Open-Vocabulary Affordance Detection using Knowledge Distillation and Text-Point Correlation

Tuan Van Vo[1], Minh Nhat Vu[2], Baoru Huang[3], Toan Nguyen[1], Ngan Le[4], Thieu Vo[5], Anh Nguyen[6]

Abstract— **Affordance detection presents intricate challenges and has a wide range of robotic applications. Previous works have faced limitations such as the complexities of 3D object shapes, the wide range of potential affordances on real-world objects, and the lack of open-vocabulary support for affordance understanding. In this paper, we introduce a new open-vocabulary affordance detection method in 3D point clouds, leveraging knowledge distillation and text-point correlation. Our approach employs pre-trained 3D models through knowledge distillation to enhance feature extraction and semantic understanding in 3D point clouds. We further introduce a new text-point correlation method to learn the semantic links between point cloud features and open-vocabulary labels. The intensive experiments show that our approach outperforms previous works and adapts to new affordance labels and unseen objects. Notably, our method achieves the improvement of 7.96% mIOU score compared to the baselines. Furthermore, it offers real-time inference which is well-suitable for robotic manipulation applications.**

I. INTRODUCTION

Intelligent robotic systems capable of interacting with objects and comprehending their affordances are of paramount importance across a wide array of real-world applications [1]. These robotic applications encompass a diverse range of tasks, including object recognition [2], [3], action anticipation [4], [5], agent's activity recognition [6], [7], and object grasping understanding [8]. In these tasks, the concept of affordance plays an important role as it refers to the potential actions or functionalities that an object can offer to its users. While affordance detection has received significant research interest in robotics, detecting object affordances poses significant challenges due to the inherent complexity and diverse shapes and functionalities of objects [9].

Classical affordance detection techniques have predominantly relied on traditional machine learning methods applied to images [11], texture-based cues [12], relational affordance models [13] and human-object interactions [14]. Deep learning, particularly Convolutional Neural Networks (CNNs) [15], has also been employed for affordance-related tasks [16]–[23]. However, these methods confront challenges arising from the variability of visual information associated with object affordances despite their shared functionalities. Although leveraging 3D point clouds has gained popularity in robotics for supplying direct 3D object and environmental

[1] FPT Software AI Center, Vietnam `tuanvv7@fpt.com`
[2] Automation & Control Institute, TU Wien, Vienna, Austria
[3] Imperial College London, UK
[4] Department of Computer Science, University of Arkansas, USA
[5] Faculty of Mathematics and Statistics, Ton Duc Thang University, Ho Chi Minh city, Vietnam
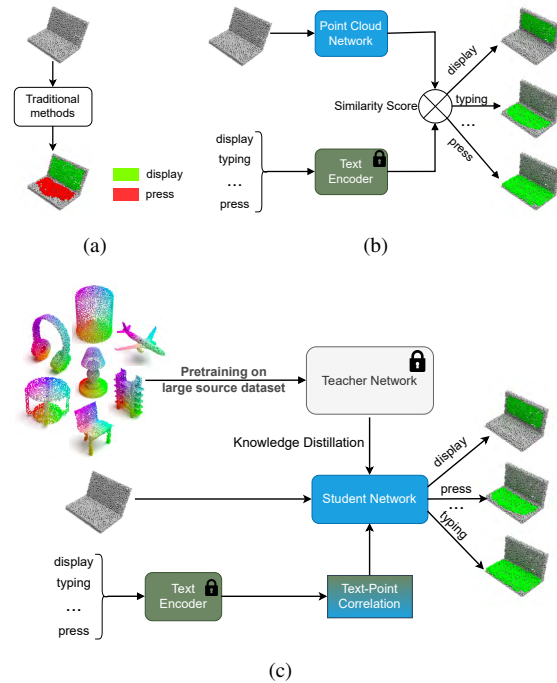[5] Department of Computer Science, University of Liverpool, UK

Fig. 1. The comparison between: (a) traditional affordance detection methods, (b) OpenAD [10], and (c) our proposed method. We leverage a point cloud teacher model and learn the text-point correlation to improve the open-vocabulary affordance detection results.

data, existing research [24]–[27] has encountered limitations imposed by a *fixed label set* tailored to specific tasks, limiting their support for broader or unsupervised inquiries. Furthermore, conventional approaches often encounter difficulties in capturing nuanced associations between localized point cloud regions and their corresponding labeled affordances [21].

To overcome the fixed label set problem, the authors in [10] introduced an *open-vocabulary* approach for 3D point cloud affordance detection that allows unrestricted natural language input, expanding the model's applicability. Despite promising strides, several limitations continue to challenge the effectiveness of existing methods for open-vocabulary affordance detection [10]. First, the inherent intricacies of 3D object shapes and the diverse range of potential affordances pose obstacles in the precise prediction and identification of object interactions [28]. Second, the predicament of entirely novel or unseen affordances in real-world scenarios remains a formidable hurdle, necessitating the reinforcement of detection models' resilience and adaptability [25]. Finally, the intricate interplay between vision and language within the 3D environment mandates a more comprehensive comprehension of object-affordance relationships [29].

To tackle these limitations, we introduce a new open-vocabulary affordance detection in 3D point clouds using knowledge distillation and text-point correlation. Our method utilizes knowledge distillation to leverage 3D models pre-trained on large-scale datasets for the affordance detection task. Knowledge distillation fuses intricate local shape intricacies and dynamic interactions in 3D point clouds, reinforcing feature extraction without class-specific guidance. This enhancement, driven by attention knowledge transfer, enriches semantic comprehension in open-vocabulary affordance detection. Moreover, we introduce the text-point correlation to refine semantic connections between point cloud features and affordance labels. This approach, employing an established attention mechanism [30], centers on relevant point cloud regions to strengthen the text-point relationships. The intensive experiment shows that our method demonstrates substantial improvements, achieving faster running time while improving 7.96 mIOU score over the baselines. Ablation studies and qualitative results further validate the effectiveness of our approach and provide insights for future research directions.

Our main contributions are summarized as follows:

- We propose a new approach to address the challenges of open-vocabulary affordance detection in 3D point clouds using knowledge distillation and text-point correlation.
- We intensively evaluate our method against prior methods and show its effectiveness in real-world robotic applications. Our code will be made available.

II. RELATED WORK

Affordance Detection. Affordance detection is commonly approached as a pixel-wise labeling task, and numerous studies have focused on this area, as evident in [18], [19], [21], [31]–[35]. The authors in [21] developed a method to detect object affordances in real-world scenes by utilizing an object detector and dense conditional random fields. In [19], the authors introduced a two-branch framework that simultaneously identifies multiple objects and their corresponding affordances from RGB images. Chen *et al.* [34] presented a multi-task dense affordance architecture. More recently, Luo *et al.* [35] proposed a cross-view knowledge transfer framework to extract invariant affordances from exocentric observations. Hassan *et al.* [14] predicted high-level affordances by exploring the mutual contexts of humans, objects, and the surrounding environment, while Chen *et al.* [36] learned meaningful affordance indicators for predicting actions in autonomous driving scenarios.

Affordance detection in the context of 3D point cloud data has been also the subject of extensive research [24]–[27]. Kim *et al.* [37] proposed a method that extracts geometric features from point cloud segments and employs logistic regression for affordance classification. Later, Kim and Sukhame [38] introduced a technique that voxelizes point cloud objects and generates an affordance map using interactive manipulation. Similarly, Kokic *et al.* [24] developed a system to model relationships between tasks, objects,

and grasping. Iriondo *et al.* [26] focused on detecting affordances for industrial bin-picking applications. Additionally, Mo *et al.* [27] learned affordance heatmaps from object-object interactions. Yang *et al.* [39] explores the task of linking 3D object affordances to 2D interactions in images. While these studies have made substantial contributions to affordance detection, the task of open-vocabulary affordance detection remains unexplored by these methods [10].

Open-Vocabulary Affordance Detection. Recently, expansive vision language models have exhibited promising outcomes in robotic tasks [40], [41]. Peng *et al.* [29] harnessed a pre-trained textual encoder from CLIP [42] to fuse 2D and 3D characteristics, aligning them with text embedding to address the challenge of open-vocabulary 3D scene comprehension. While these contributions have indeed propelled the advancements in affordance detection, they have yet to explicitly tackle the intricacies of open-vocabulary affordance detection. The authors in [10] introduced OpenAD, an open-vocabulary affordance detection method to identify a wide array of affordances in 3D point clouds. OpenAD effectively learns both textual and point-based affordance features, capitalizing on the semantic relationships among different affordances. However, a limitation of OpenAD lies in its generalization capability, particularly concerning out-of-domain 3D objects and novel affordances.

Knowledge Distillation. Knowledge distillation entails the transfer of information from one network to another [43]. Recently, there has been a shift towards cross-domain knowledge distillation [44]–[47], wherein knowledge is conveyed from a data-rich domain to one with limited diversity. For instance, Li *et al.* [45] effectively employed cross-domain and cross-modal knowledge distillation to enhance 3D point cloud semantic segmentation across diverse scenarios. While cross-domain knowledge distillation has been extensively explored, its application to open-vocabulary affordance detection and its associated techniques remain relatively uncharted. Our study focuses on 3D open-vocabulary affordance detection, with the aim of harnessing diverse 3D point cloud models to transfer this knowledge to the open-vocabulary affordance detection task.

Differing from existing approaches that primarily focus on multi-modal student-teacher frameworks [29], [45], [48], our contribution centers on refining knowledge distillation for open-vocabulary affordance detection in 3D point clouds. Our method involves training a lightweight student model using insights from a well-parameterized teacher model. By transferring attention knowledge from teacher to student, our method enhances feature extraction at the point level, amplifying differentiation capabilities independently of class-specific guidance. Additionally, we heighten semantic connections between point cloud affordances and labels through the established attention mechanism [30]. This strategy focuses on relevant point cloud regions to establish connections between point regions and labeled affordances, hence improving the point-text matching process and improving the final affordance detection results.
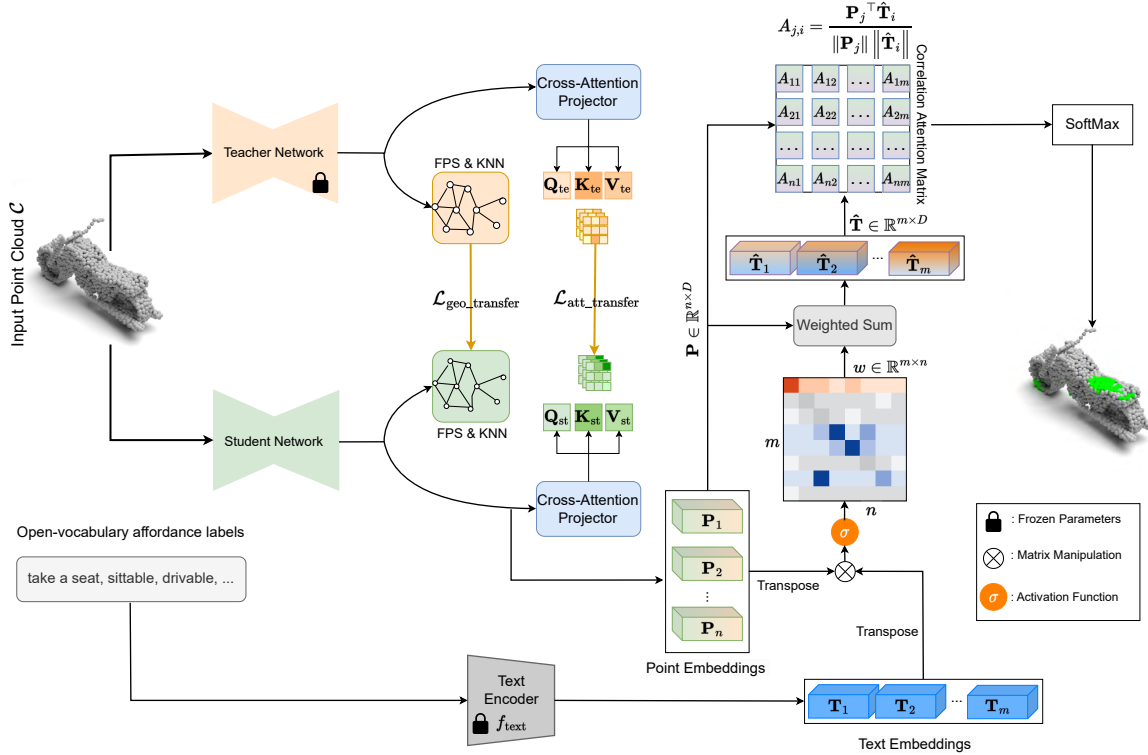
Fig. 2. An overview of our proposed open-vocabulary affordance detection method using knowledge distillation and text-point correlation.

III. OPEN-VOCABULARY AFFORDANCE DETECTION

A. Overview

Following [10], we address open-vocabulary affordance detection where the input cloud $\mathcal{C}$ contains n unordered points $\mathbf{p}_i \in \mathbb{R}^3$, and the corresponding affordance labels are represented by natural language descriptions. The number of possible labels, denoted by m, can be unlimited, allowing adaptation to various affordance labels, even unseen ones during testing. Fig. 2 shows an overview of our approach which has two branches: point-point attention with knowledge distillation and text-point correlation learning.

B. Point-Point Attention with Knowledge Distillation

We utilize a teacher model (pre-trained on large datasets [49]) to transfer its knowledge to the student model via a cross-attention distillation mechanism that minimizes the dissimilarity between the student and teacher attention maps. In particular, the point cloud network processes n input points, resulting in an embedding vector for each point, represented as $\mathbf{P}_1, \mathbf{P}_2, ..., \mathbf{P}_n \in \mathbb{R}^D$, for both the student and teacher models. Following [50], to model the geometry of the point cloud, we first use Farthest Point Sampling (FPS) on the input point cloud to uniformly sample r-proportional points (the number is $\lfloor Z = r * n \rfloor$) as anchors $\{\mathcal{C}^a\}_{a=1}^{Z}$. We then calculate the Euclidean distance between each point and the anchors and apply K-Nearest Neighbors (KNN) to sample the nearest K points $\mathcal{C}^{a,k}, k \in \mathcal{N}(a)$ to form local areas reflecting the geometric structures. Based on this, we represent the point-wise relative relationships $\mathcal{R}^a$ within the geometric neighbors, which contain the structured knowledge

for migration and can be formulated as:

$$\mathcal{R}^a = \frac{1}{K} \sum_{k \in \mathcal{N}(a)} \left(\mathbf{p}_n^{a,k} - \mathbf{p}_n^a \right) \oplus \left(\mathbf{P}_n^{a,k} - \mathbf{P}_n^a \right), \quad (1)$$

where $\mathbf{p}_n$ are xyz coordinates of points in the set $\mathcal{C}$ of n input points, $\mathbf{P}_n$ are the embedding vector for every input point $\mathbf{p}_n$, $\oplus$ indicates the concatenation operation. The point-wise feature relations of the teacher and student model can be expressed as $\mathcal{R}_{\text{te}}^a$ and $\mathcal{R}_{\text{st}}^a$ respectively. We transfer the knowledge of the teacher to the student via the MSE loss:

$$\mathcal{L}_{\text{geo\_transfer}} = \frac{1}{Z} \sum_{a=1}^{Z} \| \mathcal{R}_{\text{te}}^a - \mathcal{R}_{\text{st}}^a \|, \quad (2)$$

Subsequently, the Cross-Attention Projector transforms the feature space of both the student and teacher point clouds into the transformer attention space. It is achieved by mapping point features into query, key, and value matrices. The self-attention [51] is used to capture local relationships among objects by first generating query, key, and value embeddings from the feature matrix $\mathbf{P} \in \mathbb{R}^{n \times D}$, where $\mathbf{Q} = \mathbf{P}W_Q$, $\mathbf{K} = \mathbf{P}W_K$, and $\mathbf{V} = \mathbf{P}W_V$. Here, W_Q, W_K, and $W_V \in \mathbb{R}^{d \times d_h}$ are trainable parameters, with d denotes the query size and d_h is the output embedding dimension. We compute matrix representations $\mathbf{Q}_{\text{st}}, \mathbf{K}_{\text{st}}, \mathbf{V}_{\text{st}}$ to model the attention in the student's space, and matrix representations $\mathbf{Q}_{\text{te}}, \mathbf{K}_{\text{te}}, \mathbf{V}_{\text{te}}$ for the teacher's attention space. The self-attention mechanism for the student is computed as:

$$\Omega_{\text{st}} = \text{softmax}\left(\frac{\mathbf{Q}_{\text{st}}(\mathbf{K}_{\text{st}})^{\top}}{\sqrt{d}}\right)\mathbf{V}_{\text{st}}, \quad (3)$$

The calculation of Ω_{te} for the teacher is similar. Hence, we can minimize the distance between the attention maps of the teacher and the student to guide the student network using the following objective function:

$$\mathcal{L}_{\text{att\_transfer}} = \text{MSE}(\Omega_{\text{te}}, \Omega_{\text{st}}), \tag{4}$$

where $\text{MSE}(\cdot)$ is the mean square error function.

C. Text-Point Correlation

In our approach, following previous works [10], we use the text encoder from CLIP [42], which produces m word embeddings $\mathbf{T} = [\mathbf{T}_1, \mathbf{T}_2, ..., \mathbf{T}_m] \in \mathbb{R}^{m \times D}$ for text-affordance labels. To understand the interaction between vision and language, we focus on the correlation in the feature space of each text-affordance. Let $\mathbf{T}_i$ and $\mathbf{P}_j$ refer to the feature representation of the i-th affordance query and j-th points, respectively. We compute the correlation by calculating the $\mathbf{P}_j$'s attention weight with respect to text-affordance $\mathbf{T}_i$ as:

$$w_{i,j} = \sigma(\mathbf{T}_i^\top \mathbf{P}_j), i \in [1, m], j \in [1, n], \tag{5}$$

where σ is the activation function.

The attention features $\hat{\mathbf{T}}_i$ for the affordance text $\mathbf{T}_i$, is defined from the point feature $\mathbf{P}_j$ and a weighted summation of keypoint features, as in the following equation:

$$\hat{\mathbf{T}}_i = \frac{\sum_{j=1}^{n} \sigma(w_{i,j}\mathbf{P}_j)}{\sum_{j=1}^{n} w_{i,j}}, i \in [1, m], j \in [1, n], \tag{6}$$

The overall relevance score for the text-point correlation attention matrix is then computed as follows:

$$A_{j,i} = \frac{\mathbf{P}_j^\top \hat{\mathbf{T}}_i}{\|\mathbf{P}_j\| \|\hat{\mathbf{T}}_i\|}, j \in [1, n], i \in [1, m]. \tag{7}$$

Following [10], the point-wise softmax output of a single point i is then computed in the form:

$$s_{j,i} = \frac{\exp(A_{j,i}/\tau)}{\sum_{k=1}^{m} \exp(A_{k,i}/\tau)}, \tag{8}$$

where τ is a learnable parameter [52]. We aim to maximize the value of the entry $s_{j,i}$ that is the attention correlation of $\mathbf{P}_j$ and the text attention embedding $\hat{\mathbf{T}}_i$ corresponding to the ground-truth label $i = y_i$. This can be accomplished by optimizing the weighted negative log-likelihood loss of the point-wise softmax output over the entire point cloud in the form:

$$\mathcal{L}_{\text{point\_wise}} = -\sum_{j=1}^{n} \omega_{y_i} \log s_{j,y_i}, \tag{9}$$

where ω_{y_i} is the weighting parameter to the imbalance problem of the label classes during the training.

Finally, the overall training objective is the combination of both loss terms $\mathcal{L}_{\text{total}}$:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{point\_wise}} + \lambda_a \mathcal{L}_{\text{att\_transfer}} + \lambda_t \mathcal{L}_{\text{geo\_transfer}} \tag{10}$$

where λ_a, λ_t is hyper-parameter to balance loss.

IV. EXPERIMENTS

A. Experimental Setup

Dataset. We employ the 3D AffordanceNet dataset [25] and its open affordance labels by [10] for our experiments. This dataset is the large-scale dataset for affordance detection using 3D point clouds, containing $22,949$ instances across 23 object categories. Following [10], [25], we evaluate our approach on two tasks: full-shape and partial-view. The partial-view setup is particularly relevant in robotics, as it reflects the limited observation capability of robots, where only a partial view of the object's point cloud is available.

Baselines and Evaluation Metrics. We conduct a comparative analysis of our method with recent approaches in zero-shot learning for affordance detection in 3D point clouds, including ZSLPC [53], TZSLPC [54], 3DGenZ [55], and OpenAD [10]. To evaluate the results, we utilize three metrics commonly used in related studies [10], namely, mIoU (mean IoU over all classes), Acc (overall accuracy over all points), and mAcc (mean accuracy over all classes). During training, we keep the text encoder and pre-trained teacher model fixed, then train the rest end-to-end. Point cloud size is fixed at $n = 2048$ and D at 512 as in [10]. The hyperparameters τ, λ_a and λ_t are set to $\ln(1/0.07)$, 0.9 and 0.7, respectively. Finally, we use Adam optimizer with $\alpha = 10^{-3}$ and $\gamma = 10^{-4}$ for 200 epochs on an NVIDIA A100 40GB and batch size of 16.

TABLE I
ZERO-SHOT OPEN-VOCABULARY DETECTION RESULTS

Task	Method	mIoU	Acc	mAcc	Params	CPU(s)	GPU(s)
Full-shape	TZSLPC [54]	3.86	42.97	10.37	1.7M	0.75	0.13
	3DGenZ [55]	6.46	45.47	18.33	1.79M	0.76	0.14
	ZSLPC [53]	9.97	40.13	18.70	1.96M	0.82	0.16
	OpenAD [10]	14.37	46.31	19.51	1.8M	0.77	0.14
	Ours	**22.33**	**49.72**	**34.29**	**0.58M**	**0.43**	**0.12**
Partial-view	TZSLPC [54]	4.14	42.76	8.49	1.7M	0.75	0.13
	3DGenZ [55]	6.03	45.24	15.86	1.79M	0.76	0.14
	ZSLPC [53]	9.52	40.91	17.16	1.96M	0.82	0.16
	OpenAD [10]	12.50	45.25	17.37	1.8M	0.77	0.14
	Ours	**20.48**	**48.72**	**32.86**	**0.58M**	**0.43**	**0.12**

B. Quantitative Results

The comparison results of evaluation metrics are shown in Table I. As can be seen, our approach achieves superior results on both tasks and all three evaluation metrics. Particularly on the full-shape task, our method outperforms the runner-up model (OpenAD) by a substantial margin of 7.96% in mIoU. Additionally, our method shows significant superiority over the other approaches, surpassing OpenAD by 14.78% in mAcc and by 3.41% in Acc.

In terms of operational efficiency, our method also significantly outperforms other baselines. On CPU, we achieve a 1.5 times speedup. Moreover, the number of parameters during inference is scaled down by 3 times. Importantly, these efficiency gains do not compromise our method's performance superiority compared with other methods.
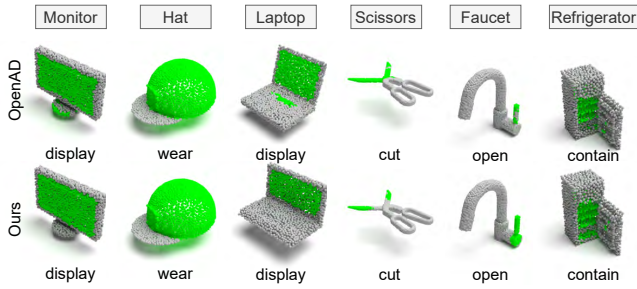
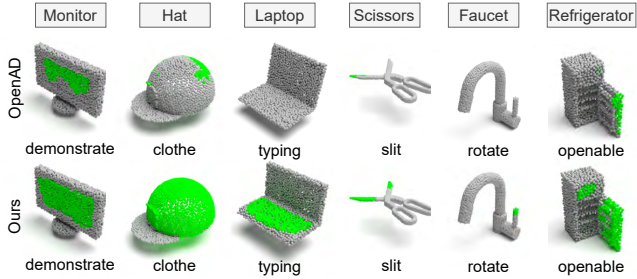Fig. 3. The visualization of our method and OpenAD [10] when detecting seen affordances.

Fig. 4. The visualization of our method and OpenAD [10] when detecting unseen affordances that do not exist in the training set.
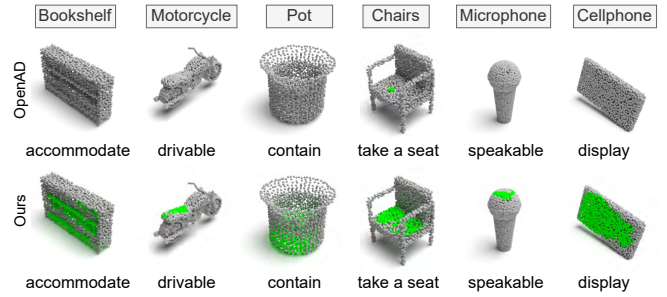

Fig. 5. Results on unseen object categories and unseen affordances.
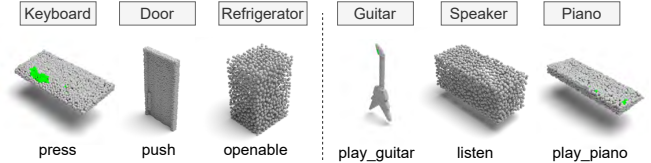

Fig. 6. Failure cases of our method.

are placed in diverse contexts, as illustrated in Fig. 6. For instance, the resemblance between a `keyboard` and a `piano`, or a `refrigerator` and a `speaker`, presents difficulties due to their similar box-like shapes and planes. This makes it challenging to pinpoint the affordance zones.

C. Quantitative Results

Seen Affordances. Fig. 3 presents a visual comparison between our method and OpenAD [10]. While detecting seen affordances is relatively straightforward, this illustration underscores OpenAD's struggles in accurately identifying known affordance areas when they overlap with other regions. For instance, for objects like `monitor`, `laptop`, `faucet`, and `refrigerator`, OpenAD often misidentifies non-affordance areas as affordance zones. In contrast, our approach consistently delivers precise results for affordance regions, avoiding confusion with other areas.

Unseen Affordance. Fig. 4 shows the comparison with unseen affordance inputs. While it is more challenging to detect unseen affordances, this figure illustrates that our method still achieves better results compared to OpenAD [10]. For instance, when considering a `laptop` object, the baseline method struggles to distinguish between the keyboard and screen areas. In contrast, our method adeptly addresses these challenges, exhibiting an enhanced ability to discern subtle differences among affordance regions.

Unseen Objects. We assess the robustness of our method in dealing with new object categories, a key evaluation criterion. Our approach outperforms the baseline model, demonstrating superior adaptability to previously unseen objects, as shown in Fig. 5. This showcases our method's effectiveness and its potential value in scenarios requiring adaptability to novel objects.

Failure Cases. While our method significantly enhances generalizability to unseen affordances and objects, challenges persist with highly semantic affordances and unfamiliar objects featuring intricate geometric structures. In some cases, the semantic information varies when these objects

TABLE II
EFFECTIVENESS OF KNOWLEDGE DISTILLATION (KD) AND
TEXT-POINT CORRELATION (TPC) IN OUR METHOD.

Task	KD	TPC	mIoU	Acc	mAcc
Full-shape			14.37	46.31	19.51
	✓		21.19	48.29	32.32
		✓	21.13	48.33	32.45
	✓	✓	**22.33**	**49.72**	**34.29**
Partial-view			12.50	45.25	17.37
	✓		19.06	47.65	28.13
		✓	19.72	48.02	29.11
	✓	✓	**20.48**	**48.72**	**32.86**

D. Knowledge Distillation and Text-Point Attention Analysis

Effectiveness of Knowledge Distillation. The impact of knowledge distillation is demonstrated in Table II. Additionally, we visually assess the effectiveness of Knowledge Distillation in Fig. 7, where the left side illustrates the student embeddings without Knowledge Distillation (KD), and the right side shows the student representations learned with KD. These results illustrate that Knowledge Distillation directs the model's attention towards interactive regions, facilitating the extraction of interaction contexts.

Effectiveness of Text-Point Correlation. Table II reports the impact of the Text-Point Correlation (TPC) in our method. Additionally, we demonstrate the representation of the input text and learned embeddings in the latent space via t-SNE visualizations [56] in Fig. 8. The results show that without TPC, the decision boundaries for most of the affordance are obscure and difficult to distinguish during training. On the other hand, applying TPC increases both the accuracy and learned features of the network.
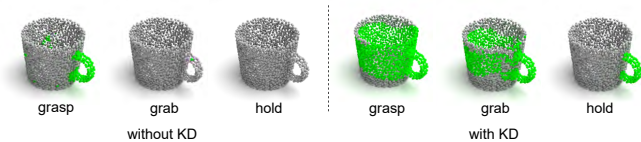
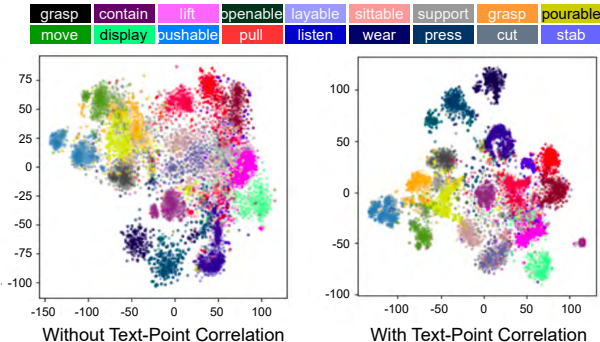Fig. 7. Effectiveness of Knowledge Distillation (KD).

Fig. 8. Effectiveness of Text-Point Correlation though t-SNE visualization of the feature maps at the last stage of student point cloud network for each related affordance.

E. Ablation Study

Pre-trained Teacher Models. Table III shows the influence of the teacher models on our method's performance. Recent state-of-the-art point cloud networks (PointTransformer [57], PointNet++ [58], DGCNN [59], PointMAE [60], and PAConv [61]) are used as the teacher model. They are all trained on the 3D segmentation task with a large source dataset [49]. This table shows that while all recent point cloud networks achieve competitive results, PointNet++ [58] shows the best performance. Therefore, we use PointNet++ in all of our experiments.

Robotic Demonstration. Figure 9 shows our robotic experimental setup. We used five key components: the KUKA LBR iiwa R820 robot, PC1 running Beckhoff TwinCAT software, an Intel RealSense D435i camera, the Robotiq 2F-85 gripper, and PC2 running ROS Noetic 20.04. PC1 controls the robot via EtherCAT, while PC2 operates the gripper and camera via USB within ROS. These PCs communicate via Ethernet. As in [10], we use an object localization method [62] to identify objects and then sample them to 2048 points from the scene point cloud. Our framework supports general input commands and generates an affordance region for useful manipulation tasks. The planning and trajectory optimization in [63], [64] is used to execute the action. Our Demonstration Video includes several demonstrations that illustrate the versatility of open-vocabulary affordance detection of our method.

F. Discussion

While our proposed framework has shown significant improvement in open-vocabulary affordance detection in comparison with recent methods, it is imperative to recognize its limitations and potential for future enhancements. Complex semantic affordances, objects with contextual geometry variations, and challenges in novel scenarios can impact our method's effectiveness as we show in the failure

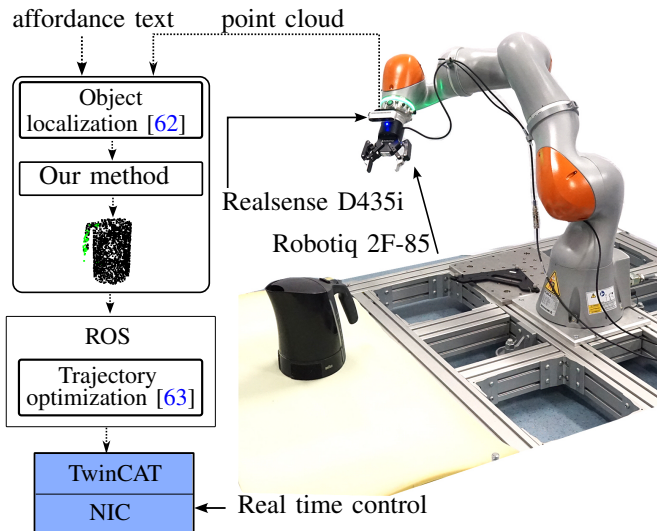| Task	Method	mIoU	Acc	mAcc
Full-shape	Point Transformer [57]	40.32	64.38	65.22
	PointNet++ [58]	**42.47**	**68.60**	**66.55**
	DGCNN [59]	41.83	67.43	64.41
	PointMAE [60]	40.17	63.52	64.28
	PAConv [61]	38.52	58.14	59.48
Partial-view	Point Transformer [57]	40.18	64.29	64.52
	PointNet++ [58]	**41.94**	**68.72**	**66.58**
	DGCNN [59]	41.52	67.01	63.22
	PointMAE [60]	39.18	63.13	62.09
	PAConv [61]	37.09	57.14	60.97

Fig. 9. Overview of the robotic experiment.

cases in Fig. 6. Addressing the gap between the semantic concept of text prompts and the geometry of the point cloud is still a challenging problem, especially when the objects' parts share the same geometry but have different affordances. Moving forward, we intend to explore open-vocabulary affordance detection in cluttered scenes to foster quantitative evaluation and direct applications on real robots. Techniques like augmentation, and cross-modal learning [65] can be useful. Furthermore, combining our open-vocabulary affordance system with long-term manipulation tasks is also an interesting direction [66]. Finally, as recognized by [10], having a new large-scale language-driven affordance dataset with natural point cloud scenes would be more beneficial to real-world robotic applications.

V. CONCLUSIONS

We have presented a new approach for open-vocabulary affordance detection in 3D point clouds. Our proposed method takes advantage of large-scale pre-trained models and text-point correlation to improve the detection results. By integrating attention mechanisms and knowledge transfer, we outperform other baselines in terms of robustness, generalization, and inference time. These enhancements hold substantial promise to apply our proposed method to different robotic applications. Our source code and trained model will be made publicly available.

REFERENCES

[1] W. Liu, A. Daruna, M. Patel, K. Ramachandruni, and S. Chernova, "A survey of semantic reasoning frameworks for robotic systems," *Robotics and Autonomous Systems*, 2023.

[2] S. Thermos, G. T. Papadopoulos, P. Daras, and G. Potamianos, "Deep affordance-grounded sensorimotor object recognition," in *CVPR*, 2017.

[3] Z. Hou, B. Yu, Y. Qiao, X. Peng, and D. Tao, "Affordance transfer learning for human-object interaction detection," in *CVPR*, 2021.

[4] A. Jain, A. R. Zamir, S. Savarese, and A. Saxena, "Structural-rnn: Deep learning on spatio-temporal graphs," in *CVPR*, 2016.

[5] D. Roy and B. Fernando, "Action anticipation using pairwise human-object interactions and transformers," *TIP*, 2021.

[6] T.-H. Vu, C. Olsson, I. Laptev, A. Oliva, and J. Sivic, "Predicting actions from static scenes," in *ECCV*, 2014.

[7] S. Qi, S. Huang, P. Wei, and S.-C. Zhu, "Predicting human activities using stochastic grammar," in *ICCV*, 2017.

[8] A. D. Vuong, M. N. Vu, H. Le, B. Huang, B. Huynh, T. Vo, A. Kugi, and A. Nguyen, "Grasp-anything: Large-scale grasp dataset from foundation models," *arXiv 2309.09818*, 2023.

[9] H. Min, R. Luo, J. Zhu, S. Bi, *et al.*, "Affordance research in developmental robotics: A survey," *IEEE Transactions on Cognitive and Developmental Systems*, 2016.

[10] T. Nguyen, M. N. Vu, A. Vuong, D. Nguyen, T. Vo, N. Le, and A. Nguyen, "Open-vocabulary affordance detection in 3d point clouds," in *IROS*, 2023.

[11] T. Hermans, J. M. Rehg, and A. Bobick, "Affordance prediction via learned object attributes," in *ICRA*, 2011.

[12] H. O. Song, M. Fritz, C. Gu, and T. Darrell, "Visual grasp affordances from appearance-based cues," in *ICCVW*, 2011.

[13] B. Moldovan and L. De Raedt, "Occluded object search by relational affordances," in *ICRA*, 2014.

[14] M. Hassan and A. Dharmaratne, "Attribute based affordance detection from human-object interaction images," in *Image and Video Technology Workshops*, 2016.

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, 2017.

[16] R. Mottaghi, C. Schenck, D. Fox, and A. Farhadi, "See the glass half full: Reasoning about liquid containers, their volume and content," in *ICCV*, 2017.

[17] C.-Y. Chuang, J. Li, A. Torralba, and S. Fidler, "Learning to act properly: Predicting and explaining affordances from images," in *CVPR*, 2018.

[18] A. Nguyen, D. Kanoulas, D. G. Caldwell, and N. G. Tsagarakis, "Detecting object affordances with convolutional neural networks," in *IROS*, 2016.

[19] T.-T. Do, A. Nguyen, and I. Reid, "Affordancenet: An end-to-end deep learning approach for object affordance detection," in *ICRA*, 2018.

[20] H. Luo, W. Zhai, J. Zhang, Y. Cao, and D. Tao, "Leverage interactive affinity for affordance learning," in *CVPR*, 2023.

[21] A. Nguyen, D. Kanoulas, D. G. Caldwell, and N. G. Tsagarakis, "Object-based affordances detection with convolutional neural networks and dense conditional random fields," in *IROS*, 2017.

[22] X. Li, S. Liu, K. Kim, X. Wang, M.-H. Yang, and J. Kautz, "Putting humans in a scene: Learning affordance in 3d indoor environments," in *CVPR*, 2019.

[23] A. Pacheco-Ortega and W. Mayol-Cuervas, "One-shot learning for human affordance detection," in *ECCVW*, 2023.

[24] M. Kokic, J. A. Stork, J. A. Haustein, and D. Kragic, "Affordance detection for task-specific grasping using deep learning," in *International Conference on Humanoid Robotics*, 2017.

[25] S. Deng, X. Xu, C. Wu, K. Chen, and K. Jia, "3d affordancenet: A benchmark for visual object affordance understanding," in *CVPR*, 2021.

[26] A. Iriondo, E. Lazkano, and A. Ansuategi, "Affordance-based grasping point detection using graph convolutional networks for industrial bin-picking applications," *Sensors*, 2021.

[27] K. Mo, Y. Qin, F. Xiang, H. Su, and L. Guibas, "O2o-afford: Annotation-free large-scale object-object affordance learning," in *Conference on Robot Learning*, 2022.

[28] A. Yu, V. Ye, M. Tancik, and A. Kanazawa, "pixelnerf: Neural radiance fields from one or few images," in *CVPR*, 2021.

[29] S. Peng, K. Genova, C. Jiang, A. Tagliasacchi, M. Pollefeys, T. Funkhouser, *et al.*, "Openscene: 3d scene understanding with open vocabularies," in *CVPR*, 2023.

[30] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *CVPR*, 2018.

[31] A. Roy and S. Todorovic, "A multi-scale cnn for affordance segmentation in rgb images," in *ECCV*, 2016.

[32] S. Thermos, P. Daras, and G. Potamianos, "A deep learning approach to object affordance segmentation," in *ICASSP*, 2020.

[33] T. Nguyen, M. N. Vu, B. Huang, T. V. Vo, V. Truong, N. Le, T. Vo, B. Le, and A. Nguyen, "Language-conditioned affordance-pose detection in 3d point clouds," *arXiv*, 2023.

[34] X. Chen, T. Liu, H. Zhao, G. Zhou, and Y.-Q. Zhang, "Cerberus transformer: Joint semantic, affordance and attribute parsing," in *CVPR*, 2022.

[35] H. Luo, W. Zhai, J. Zhang, Y. Cao, and D. Tao, "Learning affordance grounding from exocentric images," in *CVPR*, 2022.

[36] C. Chen, A. Seff, A. Kornhauser, and J. Xiao, "Deepdriving: Learning affordance for direct perception in autonomous driving," in *ICCV*, 2015.

[37] D. I. Kim and G. S. Sukhatme, "Semantic labeling of 3d point clouds with object affordance for robot manipulation," in *ICRA*, 2014.

[38] ——, "Interactive affordance map building for a robotic task," in *IROS*, 2015.

[39] Y. Yang, W. Zhai, H. Luo, Y. Cao, J. Luo, and Z.-J. Zha, "Grounding 3d object affordance from 2d interactions in images," *arXiv preprint arXiv:2303.10437*, 2023.

[40] B. Li, K. Q. Weinberger, S. Belongie, V. Koltun, and R. Ranftl, "Language-driven semantic segmentation," in *ICLR*, 2021.

[41] M. Xu, Z. Zhang, F. Wei, Y. Lin, Y. Cao, H. Hu, and X. Bai, "A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model," in *ECCV*, 2022.

[42] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *ICML*, 2021.

[43] G. Hinton, O. Vinyals, J. Dean, *et al.*, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[44] Y. Yao, Y. Zhang, Z. Yin, J. Luo, W. Ouyang, and X. Huang, "3d point cloud pre-training with knowledge distillation from 2d images," *arXiv preprint arXiv:2212.08974*, 2022.

[45] M. Li, Y. Zhang, Y. Xie, Z. Gao, C. Li, Z. Zhang, and Y. Qu, "Cross-domain and cross-modal knowledge distillation in domain adaptation for 3d semantic segmentation," in *ACM*, 2022.

[46] H. Geng, Z. Li, Y. Geng, J. Chen, H. Dong, and H. Wang, "Partmanip: Learning cross-category generalizable part manipulation policy from point cloud observations," in *CVPR*, 2023.

[47] Q. Zhang, J. Hou, and Y. Qian, "Multi-view vision-to-geometry knowledge transfer for 3d point cloud shape analysis," *arXiv preprint arXiv:2207.03128*, 2022.

[48] R. Huang, X. Pan, H. Zheng, H. Jiang, Z. Xie, S. Song, and G. Huang, "Joint representation learning for text and 3d point cloud," *arXiv preprint arXiv:2301.07584*, 2023.

[49] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, *et al.*, "Shapenet: An information-rich 3d model repository," *arXiv preprint arXiv:1512.03012*, 2015.

[50] Y. Yang, M. Hayat, Z. Jin, C. Ren, and Y. Lei, "Geometry and uncertainty-aware 3d point cloud class-incremental semantic segmentation," in *CVPR*, 2023.

[51] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *NeurIPS*, 2017.

[52] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *CVPR*, 2018.

[53] A. Cheraghian, S. Rahman, and L. Petersson, "Zero-shot learning of 3d point cloud objects," in *MVA*, 2019.

[54] A. Cheraghian, S. Rahman, D. Campbell, and L. Petersson, "Transductive zero-shot learning for 3d point cloud classification," in *WACV*, 2020.

[55] B. Michele, A. Boulch, G. Puy, M. Bucher, and R. Marlet, "Generative zero-shot learning for semantic segmentation of 3d point clouds," in *International Conference on 3D Vision*, 2021.

[56] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, 2008.

[57] H. Zhao, L. Jiang, J. Jia, P. H. Torr, and V. Koltun, "Point transformer," in *ICCV*, 2021.

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

[59] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon, "Dynamic graph cnn for learning on point clouds," *TOG*, 2019.

[60] Y. Pang, W. Wang, F. E. Tay, W. Liu, Y. Tian, and L. Yuan, "Masked autoencoders for point cloud self-supervised learning," in *ECCV*, 2022.

[61] M. Xu, R. Ding, H. Zhao, and X. Qi, "Paconv: Position adaptive convolution with dynamic kernel assembling on point clouds," in *CVPR*, 2021.

[62] Y. Zhou and O. Tuzel, "Voxelnet: End-to-end learning for point cloud based 3d object detection," in *CVPR*, 2018.

[63] M. N. Vu, F. Beck, M. Schwegel, *et al.*, "Machine learning-based framework for optimally solving the analytical inverse kinematics for redundant manipulators," *Mechatronics*, 2023.

[64] F. Beck, M. N. Vu, C. Hartl-Nesic, and A. Kugi, "Singlularity avoidance with application to online trajectory optimization for serial manipulators," *arXiv:2211.02516*, 2022.

[65] J. Zhang, R. Dong, and K. Ma, "Clip-fo3d: Learning free open-world 3d scene representations from 2d dense clip," *arXiv preprint arXiv:2303.04748*, 2023.

[66] T. Zhang, Z. McCarthy, O. Jow, D. Lee, X. Chen, K. Goldberg, and P. Abbeel, "Deep imitation learning for complex manipulation tasks from virtual reality teleoperation," in *ICRA*, 2018.