Exploiting Vision Language Model for Training-Free 3D Point Cloud OOD Detection via Graph Score Propagation

Supplementary Material

A. Details of Datasets

Partition of dataset: We begin by following the 3DOS methodology to partition the ShapeNetCore (Synthetic benchmark) and ScanObjectNN (Real benchmark) datasets. For the Real benchmark, synthetic point clouds from ModelNet40 are used for training, while testing is conducted on real-world point clouds from ScanObjectNN. For the Synthetic benchmark, the ShapeNetCore dataset is divided into three non-overlapping (i.e., semantically distinct) category sets, each containing 18 categories. The specific categories for the SR/SN sets are detailed in Tab. 2, with illustrative examples shown in Fig. 5.

The Sydney Urban Objects Dataset provides high-resolution point clouds that capture detailed geometric information about various urban objects, such as buildings, vehicles, trees, and street furniture. To demonstrate the application of out-of-distribution (OOD) methods in real-world scenarios, we split the dataset into two parts: movable objects and non-movable objects. Movable objects are treated as in-distribution (ID) classes, while non-movable objects are considered OOD classes.

The Stanford Large-Scale 3D Indoor Spaces (S3DIS) dataset is a comprehensive benchmark for 3D indoor scene understanding. To ensure sufficient geometric fidelity, we constructed the test dataset by retaining instances with over 2,048 raw points, resulting in 8,931 high-density instances.

The nuScenes dataset is a multimodal autonomous driving resource featuring 1,000 urban scenes captured via LiDAR, radar, and cameras, with 3D annotations. Given the scale of the nuScenes dataset, we utilized its trainval.mini subset to balance data diversity and computational feasibility. Point clouds with fewer than 200 raw points were discarded to ensure reliable geometric representation, resulting in a curated collection of 2205 high-resolution LiDAR frames to constructing testing dataset.

Some ID classes and OOD classes of Sydney Urban Objects, S3DIS and nuScenes dataset are shown in Fig. 6. The category classification is provided in Tab. 3.

Dataset Preprocessing: For the Sydney Urban Objects and nuScenes Dataset, the number of points in each frame of the point cloud varies, as the data is cropped from real-world scenes. To ensure consistency, we randomly sample each point cloud to standardize it to 1024 points per point cloud. Figure 1 illustrates examples of the original point clouds alongside their normalized counterparts. It is evident that certain categories, such as ”4wd” and ”biker,” contain fewer points in the original point cloud, whereas others, like ”bus,” have significantly more points.

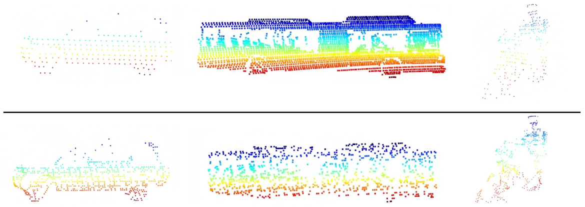

Figure 1. Illustration of point cloud interpolation results. We uniformly subsample or interpolate each sample to 2048 points.

To address these discrepancies, we implement a data augmentation strategy. For categories with fewer points, we generate synthetic points by interpolating between each original point and its nearest neighbors. Specifically, for each point, a random neighbor is selected, and a specified number of new points are created through linear interpolation. The interpolation coefficient is adjusted to distribute the new points evenly between the original point and its neighbor. For categories with excess points, we apply subsampling to reduce the point count. This approach ensures that all point clouds are consistently normalized to 1024 points before being input into the model.

Prompt templates: We follow the ULIP framework to construct prompts by applying each class name to 64 predefined templates, generating a comprehensive set of text descriptions for Prompt Clustering. The full list of prompt templates is provided in Table 4. We empirically find that including the template ”a photo of ” to be beneficial for 3D point cloud OOD detection. This is due to the fact that ULIP is built upon CLIP by aligning point cloud feature to text and image (rendered from 3D point cloud) features. Thus the text prompt of “a photo of ” is a legacy from CLIP.

B. Additional Studies

Study of Self-Training: To investigate the self-training process, we visualized the feature distributions in Fig. 2. Different rows correspond to various experimental datasets, while different columns represent varying numbers of negative prompts. The first column shows the initialization of all features, and the subsequent columns illustrate the feature distributions when the number of negative prompts is set to 10, 15, and 20, respectively. From this figure, we observe that positive prompts tend to cluster closer to ID samples, while negative prompts align more closely with OOD samples. This distinction helps assign higher scores to ID samples and lower scores to OOD samples during the la-

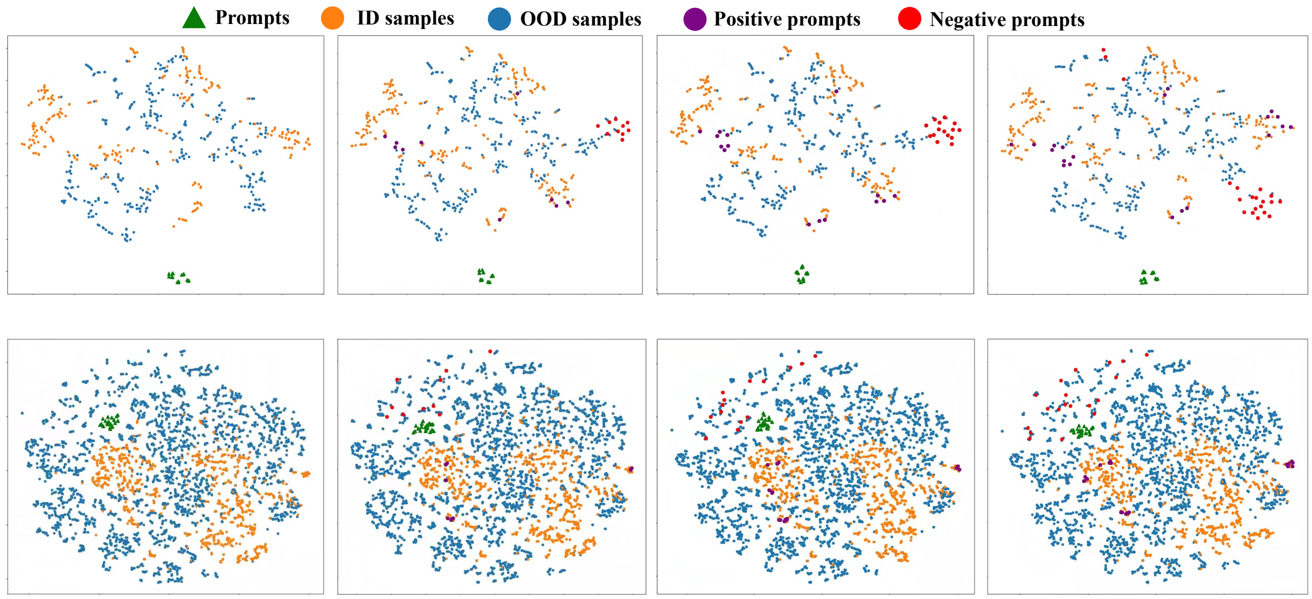

Figure 2. T-SNE Visualization on the ScanObjNN and ShapNetCore Dataset for self-training with different number of negative prompts. The first row displays the visualization on the ScanObjNN, and the second row for ShapNetCore Dataset.

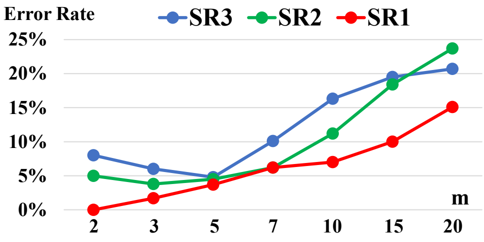

Figure 3. Error rates of pseudo prompts with different filtering ratio m.

bel propagation stage. These findings highlight the critical role of generating negative prompts through self-training in enhancing the effectiveness of the process.

VLM backbone: ULIP was chosen as the default VLM backbone due to its superior performance. However, the proposed GSP method is agnostic to the VLM backbone, as demonstrated by additional evaluations using PointClip V2 as the backbone in Tab.1. GSP consistently shows superior performance with PointClip V2 as well.

Incorrect pseudo prompts: Self-training is known to be sensitive to incorrect pseudo labels. To further analyze the impact of pseudo prompt selection, we vary m from 2 to 20 and examine the error rate on selected pseudo prompts (Fig. 3). As m decreases (stricter selection), the error rate on pseudo prompts reduces. On average, $m = 5$ proves to be an effective hyperparameter for selecting reliable pseudo prompts.

Study of negative prompt: The ablation study in Fig 4 evaluates the impact of negative prompts on ScanObjectNN Dataset, the results reveal a consistent performance gap between models with and without negative prompts.

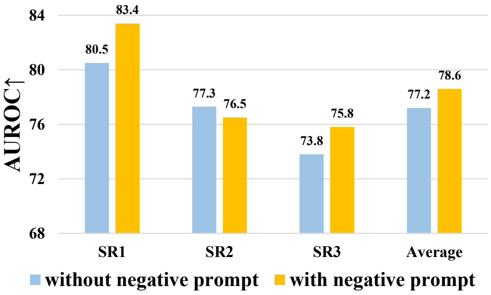

Figure 4. Ablation Study on the Impact of Negative Prompts: Blue Bars Indicate Results Without Negative Prompts, While Yellow Bars Indicate Results With Negative Prompts. All Experiments Conducted on the ScanObjNN Dataset.

C. Broader Impact and Limitations

C.1 Broader Impact

The proposed method could improve the efficacy of generalizing pre-trained 3D VLM for real-world OOD detection tasks. Adopting the techniques could benefit autonomous driving and robotics, potentially improving the safety. Potential risks include failing to differentiate OOD from ID may lead to collision and fatal consequences.

C.2 Potential Limitations

The method requires seeing a substantial amount of testing data so that a graph can be built and inference can benefit from the manifold information. If testing data arrives in a stream, the proposed method could be implemented in a batch mode or incrementally build a graph for inference.

Method	SR3		SR2		SR1		Average		MN1		MN2		MN3		Average	
	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
Cosine Dist	36.5	96.1	62.2	92.9	63.3	83.2	54.0	90.7	68.2	83.3	69.3	79.7	71.9	75.0	69.8	79.3
GSP(Ours)	48.5	81.4	63.1	93.4	67.6	81.2	59.7	85.3	71.8	68.9	73.3	64.9	69.2	74.1	71.4	69.3

Table 1. The results of distance-based and GSP on Modelnet40 and ScanObjectNN dataset with PointClip V2 as VLM backbone.

The Real benchmark(ScanobjNN)		
SR1	SR2	SR3
chair, shelf, door, sink, sofa	bed, toilet, desk, table, display	bag, bin, box, pillow, cabinet
The Synthetic benchmark(ShapeNetCore)		
SN1	SN2	SN3
mug, lamp, bed, washer, loudspeaker telephone, dishwasher, camera, birdhouse, jar bowl, bookshelf, stove, bench, display , keyboard, clock, piano	earphone, knife, chair, pillow, table, mailbox, basket, file cabinet, cabinet, sofa flowerpot, microphone, tower, bag, bathtub laptop, printer, trash bin	can, microwave, skateboard, faucet, train pistol, helmet, watercraft, airplane, bottle rocket, rifle, remote, car, bus guitar, cap, motorbike

Table 2. Classification of Categories in the ShapeNetCore and ScanobjNN Dataset.

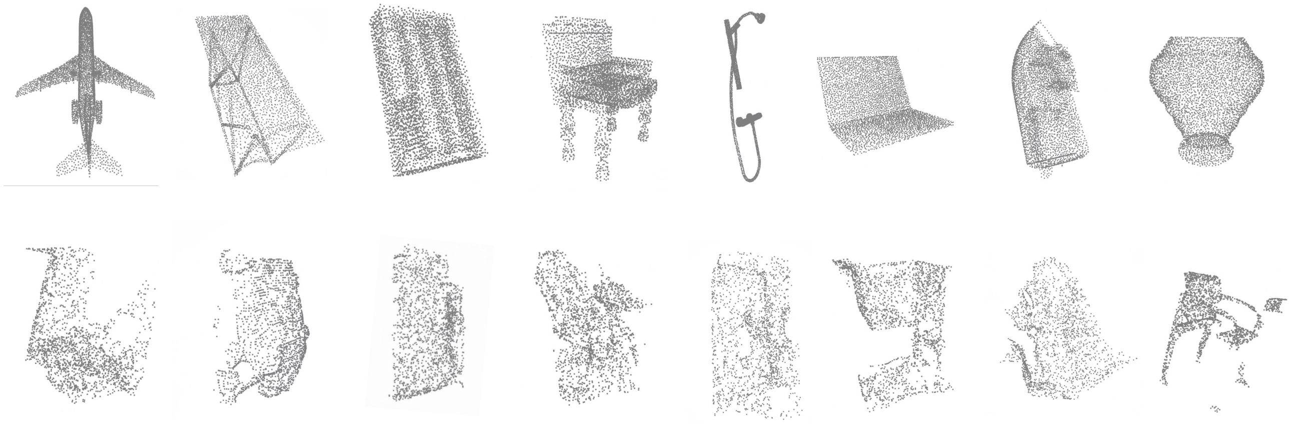

Figure 5. Visualization of some point clouds from the ScanObjNN and ShapeNetCore Dataset. The first row displays the visualization on the ShapNetCore , and the second row for ScanObjNN Dataset.

Dataset	in-distribution categories	out-of-distribution categories
The Sydney Urban Objects	bus, car, cyclist, excavator, pedestrian, scooter, 4wd, bicycle, biker, trailer, truck, van	trash, tree, trunk, umbrella, ute, pillar, pole, post, building, bench, ticket machine, traffic lights, traffic sign, vegetation
S3DIS	window, door, table, chair, clutter, sofa, bookcase	floor, ceiling, wall, beam, board, column
nuScenes	pedestrain, car, motorcycle	barrier, truck, bus, traffic_cone, construction_vehicle, trailer

Table 3. Classification of Categories in the Sydney Urban Objects Dataset, S3DIS dataset and nuScenes dataset.

Prompt templates

“a point cloud model of {}.”
“a photo of the {} in the scene.”
“itap of the {}.”
“a photo of one {}.”
“a bad photo of a {}.”
“a photo of a cool {}.”
“a photo of a small {}.”
“a photo of a clean {}.”
“a bright photo of a {}.”
“a photo of a hard to see {}.”
“a low resolution photo of the {}.”
“a jpeg corrupted photo of a {}.”
“a pixelated photo of a {}.”
“a black and white photo of a {}”
“a plushie {}.”
“an embroidered {}.”
“There is a {} in the scene.”
“a photo of one {} in the scene.”
“a photo of a {}.”
“a photo of many {}.”
“a bad photo of the {}.”
“a photo of the cool {}.”
“a photo of the small {}.”
“a photo of the clean {}.”
“a bright photo of the {}.”
“a photo of the hard to see {}.”
“a cropped photo of the {}.”
“a jpeg corrupted photo of the {}.”
“a pixelated photo of the {}.”
“the plastic {}.”
“the plushie {}.”
“the embroidered {}.”
“There is the {} in the scene.”
“itap of a {}.”
“a photo of my {}.”
“a good photo of a {}.”
“a photo of a nice {}.”
“a photo of a weird {}.”
“a photo of a large {}.”
“a photo of a dirty {}.”
“a dark photo of a {}.”
“a low resolution photo of a {}.”
“a close-up photo of a {}.”
“a blurry photo of a {}.”
“a black and white photo of the {}.”
“a toy {}.”
“a cartoon {}.”
“a painting of the {}.”
“a photo of a {} in the scene.”
“itap of my {}.”
“a photo of the {}.”
”a good photo of the {}.”
“a photo of the nice {}.”
“a photo of the weird {}.”
“a photo of the large {}.”
“a photo of the dirty {}.”
“a dark photo of the {}.”
“a cropped photo of a {}.”
“a close-up photo of the {}.”
“a blurry photo of the {}.”
“a plastic {}.”
“the toy {}.”
“the cartoon {}.”
”a painting of a {}.”

Table 4. All prompt templates for GSP.

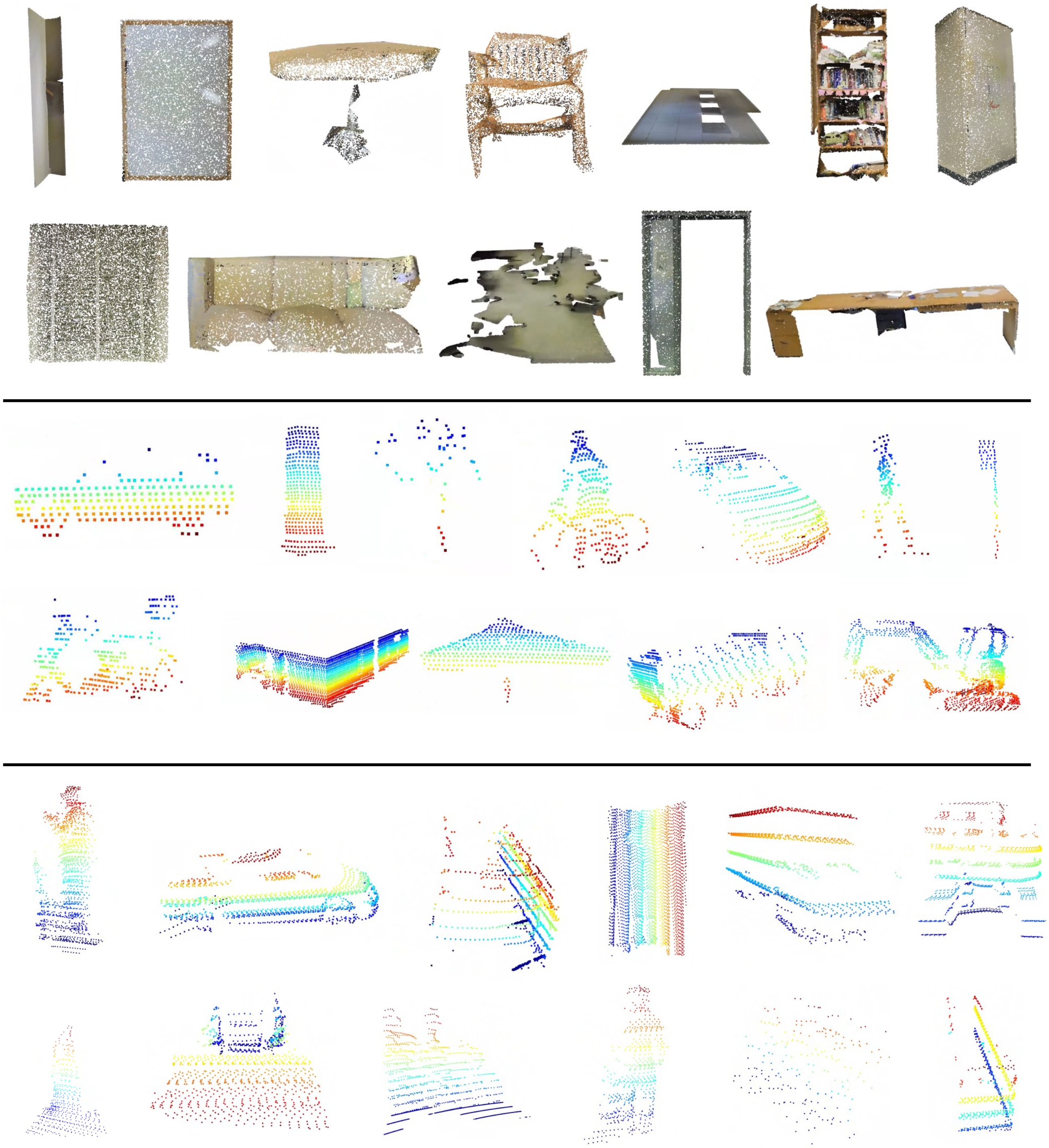

Figure 6. Visualization of some point clouds from the S3DIS, Sydney Urban Objects and nuScenes Dataset. The first row displays the visualization on the S3DIS, the second row for Sydney Urban Objects, and the third row for nuScenes Dataset.