VAPCNet: Viewpoint-Aware 3D Point Cloud Completion

Zhiheng Fu[1], Longguang Wang[2], Lian Xu[1], Zhiyong Wang[3],
Hamid Laga[4], Yulan Guo[5*], Farid Boussaid[1], Mohammed Bennamoun[1]

[1]The University of Western Australia [2]National University of Defense Technology
[3]The University of Sydney [4]Murdoch University [5]The Shenzhen Campus of Sun Yat-sen University

Abstract

Most existing learning-based 3D point cloud completion methods ignore the fact that the completion process is highly coupled with the viewpoint of a partial scan. However, the various viewpoints of incompletely scanned objects in real-world applications are normally unknown and directly estimating the viewpoint of each incomplete object is usually time-consuming and leads to huge annotation cost. In this paper, we thus propose an unsupervised viewpoint representation learning scheme for 3D point cloud completion without explicit viewpoint estimation. To be specific, we learn abstract representations of partial scans to distinguish various viewpoints in the representation space rather than the explicit estimation in the 3D space. We also introduce a Viewpoint-Aware Point cloud Completion Network (VAPCNet) with flexible adaption to various viewpoints based on the learned representations. The proposed viewpoint representation learning scheme can extract discriminative representations to obtain accurate viewpoint information. Reported experiments on two popular public datasets show that our VAPCNet achieves state-of-the-art performance for the point cloud completion task. Source code is available at https://github.com/FZH92128/VAPCNet.

1. Introduction

Incomplete shapes in 3D scans resulting from occlusion (both self-occlusion and occlusion by other objects) and low resolution of sensors often make them unsuitable for direct use in practical applications such as object grasping and Virtual Reality (VR) [16, 9, 25, 10]. To remedy this, shape completion aims to recover complete 3D shapes from partial 3D scans.

Current learning-based shape completion methods can be classified into two categories: volumetric representation The former category transforms a point cloud into 3D occupancy grids and uses 3D convolution operations to predict the complete shapes of objects. However, these volumetric representations often come with high memory expenses and restricted shape accuracy. The second category, on the other hand, operates directly on point clouds, providing a more memory-efficient way to represent 3D data. However, these point cloud based methods are difficult to recover detailed structures of objects due to the irregular and disorder nature of point clouds.

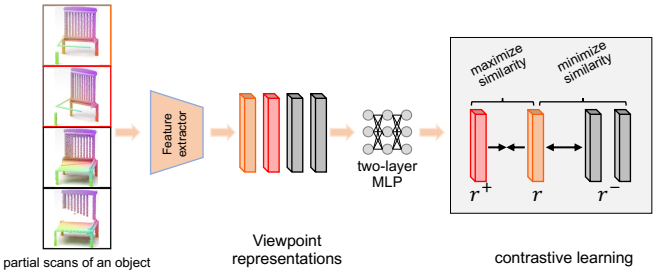

Figure 1. Example of an unsupervised viewpoint representation learning. In MVP dataset [30], all objects have been aligned, and missing parts indicate various incomplete objects scanned from varying viewpoints. In this context, incomplete objects marked with orange and red boxes are scanned from the same viewpoint, while those marked with black boxes have been scanned from different viewpoints.

To recover the full shape of partial point cloud objects, most existing methods adopt a coarse-to-fine approach to firstly predict coarse full shapes and then recover detailed complete shapes using complex refinement modules [54, 41, 52, 29, 30, 57]. More specifically, skip-connection is used to connect the detailed visible point-wise features with the missing point-wise features to preserve the detailed local patterns of objects [41]. In addition, attention mechanism is used to explore correlation between (non-)local patterns to recover detailed structures [52]. Although existing methods have made a certain progress in recovering the details of objects, they ignore the fact that the completion process is highly coupled with the scanning viewpoint.

Commonly, complete objects can be composed of visible parts and missing parts. For each object, the visible parts are closely related to corresponding viewpoints and and can vary a lot when obtained from different viewpoints. This means that the viewpoint information of each incomplete object can provide additional cues to 3D point cloud completion. However, explicit viewpoint estimation is usually time-consuming and leads to huge annotation cost. Motivated by the recent advances in contrastive learning [2, 7, 17, 20, 35], we propose to use unsupervised viewpoint representation learning by contrasting positive pairs with negative pairs within the latent space (as shown in Fig. 1), thereby generating a distinct latent representation for each incomplete object. The benefits of the proposed viewpoint representation learning are twofold: **First**, it is more practicable to learn abstract representations to distinguish between different viewpoints than extracting full representations for estimating the viewpoints of incomplete objects. Consequently, we can obtain a discriminative viewpoint representation of incomplete objects to provide auxiliary viewpoint information. **Second**, the unsupervised viewpoint representation learning does not require supervision with ground-truth viewpoints, making it more suitable for real-world applications with unknown viewpoints.

In this paper, we propose to complete partial point cloud objects under the guidance of viewpoint representation. More specifically, we first encode incomplete point clouds into viewpoint representations using contrastive learning. Then, we propose a Viewpoint-Aware Point cloud Completion Network (VAPCNet) with flexible adaptation to different viewpoints based on the learned representations. The proposed VAPCNet incorporates viewpoint information to perform feature adaptation by predicting convolutional kernels and modulated self-attention for local information aggregation from the viewpoint representation. Experimental results show that our network can handle various viewpoints and produce state-of-the-art results on both MVP [30] and PCN datasets [54]. Formally, our contributions include:

- A novel formulation of point cloud completion with unsupervised viewpoint representation learning, which reveals the usefulness of viewpoint information;
- An effective viewpoint-aware module that perform feature adaptation by predicting convolutional kernels and modulated self-attention for local information aggregation for accurate point cloud completion;
- State-of-the-art completion results on both MVP and PCN datasets.

2. Related Work

We first review the main approaches for learning-based point cloud completion, including volumetric representation based shape completion and point cloud based shape completion (Sec. 2.1). Then, we discuss the recent advances of contrastive learning (Sec. 2.2).

2.1. Learning-based Point Cloud Completion

Volumetric Representation based Shape Completion. The use of structured volumetric representations and powerful 3D convolutions has resulted in significant achievements in 3D reconstruction [4, 12] and shape completion [5, 18, 44]. However, these methods are expensive in terms of computation time and memory requirements. Although these issues can be alleviated based on sparse representations [14, 33, 37], the quantization operation in these methods still causes a significant loss of detailed information.

Point Cloud based Shape Completion. More recently, researchers have started to use unstructured point clouds as representations for 3D objects to reduce memory requirements and represent finer grained details. Nevertheless, the migration from structured 3D data understanding to point clouds analysis is non-trivial because the commonly used convolution operator is no longer suitable for unordered point clouds. PointNet and its variants [31, 31] allow the direct processing of 3D points for many downstream tasks. In the point cloud completion field, PCN-Net [54] was the first learning-based architecture. It adopts a FoldingNet [51] to map the 2D points onto a 3D surface by mimicking the deformation of a 2D plane. Following PCN-Net, Tchapmi et al. [34] proposed a tree-structured decoder to predict complete shapes. To preserve and recover local details, a number of approaches [38, 22, 41, 45] exploited local features to refine their 3D completion results. NSFA [57] recovered complete 3D shapes by combining known features and missing features. However, NSFA assumed that the ratio of the known part and the missing part is around 1:1, i.e., the visible part should be roughly half of the whole object. However, in most cases, this assumption does not hold for point cloud completion. Pan et al. [30] proposed a variational framework by leveraging the relationship between structures during the completion process. PMP-Net [42] accomplished the completion task by learning point moving paths. PMP-Net++ [43] enabled multi-step movement of the input point set and used the least total moving distance loss to mimic the earth mover distance. Xiang et al. [47] proposed a snowflake point deconvolution for point cloud completion. More recently, Yu et al. [52] reformulated point cloud completion as a set-to-set translation problem and applied transformer for missing point clouds prediction. Wang et al. [39] achieved feature extraction by matching the point features to a set of pre-trained local descriptors and designed neighbor-pooling operation that relies on adopting the feature vectors with the highest activations.

Existing methods pay more attention to designing more discriminative global features and more efficient refinement modules to recover high-quality coarse predictions and the

detailed structures of objects, respectively. Among them, [55, 13, 42, 43] are highly related to our work as they aim to do the completion along the scanning viewpoint. In [55], the ground-truth viewpoint was explicitly combined with the partial point cloud to achieve completion along the scanning viewpoint. Although it achieves better performance than previous methods, additional ground-truth viewpoints are required. In contrast to [55], [42] and [43], we propose to implicitly extract viewpoint information based on contrastive learning to get rid of the requirement of ground-truth viewpoints. We also propose a viewpoint-aware point cloud completion network by using the viewpoint representation, achieving better performance than [55, 42, 43] (see Sec. 4.2).

2.2. Contrastive Learning

In the unsupervised representation learning field, contrastive learning has demonstrated its effectiveness. Previous methods [6, 56, 11, 27] conduct representation learning by using auto-encoders to reconstruct the input itself. On the contrary, contrastive learning aims to maximize the mutual information within a representation space. Namely, the representation of a query sample should be more similar to positive counterparts while be more dissimilar to negative counterparts. The positive counterparts can be transformed versions of the input [2, 20, 46], multiple views of the input [35] and neighboring patches in the same image [21, 28].

Inspired by the success of contrastive learning in the 2D field, numerous related works [50, 58, 36, 53, 48] have been explored for 3D point cloud understanding. PointContrast [50] performs point-level invariant mapping on two transformed views of the given point cloud. Zhang et al. [58] proposed the DepthContrast to learn representations from depth scans. Wang et al. [36] proposed an encoder-decoder mechanism to reconstruct the occluded point clouds. Inspired by MAE [19] presenting a masked auto-encoder strategy for image representation learning, Yu et al. [53] proposed Point-BERT to recover the original point tokens at the masked locations under the supervision of point tokens. In this paper, we propose to learn viewpoint representation of incomplete point clouds using contrastive learning. To clarify, the partial point cloud generated with the same viewpoint (annotated with an red box in Fig. 1) is considered as positive counterpart and the partial point clouds generated from other viewpoints (annotated with an black boxes in Fig. 1) are considered as negtative counterparts.

3. Methodology

3.1. An Overview

The proposed point cloud completion framework consists of a viewpoint representation encoder and a viewpoint-aware completion network, as illustrated in Fig. 2. First, the partial scan is fed to the viewpoint representation encoder (Fig. 2(a)) to obtain a viewpoint representation. Then, this representation is incorporated in the viewpoint-aware completion network (Fig. 2(b)) to produce the complete object.

3.2. Viewpoint Representation Learning

The goal of viewpoint representation learning is to derive a discriminative representation from the partial point cloud objects in an unsupervised way, as depicted in Fig. 1. In this study, we assume that the viewpoint representation r of a partial scan is similar to its rotated version r^+ and differs for various viewpoints r^-.

In the context of viewpoint representation learning (refer to Fig. 1), a partial object is scanned from a specific viewpoint (marked by an orange box). This scan acts as the query sample, and its rotated version (indicated by a red box) is regarded as a positive sample. On the other hand, partial objects scanned from different viewpoints (represented by black boxes) are considered negative samples. To extract viewpoint representations, we employ a feature extractor to encode these query, positive, and negative samples. Drawing inspiration from SimCLR [2] and MoCo v2 [3], the resulting representations are further processed through a two-layer multi-layer perceptron (MLP) projection head to derive r, r^+ and r^-. In line with MoCo [20], we apply an InfoNCE loss to promote similarity between r and r^+ while maintaining dissimilarity with r^-. That is,

$$\mathcal{L}_r = -log\frac{exp(r \cdot r^+/\tau)}{\sum_{m=1}^{M} exp(r \cdot r_m^-/\tau) + exp(r \cdot r^+/\tau)}, \quad (1)$$

where M is the number of negative samples, τ is a temperature hyper-parameter and $\cdot$ represents the dot product between two vectors.

Each object is scanned from B viewpoints to form B partial point cloud objects (i.e., B different viewpoints). During the unsupervised training, one partial point cloud object is randomly sampled from these B partial point cloud objects as the query sample and then the remaining $B-1$ ones are selected as negative samples. The positive sample is produced by randomly rotating the query sample. Next, the query sample and these B samples are encoded into $p_q \in \mathbb{R}^{512}$ and $p_k \in \mathbb{R}^{B\times 512}$ using the feature extractor. For p_q, $p_k^1 \in \mathbb{R}^{512}$ is the latent representation of the positive sample and the remaining $B-1$ latent representations are used as negative samples. The overall loss is defined as:

$$\mathcal{L}_{view} = -log\frac{exp(p_q \cdot p_k^1/\tau)}{\sum_{j=2}^{B} exp(p_1 \cdot p_k^j/\tau) + exp(p_q \cdot p_k^1/\tau)}, \quad (2)$$

where p_k^j represents the j^{th} negative sample.

Discussion. Most existing point cloud completion methods ignore the influence of viewpoints to shape completion task.

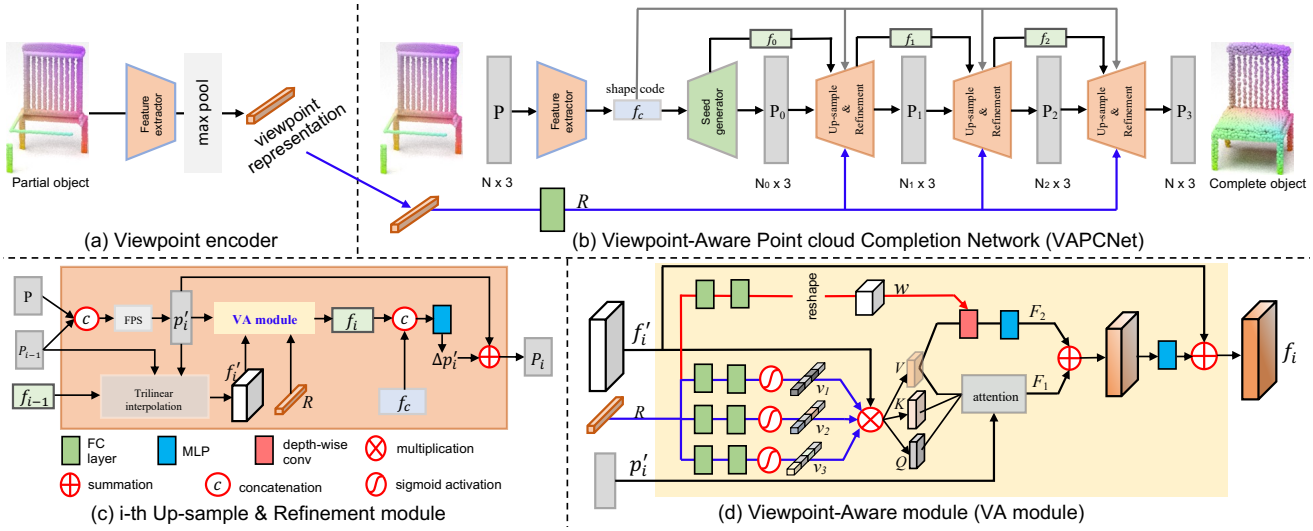

Figure 2. An overview of the proposed VAPCNet.

In this paper, we aim at learning a "good" abstract representation to distinguish a specific viewpoint representation from others rather than explicitly estimating the viewpoint. In Sec. 5, we demonstrate that our viewpoint representation learning scheme can obtain discriminative representations. Moreover, our unsupervised learning scheme does not require the supervision using ground-truth viewpoints.

3.3. Viewpoint-aware Point Cloud Completion

With unsupervised viewpoint representation learning, a Viewpoint-Aware Point cloud Completion Network (VAPCNet) is proposed to complete the incomplete object using the resultant representation. Our VAPCNet has an encoder-decoder structure, as shown in Fig. 2 (b).

Encoder-Decoder Architecture. The encoder is composed of a feature extractor and a seed generator. The decoder is composed of three up-sample and refinement modules.

Given the incomplete input point cloud $P \in \mathbb{R}^{N\times 3}$, the feature extractor uses three layers of set abstraction from [32] to aggregate point features from local to global, along which a point transformer [59] is applied to incorporate local shape context. Then, max-pooling is used to extract a shape code f_c of size $1 \times C$ from point features $p_{fe} \in \mathbb{R}^{32\times C}$ generated by the feature extractor. The aim of the seed generator is to produce a coarse but complete point cloud P_0 of size $N_0 \times 3$ that captures the geometry and structure of the target shape. More specifically, the extracted shape code f_c is integrated with the point features $p_{fe} \in \mathbb{R}^{32\times C}$ through a MLP to combine the global shape information and the visible local information. Then, a transposed convolution is used to split these aggregated features into N_0 point features f_0, which are fed to MLP to generate a coarse point cloud P_0 of size $N_0 \times 3$.

The decoder recovers high-resolution complete point clouds in a hierarchical manner. In the up-sample and refinement module (as shown in Fig. 2(c)), to preserve the detailed structure of input point cloud, we first concatenate the previous low-resolution complete point cloud P_{i-1} and the input point cloud P, and then adopt the Farthest Point Sampling (FPS) to select N_i ($N_i > N_{i-1}$) points $P_i^{'}$ from the concatenated point cloud $P_c \in \mathbb{R}^{(N+N_{i-1})\times 3}$. Next, given the previous complete point cloud P_{i-1}, the up-sampled point cloud $P_i^{'}$ and previous point features f_{i-1}, we use the trilinear interpolation [32] to obtain N_i point features $f_i^{'}$ for the up-sampled point cloud $P_i^{'}$. Subsequently, $f_i^{'}$ is fed to the Viewpoint-Aware module (VA module) to produce refined point features f_i. Finally, the refined point features f_i concatenated with the shape code f_c are fed to the MLP to predict the corresponding residual coordinates $\Delta P_i^{'}$, which are summed up with $P_i^{'}$ to obtain the refined point cloud $P_i = P_i^{'} + \Delta P_i^{'}$.

VA Module. Motivated by [59] and [36], our VA module learns to predict the kernel of a depth-wise convolution and modulated self-attention conditioned on the viewpoint representation $V_R \in \mathbb{R}^{1\times C}$, as shown in Fig. 2(d). In the VA module, the inputs are composed of the point features $f_i^{'}$, the viewpoint representation R and the up-sampled point cloud $P_i^{'}$. To be specific, the viewpoint representation R is fed to two parts. In the first part, the viewpoint representation R is fed to three branches (the blue lines) with two full-connected (FC) layers and a sigmoid activation layer in each branch, aiming to generate channel-wise modulation coefficients v_1, v_2, v_3. Then, v_1, v_2, v_3 are used to rescale different channel components of $f_i^{'}$, resulting in V, Q, K. Next, N nearest neighbour is used to select a set of points (n) and point features in a local neighborhood for each point of $p_i^{'}$ and each point feature of K. Afterwards, self-attention [59] is used to aggregate local information for each point

feature to produce point features F_1. For the second part (the red line), the viewpoint representation R is fed to another two full-connected (FC) layers and a reshape layer to produce a convolutional kernel $w \in \mathbb{R}^{M \times N_i \times n}$. Then, the point feature V is processed with a 1×1 depth-wise convolution (using w) and MLP to produce F_2. Next, F_1 is summed up with F_2 and fed to the subsequent layers to produce the refined point features f_i.

Discussion. The direct concatenation of viewpoint representation and point cloud features may not fully utilize the viewpoint information due to the domain gap between them. To more effectively integrate viewpoint information for 3D completion, we tailor a depth-wise convolutional kernel and predict channel-wise modulation coefficients (within the self-attention unit) for feature adaptation. By employing the VA module, the viewpoint representation is effectively incorporated, enhancing shape completion (as discussed in Sec. 5).

3.4. Loss Functions

Chamfer Distance (CD) and Earth Mover's Distance (EMD) are introduced [8] to measure the differences between two point clouds (P, Q). We choose the Chamfer distance due to its efficiency over EMD.

$$\mathcal{L}_{CD}(P,Q) = \frac{1}{P}\sum_{x \in P} \min_{y \in Q} \|x-y\|_2 + \frac{1}{Q}\sum_{y \in Q} \min_{x \in P} \|y-x\|_2, \tag{3}$$

where P and Q are the predicted complete point clouds and corresponding ground truth.

To explicitly constrain point clouds generated in the seed generator and the subsequent up-sampling process, we down-sampled the ground truth point clouds to the same sampling density as P_0, P_1, P_2, P_3. We define the sum of the four CD losses as the completion loss, denoted by $\mathcal{L}_{completion}$.

$$\begin{aligned} \mathcal{L}_{completion} = \mathcal{L}_{CD}(P_0, gt_0) + \mathcal{L}_{CD}(P_1, gt_1) \\ + \mathcal{L}_{CD}(P_2, gt_2) + \mathcal{L}_{CD}(P_3, gt_3), \end{aligned} \tag{4}$$

where gt_0, gt_1, gt_2, gt_3 are down-sampled ground truth point clouds corresponding to P_0, P_1, P_2, P_3.

We also exploit the partial matching loss from [40] to preserve the shape structure of the input point cloud. It is an unidirectional constraint which aims to match one shape to another without constraining the opposite direction. The partial matching loss only requires the output point cloud to partially match the input, denoted as the preservation loss $\mathcal{L}_{preservation}$. The total training loss is formulated as

$$\mathcal{L} = \mathcal{L}_{completion} + \lambda \mathcal{L}_{preservation}, \tag{5}$$

where λ is set to 0 and 1 for MVP dataset and PCN dataset, respectively. This ensures fair comparison with other methods.

4. Experiments

4.1. Datasets and Implementation Details

We tested the effectiveness of the proposed VAPCNet on MVP and PCN datasets.

MVP Dataset [30]. The MVP dataset is composed of 16 categories of 4000 CAD models. 26 camera locations are sampled for each model to simulate partial scans. In our experiments, each complete point cloud contains $2048, 4096, 8192, 16384$ points and each incomplete point cloud comprises 2048 points. Following [30], we use 62400 partial-complete point cloud pairs for training and 41600 pairs for testing. For evaluation, we also adopt the L_2 version of Chamfer distance.

PCN Dataset [54]. The PCN dataset [54] is developed based on a subset of the ShapeNet dataset [1]. In our experiments, each complete point cloud contains $16,384$ points and each incomplete point cloud comprises 2048 points. The training set includes $28,974$ different models from 8 categories. Each model has a complete point cloud with 8 partial point clouds taken from different viewpoints for data augmentation. The validation set contains 100 models. The testing set contains 1200 models with 150 models in each of the 8 categories. For evaluation, we adopt the L_1 version of Chamfer distance following [54].

Implementation Details. For the viewpoint representation encoder, we adopt three layers of set abstraction from [32] to aggregate point features from local to global, along which point transformer [59] is applied to incorporate local shape context. For the pre-training, the Adam method [23] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ was used for optimization. We first trained the viewpoint representation encoder by optimizing L_{view} for 40 epochs. The learning rate was initially set to 1×10^{-3} and then decreased to 1×10^{-4} after 20 epochs. Then, we froze the pre-trained viewpoint representation learning architecture and trained the whole network for 80 epochs and 400 epochs for MVP dataset and PCN dataset, respectively. To generate high-resolution complete point clouds, we combined the proposed VAPCNet with the Snowflake Point Deconvolution (SPD) module from [47]. The learning rate was initially set to 1×10^{-4} and then decreased to half after every 40 epochs. All experiments were run on NVIDIA 2080Ti GPUs.

4.2. Completion on the MVP Dataset

For a fair comparison, we trained all methods using the same training strategy on the MVP dataset.

Quantitative comparison. The CD loss and F-score for all evaluated methods are reported in Table 1 and Table 2, re-

Table 1. Shape completion results (CD loss multiplied by 10^4) on the multi-view partial (MVP) point cloud dataset (16,384 points). The lower, the better.

Method	airplane	cabinet	car	chair	lamp	sofa	table	watercraft	bed	bench	bookshelf	bus	guitar	motorbike	pistol	skateboard	Avg.
PCN [54]	2.95	4.13	3.04	7.07	14.93	5.56	7.06	6.08	12.72	5.73	6.91	2.46	1.02	3.53	3.28	2.99	6.02
TopNet [34]	2.72	4.25	3.40	7.95	17.01	6.04	7.42	6.04	11.56	5.62	8.22	2.37	1.37	3.90	3.97	2.09	6.36
MSN [24]	2.07	3.82	2.76	6.21	12.72	4.74	5.32	4.80	9.93	3.89	5.85	2.12	0.69	2.48	2.91	1.58	4.90
Wang et. al. [38]	1.59	3.64	2.60	5.24	9.02	4.42	5.45	4.26	9.56	3.67	5.34	2.23	0.79	2.23	2.86	2.13	4.30
ECG [29]	1.41	3.44	2.36	4.58	6.95	3.81	4.27	3.38	7.46	3.10	4.82	1.99	0.59	2.05	2.31	1.66	3.58
GRNet [49]	1.61	4.66	3.10	4.72	5.66	4.61	4.85	3.53	7.82	2.96	4.58	2.97	1.28	2.24	2.11	1.61	3.87
NSFA [57]	1.51	4.24	2.75	4.68	6.04	4.29	4.84	3.02	7.93	3.87	5.99	2.21	0.78	1.73	2.04	2.14	3.77
VRCNet [30]	1.15	3.20	2.14	3.58	5.57	3.58	4.17	2.47	6.90	2.76	3.45	1.78	0.59	1.52	1.83	1.57	3.12
Our VAPCNet	**0.78**	**3.19**	**2.10**	**3.05**	**3.16**	**3.14**	**3.26**	**2.15**	**5.36**	**1.92**	**3.08**	**1.68**	**0.33**	**1.39**	**1.34**	**0.95**	**2.40**

Table 2. Shape completion results (F-Score@1%) on the multi-view partial (MVP) point cloud dataset (16,384 points). The higher, the better.

Method	airplane	cabinet	car	chair	lamp	sofa	table	watercraft	bed	bench	bookshelf	bus	guitar	motorbike	pistol	skateboard	Avg.
PCN [54]	0.861	0.641	0.686	0.517	0.455	0.552	0.646	0.628	0.452	0.694	0.546	0.779	0.906	0.665	0.774	0.861	0.638
TopNet [34]	0.798	0.621	0.612	0.443	0.387	0.506	0.639	0.609	0.405	0.680	0.524	0.766	0.868	0.619	0.726	0.837	0.601
MSN [24]	0.879	0.692	0.693	0.599	0.604	0.627	0.730	0.696	0.569	0.797	0.637	0.806	0.935	0.728	0.809	0.885	0.710
Wang et. al. [38]	0.898	0.688	0.725	0.670	0.681	0.641	0.748	0.742	0.600	0.797	0.659	0.802	0.931	0.772	0.843	0.902	0.740
ECG [29]	0.906	0.680	0.716	0.683	0.734	0.651	0.766	0.753	0.640	0.822	0.706	0.804	0.945	0.780	0.835	0.897	0.753
GRNet [49]	0.861	0.641	0.686	0.517	0.455	0.552	0.646	0.628	0.452	0.694	0.546	0.779	0.906	0.665	0.774	0.861	0.638
NSFA [57]	0.903	0.694	0.721	0.737	0.783	0.705	0.817	0.799	0.687	0.845	0.747	0.815	0.932	0.815	0.858	0.894	0.783
VRCNet [30]	0.928	0.721	0.756	0.743	0.789	0.696	0.813	0.800	0.674	0.863	0.755	0.832	0.960	0.834	0.887	0.930	0.796
Our VAPCNet	**0.942**	**0.762**	**0.758**	**0.786**	**0.844**	**0.759**	**0.845**	**0.824**	**0.736**	**0.892**	**0.808**	**0.854**	**0.978**	**0.845**	**0.902**	**0.944**	**0.829**

Table 3. Shape completion results (CD loss multiplied by 10^4) on multi-view partial point cloud (MVP) dataset with various point cloud resolutions.

#Points	2048		4096		8192		16384	
	CD↓	F1↑	CD↓	F1↑	CD↓	F1↑	CD↓	F1↑
PCN [54]	9.77	0.320	7.96	0.458	6.99	0.563	6.02	0.638
TopNet [34]	10.11	0.308	8.20	0.440	7.00	0.533	6.36	0.601
MSN [24]	7.90	0.432	6.17	0.585	5.42	0.659	4.90	0.710
Wang et. al. [38]	7.25	0.434	5.83	0.569	4.90	0.680	4.30	0.740
ECG [29]	6.64	0.476	5.41	0.585	4.18	0.690	3.58	0.753
GRNet [49]	7.61	0.353	5.73	0.493	4.51	0.616	3.54	0.700
VRCNet [30]	5.96	0.499	4.70	0.636	3.64	0.727	3.12	0.791
PoinTr [52]	5.79	0.499	4.29	0.638	3.52	0.725	2.95	0.783
PMP-Net++[43]	-	-	-	-	-	-	3.38	0.687
Zhang et. al.[55]	-	-	-	-	-	-	2.42	0.800
Wang et. al [39]	-	-	-	-	-	-	-	0.816
SnowflakeNet[47]	5.71	0.503	4.45	0.648	3.48	0.743	2.69	0.796
Our VAPCNet	**5.40**	**0.521**	**3.96**	**0.658**	**3.02**	**0.763**	**2.40**	**0.829**

spectively. The proposed VAPCNet outperforms all existing competitive methods in terms of CD and F-score@1%. We also compare our method with existing approaches that support multi-resolution completion in Table 3. Our VAPCNet outperforms all the other methods. *Specifically*, our method using unsupervised viewpoint representation learning outperforms [55] which uses the ground-truth viewpoint as input. This demonstrates the effectiveness of our viewpoint representation learning in terms of extracting accurate viewpoint information.

Qualitative comparison. The qualitative comparison results are shown in Fig. 3. The proposed VAPCNet can generate better complete shapes with finer details than other methods. In particular, we can clearly observe the effectiveness of the proposed VA module in our complete shapes. For example, the missing legs of the chairs (the second row in Fig. 3) are recovered based on the observed legs with the awareness of the scanned viewpoint. In the third row of Fig. 3, the lamp base is reconstructed with a better shape of lampstand than other methods. With additional viewpoint information provided by the viewpoint representation, VAPCNet can effectively reconstruct complete shapes by learning structural relations using the proposed VA module from the incomplete point cloud.

4.3. Completion on the PCN Dataset

We also compare our network on the PCN dataset with several state-of-the-art baseline methods. We use the $\mathcal{L}_1$ Chamfer Distance as the evaluation metric. For the baseline methods, the results of [29, 57] are produced from the codes and pre-trained models released in their official projects at Github. The results of the other methods are copied from [47, 49, 42] and their original papers [39, 52].

Quantitative comparison. The quantitative results are shown in Table 4. Note that, our network achieves the lowest L1 CD on average. In the categories of cabinet and car, our method achieves comparable performance compared to SnowFlakeNet [47] on CD. In the other categories, our method gets better performances. Specifically, our method achieves significant improvement in the category of lamp compared to SnowFlakeNet [47], reaching 8.6%. Although [39] achieves better performance than SnowFlakeNet [47] on MVP dataset, it gets worse performance than SnowFlakeNet [47] on PCN dataset. In contrast, our method achieves consistently better performance on these two datasets.

Qualitative comparison. Fig. 4 shows qualitative comparison results. We can see that our results predict a more accu-

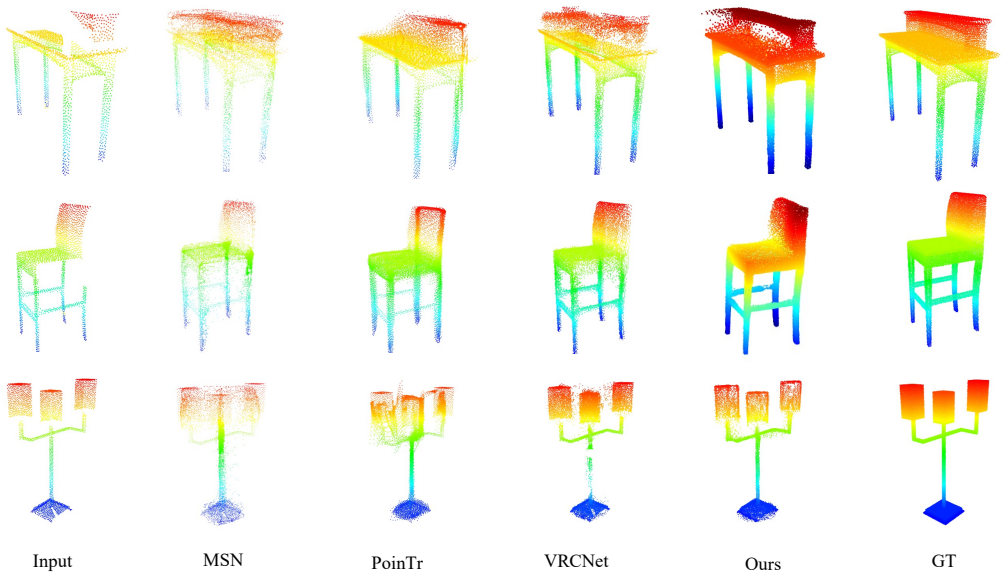

Figure 3. Visual comparisons on MVP dataset. Note that, the partial point clouds (2048 points) are sparse and self-occluded, as opposed to the reconstructed and ground truth point clouds (16,384 points) which are dense and complete.

Table 4. Quantitative comparison on PCN dataset with state-of-the-art methods in terms of L1 Chamfer Distance $\times 10^3$. The lower, the better.

Models	Avg.	airplane	cabinet	car	chair	lamp	couch	table	watercraft
AtlasNet [15]	10.58	6.37	11.94	10.10	12.06	12.37	12.99	10.33	10.61
FoldingNet [51]	14.31	9.49	15.80	12.61	15.55	16.41	15.97	13.65	14.99
PCN [54]	9.64	5.50	22.70	10.63	8.70	11.00	11.34	11.68	8.59
TopNet [34]	12.15	7.61	13.31	10.90	13.82	14.44	14.78	11.22	11.12
GRNet [49]	8.83	6.45	10.37	9.45	9.41	7.96	10.51	8.44	8.04
Wang et. al. [38]	8.51	4.79	9.97	8.31	9.49	8.94	10.69	7.81	8.05
PMP-Net [42]	8.73	5.65	11.24	9.64	9.51	6.95	10.83	8.72	7.25
ECG [29]	8.63	5.23	10.12	8.36	9.43	8.53	10.94	7.98	8.16
NSFA [57]	8.32	5.03	10.51	9.11	9.16	7.45	10.46	7.56	7.28
SK-PCN [26]	8.49	5.09	9.98	8.22	9.29	8.39	10.80	7.84	8.02
PoinTr [52]	8.38	4.75	10.47	8.68	9.39	7.75	10.93	7.78	7.29
SnowflakeNet[30]	7.21	4.29	**9.16**	**8.08**	7.89	6.07	9.23	6.55	6.40
Wang et. al [39]	7.96	-	-	-	-	-	-	-	-
Our VAPCNet	**7.02**	**4.10**	9.28	8.15	**7.51**	**5.55**	**9.18**	**6.28**	**6.10**

rate shape with detailed patterns. For instance, in the second and third row in Fig. 4, our method can better recover the detailed back of the chair and the couch. In contrast, results from other methods are very noisy. This demonstrates the ability of our network to efficiently enhance the shape with local details.

Table 5. Ablation studies for the proposed VA module, including convolutional kernels and modulated self-attention. "CD" denotes $\mathcal{L}_2$ Chamfer Distance (multiplied by 10^4) and "F1" represents F Score @1%.

Model	VA module		CD ↓	F1 ↑
	Kernel	modulated self-attention		
1	✗	✗	7.12	0.493
2	✓	✗	5.53	0.513
3	✗	✓	5.54	0.514
4	✓	✓	**5.40**	**0.521**

5. Ablation Study

In this section, we test the effectiveness of viewpoint representation and the proposed VA module.

Viewpoint Representation Learning. Our viewpoint representation aims at extracting viewpoint information from incomplete objects. To demonstrate this, we conducted experiments to study the effect of different viewpoints of each CAD model on our viewpoint representations. Specifically, there are 26 camera poses that are uniformly distributed on a unit sphere for each CAD model for training set. During testing, we randomly generated two rotation angles to rotate the incomplete objects and thus there are three incomplete objects for each viewpoint (78 incomplete objects for each CAD model). We pre-trained the viewpoint representation architecture on the training set and evaluated it on

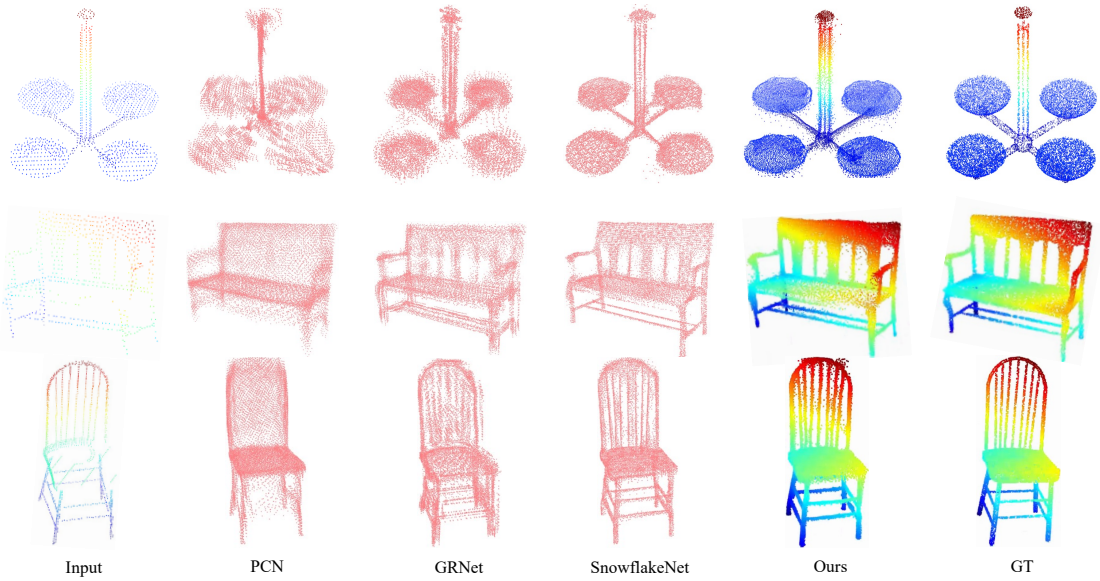

Figure 4. Visual comparisons on PCN dataset. Note that, the partial point clouds (2048 points) are sparse and self-occluded, as opposed to the reconstructed and ground truth point clouds (16,384 points) which are dense and complete.

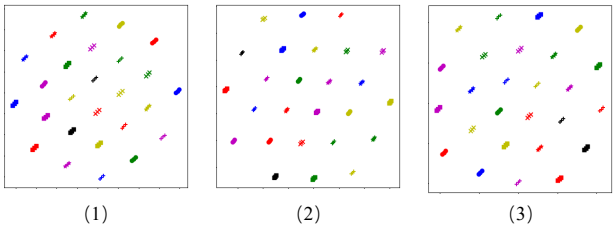

Figure 5. Examples of unsupervised viewpoint representation learning. (1), (2) and (3) represent three different CAD models. For each example, there are 26 partial objects scanned from different viewpoints for each CAD model. Each symbol represents a viewpoint representation.

the testing set. Fig. 5 shows that these 26 latent representations (three representations for each viewpoint) are far away from each other, indicating that 26 viewpoints are learned into different latent representations. This means that the pre-trained viewpoint representation architecture enables to learn a discriminative viewpoint representation, which benefits VAPCNet.

Effectiveness of the VA module. In order to evaluate the efficactiveness of the proposed VA module, which includes convolutional kernels and modulated self-attention, we carried out ablation studies on the MVP dataset (consisting of 2048 points), as presented in Table 5. In Model 1, viewpoint representation learning is omitted and MLPs are utilized to substitute the VA module in the up-sample & refinement module. Consequently, this model experiences limited accuracy. However, the inclusion of the convolutional kernel and modulated self-attention enhances the performance from 7.12 to 5.53 and 5.54, respectively. With both modules incorporated, Model 4 delivers the best outcome, thereby validating the effectiveness of our viewpoint representation learning and VA module.

6. Conclusion

In this paper, we proposed an unsupervised viewpoint representation learning scheme to achieve detailed 3D point cloud completion. Instead of explicitly estimating the viewpoints of scanned incomplete objects, we use contrastive learning to extract discriminative representations to distinguish different viewpoints. We also introduce a Viewpoint-Aware Point cloud Completion Network (VAPCNet) based on the learned representations to deal with partial objects scanned from different viewpoints. It is demonstrated that our viewpoint representation learning scheme can extract discriminative representations to obtain accurate viewpoint information and the proposed VAPCNet can recover detailed and accurate complete point clouds. Reported experimental results show that our VAPCNet achieves state-of-the-art performance on both MVP and PCN datasets. The future work will focus on generating more discriminative shape code to further improve the performance of point cloud completion for more object categories.

Acknowledgement

This research was financially supported by the Australian Research Council (ARC DP210101682, DP210102674) and received additional partial funding from the National Natural Science Foundation of China (Grant Numbers: U20A20185, 61972435), the Guangdong Basic and Applied Basic Research Foundation (Grant Number: 2022B1515020103), and the Shenzhen Science and Technology Program (Grant Number: RCYX20200714114641140).

References

[1] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 5

[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML)*, pages 1597–1607, 2020. 2, 3

[3] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 3

[4] Christopher B Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *European Conference on Computer Vision (ECCV)*, pages 628–644. Springer, 2016. 2

[5] Angela Dai, Charles Ruizhongtai Qi, and Matthias Nießner. Shape completion using 3d-encoder-predictor cnns and shape synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5868–5877, 2017. 2

[6] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1422–1430, 2015. 3

[7] Alexey Dosovitskiy, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with convolutional neural networks. *Advances in Neural Information Processing Systems (NIPS)*, 27, 2014. 2

[8] Haoqiang Fan, Hao Su, and Leonidas J Guibas. A point set generation network for 3d object reconstruction from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 605–613, 2017. 5

[9] Mingtao Feng, Haoran Hou, Liang Zhang, Yulan Guo, Hongshan Yu, Yaonan Wang, and Ajmal Mian. Exploring hierarchical spatial layout cues for 3d point cloud based scene graph prediction. *IEEE Transactions on Multimedia*, 2023. 1

[10] Zhiheng Fu, Siyu Hong, Mengyi Liu, Hamid Laga, Mohammed Bennamoun, Farid Boussaid, and Yulan Guo. Multi-stage information diffusion for joint depth and surface normal estimation. *Pattern Recognition*, 141:109660, 2023. 1

[11] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018. 3

[12] Rohit Girdhar, David F Fouhey, Mikel Rodriguez, and Abhinav Gupta. Learning a predictable and generative vector representation for objects. In *European Conference on Computer Vision (ECCV)*, pages 484–499. Springer, 2016. 2

[13] Bingchen Gong, Yinyu Nie, Yiqun Lin, Xiaoguang Han, and Yizhou Yu. Me-pcn: Point completion conditioned on mask emptiness. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12488–12497, 2021. 3

[14] Benjamin Graham, Martin Engelcke, and Laurens Van Der Maaten. 3d semantic segmentation with submanifold sparse convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9224–9232, 2018. 2

[15] Thibault Groueix, Matthew Fisher, Vladimir G Kim, Bryan C Russell, and Mathieu Aubry. A papier-mâché approach to learning 3d surface generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 216–224, 2018. 7

[16] Yulan Guo, Hanyun Wang, Qingyong Hu, Hao Liu, Li Liu, and Mohammed Bennamoun. Deep learning for 3d point clouds: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 43(12):4338–4364, 2020. 1

[17] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 1735–1742, 2006. 2

[18] Xiaoguang Han, Zhen Li, Haibin Huang, Evangelos Kalogerakis, and Yizhou 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 (ICCV)*, pages 85–93, 2017. 2

[19] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16000–16009, 2022. 3

[20] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9729–9738, 2020. 2, 3

[21] Olivier Henaff. Data-efficient image recognition with contrastive predictive coding. In *International Conference on Machine Learning (ICML)*, pages 4182–4192. PMLR, 2020. 3

[22] Zitian Huang, Yikuan Yu, Jiawen Xu, Feng Ni, and Xinyi Le. Pf-net: Point fractal network for 3d point cloud completion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7662–7670, 2020. 2

[23] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[24] Minghua Liu, Lu Sheng, Sheng Yang, Jing Shao, and Shi-Min Hu. Morphing and sampling network for dense point cloud completion. In *Proceedings of the AAAI conference on Artificial Intelligence (AAAI)*, volume 34, pages 11596–11603, 2020. 6

[25] Bo Miao, Mohammed Bennamoun, Yongsheng Gao, and Ajmal Mian. Spectrum-guided multi-granularity referring video object segmentation. *arXiv preprint arXiv:2307.13537*, 2023. 1

[26] Yinyu Nie, Yiqun Lin, Xiaoguang Han, Shihui Guo, Jian Chang, Shuguang Cui, Jian Zhang, et al. Skeleton-bridged point completion: From global inference to local adjustment. *Advances in Neural Information Processing Systems (NIPS)*, 33:16119–16130, 2020. 7

[27] Mehdi Noroozi, Hamed Pirsiavash, and Paolo Favaro. Representation learning by learning to count. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 5898–5906, 2017. 3

[28] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 3

[29] Liang Pan. Ecg: Edge-aware point cloud completion with graph convolution. *IEEE Robotics and Automation Letters*, 5(3):4392–4398, 2020. 1, 6, 7

[30] Liang Pan, Xinyi Chen, Zhongang Cai, Junzhe Zhang, Haiyu Zhao, Shuai Yi, and Ziwei Liu. Variational relational point completion network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8524–8533, 2021. 1, 2, 5, 6, 7

[31] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas 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 (CVPR)*, pages 652–660, 2017. 2

[32] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in Neural Information Processing Systems (NIPS)*, 30, 2017. 4, 5

[33] Hang Su, Varun Jampani, Deqing Sun, Subhransu Maji, Evangelos Kalogerakis, Ming-Hsuan Yang, and Jan Kautz. Splatnet: Sparse lattice networks for point cloud processing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2530–2539, 2018. 2

[34] Lyne P Tchapmi, Vineet Kosaraju, Hamid Rezatofighi, Ian Reid, and Silvio Savarese. Topnet: Structural point cloud decoder. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 383–392, 2019. 2, 6, 7

[35] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *European Conference on Computer Vision (ECCV)*, pages 776–794. Springer, 2020. 2, 3

[36] Longguang Wang, Yingqian Wang, Xiaoyu Dong, Qingyu Xu, Jungang Yang, Wei An, and Yulan Guo. Unsupervised degradation representation learning for blind super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10581–10590, 2021. 3, 4

[37] Peng-Shuai Wang, Yang Liu, Yu-Xiao Guo, Chun-Yu Sun, and Xin Tong. O-cnn: Octree-based convolutional neural networks for 3d shape analysis. *ACM Transactions On Graphics (TOG)*, 36(4):1–11, 2017. 2

[38] Xiaogang Wang, Marcelo H Ang Jr, and Gim Hee Lee. Cascaded refinement network for point cloud completion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 790–799, 2020. 2, 6, 7

[39] Yida Wang, David Joseph Tan, Nassir Navab, and Federico Tombari. Learning local displacements for point cloud completion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1568–1577, 2022. 2, 6, 7

[40] Xin Wen, Zhizhong Han, Yan-Pei Cao, Pengfei Wan, Wen Zheng, and Yu-Shen Liu. Cycle4completion: Unpaired point cloud completion using cycle transformation with missing region coding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13080–13089, 2021. 5

[41] Xin Wen, Tianyang Li, Zhizhong Han, and Yu-Shen Liu. Point cloud completion by skip-attention network with hierarchical folding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1939–1948, 2020. 1, 2

[42] Xin Wen, Peng Xiang, Zhizhong Han, Yan-Pei Cao, Pengfei Wan, Wen Zheng, and Yu-Shen Liu. Pmp-net: Point cloud completion by learning multi-step point moving paths. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7443–7452, 2021. 2, 3, 6, 7

[43] Xin Wen, Peng Xiang, Zhizhong Han, Yan-Pei Cao, Pengfei Wan, Wen Zheng, and Yu-Shen Liu. Pmp-net++: Point cloud completion by transformer-enhanced multi-step point moving paths. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 45(1):852–867, 2022. 2, 3, 6

[44] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1912–1920, 2015. 2

[45] Zijie Wu, Yaonan Wang, Mingtao Feng, He Xie, and Ajmal Mian. Sketch and text guided diffusion model for colored point cloud generation, 2023. 2

[46] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3733–3742, 2018. 3

[47] Peng Xiang, Xin Wen, Yu-Shen Liu, Yan-Pei Cao, Pengfei Wan, Wen Zheng, and Zhizhong Han. Snowflakenet: Point cloud completion by snowflake point deconvolution with skip-transformer. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 5499–5509, 2021. 2, 5, 6

[48] Aoran Xiao, Jiaxing Huang, Dayan Guan, and Shijian Lu. Unsupervised representation learning for point clouds: A survey. *arXiv preprint arXiv:2202.13589*, 2022. 3

[49] Haozhe Xie, Hongxun Yao, Shangchen Zhou, Jiageng Mao, Shengping Zhang, and Wenxiu Sun. Grnet: Gridding residual network for dense point cloud completion. In *European Conference on Computer Vision (ECCV)*, pages 365–381. Springer, 2020. 6, 7

[50] Saining Xie, Jiatao Gu, Demi Guo, Charles R Qi, Leonidas Guibas, and Or Litany. Pointcontrast: Unsupervised pre-

training for 3d point cloud understanding. In *European Conference on Computer Vision (ECCV)*. Springer, 2020. 3

[51] Yaoqing Yang, Chen Feng, Yiru Shen, and Dong Tian. Foldingnet: Point cloud auto-encoder via deep grid deformation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 206–215, 2018. 2, 7

[52] Xumin Yu, Yongming Rao, Ziyi Wang, Zuyan Liu, Jiwen Lu, and Jie Zhou. Pointr: Diverse point cloud completion with geometry-aware transformers. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 12498–12507, 2021. 1, 2, 6, 7

[53] Xumin Yu, Lulu Tang, Yongming Rao, Tiejun Huang, Jie Zhou, and Jiwen Lu. Point-bert: Pre-training 3d point cloud transformers with masked point modeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19313–19322, 2022. 3

[54] Wentao Yuan, Tejas Khot, David Held, Christoph Mertz, and Martial Hebert. Pcn: Point completion network. In *International Conference on 3D Vision (3DV)*, pages 728–737. IEEE, 2018. 1, 2, 5, 6, 7

[55] Bowen Zhang, Xi Zhao, He Wang, and Ruizhen Hu. Shape completion with points in the shadow. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–9, 2022. 3, 6

[56] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European Conference on Computer Vision (ECCV)*, pages 649–666. Springer, 2016. 3

[57] Wenxiao Zhang, Qingan Yan, and Chunxia Xiao. Detail preserved point cloud completion via separated feature aggregation. In *European Conference on Computer Vision (ECCV)*, pages 512–528. Springer, 2020. 1, 2, 6, 7

[58] Zaiwei Zhang, Rohit Girdhar, Armand Joulin, and Ishan Misra. Self-supervised pretraining of 3d features on any point-cloud. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 10252–10263, 2021. 3

[59] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. Point transformer. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 16259–16268, 2021. 4, 5