

# DDStereo: Efficient Dual Decoder Transformers for Stereo 3D Road Anomaly Detection

Shiyi Mu<sup>1</sup>, Zichong Gu<sup>1</sup>, Zhiqi Ai<sup>1</sup>, Yilin Gao<sup>1</sup>, and Shugong Xu<sup>2</sup>

<sup>1</sup> Shanghai University, Shanghai, China

<sup>2</sup> Xi'an Jiaotong-Liverpool University, Suzhou, China  
shugong.xu@xjtlu.edu.cn

**Abstract.** Stereo-based 3D obstacle perception for autonomous driving is currently constrained by an imbalanced triplet: deployment cost, detection accuracy, and open-set adaptability. While existing methods struggle to balance these three competing objectives, there is an urgent demand for high-precision, real-time algorithms capable of detecting arbitrary obstacles in the wild. In this paper, we present DDStereo, a novel Dual-Decoder Stereo Transformer that achieves a synergistic integration of 3D object detection and Out-of-Distribution (OoD) road anomaly detection. Leveraging the geometric priors of stereo disparity, our approach effectively couples 3D attribute regression with open-set foreground detection within a streamlined dual-branch decoder architecture. Conventional methods rely on complex feature-level fusion; DDStereo maintains execution efficiency by employing a decoupled decoding strategy and shared object-level queries to ensure cross-modal target alignment. Extensive evaluations of public benchmarks demonstrate that DDStereo not only achieves state-of-the-art accuracy under open-set and closed-set protocols. Our method delivers real-time performance comparable to monocular 3D detection baselines, providing a cost-effective solution for the perception of obstacles of the normal and OoD category. Code and models are available at <https://github.com/shiyi-mu/DDStereo>.

**Keywords:** 3D Object Detection · Stereo Vision · Road Anomaly

## 1 Introduction

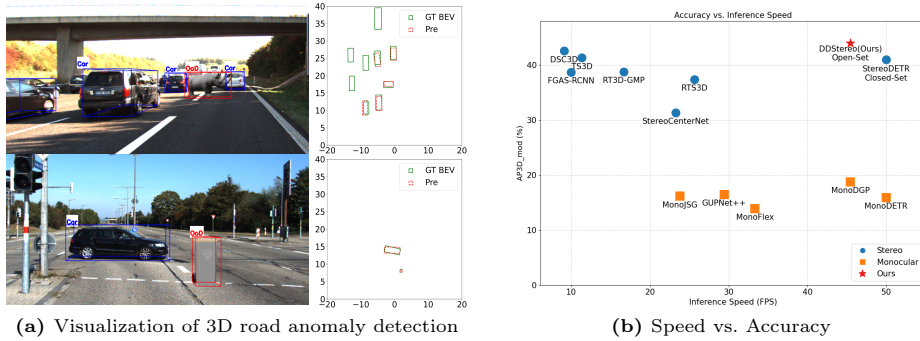
*“To sense the unseen is the threshold of safety.”*

In the context of intelligent driving, the “unseen” refers to the vast array of unforeseen, Out-of-Distribution (OoD) obstacles that lie beyond fixed training labels. Safely navigating these anomalies has long been the primary bottleneck for stereo-based systems, which must reconcile high-fidelity geometric reasoning with the low-latency demands of real-time deployment.

3D object detection has become a cornerstone technology in autonomous driving and robotic perception. While LiDAR-based methods dominate in accuracy, stereo vision offers a cost-effective alternative by providing dense depth

---

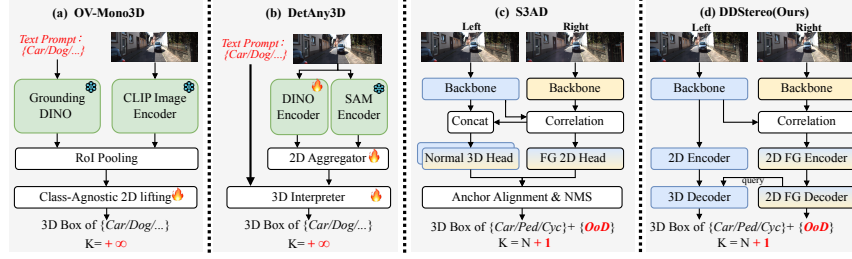
 Shugong Xu is the corresponding author.



**Fig. 1: Qualitative and quantitative results.** (a) Our model demonstrates robust 3D road anomaly detection capabilities on AR-OoD [33], providing precise 3D bounding boxes and BEV representations. (b) Performance comparison with closed-set monocular and stereo 3D detection methods on KITTI [11]. Our approach achieves a superior trade-off between inference speed, accuracy and open-set.

without expensive sensors. However, a significant gap remains: most stereo 3D detectors have reduced latency to around 40 ms, yet they remain nearly twice as slow as their monocular counterparts. StereoDETR [32] achieves a speed comparable to monocular detection, but it remains limited to closed-set detection. Simultaneously, while monocular detection has advanced toward open-vocabulary recognition [46, 50], most stereo approaches still operate under a **closed-set assumption**, rendering them blind to unseen categories such as road debris or novel construction equipment. In real-world environments, this assumption is frequently violated. Autonomous agents must not only localize known objects but also respond to OoD instances. Recent efforts have introduced stereo-based road anomaly detection; however, their inference speed exceeds 80 ms [33], far from the requirements of real-time deployment. Furthermore, unlike existing open-vocabulary algorithms that require heavy text prompts or manual cues, the immediate demand in driving scenarios is prompt-free, open-ended detection. Rather than predicting fine-grained category names, the priority is the rapid binary classification of an obstacle’s presence to ensure immediate safety. To address these challenges, we redefine stereo-based 3D perception as a unified optimization of a fundamental triplet: accurate multi-class detection, robust binary OoD discovery, and ultra-fast real-time inference. Fig. 1 illustrates the detection results and the balance between speed and accuracy.

The risk of missed detections is significantly exacerbated for unknown categories. While a small fraction of novel objects may be misclassified into known classes, the vast majority are typically overlooked due to the absence of category-specific priors and supervision. Existing open-vocabulary detectors attempt to mitigate this by requiring external guidance, such as textual prompts [46] or interactive cues (e.g., points or boxes [50]). However, in high-stakes autonomous driving, a truly robust paradigm must rely solely on visual input to perceive all potential obstacles, regardless of their semantic labels. As shown in Fig. 2,



**Fig. 2:** Compare with other methods. OV-Mono3D [46] and DetAny3D [50] rely on text prompt matching, whereas S3AD [33] uses anchor-level fusion to detect unknown foregrounds; neither approach is real-time. Our DDStereo decouples texture and disparity features to independently predict 3D attributes and foreground objectness, maintaining high computational efficiency.

while methods like OV-Mono3D [46] and DetAny3D [50] enable unconstrained recognition via language-vision alignment, they introduce significant computational overhead. Conversely, S3AD [33] extends  $N$ -class detection by treating OoD objects as an auxiliary  $(N+1)$ -th class. Despite its effectiveness, S3AD suffers from prohibitive latency ( $>80$  ms) due to dense anchor-based Non-Maximum Suppression (NMS) and high-dimensional feature concatenation.

To address these challenges, we propose **DDStereo**, a Dual-Decoder Stereo Transformer designed for end-to-end open-set 3D detection. Our core philosophy is the *decoupling of category-agnostic localization from semantic classification*. Specifically, a standard decoder regresses 3D attributes for In-Distribution (ID) objects, while a specialized foreground decoder identifies potential instances leveraging disparity-aware features. Since disparity is fundamentally driven by geometric consistency rather than texture patterns, this branch enables robust localization of novel objects based on their physical presence alone.

To ensure spatial alignment between these decoupled branches, we introduce a **shared object query** mechanism. This design maintains strict consistency between foreground boxes and class-specific predictions, allowing anomaly scores to be derived from the discrepancy between foreground confidence and semantic classification probability. Furthermore, to avoid memory-intensive feature concatenation, we decouple depth estimation from attribute regression. A lightweight convolutional head predicts object-wise depth maps, which are efficiently fused via center-based sampling. This architecture not only eliminates the need for NMS but also slashes inference latency to 24 ms.

The contributions of this work are summarized as follows:

- **Unified Real-Time Open-Set Framework:** We present DDStereo, the first 3D detector to integrate high-precision, real-time inference and open-set adaptability into a single stereo pipeline.
- **Efficient Query-Base Alignment:** We introduce a shared object query mechanism that elegantly bridges 3D regression and anomaly detection. This

simple yet efficient design ensures cross-task consistency while eliminating the need for complex feature-level fusion.

- **Generalization to the Wild:** We validate the generalization potential of DDSTereo through extensive evaluations on out-of-distribution categories and diverse real-world visualizations, demonstrating robust obstacle perception in unconstrained environments.

## 2 Related Work

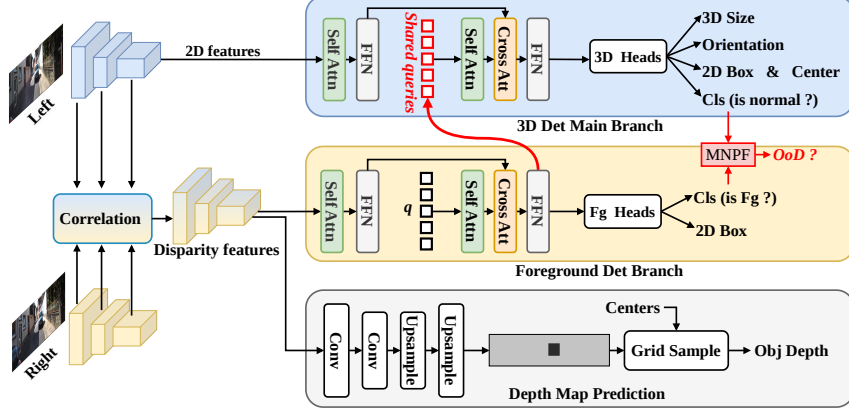
### 2.1 Stereo 3D Object Detection

Stereo-based 3D object detection methods can be broadly categorized by how they compute disparity or depth cues, typically falling into three paradigms: image-level disparity, region-level disparity, and feature-level disparity. Early approaches rely on image-level disparity estimation, such as RT3DStereo [20] and RT3D-GMP [21]. However, such pipelines often suffer from depth noise and are computationally intensive. Region-level disparity methods extract object-aligned regions using 2D detectors and compute disparity only within those areas as TLNet [38], Stereo R-CNN [23], and SIDE [34]. Recent works favor feature-level disparity learning, which leverages stereo feature matching to directly regress 3D attributes from cost volumes or concatenated features [6, 28, 40]. Although these methods have shown strong accuracy. They typically rely on heavy anchor based heads or 3D cost volumes, leading to limited runtime performance. With an extremely lightweight design, StereoDETR [32] breaks the speed bottleneck of monocular detection for the first time, yet it is restricted to closed-set scenarios.

### 2.2 Open-Set and Unknown 3D Object Detection

Open-set object detection (OSOD) and open-vocabulary object detection (OVOD) have been explored as two complementary paradigms to detect unknown objects. In 2D field, OVOD methods learn visual text alignments to localize arbitrary categories specified by text queries [8, 12, 26, 27, 44, 47, 49], while OSOD approaches aim to discover any objects absent from the training set, typically labeling them as “unknown” [9, 14, 15, 19, 48]. In 3D field, existing work has focused almost exclusively on OVOD, leveraging 2D open-vocabulary detectors to lift category-agnostic proposals into class-agnostic 3D detectors [3, 4, 31, 43, 45, 51]. Open-set 3D detection, however, has received limited attention. Cen et al. [4] first formalize the task and propose a two-stage pipeline that scores point-wise entropy to reject unseen classes. Alliegro et al. [1] introduce 3DOS, a benchmark and baseline that exposes the sensitivity of point-cloud detectors to semantic novelty. Lu et al. [31] adopt 2D open-vocabulary knowledge to generate pseudo-labels for unknown objects yet stop short of explicit open-set modeling. S3AD [33] is the first method to introduce open-set 3D object detection based on stereo vision, along with the release of the AR-OoD benchmark, aiming to detect arbitrary potential obstacles and enhance autonomous driving safety. However, as S3AD is built upon the YOLOStereo3D [28] framework, its inference time exceeds 80 ms, making it difficult to meet real-time requirements.





**Fig. 3:** Overview of the proposed DDStereo architecture. (1) The main branch predicts 3D attributes via monocular features; (2) the foreground branch outputs texture-agnostic binary foregrounds; and (3) the depth map branch estimates object-level center depth. All three branches are coupled using shared queries and Grid Sample.

### 3 Method

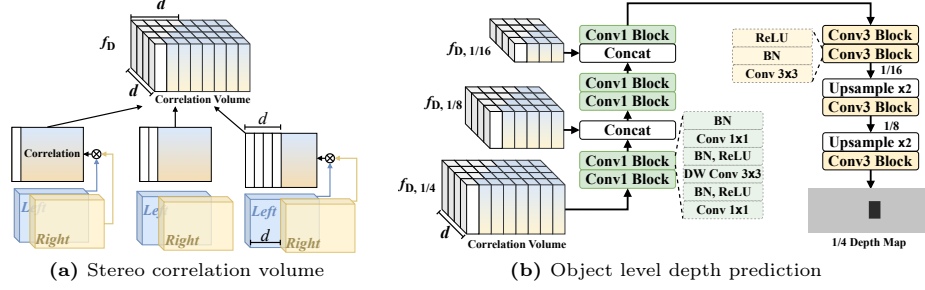
#### 3.1 Overview

We propose DDStereo, a Dual-Decoder stereo 3D detection framework tailored for open-world scenarios. As shown in Fig. 3, our model consists of three collaborative branches: a main 3D detection branch, a foreground detection branch, and an object-level depth map prediction branch. These branches are tightly integrated through query-level and point-level alignment strategies.

Let  $\mathcal{I}_L$  and  $\mathcal{I}_R$  denote the left and right stereo images, respectively. We extract 2D visual features from the left image only using a Resnet backbone, denoted as  $f_V$ . Meanwhile, to encode stereo feature, we compute a correlation volume from the features of  $\mathcal{I}_L$  and  $\mathcal{I}_R$  via cross-view cost aggregation. The resulting disparity-based feature map, denoted as  $f_D$ , encodes stereo-aware depth information.

The 3D detection main branch operates on  $f_V$  and utilizes a transformer-based encoder-decoder architecture to predict 3D object properties, including size, orientation, class scores (for known categories), and 2D bounding boxes. Concurrently, the foreground detection branch utilizes  $f_D$  to predict coarse 2D boxes and foreground probabilities via binary classification. To align the 2D texture representations and stereo-derived depth cues, both detection branches are conditioned on a shared object queries, facilitating joint reasoning across appearance and depth.

To enhance depth map estimation, we upsample the stereo disparity feature  $f_D$  to produce an object-level depth map, denoted as  $\mathbf{M}_{\text{obj}}$ . Given the centers of predicted boxes from the main 3D detection branch, we use a grid sampling operation to extract object-centric depth values, which are further used to comple-



**Fig. 4: Overview of the stereo matching and depth estimation modules.** (a) The computation process of the stereo correlation volume, constructed from concatenated left-right feature maps. (b) Our multi-scale disparity fusion strategy for high-precision depth map prediction, providing the geometric foundation for subsequent 3D detection.

ment 3D box localization. Finally, we introduce a MNPF module, where MNPF refers to the comparison between the maximum normal probability and the foreground confidence, to estimate the open-set confidence of each detection. This enables our model to identify OoD objects beyond the training categories.

### 3.2 Visual feature and Disparity feature

The left and right stereo images are first processed through a shared backbone to extract multi-scale features at  $\frac{1}{4}$ ,  $\frac{1}{8}$ ,  $\frac{1}{16}$ , and  $\frac{1}{32}$  resolutions. The multi-scale features from the left view are used as inputs to the visual decoding branch.

To construct disparity-aware representations, we compute simple correlation volumes between the left and right features at each scale. As illustrated in Fig. 4a, we generate these disparity features by shifting the right-view features along the horizontal axis up to a maximum displacement of  $d$  pixels and calculating per-pixel correlations with the left-view features. The maximum disparity steps for the four scales are set to 24, 24, 16, and 16.

### 3.3 Depth map decoder and Sampling

In MonoDETR [52] and MonoDGP [37], an explicit and coarse-grained  $\frac{1}{16}$  depth map is predicted, where the ground-truth is constructed from the 3D center distance and the corresponding 2D bounding box of each object. The predicted depth is further encoded and fused via a depth encoder and cross-attention to guide the 3D decoding process. Similar to the architecture of StereoDETR [32], we adopt a more efficient strategy that predicts a higher-resolution depth map and directly samples depth values at the predicted object centers from the detection branch. As shown in Fig. 4b, the depth prediction module adopts a simple U-shaped architecture. Multi-scale correlation volumes are concatenated after down-sampling and then upsampled through convolutional layers to reconstruct

a depth map at  $\frac{1}{4}$  resolution. The projected 2D center of each 3D object is used to sample the predicted depth as the object’s depth estimate. The ground-truth depth map and supervision loss follow the same setting as in MonoDETR [52] and MonoDGP [37].

### 3.4 Binary Foreground Detection Branch

It is difficult to determine the specific category of an object from a depth or disparity map. However, it is often sufficient to identify whether an obstacle is present. To this end, we design a foreground detection branch as a binary DETR-style decoder based on disparity features. Specifically, the multi-scale disparity features  $f_D$  are first flattened and fed into an encoder, which consists of a single-layer multi-scale deformable attention module followed by a feed-forward network (FFN). The output embeddings are denoted as  $f_D^e$ .

In the decoder, we define a learnable object query  $q \in \mathbf{R}^{N \times C}$ , where  $N$  is the maximum number of objects. This same query is also shared with the 3D detection decoder. The decoder is composed of an inter-query self-attention layer, a visual cross-attention layer, and a feed-forward network. Finally, two linear layers are used to predict the foreground classification and the 2D bounding boxes.

$$\text{SelfAttn}(q) = \text{softmax}\left(\frac{qq^\top}{\sqrt{C}}\right)q, \quad (1)$$

$$\text{CrossAttn}(q, f_D^e) = \text{softmax}\left(\frac{q(f_D^e)^\top}{\sqrt{C}}\right)f_D^e. \quad (2)$$

### 3.5 3D Detection Branch

Unlike MonoDETR [52] and MonoDGP [37], we do not adopt the depth-guided cross-attention mechanism in the 3D detection branch. Instead, we follow a structure similar to the original 2D detector to predict all attributes except depth, as the dual-decoder design already encodes geometry via shared queries, making explicit depth attention redundant and preserving computational efficiency.

Specifically, the left-view visual features  $f_V$  are fed into an encoder composed of a single-layer self-attention module followed by a FFN, producing embeddings denoted as  $f_V^e$ . In the decoder, we reuse the object query  $q$  from the foreground decoder. The decoder consists of an inter-query self-attention layer, a visual cross-attention layer, and a FFN, followed by several linear layers that respectively predict the 3D object size, orientation, normal category (with  $K$  classes), and the 2D bounding box. Notably, the 2D bounding box is parameterized using the projected 3D object center on the image plane as the reference point. The box is represented by offsets from the center to the top, bottom, left, and right boundaries, resulting in six parameters in total. This projected center will be used in the subsequent object depth sampling process.

$$\text{CrossAttn}(q, f_V^e) = \text{softmax}\left(\frac{q(f_V^e)^\top}{\sqrt{C}}\right)f_V^e. \quad (3)$$

### 3.6 Anomaly Confidence Estimation

While previous road anomaly segmentation methods primarily focus on global pixel-level OoD scoring, we propose an object-level anomaly score estimation. Specifically, the anomaly score MNPF is computed as the distance between the **M**aximum **N**ormal **P**robability and the **F**oreground probability  $S_f$ :

$$S_{\text{OoD}} = S_f - \max_{i \in K}(\sigma(S_i)), \quad (4)$$

where  $K$  is the set of normal known classes. We set  $\sigma$  as the sigmoid function. If replaced with a softmax operation, the latter term becomes equivalent to MSP [17]. If  $\sigma$  is removed (no normalization is applied), it is equivalent to MaxLogit [16].

### 3.7 Loss Function

The training loss consists of three main components: the 2D foreground loss, the 3D detection loss for in-distribution categories, and the disparity supervision loss. Both the 2D and 3D branches employ separate Hungarian matching processes for target assignment. The 2D loss includes a binary classification loss and a bounding box regression loss. The 3D prediction head regresses multiple attributes, including object category, orientation, 3D size, 3D center, and 2D bounding box.

$$\mathcal{L}_{\text{total}} = \lambda_{2D} \cdot \mathcal{L}_{2D} + \lambda_{3D} \cdot \mathcal{L}_{3D} + \lambda_D \cdot \mathcal{L}_{\text{map}} \quad (5)$$

Hungarian matching is independently applied for both  $\mathcal{L}_{2D}$  and  $\mathcal{L}_{3D}$  to associate predicted queries with ground-truth objects.

## 4 Experiments

### 4.1 Experimental Setup

**Anomaly Stereo 3D Dataset.** AR-OoD [33] is the most category-rich dataset for 3D road anomaly detection. We follow the exact same data split as S3AD [33]. The dataset contains 4,038 training and 2,347 testing samples, encompassing 105 categories in total. The training set includes 47 categories to enhance the model’s generalization ability, while the remaining 58 novel categories are used to evaluate road anomaly detection. During the testing phase, these 58 novel anomaly categories are mapped into a single "OoD" category.

**Close-set Stereo 3D Dataset.** Unlike Waymo [41] and nuScenes [2], KITTI [11] and Argoverse [5] provide paired stereo images. On KITTI [11], following the split in YOLOStereo3D [28], we use 3,712 images for training and 3,769 for validation, submitting our unlabeled test set (7,518 samples) predictions to the official server. Most baseline methods only report Car metrics, we evaluate on the Car, Pedestrian, and Cyclist categories. For the more challenging Argoverse

[5] dataset, we adopt the format and split from StereoDistill [29], using 13,122 training and 5,015 validation samples.

**Implementation Details.** We train our DDStereo using four NVIDIA RTX A6000 GPUs with 48,GB of memory. For a fair evaluation of inference efficiency, all testing is performed on a single RTX 4090D GPU. We use the Adam optimizer with an initial learning rate of 0.0002, and apply a learning rate decay factor of 0.1 at the 125 and 165 epochs. The batch size is set to 16.

## 4.2 Open-Set Evaluation on AR-OoD

**Table 1:** Anomaly evaluation on AR-OoD [33]. OoD refers to 58 novel categories mapped to a single class, while Pedestrian Car and Cyclist are known categories. Results are reported for the moderate difficulty level at an IoU threshold of 0.25. R11 and R40 denote the evaluation metric of 11 and 40 recall positions.

| Method         | Recall | OoD(Mod)     |              |              | Pedestrian(Mod) |              |              | Car(Mod)     |              |              | Cyclist(Mod) |              |              |
|----------------|--------|--------------|--------------|--------------|-----------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
|                |        | $AP_{3D}$    | $AP_{BEV}$   | $AP_{2D}$    | $AP_{3D}$       | $AP_{BEV}$   | $AP_{2D}$    | $AP_{3D}$    | $AP_{BEV}$   | $AP_{2D}$    | $AP_{3D}$    | $AP_{BEV}$   | $AP_{2D}$    |
| OV-mono3D [46] | R11    | 0.64         | 1.85         | 90.91        | 23.75           | 24.14        | 100.0        | 55.38        | 56.12        | 100.0        | 20.03        | 21.16        | 100.0        |
| S3AD [33]      | R11    | <b>74.35</b> | 76.39        | 90.06        | 48.26           | 49.07        | 64.80        | 80.20        | 80.41        | 88.54        | 25.26        | 25.25        | <b>37.12</b> |
| DDStereo(Ours) | R11    | 74.23        | <b>83.01</b> | <b>98.36</b> | <b>51.86</b>    | <b>52.16</b> | 60.33        | <b>87.23</b> | <b>87.31</b> | <b>89.85</b> | <b>32.82</b> | <b>32.82</b> | 33.39        |
| Improvement    | R11    | -0.12        | +6.62        | +8.30        | +3.60           | +3.09        | -4.47        | +7.03        | +6.90        | +1.31        | +7.56        | +7.57        | -3.73        |
| OV-mono3D [46] | R40    | 0.53         | 1.02         | 97.50        | 18.84           | 20.01        | <b>100.0</b> | 56.54        | 57.50        | <b>100.0</b> | 17.95        | 19.60        | <b>100.0</b> |
| S3AD [33]      | R40    | 73.69        | 76.76        | 95.53        | 45.74           | 46.39        | 63.96        | 80.25        | 82.23        | 89.01        | 20.82        | 22.02        | <b>34.40</b> |
| DDStereo(Ours) | R40    | <b>78.09</b> | <b>82.44</b> | <b>99.14</b> | <b>52.41</b>    | <b>52.89</b> | 62.39        | <b>87.47</b> | <b>87.68</b> | 93.40        | <b>31.68</b> | <b>31.69</b> | 32.75        |
| Improvement    | R40    | +4.40        | +5.68        | +3.61        | +6.67           | +6.50        | -1.57        | +7.22        | +5.45        | +4.39        | +10.86       | +9.67        | -1.65        |

We evaluate our method under the open-set setting on the AR-OoD [33], and report the results in Table 1. Following the S3AD [33], we focus on the moderate difficulty level and set the confidence threshold to 0.25. We report results under both R11 and R40 settings. DDStereo achieves consistently higher accuracy for both known and unknown object categories. On the unknown OoD class, DDStereo attains 74.23%  $AP_{3D}$ , 83.01%  $AP_{BEV}$ , 98.36%  $AP_{2D}$  under R11 and 78.09%, 82.44%, 99.14% under R40, surpassing S3AD [33] especially in BEV localization. Moreover, DDStereo also achieves stronger performance on the known classes Pedestrian, Car, and Cyclist, demonstrating its balanced capability in both open-set and close-set 3D detection. Fig. 5 presents selected qualitative results. The rhino and the bench shown in the figure belong to categories not present in the training set, but DDStereo successfully estimates their 3D positions and dimensions. For OV-Mono3D [46], we provided ground truth text prompts during testing like "Dog". Benefiting from the broad 2D open-vocabulary coverage, it achieved 100% 2D detection accuracy (IoU > 0.25) on cars, pedestrians, and cyclists. However, its 3D detection performance is inferior to our method.

**Table 2:** 3D Object Detection of Car Category on KITTI. (IoU>0.7)

| Method                          | Input     | OpenSet | $AP_{3D}(Test)$ |              |              | $AP_{BEV}(Test)$ |              |              | Time (ms) |
|---------------------------------|-----------|---------|-----------------|--------------|--------------|------------------|--------------|--------------|-----------|
|                                 |           |         | Easy            | Mod          | Hard         | Easy             | Mod          | Hard         |           |
| MonoFlex [53]                   | Monocular | ×       | 19.94           | 13.89        | 12.07        | 28.23            | 19.75        | 16.89        | 30        |
| MonoJSG [25]                    | Monocular | ×       | 24.69           | 16.14        | 13.64        | 32.59            | 21.26        | 18.18        | 42        |
| GUPNet++ [30]                   | Monocular | ×       | 24.99           | 16.48        | 14.58        | -                | -            | -            | 34        |
| MonoDGP [37]                    | Monocular | ×       | 26.35           | 18.72        | 15.97        | 35.24            | 25.23        | 22.02        | 22        |
| MonoDETR [52]                   | Monocular | ×       | 25.00           | 16.47        | 13.58        | 33.60            | 22.11        | 18.60        | 20        |
| <hr/>                           |           |         |                 |              |              |                  |              |              |           |
| OC-Stereo [36]                  | Stereo    | ×       | 55.15           | 37.60        | 30.25        | 68.89            | 51.47        | 42.97        | 350       |
| SIDE [34]                       | Stereo    | ×       | 47.69           | 30.82        | 25.68        | -                | -            | -            | 260       |
| StereoRCNN [23]                 | Stereo    | ×       | 47.58           | 30.23        | 23.72        | 61.92            | 41.31        | 33.42        | 200       |
| DSC3D [6]                       | Stereo    | ×       | <b>66.46</b>    | 42.54        | 34.04        | 74.56            | 51.21        | 42.07        | 110       |
| TLNet [38]                      | Stereo    | ×       | 7.46            | 4.37         | 3.74         | 13.71            | 7.69         | 6.73         | 100       |
| FGAS RCNN [42]                  | Stereo    | ×       | 58.02           | 38.68        | 32.53        | 72.56            | 58.31        | 46.24        | 100       |
| TS3D [40]                       | Stereo    | ×       | 64.61           | 41.29        | 30.68        | 73.34            | 48.59        | 36.98        | 88        |
| YoloStereo3D [28]               | Stereo    | ×       | 65.68           | 41.25        | 30.42        | 76.10            | 50.28        | 36.86        | 80        |
| RT3DStereo [20]                 | Stereo    | ×       | 29.90           | 23.28        | 18.96        | 58.81            | 46.82        | 38.38        | 79        |
| ESGN [10]                       | Stereo    | ×       | 65.80           | <b>46.39</b> | <b>38.42</b> | <b>78.10</b>     | 58.12        | <b>49.28</b> | 62        |
| RT3D-GMP [21]                   | Stereo    | ×       | 45.79           | 38.76        | 30.00        | 69.14            | <b>59.00</b> | 45.49        | <b>60</b> |
| <hr/>                           |           |         |                 |              |              |                  |              |              |           |
| <i>Real-time Stereo Methods</i> |           |         |                 |              |              |                  |              |              |           |
| StereoCenterNet [39]            | Stereo    | ×       | 49.94           | 31.30        | 25.62        | 62.97            | 42.12        | 35.37        | 43        |
| RTS3D [24]                      | Stereo    | ×       | 58.51           | 37.38        | 31.12        | 72.17            | 51.79        | 43.19        | 39        |
| StereoDETR [32]                 | Stereo    | ×       | 59.45           | 41.17        | 35.13        | 72.77            | <b>54.53</b> | <b>46.41</b> | <b>18</b> |
| DDStereo(Ours)                  | Stereo    | ✓       | <b>62.44</b>    | <b>43.97</b> | <b>36.16</b> | <b>73.63</b>     | 53.60        | 45.02        | 24        |
| <hr/>                           |           |         |                 |              |              |                  |              |              |           |
| <i>Improvement</i>              |           |         | +2.99           | +2.80        | +1.03        | +0.86            | -0.93        | -1.39        | +6        |

### 4.3 Close-Set Results on KITTI Benchmark

We report the 3D object detection performance on the KITTI benchmark in Table 2, including both the official *test server* results (for Car) and offline *validation set* evaluations. Our DDStereo achieves state-of-the-art results on the *Moderate* and *Hard* difficulty levels of  $AP_{3D}$  with 43.97% and 36.16% respectively, outperforming all existing stereo-based approaches, including DSC3D [6] and YOLOStereo3D [28]. Since the training set contains only 3K stereo pairs, it is insufficient for effectively training a Transformer-based architecture. Consequently, the proposed method does not surpass the YOLOStereo3D-based approaches [6, 28, 40] on the Easy subset. However, it still maintains an accuracy advantage over existing real-time methods [24, 39]. Moreover, the Easy, Mod, and Hard splits are defined by difficulty thresholds, where the Hard set includes the Easy and Mod, thus better reflecting the overall average accuracy. In terms of efficiency, DDStereo runs at 24ms per frame significantly outpacing TS3D (88ms) [40], RTS3D (39ms) [24], and YOLOStereo3D (80ms) [28], while maintaining superior accuracy. Relative to the closed-set baseline StereoDETR [32], DDStereo extends open-set detection capability with only 6ms extra latency, while simultaneously achieving better closed-set performance.

As shown in Table 3, our method DDStereo achieves the best performance across all difficulty levels of the Pedestrian class, surpassing all previous stereo

**Table 3:** 3D Detection of Pedestrian and Cyclist on The KITTI Test Server.

|                   | Pedestrian   |              |              |              |              |              | Cyclist      |              |              |              |              |              |
|-------------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
|                   | Easy         |              | Mod          |              | Hard         |              | Easy         |              | Mod          |              | Hard         |              |
| Method            | $AP_{3D}$    | $AP_{BEV}$   | $AP_{3D}$    | $AP_{BEV}$   | $AP_{3D}$    | $AP_{BEV}$   | $AP_{3D}$    | $AP_{BEV}$   | $AP_{3D}$    | $AP_{BEV}$   | $AP_{3D}$    | $AP_{BEV}$   |
| MonoFlex [53]     | 9.43         | 10.36        | 6.31         | 7.36         | 5.26         | 6.29         | 4.17         | 4.41         | 2.35         | 2.35         | 2.04         | 2.50         |
| DEVIANT [22]      | 13.43        | 14.49        | 8.65         | 9.77         | 7.69         | 8.28         | 5.05         | 6.42         | 3.13         | 3.97         | 2.59         | 3.51         |
| MonoDTR [18]      | 15.33        | 16.66        | 10.18        | 10.59        | 8.61         | 9.00         | 5.05         | 5.84         | 3.27         | 4.11         | 3.19         | 3.48         |
| GUPNet++ [30]     | 12.45        | —            | 8.13         | —            | 6.91         | —            | 6.71         | —            | 3.91         | —            | 3.80         | —            |
| RT3DStereo [20]   | 3.28         | 4.72         | 2.45         | 3.65         | 2.35         | 3.00         | —            | —            | —            | —            | —            | —            |
| ESGN [10]         | 14.05        | 17.94        | 10.27        | 13.03        | 9.02         | 11.54        | 13.84        | 15.78        | 7.69         | 9.02         | 6.75         | 7.96         |
| RT3D-GMP [21]     | 16.23        | 19.92        | 11.41        | 14.22        | 10.12        | 12.83        | 18.31        | 20.59        | 12.99        | 13.92        | 10.63        | 12.74        |
| OC-Stereo [36]    | 24.48        | 29.79        | 17.58        | 20.80        | 15.60        | 18.62        | 29.40        | 32.47        | 16.63        | 19.23        | 14.72        | 17.11        |
| YoloStereo3D [28] | 28.49        | 31.01        | 19.75        | 20.76        | 16.48        | 18.41        | —            | —            | —            | —            | —            | —            |
| DSC3D [6]         | 29.60        | 32.25        | 20.43        | 22.57        | 17.92        | 19.01        | —            | —            | —            | —            | —            | —            |
| DDStereo(Ours)    | <b>32.14</b> | <b>35.80</b> | <b>21.92</b> | <b>24.72</b> | <b>19.43</b> | <b>22.06</b> | <b>34.59</b> | <b>37.92</b> | <b>21.80</b> | <b>24.36</b> | <b>18.59</b> | <b>20.87</b> |

methods. DDStereo achieves consistent improvements over the previous best-performing stereo method DSC3D [6] across all difficulty levels. For the Cyclist category, which poses greater challenges, only a few existing works have reported results on the KITTI test server. Among the published stereo-based methods, OC-Stereo [36] achieves the strongest 3D detection performance. Compared to it, our DDStereo achieves consistent improvements across all difficulty levels. Other methods listed in Table 2 did not report results or release code for the Pedestrian and Cyclist classes, and thus are not included in the comparison.

On the more challenging Argoverse [5] dataset, we follow the same evaluation metrics as StereoDistill [29] and present BEV and 3D detection results for the moderate Car and Pedestrian categories at 3D IoU thresholds of 0.5 and 0.25. As shown in Table 4, our method demonstrates a clear accuracy advantage over the pure stereo approach YOLOStereo3D [28], although it remains inferior to LiDAR-distillation-based methods.

**Table 4:** Close-set performance comparison on Argoverse [5].

| Method             | Lidar Distill | $AP_{3D}^{Car}$ | $AP_{BEV}^{Car}$ | $AP_{3D}^{Ped}$ | $AP_{BEV}^{Ped}$ |
|--------------------|---------------|-----------------|------------------|-----------------|------------------|
| DSGN [7]           | ✓             | 33.68           | 42.76            | 8.58            | 8.95             |
| LIGA [13]          | ✓             | 34.37           | 45.47            | 10.76           | 11.01            |
| StereoDistill [29] | ✓             | 37.55           | 46.99            | 13.70           | 14.04            |
| YOLOStereo3D [28]  | ×             | 16.65           | 23.87            | 3.65            | 3.94             |
| DDStereo (Ours)    | ×             | <b>25.74</b>    | <b>30.87</b>     | <b>8.52</b>     | <b>8.78</b>      |
| <i>Improvement</i> | ×             | <b>+9.09</b>    | <b>+7.00</b>     | <b>+4.87</b>    | <b>+4.84</b>     |

#### 4.4 Ablation Study

**Shared Queries.** The dual-branch architecture separately predicts binary foreground and multi-class 3D attributes. In the CNN-based framework, S3AD [33] aligns anchors in the decoding stage to compute anomaly scores. In this work, a Transformer-based end-to-end framework is adopted, where shared queries are used for query-level alignment. As shown in Tables 5 and 6, shared queries consistently boost both closed-set and open-set performance. Especially for open-set detection, employing an independent query design leads to a decrease of approximately 27% in  $AP_{3D}^{OoD}$ , primarily because decoding order misalignment interferes with confidence estimation.

**Table 5:** Queries Ablation close-set KITTI [11] (IoU > 0.7/0.5).

|         | $AP_{3D}^{Car}$ (0.7) |             |             | $AP_{3D}^{Cyc}$ (0.5) |             |             |
|---------|-----------------------|-------------|-------------|-----------------------|-------------|-------------|
| Queries | Easy                  | Mod         | Hard        | Easy                  | Mod         | Hard        |
| Indep.  | 61.9                  | 44.3        | 37.6        | 34.9                  | 19.0        | 17.2        |
| Shared  | <b>66.9</b>           | <b>48.2</b> | <b>41.1</b> | <b>37.8</b>           | <b>20.4</b> | <b>18.9</b> |

**Table 6:** Queries Ablation on open-set AR-OoD [33] (MoD, IoU > 0.25).

| Queries | $AP_{3D}^{OoD}$ | $AP_{3D}^{Car}$ | $AP_{3D}^{Ped}$ | $AP_{3D}^{Cyc}$ |
|---------|-----------------|-----------------|-----------------|-----------------|
| Indep.  | 51.19           | 72.83           | 41.20           | 26.01           |
| Shared  | <b>78.09</b>    | <b>87.43</b>    | <b>52.51</b>    | <b>31.69</b>    |

**Decoder Design and Depth Prediction Strategy.** We conduct an ablation study to investigate the effectiveness of the dual-decoder structure and the depth prediction strategy, as shown in Table 7. The table also reports the performance on OoD and normal categories at a threshold of 0.25. Replacing the dual-decoder decoder with a single design brings a significant drop of 11.72%  $AP_{3D}^{OoD}$ , demonstrating the benefit of decoupling foreground detection and 3D attribute estimation. For depth prediction, we adopt the average strategy from MonoDETR [52], which computes the mean over multiple depth predictions. When replacing the Grid Sample operation with the Average strategy in MonoDETR [52], the  $AP_{3D}^{OoD}$  decreases by 8.70%.

**Table 7:** Ablation of decoder and depth strategy (Mod, IoU>0.25).

| Setting          | $AP_{3D}^{OoD}$ | $AP_{3D}^{Car}$ | $AP_{3D}^{Ped}$ | $AP_{3D}^{Cyc}$ |
|------------------|-----------------|-----------------|-----------------|-----------------|
| DDStereo (Ours)  | <b>78.09</b>    | <b>87.43</b>    | <b>52.51</b>    | <b>31.69</b>    |
| w/o Dual Decoder | 66.37           | 85.21           | 49.70           | 27.10           |
| w/o Grid Sample  | 69.39           | 86.31           | 50.66           | 30.34           |

**Table 8:** OoD scoring strategy comparison (Mod, IoU>0.25).

| Method        | $AP_{3D}^{OoD}$ | $AP_{BEV}^{OoD}$ | $AP_{2D}^{OoD}$ |
|---------------|-----------------|------------------|-----------------|
| MaxLogit [16] | 67.71           | 71.52            | 84.51           |
| MSP [17]      | 75.16           | 79.28            | 92.22           |
| MNPF (Ours)   | <b>78.09</b>    | <b>82.44</b>     | <b>99.14</b>    |

**OoD Score.** As shown in Table 8, we compare different scoring strategies for open-set detection. Compared to MaxLogit [16], which does not use normalization, and MSP [17], which utilizes Softmax normalization, the Sigmoid normalization used DDStereo achieves higher performance. This could be attributed to



the use of Sigmoid-based Focal loss classification loss during the training phase. Our strategy of subtracting the highest known-class score from the binary foreground score is simple yet effective.

**Generalization with Limited Extra Data.** S3AD [33] proposed using synthetic datasets to generalize scale prediction capabilities, with its open-set detection performance improving as more supplementary data is introduced. In Table 9, we compare the performance of DDSTereo and S3AD under limited additional data. When only 10% (400 images) of the extra synthetic dataset is added, our method demonstrates a more pronounced performance advantage.

**Table 9:** Generalization with Limited Extra Data.

| Method    | AR-ExD | $AP_{3D}^{OoD}$ | $AP_{3D}^{Ped}$ | $AP_{3D}^{Cyc}$ |
|-----------|--------|-----------------|-----------------|-----------------|
| S3AD [33] | ×10%   | 24.45           | 42.88           | 21.58           |
| DDSTereo  | ×10%   | <b>54.65</b>    | <b>47.76</b>    | <b>29.27</b>    |
| S3AD [33] | ×100%  | 73.69           | 45.74           | 20.82           |
| DDSTereo  | ×100%  | <b>78.09</b>    | <b>52.51</b>    | <b>31.69</b>    |

**Table 10:** Comparison of computational load. Values in parentheses is our tested speeds.

| Method                 | OoD | Time(ms) | GFlops | Params  |
|------------------------|-----|----------|--------|---------|
| YOLOStereo3D [28]      | ×   | 80 (71*) | 177.82 | 107.6 M |
| StereoDETR [32]        | ×   | 18 (18*) | 59.80  | 10.2 M  |
| MonoDETR [52]          | ×   | 20 (19*) | 62.96  | 37.7 M  |
| MonoDGP [37]           | ×   | 22 (20*) | 71.79  | 42.2 M  |
| S3AD [33]              | ✓   | 75 (75*) | 180.71 | 109.5 M |
| <b>DDSTereo (Ours)</b> | ✓   | 24       | 62.65  | 19.6 M  |

#### 4.5 Model Efficiency

We compare the computational efficiency and parameter size of our model with both monocular and stereo-based baselines in Table 10. In terms of computational cost, DDSTereo requires only 62.65 GFLOPs, lower than the open-set stereo method S3AD [33] and even slightly lower than MonoDETR [52] and MonoDGP [37]. Furthermore, our model contains only 19.6 M parameters, which is nearly half the size of the monocular models, and less than one fifth of S3AD [33]. These results demonstrate that our dual-decoder stereo architecture achieves a favorable balance between accuracy and efficiency, making it suitable for real-time deployment in autonomous driving systems. Compared to StereoDETR [32], the fastest existing closed-set stereo detector, DDSTereo expands its capability to arbitrary OoD obstacle detection with a marginal overhead of only 6 ms in latency and 2.75 GFLOPs.

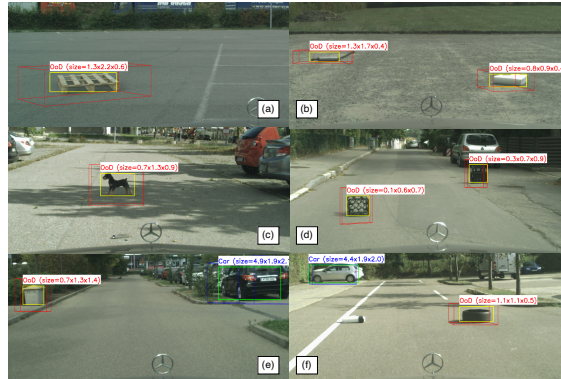
#### 4.6 Visualization

**Image-editing-based synthetic datasets.** Fig. 5 visualizes the detection of anomaly classes on the AR-OoD [33] dataset. Despite these categories and similar ones being entirely absent during training, our approach leverages stereo disparity to accurately estimate the BEV layout of novel classes, though some scale estimation errors persist.



**Fig. 5:** Visualization results with 3D boxes and BEV map. In the RGB images, **red** 3D boxes represent predicted anomaly categories, **blue** 3D boxes represent predicted known categories, and **green** boxes indicate the ground truth.

**In-the-Wild Validation.** To further evaluate real-world performance, we perform cross-camera testing on LaF [35], a road anomaly segmentation dataset providing paired stereo images. Our model is trained solely on 7.7K samples from AR-OoD and KITTI. As visualized in Fig. 6, despite misaligned camera parameters, the disparity-based method successfully detects unknown obstacles. Yellow and red boxes represent predicted 2D and 3D anomaly bounding boxes, respectively. Subfigures (e) and (f) present failure cases: a false positive detection of a roadside trash can and a missed white obstacle, respectively. Overall, with merely 7K training samples, our model demonstrates reliable generalization for stereo 3D anomaly detection.



**Fig. 6:** Zero-shot test on LaF [35]. **yellow**: 2D predictions; **red**: 3D projections.

## 5 Conclusion and Limitations

We propose DDStereo, a dual-decoder transformer that unifies open-set detection with real-time stereo 3D detection. By leveraging shared queries and a streamlined architecture, DDStereo achieves a superior accuracy-efficiency trade-off, marking the first stereo anomaly detector to match monocular close-set baselines in speed. However, limitations remain: the model’s scalability across larger, more diverse datasets requires further validation, and its generalization to varied camera parameters is constrained by geometric sensitivities. Addressing these via robust rectification and large-scale pre-training will be the focus of future work.

## Acknowledgements

This work was supported in part by the 6G Science and Technology Innovation and Future Industry Cultivation Special Project of Shanghai Municipal Science and Technology Commission under Grant 24DP1501001, in part by the National High Quality Program under Grant TC220H07D, and in part by the Xi’an Jiaotong-Liverpool University under Grant for ILAI.

## References

1. Alliegro, A., Cappio Borlino, F., Tommasi, T.: 3dos: Towards 3d open set learning - benchmarking and understanding semantic novelty detection on point clouds. In: *Advances in Neural Information Processing Systems*. vol. 35, pp. 21228–21240. Curran Associates, Inc. (2022)
2. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2020)
3. Cao, Y., Yihan, Z., Xu, H., Xu, D.: Coda: Collaborative novel box discovery and cross-modal alignment for open-vocabulary 3d object detection. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) *Advances in Neural Information Processing Systems*. vol. 36, pp. 71862–71873. Curran Associates, Inc. (2023)
4. Cen, J., Yun, P., Cai, J., Wang, M.Y., Liu, M.: Open-set 3d object detection. In: *2021 International Conference on 3D Vision (3DV)*. pp. 869–878 (2021)
5. Chang, M.F., Lambert, J., Sangkloy, P., Singh, J., Bak, S., Hartnett, A., Wang, D., Carr, P., Lucey, S., Ramanan, D., Hays, J.: Argoverse: 3d tracking and forecasting with rich maps. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2019)
6. Chen, J., Song, Q., Guo, W., Huang, R.: Dsc3d: Deformable sampling constraints in stereo 3d object detection for autonomous driving. *IEEE Transactions on Circuits and Systems for Video Technology* **35**(3), 2794–2805 (2025)
7. Chen, Y., Liu, S., Shen, X., Jia, J.: Dsgn: Deep stereo geometry network for 3d object detection. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2020)

8. Cheng, T., Song, L., Ge, Y., Liu, W., Wang, X., Shan, Y.: Yolo-world: Real-time open-vocabulary object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16901–16911 (2024)
9. Dhamija, A., Gunther, M., Ventura, J., Boulton, T.: The overlooked elephant of object detection: Open set. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 1021–1030 (March 2020)
10. Gao, A., Pang, Y., Nie, J., Shao, Z., Cao, J., Guo, Y., Li, X.: Esgn: Efficient stereo geometry network for fast 3d object detection. *IEEE Transactions on Circuits and Systems for Video Technology* **34**(4), 2000–2009 (2024)
11. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. pp. 3354–3361 (Jun 2012)
12. Gu, X., Lin, T.Y., Kuo, W., Cui, Y.: Open-vocabulary object detection via vision and language knowledge distillation. In: International Conference on Learning Representations (2022)
13. Guo, X., Shi, S., Wang, X., Li, H.: Liga-stereo: Learning lidar geometry aware representations for stereo-based 3d detector. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3153–3163 (October 2021)
14. Gupta, A., Narayan, S., Joseph, K.J., Khan, S., Khan, F.S., Shah, M.: Ow-detr: Open-world detection transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9235–9244 (June 2022)
15. Han, J., Ren, Y., Ding, J., Pan, X., Yan, K., Xia, G.S.: Expanding low-density latent regions for open-set object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9591–9600 (June 2022)
16. Hendrycks, D., Basart, S., Mazeika, M., Zou, A., Kwon, J., Mostajabi, M., Steinhardt, J., Song, D.: Scaling out-of-distribution detection for real-world settings. In: Proceedings of the 39th International Conference on Machine Learning. vol. 162, pp. 8759–8773. PMLR (Jul 2022)
17. Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. In: International Conference on Learning Representations (2017)
18. Huang, K.C., Wu, T.H., Su, H.T., Hsu, W.H.: Monodtr: Monocular 3d object detection with depth-aware transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4012–4021 (June 2022)
19. Joseph, K.J., Khan, S., Khan, F.S., Balasubramanian, V.N.: Towards open world object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5830–5840 (June 2021)
20. Konigshof, H., Salscheider, N.O., Stiller, C.: Realtime 3d object detection for automated driving using stereo vision and semantic information. In: 2019 IEEE Intelligent Transportation Systems Conference (ITSC). pp. 1405–1410 (Oct 2019)
21. Konigshof, H., Stiller, C.: Learning-based shape estimation with grid map patches for realtime 3d object detection for automated driving. In: 2020 IEEE 23rd International conference on intelligent transportation systems (ITSC). pp. 1–6. IEEE (2020)
22. Kumar, A., Brazil, G., Corona, E., Parchami, A., Liu, X.: Deviant: Depth equivariant network for monocular 3d object detection. In: European Conference on Computer Vision. pp. 664–683. Springer (2022)

23. Li, P., Chen, X., Shen, S.: Stereo r-cnn based 3d object detection for autonomous driving. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (Jun 2019)
24. Li, P., Su, S., Zhao, H.: Rts3d: Real-time stereo 3d detection from 4d feature-consistency embedding space for autonomous driving. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 1930–1939 (2021)
25. Lian, Q., Li, P., Chen, X.: Monojs: Joint semantic and geometric cost volume for monocular 3d object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1070–1079 (June 2022)
26. Lin, C., Jiang, Y., Qu, L., Yuan, Z., Cai, J.: Generative region-language pretraining for open-ended object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13958–13968 (June 2024)
27. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In: European conference on computer vision. pp. 38–55. Springer (2024)
28. Liu, Y., Wang, L., Liu, M.: Yolostereo3d: A step back to 2d for efficient stereo 3d detection. In: 2021 IEEE International Conference on Robotics and Automation (ICRA). pp. 13018–13024 (2021)
29. Liu, Z., Ye, X., Tan, X., Ding, E., Bai, X.: Stereodistill: Pick the cream from lidar for distilling stereo-based 3d object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 1790–1798 (2023)
30. Lu, Y., Ma, X., Yang, L., Zhang, T., Liu, Y., Chu, Q., He, T., Li, Y., Ouyang, W.: Gupnet++: Geometry uncertainty propagation network for monocular 3d object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **47**(2), 900–915 (2025)
31. Lu, Y., Xu, C., Wei, X., Xie, X., Tomizuka, M., Keutzer, K., Zhang, S.: Open-vocabulary point-cloud object detection without 3d annotation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1190–1199 (June 2023)
32. Mu, S., Gu, Z., Ai, Z., Liu, A., Gao, Y., Xu, S.: Stereodetr: Stereo-based transformer for 3d object detection. *IEEE Transactions on Circuits and Systems for Video Technology* **36**(4), 4371–4383 (2026)
33. Mu, S., Gu, Z., Lyu, H., Gao, Y., Xu, S.: Stereo-based 3d anomaly object detection for autonomous driving: A new dataset and baseline. *IEEE Internet of Things Journal* pp. 1–1 (2026)
34. Peng, X., Zhu, X., Wang, T., Ma, Y.: Side: Center-based stereo 3d detector with structure-aware instance depth estimation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 119–128 (January 2022)
35. Pinggera, P., Ramos, S., Gehrig, S., Franke, U., Rother, C., Mester, R.: Lost and found: detecting small road hazards for self-driving vehicles. In: 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 1099–1106 (2016)
36. Pon, A.D., Ku, J., Li, C., Waslander, S.L.: Object-centric stereo matching for 3d object detection. In: 2020 IEEE International Conference on Robotics and Automation (ICRA). pp. 8383–8389 (2020)
37. Pu, F., Wang, Y., Deng, J., Yang, W.: Monodgp: Monocular 3d object detection with decoupled-query and geometry-error priors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6520–6530 (June 2025)

38. Qin, Z., Wang, J., Lu, Y.: Triangulation learning network: From monocular to stereo 3d object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7615–7623 (June 2019)
39. Shi, Y., Guo, Y., Mi, Z., Li, X.: Stereo centernet-based 3d object detection for autonomous driving. *Neurocomputing* **471**, 219–229 (2022)
40. Sun, H., Pang, Y., Cao, J., Xie, J., Li, X.: Transformer-based stereo-aware 3d object detection from binocular images. *IEEE Transactions on Intelligent Transportation Systems* **25**(12), 19675–19687 (2024)
41. Sun, P., Kretschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., Vasudevan, V., Han, W., Ngiam, J., Zhao, H., Timofeev, A., Ettinger, S., Krivokon, M., Gao, A., Joshi, A., Zhang, Y., Shlens, J., Chen, Z., Anguelov, D.: Scalability in perception for autonomous driving: Waymo open dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2446–2454 (June 2020)
42. Tao, C., Cao, C., Cheng, H., Gao, Z., Luo, X., Zhang, Z., Zheng, S.: An efficient 3d object detection method based on fast guided anchor stereo rcnn. *Advanced Engineering Informatics* **57**, 102069 (2023)
43. Wang, Z., Li, Y., Liu, T., Zhao, H., Wang, S.: Ov-uni3detr: Towards unified open-vocabulary 3d object detection via cycle-modality propagation. In: European Conference on Computer Vision. pp. 73–89. Springer (2024)
44. Wu, X., Zhu, F., Zhao, R., Li, H.: Cora: Adapting clip for open-vocabulary detection with region prompting and anchor pre-matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7031–7040 (June 2023)
45. Xue, L., Gao, M., Xing, C., Martín-Martín, R., Wu, J., Xiong, C., Xu, R., Niebles, J.C., Savarese, S.: Ulip: Learning a unified representation of language, images, and point clouds for 3d understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1179–1189 (June 2023)
46. Yao, J., Gu, H., Chen, X., Wang, J., Cheng, Z.: Open vocabulary monocular 3d object detection. In: 2026 International Conference on 3D Vision (3DV). pp. 1801–1811 (2026)
47. Yao, L., Han, J., Liang, X., Xu, D., Zhang, W., Li, Z., Xu, H.: Detclipv2: Scalable open-vocabulary object detection pre-training via word-region alignment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 23497–23506 (June 2023)
48. Yao, L., Han, J., Wen, Y., Liang, X., Xu, D., Zhang, W., Li, Z., Xu, C., Xu, H.: Detclip: Dictionary-enriched visual-concept paralleled pre-training for open-world detection. *Advances in Neural Information Processing Systems* **35**, 9125–9138 (2022)
49. Yao, L., Pi, R., Han, J., Liang, X., Xu, H., Zhang, W., Li, Z., Xu, D.: Detclipv3: Towards versatile generative open-vocabulary object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 27391–27401 (June 2024)
50. Zhang, H., Jiang, H., Yao, Q., Sun, Y., Zhang, R., Zhao, H., Li, H., Zhu, H., Yang, Z.: Detect anything 3d in the wild. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 5048–5059 (October 2025)
51. Zhang, R., Guo, Z., Zhang, W., Li, K., Miao, X., Cui, B., Qiao, Y., Gao, P., Li, H.: Pointclip: Point cloud understanding by clip. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8552–8562 (June 2022)

52. 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 (ICCV). pp. 9155–9166 (October 2023)
53. Zhang, Y., Lu, J., Zhou, J.: Objects are different: Flexible monocular 3d object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3289–3298 (June 2021)