2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

Language-Conditioned Affordance-Pose Detection in 3D Point Clouds

Toan Nguyen[1,2], Minh Nhat Vu[*,3,4], Baoru Huang[5], Tuan Van Vo[1], Vy Truong[1], Ngan Le[6]
Thieu Vo[7], Bac Le[2,8], Anh Nguyen[9]

***Abstract*— Affordance detection and pose estimation are of great importance in many robotic applications. Their combination helps the robot gain an enhanced manipulation capability, in which the generated pose can facilitate the corresponding affordance task. Previous methods for affodance-pose joint learning are limited to a predefined set of affordances, thus limiting the adaptability of robots in real-world environments. In this paper, we propose a new method for language-conditioned affordance-pose joint learning in 3D point clouds. Given a 3D point cloud object, our method detects the affordance region and generates appropriate 6-DoF poses for any unconstrained affordance label. Our method consists of an open-vocabulary affordance detection branch and a language-guided diffusion model that generates 6-DoF poses based on the affordance text. We also introduce a new high-quality dataset for the task of language-driven affordance-pose joint learning. Intensive experimental results demonstrate that our proposed method works effectively on a wide range of open-vocabulary affordances and outperforms other baselines by a large margin. In addition, we illustrate the usefulness of our method in real-world robotic applications. Our code and dataset are publicly available at https://3DAPNet.github.io.**

I. INTRODUCTION

In robotic research, affordance detection and pose estimation are among the most important and well-concerned problems [1], [2]. Understanding object affordance helps robots decide the inherent possibilities and potential actions within an environment, while pose estimation is considered a prerequisite for robots to interact with and manipulate their surrounding objects effectively. Combining affordance detection and pose estimation holds the potential to help robots gain a more comprehensive understanding of their environment's possibilities and, at the same time, achieve enhanced manipulation abilities [3]. However, prior research has predominantly focused on solving these problems independently, while few works tackled both tasks simultaneously [4]–[6]. This is because the concept of affordance can be arbitrary, and without extra information (e.g., text input), it is challenging to detect the associated pose.

[1] FPT Software AI Center, Vietnam toannt28@fpt.com
[2] Faculty of Information Technology, University of Science, Ho Chi Minh City, Viet Nam
[3] Automation & Control Institute, TU Wien, Vienna, Austria
[4] AIT Austrian Institute of Technology, Vienna, Austria
[5] Imperial College London, UK
[6] Department of Computer Science & Computer Engineering, University of Arkansas, USA
[7] Faculty of Mathematics and Statistics, Ton Duc Thang University, Ho Chi Minh City, Vietnam
[8] Vietnam National University, Ho Chi Minh City, Vietnam
[9] Department of Computer Science, University of Liverpool, UK
[*] Corresponding author minh.vu@ait.ac.at

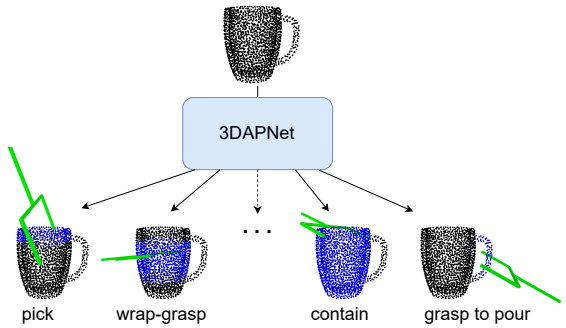

Fig. 1. Our framework allows the simultaneous detection of affordance region and corresponding supporting poses given the input point cloud object and an arbitrary affordance text.

Recently, with the availability of depth cameras, several works have addressed the task of affordance detection in 3D point clouds [5], [7]–[10]. Most of them treated the problem as a supervised task of labeling predefined affordance labels for each point in the point cloud [5], [8]. Lately, the authors in [10] explored the open-vocabulary affordance detection task, a new research direction liberating the constraint of a predefined affordance label set with the utilization of language models [11], [12]. The work in [10] increased the flexibility of the affordance learning process, getting closer to universal affordance detection, however, it does not provide 6-DoF poses that support the corresponding affordance. As a result, the task remains a visionary problem and currently hinders its practical application on real robots. Other works exhibit a combination of affordance detection and pose estimation [5], [13]–[15], yet their methods are still limited to a predefined set of affordance tasks.

In this research, we take a step further by integrating the tasks of *open-vocabulary* affordance detection and *language-driven pose estimation*. Given a 3D point cloud, our goal is to simultaneously detect the unconstrained affordance and generate poses based on the input text query. To realize that objective, we first establish a new dataset for the task of 3D Affordance-Pose joint learning, namely 3DAP dataset. Our dataset is composed of several triplets of a 3D point cloud, an affordance label in the form of the natural text, and a set of multiple 6-DoF poses associated with the affordance. We then present a joint learning framework consisting of a language-driven affordance detection branch and a pose estimation branch which is a guided diffusion model that generates 6-DoF poses conditioned on the given point cloud object and the affordance text. Our choice of the diffusion model is motivated by its recent remarkable results in generating diverse data modalities from multiple conditions [16]–[18], yet its application to pose estimation

remains limitedly explored [19]. Our method is an end-to-end pipeline where via a text prompt, the robot can perform a manipulation task using the affordance and the detected pose. Figure 1 shows the main concept of our work.

Our contributions are summarized as follows:

- We introduce 3DAP, a new dataset of 3D point cloud objects with affordance language labels and affordance-specific 6-DoF poses.
- We propose 3DAPNet, a new method that effectively tackles the task of affordance-pose joint learning.
- We validate our method through intensive experiments and demonstrate the usefulness of 3DAPNet in several real-world robotic manipulation tasks.

II. Related Work

Affordance Detection. Many works tackled the task of affordance detection in RGB images [20]–[24] and 3D point clouds [5], [8]–[10]. In particular, Luo *et al.* [22] leveraged affinity from human-object interaction to detect affordances of non-interactive objects in 2D images. Authors in [7] detected affordance maps on 3D point cloud scenes through interactive manipulation. Also working on point clouds, authors in [9] proposed a framework that detects affordance maps from object-object interaction. Most of these works focus on detecting a set of predefined affordances rather than open-vocabulary setting. Lately, Nguyen *et al.* [10] introduced a framework that allows the detection of arbitrary affordance given in form of a text description. While achieving promising results, the common shortcoming of previous methods is that they solely detect the affordance regions while usually neglecting the corresponding poses that support the detected affordances. This limitation poses a challenge for the robot to effectively execute necessary affordance tasks in real-world manipulation settings.

3D Pose Generation. Given a single object or multiple objects in a cluttered environment, the goal of 3D pose generation algorithms is to find a pose configuration that can support manipulation tasks [2], [3]. Initial works addressed the problem by employing analytical approaches [25], [26], which are practically limited since they assume complete knowledge of object properties like shape, geometry, and material. The rapid development of grasping simulators [27]–[29] in the following years led to the rise of data-driven approaches. Early data-driven methods primarily used whether hand-crafted features [30]–[32] or traditional machine learning algorithms [33]–[35]. Recently witnessed the groundbreaking performances of deep learning methods for pose estimation [5], [36]–[40]. In particular, Lou *et al.* [38] presented a method that can generate collision-free poses in challenging environments, while authors in [37] synthesized poses using a variational autoencoder network. With the recent remarkable results in various generation tasks, diffusion models have also been applied for the task of pose generation [19], [41]. Different from these approaches which are affordance-agnostic, our proposed diffusion model tackles the task of affordance-specific pose generation. Some other earlier works leveraged affordance learning for the problem of task-specific grasping [5], [13], [14]. Nonetheless, their methods are restricted to a predetermined set of affordance tasks. In comparison, our method focuses on the open-vocabulary setting.

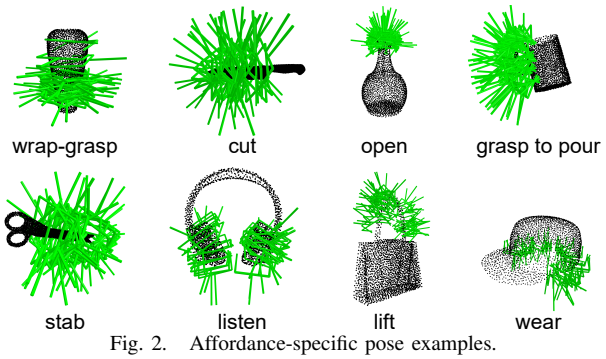

Fig. 2. Affordance-specific pose examples.

Language-Conditioned Robotic Manipulation. With the stunning advancements of large language models [12], [42]–[45], several recent works [46]–[52] have utilized the rich semantics of language for the tasks of robotic manipulation. For instance, Ahn *et al.* [48] proposed a method that constrains the language model to recommend actions that are both plausible to the robot and contextually appropriate. Silva *et al.* [47] proposed to use language to support generalization in multi-task manipulation. The authors in [49] presented a framework that can learn meaningful skill abstractions from language-based expert demonstrations. More recently, Ren *et al.* [51] introduced a language-conditioned and meta-learning approach that learns efficient policies adaptable to novel tools from text descriptions. Different from these works, our method addresses the task of language-conditioned affordance-pose joint learning, where the affordance language simultaneously grounds the affordance region and 6-DoF pose configurations.

III. The 3D Affordance-Pose Dataset

We present the 3D Affordance-Pose dataset (3DAP) as a dataset for affordance-pose joint learning. To construct this dataset, we apply a semi-automatic pipeline in which we first collect affordance-annotated 3D point clouds from 3D AffordanceNet [8], a widely-used and currently the largest dataset for affordance detection in 3D point clouds. Next, we leverage 6-DoF GraspNet [37] method to generate a large number of 6-DoF pose candidates. Afterwards, we manually select the affordance-specific poses for each affordance that the object affords.

Point Cloud Collection. We collect affordance-annotated point clouds from the recent 3D AffordanceNet dataset [8]. Each point cloud represents a single object and is an unordered set of $2,048$ points. Each point is represented by its Euclidean coordinate. The point coordinate of every point cloud are normalized to be in $[0, 1]$. In order to well represent the real-world objects, we scale the point clouds by different scale factors so that the longest side of an axis-aligned bounding box for each object is from $5\,\mathrm{cm}$ to $30\,\mathrm{cm}$. The collected objects are of well-used categories in the daily manipulation tasks, such as `knife`, `bottle`, or `mug`, etc.

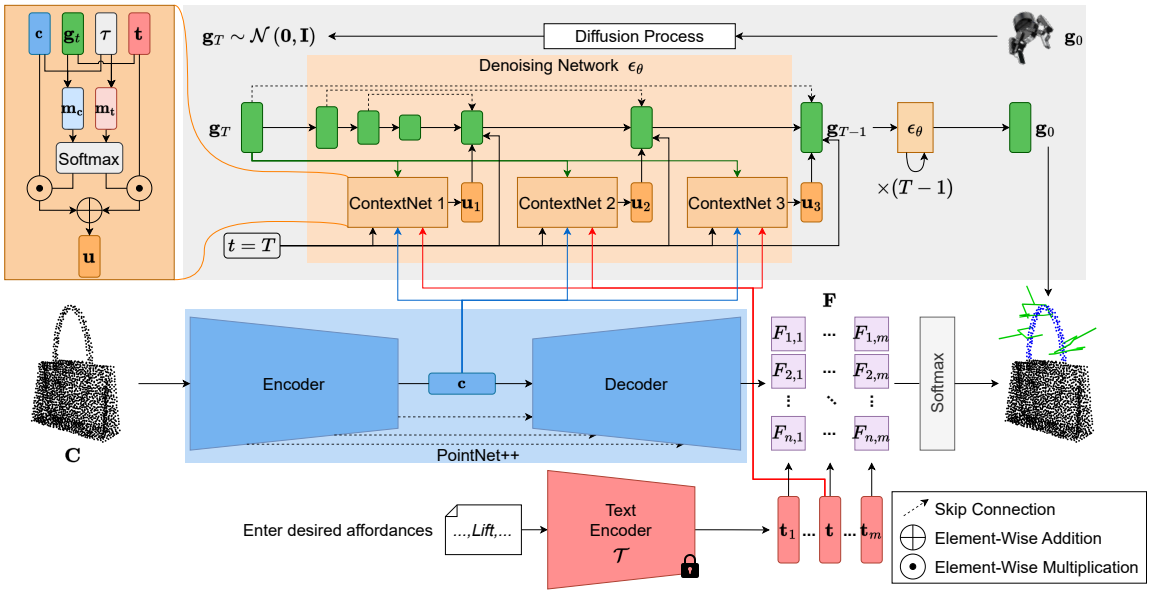

Fig. 3. The overview of our 3DAPNet. Our network includes two branches: one for affordance detection and one for pose generation. The unrestricted affordance label represented in natural language form enables the open-vocabulary setting. In inference, the predicted affordance map and the generated pose are combined to further support the appropriate task.

We express affordance labels as natural language descriptors. This facilitates open-vocabulary affordance detection, so that methods trained on our 3DAP dataset can potentially generalize to unseen affordances.

Poses Collection. We utilize 6-DoF GraspNet [37] to automatically generate a large number of pose candidates for each collected point cloud. In particular, for each object, we pick $1,000$ successful parallel-jaw poses with the highest evaluating scores. Following the Robotiq 2F-85 setting [53], the collected poses have the maximum grip aperture of $85\,\mathrm{mm}$. From the generated poses, we manually select the affordance-specific poses for each object. Given an object and an affordance, we select among $1,000$ candidates ones that best support the affordance task. For example, with a `bottle` and affordance `open`, the poses whose contact points lie on the lid are curated. In total, our dataset contains 28K gripper poses for a wide variety of affordance tasks. Examples of affordance-specific poses in our dataset are presented in Figure 2.

IV. Affordance-Pose Joint Learning

A. Problem Formulation

We present 3DAPNet, a new method for affordance-pose joint learning. Given the captured 3D point cloud of an object and a set of affordance labels conveyed through natural language texts, our objective is to jointly produce both the relevant affordance regions and the appropriate pose configurations that facilitate the affordances. Particularly, 3DAPNet takes as input a point cloud denoted by $\mathbf{C} = \{\mathbf{p}_1, \mathbf{p}_2, ..., \mathbf{p}_n\}$, containing n points in 3D Euclidean space, alongside m arbitrary affordance labels articulated in natural language. The desired output from our framework encompasses an affordance map $\mathbf{A} = \{a_1, a_2, ..., a_n\}$ that assigns an affordance label to each point, and m sets of 6-DoF poses that facilitate corresponding affordances. We consider a 6-DoF pose as configured by $[\mathbf{g}_{\text{qu}}, \mathbf{g}_{\text{tr}}]$, in which $\mathbf{g}_{\mathbf{qu}}$ is a unit-norm quaternion representing the rotation and $\mathbf{g}_{\mathbf{tr}}$ is a translation vector. The overview of our network is illustrated in Figure 3.

B. Open-Vocabulary Affordance Detection

We follow the recent work [10] to detect affordances with open-vocabulary setting. The input point cloud $\mathbf{C}$ is plugged into a PointNet++ model [54] to extract n point-wise feature vectors $\mathbf{P}_1, \mathbf{P}_2, ..., \mathbf{P}_n$. Next, the m affordance language labels are fed into a text encoder $\mathcal{T}$ to extract m text embeddings $\mathbf{t}_1, \mathbf{t}_2, ..., \mathbf{t}_m$. Similar to other works [10], [55], [56], the choices for the text encoder are versatile.

To enable open-vocabulary affordance detection, we compute the semantic relations between the point cloud affordance and its potential labels by correlating the text embeddings and point features using cosine similarity function. Concretely, the $F_{i,j}$ element at the i-th row and j-th column of the correlation matrix $\mathbf{F} \in \mathbb{R}^{n \times m}$, which is the correlation score of the point feature $\mathbf{P}_i$ and the affordance text embedding $\mathbf{t}_j$, is computed as:

$$F_{i,j} = \frac{\mathbf{P}_i^\top \mathbf{t}_j}{\|\mathbf{P}_i\| \, \|\mathbf{t}_j\|} \ . \tag{1}$$

During the training, we optimize the PointNet++ to provide point embeddings that are close to the corresponding label text embeddings. The point-wise softmax output of every point i is then computed as:

$$S_{i,j} = \frac{\exp\left(F_{i,j}/\eta\right)}{\sum_{k=1}^{m} \exp(F_{i,k}/\eta)} \ , \tag{2}$$

where η is a learned parameter. The loss function for affordance detection is computed as the negative log-likelihood of the softmax output over the entire point cloud:

$$\mathcal{L}_{\text{aff}} = -\sum_{i=1}^{n} \log S_{i,a_i}. \tag{3}$$

C. Language-Conditioned Pose Generation

Our key contribution is a new guided diffusion model to address the task of affordance-specific pose generation. Our diffusion model is designed to produce poses that not only based on the point cloud, but also facilitate the affordance task by conditioning on the input text.

Forward Process. Given a pose from the dataset $\mathbf{g}_0 \sim q(\mathbf{g})$, in the forward process, we gradually add to the pose small amounts of Gaussian noise in T steps, creating a sequence of noisy poses $\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_T$. When $T \to \infty$, $\mathbf{g}_T$ is equivalent to $\mathcal{N}(\mathbf{0}, \mathbf{I})$ [57]. The noise step sizes are specified by a predefined variance schedule $\{\beta_t \in (0,1)\}_{t=1}^{T}$. From that, the forward process is formulated as $q(\mathbf{g}_t \mid \mathbf{g}_{t-1}) = \mathcal{N}\left(\sqrt{1-\beta_t}\mathbf{g}_{t-1}, \beta_t \mathbf{I}\right)$. The noisy sample at any arbitrary time step t can be obtained in a closed form of:

$$\mathbf{g}_t = \sqrt{\bar{\alpha}_t}\mathbf{g}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, \tag{4}$$

where $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_t$ with $\alpha_t = 1 - \beta_t$ and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

Reverse Process. The reverse process allows us to generate a pose from the Gaussian noise by gradually denoising through T steps via the reverse probability $q(\mathbf{g}_{t-1} \mid \mathbf{g}_t, \mathbf{c}, \mathbf{t})$. In this probability, $\mathbf{c}$ is the point cloud feature produced by the PointNet++ encoder and $\mathbf{t}$ is the text embedding of the affordance of interest. $\mathbf{c}$ and $\mathbf{t}$ represent two guidances that our model needs to condition on, i.e., the point cloud object and the affordance text. As $q(\mathbf{g}_{t-1} \mid \mathbf{g}_t, \mathbf{c}, \mathbf{t})$ is intractable [57], we approximate it with a neural network. More particularly, we approximate the noise $\boldsymbol{\epsilon}$ at every timestep t by a denoising network $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{g}_t, \mathbf{c}, \mathbf{t}, t)$. $\boldsymbol{\epsilon}$ is updated to minimize the difference between the real and approximated noises. The loss function for pose generation is therefore computed as:

$$\mathcal{L}_{\text{pose}} = \mathbb{E}_{\boldsymbol{\epsilon}, \mathbf{g}_0, \mathbf{c}, \mathbf{t}, t}\left[\left\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{g}_t, \mathbf{c}, \mathbf{t}, t)\right\|^2\right]. \tag{5}$$

Following other works [58], to balance between the quality and the diversity of the generated poses, we randomly drop the conditions $\mathbf{c}$ and $\mathbf{t}$ to train unconditionally with a probability p_{uncond}. Our design allows conditional training and unconditional training to use a single network.

Subsequently, we detail the design of our denoising network $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}$. Kindly refer to Figure 3 for an illustrative demonstration. In particular, following other works of diffusion models [57], we employ a downscale-upscale U-Net architecture [59] for the network. The noisy pose $\mathbf{g}_t$ at timestep t is first plugged into three consecutive downscaling MLPs, and then, three other consecutive MLPs are used in the upscaling phase. To form the input to each upscaling MLP, we combine the output of the preceding one, the feature from skip connection, the time embedding τ computed from the timestep t, and the unified context $\mathbf{u}$ combining the point cloud feature $\mathbf{c}$ and the text feature $\mathbf{t}$. The unified context $\mathbf{u}$ is obtained via a ContextNet module. In this ContextNet, we first compute the point cloud influence mask $\mathbf{m_c}$ and text influence mask $\mathbf{m_t}$ using two MLPs and a softmax layer. The influence masks are at the same size as the two features. The unified context $\mathbf{u}$ is then computed as:

$$\mathbf{u} = \mathbf{c} \odot \mathbf{m_c} + \mathbf{t} \odot \mathbf{m_t}, \tag{6}$$

where $\odot$ represents the element-wise multiplication.

Pose Sampling. When finishing the model training, we can sample poses from Gaussian noise by applying the reverse process from timestep T to 0 using the formulation:

$$\mathbf{g}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{g}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\bar{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}}(\mathbf{g}_t, \mathbf{c}, \mathbf{t}, t)\right) + \sqrt{\beta_t}\mathbf{z}, \tag{7}$$

where $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $t > 1$, else $\mathbf{z} = \mathbf{0}$, and $\bar{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}}(\mathbf{g}_t, \mathbf{c}, \mathbf{t}, t)$ is calculated as:

$$\bar{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}}(\mathbf{g}_t, \mathbf{c}, \mathbf{t}, t) = (w+1)\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{g}_t, \mathbf{c}, \mathbf{t}, t) - w\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{g}_t, t), \tag{8}$$

where w is a guidance scale hyperparameter and $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{g}_t, t)$ is the predicted noise when the conditions are discarded.

D. Training and Inference

We define the overall loss function as $\mathcal{L} = \mathcal{L}_{\text{aff}} + \mathcal{L}_{\text{pose}}$. The number of points in each point cloud is fixed to $n = 2,048$. We utilize the state-of-the-art CLIP text encoder [11] and freeze it during training. For the diffusion model, we set $T = 1,000$, and set the forward diffusion variances to constants increasing linearly from $\beta_1 = 10^{-4}$ to $\beta_T = 0.02$. The unconditional training probability is set to $p_{\text{uncond}} = 0.05$. The whole network is trained end-to-end over 200 epochs on a 24GB-RAM NVIDIA GeForce RTX 3090 Ti with a batch size of 32. The Adam optimizer [60] with the learning rate 10^{-3} and the weight decay 10^{-4} is used. When sampling poses, we set the guidance scale to $w = 0.2$. Our framework takes 180 ms to detect affordances and generate $2,000$ corresponding 6-DoF poses for one point cloud.

V. Experiments

In this section, we conduct several experiments to demonstrate the effectiveness of our proposed 3DAPNet trained on our 3DAP dataset. We start by comparing our method with other baselines. Second, we present 3DAPNet's notable qualitative results. Third, we provide different ablation studies for a more in-depth investigation of our method. Finally, we validate our framework in real robotic experiments.

A. Quantitative Comparisons

Baselines. We compare our 3DAPNet with the following methods: 6D-TGD [5], OpenAD [10], 6D-GraspNet [37]. Note that, OpenAD does not support pose estimation, while 6D-GraspNet does not support affordance detection. We tailor 6D-GraspNet [37] to open-vocabulary setting by incorporating the affordance text branch into the network input. All methods are trained on our dataset with the splitting ratio for training, evaluation, and testing of 7:1:2.

Metrics. For affordance detection, following [10], we evaluate the methods using three metrics, i.e., mIoU (mean IoU over all affordance classes), Acc (overall accuracy of all points), and mAcc (mean accuracy over all affordances). For pose generation, we use two metrics as in recent works: the mean evaluated similarity metric (mESM) [5] and the mean

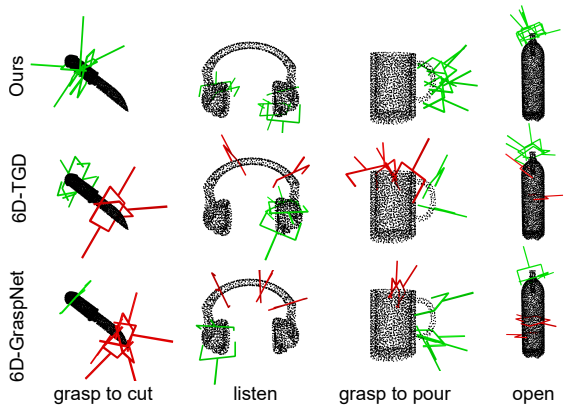

Fig. 4. Qualitative comparison. Green color denotes poses that are related to the input text, while red indicates poses not related to the input text.

TABLE I
BASELINE COMPARISONS

Method	Affordance Detection			Pose Estimation	
	mIoU↑	Acc↑	mAcc↑	mESM↓	mCR↑
OpenAD [10]	55.20	58.89	58.22	–	–
6D-GraspNet [37]	–	–	–	0.373	22.81
6D-TGD [5]	55.38	60.14	57.99	0.219	30.10
3DAPNet (ours)	**56.18**	**61.77**	**59.26**	**0.120**	**44.63**

coverage rate (mCR) [37]. To validate the pose estimation results, we generate 200 poses for each pair of object-affordance during the testing.

Result. Table I illustrates the performance of our 3DAPNet across both tasks, consistently achieving the highest scores across all five metrics when compared to alternative approaches. Specifically, in the affordance detection task, 3DAPNet exhibits improvements of 0.8% on mIoU, 1.63% on Acc, and 1.04% on mAcc relative to its closest competitors. Regarding pose estimation, 3DAPNet is nearly twice as good as the runner-up 6D-TGD on mESM metric (0.120 compared to 0.219). Furthermore, 3DAPNet significantly outperforms competitors in the mCR metric, with a score of 44.63% compared to the second-best 6D-TGD's 30.10%.

B. Qualitative Results

Qualitative Comparison. We present qualitative results to compare our 3DAPNet with other methods in pose generation capability. Particularly, we select poses generated by our method, 6D-GraspNet [37], and 6D-TGD [5]. The example poses are shown in Figure 4. We observe that our method produces more poses that directly support the affordance tasks, while in contrast, the other two baselines generate a large number of poses that do not facilitate them. This result further highlights the enhanced effectiveness of our approach.

Generalization to Unseen Affordances. We present several examples demonstrating 3DAPNet's ability to generalize to unseen affordances in Figure 5. For affordances in the training set, 3DAPNet yields high-quality results both in affordance detection and pose generation. With the reference of seen affordances, when evaluating on unseen affordances, our method still succeeds in detecting the associated regions and generating the corresponding appropriate poses, though those affordances do not appear in the training set.

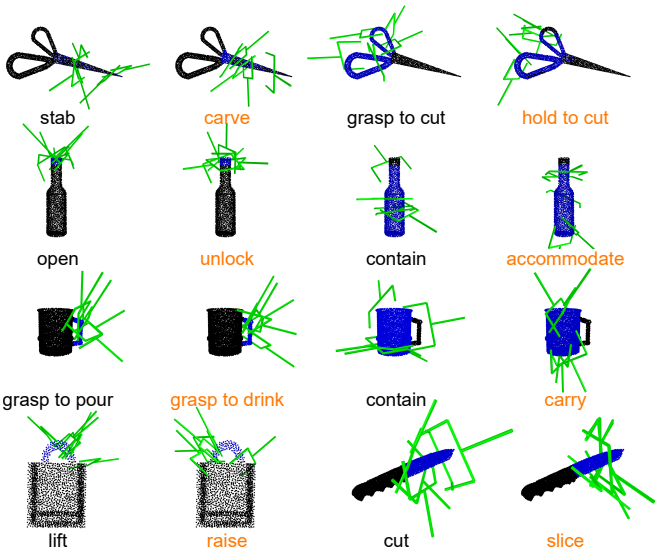

Fig. 5. Qualitative results of 3DAPNet's generalization to unseen affordances. The unseen affordances are shown in orange.

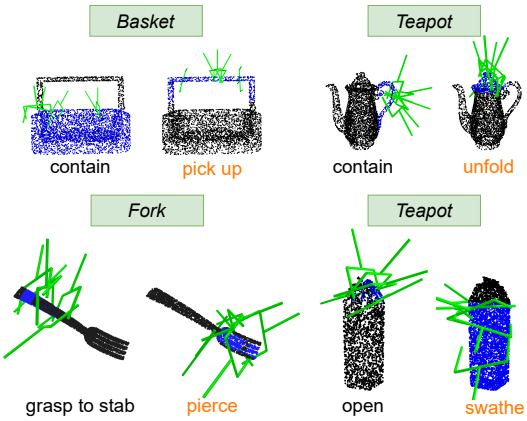

Fig. 6. Qualitative results of 3DAPNet's generalization to unseen object categories. The unseen affordances are shown in orange.

Generalization to Unseen Objects. We extend our assessment to the broader context of unseen object categories. Concretely, we curate new objects from the ShapeNetCore dataset [61] and feed both seen and unseen affordances to the model. The reasonable outcomes shown in Figure 6 reaffirm the generalization of our 3DAPNet.

C. Ablation Study

Single branch vs. jointly learned network. We investigate the performance of each branch in our 3DAPNet on its corresponding task while excluding the other. The results are detailed in Table II. In the case of pose generation only, we retain the usage of the point cloud and text embeddings by keeping the PointNet++ encoder and the CLIP text encoder, while removing the PointNet++ decoder and the correlation head. The results indicate that the combination of the two branches yields the highest performance on both tasks when compared to each branch operating individually. This further validates the efficacy of our design, where the learning processes of the two tasks mutually benefit each other.

Effectiveness of ContextNet. We further validate the effectiveness of our framework design by performing an

TABLE II

SINGLE-BRANCH VS. JOINTLY LEARNED NETWORK

Method	Affordance Detection			Pose Estimation	
	mIoU↑	Acc↑	mAcc↑	mESM↓	mCR↑
Affordance Only	55.65	60.73	58.80	–	–
Pose Only	–	–	–	0.147	41.29
Both	**56.18**	**61.77**	**59.26**	**0.120**	**44.63**

TABLE III

THE EFFECTIVENESS OF CONTEXTNET

ContextNet	Affordance Detection			Pose Estimation	
	mIoU↑	Acc↑	mAcc↑	mESM↓	mCR↑
✗	53.97	60.20	58.49	0.433	12.61
✓	**56.18**	**61.77**	**59.26**	**0.120**	**44.63**

ablation experiment on the ContextNet. Specifically, we report the performances of our framework with and without the ContextNet in Table III. In the case of ContextNet being removed, we combine the point cloud and affordance text conditions naively by adding them. The empirical result shows that our framework with ContextNet completely dominates the other one, with a very large gap on the task of pose generation, which is the main target of our design.

TABLE IV

TEXT ENCODER ANALYSIS

Text Encoder	Affordance Detection			Pose Estimation	
	mIoU↑	Acc↑	mAcc↑	mESM↓	mCR↑
BERT [12]	52.89	59.77	59.25	0.182	30.46
RoBERTa [62]	53.92	58.13	57.01	0.277	30.11
CLIP [11]	**56.18**	**61.77**	**59.26**	**0.120**	**44.63**

Text Encoder. As the text encoder is critical in our framework, we conduct an additional study to investigate the performances of different text encoders. In particular, we use three state-of-the-art text encoder, which are BERT [12], RoBERTa [62], and CLIP [11]. The result is shown in Table IV. We observe that the CLIP encoder significantly outperforms its counterparts on both tasks, especially on pose generation. This result demonstrates the superiority of CLIP in language-vision understanding.

D. Robotic Demonstration

The experiment setup, shown in Figure 7, comprises three main modules, i.e., the inference module, the ROS module, and the real-time controller module. In the inference module, after receiving point cloud data of the environment from the RealSense D435i camera, we utilize the state-of-the-art object localization method [64] to identify the object, then perform point sampling to get $2,048$ points. We then feed this point cloud with a text affordance command into our 3DAPNet. Then, the generated affordance pose from 3DAPNet is sent to ROS for planning and trajectory generation. Analytical inverse kinematics [63] and trajectory optimization [65] are employed to compute optimal trajectories of the robot to reach the pose provided by our network. Note that, using our 3DAPNet, we can have a general input command and are not limited to predefined affordance labels. Several demonstrations can be found in our Demonstration Video.

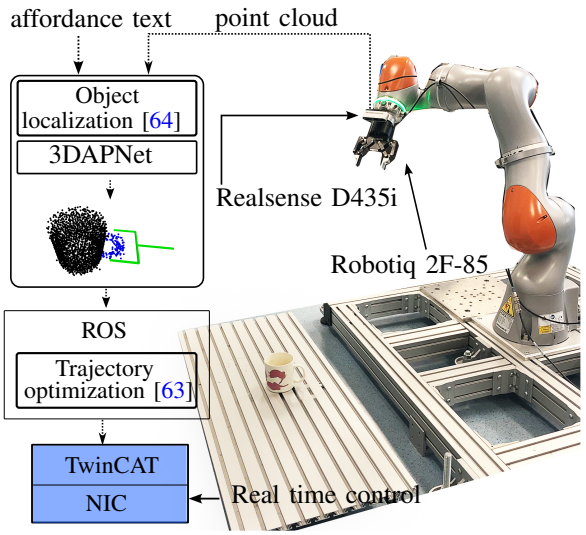

Fig. 7. The overview of the robot experiment setup. More qualitative results can be found in our Demonstration Video.

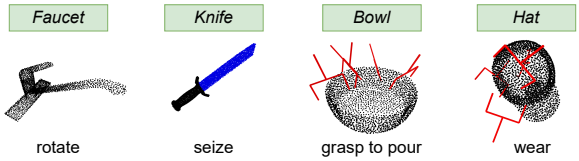

Fig. 8. Some wrong cases of our method.

E. Discussion

Despite promising results, it is important to acknowledge that our method has not fulfilled the perfect ability in universal affordance detection and pose estimation. There are cases where our method shows its limitations, which are presented in Figure 8. Particularly, on the left are two cases of fail and false-positive detection of unseen affordances. In two cases on the right, we show examples where our method generates poses that do not facilitate the corresponding affordances. Furthermore, our method can only detect affordance from single objects due to the dataset limitation. This leads to the fact that it is not straightforward to perform evaluation on real robots with other methods. Therefore, having a large-scale dataset with cluttered point cloud scenes would enable more qualitative comparisons and applications.

VI. CONCLUSIONS

We have tackled the task of open-vocabulary affordance detection and pose estimation in 3D point clouds. In particular, we have presented the 3DAP dataset for affordance-pose joint learning and proposed the 3DAPNet method that can simultaneously detect open-vocabulary affordances and generate affordance-specific 6-DoF poses. Experimental results show that our approach outperforms other methods by a large margin on both tasks. We extensively demonstrated the effectiveness of 3DAPNet in real-world robotic manipulation applications. We hope that the prospective results of our 3DAPNet could encourage more future researchers to further investigate this important yet challenging problem. Our code and trained model will be made publicly available.

REFERENCES

[1] M. Hassanin, S. Khan, and M. Tahtali, "Visual affordance and function understanding: A survey," *CSUR*, 2021.

[2] H. Duan, P. Wang, Y. Huang, G. Xu, W. Wei, and X. Shen, "Robotics dexterous grasping: The methods based on point cloud and deep learning," *Frontiers in Neurorobotics*, 2021.

[3] R. Newbury, M. Gu, L. Chumbley, A. Mousavian, C. Eppner, J. Leitner, J. Bohg, A. Morales, T. Asfour, D. Kragic, *et al.*, "Deep learning approaches to grasp synthesis: A review," *T-RO*, 2023.

[4] W. Liu, A. Daruna, and S. Chernova, "Cage: Context-aware grasping engine," in *ICRA*, 2020.

[5] W. Chen, H. Liang, Z. Chen, F. Sun, and J. Zhang, "Learning 6-dof task-oriented grasp detection via implicit estimation and visual affordance," in *IROS*, 2022.

[6] A. Guo, B. Wen, J. Yuan, J. Tremblay, S. Tyree, J. Smith, and S. Birchfield, "Handal: A dataset of real-world manipulable object categories with pose annotations, affordances, and reconstructions," in *IROS*, 2023.

[7] D. I. Kim and G. S. Sukhatme, "Interactive affordance map building for a robotic task," in *IROS*, 2015.

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

[9] K. Mo, Y. Qin, F. Xiang, H. Su, and L. Guibas, "O2O-Afford: Annotation-free large-scale object-object affordance learning," in *CoRL*, 2022.

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

[11] 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.

[12] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *NAACL*, 2019.

[13] M. Kokic, J. A. Stork, J. A. Haustein, and D. Kragic, "Affordance detection for task-specific grasping using deep learning," in *Humanoids*, 2017.

[14] Y. Zhao, R. Wu, Z. Chen, Y. Zhang, Q. Fan, K. Mo, and H. Dong, "Dualafford: Learning collaborative visual affordance for dual-gripper manipulation," in *ICLR*, 2022.

[15] Z. He, N. Chavan-Dafle, J. Huh, S. Song, and V. Isler, "Pick2place: Task-aware 6dof grasp estimation via object-centric perspective affordance," in *ICRA*, 2023.

[16] S. Luo and W. Hu, "Diffusion probabilistic models for 3d point cloud generation," in *CVPR*, 2021.

[17] X. Li, J. Thickstun, I. Gulrajani, P. S. Liang, and T. B. Hashimoto, "Diffusion-lm improves controllable text generation," *NeurIPS*, 2022.

[18] A. Blattmann, R. Rombach, H. Ling, T. Dockhorn, S. W. Kim, S. Fidler, and K. Kreis, "Align your latents: High-resolution video synthesis with latent diffusion models," in *CVPR*, 2023.

[19] J. Urain, N. Funk, J. Peters, and G. Chalvatzaki, "Se (3)-diffusionfields: Learning smooth cost functions for joint grasp and motion optimization through diffusion," in *ICRA*, 2023.

[20] 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.

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

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

[23] J. Chen, D. Gao, K. Q. Lin, and M. Z. Shou, "Affordance grounding from demonstration video to target image," in *CVPR*, 2023.

[24] L. Mur-Labadia, R. Martinez-Cantin, and J. J. Guerrero, "Bayesian deep learning for affordance segmentation in images," in *ICRA*, 2023.

[25] V.-D. Nguyen, "Constructing force-closure grasps," *IJRR*, 1988.

[26] A. Bicchi and V. Kumar, "Robotic grasping and contact: A review," in *ICRA*, 2000.

[27] A. T. Miller and P. K. Allen, "Graspit! a versatile simulator for robotic grasping," *IEEE Robotics & Automation Magazine*, 2004.

[28] N. Koenig and A. Howard, "Design and use paradigms for gazebo, an open-source multi-robot simulator," in *IROS*, 2004.

[29] E. Todorov, T. Erez, and Y. Tassa, "Mujoco: A physics engine for model-based control," in *IROS*, 2012.

[30] F. Kyota, T. Watabe, S. Saito, and M. Nakajima, "Detection and evaluation of grasping positions for autonomous agents," in *International Conference on Cyberworlds*, 2005.

[31] C. Michel, V. Perdereau, and M. Drouin, "An approach to extract natural grasping axes with a real 3d vision system," in *ISIE*, 2006.

[32] S. El-Khoury, A. Sahbani, and V. Perdereau, "Learning the natural grasping component of an unknown object," in *IROS*, 2007.

[33] A. Saxena, J. Driemeyer, J. Kearns, C. Osondu, and A. Y. Ng, "Learning to grasp novel objects using vision," in *ISER*, 2008.

[34] A. Saxena, J. Driemeyer, and A. Y. Ng, "Robotic grasping of novel objects using vision," *IJRR*, 2008.

[35] Y. Jiang, S. Moseson, and A. Saxena, "Efficient grasping from rgbd images: Learning using a new rectangle representation," in *ICRA*, 2011.

[36] A. Nguyen, D. Kanoulas, D. G. Caldwell, and N. G. Tsagarakis, "Preparatory object reorientation for task-oriented grasping," in *IROS*, 2016.

[37] A. Mousavian, C. Eppner, and D. Fox, "6-dof graspnet: Variational grasp generation for object manipulation," in *ICCV*, 2019.

[38] X. Lou, Y. Yang, and C. Choi, "Collision-aware target-driven object grasping in constrained environments," in *ICRA*, 2021.

[39] B. Sen, A. Agarwal, G. Singh, B. Brojeshwar, S. Sridhar, and M. Krishna, "Scarp: 3d shape completion in arbitrary poses for improved grasping," in *ICRA*, 2023.

[40] G. Schiavi, P. Wulkop, G. Rizzi, L. Ott, R. Siegwart, and J. J. Chung, "Learning agent-aware affordances for closed-loop interaction with articulated objects," in *ICRA*, 2023.

[41] S. Huang, Z. Wang, P. Li, B. Jia, T. Liu, Y. Zhu, W. Liang, and S.-C. Zhu, "Diffusion-based generation, optimization, and planning in 3d scenes," in *CVPR*, 2023.

[42] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.*, "Language models are few-shot learners," *NeurIPS*, 2020.

[43] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, *et al.*, "Palm: Scaling language modeling with pathways," *arXiv preprint arXiv:2204.02311*, 2022.

[44] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[45] D. Driess, F. Xia, M. S. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu, *et al.*, "Palm-e: An embodied multimodal language model," *ICML*, 2023.

[46] S. G. Venkatesh, R. Upadrashta, and B. Amrutur, "Translating natural language instructions to computer programs for robot manipulation," in *IROS*, 2021.

[47] A. Silva, N. Moorman, W. Silva, Z. Zaidi, N. Gopalan, and M. Gombolay, "Lancon-learn: Learning with language to enable generalization in multi-task manipulation," *RA-L*, 2021.

[48] M. Ahn, A. Brohan, N. Brown, Y. Chebotar, O. Cortes, B. David, C. Finn, C. Fu, K. Gopalakrishnan, K. Hausman, *et al.*, "Do as i can, not as i say: Grounding language in robotic affordances," *CoRL*, 2023.

[49] D. Garg, S. Vaidyanath, K. Kim, J. Song, and S. Ermon, "Lisa: Learning interpretable skill abstractions from language," *NeurIPS*, 2022.

[50] O. Mees, J. Borja-Diaz, and W. Burgard, "Grounding language with visual affordances over unstructured data," in *ICRA*, 2023.

[51] A. Z. Ren, B. Govil, T.-Y. Yang, K. R. Narasimhan, and A. Majumdar, "Leveraging language for accelerated learning of tool manipulation," in *CoRL*, 2023.

[52] C. Tang, D. Huang, L. Meng, W. Liu, and H. Zhang, "Task-oriented grasp prediction with visual-language inputs," *IROS*, 2023.

[53] Robotiq. (2018) Robotiq 2f-85. [Online]. Available: https://robotiq.com/products/2f85-140-adaptive-robot-gripper

[54] 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.

[55] M. Xu, Z. Zhang, F. Wei, H. Hu, and X. Bai, "Side adapter network for open-vocabulary semantic segmentation," in *CVPR*, 2023.

[56] H. Luo, J. Bao, Y. Wu, X. He, and T. Li, "Segclip: Patch aggregation with learnable centers for open-vocabulary semantic segmentation," in *ICML*, 2023.

[57] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *NeurIPS*, 2020.

[58] J. Ho and T. Salimans, "Classifier-free diffusion guidance," *NeurIPS workshop*, 2021.
[59] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*, 2015.
[60] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *ICLR*, 2015.
[61] 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.
[62] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.
[63] M. N. Vu, F. Beck, M. Schwegel, C. Hartl-Nesic, A. Nguyen, and A. Kugi, "Machine learning-based framework for optimally solving the analytical inverse kinematics for redundant manipulators," *Mechatronics*, 2023.
[64] Y. Zhou and O. Tuzel, "Voxelnet: End-to-end learning for point cloud based 3d object detection," in *CVPR*, 2018.
[65] F. Beck, M. N. Vu, C. Hartl-Nesic, and A. Kugi, "Singularity avoidance with application to online trajectory optimization for serial manipulators," *IFAC World Congress*, 2023.