrical kit and wire detection are obtained when UniDepthV2 is employed for both MDE and camera intrinsic parameter estimation. For pump detection, the best performance is achieved using the integration of Depth Anything V2 and GeoCalib. Tank detection attains the highest mIoU with Depth Anything V2 and PerspectiveField, while valve detection performs best using UniDepthV2 and GeoCalib. These findings demonstrate that the choice of depth and camera intrinsic parameter estimators is crucial, as their performance varies across asset types and geometric complexities.

Table 4 depicts the accuracy of depth estimation and 3D bounding box orientation for different combinations of MDE models and camera intrinsic parameter estimators. To assess the accuracy of depth estimation, the actual camera intrinsic parameters were used to generate the ground-truth depth maps. In this evaluation, UniDepthV2 employed for both camera parameter estimation and depth estimation achieves the lowest RMSE and MAE. In contrast, PerspectiveField and GeoCalib exhibit higher depth errors. PerspectiveField produces the largest deviations, resulting in the highest MAE when paired with UniDepthV2 and the highest RMSE when combined with Depth Anything V2. Depth

Anything V2 shows a stronger dependency on the choice of camera intrinsic parameter estimator.

Orientation accuracy is measured by the mean symmetry-aware principal-axis error $\bar{e}_{axis}$, defined as the minimum angular deviation between corresponding PCA axes of predicted and ground-truth geometries. The lowest mean orientation error is achieved when UniDepthV2 is used for both depth and camera intrinsic parameter estimation. PerspectiveField exhibits substantially higher orientation errors, indicating reduced robustness in camera intrinsic parameter estimation. While GeoCalib improves orientation accuracy compared with PerspectiveField, its performance remains less accurate than UniDepthV2.

Table 5 illustrates the effectiveness of the adaptive noise filtering method in applying size-aware filtering across different asset categories. The method consistently achieves high removal rates (ranging from 87% to 99%) while preserving essential geometric features through adaptive parameter tuning. For small objects (<0.3 m), conservative voxel downsampling combined with distance-based filtering effectively reduces redundancy while retaining fine geometrical details. Medium-sized assets (0.3–2.0 m) are filtered through a three-stage process combining radius-based methods and connectivity analysis, resulting in removal rates ranging from 82% to 96%. For large objects (>2.0 m), aggressive distance thresholds and clustering are applied, achieving up to 97% removal while maintaining geometrical integrity.

Table 6 shows the 3D mIoU results for M3OD across different noise filtering methods. Overall, the adaptive noise filtering method achieves the highest mIoU for most asset types, demonstrating its effectiveness in preserving critical geometric features. It outperforms other methods for electrical kit, valve, pump, and wire, achieving the highest mIoU values in each asset category. However, for tank detection, DBSCAN yields the best result with an mIoU of 0.4086, outperforming both voxel-based and adaptive methods. This outcome can be attributed to the distinct size and geometric characteristics of tanks, which are typically large in scale with smooth surfaces and minimal fine-grained details. In such cases, aggressive filtering methods, including voxel-based and adaptive approaches, may remove geometrically important points along curved surfaces or boundaries. By contrast, DBSCAN preserves the overall geometry and spatial structure of tanks, which is essential for accurate 3D bounding box estimation. Voxel downsampling with statistical outlier removal (SOR) achieves moderate performance but often sacrifices important geometric details, especially for smaller or intricate assets. These findings indicate that filtering strategies need to be adapted to the scale and geometric complexity of assets in industrial environments.

Table 7 summarises the sensitivity of the proposed filtering strategy to variations in key parameters by reporting the relative change in mIoU. The results indicate that the selected default parameters lie within a robust parameter range, as most variations result in limited reductions in performance rather than improvement. This behaviour demonstrates that the filtering pipeline maintains stable performance under different parameter adjustments, while the adopted default values provide a balanced trade-off between noise removal and geometric preservation across different asset types.

4.2.2. Qualitative results

Fig. 4 presents a qualitative comparison of OVOD results for industrial assets. Due to the diversity of industrial assets, such as variations in valves, OVMono3D-GEO fails to detect certain types of these assets. For wires, the method misclassifies control panels as wires, leading to false positives. In contrast, our approach demonstrates improved asset type recognition and accurate localisation. Besides, in dense areas with a high concentration of assets, Grounding DINO fails to detect multiple asset instances, whereas incorporating GPT-4o for asset description extraction improves the accuracy of 2D detection.

Fig. 5 illustrates the MDE results obtained using the pre-trained metric depth estimators, including UniDepthV2, Depth Anything V2, and Depth Pro. The results indicate that these models generate accurate depth maps that effectively capture the geometric structure of diverse asset configurations and spatial layouts. The clear delineation of object edges and depth discontinuities demonstrates their ability to generalise effectively to complex industrial environments without relying on scene-specific training.

Fig. 6 shows the M3OD results for different industrial assets. For the electrical kits and control panels, OVMono3D-LIFT detects the first instance with accurate 3D bounding box orientation, although the asset dimensions are not estimated precisely. This method fails to detect the control panel in the second image, whereas our method detects the electrical kit and the control panel in both cases. For pumps, OVMono3D-LIFT detects the assets but produces inaccurate size estimation. By contrast, the proposed method detects these assets, with PCA improving the orientation alignment of the 3D bounding boxes. In the case of tanks, OVMono3D-LIFT achieves accurate orientation and size estimation. However, this method fails to detect valves and wires, whereas our method successfully identifies these assets. In both cases, PCA enhances the alignment of 3D bounding box orientation.

To evaluate the generalisation capability of the proposed framework, Fig. 7 presents qualitative results from different scenes. These scenes exhibit diverse spatial layouts, asset configurations, and background conditions. Despite cross-scene variability, the framework produces geometrically consistent 3D bounding boxes, indicating robustness of the proposed approach beyond the original data collection site.

5. Discussion

Although M3OD achieves lower accuracy than multi-view images or LiDAR-based detection methods, it offers notable advantages in affordability and simplicity, relying solely on a single RGB image. The proposed training-free open-vocabulary monocular 3D object detection approach aims to enhance 3D scene understanding in industrial environments where domain-specific labelled datasets, camera parameters, and depth information are typically unavailable. However, the proposed approach is not intended for high-precision 3D reconstruction or CAD modelling but can be effectively employed in O&M activities such as asset management and industrial inspection. In such contexts, the ability to automatically detect and localise assets within large-scale industrial environments provides valuable input for rapid scene interpretation and remote monitoring.

The 3D mIoU values, ranging from 0.22 to 0.43, demonstrate that the proposed approach achieves competitive accuracy compared with existing M3OD methods, while eliminating the need for training data and camera calibration parameters. Although these mIoU values are moderate in absolute terms, they are sufficient for different O&M scenarios, where the primary requirements are reliable asset localisation, geometric extent, and spatial relationships. For instance, an mIoU of 0.43 for electrical kits indicates consistent spatial alignment effective for asset localisation and 3D scene understanding. While lower values, such as 0.22 for wires, indicate reduced precision for fine-scale objects, the results remain adequate for O&M tasks where approximate geometry is sufficient.

The primary aim of this paper is to enhance the accuracy and generalisation of training-free M3OD in complex industrial environments. Nevertheless, a comparative analysis of computational efficiency was conducted to evaluate the inference performance of the proposed method relative to existing approaches. Table 8 presents the average inference time and processing speed for each method. The results indicate that OVMono3D-LIFT achieves the highest processing speed due to its single-stage design, whereas the proposed method attains a shorter average inference time (560 ms per image) than OVMono3D-GEO. Although it may not be suitable for real-time applications, the approach remains effective for offline 3D scene understanding, which only needs to be performed once for each industrial environment. In terms of scalability to large-scale scenes, the computational cost of the proposed pipeline scales primarily with the number of detected assets rather than the spatial extent of the scene. Each asset is processed

independently after 2D detection, including depth extraction, point cloud generation, and adaptive filtering, which enables batch processing and parallelisation across assets and images. The localised assets can then be integrated with various data sources in digital twin environments to enhance operational efficiency and support informed decision-making.

5.1. Limitations

While the proposed approach demonstrates strong potential for training-free open-vocabulary monocular 3D object detection, it has certain limitations that should be acknowledged:

- **Generalisation:** The diversity of industrial assets in both type and size poses significant challenges for enhancing generalisation across different industrial settings. High asset density in complex scenes can lower the performance of asset detection and localisation. Moreover, the collected dataset includes a restricted number of assets, which may constrain its applicability to asset types that are not effectively detected through vision language foundation models. Environmental variations such as lighting conditions and accessibility constraints can affect detection accuracy. In addition, the dataset used in this paper was collected from a single industrial site, which may limit the transferability of the reported results to other industrial environments with different layouts, asset configurations, or operational conditions.
- **M3OD accuracy**: Although the proposed method demonstrates strong performance in recognising assets and generating corresponding 2D bounding boxes, the accuracy of 3D bounding boxes derived from M3OD remains limited compared with multi-view and LiDAR-based approaches. The intricate geometry and spatial complexity of certain assets, particularly wires and valves, present challenges for accurate 3D object detection in industrial scenes. Moreover, in occluded scenes, the full geometry of assets cannot be captured, leading to incomplete 3D bounding box generation and reduced localisation performance. This limitation becomes more evident in dense industrial environments where assets frequently overlap or extend beyond the camera's field of view.

6. Conclusions and future work

This paper presented a training-free open-vocabulary monocular 3D object detection approach that leverages vision language foundation models, MDE, and adaptive noise filtering to detect and localise assets in industrial environments. In contrast to existing approaches, the proposed method eliminates the reliance on task-specific training datasets and auxiliary information, such as depth maps and camera intrinsic parameters. By integrating GPT-4o with Grounding DINO, the framework enhances OVOD through asset description extraction and prompt-guided object detection. Depth estimation and camera intrinsic parameter prediction are performed using pre-trained models, enabling accurate 2D-to-3D transformation without manual calibration. The proposed method outperforms existing open-vocabulary monocular 3D detection approaches, achieving higher mIoU across diverse asset types, with values of 0.4251 for electrical kits, 0.3205 for pumps, 0.3659 for tanks, 0.2823 for valves, and 0.2217 for wires.

Future research could explore several directions to further enhance the accuracy and robustness of the proposed approach:

- Context-guided 3D bounding box refinement: The spatial context of neighbouring assets can be utilised to refine the orientation and dimensions of 3D bounding boxes. The relative positioning and alignment of surrounding objects provide geometric cues that enhance localisation accuracy and consistency in complex industrial environments. Vision language foundation models can be employed to enable spatial reasoning by deriving the orientation and dimensions of assets in relation to their surroundings, thereby refining the detected 3D bounding boxes.
- Novel view synthesis: Novel view generation methods may be employed to reconstruct additional perspectives of assets from single images. These approaches utilise learned representations of scene geometry to generate novel viewpoints beyond the original camera perspective. By synthesising new viewpoints, occluded or unseen parts of assets can be recovered to improve the prediction of asset dimensions and orientation.

CRediT authorship contribution statement

Masoud Kamali: Writing – review & editing, Writing – original draft, Methodology, Formal analysis, Data curation, Conceptualization. **Behnam Atazadeh:** Writing – review & editing, Supervision, Methodology, Funding acquisition, Conceptualization. **Abbas Rajabifard:** Writing – review & editing, Supervision, Project administration, Funding acquisition. **Yiqun Chen:** Writing – review & editing, Supervision, Project administration, Methodology, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests.

Masoud Kamali reports financial support was provided by Australian Research Council. Behnam Atazadeh reports financial support was provided by Australian Research Council. Abbas Rajabifard reports financial support was provided by Australian Research Council. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research was supported by Australian Research Council [grant numbers: IH210100048, DE220100094], the University of Melbourne's Research Computing Services, and Petascale Campus Initiative. The authors acknowledge the support of industry partners: Emerson and Rockfield. The authors emphasise that the views expressed in this article are the authors' alone.

Data availability

Some or all data, models, and code that support the findings of this study are available from the corresponding author upon request.

References

[1] Z.-Z. Hu, P.-L. Tian, S.-W. Li, J.-P. Zhang, BIM-based integrated delivery technologies for intelligent MEP management in the operation and maintenance phase, Adv. Eng. Softw. 115 (2018) 1–16, https://doi.org/10.1016/j.advengsoft.2017.08.007.
[2] R.E. Chapman, Inadequate Interoperability: a closer look at the costs, in: Presented at the 22nd International Symposium on Automation and Robotics in Construction (ISARC 2005), Ferrara, Italy, September 11–14, 2005, https://doi.org/10.22260/ISARC2005/0087.
[3] P. Reliability, Achieving a 26% Reduction in Annual O&M Costs Through Reliability-Centered Maintenance. https://pinnaclereliability.com/learn/case-studies/achieving-a-26-reduction-in-annual-om-costs-through-reliability-centered-maintenance/, 2026 (accessed August 13, 2025).
[4] L.C. Lemes, L. Hvam, Maintenance costs in the process industry: A literature review, in: In 2019 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM), IEEE, Macau, December 15–18 2019, pp. 1481–1485, https://doi.org/10.1109/ieem44572.2019.8978559.
[5] D. Thomas, B. Weiss, Maintenance costs and advanced maintenance techniques in manufacturing machinery: survey and analysis, Int. J. Prognost. Health Manag. 12 (1) (2021), https://doi.org/10.36001/ijphm.2021.v12i1.2883.
[6] Z. Jiang, J.I. Messner, Computer vision applications in construction and asset management phases: a literature review, J. Informat. Technol. Construct. 28 (2023) 176–199, https://doi.org/10.36680/j.itcon.2023.009.

[7] C. Kanellakis, E. Fresk, S.S. Mansouri, D. Kominiak, G. Nikolakopoulos, Autonomous visual inspection of large-scale infrastructures using aerial robots, arXiv preprint (2019), https://doi.org/10.48550/arXiv.1901.05510 arXiv: 1901.05510.

[8] M. Kamali, B. Atazadeh, A. Rajabifard, Y. Chen, Advancements in 3D digital model generation for digital twins in industrial environments: knowledge gaps and future directions, Adv. Eng. Inform. 62 (2024) 102929, https://doi.org/10.1016/j.aei.2024.102929.

[9] A. Rahimi, M. Anvaripour, K. Hayat, Object detection using deep learning in a manufacturing plant to improve manual inspection, in: 2021 IEEE International Conference on Prognostics and Health Management (ICPHM), Online, June 7–9, IEEE, 2021, pp. 1–7, https://doi.org/10.1109/ICPHM51084.2021.9486529.

[10] H. Son, C. Kim, C. Kim, 3D reconstruction of as-built industrial instrumentation models from laser-scan data and a 3D CAD database based on prior knowledge, Autom. Constr. 49 (2015) 193–200, https://doi.org/10.1016/j.autcon.2014.08.007.

[11] D.G. Schneider, M.R. Stemmer, CNN-based multi-object detection and segmentation in 3D LiDAR data for dynamic industrial environments, Robotics 13 (12) (2024) 174, https://doi.org/10.3390/robotics13120174.

[12] Y. Shang, W. Yu, G. Zeng, H. Li, Y. Wu, StereoYOLO: a stereo vision-based method for maritime object recognition and localization, J. Mar. Sci. Eng. 12 (1) (2024) 197, https://doi.org/10.3390/jmse12010197.

[13] Z. Fan, Y. Zhu, Y. He, Q. Sun, H. Liu, J. He, Deep learning on monocular object pose detection and tracking: a comprehensive overview, ACM Comput. Surv. 55 (4) (2022) 1–40, https://doi.org/10.1145/3524496.

[14] Q. Lian, P. Li, X. Chen, MonoJSG: Joint semantic and geometric cost volume for monocular 3D object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, June 19–24, 2022, pp. 1070–1079, https://doi.org/10.1109/CVPR52688.2022.00114.

[15] A. Geiger, P. Lenz, R. Urtasun, Are we ready for autonomous driving? The KITTI vision benchmark suite, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, Providence, Rhode Island, USA, June , pp. 3354–3361, https://doi.org/10.1109/CVPR.2012.6248074.

[16] H. Caesar, V. Bankiti, A.H. Lang, S. Vora, V.E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, O. Beijbom, nuScenes: A multimodal dataset for autonomous driving, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 13–19 2020, pp. 11621–11631, https://doi.org/10.1109/CVPR42600.2020.01164.

[17] G. Brazil, A. Kumar, J. Straub, N. Ravi, J. Johnson, G. Gkioxari, Omni3D: A large benchmark and model for 3D object detection in the wild, in: In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 2023, pp. 13154–13164. https://doi.org/10.1109/CVPR52729.2023.01264.

[18] Y. Wu, Y. Wang, S. Zhang, H. Ogai, Deep 3D object detection networks using LiDAR data: a review, IEEE Sensors J. 21 (2) (2020) 1152–1171, https://doi.org/10.1109/JSEN.2020.3020626.

[19] S.Y. Alaba, J.E. Ball, A survey on deep-learning-based lidar 3d object detection for autonomous driving, Sensors 22 (24) (2022) 9577, https://doi.org/10.3390/s22249577.

[20] X. Ma, W. Ouyang, A. Simonelli, E. Ricci, 3D object detection from images for autonomous driving: a survey, IEEE Trans. Pattern Anal. Mach. Intell. 46 (5) (2023) 3537–3556, https://doi.org/10.1109/TPAMI.2023.3346386.

[21] S.Y. Alaba, J.E. Ball, Deep learning-based image 3-d object detection for autonomous driving, IEEE Sensors J. 23 (4) (2023) 3378–3394, https://doi.org/10.1109/JSEN.2023.3235830.

[22] Y. Xie, C. Xu, M.-J. Rakotosaona, P. Rim, F. Tombari, K. Keutzer, M. Tomizuka, W. Zhan, SparseFusion: Fusing multi-modal sparse representations for multi-sensor 3d object detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, October 1–6 2023, pp. 17591–17602, https://doi.org/10.1109/ICCV51070.2023.01613.

[23] L. Wang, X. Zhang, Z. Song, J. Bi, G. Zhang, H. Wei, L. Tang, L. Yang, J. Li, C. Jia, Multi-modal 3D object detection in autonomous driving: a survey and taxonomy, IEEE Trans. Intelligent Veh. 8 (7) (2023) 3781–3798, https://doi.org/10.1109/TIV.2023.3264658.

[24] L. Peng, X. Wu, Z. Yang, H. Liu, D. Cai, DID-M3D: Decoupling instance depth for monocular 3d object detection, in: 17th European Conference on Computer Vision, Tel Aviv, Springer, Israel, October , pp. 71–88, https://doi.org/10.1007/978-3-031-19769-7_5.

[25] R. Zhang, H. Qiu, T. Wang, Z. Guo, Z. Cui, Y. Qiao, H. Li, P. Gao, MonoDETR: Depth-guided transformer for monocular 3D object detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, October 1–6, 2023, pp. 9155–9166, https://doi.org/10.1109/ICCV51070.2023.00840.

[26] Y. Li, Y. Chen, J. He, Z. Zhang, Densely constrained depth estimator for monocular 3d object detection, in: In 17th European Conference on Computer Vision, Tel-Aviv, Springer, Israel, October , pp. 718–734, https://doi.org/10.48550/arXiv.2207.10047.

[27] Y. Zhang, J. Lu, J. Zhou, Objects are different: Flexible monocular 3d object detection, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, June 20–25, 2021, pp. 3289–3298, https://doi.org/10.1109/CVPR46437.2021.00330.

[28] A. Kumar, G. Brazil, E. Corona, A. Parchami, X. Liu, DEVIANT: Depth equivariant network for monocular 3d object detection, in: In 17th European Conference on Computer Vision, Tel Aviv, Springer, Israel, October , pp. 664–683, https://doi.org/10.1007/978-3-031-20077-9_39.

[29] J. Shen, L. Jiao, C. Zhang, K. Peng, Monocular 3D object detection for construction scene analysis, Comput. Aided Civ. Inf. Eng. 39 (9) (2024) 1370–1389, https://doi.org/10.1111/mice.13143.

[30] J. Shen, W. Yan, P. Li, X. Xiong, Deep learning-based object identification with instance segmentation and pseudo-LiDAR point cloud for work zone safety, Comput. Aided Civ. Inf. Eng. 36 (12) (2021) 1549–1567, https://doi.org/10.1111/mice.12749.

[31] H. Kim, H. Kim, Y.W. Hong, H. Byun, Detecting construction equipment using a region-based fully convolutional network and transfer learning, J. Comput. Civ. Eng. 32 (2) (2018) 04017082, https://doi.org/10.1061/(ASCE)CP.1943-5487.0000731.

[32] Y. Ding, Q. Liu, A. Ji, H. Li, X. Luo, Monocular three-dimensional object detection for proximity monitoring in human-machine collision warning systems on construction sites, Eng. Appl. Artif. Intell. 159 (2025) 111722, https://doi.org/10.1016/j.engappai.2025.111722.

[33] B. Drost, M. Ulrich, P. Bergmann, P. Hartinger, C. Steger, Introducing MVTec ITODD - A dataset for 3d object recognition in industry, in: IEEE International Conference on Computer Vision Workshops, Venice, Italy, October 22–29, 2017, pp. 2200–2208, https://doi.org/10.1109/ICCVW.2017.257.

[34] T. Hodan, P. Haluza, Š. Obdržálek, J. Matas, M. Lourakis, X. Zabulis, T-LESS: An RGB-D dataset for 6D pose estimation of texture-less objects, in: In 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, Santa Rosa, CA, USA, March , pp. 880–888, https://doi.org/10.1109/WACV.2017.103.

[35] Y. Wang, W.-L. Chao, D. Garg, B. Hariharan, M. Campbell, K.Q. Weinberger, Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, June 15–20, 2019, pp. 8445–8453, https://doi.org/10.1109/CVPR.2019.00864.

[36] B. Kerbl, G. Kopanas, T. Leimkühler, G. Drettakis, 3D gaussian splatting for real-time radiance field rendering, ACM Trans. Graph. 42 (4) (2023) 1–14, https://doi.org/10.1145/3592433.

[37] S.F. Bhat, R. Birkl, D. Wofk, P. Wonka, M. Müller, Zoedepth: Zero-shot transfer by combining relative and metric depth, arXiv preprint (2023), https://doi.org/10.48550/arXiv.2302.12288 arXiv:2302.12288.

[38] L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng, H. Zhao, Depth Anything V2, arXiv preprint (2024), https://doi.org/10.48550/arXiv.2406.09414 arXiv: 2406.09414.

[39] A. Bochkovskii, A. Delaunoy, H. Germain, M. Santos, Y. Zhou, S.R. Richter, V. Koltun, Depth Pro: Sharp monocular metric depth in less than a second, arXiv preprint (2024), https://doi.org/10.48550/arXiv.2410.02073 arXiv:2410.02073.

[40] L. Piccinelli, Y.-H. Yang, C. Sakaridis, M. Segu, S. Li, L. Van Gool, F. Yu, UniDepth: Universal monocular metric depth estimation, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, June 16–22, 2024, pp. 10106–10116, https://doi.org/10.1109/CVPR52733.2024.00963.

[41] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, Q. Jiang, C. Li, J. Yang, H. Su, Grounding DINO: Marrying dino with grounded pre-training for open-set object detection, in: In 18th European Conference on Computer Vision, Springer, Milan, Italy, September , pp. 38–55, https://doi.org/10.1007/978-3-031-72970-6_3.

[42] D. Zhang, C. Li, R. Zhang, S. Xie, W. Xue, X. Xie, S. Zhang, FM-OV3D: foundation model-based cross-modal knowledge blending for open-vocabulary 3d detection, Proceed. AAAI Conferen. Artifi. Intellig. 38 (15) (2024) 16723–16731, https://doi.org/10.1609/aaai.v38i15.29612.

[43] T. Ren, S. Liu, A. Zeng, J. Lin, K. Li, H. Cao, J. Chen, X. Huang, Y. Chen, F. Yan, Grounded SAM: Assembling open-world models for diverse visual tasks, arXiv preprint (2024), https://doi.org/10.48550/arXiv.2401.14159 arXiv:2401.14159.

[44] P. Jiao, N. Zhao, J. Chen, Y.-G. Jiang, Unlocking textual and visual wisdom: Open-vocabulary 3d object detection enhanced by comprehensive guidance from text and image, in: In 18th European Conference on Computer Vision, Springer, Milan, Italy, September , pp. 376–392, https://doi.org/10.1007/978-3-031-73195-2_22.

[45] R. Mohiuddin, S.M. Prakhya, F. Collins, Z. Liu, A. Borrmann, OpenSU3D: Open world 3d scene understanding using foundation models, in: 2025 IEEE International Conference on Robotics and Automation (ICRA), Atlanta, GA, USA, IEEE, 2025, pp. 13560–13566, https://doi.org/10.1109/ICRA55743.2025.11127896.

[46] Y. Lu, C. Xu, X. Wei, X. Xie, M. Tomizuka, K. Keutzer, S. Zhang, Open-vocabulary point-cloud object detection without 3D annotation, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, June 17–24, 2023, pp. 1190–1199, https://doi.org/10.1109/CVPR52729.2023.00121.

[47] R. Huang, H. Zheng, Y. Wang, Z. Xia, M. Pavone, G. Huang, Training an open-vocabulary monocular 3d detection model without 3d data, in: International Conference on Neural Information Processing Systems, Vancouver, Canada vol. 37, December 1-15 2024, pp. 72145–72169, https://doi.org/10.52202/079017-2303.

[48] J. Yao, H. Gu, X. Chen, J. Wang, Z. Cheng, Open Vocabulary Monocular 3D Object Detection, arXiv preprint (2024), https://doi.org/10.48550/arXiv.2411.16833 arXiv:2411.16833.

[49] X. Liu, N. Xue, T. Wu, Learning auxiliary monocular contexts helps monocular 3d object detection, in: Proceedings of the AAAI Conference on Artificial Intelligence, February 22 – March 1 36, no. 2, 2022, pp. 1810–1818, https://doi.org/10.1609/aaai.v36i2.20074.

[50] Y. Oh, H.-I. Kim, S.T. Kim, J.U. Kim, MonoWAD: Weather-adaptive diffusion model for robust monocular 3d object detection, in: In 18th European Conference on Computer Vision, Springer, Milan, Italy, September , pp. 326–345, https://doi.org/10.1007/978-3-031-72684-2_19.

[51] Y. Gao, X. Miao, G. Zhang, Monocular 3D object detection for occluded targets based on spatial relationships and decoupled depth predictions, Front. Comp. Sci. 6 (2025) 1382080, https://doi.org/10.3389/fcomp.2024.1382080.

[52] R. Zhang, H.-S. Choi, D. Jung, P.H.N. Anh, S.-K. Jeong, Z. Zhu, Auxdepthnet: real-time monocular 3D object detection with depth-sensitive features, Appl. Sci. 15 (13) (2025), https://doi.org/10.3390/app15137538. Art no. 7538.
[53] S. Peng, K. Genova, C. Jiang, A. Tagliasacchi, M. Pollefeys, T. Funkhouser, OpenScene: 3D scene understanding with open vocabularies, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, June 17–24, 2023, pp. 815–824, https://doi.org/10.1109/CVPR52729.2023.00085.
[54] Z. Wang, Y. Li, T. Liu, H. Zhao, S. Wang, Ov-Uni3DETR: Towards unified open-vocabulary 3d object detection via cycle-modality propagation, in: European Conference on Computer Vision, Milan, Italy, September 29-October 4, Springer, 2024, pp. 73–89, https://doi.org/10.1007/978-3-031-72970-6_5.
[55] OpenAI, ChatGPT-4o (September 15 version). https://chat.openai.com/chat (accessed 10 January, 2026).
[56] A. Veicht, P.-E. Sarlin, P. Lindenberger, M. Pollefeys, Geocalib: Learning single-image calibration with geometric optimization, in: 18th European Conference on Computer Vision, Milan, Italy, September 29–October 4, Springer, 2024, pp. 1–20, https://doi.org/10.1007/978-3-031-73661-2_1.
[57] L. Jin, J. Zhang, Y. Hold-Geoffroy, O. Wang, K. Blackburn-Matzen, M. Sticha, D. F. Fouhey, Perspective fields for single image camera calibration, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, June 17–24, 2023, pp. 17307–17316, https://doi.org/10.1109/CVPR52729.2023.01660.