

3DSGrasp: 3D Shape-Completion for Robotic Grasp

Seyed S. Mohammadi^{2,3} Nuno F. Duarte¹ Dimitrios Dimou¹ Yiming Wang^{3,4} Matteo Taiana³ Pietro Morerio³
Atabak Dehban¹ Plinio Moreno¹ Alexandre Bernardino¹ Alessio Del Bue³ José Santos-Victor¹

Abstract—Real-world robotic grasping can be done robustly if a complete 3D Point Cloud Data (PCD) of an object is available. However, in practice, PCDs are often incomplete when objects are viewed from few and sparse viewpoints before the grasping action, leading to the generation of wrong or inaccurate grasp poses. We propose a novel grasping strategy, named *3DSGrasp*, that predicts the missing geometry from the partial PCD to produce reliable grasp poses. Our proposed PCD completion network is a Transformer-based encoder-decoder network with an Offset-Attention layer. Our network is inherently invariant to the object pose and point's permutation, which generates PCDs that are geometrically consistent and completed properly. Experiments on a wide range of partial PCD show that *3DSGrasp* outperforms the best state-of-the-art method on PCD completion tasks and largely improves the grasping success rate in real-world scenarios. The code and dataset are available at: <https://github.com/NunoDuarte/3DSGrasp>.

I. INTRODUCTION

Robotic grasping has recently gained increasing attention because of its essential role in many real-world applications, such as domestic and collaborative robotics. The seminal work of Pas et al. [1] uses 3D Point Cloud Data (PCD) to generate grasp poses directly on the available 3D object structure. However, in real practical scenarios, we often have to rely on incomplete geometric information acquired from single or few viewpoints, which leads to a drastic reduction of grasping success rate.

Researchers bypassed this problem by acquiring complete 3D object scans [2] but this requires a feasible camera path around the object, which is time-consuming to obtain and not always feasible. Another strategy is to place additional sensors around the object of interest [3], but this is not cost-effective and it requires careful calibration.

Instead, this paper aims to improve single-view grasping by predicting the missing geometrical structure from a partial PCD. 3D shape completion is an inherently ambiguous problem but recent learning-based approaches have provided

*This work has partially received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 964854; the FCT funding to the ISR/LARSyS Associated Laboratory UID/EEA/50009/2020 and LA/P/0083/2020 N. F. Duarte is supported by FCT-IST fellowship grant PD/BD/135116/2017.

¹Vislab, ISR|Lisboa, Instituto Superior Técnico, Universidade de Lisboa, Portugal. Email:{nferreiraduarte, plinio, alex, jasv}@isr.tecnico.ulisboa.pt

²Department of Marine, Electrical, Electronic and Telecommunications Engineering, University of Genoa, Italy.

³Pattern Analysis & Computer Vision (PAVIS), Istituto Italiano di Tecnologia (IIT), Genoa, Italy. Email:{seyed.mohammadi, yiming.Wang, matteo.taiana, pietro.morerio, alessio.delbue}@iit.it

⁴Deep Visual Learning (DVL), Fondazione Bruno Kessler, Trento, Italy.

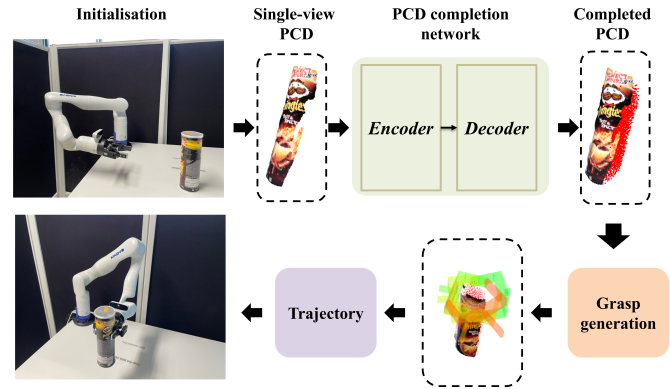


Fig. 1. Overall pipeline of the proposed 3D robotic grasping strategy. We first capture a partial PCD from a single view of the object using a depth sensor located on the Kinova robotic arm. We then feed the single-view PCD to the completion network and produce a completed PCD. Finally, we generate the grasp pose and execute the grasp with a feasible trajectory for the robot.

encouraging results on different classes of objects. Initial shape completion solutions [4], [5] converted the 3D point cloud to a voxel grid with the rendering of additional data that increases processing time and memory requirements. More efficient networks [6], [7] were inspired by the PointNet [8] architecture that directly processes unordered PCDs. However, most of these methods have been evaluated on synthetic, noise-free datasets, far from real-world scenarios. Differently, this work proposes a new model for 3D point completion that can operate in a realistic scenario for robot grasping with arbitrary object classes. Our method adopts a transformer-based network [9], and it proposes a modification of an Offset-Attention layer [10], [11] with the introduction of skip-connections that is able to complete the partial PCD as extracted from just a single depth camera frame. By completing the point cloud, the computation of the grasp poses can leverage the additional information of a full PCD.

Our proposed grasping pipeline is shown in Figure 1. With the calibrated camera equipped on the robotic arm, we first acquire the PCD and segment the background information using PointNet++ [12]. The segmented partial PCD of the object is then normalised, i.e. scaled and centered, and fed to the PCD completion network to predict the missing geometry of the object. We then map back the predicted point cloud in the real-world scene reference in order to merge the predicted missing PCD with the observed partial input. Furthermore, we generate the grasp pose on the top of a

virtually completed point cloud using the method proposed in Grasp Pose Detection in Point Clouds (GPD) [13]. Finally, we utilise MoveIt! [14] to plan the arm trajectory that moves the gripper to the pose estimated by GPD.

We first evaluate our PCD completion method on a PCD completion benchmark dataset [4] that has been generated on the top of YCB dataset [15], by training all the state-of-the-art methods (from scratch) using the same dataset (and split), and outperform the reconstruction error of the best state-of-the-art methods. Then, we test the proposed grasping pipeline in a real scenario using a Kinova arm, our completion network, and GPD. Our method provides accurate completions for successful grasp poses, which enclose the self-occluded parts of the object. Thus, the set of promising grasp hypothesis is larger, which improves the overall success rate score.

To summarise, these are our main contributions:

- We propose a novel partial PCD completion network based on the Offset-Attention encoder-decoder Transformer, that achieves state-of-the-art PCD completion performance when evaluated on the partial version of the YCB dataset proposed in [4].
- We integrate and test our grasping pipeline with a Kinova arm, showing a significant improvement in robotic grasping success rate.
- We present extensive ablation studies on the architecture of our proposed completion network to best justify our design choices.

II. RELATED WORK

We mainly cover related works addressing shape completion with 3D data and robotic grasping.

3D shape completion. In environments where objects are not placed on top of others, such as cupboards and shelves, object shape completion can provide additional grasp poses that augment the selection range.

Given the incomplete partial 3D data as the input, the aim is to predict an approximation of the complete shape. 3D shape completion methods can be categorised into geometric and data-driven approaches [16]. Geometry-based methods [17], [18] assume the presence of shape priors, such as geometric primitives, symmetry and structural repetition [19]. However, the application of these priors may lead to less accurate reconstructions for large-scale datasets and real-world 3D data. Data-driven (i.e. learning-based) approaches rely on deep neural networks that discover the shape completion priors from the data both at local and global point cloud level [20], [16].

In earlier works, the irregular 3D data (i.e., raw point cloud) is converted to a regular data representation (i.e., voxel grid), where 3D CNNs applied on voxelized inputs have been widely adopted for the pure 3D shape completion task [16] and for shape completion for improving grasp estimation [5], [4]. However, the cost of memory usage and computational time for such methods is very large [8].

Instead, PCN [6] directly uses raw PCD for shape completion tasks, and it is based on an encoder-decoder architecture.

The encoder is a PointNet-based backbone network that provides global features. The decoder has two stages, one estimates a coarse point cloud by applying an MLP. After that, FoldingNet [21] is used to generate the detailed and completed point cloud. Following PCN, a range of learning-based methods for pure 3D shape completion tasks from PCD were proposed [7], [22], [23], [24] to improve the resolution and robustness of the reconstructed PCD, while others [25], [26], [27], were proposed for improving the performance of grasp success rate by directly processing 3D PCD for completing the shape of the object. PoinTr [24] was the first PCD completion system to adopt the Transformer architecture [9], leading to a significant improvement in performance. Later, [27] introduced a transformer-based network for object completion that consists of an encoder-decoder architecture, where the encoder is a conventional Multi-Head Self-Attention module, and the decoder is based on the AtlasNet [28]. Although the authors improve the reconstruction result of the GRnet [29] network that uses 3D grids to regularize unordered PCD for point cloud completion, they do not compare their results with PoinTr [24], which consistently outperforms GRnet. In addition, the alignment between the partial point cloud and the reconstructed one requires a 6D pose estimation module. In contrast, our method accurately aligns the observed point cloud with the reconstructed one. Additionally, according to our experiments, we improve the reconstruction results compared to the state-of-the-art and provide more promising grasp poses.

Vision based Robotic Grasping. Robotic Grasping aims to find the optimal pose of the robot's end-effector that leads to a successful grasp of an object. In one way, model-based methods consider contact points and exerted forces to select the grasps that provide more stability, but the evaluation is usually in simulation, which suffers from a large reality gap [30]. On the other, data-driven approaches aim to map directly perceptual input such as RGB [31], [32] and RGB-D images [33], [34], [35], to the grasp success. Recent methods take advantage of model-based and data-driven approaches by generating data samples and labels from simulations using domain randomization [33], [34], [35].

Current approaches are able to map 6DoF pose candidates to point clouds [36], [13], [35], [37], [12], addressing successful grasping in cluttered scenarios. From the perceptual point of view, segmentation of the objects is very challenging, so these approaches start by sampling grasp poses, followed by grasp pose score computation and finally a refinement pose procedure. Amongst the 6DoF grasping approaches, we select Grasp Pose Detection in Point Clouds (GPD) [13] to be used in our system, due to the computational efficiency of the grasp sampling and score computation [37]. The main steps of GPD are: (i) heuristics-based grasp candidate sampling and (ii) binary classification of candidates by a CNN. A detailed description of GPD is in Section III-C.

III. APPROACH

We assembled the setup as a Kinova robotic arm equipped with a RealSense depth camera and Robotiq gripper; and an

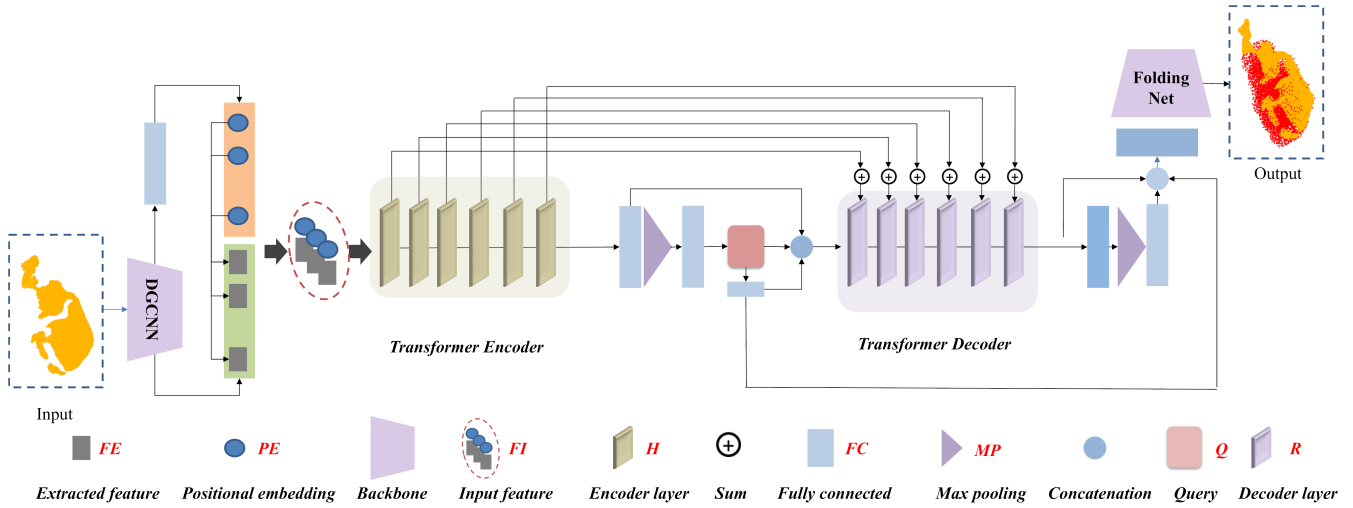


Fig. 2. Architecture of our point cloud completion network. Given the partial PCD as the input, we first apply *FPS* to the subset of the points representing the center point *CR* of each local region *LR*. Then we use *KNN* to gather the points around each *CR* and send them to *DGCNN* to extract embedding feature *FE*. We then send the *CR* to a *FC* layer to learn the Positional Embedding *PE*. Furthermore, we concatenate *PE* with the corresponding *FE* to be the input of the Transformer. As the output we predict the shape feature for a missing PCD *PM* and feed to the FoldingNet to generate high-resolution PCD, then we merge the input PCD with the predicted output PCD to shape the completed PCD *PC*.

object *O* to be grasped. We then utilise the depth camera to capture a depth image from a single viewpoint of the scene. Furthermore, we convert the depth image to PCD using camera parameters. The reconstructed PCD contains only the visible part of the object from the camera's point of view (i.e. partial 3D scan).

Given a partial 3D scan, containing background information and a colourless partial PCD, we first segment the partial PCD *PP*, $PP = \{PP^i \mid PP^i \in \mathbb{R}^3, i = 1 \dots N\}_{N=2048}$, using the PCD segmentation network presented in [12]. Then, we use our proposed completion network for processing *PP* to predict the missing PCD *PM*, $PM = \{PM^i \mid PM^i \in \mathbb{R}^3, i = 1 \dots M\}_{M=6144}$, representing the missing point cloud of the complete shape. Furthermore, we map back the predicted missing PCD to the real scene and merge it with the partial PCD. Finally, we generate a grasp candidate *GK* on the *completed* PCD using the Grasp Pose Detection (GPD) network [13], which outputs a set of grasp poses $\{G_k\}$, $GK = \{GK_1, GK_2, \dots, GK_V\}_{V=5}$ with their corresponding classification scores *CS*. Lastly, the grasp with the best classification score GK_{BCS} that is considered feasible by MoveIt! [14], is executed on a real robot.

In Section III-A, we introduce a dataset pre-processing step and address the PCD alignment problem for the PCD completion task. Section III-B describes the proposed point cloud completion network in detail and the defined loss functions used for training the network. Finally, Section III-C describes the grasp pose generation and evaluation network.

A. PCD alignment pre-processing

Data normalisation is a primary stage for improving the generalisation of deep models on the learning process [38]. However, standard PCD completion approaches use a data normalisation that is not applicable to grasping problems.

The centroid of each PCD in training is given by the centroid of the completed (full PCD) object either from CAD model [39], [40] or the Ground-truth (GT) PCD [6]. This is not an issue in general, as PCD completion protocols during testing provide the partial shape aligned with the centroid of GT. Differently, in a real testing scenario, the shape centroid can only be computed from partial PCDs and thus being different from the GT one. For this reason, the pre-processing of PCD in training has to take into account that the centroid available is only related to the partial PCD. Otherwise, the completed PCD would be misaligned as shown in the ablation studies in Section IV-A.

In this work, we have proposed a simple but effective technique to solve this problem without using GT information. Given the partial PCD $\{PP\}$, we first calculate the translational offset vector $\{T_p \in \mathbb{R}^3\}$, where $T_p = \frac{1}{N} \sum_{i=1}^N PP^i$. We then calculate the centered PCD $\{PP\}^c$ as; $PP^c = PP - T_p$. Furthermore, we normalise the scale $\{S_p \in \mathbb{R}\}$ as; $S_p = \max_i \|PP^i - T_p\|_2$, where $\|\cdot\|_2$ is norm-2 and the final normalise PCD $\{PP^n\}$ is defined as: $PP^n = PP^c / S_p$.

T_p and S_p are the normalisation parameters calculated from the partial point cloud. However, to avoid the misalignment phenomenon, instead of separately calculating the offset and the scaling for ground-truth point cloud *PGT*, we simply apply the same parameters (i.e., T_p , S_p) achieved by the normalisation of *PP* on P_{gt} such that: $PGT^c = PGT - T_p$ and $PGT^n = PGT^c / S_p$. In this way, we precisely align the GT PCD with the partial one, and we also consider the partial PCD as a reference PCD which is the application for a real-world scenario. After normalising the dataset, we apply *FarthestPointSampling (FPS)* to sample 2048 points for a partial PCD and 8192 points for the GT PCD.

B. Point cloud completion network

This section illustrates in detail how the proposed Transformer completion network predicts the missing geometry of the 3D data. The architecture is inspired from [24], but using an Offset-Attention [11] instead of the usual Self-Attention encoder-decoder block, which was shown to be more suitable to process PCD given its intrinsic invariance to rigid transformation. Moreover, we propose Skip-Connections among the layer of the encoder and decoder for better generalisation of the network. The network is composed of three main blocks: The PCD embedding, the Transformer block consisting of the Offset-Attention encoder-decoder layer, and the block that generates the PCD for the missing part.

1) **Point cloud embedding:** The Transformer architecture requires an ordered sequence of vectors (*e.g.* like words in a sentence). However, PCD is invariant to permutations, (*i.e.*, by changing the point sequence order there should be no difference in the description of the shape of the object). To address this property of PCD, in this work, we follow the pipeline as in [24]. We divide partial PCD PP into the set of Local Regions LR , $LR = \{LR_1, LR_2, \dots, LR_R\}_{R=128}$ by applying FPS [12] and then we represent the centroid as CR , $\{CR^i \mid CR^i \in \mathbb{R}^3, i = 1 \dots B\}_{B=128}$, of each LR . We then apply KNN [41] to find the points around each CR . Furthermore, we feed the points in each LR into the PCD-backbone network [41] (DGCNN) to compute the embedding feature FE , $FE = \{FE_1, FE_2, \dots, FE_B\}_{B=128}$. We also feed each CR to the *fully-connected* (FC) layer to extract the positional embedding PE (*i.e.*, describe the location of each subset of the points in each LR [9]) for each FE . Finally, we concatenate the PE with the corresponding FE to be the input FI , $FI = \{FI_1, FI_2, \dots, FI_J\}_{J=128}$ of the Offset-attention Transformer encoder network.

2) **Transformer architecture:** We propose to use a multi-head Offset-Attention encoder-decoder Transformer layer [11], [10] for PCD completion task since Offset-Attention layers have been shown to be advantageous over the usual self-attention layer on point cloud segmentation and classification. This is especially important in robotic grasping contexts where the relative pose between the object and end-effector is arbitrary. The real-world point cloud completion task must be independent from the initial pose of the object as the camera can see the object from different positions. By using the Offset-Attention layer, we take advantage of its invariance to rigid transformations, resulting in a more robust object completion. Fig. 3 shows the architecture of the Offset-Attention layer, where the offset is calculated by measuring the difference between the input features FI and Self-attention features SA , $SA = \{SA_1, SA_2, \dots, SA_J\}$ by subtracting one from the other $FI_J - SA_J$.

Given a sequence of the input features FI , we formulate the encoder as: $AE = E(FI)$, where E is the encoder and $AE = \{AE_1, AE_2, \dots, AE_w\}_{w=1024}$ is the output feature vector of the encoder. The Offset-Attention in the encoder layer first updates the input features FI . Then, we feed the output of the encoder to the FC layer, followed by a *Max-Pooling* (MP) operation. Moreover, to force the encoder

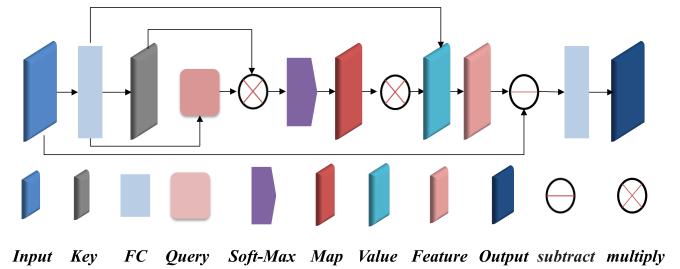


Fig. 3. The Offset-Attention layer measure the difference between the Attention and the input feature.

to learn and generalise better about the global complete shape information, we predict the sparse PCD PS , $PS = \{PS^i \mid PS^i \in \mathbb{R}^3, i = 1 \dots S\}_{S=192}$, where PS is the predicted PCD, representing the complete shape of the object with a lower number of points. We predict PS by passing the generated feature vector AE (*i.e.*, the output of the encoder) to the queries Q layer that contains FC . Then we reshape the output of the FC layer to $S \times 3$ to create PS .

On the other hand, the Offset-Attention decoder layer D shares the exact architecture of the encoder network except for having cross-attention mechanisms [42]. We formulate the decoder architecture as $AD = D(Q, H)$, where $Q = \{Q_1, Q_2, \dots, Q_X\}_{X=192}$ is a set of queries and $AD = \{AD_1, AD_2, \dots, AD_Y\}_{Y=512}$ are the predicted output features representing the feature vector of the missing point cloud.

3) **Point Cloud Generation:** The main objective of our proposed PCD completion network is to predict the missing point cloud representing the unseen part of the object. To do that, we feed the predicted feature vector AD (*i.e.*, the output of the decoder) to the FC layer, followed by Max-pooling and another FC layer. Furthermore, the output of the last FC layer will be concatenated with the predicted sparse point cloud PS reconstructed by query Q , and pass through another FC layer. Then, we utilise FoldingNet [21] FN which is able to output a high-resolution PCD by applying Fold operation on the output predicted feature vector of the missing PCD from the decoder. We can define the point cloud generation process as: $PM = FN(AD) + PS$, (symbol '+' represents set concatenation) where PM is the prediction of the missing parts of the point cloud. The predicted missing point cloud PM will be merged with the partial input point cloud PP to shape the final complete point cloud PC where $PC = \{PC^i \in \mathbb{R}^3, i = 1 \dots Z\}_{Z=8192}$. We also fed the output feature vector of each encoder layer (by an element-wise summation) to the corresponding decoder layer using Skip-Connections (see Table II for our design choice).

4) **Network training:** Training is achieved by summation of the *Chamfer-Distance* (CD) loss between the sparse and completed point cloud and Ground-truth point cloud:

$$L = \mathcal{L}_{cd}(PS, PGT) + \mathcal{L}_{cd}(PC, PGT),$$

where \mathcal{L}_{cd} is the *Chamfer-Distance* loss [6], PGT is the ground-truth PCD, PC is the the completed PCD, and PS

TABLE I
COMPARISON OF $L2$ CD LOSS IN DIFFERENT POINT CLOUD COMPLETION MODELS ON YCB DATASET. WE REPORT THE RESULT OF 13 SEEN CATEGORIES AND 4 UNSEEN CATEGORIES.

Method	Avg	Drill box	Ball	Tomato Soup	Cleanser	Comet Bleach	Sugar box	Mustard	Lemon	Morton Salt	Pringles	Pitcher	Sponge	Cup	Block	Cracker Box	Banana	Stack Blocks
TopNet [43]	2.51	2.18	2.24	1.98	1.84	1.84	1.79	2.01	2.51	1.87	1.70	1.99	2.24	2.84	3.32	3.56	3.90	7.52
FoldingNet [21]	2.28	2.01	2.19	1.81	1.66	1.59	1.48	1.87	2.29	1.63	1.55	1.82	2.18	2.53	3.10	3.14	3.33	7.01
PCN [6]	2.07	1.86	1.95	1.59	1.48	1.41	1.62	1.98	1.42	1.50	1.46	1.69	2.00	2.05	2.89	2.93	2.99	6.53
MSN [7]	1.98	1.81	2.0	1.49	1.39	1.44	1.51	1.88	1.25	1.38	1.52	1.63	1.84	1.91	2.92	2.65	2.78	6.41
PoinTr [24]	1.15	0.83	0.95	0.99	0.83	0.68	0.64	0.79	0.91	0.83	0.61	0.72	0.87	0.98	1.46	1.32	1.5	5.91
<i>3DSGrasp</i>	0.92	0.64	1.00	0.81	0.52	0.49	0.48	0.53	0.99	0.65	0.40	0.68	0.51	0.70	1.18	1.15	0.98	4.89

TABLE II
ABLATION STUDY ON THE NETWORK DESIGNS.

Model	Skip-connection	Offset-Attention	$L2$ CD
A			1.15
B	✓		1.02
C		✓	0.98
D	✓	✓	0.92

TABLE III
ABLATION STUDY ON THE NORMALISATION TECHNIQUE EFFECT IN $L2$ CD LOSS

Method	Baseline norm.	Our norm.
PCN	2.59	2.07
PoinTr	1.66	1.15
Ours	1.44	0.92

is the predicted sparse PCD. We supervise both PS and PC using Ground-truth completed point cloud during training to force both the encoder and decoder about the complete shape of the GT PCD.

C. Grasp pose generation

To generate the 6DoF grasp pose candidate for the two-fingered gripper, we use the Grasp Pose Detection (GPD) network introduced in [13]. GPD uniformly samples points in a Region of Interest (ROI) at random. ROIs are selected from an image-based object detection algorithm, but the algorithm can be tailored to the application’s constraints. On the randomly sampled points of the ROIs, a local search heuristic is applied to find suitable orientations in the vicinity of each point, so a grasp candidate GK corresponds to the sampled point and selected orientation. Then, the candidates GK are classified as graspable by a four-layer CNN. The input of the CNN is a multiple-view representation of the (clipped by the gripper) point cloud. To obtain the views, the PCD is voxelized, and the voxels are projected onto orthogonal axes. Finally, the grasps are ranked according to the output of the last layer of the CNN before the application of the Soft-max function. Thus, the classification score CS provides the ranking of GK . In real-world experiments, grasps are executed according to their ranks. Each grasp candidate corresponds to the goal pose of the end-effector of the robotic arm. We use MoveIt to compute a collision-free trajectory for the arm to reach the target pose.

IV. EXPERIMENTS

The experiments are divided into two parts. In Section IV-A, we first evaluate the performance of our proposed PCD completion network against a range of state-of-the-art methods on the partial YCB dataset [4]. We also perform

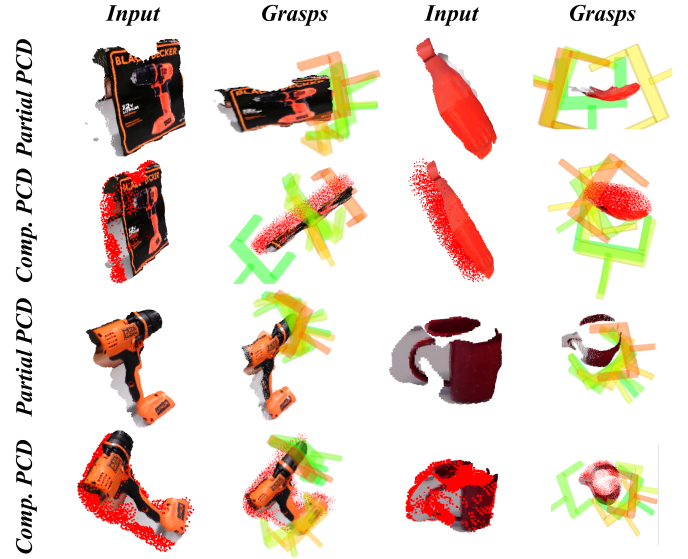


Fig. 4. Qualitative results of generated grasp proposal on the top of partial and completed PCD of 4 objects using our PCD completion network. The partial PCDs are acquired by the real sensor on the Kinova robotic arm. Each candidate grasp pose generated by GPD is color-coded with green to red representing the score from high to low.

extensive ablation studies to justify the design choices of our PCD completion network architecture in terms of the Offset-Attention encoder-decoder and the skip-connection. We then present real robotic experiments in Section IV-B utilising a Kinova robotic arm equipped with an RGB-D sensor. We evaluate the grasp success rate of *3DSGrasp* (i.e., GPD with our PCD completion network) in comparison to the one using only partial PCD. A grasp is considered successful if the object is held (i.e. does not fall) after lifting it up.

A. Evaluation on 3D shape completion

Dataset. We use the partial version of the YCB dataset [4], which is a popular choice in PCD completion for robotic grasping [5]. We randomly sample 50 views from the training set (Training Views), 50 views from the holdout view set (Holdout Views), and 50 views from the holdout models set (Holdout Models). We evaluate the completion network on holdout views and holdout model sets and the training is achieved with only the training set split. As the exact train/test split is unspecified in [4], for a fair comparison with the state-of-the-art methods on PCD shape completion, we train all compared methods using our own split dataset. All PCDs are pre-processed as described in Section III-A for normalisation and sampling.

TABLE IV
REAL ROBOT EXPERIMENT RESULT.

Method	Avg	Pringles	Drill box	Mustard	Mug	Cleanser	Clamps (biggest)	Drill	Jell-o	Baseball	Pitcher with lid
GPD [13]	46%	50%	0%	50%	70%	60%	30%	40%	80%	60%	20%
GPD + ours	76%	80%	80%	80%	80%	70%	70%	80%	90%	90%	40%

Comparison. We compare our proposed PCD completion network against a range of state-of-the-art methods for shape completion in terms of the $L2$ Chamfer-Distance loss [6] (multiplied by 1000) between the reconstructed, and the Ground truth PCD. For a fair comparison, we train from scratch (all the mentioned methods using the same dataset and split) and test against the existing PCD completion networks such as FoldingNet [21], PCN [6], MSN [7], and PoinTr [24] on the partial YCB dataset using their open-source code with their best hyper-parameters. We are unable to fairly compare against [27] as the code is unavailable. As shown in Table I, on average, our completion network achieves the lowest reconstruction loss among the competitors, outperforming the state-of-the-art method (i.e., PoinTr) by +0.23.

Ablation. We perform extensive experiments to justify our network design choices in terms of Offset-attention and skip-connection using the partial YCB dataset. Moreover, we evaluate the effect of our proposed PCD alignment processing technique on PCD completion performance.

Does the Offset-attention and skip-connection improve the PCD completion accuracy? We evaluate the impact of our proposed Offset-Attention encoder-decoder layer and skip-connection on the PCD reconstruction error. A set of variant models are studied: model *A* is the baseline Transformer with Self-Attention encoder-decoder layer, model *B* adds Skip-connection between the encoder and decoder to the baseline model, model *C* replaces Self-Attention with Offset-Attention layer and model *D* adds both skip-connection and Offset-Attention layer to the transformer. As shown in Table II, we observed that using skip-connection can improve the performance of the baseline model by +0.13. When using the Offset-Attention layer (model *C*) instead of the Offset-Attention layer (model *A*), we observe an improvement to 0.98. The best result is achieved by model *D* when adding both skip-connection and Offset-Attention layers to the transformer.

Does the PCD pre-processing help with PCD completion?

We evaluate the effect of our proposed PCD pre-processing technique on point cloud completion using our completion network, PCN [6] and PoinTr [24]. The *Baseline norm.* stands for normalising the GT and partial PCD with the same formula but different parameters and *Our norm.* use the parameters of partial PCD to normalise GT PCD (See Section III-A). As shown in Table III, our network achieved the lowest reconstruction error compared to the other two methods using both PCD processing techniques. Moreover, there is a large reduction in PCD reconstruction error when applying our proposed pre-processing technique to all methods.

B. Evaluation on robotic grasping

We perform the real-world experiments utilising a Kinova Gen3 robot equipped with a Robotiq 2F-85 gripper for grasping and an Intel RealSense D430 depth camera to capture the point cloud. During the experiments, an object is placed on the table and the robot starts at a predefined initial pose facing the object as shown in Figure 1. For each test, the partial PCD of the object is extracted by removing the background information using a PCD segmentation network [12]. Then, the segmented point cloud is fed to our completion network. Finally, the GPD [13] network generates and ranks grasp candidates for both the completed and the partial PCD. The final grasp is chosen as the best ranked and with a feasible solution when sent to the MoveIt! motion planner. Our method takes about 2 seconds for segmentation, 2 seconds for completion and 6 seconds for planning a grasp. Each object is grasped 10 times at different poses w.r.t. the arm in its workspace. We compare GPD vs. GPD with *3DSGrasp* in Table IV, where GPD with *3DSGrasp* consistently outperform GPD without object PCD completion. For example, the drill box dimensions (specially width, see Fig. 4) are not captured by the partial PCD, which results in a low success rate due to collisions with the object. With our method, the completed PCD can better address this issue after correctly reconstructing its shape. Another hard case is the Pitcher with a lid given the size of the object and the gripper’s maximum aperture. Since the lid reduces the available grasp poses from the top, only grasps from the handle are feasible. Nevertheless, our *3DSGrasp* doubles the successful grasps compared to the baseline with partial PCD.

V. CONCLUSIONS

In this work, we propose a new system called *3DSGrasp* for improving the robotic grasp success rate in a real-world experiment. The central core of the proposed system is the PCD completion head with the ability to complete accurately the missing geometry of the 3D objects that have not been observed before and without moving the camera to extract more information. We also proposed a new way to normalise partial views of PCD, solving the misalignment problem that improves the robotic grasp success rate and reduces PCD completion error. With the experiment, we show that our network achieves a state-of-the-art result on PCD completion tasks and improves the average-grasp success rate by a large margin. In future work, we will extend our work to multi-object shape completion and grasping. Moreover, we will investigate the possibility of a fusion between single-view RGB image and 3D PCD within the framework of *3DSGrasp*, to boost the 3D completion accuracy and grasp success rate.

REFERENCES

- [1] A. Ten Pas and R. Platt, "Using geometry to detect grasp poses in 3d point clouds," in *Robotics research*, pp. 307–324, Springer, 2018.
- [2] S. Lu, R. Wang, Y. Miao, C. Mitash, and K. Bekris, "Online object model reconstruction and reuse for lifelong improvement of robot manipulation," in *2022 International Conference on Robotics and Automation (ICRA)*, pp. 1540–1546, IEEE, 2022.
- [3] H.-Y. Lin, S.-C. Liang, and Y.-K. Chen, "Robotic grasping with multi-view image acquisition and model-based pose estimation," *IEEE Sensors Journal*, vol. 21, no. 10, pp. 11870–11878, 2020.
- [4] J. Varley, C. DeChant, A. Richardson, J. Ruales, and P. Allen, "Shape completion enabled robotic grasping," in *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pp. 2442–2447, IEEE, 2017.
- [5] J. Lundell, F. Verdoja, and V. Kyrki, "Robust grasp planning over uncertain shape completions," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1526–1532, IEEE, 2019.
- [6] W. Yuan, T. Khot, D. Held, C. Mertz, and M. Hebert, "Pcn: Point completion network," in *2018 International Conference on 3D Vision (3DV)*, pp. 728–737, IEEE, 2018.
- [7] M. Liu, L. Sheng, S. Yang, J. Shao, and S.-M. Hu, "Morphing and sampling network for dense point cloud completion," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, pp. 11596–11603, 2020.
- [8] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 652–660, 2017.
- [9] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.
- [10] J. Wang, X. Lin, and H. Yu, "Poat-net: Parallel offset-attention assisted transformer for 3d object detection for autonomous driving," *IEEE Access*, vol. 9, pp. 151110–151117, 2021.
- [11] M.-H. Guo, J.-X. Cai, Z.-N. Liu, T.-J. Mu, R. R. Martin, and S.-M. Hu, "Pct: Point cloud transformer," *Computational Visual Media*, vol. 7, no. 2, pp. 187–199, 2021.
- [12] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," *Advances in neural information processing systems*, vol. 30, 2017.
- [13] A. ten Pas, M. Gualtieri, K. Saenko, and R. Platt, "Grasp pose detection in point clouds," *The International Journal of Robotics Research*, vol. 36, no. 13-14, pp. 1455–1473, 2017.
- [14] D. Coleman, I. Sukan, S. Chitta, and N. Correll, "Reducing the barrier to entry of complex robotic software: a moveit! case study," *Journal of Software Engineering for Robotics*, vol. 5, no. 1, pp. 3–16, 2014.
- [15] B. Calli, A. Singh, A. Walsman, S. Srinivasa, P. Abbeel, and A. M. Dollar, "The ycb object and model set: Towards common benchmarks for manipulation research," in *2015 international conference on advanced robotics (ICAR)*, pp. 510–517, IEEE, 2015.
- [16] X. Han, Z. Li, H. Huang, E. Kalogerakis, and Y. Yu, "High-resolution shape completion using deep neural networks for global structure and local geometry inference," in *Proceedings of the IEEE international conference on computer vision*, pp. 85–93, 2017.
- [17] M. Kazhdan, M. Bolitho, and H. Hoppe, "Poisson surface reconstruction," in *Proceedings of the fourth Eurographics symposium on Geometry processing*, vol. 7, 2006.
- [18] R. P. de Figueiredo, P. Moreno, and A. Bernardino, "Efficient pose estimation of rotationally symmetric objects," *Neurocomputing*, vol. 150, pp. 126–135, 2015.
- [19] M. Berger, A. Tagliasacchi, L. Seversky, P. Alliez, J. Levine, A. Sharf, and C. Silva, "State of the art in surface reconstruction from point clouds," *Eurographics 2014-State of the Art Reports*, vol. 1, no. 1, pp. 161–185, 2014.
- [20] S. Gurumurthy and S. Agrawal, "High fidelity semantic shape completion for point clouds using latent optimization," in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1099–1108, IEEE, 2019.
- [21] Y. Yang, C. Feng, Y. Shen, and D. Tian, "Foldingnet: Point cloud auto-encoder via deep grid deformation," in *Proceedings of the IEEE Conf. on computer vision and pattern recognition*, pp. 206–215, 2018.
- [22] T. Groueix, M. Fisher, V. G. Kim, B. C. Russell, and M. Aubry, "A papier-mâché approach to learning 3d surface generation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 216–224, 2018.
- [23] L. Pan, T. Wu, Z. Cai, Z. Liu, X. Yu, Y. Rao, J. Lu, J. Zhou, M. Xu, X. Luo, *et al.*, "Multi-view partial (mvp) point cloud challenge 2021 on completion and registration: Methods and results," *arXiv preprint arXiv:2112.12053*, 2021.
- [24] X. Yu, Y. Rao, Z. Wang, Z. Liu, J. Lu, and J. Zhou, "Pointnet: Diverse point cloud completion with geometry-aware transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 12498–12507, 2021.
- [25] D. Yang, T. Tosun, B. Eisner, V. Isler, and D. Lee, "Robotic grasping through combined image-based grasp proposal and 3d reconstruction," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6350–6356, IEEE, 2021.
- [26] M. Van der Merwe, Q. Lu, B. Sundaralingam, M. Matak, and T. Hermans, "Learning continuous 3d reconstructions for geometrically aware grasping," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 11516–11522, 2020.
- [27] W. Chen, H. Liang, Z. Chen, F. Sun, and J. Zhang, "Improving object grasp performance via transformer-based sparse shape completion," *Journal of Intelligent & Robotic Systems*, vol. 104, no. 3, pp. 1–14, 2022.
- [28] T. Groueix, M. Fisher, V. Kim, B. Russell, and M. Aubry, "Atlasnet: A papier-mâché approach to learning 3d surface generation. arXiv 2018," *arXiv preprint arXiv:1802.05384*, 1802.
- [29] H. Xie, H. Yao, S. Zhou, J. Mao, S. Zhang, and W. Sun, "Grnet: Gridding residual network for dense point cloud completion," in *European Conference on Computer Vision*, pp. 365–381, Springer, 2020.
- [30] G. Du, K. Wang, S. Lian, and K. Zhao, "Vision-based robotic grasping from object localization, object pose estimation to grasp estimation for parallel grippers: a review," *Artificial Intelligence Review*, vol. 54, no. 3, pp. 1677–1734, 2021.
- [31] I. Lenz, H. Lee, and A. Saxena, "Deep learning for detecting robotic grasps," *The International Journal of Robotics Research*, vol. 34, no. 4-5, pp. 705–724, 2015.
- [32] J. Redmon and A. Angelova, "Real-time grasp detection using convolutional neural networks," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1316–1322, 2015.
- [33] H.-S. Fang, C. Wang, M. Gou, and C. Lu, "Graspnet-1billion: A large-scale benchmark for general object grasping," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11444–11453, 2020.
- [34] J. Mahler, M. Matl, V. Satish, M. Danielczuk, B. DeRose, S. McKinley, and K. Goldberg, "Learning ambidextrous robot grasping policies," *Science Robotics*, vol. 4, no. 26, p. eaau4984, 2019.
- [35] M. Sundermeyer, A. Mousavian, R. Triebel, and D. Fox, "Contact-graspnet: Efficient 6-dof grasp generation in cluttered scenes," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 13438–13444, IEEE, 2021.
- [36] A. Mousavian, C. Eppner, and D. Fox, "6-dof graspnet: Variational grasp generation for object manipulation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2901–2910, 2019.
- [37] B. Zhao, H. Zhang, X. Lan, H. Wang, Z. Tian, and N. Zheng, "Regnet: Region-based grasp network for end-to-end grasp detection in point clouds," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 13474–13480, IEEE, 2021.
- [38] Y. Bengio, "Practical recommendations for gradient-based training of deep architectures," in *Neural networks: Tricks of the trade*, pp. 437–478, Springer, 2012.
- [39] S. S. Mohammadi, Y. Wang, and A. Del Bue, "Pointview-gcn: 3d shape classification with multi-view point clouds," in *2021 IEEE International Conference on Image Processing (ICIP)*, pp. 3103–3107, IEEE, 2021.
- [40] S. S. Mohammadi, Y. Wang, M. Taiana, P. Morerio, and A. Del Bue, "Svp-classifier: Single-view point cloud data classifier with multi-view hallucination," in *Image Analysis and Processing-ICIAP 2022: 21st International Conference, Lecce, Italy, May 23–27, 2022, Proceedings, Part II*, pp. 15–26, Springer, 2022.
- [41] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon, "Dynamic graph cnn for learning on point clouds," *ACM Transactions on Graphics (TOG)*, 2019.

- [42] H. Lin, X. Cheng, X. Wu, and D. Shen, "Cat: Cross attention in vision transformer," in *2022 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, IEEE, 2022.
- [43] L. P. Tchampi, V. Kosaraju, H. Rezatofighi, I. Reid, and S. Savarese, "Topnet: Structural point cloud decoder," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 383–392, 2019.