3D Semantic Novelty Detection via Large-Scale Pre-Trained Models

IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received 22 August 2024, accepted 17 September 2024, date of publication 19 September 2024, date of current version 30 September 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3464334

RESEARCH ARTICLE

3D Semantic Novelty Detection via Large-Scale Pre-Trained Models

PAOLO RABINO[1,2], (Student Member, IEEE), ANTONIO ALLIEGRO[1], (Student Member, IEEE), AND TATIANA TOMMASI[1], (Member, IEEE)

[1]Department of Control and Computer Engineering, Polytechnic University of Turin, 10129 Turin, Italy
[2]Italian Institute of Technology, 16163 Genoa, Italy

Corresponding author: Paolo Rabino (paolo.rabino@polito.it)

The work of Antonio Alliegro was supported by the EU Project ELSA–European Lighthouse on Secure and Safe AI. The work of Tatiana Tommasi was supported by the FAIR–Future Artificial Intelligence Research funded by the European Union Next-GenerationEU (Piano Nazionale di Ripresa e Resilienza (PNRR)–Missione 4 Componente 2, Investimento 1.3–D.D. 1555 11/10/2022) under Grant PE00000013.

ABSTRACT Shifting deep learning models from lab environments to real-world settings entails preparing them to handle unforeseen conditions, including the chance of encountering novel objects from classes that were not included in their training data. Such occurrences can pose serious threats in various applications. The task of Semantic Novelty detection has attracted significant attention in the last years mainly on 2D images, overlooking the complex 3D nature of the real-world. In this study, we address this gap by examining the geometric structures of objects within 3D point clouds to detect semantic novelty effectively. We advance the field by introducing 3D-SeND, a method that harnesses a large-scale pre-trained model to extract patch-based object representations directly from its intermediate feature representation. These patches are used to characterize each known class precisely. At inference, a normalcy score is obtained by assessing whether a test sample can be reconstructed predominantly from patches of a single known class or from multiple classes. We evaluate 3D-SeND on real-world point cloud samples when the reference known data are synthetic and demonstrate that it excels in both standard and few-shot scenarios. Thanks to its patch-based object representation, it is possible to visualize 3D-SeND's predictions with a valuable explanation of the decision process. Moreover, the inherent training-free nature of 3D-SeND allows for its immediate application to a wide array of real-world tasks, offering a compelling advantage over approaches that require a task-specific learning phase. Our code is available at https://paolotron.github.io/3DSend.github.io.

INDEX TERMS 3D point clouds, semantic novelty detection, out-of-distribution detection, training-free.

I. INTRODUCTION

Semantic Novelty Detection (SeND) consists of identifying instances of object categories not previously observed in a reference dataset. This task involves analyzing the semantic image content to determine whether it contains unfamiliar information. Despite the widely acknowledged success of deep learning models, SeND offers them several challenges when shifting from the constrained laboratory setting to the open world. In this scenario, recognizing novelty is crucial in ensuring the model's reliability and safety.

Traditional Out-of-Distribution (OOD) detection approaches focus on mitigating classification over-confidence but they often overlook the distinction between domain and semantic novelty which leads to rejecting instances of known classes appearing with a different visual style [1], [2], [3], [4]. Moreover, these approaches are not suitable for many practical applications where the learning agent has access to limited computational resources that constrain the training phase, and the reference dataset that exemplifies normalcy is composed of only a few object exemplars.

A recent trend consists of leveraging large-scale pre-trained models and fine-tuning them on the reference set. However, this process may be suboptimal as fine-tuning can lead to catastrophic forgetting, reducing the generalization

The associate editor coordinating the review of this manuscript and approving it for publication was Tallha Akram.

ability of the original model rather than supporting novelty recognition [5], [6]. Alternative solutions adopt a zero-shot approach by exploiting vision-language foundation models [7], [8]. This strategy has the advantage of avoiding any further learning process as the test samples are compared directly to textual class names used as prototypes in the shared embedding space. More recently, prompt engineering has been proposed to add refined linguistic descriptions of the known classes, which however either require manual tuning or re-introduces a training phase [9], [10], [11].

Overall, research in the field of novelty detection has attracted a lot of attention in the last years mainly for 2D data types [12], [13]. Only recently the introduction of the 3DOS benchmark [14] for 3D Open-Set recognition and the described challenges of adapting 2D OOD detection methods to 3D data started to call for new 3D solutions. At the same time, the emergence of vast multi-modal datasets featuring 3D modality, such as Objaverse [15], is transforming the 3D research landscape, showcasing how large deep learning models can support reasoning on 3D data [16], [17]. This shift paves the way to explore foundation-model-based training-free strategies for OOD detection on 3D data. Indeed, strategies that do not require learning on tailored task-specific data collections sound well suited for 3D data as the costs of data collection and model training increase with data dimensionality.

In this work, we present a comprehensive examination of large-scale pre-trained models, discussing how their latent representation can be efficiently and effectively exploited for 3D semantic novelty detection. We focus on a challenging SeND setting that mimics a typical industrial scenario where the reference set consists of a few curated 3D synthetic data, while the test samples are derived from real-world scans captured by on-site 3D sensors. These scans exhibit a different visual domain compared to the curated synthetic data and encompass samples from both known categories present in the support set and novel unknown ones. To tackle the task **we introduce 3D-SeND, an innovative method that utilizes large-scale pre-trained 3D feature encoders for extracting patch-based representations of 3D objects**. These representations capture both local and global attributes of objects and we use their co-occurrence to devise a tailored novelty score. We perform an extensive analysis assessing the role of different pre-training networks and objectives, further considering score variants, and show experimentally that 3D-SeND excels in the separation of known and unknown test samples without requiring a tailored fine-tuning phase on the reference set. Remarkably, 3D-SeND surpasses competing methods, demonstrating exceptional performance even in scenarios where the reference data is scarce. Furthermore, its design based on discovered semantically relevant (*i.e.*, a *leg* of a table) and geometrically significant (*i.e.*, a *cylinder*) patches, provides visualizations that make it inherently explainable.

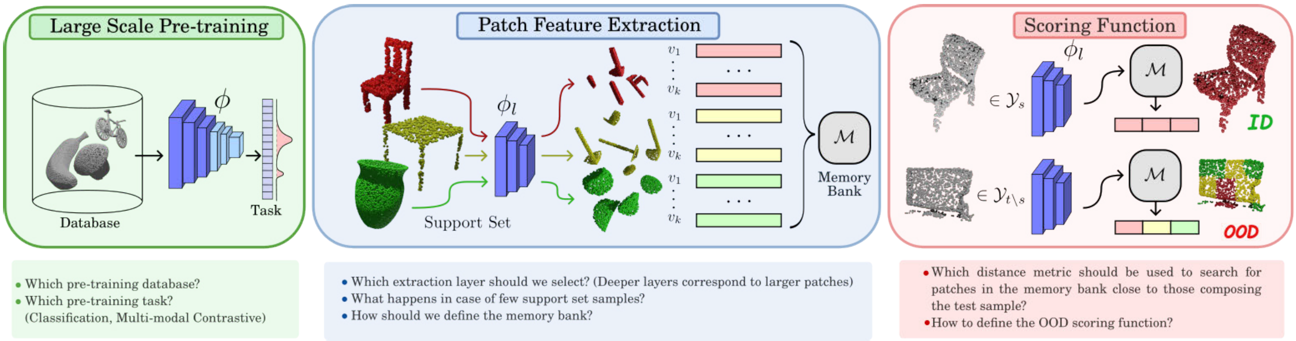
We believe that our new solution to leverage the knowledge captured by 3D foundation models will serve as a stepping stone to enhance the dependability of artificial intelligence approaches for open-world applications.

II. RELATED WORKS

Out-of-distribution detection is an umbrella term for many subcategories of methods designed to identify novelty at inference time. Part of the differences among these categories originate from the source of novelty (*i.e.* due to covariate or semantic shift) while others relate to the exact experimental setting. The basic OOD detection framework consists of a simple binary task that separates samples belonging to a *known* reference distribution from samples drawn from a different *unknown* one. The instances of the first group are often identified as In-Distibution (ID) samples and they may be structured in multiple classes. Discriminating among these classes while rejecting novelty is indicated as *Open Set Recognition*. Finally, the focus of *Anomaly Detection* is on locating abnormal parts within a scene or an object instance. In industrial applications, this means training a model for each object class to spot possible components (*e.g.* defective parts) that deviate from the reference normalcy.

In this work we are interested in Semantic Novelty detection, thus we overview those methods in the OOD detection literature that can be used with the binary objective of recognizing whether a new sample belongs to one of the known classes or not, neglecting domain or style variations. A simple strategy consists of relying on the Maximum Softmax Prediction (MSP) of a classifier trained on the ID reference data [1]. Other approaches have followed the same *post-hoc paradigm* exploiting a classifier output by introducing temperature scaling to reduce overconfidence [2], energy scores that estimate the probability density of the input [3], leveraging the norm of the network gradients [12], or rectifying the network activations [13]. The *outlier exposure* methods [18], [19], [20], [21] assume the availability of either real or synthetically generated OOD examples during training but present limited generalization abilities. *Density and reconstruction methods* explicitly model the distribution of known data. This can involve learning a generative model for input reconstruction [22] or exploiting a likelihood regret strategy [23]. *Distance-based methods* exploit a learned feature embedding and evaluate sample distances by using the L^2 norm [24], layer-wise Mahalanobis [25] or similarity metrics based on Gram [26] matrices.

In real-world applications, efficiency and robustness are pivotal, and three-dimensional reasoning is essential for agents interacting with their environment. Therefore, methods that can elaborate on 3D data for detecting novelty, requiring minimal learning effort, and exhibiting broad transferability and generalization across tasks should be prioritized. A few recent works have started to propose training-free OOD detection strategies in the 2D literature by exploiting the representation learned by large-scale pre-trained models [7], [27]. This logic has been also applied for 2D anomaly detection with promising results [28], [29],



- Which pre-training database?
- Which pre-training task? (Classification, Multi-modal Contrastive)

- Which extraction layer should we select? (Deeper layers correspond to larger patches)
- What happens in case of few support set samples?
- How should we define the memory bank?

- Which distance metric should be used to search for patches in the memory bank close to those composing the test sample?
- How to define the OOD scoring function?

FIGURE 1. Schematic visualization of the three main components of 3D-SeND and the associated research questions. Left: we start from a model pre-trained on a large-scale dataset capable of extracting semantically and geometrically relevant patch embeddings from point clouds. Middle: the embeddings extracted from the support set are collected into a memory bank encoding known classes. Right: at test time we extract patch embeddings from any new sample and compare them with the memory bank. If the nearest neighbors of the test patches in the memory bank are far away and the associated class labels show high entropy the score will suggest novelty (OOD), while a sample composed of patches with low distances and low entropy will be recognized as belonging to a known class (ID).

[30]. Despite these progresses in 2D, the research on OOD detection and SeND on 3D data is still in its infancy and deserves much more attention [31], [32], [33], [34], [35]. Indeed, as shown by the thorough OOD detection benchmark pursued in [14], 2D methods extended to 3D data are only mildly effective. A 3D approach based on part composition learning with outlier exposure for Open Set Recognition recently appeared in [36]. Although reasoning on object parts sounds promising, it falls short in addressing cross-domain scenarios and still necessitates ad-hoc training on task-specific ID reference data.

With our work, we explain how relevant 3D object patches and their relations can be extracted from the latent representation of large-scale pre-trained models and can be effectively used for SeND without requiring any further training.

III. METHOD

In SeND we are provided with annotated samples $\mathcal{S} = \{\boldsymbol{x}_i^s, y_i^s\}_{i=1}^N$ where $y_i^s \in \mathcal{Y}_s = \{1, \ldots, C\}$ indicate the label set, and we are asked to evaluate whether a test sample $\boldsymbol{x}^t$ belongs to an observed object class in $\mathcal{Y}_s$ or not. We name the reference annotated dataset $\mathcal{S}$ as *support set* and $\mathcal{Y}_s$ as *known classes*. The unlabeled data $\mathcal{T} = \{\boldsymbol{x}_j^t\}_{j=1}^J$ define the *test set* with $y_j^t \in \mathcal{Y}_t = \{1, \ldots, C, C+1, \ldots, K\}$ where all the classes in $\mathcal{Y}_{t\setminus s}$ are indicated as *unknown*.

Our method, named 3D-SeND, comprises three key components, as illustrated in Figure 1. The first is the **pre-training**. Unlike standard approaches that train a model on the support set $\mathcal{S}$, we leverage the expressive feature representation from a large-scale pre-trained model, eliminating the need for task-specific training. The choice of dataset, learning objective, and network architecture for pre-training may have varying effects on the downstream SeND task. The second component is the procedure for **patch feature extraction** which is used for both the support set and the test samples. This involves using the large-scale pre-trained feature encoder and selecting a specific layer within the network hierarchy to extract embeddings that represent portions of the input point cloud. In particular, the *patch embeddings* collected from the support set are organized in a memory bank that embodies the concept of known classes for the task at hand. Finally, the third component is the **scoring function**. For each of the patch features extracted from a test sample, we evaluate the distance to its nearest neighbor in the memory bank and its label. These two pieces of information are combined into a score that reflects the level of confidence in assigning the test sample to one of the known classes.

In the following subsections, we describe each of these components in more detail.

A. LARGE-SCALE PRE-TRAINING

3D-SeND leverages models that have encoded comprehensive knowledge about the structure of object point clouds within their latent representation. We consider two families of models, that are respectively trained on single-modal and multi-modal data. The first includes two point cloud encoding backbones trained for object classification with standard cross-entropy loss function on the Objaverse-LVIS dataset [37] containing 47K samples from 1156 semantic categories. Specifically, the architectures are PointNet++ [38] (1.50M backbone parameter) and EPN [39] (8.10M), with the latter chosen for its peculiar ability to learn SE(3)-equivariant features. The second family includes OpenShape [16] and Uni3D [17] which mainly differ from each other for the number of backbone parameters, respectively 32.33M and 88.96M. They are both pre-trained with a contrastive objective across three distinct input modalities (point cloud, image, and text) on an ensemble of four datasets (Objaverse [15], ShapeNet [40], ABO [41], 3D-Future [42]), resulting in 876K training shapes from 21K semantic categories.

B. PATCH FEATURE EXTRACTOR AND MEMORY BANK

Point-based backbone architectures employ a hierarchical approach to encode point clouds. They begin by capturing local features that represent detailed geometric structures from small neighborhood areas within the point cloud. These local features are then progressively aggregated into larger units, forming more semantic, higher-level features. As the network delves deeper, the receptive field of its layers broadens, enabling the deeper layers to model increasingly larger segments of the input point cloud. 3D-SeND is designed to extract local geometric features, also referred to as *patch embeddings*, from the internal feature representations of a specific network layer. Selecting a particular layer for this extraction means choosing the granularity of the patches.

More formally, given an input point cloud $\boldsymbol{x}$ and a 3D network ϕ, we denote the output feature map from its l-th layer as $\phi_l(\boldsymbol{x}) \in \mathbb{R}^{P_l \times C_l}$. This tensor can be interpreted as a collection of patch embeddings $\{\boldsymbol{v}_k\}_{k=1}^{P_l}$, where P_l is the number of patches extracted at the l-th layer, each with a feature descriptor with a dimension of C_l. Depending on the chosen architecture and the specific value of l, each vector $\boldsymbol{v}_k$ captures information about a distinct-sized portion of the 3D shape. For PointNet++, we use the multi-scale grouping classification backbone and extract patch embeddings after the l-th Set Abstraction (SA) layer. In this case, the number of patch embedding P_l obtained from each input point cloud is equal to the number of FPS points at the chosen SA layer. EPN exploits a point convolutional operator functioning within a discretized space of SO(3) rotations. Each convolutional layer produces a feature map of size $(P_l \times R \times C_l)$, where P_l represents the number of FPS points at the l-th layer, R denotes the fixed number of explicit rotations, and C_l represents the number of output channels. To extract patch embeddings from a specific layer, we employ a symmetric max function to aggregate information across the rotation dimension R, thus obtaining a $(P_l \times C_l)$ output tensor. Unlike PointNet++ and EPN, the patch embeddings of OpenShape and Uni3D correspond to the tokens output of the transformer blocks, and their number P_l remains constant throughout the network's depth. Due to the self-attention mechanism, as tokens pass through successive layers (or transformer blocks), they gather and integrate information from other tokens, modeling increasing portions of the input shape.

By feeding the support set $\mathcal{S}$ to any of the described frozen large-scale pre-trained models, having chosen a specific layer l, we obtain a set of patch-class pairs $\{\boldsymbol{v}_k^s, y_k^s\}_{k=1}^{P_l \times N}$. Here each patch embeddings is annotated with the label of the object sample from which it has been extracted. These pairs are aggregated into a unified memory bank, denoted as $\mathcal{M}$, and used as a reference to evaluate whether a test sample belongs to a known class.

C. SCORING FUNCTION

For each test sample $\boldsymbol{x}^t$, we extract a set of patch embeddings $\phi_l(\boldsymbol{x}^t) = \{\boldsymbol{v}_k\}_{k=1}^{P_l}$. Subsequently, for each patch, we perform nearest neighbor matching with samples stored in the memory bank and we define:

$$\delta(\boldsymbol{v}_k) = \min_{\boldsymbol{v}^s \in \mathcal{M}} d(\boldsymbol{v}_k, \boldsymbol{v}^s), \tag{1}$$

$$\lambda(\boldsymbol{v}_k) = y_{v^*}^s, \text{ where } v^* = \arg\min_{\boldsymbol{v}^s \in \mathcal{M}} d(\boldsymbol{v}_k, \boldsymbol{v}^s) \,. \tag{2}$$

Here, $\delta(\boldsymbol{v}_k)$ is the Euclidean distance of $\boldsymbol{v}_k$ to the nearest patch in the memory bank $\mathcal{M}$, and $\lambda(\boldsymbol{v}_k)$ denotes the class label of the nearest patch v^* in the memory bank $\mathcal{M}$. We use them to obtain the average distance of sample $\boldsymbol{x}^t$ to the support set class y^s, aggregating distances of patches whose nearest neighbors share the same label:

$$D_{y^s}(\boldsymbol{x}^t) = \frac{1}{P_l} \sum_{k=1}^{P_l} \delta(\boldsymbol{v}_k) \mathbb{1}_{\lambda(\boldsymbol{v}_k)=y^s} \,, \tag{3}$$

we indicate this quantity as *Class Average Distance*. We also quantify the fraction of patches from $\boldsymbol{x}^t$ that are assigned to the class y^s:

$$L_{y^s}(\boldsymbol{x}^t) = \frac{1}{P_l} \sum_{k=1}^{P_l} \mathbb{1}_{\lambda(\boldsymbol{v}_k)=y^s} \,, \tag{4}$$

that we name *Class Assignment*. Using these metrics, we derive a *normalcy score* based on the inverse entropy of the patch class assignments:

$$H(\boldsymbol{x}^t) = \sum_{y^s=1}^{C} L_{y^s}(\boldsymbol{x}^t) \log L_{y^s}(\boldsymbol{x}^t) \,. \tag{5}$$

This function yields low *normalcy scores* when class assignments are spread across different classes, indicative of high entropy. However, it overlooks patches' embedding distances, which can provide valuable insights into sample normalcy. To address this limitation, we draw inspiration from the weighted entropy [43] formulation and augment the entropy-based normalcy score with class-level aggregated embedding distances (L_{y^s}):

$$H_w(\boldsymbol{x}^t) = \sum_{y^s=1}^{C} D_{y^s}(\boldsymbol{x}^t) L_{y^s}(\boldsymbol{x}^t) \log L_{y^s}(\boldsymbol{x}^t) \,. \tag{6}$$

Incorporating class-level embedding distances improves robustness by resolving the ambiguity in OOD samples whose patch class assignments match only a few support set classes (resulting in low entropy) yet exhibit a large embedded distance.

IV. EXPERIMENTS

In this section, we present a thorough experimental analysis of 3D-SeND on a realistic and challenging scenario in which the support set is composed of clean, synthetic point clouds, while the test set is drawn from a collection of real-world 3D scans affected by acquisition artifacts such as vertex noise, non-uniform density, missing parts, and occlusions. This setting encompasses a combination of covariate and semantic shifts, and the goal is to identify semantic novelty regardless of the domain gap. We show how 3D-SeND outperforms several state-of-the-art competitors, both training-free and traditional training-based ones. 3D-SeND also excels when the support set contains only a very limited amount of samples (few-shot).

We complete the experimental evaluation by considering also the in-domain scenario involving only the semantic shift, when the support set and test data are drawn from the real-world distribution. Finally, we provide an analysis of the role of the different components of the proposed 3D-SeND.

A. EXPERIMENTAL SETUP

1) PRE-TRAINING DETAILS

3D-SeND extracts patches as the internal feature representation at a chosen layer within the network hierarchy to obtain $\boldsymbol{v}_k$. For the EPN backbone, we select the last convolutional block (out of a total of 4 blocks) designated as $l = 4$. This choice yields for each point cloud $P_l = 256$ patches with $C_l = 256$ channels. In the case of the PointNet++ backbone, we opt for the second Set Abstraction layer (out of a total of 2 layers), labeled as $l = 2$. This results in $P_l = 128$ patches and $C_l = 640$ channels. During training, we augment the point clouds with jittering, SO(3) rotation, random rescaling, random translation, and random crop of a small neighborhood of points.

For OpenShape and Uni3D we leverage the pre-trained weights made publicly available by the authors. In the case of OpenShape, we extract patches from the final transformer block ($l = 11$), resulting in a total of $P_l = 512$ patches and $C_l = 256$ channels. Similarly, for the Uni3D model, we utilize the last transformer block ($l = 11$), which yields $P_l = 512$ patches, and $C_l = 1024$.

2) TESTBED DATASET

We run our experiments on the 3DOS benchmark [14] that offers several tracks. We focus mainly on the *Synthetic to Real* one composed by synthetic point clouds from ModelNet40 [44] for the support set, and real-world point clouds from ScanObjectNN [45] for the test set. It features three distinct groups of categories: SR1 (chair, shelf, door, sink, sofa), SR2 (bed, toilet, desk, table, display), and SR3 (bag, bin, box, pillow, cabinet). Either of the first two is designed as the known class set, while the other two sets are labeled as unknown. In this way we obtain two experimental sets of different difficulty (easy/hard): we report their separate results as well as the overall average.

We also consider the *Real to Real* track based on the same SR category sets created from ScanObjectNN described above. Specifically, each of them is used as unknown in the test set, while the other two are divided into train and test and used as known classes. This provides three experimental sets of different difficulties (easy/med/hard) and we report the obtained average results.

3) REFERENCE METHODS AND EMBEDDING DISTANCES

In collecting the 3D-SeND competitors to be used as references, we considered the literature on large-scale pre-trained models, as well as more specific OOD detection approaches.

For the former, the main objective is to obtain a rich and reliable representation, reusable for diverse downstream tasks. In such learned embeddings, sample similarity is directly expressed by their feature vector closeness, thus a simple way to probe them for SeND is by defining the normalcy score for a test sample as the inverse of its Euclidean distance to the nearest neighbor within the support set. We indicate this approach as *1NN*. We also evaluate alternative techniques based on a parametric estimate of the per-class feature distributions. *EVM* [46] assumes the embeddings to conform to a Weibull distribution, while *Mahalanobis* [25] assumes a multivariate Gaussian distribution. During testing, the first method calculates the likelihood of the test sample belonging to each known class and selects the maximum likelihood as the measure of normalcy, whereas the second method utilizes the inverse Mahalanobis distance from the nearest class distribution as the indicator of normalcy.

In the case of multi-modal pre-trained models, we opt for the cosine distance rather than the Euclidean one as it better aligns with their pre-training objective. Moreover, as these models involve language, the names of the known categories may serve as class prototypes. Thus, at deployment time it is possible to evaluate the distance of a test sample from them and apply Maximum Concept Matching (*MCM*, [7]).

All the aforementioned approaches exploit pre-trained models without any additional learning stage, exactly as our 3D-SeND. Differently, standard OOD detection models require training or at least fine-tuning on the support set. This is considered an essential step to capture discriminative information on the available classes and then use this knowledge to reject novelty. Despite the evident inefficiency of these post-hoc solutions, we include three of them in our analysis to contextualize 3D-SeND within the broader OOD detection literature. *MSP* [1] exploits the maximum softmax probability as a normalcy score, assuming that unknown samples will be classified with lower confidence. *MLS* [47] proposes to discard the normalization step provided by the softmax application, and uses the maximum logit value directly. *ReAct* [13] improves the known-unknown separation by applying a rectification on the network activations.

4) PERFORMANCE METRICS

As SeND is inherently a binary task, we employ *AUROC* and *FPR95* [1] as evaluation metrics. We refer to the samples belonging to known and unknown classes respectively as *positive* and *negative* The AUROC (higher is better) is the Area Under the Receiver Operating Characteristics curve, which plots the True Positive Rate against the False Positive Rate when varying a threshold applied to the predicted positive scores. As a result, this metric is threshold-independent and can be interpreted as the probability for a nominal sample to have a greater score than an unknown one. The FPR95 (lower is better) is the False Positive Rate computed when the threshold is set at the value that corresponds to a True Positive Rate of 95%. Although AUROC is the metric that better reflects the potential ability of a method to correctly

TABLE 1. Training-free SeND results on the 3DOS Synthetic to Real benchmark [14], where SR1 and SR2 are used as support sets defining two tasks of increasing complexity. The top part of the table presents the performance of different methods that leverage a classification pre-training on point clouds from Objaverse-LVIS when using respectively the EPN (top-left) and PointNet++ (top-right) backbones. For the results in the bottom part of the table, the pre-training was executed on four multi-modal datasets with a contrastive objective by using the OpenShape-PointBERT (bottom-left) and Uni3D-Base (bottom-right) architectures. 3D-SeND shows top results in all the settings.

Pre-training Db.	**Objaverse-LVIS [37]**											
Architecture	**EPN [39]**						**PointNet++ [38]**					
Setting	SR1 (easy)		SR2 (hard)		Avg		SR1 (easy)		SR2 (hard)		Avg	
Method	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
1NN	57.59	95.11	56.94	90.15	57.27	92.63	57.79	89.97	59.96	93.67	58.87	91.82
EVM [46]	71.58	80.00	60.60	91.48	66.09	85.74	60.69	92.23	61.30	87.70	61.00	89.97
Mahalanobis [25]	61.07	92.04	60.18	91.20	60.63	91.62	61.15	92.53	58.20	85.68	59.68	89.11
3D-SeND	71.61	82.87	72.73	80.97	**72.17**	**81.92**	68.55	87.83	63.26	89.68	**65.90**	**88.75**

Pre-training Db.	**Objaverse [15], ShapeNet [40], ABO [41], 3D-Future [42])**											
Architecture	**OpenShape [16]**						**Uni3D [17]**					
Setting	SR1 (easy)		SR2 (hard)		Avg		SR1 (easy)		SR2 (hard)		Avg	
Method	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
1NN	87.08	64.90	53.32	98.04	70.20	81.47	90.79	49.42	81.59	67.46	86.19	58.44
EVM [46]	76.00	88.64	58.91	93.48	67.46	91.06	58.40	93.98	54.70	94.47	56.55	94.23
Mahalanobis [25]	73.80	89.70	65.25	83.25	69.53	86.48	85.56	71.19	74.99	74.11	80.28	72.65
MCM [7]	87.10	66.70	64.40	90.10	75.75	78.40	84.70	69.90	64.40	92.30	74.55	81.10
3D-SeND	83.35	55.72	73.73	71.69	**78.54**	**63.71**	91.85	47.80	80.79	59.90	**86.32**	**53.85**

TABLE 2. Training-free SeND results on the 3DOS Synthetic to Real benchmark [14] with OpenShape. In these experiments, the pre-training dataset is reduced to avoid label overlap with the support set. EVM is discarded as it produces the worst results in the more favorable setting. Notably, 3D-SeND presents the best performance.

Pre-training Db.	**Objaverse [15] excluding Objaverse-LVIS [37]**					
Architecture	**OpenShape [16]**					
Setting	SR1 (easy)		SR2 (hard)		Avg	
Method	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
1NN	83.53	69.36	57.95	90.80	70.74	80.08
Mahalanobis [25]	78.73	77.98	58.71	89.53	68.72	83.76
MCM [7]	82.63	80.28	58.66	94.81	70.65	87.55
3D-SeND	78.56	74.01	69.58	80.97	**74.07**	**77.49**

TABLE 3. Training-based vs Training-free SeND results on the 3DOS Synthetic to Real benchmark [14]. The difference between the two settings is specified by the second column that indicates whether the approach undergoes training on the support set $\mathcal{S}$ (✓), or not (✗). The top part of the table contains results obtained with the EPN architecture and the training when needed is executed for classification with a standard cross-entropy loss. The bottom part of the table shows results obtained with Uni3D backbone pre-trained on four multi-modal datasets, fine-tuned $\mathcal{S}$ (✓) or not (✗) on the support set data.

Pre-train. Db.	train on $\mathcal{S}$	Setting / Method	SR1 (easy) AUROC↑	SR1 (easy) FPR95↓	SR2 (hard) AUROC↑	SR2 (hard) FPR95↓	Avg AUROC↑	Avg FPR95↓
		EPN [39]						
✗	✓	1NN [24]	72.49	92.91	69.78	85.78	71.14	89.35
		MSP [1]	74.00	89.10	69.10	89.8	71.60	89.50
		MLS [47]	72.80	92.80	71.70	79.00	72.30	85.90
		ReAct [13]	76.60	92.50	72.20	76.70	**74.40**	**84.60**
Obj-LVIS [37]	✓	1NN [24]	71.80	89.72	68.06	89.87	69.93	89.80
		MSP [1]	74.30	88.50	71.30	83.40	72.80	85.90
		MLS [47]	72.80	87.70	73.40	79.90	73.10	83.80
		ReAct [13]	73.60	90.40	73.90	76.10	**73.70**	**83.20**
	✗	3D-SeND	71.61	82.87	72.73	80.97	**72.17**	**81.92**
		Uni3D [17]						
Objaverse [15] ShapeNet [40] ABO [41] 3D-Future [42]	✓	1NN [24]	82.99	66.37	72.24	83.30	**80.16**	**74.84**
		MSP [1]	82.00	70.88	75.17	86.36	78.59	78.62
		MLS [47]	82.49	68.51	76.40	82.70	79.45	75.61
		ReAct [13]	82.08	68.55	73.86	97.61	77.97	83.08
	✗	3D-SeND	91.85	47.80	80.79	59.90	**86.32**	**53.85**

detect novelty, FPR95 offers a more concrete idea of the operational performance of an OOD detection method: it provides a guarantee about the safe recognition of known data and gauges the risk of mistakenly accept as known a sample of an unseen object class.

B. TRAINING-FREE RESULTS

In the following, we present and discuss the results obtained by leveraging large-scale pre-trained models without any further learning stage. **The top part of Tab. 1** shows the performance of the first family of models based on single-modal data pre-training and medium-sized architectures. Here 3D-SeND surpasses all competitors in both SR1 and SR2 tracks by a large margin, indicating that 3D-SeND is better able to exploit the internal representation of the models to evaluate object similarity. Moreover, the advantage of EPN on PointNet++ reveals the suitability of a rotation-invariant backbone for the SeND task. **The bottom part of Tab. 1** reports the result of the second family of models based on multi-modal data pre-training and very large architectures. Even in this case, 3D-SeND effectively leverages the robust feature embeddings to outperform competitor methods, including MCM that exploits language at test time. Interestingly, when using Uni3D as pre-training, the advantage of 3D-SeND over 1NN is thin in terms of AUROC, indicating that the learned embedding is particularly expressive and simple distances among samples provide useful information to detect novelty. Still, the enhancement brought by 3D-SeND remains evident in terms of FPR95.

We further extended our analysis to address scenarios involving **two possible constraints: one involving multi-modal data used for pre-training, and another involving the available support set at deployment time.**

In the first case, we consider OpenShape with restricted access to only one dataset for pre-training, rather than four datasets as considered before. In particular, the pre-training is executed on Objaverse, deliberately excluding the LVIS subset to minimize category overlap with the downstream task. Tab. 2 shows the obtained results: by comparing them with those in the bottom-left part of Tab. 1 we notice a general drop in performance indicating the relevance of the chosen pre-training set. Nevertheless, 3D-SeND (AUROC:74.07, FPR95:77.49) confirms its superiority to the other reference methods, even remaining competitive with MCM defined on the ensemble of four datasets (AUROC:75.75, FPR95:78.40).

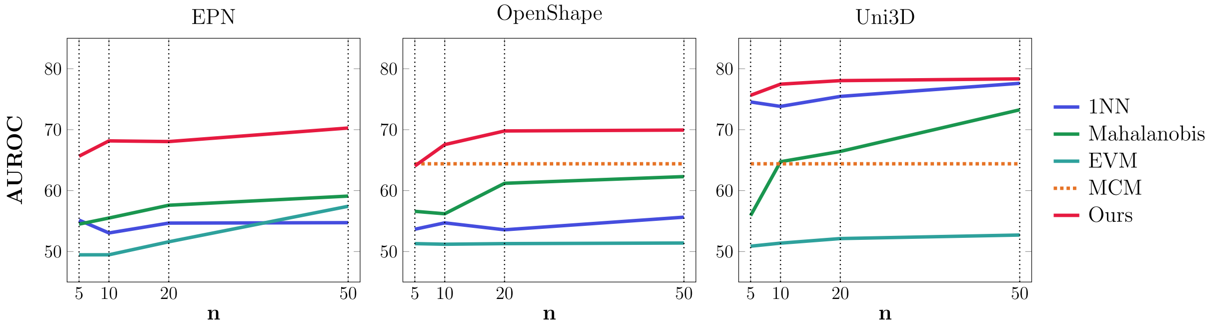


FIGURE 2. Few-shot Training-free SeND results on the Synthetic to Real SR2 (hard) benchmark. *n* represents the number of support set samples for each known class. The plot titles refer to the same architectures (and corresponding pre-training databases) already adopted for the experiments presented in Tab. 1. Here we discard PointNet++ as it was producing the worst results in the more favorable full-shot setting. 3D-SeND consistently surpasses all competing approaches even with low *n* values.

The second constrained scenario aims to mimic real-world applications where it is challenging to collect a sizable and varied support set of 3D models that accurately represent normalcy and can be used for training purposes. To simulate this few-shot condition, we concentrate on the SR2 (hard) benchmark track and create splits with n randomly selected samples for each class in the support set with $n \in \{5, 10, 20, 50\}$. We repeat the sampling process 10 times for each n and report the average results of our experiments in Fig. 2. Across all evaluated pre-trainings and methods, 3D-SeND consistently surpasses all competing approaches even with low n values.

C. TRAINING-BASED VS TRAINING-FREE RESULTS

To position 3D-SeND within the broader landscape of the OOD detection literature we consider a benchmark with conventional methodologies that adopt support set data for either training or fine-tuning. The top part of Tab. 3 shows results obtained with the EPN architecture. When learning on the support set is possible, ReAct [13] delivers the best performance. This holds both if the training on the support set data is done from scratch or through fine-tuning, transferring knowledge from an initial pre-training phase on Objaverse-LVIS. However, this second strategy might backfire: transfer learning causes a detrimental effect on performance, with ReAct showing a slight AUROC decrease. The results of 3D-SeND are close to those of MSP with pre-training and MLS without pre-training, both of which anyway require learning on the support set. Moreover, 3D-SeND gets a lower AUROC with respect to the top training-based ReAct method, but remarkably it shows a better FPR95.

The bottom part of Tab. 3 shows results obtained with Uni3D. In this case, we are dealing with a very large model (88.96M parameters) pre-trained on four multi-modal datasets. As previously explored, leveraging this comprehensive knowledge base enables 3D-SeND to achieve top results even without an application-specific training. Still, to investigate whether a learning phase on the support set could provide further improvement, we devised a tailored fine-tuning strategy. Specifically, we use low-rank adaptation with LoRA [48] which freezes the pre-trained model weights and injects trainable rank decomposition matrices into each layer of the Uni3D transformer architecture. Additionally, we substitute the contrastive training objective with a standard cross-entropy classification objective. The classification head and the fine-tuning training regime replicate those employed for EPN fine-tuning. The obtained results produced with post-hoc OOD detection approaches indicate that fine-tuning the original large-scale model may be detrimental even if performed with a state-of-the-art technique.

TABLE 4. Training-free SeND results on the 3DOS Real to Real benchmark [14]. In this scenario, there is no Domain Shift between the support set and test data. Our 3D-SeND provides the best results with OpenShape while ranking second in terms of FPR95 and third in terms of AUROC with Uni3D. .

Real to Real with Pre-train. Db.: Objaverse [15], ShapeNet [40], ABO [41], 3D-Future [42])				
Architecture	OpenShape [16]		Uni3D [17]	
Setting / Method	Avg		Avg	
	AUROC↑	FPR95↓	AUROC↑	FPR95↓
1NN [24]	70.27	81.12	**90.10**	**40.96**
EVM [46]	71.37	90.43	80.69	85.71
Mahalanobis [25]	64.31	85.21	87.78	54.40
MCM [7]	67.52	90.45	68.43	88.32
3D-SeND	**74.50**	**76.17**	86.36	53.91

D. IN-DOMAIN EXPERIMENTS

To assess the performance of 3D-SeND in an in-domain scenario involving only semantic shift, we benchmark training-free methods on the 3DOS Real to Real benchmark, where both the support set and test data are sourced from the same real-world distribution. The results are presented in Tab. 4. In this simpler setting, 1NN combined with the Uni3D feature encoding is competitive with 3D-SeND which still delivers strong performance with both OpenShape and Uni3D encodings. We can conclude that reasoning on object patches becomes particularly relevant in case of domain shift, while leveraging a global shape representation with 1NN may be effective otherwise.

E. COMPONENT ANALYSIS

To better understand the peculiarities and limits of 3D-SeND we provide a comprehensive analysis of the technical choices made for its three main components: the patch feature

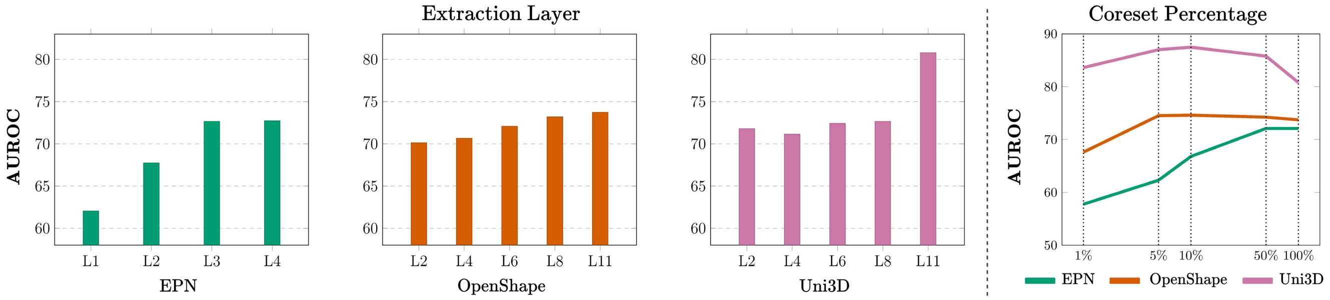


FIGURE 3. Component Analysis of 3D-SeND on the SR2 benchmark. Left: AUROC trends of 3D-SeND for EPN, OpenShape, and Uni3D backbones when varying the patches *Extraction Layer*. Right: Evaluation of 3D-SeND with different portions of the total patches retained for each support set class (Coreset Percentage).

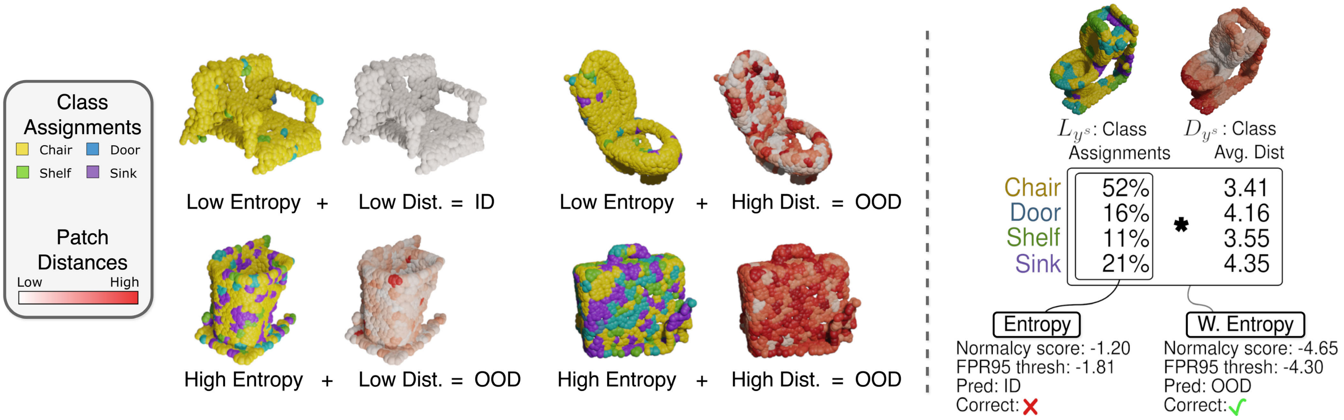


FIGURE 4. ***Left***: Qualitative results of 3D-SeND results with the EPN Backbone. We offer a visual inspection of one ID (known) and three OOD (novel) test samples, showcasing for each instance a pair with the patches' class assignments (left) and their distances to support data patches in the memory banks (right). These are the key components in calculating our Weighted Entropy scoring. ***Right***: We compare Entropy and Weighted Entropy for normalcy scoring of an OOD test sample. After calculating the normalcy scores for both strategies, a binary prediction is made using the FPR95 threshold. Entropy scoring misclassifies the sample as ID, as its normalcy score (−1.20) exceeds the FPR95 threshold (−1.81). In contrast, Weighted Entropy scoring, which integrates Class Average Distances into the normalcy score computation, correctly identifies the sample as OOD, with its normalcy score (−4.65) falling below the FPR95 threshold (−4.30). .

extraction, the memory bank, and the scoring function. For this analysis, we run several evaluations considering the EPN, OpenShape, and Uni3D architectures with the same pre-trainings already presented in the previous sections and focusing on the most challenging Synthetic to Real (SR2) scenario.

1) EXTRACTION LAYER

The layer from which the patch embeddings are extracted impacts the representation's detail and subsequent sample similarity evaluation. The histogram bars on the left side of Fig. 3 demonstrate 3D-SeND performance enhancement as we move from shallow to deeper layers. This trend holds for all the architectures and signifies the growing semantic information inherent in the patch embeddings when progressing toward the network's output.

2) MEMORY BANK SUBSAMPLING

By removing the need for a learning phase on the support set, 3D-SeND already provides a significant efficiency advantage over the post-hoc OOD detection methods. Still, the process of matching patch embeddings of a test sample to elements in the memory bank may be cumbersome and can be optimized by reducing the cardinality of those reference elements extracted from the support set. Specifically, we can minimize their redundancy with a tailored sub-selection strategy such as *greedy coreset* [30], [49], [50]. This procedure is applied separately to each class in the memory bank, reducing the number of per-class patches to a specific fraction indicated as *Coreset Percentage*. The right part of Fig. 3 depicts the variation in AUROC at 1%, 5%, 10%, 50%, and 100% of the overall patch count. When based on OpenShape our method shows almost stable performance. Interestingly, for Uni3D, patch redundancy in the memory bank appears counterproductive, with optimal AUROC observed when only 10% of patches at each class are retained. EPN appears instead more sensitive to the Coreset Percentage with an almost linear performance decrease.

3) SCORING FUNCTIONS

In Tab. 5 we present the results obtained by evaluating 3D-SeND with different scoring functions. We compare the effect of the chosen Weighted Entropy (H_w) with that of the naïve entropy (H) and considering purely distance-based scoring functions such as max and mean respectively defined as $Max = max_{k=1,\dots,P_l}(\delta(\mathbf{v}_k))$ and $Mean = \sum_k^{P_l}(\delta(\mathbf{v}_k))/P_l$. The

TABLE 5. 3D-SeND results on top of EPN, OpenShape and Uni3D feature encoders for the SR2 benchmark when varying the scoring function.

Scoring	EPN		OpenShape		Uni3D	
Functions	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
Max	60.93	92.71	66.79	77.59	65.85	91.53
Mean	68.34	84.90	60.45	85.35	75.03	71.50
H	70.80	86.82	**74.71**	**68.65**	78.65	67.20
H_w	**72.73**	**80.97**	73.73	71.69	**80.79**	**59.90**

results confirm the effectiveness of the proposed weighted-entropy-based solution but also indicate that using the entropy can be slightly preferable with the OpenShape architecture.

4) QUALITITATIVE ANALYSIS

As 3D-SeND operates by leveraging self-discovered relevant object patches and their combination, its predictions can be easily interpreted by visualization. A test sample will be labeled as unknown when its component patches are recognized as belonging to a large variety of classes (high entropy), or possibly a limited number of classes (low entropy), but with a highly uncertain prediction (large distance between the test sample patches and those in the memory bank). Using the EPN backbone, we present in Fig. 4 (left) four test samples and show the distribution of patches' class assignments and distances with color coding - each point in the point cloud inherits the color from its patch. In the right part of Fig. 4, we present a detailed comparison between Entropy and Weighted Entropy normalcy scoring strategies for an OOD test sample. The pure Entropy scoring method fails to identify the test sample as OOD, while the Weighted Entropy scoring method, by incorporating Class Average Distances, successfully detects the sample as OOD.

V. CONCLUSION

We introduced 3D-SeND, a model that effectively detects semantic novelty in 3D data without requiring to be trained or fine-tuned on task-specific support set data. We proposed a strategy to extract generalizable patch features from a pre-trained 3D deep learning architecture and we designed an innovative approach that combines semantic and relative distance information to accurately identify test samples belonging to novel classes.

What sets 3D-SeND apart is its remarkable ability to operate in limited data scenarios and its resilience to domain shifts regarding the Synthetic to Real scenario. 3D-SeND proves to be a flexible solution for real-world applications and provides clear and intuitive visualization to understand its inner functioning. It can be effectively deployed in data-constrained environments, eliminating the need for running custom data collections and training expensive task-specific models.

ACKNOWLEDGMENT

The authors acknowledge the CINECA Award under the ISCRA initiative, for the availability of high performance computing resources and support through the IscrB_MAD-2D3D Project.

REFERENCES

[1] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," in *Proc. ICLR*, 2017, pp. 1–16.

[2] S. Liang, Y. Li, and R. Srikant, "Enhancing the reliability of out-of-distribution image detection in neural networks," 2018, *arXiv:1706.02690*.

[3] W. Liu, X. Wang, J. Owens, and Y. Li, "Energy-based out-of-distribution detection," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 21464–21475.

[4] R. Huang and Y. Li, "MOS: Towards scaling out-of-distribution detection for large semantic space," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 8706–8715.

[5] M. Wortsman, G. Ilharco, J. W. Kim, M. Li, S. Kornblith, R. Roelofs, R. G. Lopes, H. Hajishirzi, A. Farhadi, H. Namkoong, and L. Schmidt, "Robust fine-tuning of zero-shot models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 7949–7961.

[6] A. Kumar, A. Raghunathan, R. M. Jones, T. Ma, and P. Liang, "Fine-tuning can distort pretrained features and underperform out-of-distribution," in *Proc. ICLR*, 2022, pp. 1–20.

[7] Y. Ming, Z. Cai, J. Gu, Y. Sun, W. Li, and Y. Li, "Delving into out-of-distribution detection with vision-language representations," in *Proc. NeurIPS*, 2022, pp. 1–11.

[8] X. Jiang, F. Liu, Z. Fang, H. Chen, T. Liu, F. Zheng, and B. Han, "Negative label guided OOD detection with pretrained vision-language models," 2024, *arXiv:2403.20078*.

[9] A. Miyai, Q. Yu, G. Irie, and K. Aizawa, "Locoop: Few-shot out-of-distribution detection via prompt learning," in *Proc. NeurIPS*, 2023, p. 36.

[10] J. Nie, Y. Zhang, Z. Fang, T. Liu, B. Han, and X. Tian, "Out-of-distribution detection with negative prompts," in *Proc. 12th Int. Conf. Learn. Represent.*, 2024, pp. 1–27.

[11] H. Wang, Y. Li, H. Yao, and X. Li, "CLIPN for zero-shot OOD detection: Teaching CLIP to say no.," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, vol. 15, Oct. 2023, pp. 1802–1812.

[12] R. Huang, A. Geng, and Y. Li, "On the importance of gradients for detecting distributional shifts in the wild," in *Proc. NeurIPS*, 2021, pp. 1–26.

[13] Y. Sun, C. Guo, and Y. Li, "ReAct: Out-of-distribution detection with rectified activations," in *Proc. NeurIPS*, 2021, pp. 144–157.

[14] A. Alliegro, F. C. Borlino, and T. Tommasi, "Towards open set 3D learning: Benchmarking and understanding semantic novelty detection on pointclouds," in *Proc. NeurIPS Datasets Benchmarks Track*, 2022, pp. 21228–21240.

[15] M. Deitke, D. Schwenk, J. Salvador, L. Weihs, O. Michel, E. VanderBilt, L. Schmidt, K. Ehsanit, A. Kembhavi, and A. Farhadi, "Objaverse: A universe of annotated 3D objects," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 1–28.

[16] M. Liu, R. Shi, K. Kuang, Y. Zhu, X. Li, S. Han, H. Cai, F. Porikli, and H. Su, "Openshape: Scaling up 3D shape representation towards open-world understanding," in *Proc. NeurIPS*, 2023, p. 36.

[17] J. Zhou, J. Wang, B. Ma, Y.-S. Liu, T. Huang, and X. Wang, "Uni3D: Exploring unified 3D representation at scale," 2024, *arXiv:2310.06773*.

[18] D. Hendrycks, M. Mazeika, and T. Dietterich, "Deep anomaly detection with outlier exposure," in *Proc. ICLR*, 2019, pp. 1–28.

[19] X. Du, Z. Wang, M. Cai, and Y. Li, "Vos: Learning what you don't know by virtual outlier synthesis," in *Proc. ICLR*, 2021, pp. 1–19.

[20] L. Tao, X. Du, J. Zhu, and Y. Li, "Non-parametric outlier synthesis," 2023, *arXiv:2303.02966*.

[21] X. Du, Y. Sun, J. Zhu, and Y. Li, "Dream the impossible: Outlier imagination with diffusion models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2024, pp. 1–32.

[22] D. Abati, A. Porrello, S. Calderara, and R. Cucchiara, "Latent space autoregression for novelty detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 481–490.

[23] Z. Xiao, Q. Yan, and Y. Amit, "Likelihood regret: An out-of-distribution detection score for variational auto-encoder," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 20685–20696.

[24] Z. Bukhsh and A. Saeed, "On out-of-distribution detection for audio with deep nearest neighbors," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, vol. 33, Jun. 2023, pp. 1–5.

[25] K. Lee, K. Lee, H. Lee, and J. Shin, "A simple unified framework for detecting out-of-distribution samples and adversarial attacks," in *Proc. NeurIPS*, 2018, pp. 1–9.

[26] C. S. Sastry and S. Oore, "Detecting out-of-distribution examples with Gram matrices," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 8491–8501.

[27] F. Cappio Borlino, S. Bucci, and T. Tommasi, "Semantic novelty detection via relational reasoning," in *Proc. ECCV*, 2022, pp. 183–200.

[28] N. Cohen and Y. Hoshen, "Sub-image anomaly detection with deep pyramid correspondences," 2020, *arXiv:2005.02357*.

[29] T. Defard, A. Setkov, A. Loesch, and R. Audigier, "Padim: A patch distribution modeling framework for anomaly detection and localization," in *Proc. ICPR*, 2021, pp. 1–22.

[30] K. Roth, L. Pemula, J. Zepeda, B. Schölkopf, T. Brox, and P. Gehler, "Towards total recall in industrial anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 14298–14308.

[31] M. Masuda, R. Hachiuma, R. Fujii, H. Saito, and Y. Sekikawa, "Toward unsupervised 3D point cloud anomaly detection using variational autoencoder," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2021, pp. 3118–3122.

[32] A. Bhardwaj, S. Pimpale, S. Kumar, and B. Banerjee, "Empowering knowledge distillation via open set recognition for robust 3D point cloud classification," *Pattern Recognit. Lett.*, vol. 151, pp. 172–179, Nov. 2021.

[33] J. Cen, P. Yun, J. Cai, M. Wang, and M. Liu, "Open-set 3D object detection," in *Proc. Int. Conf. 3D Vision*, 2021, pp. 869–878.

[34] K. Wong, S. Wang, M. Ren, M. Liang, and R. Urtasun, "Identifying unknown instances for autonomous driving," in *Proc. Conf. Robot Learn.*, 2019, pp. 384–393.

[35] L. Riz, C. Saltori, E. Ricci, and F. Poiesi, "Novel class discovery for 3D point cloud semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 1–20.

[36] T. Weng, J. Xiao, H. Pan, and H. Jiang, "PartCom: Part composition learning for 3D open-set recognition," *Int. J. Comput. Vis.*, vol. 132, no. 4, pp. 1393–1416, Apr. 2024.

[37] A. Gupta, P. Dollár, and R. Girshick, "LVIS: A dataset for large vocabulary instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5351–5359.

[38] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," in *Proc. NeurIPS*, 2017, pp. 1–27.

[39] H. Chen, S. Liu, W. Chen, H. Li, and R. Hill, "Equivariant point network for 3D point cloud analysis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14509–14518.

[40] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, J. Xiao, L. Yi, and F. Yu, "ShapeNet: An information-rich 3D model repository," 2015, *arXiv:1512.03012*.

[41] J. Collins, S. Goel, K. Deng, A. Luthra, L. Xu, E. Gundogdu, X. Zhang, T. F. Y. Vicente, T. Dideriksen, H. Arora, M. Guillaumin, and J. Malik, "ABO: Dataset and benchmarks for real-world 3D object understanding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 21094–21104.

[42] H. Fu, R. Jia, L. Gao, M. Gong, B. Zhao, S. Maybank, and D. Tao, "3D-FUTURE: 3D furniture shape with TextURE," *Int. J. Comput. Vis.*, vol. 129, no. 12, pp. 3313–3337, Dec. 2021.

[43] S. Guiaşu, "Weighted entropy," *Rep. Math. Phys.*, vol. 2, no. 3, pp. 165–179, 1971.

[44] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao, "3D ShapeNets: A deep representation for volumetric shapes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1912–1920.

[45] M. A. Uy, Q.-H. Pham, B.-S. Hua, T. Nguyen, and S.-K. Yeung, "Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1588–1597.

[46] E. M. Rudd, L. P. Jain, W. J. Scheirer, and T. E. Boult, "The extreme value machine," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 3, pp. 762–768, Mar. 2018.

[47] S. Vaze, K. Han, A. Vedaldi, and A. Zisserman, "Open-set recognition: A good closed-set classifier is all you need," in *Proc. ICLR*, 2022, pp. 1–11.

[48] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in *Proc. ICLR*, 2022, pp. 1–28.

[49] O. Sener and S. Savarese, "Active learning for convolutional neural networks: A core-set approach," in *Proc. ICLR*, 2018, pp. 1–25.

[50] S. Sinha, H. Zhang, A. Goyal, Y. Bengio, H. Larochelle, and A. Odena, "Small-GAN: Speeding up GAN training using core-sets," in *Proc. ICML*, 2020, pp. 1–29.

PAOLO RABINO (Student Member, IEEE) received the B.S. and M.S. degrees in computer engineering from the Polytechnic University of Turin, in 2020 and 2022 respectively. He is currently pursuing the Ph.D. degree with the Visual and Multimodal Applied Learning Laboratory, Turin, working under the supervision of Prof. Tatiana Tommasi. He is also affiliated with the Italian Institute of Technology. His research interests lie at the intersection of robot learning and 3D vision.

ANTONIO ALLIEGRO (Student Member, IEEE) is currently a Postdoctoral Researcher with the Polytechnic University of Turin. His research focuses on 3-D shape understanding and its application to real-world scenarios, including research on reducing the synth-to-real domain gap and open-set 3-D shape recognition. He has published multiple papers presented at prestigious computer vision conferences and journals such as CVPR, IROS, NeurIPS, and RA-L. Additionally, he has contributed as a reviewer at various academic events.

TATIANA TOMMASI (Member, IEEE) received the Ph.D. degree from EPFL Lausanne, in 2013. Subsequently, she undertook a postdoctoral roles in both Belgium and the USA. She holds the position of an Associate Professor with the Control and Computer Engineering Department, Polytechnic University of Turin, and the Director of the ELLIS Unit, Turin. Before her current position, she was an Assistant Professor with Sapienza University in Rome, Italy. Her publication record boasts over 50 papers in top conferences and journals, specializing in machine learning and computer vision. Her expertise lies in the development of theoretically grounded algorithms for automatic learning from images, particularly within the realms of robotics, medical applications, and human–machine interaction. She pioneered the field of transfer learning in computer vision and possesses extensive experience in areas, such as domain adaptation, generalization, multimodal learning, and open-set learning. She is an Associate Editor of IEEE Transactions on Pattern Analysis and Machine Intelligence and IEEE Transactions on Emerging Topics in Computing.

Open Access funding provided by 'Politecnico di Torino' within the CRUI CARE Agreement