Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Depth-Guided RGB-D Detection for Tiny Pavement Debris in Complex Warehouse Inspection Scenes

Jun Wu

rxguo_cauc@163.com

Civil Aviation University of China

Jiaxin Li

Civil Aviation University of China

Runxia Guo

Civil Aviation University of China

Jiusheng Chen

Engineering Technology Training Center

Guihang Liu

Engineering Technology Training Center

Research Article

Keywords:

Posted Date: November 17th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11594712/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Depth-Guided RGB-D Detection for Tiny Pavement Debris in Complex Warehouse Inspection Scenes

Jun Wu[1,2*], Jiaxin Li[1], Runxia Guo[2], Jiusheng Chen[3], Guihang Liu[3]

[1*]School of Aeronautical Engineering, Civil Aviation University of China, Tianjin, China.
[2]College of Electronic Information and Automation, Civil Aviation University of China, Tianjin, China.
[3]Engineering Technology Training Center, Civil Aviation University of China, Tianjin, China.

*Corresponding author(s). E-mail(s): rxguo_cauc@163.com;
Contributing authors: lijiaxin020615@163.com;

Acknowledgements. The authors would like to thank the National Natural Science Foundation of China, the Science and Technology Program of Tianjin, the Foundation of Tianjin Educational Committee, and the Basic Science-Research Funds of National University for their financial support.

Abstract

Autonomous guided vehicles (AGVs) are increasingly used for inspection and transportation tasks in aviation warehouses. However, dense shelving, local shadows, and uneven illumination often degrade RGB image quality, reducing the reliability of vision-based perception. Tiny pavement debris, such as packaging fragments, plastic bags, cigarette butts, and small hardware items, may interfere with AGV locomotion and local perception, posing risks to warehouse inspection safety. To address this problem, this paper proposes a depth-guided RGB-depth (RGB-D) detection framework, termed PDE-RT-DETR, for robust tiny debris detection in low-illumination aviation warehouse scenes. The framework introduces a depth branch to provide auxiliary geometric cues, designs a Depth Enhancement Module (DEM) to improve depth-related feature representation through nonlinear depth discretization and re-encoding, and incorporates Continuous Depth Position Encoding (CDPE) to inject continuous depth-aware positional information into the Transformer feature space. A self-built RGB-D pavement debris dataset containing 5,635 aligned image pairs was constructed under multiple illumination conditions. Experimental results show that PDE-RT-DETR achieves 86.4% precision, 87.6% recall, 91.4% $mAP_{0.50}$, and 66.8% $mAP_{0.50:0.95}$. Compared with the RGB-D RT-DETR-R18 baseline, the proposed method improves $mAP_{0.50}$ and $mAP_{0.50:0.95}$ by 9.2 and 13.3 percentage points, respectively. Under extreme low illumination, PDE-RT-DETR maintains 82.1% $mAP_{0.50}$, with a degradation rate of only 16.14%. These results demonstrate that depth-guided RGB-D perception improves the accuracy and robustness of tiny pavement debris detection in complex warehouse inspection scenes.

Keywords: RGB-D perception, tiny debris detection, low-illumination vision, depth-guided detection, warehouse inspection

1 Introduction

With the rapid development of intelligent logistics and digital warehouse management, autonomous guided vehicles (AGVs) have been increasingly used in aviation warehouses for material transportation, inventory inspection, and routine patrol tasks [1, 2]. Compared with manual inspection [3], AGV-based systems can improve operational efficiency, reduce labor costs, and support long-duration autonomous operation. As AGVs become an important component of aviation warehouse management, reliable environmental perception has become a key prerequisite for stable and safe warehouse inspection. In particular, AGVs operating in densely arranged warehouse aisles must continuously perceive pavement conditions and identify abnormal objects that may affect locomotion or navigation.

Tiny pavement debris is a common but easily overlooked risk factor in aviation warehouse environments. During cargo handling, equipment maintenance, and daily warehouse operations, small objects such as paper balls, plastic bags, cigarette butts, packaging fragments, and small hardware items may appear on the pavement. Although these objects are usually small in size, they may interfere with AGV wheels, onboard sensors, or local perception modules. For example, tiny debris may cause

wheel jamming or slippage, trigger sensor misreadings, and lead to path deviation, emergency braking, or collisions with nearby racks and equipment [4–6]. Since aviation warehouses often contain high-value components and densely arranged storage facilities, even minor perception failures may result in operational interruption or equipment damage. Therefore, reliable detection of tiny pavement debris is important for AGV inspection and warehouse safety management.

Despite the rapid progress of deep-learning-based object detection methods [7], tiny debris detection remains a challenging machine vision task in aviation warehouse scenes. First, debris targets usually occupy only a small number of pixels in RGB images, resulting in limited texture and edge information. Second, the visual appearance of debris is often similar to that of warehouse pavements, especially for low-contrast objects such as plastic bags, paper fragments, or cigarette butts. Third, aviation warehouses are commonly affected by dense shelving, partial occlusion, uneven illumination, shadows, and local low-light regions. Under such conditions, RGB image quality degrades significantly due to reduced signal-to-noise ratio, weakened texture details, and blurred object boundaries [8, 9]. These factors make tiny debris visually indistinguishable from complex pavement backgrounds and lead to missed detections or inaccurate localization in RGB-only detection methods [10].

Existing object detection methods have achieved considerable progress in general vision tasks. Convolutional neural network (CNN)-based detectors, such as Faster R-CNN and the YOLO series, have been widely applied in industrial inspection and robotic perception due to their strong feature extraction capability and high detection efficiency [11]. Transformer-based detectors further improve object detection by modeling global dependencies and reducing the reliance on hand-designed post-processing procedures. DETR introduced an end-to-end detection paradigm based on object queries, and subsequent variants such as Deformable DETR improved convergence and computational efficiency by introducing deformable attention mechanisms [12]. RT-DETR further enhances real-time detection performance through an efficient hybrid encoder and query selection strategy [13]. However, most of these detectors are mainly designed for RGB images. When the visual modality is degraded by low illumination, shadows, and weak texture, the semantic features extracted from RGB images become unstable, which limits their robustness in tiny debris detection for aviation warehouse inspection.

To overcome the limitations of RGB-only perception, multisensor machine vision has attracted increasing attention in robotic inspection and autonomous navigation applications. By combining complementary information from different sensing modalities, multimodal perception can improve detection reliability under challenging environmental conditions. Existing studies have explored RGB-infrared or RGB-thermal fusion to enhance perception under poor lighting conditions [14, 15]. These methods demonstrate that auxiliary sensing modalities can compensate for the degradation of visible images. However, for pavement debris detection, thermal or infrared information may not always provide sufficient geometric distinction between small objects and the ground surface. In contrast, depth sensing can directly describe the spatial structure of the scene and provide geometric cues that are less sensitive to illumination changes.

Depth information is particularly useful for tiny pavement debris detection because it can represent height differences and geometric discontinuities between debris and the ground plane. Unlike RGB images, which mainly describe color and texture appearance, depth maps provide spatial relationships between objects and surrounding surfaces. For weak-textured, low-contrast, or partially transparent debris, depth cues can offer complementary information that is difficult to obtain from visual appearance alone. RGB-D perception has therefore been widely studied in object detection, scene understanding, robot navigation, and autonomous inspection. Nevertheless, directly applying RGB-D fusion to low-illumination aviation warehouse scenes still faces several problems. Raw depth observations may be affected by measurement noise, reflective materials, sensing distance, and missing depth values, resulting in depth holes, discontinuities, and inaccurate geometric responses. If such unreliable depth information is directly fused with RGB features, it may introduce artifacts into the detection pipeline and reduce feature representation quality.

Another limitation of existing RGB-D detection methods is that depth information is often used only as an additional input feature, while the geometric relationships encoded in depth maps are not sufficiently modeled. For tiny debris detection, reliable localization depends not only on semantic recognition but also on accurate spatial representation. Therefore, it is necessary to enhance the reliability of raw depth observations and to incorporate depth-aware geometric cues into the detection framework in a more explicit manner. This motivates the development of a depth-guided RGB-D detection method for tiny pavement debris perception in low-illumination aviation warehouse inspection scenes.

To address these challenges, this paper presents a depth-guided RGB-D detection framework, termed PDE-RT-DETR, for tiny pavement debris detection in complex aviation warehouse scenes. The proposed framework is built upon RT-DETR and uses RGB images for semantic feature extraction and depth observations for geometric compensation. A Depth Enhancement Module (DEM) is introduced to improve the representation quality of raw depth information by suppressing unstable depth responses and enhancing geometric structure cues. In addition, a Continuous Depth Position Encoding (CDPE) strategy is designed to inject depth-aware geometric information into the positional representation space of the Transformer-based detector, thereby improving spatial representation learning and object localization under low-illumination conditions.The main contributions of this work are summarized as follows:

(1) A depth-guided RGB-D detection framework, termed PDE-RT-DETR, is developed for tiny pavement debris detection in low-illumination aviation warehouse inspection scenes. The framework exploits the complementary advantages of RGB semantic information and depth-based geometric structure to improve detection reliability under visual degradation.

(2) A Depth Enhancement Module (DEM) and a Continuous Depth Position Encoding (CDPE) strategy are introduced into the RT-DETR-based detector. DEM improves the usability of raw depth observations by enhancing geometric structure representation, while CDPE incorporates depth-aware geometric cues into the positional

representation space, thereby improving the localization capability for weak-textured and low-contrast tiny objects.

(3) An RGB-D pavement debris dataset containing 5,635 image pairs is constructed under multiple illumination conditions in aviation warehouse inspection scenes. Comparative experiments, ablation studies, and illumination robustness evaluations are conducted to verify the effectiveness of the proposed method.

2 Related work

2.1 Real-time object detection for robotic inspection

Real-time object detection is a fundamental component of robotic inspection and autonomous perception systems. In recent years, convolutional neural network (CNN)-based detectors have been widely used in industrial and robotic vision applications because of their strong feature representation capability and high inference efficiency. Two-stage detectors, such as Faster R-CNN, improve detection accuracy by generating region proposals before classification and regression [11]. However, their relatively high computational cost may limit their applicability in time-sensitive robotic inspection scenarios. In contrast, one-stage detectors represented by the YOLO series directly predict object categories and bounding boxes from dense feature maps, enabling high-speed inference and making them widely used in real-time perception tasks.

Although YOLO-based methods achieve favorable detection speed, they usually rely on post-processing procedures such as non-maximum suppression (NMS) to remove redundant predictions. In complex scenes with small, dense, or partially overlapping targets, improper suppression may lead to missed detections or unstable localization [16]. To reduce the dependence on hand-crafted post-processing, Transformer-based detectors have introduced an end-to-end detection paradigm. DETR formulates object detection as a set prediction problem and uses object queries to directly predict detection results [17]. Subsequent variants, such as Deformable DETR, improve convergence speed and computational efficiency by introducing deformable attention mechanisms [12]. RT-DETR further improves the balance between accuracy and efficiency through an efficient hybrid encoder and a query selection strategy [13]. Owing to its end-to-end structure and lightweight variants, RT-DETR provides a suitable baseline for real-time machine vision applications.

In this study, RT-DETR with a ResNet-18 backbone is selected as the baseline detector because it provides a relatively lightweight structure while maintaining the advantages of Transformer-based global feature modeling [18]. Nevertheless, the original RT-DETR framework is mainly designed for RGB-based object detection and does not explicitly exploit geometric information from depth sensors. When RGB appearance features are degraded by low illumination, shadows, or weak texture, the robustness of RGB-only detection is limited. Therefore, additional depth-guided representation is required for tiny pavement debris detection in aviation warehouse inspection scenes.

2.2 RGB-D and depth-guided machine vision

Multisensor machine vision has attracted increasing attention in robotic perception because different sensing modalities can provide complementary information under challenging environmental conditions. Compared with RGB-only perception, RGB-D sensing introduces depth measurements that describe the spatial structure of the scene. RGB images provide color, texture, and semantic appearance cues, whereas depth maps provide geometric information such as object distance, height variation, and surface discontinuity. Such complementary information is particularly valuable in scenes where visual appearance is unreliable.

Existing studies have demonstrated the effectiveness of incorporating structural or depth-related information into visual perception models. For example, MonoDETR uses depth-related geometric cues to enhance spatial reasoning in monocular 3D object detection, indicating the importance of geometric priors for visual understanding [19]. Other studies have explored Transformer-based fusion strategies for heterogeneous sensors, such as cameras and millimeter-wave radar, to model cross-modal spatial dependencies and improve obstacle perception in complex environments [20, 21]. These works show that auxiliary sensing modalities can improve perception robustness by compensating for the limitations of visible-light images.

For pavement debris detection, depth information is especially useful because tiny objects often differ from the ground plane in local geometric structure even when their RGB appearance is weak or similar to the background. Depth maps can highlight height discontinuities and spatial relationships that are difficult to capture from RGB images alone. Therefore, RGB-D perception provides a promising solution for detecting weak-textured and low-contrast debris in warehouse inspection scenes. However, the effectiveness of RGB-D perception depends heavily on the quality and utilization of depth information.

2.3 Low-illumination tiny object detection in warehouse scenes

Tiny pavement debris detection in aviation warehouses is more challenging than general object detection. First, debris objects are usually small and occupy only a limited number of pixels in the image, which makes their visual features easily weakened during feature downsampling and multi-scale fusion [7]. Second, many debris categories, such as plastic bags, paper fragments, cigarette butts, and small hardware items, have low contrast with warehouse pavements. In particular, transparent or semi-transparent objects may appear highly similar to the background in RGB images, making their boundaries difficult to distinguish under complex illumination conditions [10]. Third, aviation warehouse environments often contain dense shelves, uneven illumination, shadows, and local low-light regions, which further reduce RGB image quality and cause blurred boundaries or missing texture details [8, 9].

Under low-illumination conditions, RGB-based detectors may suffer from unstable semantic feature extraction, missed detections, and inaccurate localization. Low-light image enhancement and attention-based feature fusion strategies can partially improve visual representation by enhancing image contrast or strengthening discriminative features [14, 15]. However, these methods still mainly rely on visible appearance

information and cannot fully recover geometric cues that are absent or unreliable in degraded RGB images. RGB-D sensing can provide complementary spatial information for object detection and scene understanding by introducing depth-based geometric structure [19, 20]. Nevertheless, raw depth observations are also affected by sensor noise, missing measurements, reflective materials, sensing distance, and geometric discontinuities. Directly fusing noisy depth maps with RGB features may introduce unreliable responses into the detection pipeline and degrade feature representation quality [9].

Moreover, many existing RGB-D detection methods mainly use depth maps as an additional input feature and pay insufficient attention to depth quality enhancement and depth-aware geometric representation. For tiny debris detection, reliable localization requires not only semantic discrimination but also accurate spatial representation. Depth information can describe height variations and surface discontinuities between tiny objects and the ground plane, but its effectiveness depends on whether reliable geometric cues can be extracted and properly incorporated into the detection framework [19]. Therefore, it is necessary to improve the usability of raw depth observations and explicitly incorporate depth-guided geometric cues into the detection framework.

To address these issues, this paper develops a depth-guided RGB-D detection framework based on RT-DETR for tiny pavement debris detection in low-illumination aviation warehouse inspection scenes. The proposed method introduces a Depth Enhancement Module (DEM) to improve the representation quality of raw depth information and a Continuous Depth Position Encoding (CDPE) strategy to incorporate depth-aware geometric cues into Transformer-based feature learning. In this way, the framework aims to improve the robustness of tiny object detection under visual degradation.

3 Proposed model

This section presents the proposed depth-guided RGB-D detection framework for tiny pavement debris detection in low-illumination aviation warehouse inspection scenes. The framework is built upon RT-DETR and introduces depth-guided representation learning to improve detection robustness when RGB appearance features are degraded by shadows, weak texture, and uneven illumination. First, the baseline RT-DETR detector is briefly introduced. Then, the overall architecture of the proposed PDE-RT-DETR is described. Finally, the Depth Enhancement Module (DEM) and the Continuous Depth Position Encoding (CDPE) strategy are presented in detail.

3.1 Baseline RT-DETR detector

RT-DETR is an end-to-end real-time object detector based on a Transformer detection framework [13]. Its architecture mainly consists of a backbone network, an efficient hybrid encoder, a query selection module, and a Transformer decoder with prediction heads, as shown in Fig. 1.

In the baseline detector, the backbone network extracts multi-scale visual features from the input RGB image. In this study, ResNet-18 is adopted as the backbone because of its relatively lightweight structure and stable feature extraction capability

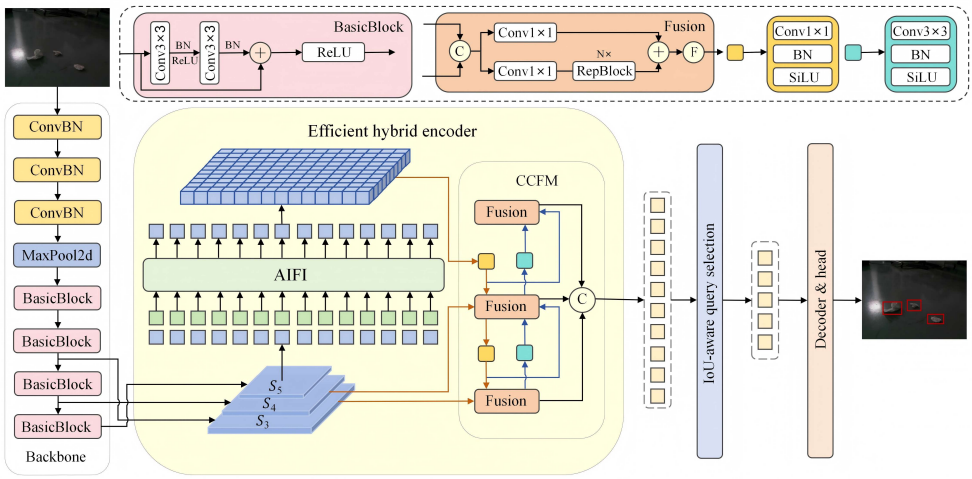


Fig. 1 RT-DETR model structure

[18]. The extracted multi-scale features are denoted as S_3, S_4, and S_5. These features are then fed into the efficient hybrid encoder, which contains an attention-based intra-scale feature interaction module and a convolution-based cross-scale feature fusion module. The attention-based module captures global semantic dependencies from high-level features, whereas the cross-scale fusion module aggregates information from different feature resolutions to enhance multi-scale representation.

After feature encoding, RT-DETR uses an IoU-aware query selection strategy to generate high-quality object queries for the Transformer decoder. The decoder further refines the selected queries and outputs object categories and bounding boxes through the prediction heads. Since this detection paradigm avoids hand-crafted non-maximum suppression, it provides a compact end-to-end framework for real-time machine vision applications.

Although RT-DETR provides an effective detection framework, the original model is mainly designed for RGB-based object detection. It relies heavily on visible appearance features, such as color, texture, and object boundaries. In low-illumination warehouse scenes, these RGB features may become unstable or indistinguishable from the pavement background. Moreover, the original framework does not explicitly use geometric information from depth sensors. Therefore, it is necessary to introduce depth-guided representation learning to improve the robustness of tiny pavement debris detection under visual degradation.

3.2 Overall architecture of PDE-RT-DETR

To address the limitations of RGB-only perception under low-illumination and weak-texture conditions, this paper proposes a depth-guided RGB-D detection framework, termed PDE-RT-DETR. The overall architecture is shown in Fig. 2. The proposed

framework uses RGB images to extract semantic appearance features and depth observations to provide complementary geometric cues, aiming to improve the detection reliability of tiny pavement debris in complex aviation warehouse inspection scenes.

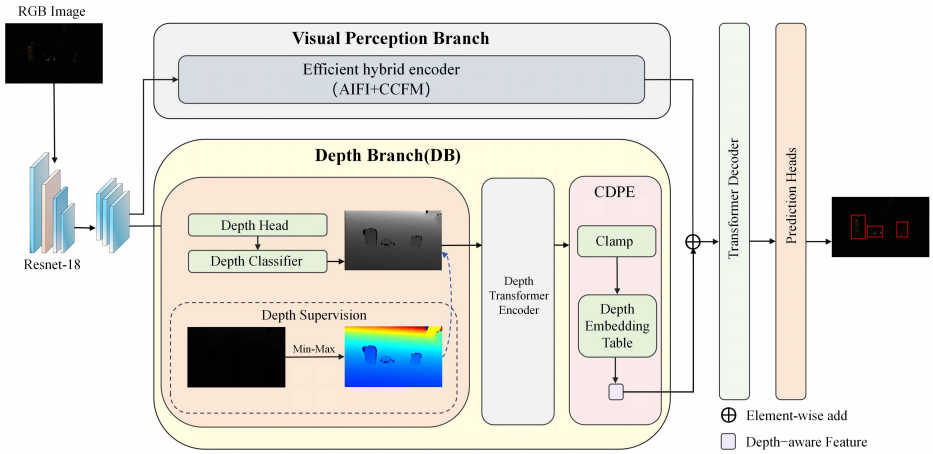


Fig. 2 Overall architecture of the proposed PDE-RT-DETR framework

Given an aligned RGB-D image pair, the RGB image is first processed by the ResNet-18 backbone to extract multi-scale visual features. These features are then forwarded to two parallel branches: the visual perception branch and the depth-guided branch. In this study, the depth-guided branch is denoted as the depth branch (DB), which refers to the auxiliary pathway that introduces aligned depth observations into the RT-DETR framework to provide geometric information. In the ablation study, DB represents the basic depth-information introduction setting without the proposed DEM and CDPE modules.

The visual perception branch follows the efficient hybrid encoder of RT-DETR and preserves the original semantic feature extraction capability of the baseline detector. This branch provides discriminative visual representations for object category recognition, especially under normal or relatively favorable illumination conditions.

The depth-guided branch is introduced to enhance geometric representation and compensate for the degradation of RGB appearance features. Instead of directly concatenating raw depth maps with RGB features, the proposed framework first improves the reliability of depth-related representations through the Depth Enhancement Module (DEM). DEM reconstructs depth responses from aggregated multi-scale features under depth supervision and generates an enhanced depth response containing more stable geometric structure information. This design helps reduce the influence of raw depth noise, missing measurements, and unstable local depth responses.

Based on the enhanced depth representation, Continuous Depth Position Encoding (CDPE) is further used to incorporate depth-aware geometric cues into the positional representation space. CDPE maps continuous depth values into learnable depth embeddings and injects the resulting depth-aware positional information into

the Transformer-based feature representation. In this way, the detector can jointly exploit RGB semantic information and depth-based geometric structure for tiny object localization.

Finally, the depth-aware features are fused with the visual features through element-wise addition and then fed into the Transformer decoder and prediction heads. This fusion strategy preserves the semantic representation capability of RGB features while introducing complementary geometric information from depth observations. Compared with direct RGB-D feature concatenation, the proposed depth-guided design focuses on improving the usability of depth information and its contribution to spatial localization. Therefore, PDE-RT-DETR is better suited for tiny pavement debris detection under low-illumination and weak-texture conditions.

3.3 Depth Enhancement Module

Depth information provides important geometric cues for tiny pavement debris detection, especially when RGB appearance features are degraded by low illumination, shadows, or weak texture. However, depth observations in aviation warehouse scenes are often affected by measurement noise, missing values, reflective materials, sensing distance, and local depth discontinuities. Directly using unreliable depth information may introduce unstable geometric responses into the detection framework. To improve the quality and usability of depth-guided representation, a Depth Enhancement Module (DEM) is introduced into the proposed framework. The structure of DEM is shown in Fig. 3.

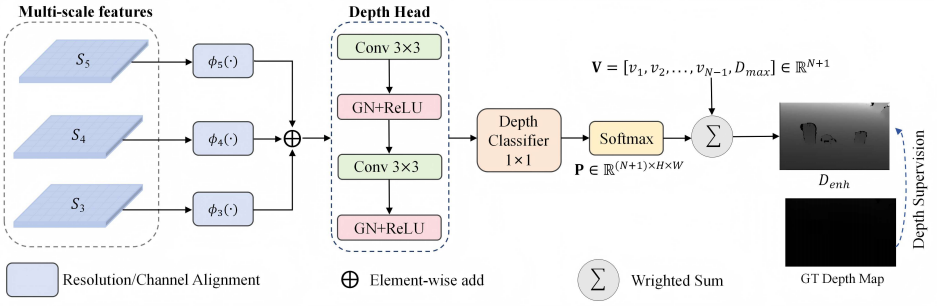


Fig. 3 Structure of the DEM

As shown in Fig.3, DEM first aggregates multi-scale feature maps extracted from the backbone network. Let the multi-scale features be denoted as S_3, S_4, and S_5. These feature maps contain spatial details and semantic information at different resolutions. To combine local geometric cues and high-level contextual information, the feature maps are first aligned to a unified spatial resolution and then fused through element-wise aggregation. The aggregated feature map can be written as:

$$\mathbf{F}_{\text{agg}} = \frac{1}{3}\left(\phi_3(S_3) + \phi_4(S_4) + \phi_5(S_5)\right) \tag{1}$$

where $\Phi_3(\cdot)$, $\Phi_4(\cdot)$, and $\Phi_5(\cdot)$ denote the corresponding resolution and channel alignment operations for different feature levels. Through multi-scale aggregation, DEM integrates fine-grained spatial information and high-level semantic features before depth response prediction.

The aggregated feature map $\mathbf{F}_{\text{agg}}$ is then fed into a depth head composed of convolutional blocks and nonlinear activation functions. The depth head enhances local structural representation and generates depth-aware features. Subsequently, a depth classifier predicts interval-wise depth logits, which can be expressed as:

$$\mathbf{Z} = \mathcal{C}_d\left(\mathcal{H}_d\left(\mathbf{F}_{\text{agg}}\right)\right) \tag{2}$$

where $\mathcal{H}_d(\cdot)$ denotes the depth head, $\mathcal{C}_d(\cdot)$ denotes the depth classifier, and $\mathbf{Z} \in \mathbb{R}^{(N+1)\times H\times W}$ represents the predicted logits over the predefined depth support values.

To transform the continuous depth estimation problem into a structured probability estimation problem, DEM adopts a nonlinear depth discretization strategy inspired by ordinal depth representation [22, 23]. The normalized depth range $[D_{\min}, D_{\max}]$ is divided into N nonlinearly distributed bins. The interval step size S_{bin} is defined as:

$$S_{\text{bin}} = \frac{2\left(D_{\max} - D_{\min}\right)}{N(1+N)} \tag{3}$$

The center value v_i of the i-th depth bin is calculated as:

$$v_i = \frac{(i+0.5)^2 S_{\text{bin}}}{2} - \frac{S_{\text{bin}}}{8} + D_{\min} \tag{4}$$

where $i \in \{0, 1, \ldots, N-1\}$ denotes the depth interval index. By concatenating $D_{\max}$ to the bin centers, the complete depth support vector is obtained as:

$$\mathbf{V} = [v_0, v_1, \ldots, v_{N-1}, D_{\max}] \in \mathbb{R}^{N+1} \tag{5}$$

Compared with uniform discretization, the nonlinear binning strategy assigns denser depth support values to near-range regions. This design is beneficial for representing subtle depth variations between tiny pavement debris and the ground plane.

The depth logits are normalized by the softmax function to obtain the probability distribution over the depth support values:

$$\mathbf{P} = \text{Softmax}(\mathbf{Z}) \tag{6}$$

where $\mathbf{P} \in \mathbb{R}^{(N+1)\times H\times W}$. The enhanced depth response D_{enh} is reconstructed by calculating the expectation of the predefined depth support values:

$$D_{\text{enh}}(h, w) = \sum_{i=0}^{N} P_i(h, w) V_i \tag{7}$$

where $P_i(h, w)$ denotes the probability assigned to the i-th depth support value at spatial position (h, w), and V_i denotes the corresponding depth support value.

Through this probabilistic reconstruction process, DEM converts depth-guided representation learning into interval-wise depth distribution estimation. Instead of

directly concatenating raw depth maps with RGB features, DEM produces a structured enhanced depth response under depth supervision. The aligned ground-truth depth map is used to constrain the predicted depth response during training, guiding the module to learn geometric representations that are consistent with the spatial structure of the scene.

The enhanced depth response provides more stable geometric guidance for subsequent depth-aware positional encoding and multimodal feature fusion. In low-illumination aviation warehouse scenes, where RGB texture and boundary information are often weakened, this depth-enhanced representation helps the detector preserve spatial cues of tiny pavement debris and improves localization robustness.

3.4 Continuous Depth Position Encoding

Transformer-based detectors rely on positional information to distinguish spatial locations during feature interaction [24]. Conventional two-dimensional positional encoding mainly describes the horizontal and vertical locations of image features, but it does not explicitly represent depth-related geometric relationships [16]. Previous studies have shown that depth-aware or 3D-aware positional encoding can provide useful spatial priors for Transformer-based visual perception [25]. For tiny pavement debris detection in low-illumination warehouse scenes, depth information provides useful spatial cues because debris objects usually introduce local geometric discontinuities relative to the ground plane. Therefore, a Continuous Depth Position Encoding (CDPE) strategy is introduced to incorporate enhanced depth responses into the positional representation space.

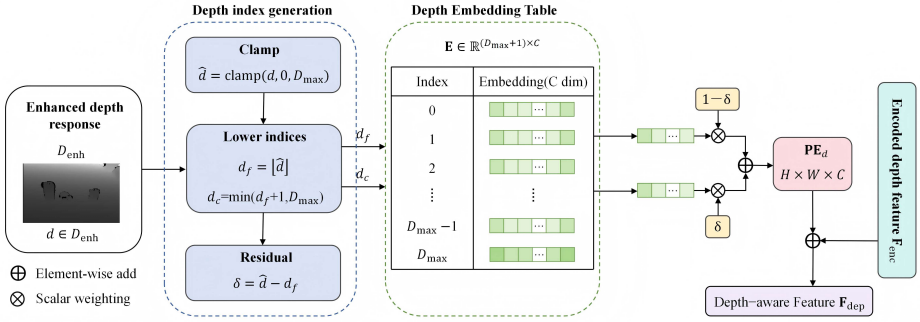


Fig. 4 Structure of the CDPE

Given the enhanced depth response D_{enh} generated by DEM, CDPE first constrains each continuous depth value into the valid depth range. For a depth value $d \in D_{\text{enh}}$, the clamped depth value is calculated as

$$\hat{d} = \text{clamp}(d, 0, D_{\max}) \tag{8}$$

The lower integer depth index d_f is obtained by applying the floor operation to $\hat{d}$:

$$d_f = \left\lfloor \hat{d} \right\rfloor \tag{9}$$

The upper adjacent depth index d_c is then defined as

$$d_c = \min(d_f + 1, D_{\max}) \tag{10}$$

Here, $D_{\max}$ denotes the maximum valid depth index after depth normalization. The residual δ between the continuous depth value and the lower integer index is calculated as

$$\delta = \hat{d} - d_f \tag{11}$$

To generate continuous depth-aware positional embeddings, a learnable depth embedding table $\mathbf{E} \in \mathbb{R}^{(D_{\max}+1)\times C}$ is constructed, where C denotes the embedding dimension.

For each depth value, the embeddings corresponding to d_f and d_c are linearly interpolated according to δ:

$$\mathbf{PE}_d = \mathbf{E}(d_f)(1-\delta) + \mathbf{E}(d_c)\delta \tag{12}$$

Here, $\mathbf{E}(d_f)$ and $\mathbf{E}(d_c)$ denote the learnable embeddings of the lower and upper adjacent depth indices, respectively. Compared with directly assigning each depth value to a discrete embedding, the interpolation strategy preserves the continuity of depth variations and avoids abrupt changes caused by hard quantization.

After interpolation, the generated depth-aware positional embedding $\mathbf{PE}_d$ is reshaped to match the spatial dimension of the depth feature map. Let $\mathbf{F}_{\text{enc}}$ denote the feature representation output by the depth Transformer encoder. The depth-aware feature is obtained by adding the continuous depth positional embedding to the encoded depth feature:

$$\mathbf{F}_{\text{dep}} = \mathbf{F}_{\text{enc}} + \mathbf{PE}_d \tag{13}$$

Through this additive integration, CDPE introduces depth-related geometric cues into the feature representation without disrupting the original semantic feature extraction process. The resulting depth-aware feature provides complementary spatial information for subsequent multimodal fusion with the visual branch.In low-illumination and weak-texture scenes, this strategy helps preserve depth-related spatial relationships between tiny pavement debris and the ground plane. As a result, the detector can obtain more stable localization cues when RGB texture and boundary information are degraded.

4 Experimental and discussion

4.1 Datasets

4.1.1 Self-built RGB-D multi-illumination dataset

Reliable detection of tiny pavement debris is important for autonomous guided vehicle (AGV) inspection in aviation warehouse environments. Small objects on the ground

may interfere with robot locomotion, local perception, and path execution, especially when the illumination condition is poor. Although several public RGB-D datasets, such as SUN RGB-D [26] and NYU Depth V2 [27], have been widely used for indoor scene understanding, they are mainly designed for general indoor perception tasks rather than tiny pavement debris detection in warehouse inspection scenes [28]. Existing RGB-D warehouse-related datasets mainly focus on object detection or pose estimation for regular objects in pick-and-place scenarios, and provide limited coverage of small-scale pavement debris under low-illumination conditions [9]. Therefore, these datasets are not specifically suitable for evaluating RGB-D tiny debris detection in complex aviation warehouse inspection scenes.

To address this limitation, a self-built RGB-D pavement debris dataset was constructed for multi-illumination aviation warehouse inspection scenarios. An Angstrong HP60C industrial RGB-D camera was used for data acquisition. The camera provides RGB and depth observations, which were spatially aligned during preprocessing to ensure pixel-level correspondence between the two modalities. A total of 5,635 RGB-D image pairs were collected at a resolution of 1920×1080 pixels. The dataset covers eight typical categories of pavement debris commonly found in warehouse environments, including plastic bottles, paper balls, cans, plastic bags, cigarette butts, food packaging, paperboard, and small hardware items.

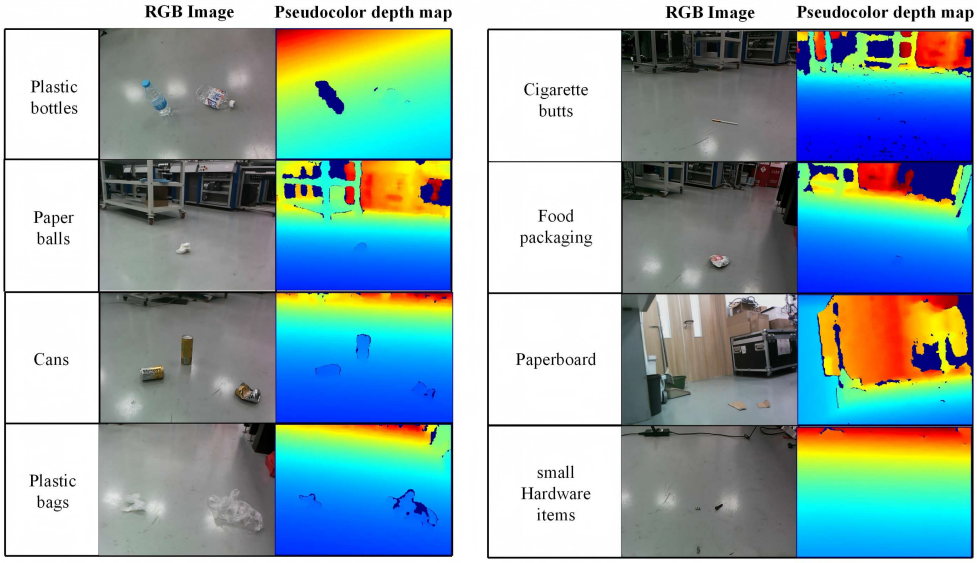


Fig. 5 Representative RGB-D samples of eight pavement debris categories in the self-built dataset

The raw depth image is stored as a 16-bit single-channel image. Since the gray-level differences between tiny debris and the ground plane are often subtle, direct visualization of the raw depth map may fail to clearly present local geometric variations. Therefore, the raw depth values were normalized using Min-Max normalization

and mapped to the range of 0–255. For visualization and sample presentation, a JET colormap was applied to generate pseudocolor depth maps, transforming small depth variations into more distinguishable color gradients. This preprocessing helps visualize the geometric differences between small debris and the pavement background while preserving the original depth structure for RGB-D perception.

4.1.2 Multi-grade illumination samples

Aviation warehouses usually contain dense shelves, large indoor spaces, and non-uniform lighting layouts. Local shadows, shelf occlusion, partially illuminated regions, and energy-saving lighting modes may lead to significant illumination variation on warehouse pavements. Under such conditions, RGB image quality may degrade due to reduced signal-to-noise ratio, weakened texture, and blurred object boundaries [29, 30]. Low-illumination degradation is particularly challenging for tiny objects because their visual features are already limited by small scale. In addition, transparent or semi-transparent objects, such as plastic bags, may exhibit weak boundaries, specular reflection, or low contrast with the background, which further increases the difficulty of visual detection [10].

To evaluate the robustness of the proposed method under different illumination conditions, the collected samples were divided into three illumination groups: normal illumination, low illumination, and extreme low illumination. The three groups were designed to simulate typical lighting states in aviation warehouse inspection scenes.

(1) Normal illumination. This group simulates warehouse scenes in which the main lighting system is fully activated. RGB images usually contain relatively clear texture, color, and boundary information, and the contrast between debris and the pavement background is comparatively high. This group is used to evaluate the basic detection capability of the model under favorable visual conditions.

(2) Low illumination. This group simulates partially illuminated warehouse regions, such as areas affected by shelf shadows or auxiliary lighting. In this condition, RGB images exhibit reduced brightness and weakened boundary information. Small debris with weak texture or low contrast becomes more difficult to distinguish from the ground background. Depth information can provide complementary geometric cues for such visually degraded cases.

(3) Extreme low illumination. This group simulates severely degraded lighting conditions, where the main illumination is unavailable or only weak ambient light remains. RGB images contain limited effective appearance information, and tiny debris may be difficult to locate using visible cues alone. This group is used to evaluate whether depth-guided representation can improve robustness when RGB appearance features are severely degraded.

Representative samples of transparent or weak-textured objects under different illumination conditions are shown in Fig. 6.

To increase the coverage of visually degraded scenes, the dataset was constructed with a larger proportion of low-illumination and extreme low-illumination samples. The approximate ratio of normal, low, and extreme low-illumination samples is 1:1:2. This design increases the representation of difficult low-light cases and provides sufficient samples for evaluating the robustness of RGB-D detection methods

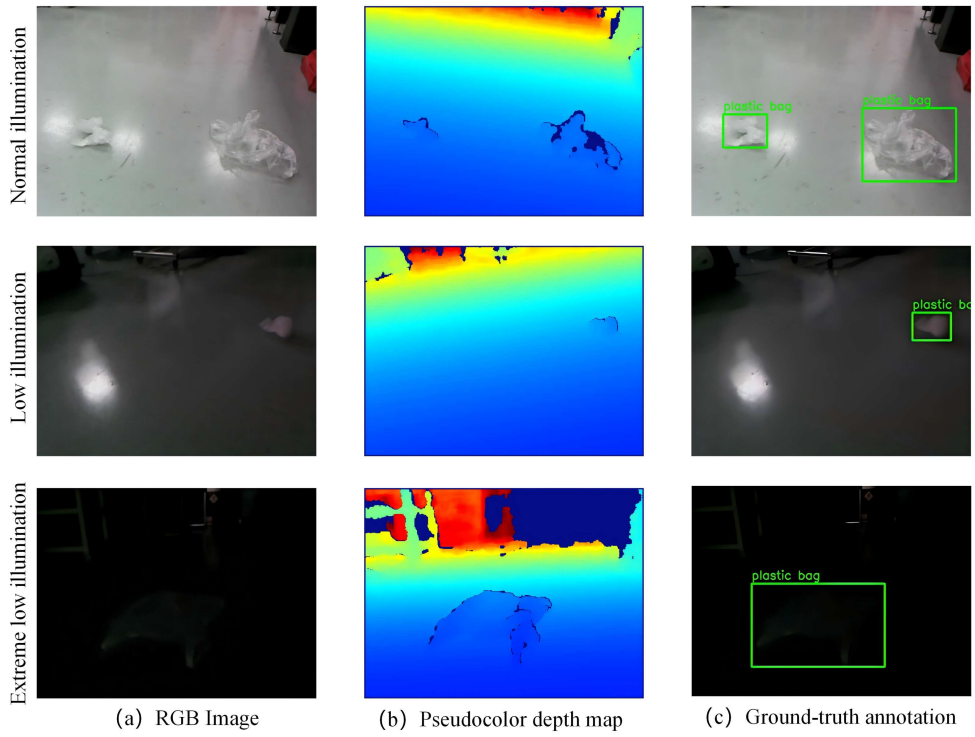


Fig. 6 Representative transparent or weak-textured debris samples under different illumination conditions

under illumination degradation. The distribution of target instances across different illumination conditions is shown in Fig. 7.

For annotation, LabelImg was used to generate bounding box labels for the eight pavement debris categories. To support RGB-D feature learning, RGB images and depth maps were spatially aligned during preprocessing so that corresponding pixels in the two modalities represent the same scene location. To reduce information leakage caused by adjacent frames or highly similar samples from the same scene, the collected data were organized according to illumination conditions and capture scenes before partitioning. The final dataset was divided into training, validation, and test sets at a ratio of 7:1.5:1.5, providing a reliable basis for evaluating the effectiveness and illumination robustness of the proposed PDE-RT-DETR framework.

4.2 Experimental environment

All experiments were conducted on a workstation equipped with an Intel Xeon W-3245 CPU running at 3.20 GHz, 128 GB of system RAM, and two NVIDIA RTX 3090 Turbo GPUs, each with 24 GB of GPU memory. The operating system was Ubuntu 20.04. All models were implemented using Python 3.7 and PyTorch 1.8.0, with CUDA 11.1 for GPU acceleration.

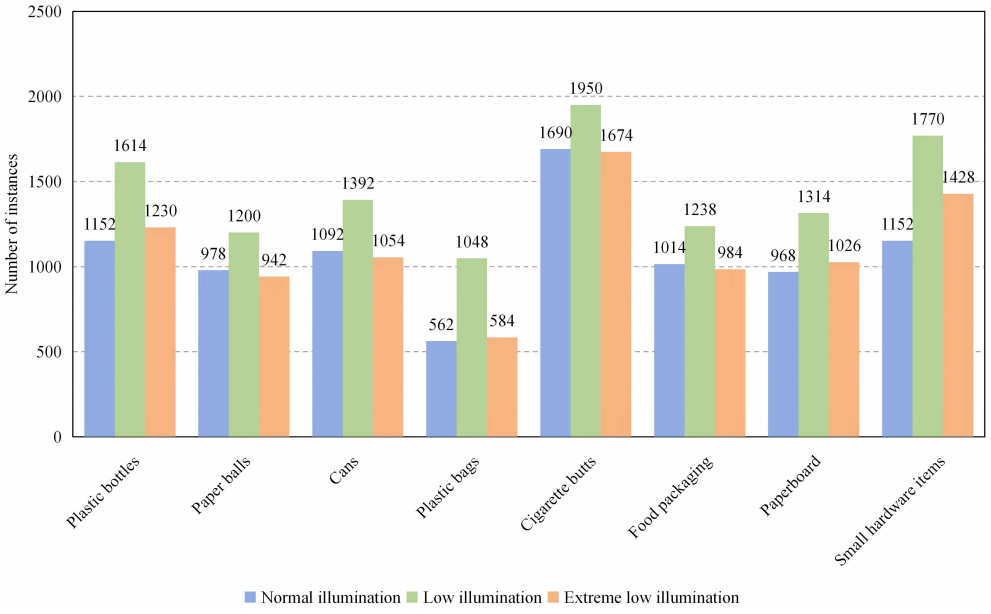


Fig. 7 Distribution of target instances under different illumination conditions

Table 1 Experimental platform and software configuration

Category	Item	Specification
Hardware	CPU	Intel Xeon W-3245 @ 3.20 GHz
	GPU	2 × NVIDIA RTX 3090 Turbo
	GPU memory	24 GB per GPU
	System memory	128 GB RAM
	Storage	2 TB
Software	OS	Ubuntu 20.04
	CUDA	11.1
	Python	3.7
	PyTorch	1.8.0

Unless otherwise specified, all models in the ablation experiments, comparison experiments, and illumination robustness evaluation were trained and tested using the same dataset split and evaluation protocol to ensure fair comparison. The input images were resized to 640×640 during training. AdamW was adopted as the optimizer, with an initial learning rate of 0.0002, a weight decay of 0.0005, and a momentum coefficient of 0.9. The batch size was set to 16, and all models were trained for 100 epochs.

To improve reproducibility, the random seed was fixed during training, and deterministic settings were enabled where applicable. The experimental platform is summarized in Table 1, and the main training hyperparameters are listed in Table 2.

Table 2 Training hyperparameters used in the experiments

Hyperparameter	Value
Initial learning rate	0.0002
Momentum	0.900
Weight decay	0.0005
Batch size	16
Optimizer	AdamW
Input image size	640×640
Epochs	100

4.3 Evaluation metrics

To objectively evaluate the performance of different models in tiny pavement debris detection, precision (P), recall (R), $\mathrm{mAP}_{0.50}$, and $\mathrm{mAP}_{0.50:0.95}$ are adopted as the main evaluation metrics. Precision measures the proportion of correctly detected objects among all predicted positive samples, while recall measures the proportion of correctly detected objects among all ground-truth objects. They are defined as follows:

$$P = \frac{TP}{TP + FP} \tag{14}$$

$$R = \frac{TP}{TP + FN} \tag{15}$$

where TP, FP, and FN denote true positives, false positives, and false negatives, respectively.

Average precision (AP) is calculated from the precision–recall curve for each category. The mean average precision at an Intersection over Union (IoU) threshold of 0.50 is defined as

$$\mathrm{mAP}_{0.50} = \frac{1}{K} \sum_{i=1}^{K} \mathrm{AP}_{i,0.50} \tag{16}$$

where K denotes the number of debris categories, and $\mathrm{AP}_{i,0.50}$ represents the average precision of the i-th category when the Intersection over Union (IoU) threshold is set to 0.50.

To further evaluate localization accuracy under stricter overlap requirements, $\mathrm{mAP}_{0.50:0.95}$ is also adopted following the COCO evaluation protocol. It is calculated by averaging AP values over 10 IoU thresholds from 0.50 to 0.95 with a step size of 0.05:

$$\mathrm{mAP}_{0.50:0.95} = \frac{1}{10K} \sum_{j=0}^{9} \sum_{i=1}^{K} \mathrm{AP}_{i,\,0.50+0.05j} \tag{17}$$

Compared with $\mathrm{mAP}_{0.50}$, $\mathrm{mAP}_{0.50:0.95}$ provides a more comprehensive evaluation of both detection accuracy and bounding-box localization quality. Therefore, these

metrics are used together to assess the effectiveness of the proposed PDE-RT-DETR and the compared detection models.

4.4 Ablation experiment

To verify the effectiveness of the main components in the proposed PDE-RT-DETR framework, ablation experiments were conducted based on the RT-DETR-R18 baseline. The baseline model uses only RGB images as input. Based on this baseline, the depth branch (DB), Depth Enhancement Module (DEM), and Continuous Depth Position Encoding (CDPE) were progressively introduced to evaluate their individual contributions. The experimental results are summarized in Table 3.

Table 3 Ablation study of PDE-RT-DETR components

No.	DB	DEM	CDPE	$mAP_{0.50}$ (%)	$mAP_{0.50:0.95}$ (%)
1	–	–	–	77.5	49.6
2	✓	–	–	82.2	53.5
3	✓	✓	–	87.4	58.2
4	✓	✓	✓	91.4	66.8

As shown in Table 3, the RGB-only baseline achieves 77.5% $mAP_{0.50}$ and 49.6% $mAP_{0.50:0.95}$. After introducing the depth branch, $mAP_{0.50}$ increases from 77.5% to 82.2%, and $mAP_{0.50:0.95}$ increases from 49.6% to 53.5%. This improvement indicates that the aligned depth modality provides complementary geometric information for tiny pavement debris detection, especially when RGB appearance features are degraded by low illumination or weak texture.

When DEM is further incorporated, the model achieves 87.4% $mAP_{0.50}$ and 58.2% $mAP_{0.50:0.95}$, with improvements of 5.2 and 4.7 percentage points over the DB-only setting, respectively. This result demonstrates that simply introducing depth information is not sufficient, while enhancing the depth representation through nonlinear depth discretization and re-encoding can improve the usability of depth-related features and provide more reliable geometric cues for multimodal detection.

After integrating CDPE, the complete PDE-RT-DETR achieves the best performance, with 91.4% $mAP_{0.50}$ and 66.8% $mAP_{0.50:0.95}$. Compared with the configuration using DB and DEM, CDPE brings further improvements of 4.0 percentage points in $mAP_{0.50}$ and 8.6 percentage points in $mAP_{0.50:0.95}$. The larger improvement in $mAP_{0.50:0.95}$ suggests that CDPE contributes not only to object recognition but also to more accurate bounding-box localization. This is because CDPE introduces continuous depth-aware positional information into the Transformer feature representation, strengthening the model's spatial geometric modeling ability.

Overall, compared with the RGB-only baseline, the complete PDE-RT-DETR improves $mAP_{0.50}$ from 77.5% to 91.4% and $mAP_{0.50:0.95}$ from 49.6% to 66.8%. These results verify that the three components are complementary: DB introduces depth-based geometric information, DEM improves the quality of depth representation,

and CDPE further enhances depth-aware spatial localization. Together, they effectively improve the detection performance of PDE-RT-DETR in complex warehouse inspection scenes.

4.5 Comparison experiment

To further evaluate the effectiveness of the proposed PDE-RT-DETR, representative RGB-based and RGB-D detection methods were selected for comparison. The compared methods include Faster R-CNN [11], DEYOLO [31], SSD [32], YOLOv8n, YOLOv10n, YOLOv11n, and RT-DETR-R18. Specifically, Faster R-CNN and SSD were used as representative classical two-stage and one-stage detectors, respectively. YOLOv8n, YOLOv10n, YOLOv11n, and RT-DETR-R18 were selected as representative real-time detection models. DEYOLO was introduced as a representative RGB-D cross-modal detection method. All models were trained and evaluated on the same self-built RGB-D pavement debris dataset using the evaluation metrics defined in Section 4.3. The quantitative results are shown in Table 4.

Table 4 Performance comparison of different detection models on the self-built dataset

Model	Modality	Precision (%)	Recall (%)	$mAP_{0.50}$ (%)	$mAP_{0.50:0.95}$ (%)
Faster R-CNN	RGB	59.2	67.6	70.2	44.1
SSD	RGB	64.2	61.6	68.6	43.9
YOLOv8n	RGB	68.4	74.1	73.8	43.7
	RGB-D	76.9	80.5	82.5	51.1
YOLOv10n	RGB	72.8	67.0	75.6	53.1
	RGB-D	76.6	66.1	76.2	54.8
YOLOv11n	RGB	79.7	78.5	84.6	59.4
	RGB-D	80.1	82.7	88.4	60.5
DEYOLO	RGB-D	74.5	82.7	82.3	54.6
RT-DETR-R18	RGB	72.7	73.5	77.5	49.6
	RGB-D	77.4	74.9	82.2	53.5
PDE-RT-DETR	RGB-D	**86.4**	**87.6**	**91.4**	**66.8**

The comparison between RGB and RGB-D variants also demonstrates the usefulness of depth information for tiny pavement debris detection. For example, YOLOv8n improves from 73.8% to 82.5% in $mAP_{0.50}$ after introducing depth information, and RT-DETR-R18 improves from 77.5% to 82.2%. These results suggest that depth observations provide complementary geometric cues for visually degraded or weak-textured debris. However, direct RGB-D input does not always lead to large performance gains, as shown by the limited improvement of YOLOv10n. This indicates that the effectiveness of RGB-D detection depends not only on the availability of depth information but also on how depth features are represented and fused.

Compared with the strongest RGB-D YOLO variant, YOLOv11n RGB-D, PDE-RT-DETR improves $mAP_{0.50}$ and $mAP_{0.50:0.95}$ by 3.0 and 6.3 percentage points, respectively. Compared with DEYOLO, PDE-RT-DETR achieves improvements of

9.1 and 12.2 percentage points in $mAP_{0.50}$ and $mAP_{0.50:0.95}$, respectively. The larger gain in $mAP_{0.50:0.95}$ shows that the proposed method improves not only object classification and detection confidence but also bounding-box localization accuracy under complex warehouse conditions.



Fig. 8 Qualitative comparison of detection results obtained by different models in a low-illumination warehouse scene

The qualitative results in Fig. 8 further illustrate the detection behavior of different methods. RGB-based detectors tend to produce missed detections or inaccurate localization for low-contrast and weak-textured debris, whereas RGB-D methods generally provide more stable detection results. Among them, PDE-RT-DETR achieves more complete and accurate bounding boxes, demonstrating the effectiveness of combining enhanced depth representation with continuous depth-aware positional encoding for tiny pavement debris detection.

4.6 Robustness evaluation under complex illumination conditions

To further evaluate the robustness of the proposed PDE-RT-DETR under illumination variations, we compare its detection performance with YOLOv8n and RT-DETR-R18 under three illumination conditions: normal illumination, low illumination, and extreme low illumination. RT-DETR-R18 is used as the RGB baseline, YOLOv8n is selected as a representative lightweight real-time detector, and PDE-RT-DETR represents the proposed RGB-D detection framework. The comparison is conducted on the self-built RGB-D pavement debris dataset.

To quantitatively measure the sensitivity of different models to illumination degradation, a performance degradation rate Δ is introduced. It measures the relative decrease in $\mathrm{mAP}_{0.50}$ from normal illumination to extreme low illumination. A smaller Δ indicates lower performance fluctuation and stronger robustness under illumination changes. It is defined as

$$\Delta = \frac{\mathrm{mAP}_{0.50}^{\mathrm{Normal}} - \mathrm{mAP}_{0.50}^{\mathrm{Extreme}}}{\mathrm{mAP}_{0.50}^{\mathrm{Normal}}} \times 100\% \tag{18}$$

As shown in Table 5, all models exhibit performance degradation as the illumination condition becomes worse. RT-DETR-R18 shows the largest performance drop, with $\mathrm{mAP}_{0.50}$ decreasing from 74.5% under normal illumination to 44.9% under extreme low illumination, corresponding to a degradation rate of 39.73%. YOLOv8n achieves 84.1% $\mathrm{mAP}_{0.50}$ under normal illumination, but decreases to 64.3% under extreme low illumination, with a degradation rate of 23.54%. These results indicate that RGB-only detectors are sensitive to low illumination, because weakened texture, reduced contrast, and blurred boundaries make tiny debris difficult to distinguish from the pavement background.

Table 5 Robustness comparison of detection models under different illumination conditions

Model	Modality	Normal illumination	Low illumination	Extreme low illumination	Δ (%)
RT-DETR-R18	RGB	74.5	68.2	44.9	39.73
YOLOv8n	RGB	84.1	68.5	64.3	23.54
PDE-RT-DETR	RGB-D	**97.9**	**94.2**	**82.1**	**16.14**

In contrast, the proposed PDE-RT-DETR maintains higher detection performance under all three illumination conditions. It achieves 97.9%, 94.2%, and 82.1% $\mathrm{mAP}_{0.50}$ under normal, low, and extreme low illumination, respectively. The corresponding degradation rate is only 16.14%, which is lower than those of RT-DETR-R18 and YOLOv8n. Notably, the $\mathrm{mAP}_{0.50}$ of PDE-RT-DETR under extreme low illumination remains higher than that of RT-DETR-R18 under normal illumination, indicating that the proposed RGB-D framework has stronger robustness against illumination degradation.

The improved robustness mainly benefits from the use of depth-guided geometric information. When RGB appearance features are severely degraded, the depth branch provides complementary geometric cues that are less affected by illumination changes. Furthermore, DEM enhances the reliability of depth-related representations, while CDPE introduces continuous depth-aware positional information into the Transformer feature space. These mechanisms help the model maintain more stable object localization and recognition under weak-texture and low-contrast conditions.

Overall, the robustness evaluation demonstrates that PDE-RT-DETR can effectively reduce performance degradation caused by illumination changes. This confirms the advantage of combining enhanced depth representation and depth-aware positional encoding for tiny pavement debris detection in complex warehouse inspection scenes.

5 Conclusion

This paper focuses on the problem of tiny pavement debris detection in aviation warehouse inspection scenes, where RGB-based perception is easily affected by low illumination, local shadows, weak texture, and low object-background contrast. To improve detection reliability under such visually degraded conditions, a depth-guided RGB-D detection framework, termed PDE-RT-DETR, is proposed based on RT-DETR-R18. The proposed framework introduces a depth branch to provide auxiliary geometric information, designs a Depth Enhancement Module (DEM) to improve the representation quality of depth-related features, and further incorporates Continuous Depth Position Encoding (CDPE) to embed continuous depth-aware positional information into the Transformer feature space. Through the combination of semantic information from RGB images and geometric cues from depth observations, the model enhances both tiny object recognition and spatial localization capability.

A self-built RGB-D pavement debris dataset containing 5,635 aligned image pairs was constructed under multiple illumination conditions. Experimental results demonstrate that PDE-RT-DETR achieves 86.4% precision, 87.6% recall, 91.4% $mAP_{0.50}$, and 66.8% $mAP_{0.50:0.95}$ on the self-built dataset. Compared with the RGB-D RT-DETR-R18 baseline, the proposed method improves $mAP_{0.50}$ and $mAP_{0.50:0.95}$ by 9.2 and 13.3 percentage points, respectively. Compared with the RGB-only RT-DETR-R18 baseline, the improvements reach 13.9 and 17.2 percentage points, respectively. These results indicate that the proposed depth-guided design effectively improves both overall detection accuracy and bounding-box localization quality.

The ablation experiments further verify the contribution of each component. Introducing the depth branch improves $mAP_{0.50}$ from 77.5% to 82.2%, showing that depth information provides useful complementary geometric cues. Incorporating DEM further increases $mAP_{0.50}$ to 87.4%, indicating that depth representation enhancement improves the usability of depth-related features. With CDPE, the complete PDE-RT-DETR achieves the best performance, especially improving $mAP_{0.50:0.95}$ to 66.8%, which suggests that continuous depth-aware positional encoding contributes to more accurate spatial localization.

Robustness experiments under different illumination conditions show that PDE-RT-DETR maintains greater stability than RGB-based comparison models. Under extreme low illumination, the proposed method still achieves 82.1% $mAP_{0.50}$, with a performance degradation rate of 16.14%, which is lower than those of RT-DETR-R18 and YOLOv8n. This demonstrates that depth-guided geometric information can effectively compensate for the degradation of RGB appearance features under weak-illumination and weak-texture conditions.

Overall, the proposed PDE-RT-DETR provides an effective RGB-D perception solution for tiny pavement debris detection in complex warehouse inspection scenes. It

shows practical potential for AGV-based inspection and safety monitoring in aviation warehouses. Future work will focus on further reducing model complexity, evaluating inference speed on edge devices, and extending the method to more complex real-world inspection scenarios with dynamic obstacles and larger-scale data.

Statements and Declarations

Funding. This project is supported by National Natural Science Foundation of China (52375557); Science and Technology Program of Tianjin(23JCYBJC00060; 24JCYBJC00160); Foundation of Tianjin educational committee (2023KJ222; 2025KJ162); The Basic Science-research Funds of National University (3122023PY06; 3122024PT07; 3122024052; 3122022PY15; 3122024044).

Competing interests. The authors declare that they have no competingfinancial interests or personal relationships that could have influencedthe work reported in this paper.

Consent for publication. All authors consent to the publication of thiswork.

Data availability. The datasets generated and analysed during the current study are available from the corresponding author on reasonable request.

Code availability. The custom code utilized in this study is not publicly available to protect the confidentiality of an ongoing project, but further information can be obtained from the corresponding author upon reasonable request.

Author contribution. Jun Wu: Conceptualization, Supervision, Funding acquisition, Writing - review & editing. Jiaxin Li: Methodology, Software, Validation, Writing - original draft. Runxia Guo: Formal analysis, Supervision, Funding acquisition. Jiusheng Chen: Investigation, Data curation, Funding acquisition. Guihang Liu: Data curation, Visualization, Validation.

References

[1] Zhu, M., Zhang, B., Zhou, C., Zou, H., Wang, X.: Target recognition of multi source machine vision pan tilt integrated inspection robot for power inspection. *IEEE ACCESS* **12**, 45693–45708 (2024) https://doi.org/10.1109/ACCESS.2024.3378580

[2] Dong, L., Chen, N., Liang, J., Li, T., Yan, Z., Zhang, B.: A review of indoor-orbital electrical inspection robots in substations. *Industrial Robot: the international journal of robotics research and application* **50**, 337–352 (2023) https://doi.org/10.1108/ir--06--2022--0162

[3] Fragapane, G., Koster, R., Sgarbossa, F., Strandhagen, J.O.: Planning and control of autonomous mobile robots for intralogistics: Literature review and research agenda. *EUROPEAN JOURNAL OF OPERATIONAL RESEARCH* **294**, 405–426 (2021) https://doi.org/10.1016/j.ejor.2021.01.019

[4] Jahan, F., Sun, W., Niyaz, Q., Alam, M.: Security modeling of autonomous systems: A survey. ACM COMPUTING SURVEYS **52**(5) (2019) https://doi.org/10.1145/3337791

[5] Wang, J., Ma, Y., Zhang, L., Gao, R.X., Wu, D.: Deep learning for smart manufacturing: Methods and applications. *JOURNAL OF MANUFACTURING SYSTEMS* **48**, 144–156 (2018) https://doi.org/10.1016/j.jmsy.2018.01.003

[6] Katona, K., Neamah, H.A., Korondi, P.: Obstacle avoidance and path planning methods for autonomous navigation of mobile robot. Sensors **24**(11) (2024) https://doi.org/10.3390/s24113573

[7] Hoang, D.-C., Tan, P.X., Nguyen, A.-N., Bui, S.-A., Duong, T.H.A., Huynh, T.-M., Nguyen, D.-M., Trinh, V.-A., Ha, Q.-H., Long, N.D.B., Tran, D.-T., Dinh, X.-T., Duong, V.-H., Trang, T.T.T.: Image-based anomaly detection in low-light industrial environments with feature enhancement. *Results in Engineering* **25**, 104309 (2025) https://doi.org/10.1016/j.rineng.2025.104309

[8] Tang, H., Zhu, H., Fei, L., Wang, T., Cao, Y., Xie, C.: Low-illumination image enhancement based on deep learning techniques: A brief review. PHOTONICS **10**(2) (2023) https://doi.org/10.3390/photonics10020198

[9] Rennie, C., Shome, R., Bekris, K.E., De Souza, A.F.: A dataset for improved rgbd-based object detection and pose estimation for warehouse pick-and-place. *IEEE Robotics and Automation Letters* **1**, 1179–1185 (2016) https://doi.org/10.1109/LRA.2016.2532924

[10] Xie, E., Wang, W., Wang, W., Ding, M., Shen, C., Luo, P.: Segmenting transparent objects in the wild. In: *Computer Vision – ECCV 2020*, pp. 696–711 (2020). https://doi.org/10.1007/978--3--030--58601--0_41

[11] Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **39**(6), 1137–1149 (2017) https://doi.org/10.1109/TPAMI.2016.2577031

[12] Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. ArXiv **abs/2010.04159** (2020)

[13] Zhao, Y., Lv, W., Xu, S., Wei, J., Wang, G., Dang, Q., Liu, Y., Chen, J.: Detrs beat yolos on real-time object detection. In: *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16965–16974 (2024). https://doi.org/10.1109/CVPR52733.2024.01605

[14] Huang, Y., Pan, K., Wei, K., Wu, R.: Swin detr: Pedestrian and vehicle detection under low light conditions based on swin detr. In: *Proceedings of the 2024 3rd International Conference on Artificial Intelligence and Intelligent Information*

Processing, pp. 292–297 (2024)

[15] Wu, H., Chen, S., Chen, G., Wang, W., Lei, B., Wen, Z.: Fat-net: Feature adaptive transformers for automated skin lesion segmentation. *Medical Image Analysis* **76**, 102327 (2022) https://doi.org/10.1016/j.media.2021.102327

[16] Bodla, N., Singh, B., Chellappa, R., Davis, L.S.: Soft-nms – improving object detection with one line of code. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (2017)

[17] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: *European Conference on Computer Vision*, pp. 213–229 (2020)

[18] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90

[19] Zhang, R., Qiu, H., Wang, T., Guo, Z., Cui, Z., Qiao, Y., Li, H., Gao, P.: Monodetr: Depth-guided transformer for monocular 3d object detection. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9155–9166 (2023)

[20] Prakash, A., Chitta, K., Geiger, A.: Multi-modal fusion transformer for end-to-end autonomous driving. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7077–7087 (2021)

[21] Qingyun, F., Dapeng, H., Zhaokui, W.: Cross-Modality Fusion Transformer for Multispectral Object Detection (2022). https://arxiv.org/abs/2111.00273

[22] Fu, H., Gong, M., Wang, C., Batmanghelich, K., Tao, D.: Deep ordinal regression network for monocular depth estimation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2002–2011 (2018)

[23] Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. Advances in neural information processing systems **27** (2014)

[24] Li, F., Zhang, H., Liu, S., Guo, J., Ni, L.M., Zhang, L.: Dn-detr: Accelerate detr training by introducing query denoising. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13619–13627 (2022)

[25] Wu, H., Chen, S., Chen, G., Wang, W., Lei, B., Wen, Z.: Attention-based depth distillation with 3d-aware positional encoding for monocular 3d object detection. *Proceedings of the AAAI Conference on Artificial Intelligence* **37**(3), 2892–2900 (2022) https://doi.org/10.1609/aaai.v37i3.25391

[26] Song, S., Lichtenberg, S.P., Xiao, J.: Sun rgb-d: A rgb-d scene understanding benchmark suite. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 567–576 (2015)

[27] Ignatov, D., Ignatov, A., Timofte, R.: Virtually Enriched NYU Depth V2 Dataset for Monocular Depth Estimation: Do We Need Artificial Augmentation? (2024). https://arxiv.org/abs/2404.09469

[28] Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: *European Conference on Computer Vision*, pp. 746–760 (2012)

[29] Li, C., Guo, J., Porikli, F., Pang, Y.: Lightennet: A convolutional neural network for weakly illuminated image enhancement. *Pattern recognition letters* **104**, 15–22 (2018)

[30] Qiqi, K., Heng, Z., Quanlin, W., Liangliang, C., Deqiang, C.: A review and prospect of image enhancement methods in low-luminance environments of mines. *COAL SCIENCE AND TECHNOLOGY* **53**(11), 82–100 (2025)

[31] Chen, Y., Wang, B., Guo, X., Zhu, W., He, J., Liu, X., Yuan, J.: Deyolo: Dual-feature-enhancement yolo for cross-modality object detection. In: *Pattern Recognition*, pp. 236–252 (2024). http://dx.doi.org/10.1007/978-3-031-78447-7_16

[32] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., Berg, A.C.: Ssd: Single shot multibox detector. In: *European Conference on Computer Vision*, pp. 21–37 (2016). Springer