Variational Relational Point Completion Network

Liang Pan[1*] Xinyi Chen[1,2] Zhongang Cai[2,3] Junzhe Zhang[1,2]
Haiyu Zhao[2,3] Shuai Yi[2,3] Ziwei Liu[1✉]
[1]S-Lab, Nanyang Technological University [2]SenseTime Research [3]Shanghai AI Laboratory
{liang.pan,ziwei.liu}@ntu.edu.sg, {xchen032,junzhe001}@e.ntu.edu.sg
{caizhongang,zhaohaiyu,yishuai}@sensetime.com

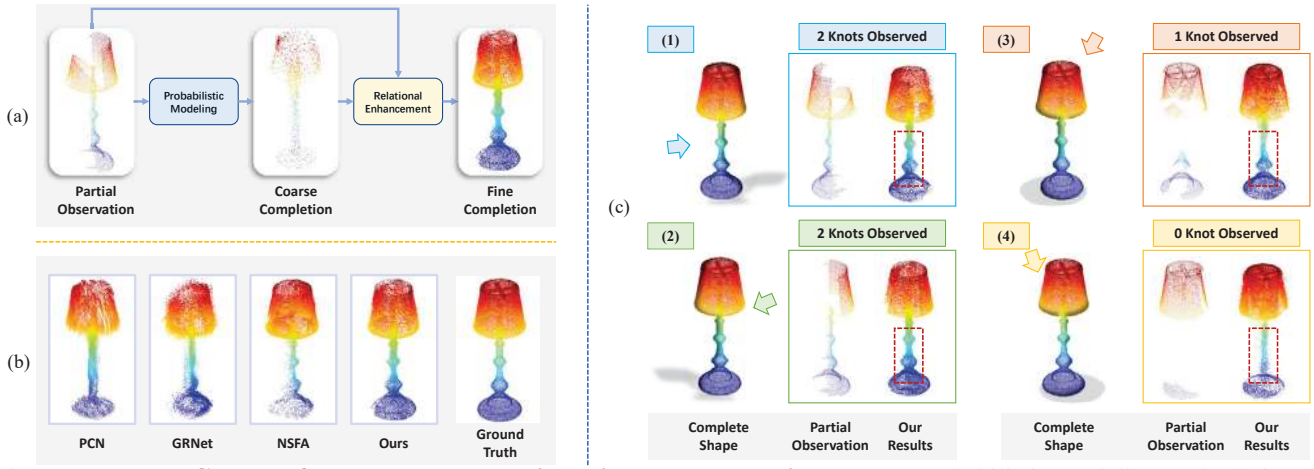


Figure 1: (a) **VRCNet performs shape completion with two consecutive stages**: probabilistic modeling and relational enhancement. (b) **Qualitative Results** show that VRCNet generates better shape details than the other works [29, 27, 30]. (c) **Our completion results conditioned on partial observations.** The arrows indicate the viewing angles. In (1) and (2), 2 knots are partially observed for the pole of the lamp, and hence we generate 2 complete knots. In (3), only 1 knot is observed, and then we reconstruct 1 complete knot. If no knots are observed (see (4)), VRCNet generates a smooth pole without knots.

Abstract

*Real-scanned point clouds are often incomplete due to viewpoint, occlusion, and noise. Existing point cloud completion methods tend to generate global shape skeletons and hence lack fine local details. Furthermore, they mostly learn a deterministic partial-to-complete mapping, but overlook structural relations in man-made objects. To tackle these challenges, this paper proposes a variational framework, **V**ariational **R**elational point **C**ompletion network (VRCNet) with two appealing properties: **1) Probabilistic Modeling.** In particular, we propose a dual-path architecture to enable principled probabilistic modeling across partial and complete clouds. One path consumes complete point clouds for reconstruction by learning a point VAE. The other path generates complete shapes for partial point clouds, whose embedded distribution is guided by distribution obtained from the reconstruction path during training. **2) Relational Enhancement.** Specifically, we carefully design point self-attention kernel and point selective kernel module to exploit relational point features, which refines local shape details conditioned on the coarse completion. In addition, we contribute **a multi-view partial point cloud dataset (MVP dataset)** containing over 100,000 high-quality scans, which renders partial 3D shapes from 26 uniformly distributed camera poses for each 3D CAD model. Extensive experiments demonstrate that VRCNet outperforms state-of-the-art methods on all standard point cloud completion benchmarks. Notably, VRCNet shows great generalizability and robustness on real-world point cloud scans.*

1. Introduction

3D point cloud is an intuitive representation of 3D scenes and objects, which has extensive applications in various vision and robotics tasks. Unfortunately, scanned 3D point clouds are usually incomplete owing to occlusions and missing measurements, hampering practical usages. Therefore, it is desirable and important to predict the complete

* Work partially done while working at NUS.
Our project website: https://paul007pl.github.io/projects/VRCNet

3D shape from a partially observed point cloud.

The pioneering work PCN [29] uses PointNet-based encoder to generate global features for shape completion, which cannot recover fine geometric details. The follow-up works [14, 23, 15, 27] provide better completion results by preserving observed geometric details from the incomplete point shape using local features. However, they [29, 14, 23, 15, 27] mostly generate complete shapes by learning a deterministic partial-to-complete mapping, lacking the conditional generative capability based on the partial observation. Furthermore, 3D shape completion is expected to recover plausible yet fine-grained complete shapes by learning relational structure properties, such as geometrical symmetries, regular arrangements and surface smoothness, which existing methods fail to capture.

To this end, we propose Variational Relational Point Completion network (entitled as VRCNet), which consists of two consecutive encoder-decoder sub-networks that serve as "probabilistic modeling" (PMNet) and "relational enhancement" (RENet), respectively (shown in Fig. 1 (a)). The first sub-network, PMNet, embeds global features and latent distributions from incomplete point clouds, and predicts the overall skeletons (*i.e*. coarse completions, see Fig. 1 (a)) that are used as 3D adaptive anchor points for exploiting multi-scale point relations in RENet. Inspired by [32], PMNet uses smooth complete shape priors to improve the generated coarse completions using a dual-path architecture consisting of two parallel paths: 1) a reconstruction path for complete point clouds, and 2) a completion path for incomplete point clouds. During training, we regularize the consistency between the encoded posterior distributions from partial point clouds and the prior distributions from complete point clouds. With the help of the generated coarse completions, the second sub-network RENet strives to enhance structural relations by learning multi-scale local point features. Motivated by the success of local relation operations in image recognition [31, 7], we propose the Point Self-Attention Kernel (PSA) as a basic building block for RENet. Instead of using fixed weights, PSA interleaves local point features by adaptively predicting weights based on the learned relations among neighboring points. Inspired by the Selective Kernel (SK) unit [12], we propose the Point Selective Kernel Module (PSK) that utilizes multiple branches with different kernel sizes to exploit and fuse multi-scale point features, which further improves the performance.

Moreover, we create a large-scale Multi-View Partial point cloud (MVP) dataset with over 100,000 high-quality scanned partial and complete point clouds. For each complete 3D CAD model selected from ShapeNet [26], we randomly render 26 partial point clouds from uniformly distributed camera views on a unit sphere, which improves the data diversity. Experimental results on our MVP and Completion3D benchmark [21] show that VRCNet outperforms SOTA methods. In Fig. 1 (b), VRCNet reconstructs richer details than the other methods by implicitly learning the shape symmetry from this incomplete lamp. Given different partial observations, VRCNet can predict different plausible complete shapes (Fig. 1 (c)). Furthermore, VRCNet can generate impressive complete shapes for incomplete real-world scans from KITTI [3] and ScanNet [2], which reveals its remarkable robustness and generalizability.

The key contributions can be summarized as: **1)** We propose a novel **V**ariational **R**elational point **C**ompletion **Net**work (VRCNet), and it first performs probabilistic modeling using a novel dual-path network followed by a relational enhancement network. **2)** We design multiple relational modules that can effectively exploit and fuse multi-scale point features for point cloud analysis, such as the Point Self-Attention Kernel and the Point Selective Kernel Module. **3)** Furthermore, we contribute a large-scale multi-view partial point cloud (MVP) dataset with over 100,000 high-quality 3D point shapes. Extensive experiments show that VRCNet outperforms previous SOTA methods on all evaluated benchmark datasets.

2. Related Works

Multi-scale Features Exploitation. Convolutional operations have yielded impressive results for image applications [11, 6, 19]. However, conventional convolutions cannot be directly applied to point clouds due to the absence of regular grids. Previous networks mostly exploit local point features by two operations: local pooling [24, 18, 16] and flexible convolution [4, 22, 13, 25]. Self-attention often uses linear layers, such as fully-connected (FC) layers and shared multilayer perceptron (shared MLP) layers, which are appropriate for point clouds. In particular, recent works [31, 7, 17] have shown that local self-attention (*i.e*. relation operations) can outperform their convolutional counterparts, which holds the exciting prospect of designing networks for point clouds.

Point Cloud Completion. The target of point cloud completion is to recover a complete 3D shape based on a partial point cloud observation. PCN [29] first generates a coarse completion based on learned global features from the partial input point cloud, which is upsampled using folding operations [28]. Following PCN, TopNet [21] proposes a tree-structured decoder to predict complete shapes. To preserve and recover local details, previous approaches [23, 15, 27] exploit local features to refine their 3D completion results. Recently, NSFA [30] recovers complete 3D shapes by combining known features and missing features. However, NSFA assumes that the ratio of the known part and the missing part is around $1:1$ (*i.e*., the visible part should be roughly a half of the whole object), which does not hold for point clouds completion in most cases.

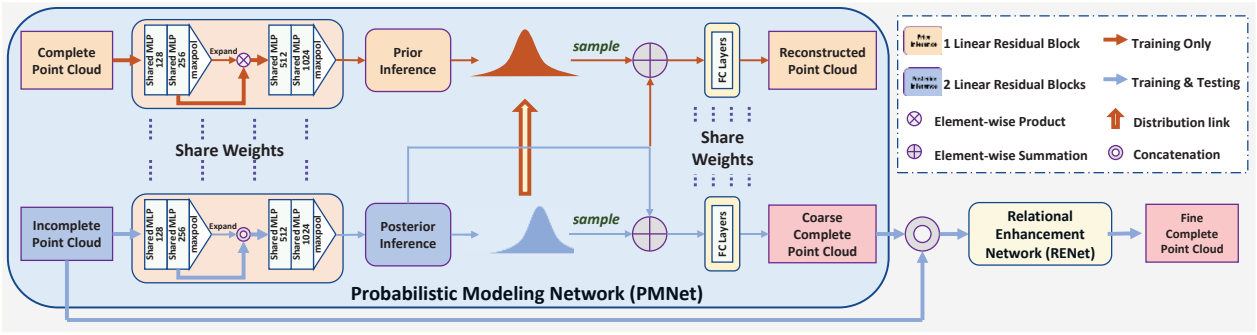


Figure 2: **Our PMNet (light blue block) consists of two parallel paths**, the upper construction path (orange line) and the lower completion path (blue line). The reconstruction path is only used in training, and the completion path generates a coarse complete point cloud based on the inferred distribution and global features. Subsequently, our RENet (Fig 4) adaptively exploits relational structure properties to predict the fine complete point cloud.

3. Our Approach

We define the incomplete point cloud $\mathbf{X}$ as a partial observation for a 3D object, and a complete point cloud $\mathbf{Y}$ is sampled from the surfaces of the object. Note that $\mathbf{X}$ need not to be a subset of $\mathbf{Y}$, since $\mathbf{X}$ and $\mathbf{Y}$ are generated by two separate processes. The point cloud completion task aims to predict a complete shape $\mathbf{Y}'$ conditioned on $\mathbf{X}$. VRCNet generate high-quality complete point clouds in a coarse-to-fine fashion. Firstly, we predict a coarse completion $\mathbf{Y}'_c$ based on embedded global features and an estimated parametric distribution. Subsequently, we recover relational geometries for the fine completion $\mathbf{Y}'_f$ by exploiting multi-scale point features with novel self-attention modules.

3.1. Probabilistic Modeling

Previous networks [29, 21] tend to decode learned global features to predict overall shape skeletons as their completion results, which cannot recover fine-grained geometric details. However, it is still beneficial to first predict the shape skeletons before refining local details for the following reasons: 1) shape skeletons describe the coarse complete structures, especially for those areas that are entirely missing in the partial observations; 2) shape skeletons can be regarded as adaptive 3D anchor points for exploiting local point features in incomplete point clouds [15]. With these benefits, we propose the Probabilistic Modeling network (PMNet) to generate the overall skeletons (*i.e.* coarse completions) for incomplete point clouds.

In contrast to previous methods, PMNet employs probabilistic modelling to predict the coarse completions based on both embedded global features and learned latent distributions. Moreover, we employ a dual-path architecture (shown in Fig. 2) that contains two parallel pipelines: the upper reconstruction path for complete point clouds $\mathbf{Y}$ and the lower completion path for partial point clouds $\mathbf{X}$. The reconstruction path follows a variational auto-encoder (VAE) scheme. It first encodes global features $\mathbf{z_g}$ and latent distributions $q_\phi(\mathbf{z_g}|\mathbf{Y})$ for the complete shape $\mathbf{Y}$, and then it uses a decoding distribution $p^r_\theta(\mathbf{Y}|\mathbf{z_g})$ to recover a complete shape $\mathbf{Y}'_r$. The objective function for the reconstruction path can be formulated as:

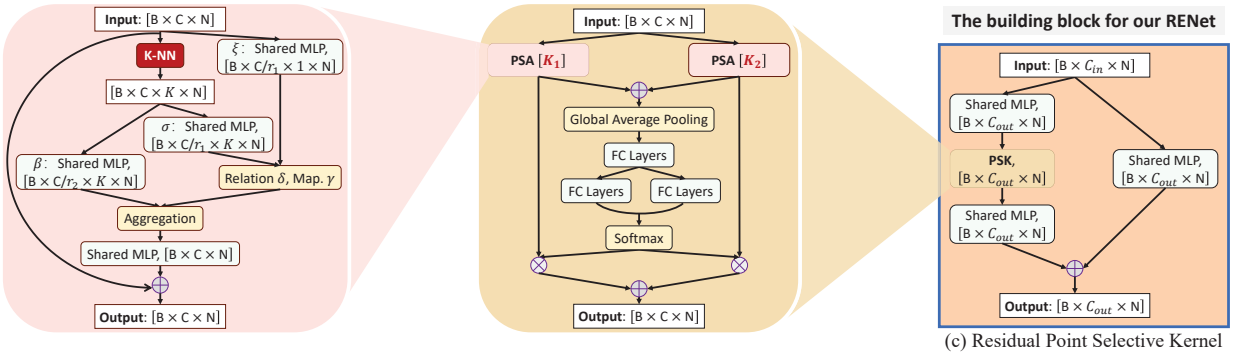
$$\begin{aligned}\mathcal{L}_{rec} = &-\lambda \mathbf{KL}\big[q_\phi(\mathbf{z_g}|\mathbf{Y}) \,\big\|\, p(\mathbf{z_g})\big] \\ &+ \mathbb{E}_{p_{data}(\mathbf{Y})}\mathbb{E}_{q_\phi(\mathbf{z_g}|\mathbf{Y})}\big[\log p^r_\theta(\mathbf{Y}|\mathbf{z_g})\big],\end{aligned} \tag{1}$$

where $\mathbf{KL}$ is the KL divergence, $\mathbb{E}$ denotes the estimated expectations of certain functions, $p_{data}(\mathbf{Y})$ denotes the true underlying distribution of data, and $p(\mathbf{z_g}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$ is the conditional prior predefined as a Gaussian distribution, and λ is a weighting parameter.

The completion path has a similar structure as the constructive path, and both two paths share weights for their encoder and decoder except the distribution inference layers. Likewise, the completion path aims to reconstruct a complete shape $\mathbf{Y}'_c$ based on global features $\mathbf{z_g}$ and latent distributions $p_\psi(\mathbf{z_g}|\mathbf{X})$ from an incomplete input $\mathbf{X}$. To exploit the most salient features from the incomplete point cloud, we use the learned conditional distribution $q_\phi(\mathbf{z_g}|\mathbf{Y})$ encoded by its corresponding complete 3D shapes $\mathbf{Y}$ to regularize latent distributions $p_\psi(\mathbf{z_g}|\mathbf{X})$ during training (shown as the Distribution Link in Fig. 2, the arrow indicates that we regularize $p_\psi(\mathbf{z_g}|\mathbf{X})$ to approach $q_\phi(\mathbf{z_g}|\mathbf{Y})$). Hence, $q_\phi(\mathbf{z_g}|\mathbf{Y})$ constitutes the prior distributions, $p_\psi(\mathbf{z_g}|\mathbf{X})$ is the posterior importance sampling function, and the objective function for completion path is defined as follows:

$$\begin{aligned}\mathcal{L}_{com} = &-\lambda \mathbf{KL}\big[q_\phi(\mathbf{z_g}|\mathbf{Y}) \,\big\|\, p_\psi(\mathbf{z_g}|\mathbf{X})\big] \\ &+ \mathbb{E}_{p_{data}(\mathbf{X})}\mathbb{E}_{p_\psi(\mathbf{z_g}|\mathbf{X})}\big[\log p^c_\theta(\mathbf{Y}|\mathbf{z_g})\big],\end{aligned} \tag{2}$$

where ϕ, ψ and θ represent different network weights of their corresponding functions. Notably, the reconstruction path is only used in training, and hence the dual-path architecture does not influence our inference efficiency.



(a) Point Self-Attention Kernel (PSA) (b) Point Selective Kernel Module (PSK) (c) Residual Point Selective Kernel Module (R-PSK)

Figure 3: **Our proposed point kernels.** (a) Our PSA adaptively aggregate neighboring point features. (b) Using selective kernel unit, our PSK can adaptively adjust receptive fields to exploit and fuse multi-scale point features. (c) By adding a residual connection, we construct our RPSK that is an important building block for our RENet.

3.2. Relational Enhancement

After obtaining coarse completions $\mathbf{Y}'_c$, the Relational Enhancement network (RENet) targets at enhancing structural relations to recover local shape details. Although previous methods [23, 30, 15] can preserve observed geometric details by exploiting local point features, they cannot effectively extract structural relations (*e.g.* geometric symmetries) to recover those missing parts conditioned on the partial observations. Inspired by the relation operations for image recognition [7, 31], we propose the Point Self-Attention kernel (PSA) to adaptively aggregate local neighboring point features with learned relations in neighboring points (Fig. 3 (a)). The operation of PSA is formulated as:

$$\mathbf{y}_i = \sum_{j \in \mathcal{N}(i)} \alpha(\mathbf{x}_{\mathcal{N}(i)})_j \odot \beta(\mathbf{x}_j), \tag{3}$$

where $\mathbf{x}_{\mathcal{N}(i)}$ is the group of point feature vectors for the selected K-Nearest Neighboring (K-NN) points $\mathcal{N}(i)$. $\alpha(\mathbf{x}_{\mathcal{N}(i)})$ is a weighting tensor for all selected feature vectors. $\beta(\mathbf{x}_j)$ is the transformed features for point j, which has the same spatial dimensionality with $\alpha(\mathbf{x}_{\mathcal{N}(i)})_j$. Afterwards, we obtain the output $\mathbf{y}_i$ using an element-wise product $\odot$, which performs a weighted summation for all points $j \in \mathcal{N}(i)$. The weight computation $\alpha(\mathbf{x}_{\mathcal{N}(i)})$ can be decomposed as follows:

$$\begin{aligned} \alpha(\mathbf{x}_{\mathcal{N}(i)}) &= \gamma\big(\delta(\mathbf{x}_{\mathcal{N}(i)})\big), \\ \delta(\mathbf{x}_{\mathcal{N}(i)}) &= \big[\sigma(\mathbf{x}_i), [\xi(\mathbf{x}_j)]_{\forall j \in \mathcal{N}(i)}\big], \end{aligned} \tag{4}$$

where γ, σ and ξ are all shared MLP layers (Fig. 3 (a)), and the relation function δ combines all feature vectors $\mathbf{x}_j \in \mathbf{x}_{\mathcal{N}(i)}$ by using concatenation operations.

Observing that different relational structures can have different scales, we enable the neurons to adaptively adjust their receptive field sizes by leveraging the selective kernel unit [12]. Hence, we construct the Point Selective Kernel module (PSK), which adaptively fuses learned structural relations from different scales. In Fig. 3 (b), we show a two-branch case, which has two PSA kernels with different kernel (*i.e.* K-NN) sizes. The operations of the PSK are formulated as:

$$\begin{cases} \mathbf{V}_c = \tilde{\mathbf{U}}_c \cdot a_c + \hat{\mathbf{U}}_c \cdot b_c\,, \\ a_c = \dfrac{e^{\mathbf{A}_c \mathbf{z}}}{e^{\mathbf{A}_c \mathbf{z}} + e^{\mathbf{B}_c \mathbf{z}}}, \quad b_c = \dfrac{e^{\mathbf{B}_c \mathbf{z}}}{e^{\mathbf{A}_c \mathbf{z}} + e^{\mathbf{B}_c \mathbf{z}}}, \\ \mathbf{U} = \tilde{\mathbf{U}} + \hat{\mathbf{U}}, \quad s_c = \dfrac{1}{N}\sum_{i=1}^{N} \mathbf{U}_c(i), \quad \mathbf{z} = \eta(\mathbf{W}\mathbf{s}), \end{cases} \tag{5}$$

where $\hat{\mathbf{U}}, \tilde{\mathbf{U}} \in \mathbb{R}^{N \times C}$ are point features encoded by two kernels respectively, $\tilde{\mathbf{V}} \in \mathbb{R}^{N \times C}$ is the final fused features, $\mathbf{s}$ is obtained by using element-wise average pooling over all N points for each feature $c \in C$, η is a FC layer, $\mathbf{W} \in \mathbb{R}^{d \times C}$, $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{C \times d}$, and d is a reduced feature size.

Furthermore, we add an residual path besides the main path (shown in Fig. 3 (c)) and then construct the Residual Point Selective Kernel module (R-PSK) that is used as a building block for RENet. As shown in Fig. 4, RENet follows a hierarchical encoder-decoder architecture by using Edge-preserved Pooling (EP) and Edge-preserved Unpooling (EU) modules [16]. We use an Edge-aware Feature Expansion (EFE) module [15] to expand point features, which

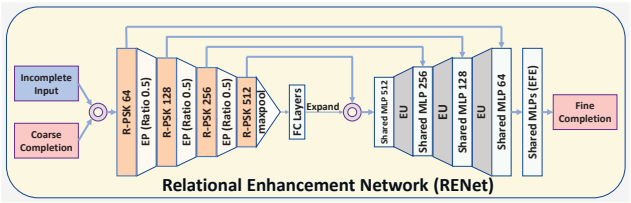
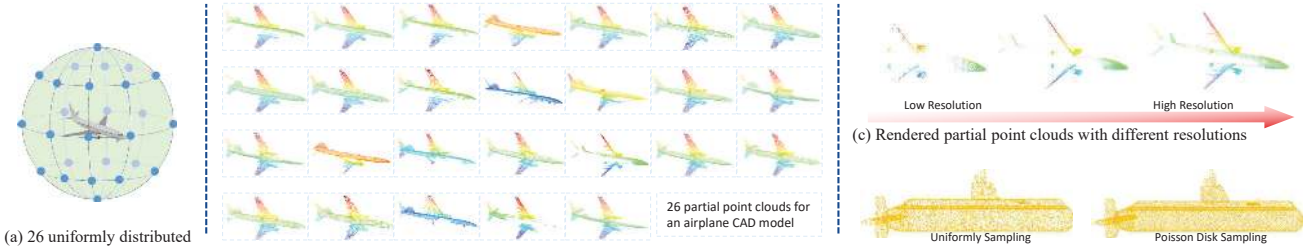


Figure 4: **Our Relational Enhancement Network (RENet)** uses a hierarchical encoder-decoder architecture, which effectively learns multi-scale structural relations.

Table 1: **Comparing MVP with existing datasets.** MVP has many appealing properties, such as 1) diversity of uniform views; 2) large-scale and high-quality; 3) rich categories. Note that both PCN and C3D only randomly render **One** incomplete point cloud for each CAD model to construct their testing sets. (C3D: Completion3D; Cat.: Categories; Distri.: Distribution; Reso.: Resolution; PC: Point Cloud; FPS: Farthest Point Sampling; PDS: Poisson Disk Sampling. Point cloud resolution is shown as multiples of 2048 points.)

	#Cat.	Training Set		Testing Set		Virtual Camera			Complete PC		Incomplete PC	
		#CAD	#Pair	#CAD	#Pair	Num.	Distri.	Reso.	Sampling	Reso.	Sampling	Reso.
PCN [29]	8	28974	∼200k	1200	1200	8	Random	160×120	Uniform	8×	Random	∼3000
C3D [21]	8	28974	28974	1184	1184	1	Random	160×120	Uniform	1×	Random	1×
MVP	**16**	2400	62400	1600	**41600**	**26**	**Uniform**	**1600×1200**	**PDS**	**1/2/4/8×**	**FPS**	1×



(a) 26 uniformly distributed camera poses

(b) The 26 rendered incomplete point clouds for this 3D airplane

(c) Rendered partial point clouds with different resolutions

(d) Sampled complete point clouds with different sampling methods

Figure 5: Our **M**ulti-**V**iew **P**artial point cloud dataset (**MVP**). (a) shows an example for our 26 uniformly distributed camera poses on a unit sphere. (b) presents the 26 partial point clouds for the airplane from our uniformly distributed virtual cameras. (c) compares the rendered incomplete point clouds with different camera resolutions. (d) shows that Poisson disk sampling generates complete point clouds with a higher quality than uniform sampling.

generates high-resolution complete point clouds with predicted fine local details. Consequently, multi-scale structural relations can be exploited for fine details generation.

3.3. Loss Functions

Our VRCNet is trained end-to-end, and the training loss consists of three parts: $\mathcal{L}_{rec}$ (reconstruction path), $\mathcal{L}_{com}$ (completion path) and $\mathcal{L}_{fine}$ (relational enhancement). $\mathcal{L}_{rec}$ and $\mathcal{L}_{com}$ have two loss items, a $\mathbf{KL}$ divergence loss and a reconstruction loss, while $\mathcal{L}_{fine}$ only has a reconstruction loss. The KL divergence is defined as:

$$\mathcal{L}_{\mathbf{KL}}(q,p) = -\mathbf{KL}\big[q(\mathbf{z}) \,\big\|\, p(\mathbf{z})\big]. \tag{6}$$

Considering the training efficiency, we choose the symmetric Chamfer Distance (CD) as the reconstruction loss:

$$\mathcal{L}_{\mathbf{CD}}(\mathbf{P},\mathbf{Q}) = \frac{1}{|\mathbf{P}|}\sum_{x\in\mathbf{P}}\min_{y\in\mathbf{Q}}\|x-y\|^2 + \frac{1}{|\mathbf{Q}|}\sum_{y\in\mathbf{Q}}\min_{x\in\mathbf{P}}\|x-y\|^2, \tag{7}$$

where x and y denote points that belong to two point clouds $\mathbf{P}$ and $\mathbf{Q}$, respectively. Consequently, the joint loss function can be formulated as:

$$\begin{aligned}\mathcal{L} =& \lambda_{rec}\mathcal{L}_{rec} + \lambda_{com}\mathcal{L}_{com} + \lambda_{fine}\mathcal{L}_{fine} \\ =& \lambda_{rec}\big[\mathcal{L}_{\mathbf{KL}}(q_\phi(\mathbf{z_g}|\mathbf{Y}),\, \mathcal{N}(\mathbf{0},\mathbf{I})) + \mathcal{L}_{\mathbf{CD}}(\mathbf{Y}'_r,\mathbf{Y})\big] \\ &+\lambda_{com}\big[\mathcal{L}_{\mathbf{KL}}(p_\psi(\mathbf{z_g}|\mathbf{X}),\, q_\phi(\mathbf{z_g}|\mathbf{Y})) + \mathcal{L}_{\mathbf{CD}}(\mathbf{Y}'_c,\mathbf{Y})\big] \\ &+\lambda_{fine}\mathcal{L}_{\mathbf{CD}}(\mathbf{Y}'_f,\mathbf{Y}),\end{aligned} \tag{8}$$

where λ_f, λ_r and λ_c are the weighting parameters.

4. Multi-View Partial Point Cloud Dataset

Towards an effort to build a more unified and comprehensive dataset for incomplete point clouds, we contribute the MVP dataset, a high-quality multi-view partial point cloud dataset, to the community. We compare the MVP dataset to previous partial point cloud benchmarks, PCN [29] and Completion3D [21] in Table 1. The MVP dataset has many advantages over the other datasets.

Diversity & Uniform Views. First, the MVP dataset consists of diverse partial point clouds. Instead of rendering partial shapes by using randomly selected camera poses [29, 21], we select 26 camera poses that are uniformly distributed on a unit sphere for each CAD model (Fig. 5 (a)). Notably, the relative poses between our 26 camera poses are fixed, but the first camera pose is randomly selected, which is equivalent to performing a random rotation to all 26 camera poses. The major advantages of using uniformly distributed camera views are threefold: **1)** The MVP dataset has fewer similar rendered partial 3D shapes than the other datasets. **2)** Its partial point clouds rendered by uniformly distributed camera views can cover most parts of a complete 3D shape. **3)** We can generate sufficient incomplete-complete 3D shape pairs with a relatively small number of 3D CAD models. According to Tatarchenko et. al. [20], many 3D reconstruction methods rely primarily on shape recognition; they essentially perform shape retrieval from the massive training data. Hence, using fewer complete shapes during training can better evaluate the capability of generating complete 3D shapes conditioned on the partial

Table 2: Shape completion results (CD loss multiplied by 10^4) on our multi-view partial point cloud dataset (16,384 points). VRCNet outperforms all existing methods by convincing margins. Note that besides the conventional 8 categories in existing datasets, MVP allows evaluation on 8 additional categories.

Method	*airplane*	*cabinet*	*car*	*chair*	*lamp*	*sofa*	*table*	*watercraft*	*bed*	*bench*	*bookshelf*	*bus*	*guitar*	*motorbike*	*pistol*	*skateboard*	Avg.
PCN [29]	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 [21]	2.72	4.25	3.40	7.95	17.01	6.04	7.42	6.04	11.60	5.62	8.22	2.37	1.33	3.90	3.97	2.09	6.36
MSN [14]	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. [23]	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 [15]	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 [27]	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 [30]	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 (Ours)	**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**	**3.06**

observation, rather than naively retrieving a known similar complete shape. An example of 26 rendered partial point clouds are shown in Fig. 5 (b).

Large-Scale & High-Resolution. Second, the MVP dataset consists of over 100,000 high-quality incomplete and complete point clouds. Previous methods render incomplete point clouds by using small virtual camera resolutions (*e.g.* 160 $\times$ 120), which is much smaller than real depth cameras (*e.g.* both Kinect V2 and Intel RealSense are 1920 $\times$ 1080). Consequently, the rendered partial point clouds are unrealistic. In contrast, we use the resolution 1600 $\times$ 1200 to render partial 3D shapes of high quality (Fig. 5 (c)). For ground truth, we employ Poisson Disk Sampling (PDS) [1, 8] to sample non-overlapped and uniformly spaced points for complete shapes (Fig. 5 (d)). PDS yields smoother complete point clouds than uniform sampling, making them a better representation of the underlying object CAD models. Hence, we can better evaluate network capabilities of recovering high-quality geometric details. Previous datasets provide complete shapes with only one resolution. Unlike those datasets, we create complete point clouds with different resolutions, including 2048(1x), 4096(2x), 8192(4x) and 16384(8x) for precisely evaluating the completion quality at different resolutions.

Rich Categories. Third, the MVP dataset consists of 16 shape categories of partial and complete shapes for training and testing. Besides the 8 categories (airplane, cabinet, car, chair, lamp, sofa, table and watercraft) included in previous datasets [29, 21], we add another 8 categories (bed, bench, bookshelf, bus, guitar, motorbike, pistol and skateboard). By using more categories of shapes, it becomes more challenging to train and evaluate networks on the MVP dataset.

To sum up, our MVP dataset consists of a large number of high-quality synthetic partial scans for 3D CAD models, which imitates real-scanned incomplete point clouds caused by self-occlusion. Besides 3D shape completion, our MVP dataset can be used in many other partial point cloud tasks, such as classification, registration and keypoints extraction. Compared to previous partial point cloud datasets, MVP dataset has many favorable properties. More detailed comparisons between our dataset and previous datasets are reported in our supplementary materials.

5. Experiments

Evaluation Metrics. In line with previous methods [21, 27, 30], we evaluate the reconstruction accuracy by computing the Chamfer Distance (Eq. (7)) between the predicted complete shapes $\mathbf{Y}'$ and the ground truth shapes $\mathbf{Y}$. Based on the insight that CD can be misleading due to its sensitivity to outliers [20], we also use F-score [10] to evaluate the distance between object surfaces, which is defined as the harmonic mean between precision and recall.

Implementation Details. Our networks are implemented using PyTorch. We train our models using the Adam optimizer [9] with initial learning rate $1e^{-4}$ (decayed by 0.7 every 40 epochs) and batch size 32 by NVIDIA TITAN Xp GPU. Note that VRCNet does not use any symmetry tricks, such as reflection symmetry or mirror operations.

5.1. Shape Completion on Our MVP Dataset

Quantitative Evaluation. As introduced in Sec. 4, our MVP dataset consists of 16 categories of high-quality partial/complete point clouds that are generated by CAD models selected from the ShapeNet [26] dataset. We split our models into a training set (62,400 shape pairs) and a test set (41,600 shape pairs). Note that none of the complete shapes in our test set are included in our training set. To achieve a fair comparison, we train all methods using the same training strategy on our MVP dataset. The evaluated CD loss and F-score for all evaluated methods (16,384 points) are reported in Table 2 and Table 3, respectively. VRCNet outperforms all existing competitive methods in terms of CD and F-score@1%. Moreover, VRCNet can generate complete point clouds with various resolutions (N = 2048, 4096, 8192 and 16384). We compare our methods with existing

Table 3: Shape completion results (F-Score@1%) on our multi-view partial (MVP) point cloud dataset (16,384 points).

Method	*airplane*	*cabinet*	*car*	*chair*	*lamp*	*sofa*	*table*	*watercraft*	*bed*	*bench*	*bookshelf*	*bus*	*guitar*	*motorbike*	*pistol*	*skateboard*	Avg.
PCN [29]	0.816	0.614	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 [21]	0.789	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 [14]	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. [23]	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 [15]	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 [27]	0.853	0.578	0.646	0.635	0.710	0.580	0.690	0.723	0.586	0.765	0.635	0.682	0.865	0.736	0.787	0.850	0.692
NSFA [30]	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 (Ours)	**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**

Table 4: Shape completion results (CD loss multiplied by 10^4) with various resolutions.

# Points	2,048		4,096		8,192		16,384	
	CD	F1	CD	F1	CD	F1	CD	F1
PCN [29]	9.77	0.320	7.96	0.458	6.99	0.563	6.02	0.638
TopNet [21]	10.11	0.308	8.20	0.440	7.00	0.533	6.36	0.601
MSN [14]	7.90	0.432	6.17	0.585	5.42	0.659	4.90	0.710
Wang et. al. [23]	7.25	0.434	5.83	0.569	4.90	0.680	4.30	0.740
ECG [15]	6.64	0.476	5.41	0.585	4.18	0.690	3.58	0.753
VRCNet (Ours)	**5.96**	**0.499**	**4.70**	**0.636**	**3.64**	**0.727**	**3.12**	**0.791**

Table 5: Ablation studies (2,048 points) for the proposed network modules, including Point Self-Attention Kernel, Dual-path Architecture and Point Selective Kernel Module.

Point Self-Attention	Dual-path Architecture	Kernel Selection	CD	F1
			6.64	0.476
✓			6.35	0.484
✓	✓		6.15	0.492
✓	✓	✓	5.96	0.499

approaches that support multi-resolution completion in Table 4, and VRCNet outperforms all the other methods.

Qualitative Evaluation. The qualitative comparison results are shown in Fig. 6. The proposed VRCNet can generate better complete shapes with fine details than the other methods. In particular, we can clearly observe the learned relational structures in our complete shapes. For example, the missing legs of the chairs (the second row and the fourth row in Fig. 6) are recovered based on the observed legs with the learned shape symmetry. In the third row of Fig. 6, we reconstruct the incomplete lamp base with a smooth round bowl shape, which makes it a more plausible completion than the others. The partially observed motorbike in the last row does not contain its front wheel, and VRCNet reconstructs a complete wheel by learning the observed back wheel. Consequently, VRCNet can effectively reconstruct complete shapes by learning structural relations, including geometrical symmetries, regular arrangements and surface smoothness, from the incomplete point cloud.

Table 6: Shape completion results (CD loss multiplied by 10^4) on the Completion3D benchmark (2,048 points). Our VRCNet outperforms all SOTAs by significant margins.

Method	*airplane*	*cabinet*	*car*	*chair*	*lamp*	*sofa*	*table*	*watercraft*	Avg.
AtlasNet [5]	10.36	23.40	13.40	24.16	20.24	20.82	17.52	11.62	17.77
PCN [29]	9.79	22.70	12.43	25.14	22.72	20.26	20.27	11.73	18.22
TopNet [21]	7.32	18.77	12.88	19.82	14.60	16.29	14.89	8.82	14.25
GRNet [27]	6.13	16.90	8.27	12.23	10.22	14.93	10.08	5.86	10.64
VRCNet (Ours)	**3.94**	**10.93**	**6.44**	**9.32**	**8.32**	**11.35**	**8.60**	**5.78**	**8.12**

Ablation Study. The ablation studies for all our proposed modules, Point Self-Attention Kernel (PSA), Dual-path Architecture and Kernel Selection (two-branch PSK), are presented in Table 5. We use ECG [15] as our baseline model and evaluate the completion results with 2048 points. By adding the proposed modules, better completion results can be achieved, which validates their effectiveness.

5.2. Shape Completion on Completion3D

The Completion3D benchmark is an online platform for evaluating 3D shape completion approaches. Following their instructions, we train VRCNet using their prepared training data and upload our best completion results (2,048 points). As reported in the online leaderboard[1], also shown in Table 6, VRCNet significantly outperforms SOTA methods and is ranked first on the Completion3D benchmark.

5.3. Shape Completion on Real-world Partial Scans

We further evaluate VRCNet (trained on MVP with all categories) on real scans, including cars from the KITTI [3] dataset, chairs and tables from the ScanNet dataset [2]. It is noteworthy that the KITTI dataset captured point clouds by using a LiDAR whereas the ScanNet dataset uses a depth camera. For sparse LiDAR data, we fine-tune all trained models on ShapeNet-car dataset, but no fine-tuning is needed for chairs and tables. The qualitative completion results are shown in Fig. 7. For those sparse point clouds of cars, VRCNet can predict complete and smooth surfaces

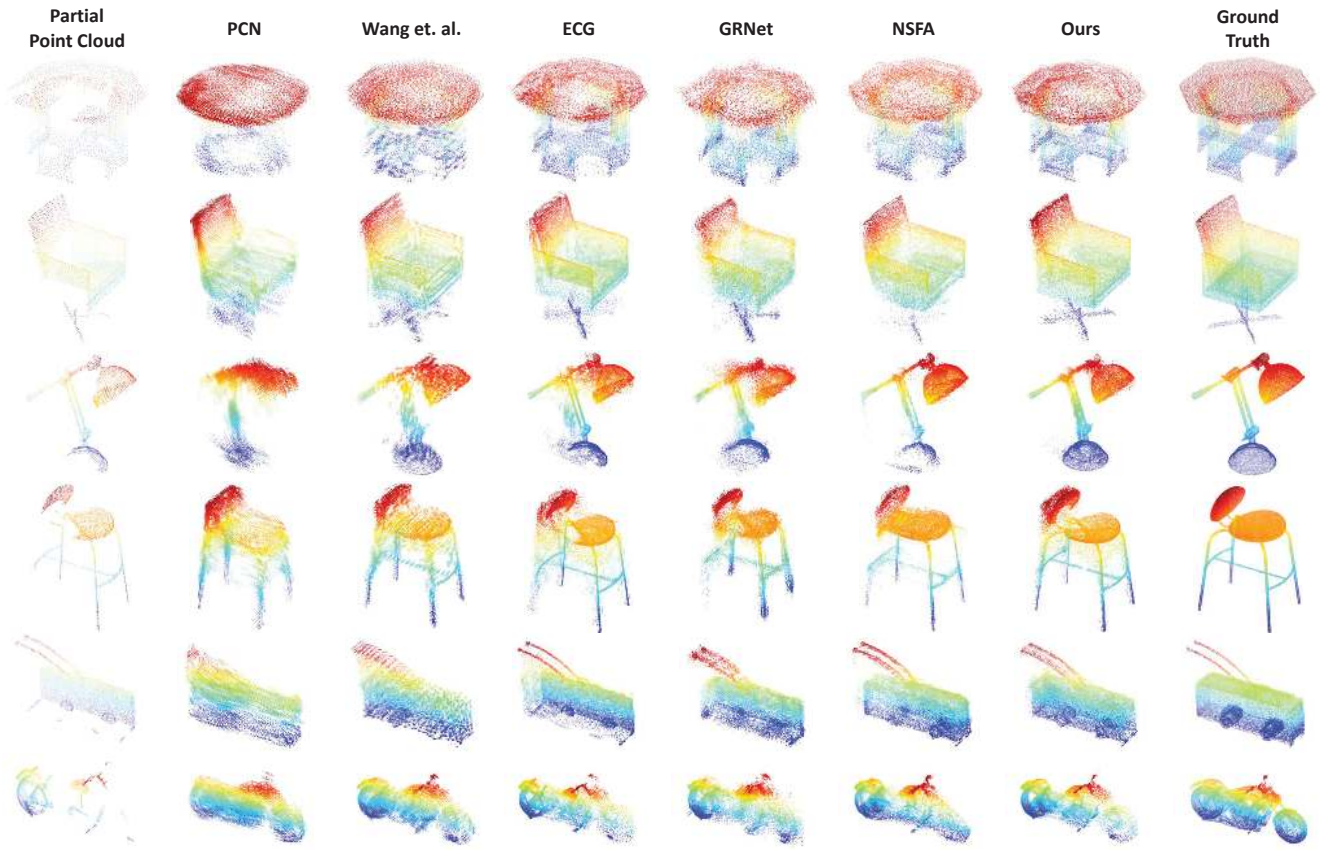


Figure 6: Qualitative completion results (16,384 points) on the MVP dataset by different methods. VRCNet can generate better complete point clouds than the other methods by learning geometrical symmetries.

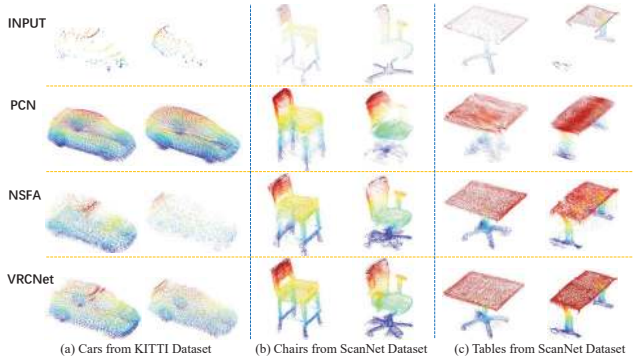


Figure 7: VRCNet generates impressive complete shapes for real-scanned point clouds by learning and predicting shape symmetries. (a) shows completion results for cars from Kitti dataset [3]. (b) and (c) show completion results for chairs and tables from ScanNet dataset [2], respectively.

that also preserves the observed shape details. In comparison, PCN [29] suffers a loss of fine shape details and NSFA [30] cannot generate high-quality complete shapes due to large missing ratios. For those incomplete chairs and tables, VRCNetgenerates appealing complete point clouds by exploiting the shape symmetries in the partial scans.

User Study. We conduct a user study on completion results for real-scanned point clouds by PCN, NSFA and VRCNet, where our VRCNet is the most preferred method overall. More details are reported in our supplementary materials.

6. Conclusion

In this paper, we propose VRCNet, a variational relational point completion network, which effectively exploits 3D structural relations to predict complete shapes. Novel self-attention modules, such as PSA and PSK, are proposed for adaptively learning point cloud features, which can be conveniently used in other point cloud tasks. In addition, we contribute a large-scale MVP dataset, which consists of over 100,000 high-quality 3D point clouds. We highly encourage researchers to use our proposed novel modules and the MVP dataset for future studies on partial point clouds.

Acknowledgement

This research was conducted in collaboration with SenseTime. This work is supported by NTU NAP and A*STAR through the Industry Alignment Fund - Industry Collaboration Projects Grant.

References

[1] Robert Bridson. Fast poisson disk sampling in arbitrary dimensions. *SIGGRAPH sketches*, 10:1, 2007.

[2] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017.

[3] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[4] Fabian Groh, Patrick Wieschollek, and Hendrik Lensch. Flex-convolution (deep learning beyond grid-worlds). *arXiv preprint arXiv:1803.07289*, 2018.

[5] 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*, pages 216–224, 2018.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[7] Han Hu, Zheng Zhang, Zhenda Xie, and Stephen Lin. Local relation networks for image recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3464–3473, 2019.

[8] Michael Kazhdan and Hugues Hoppe. Screened poisson surface reconstruction. *ACM Transactions on Graphics (ToG)*, 32(3):1–13, 2013.

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

[10] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics (ToG)*, 36(4):1–13, 2017.

[11] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

[12] Xiang Li, Wenhai Wang, Xiaolin Hu, and Jian Yang. Selective kernel networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 510–519, 2019.

[13] Yangyan Li, Rui Bu, Mingchao Sun, and Baoquan Chen. Pointcnn. *arXiv preprint arXiv:1801.07791*, 2018.

[14] 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*, volume 34, pages 11596–11603, 2020.

[15] Liang Pan. Ecg: Edge-aware point cloud completion with graph convolution. *IEEE Robotics and Automation Letters*, 2020.

[16] Liang Pan, Chee-Meng Chew, and Gim Hee Lee. Pointatrousgraph: Deep hierarchical encoder-decoder with atrous convolution for point clouds. *arXiv preprint arXiv:1907.09798*, 2019.

[17] Niki Parmar, Prajit Ramachandran, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jon Shlens. Stand-alone self-attention in vision models. In *Advances in Neural Information Processing Systems*, pages 68–80, 2019.

[18] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in Neural Information Processing Systems*, pages 5099–5108, 2017.

[19] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[20] Maxim Tatarchenko, Stephan R Richter, René Ranftl, Zhuwen Li, Vladlen Koltun, and Thomas Brox. What do single-view 3d reconstruction networks learn? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3405–3414, 2019.

[21] 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*, pages 383–392, 2019.

[22] Hugues Thomas, Charles R Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, François Goulette, and Leonidas J Guibas. Kpconv: Flexible and deformable convolution for point clouds. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6411–6420, 2019.

[23] Xiaogang Wang, Marcelo H Ang Jr, and Gim Hee Lee. Cascaded refinement network for point cloud completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 790–799, 2020.

[24] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. *Acm Transactions On Graphics (tog)*, 38(5):1–12, 2019.

[25] Wenxuan Wu, Zhongang Qi, and Li Fuxin. Pointconv: Deep convolutional networks on 3d point clouds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9621–9630, 2019.

[26] 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*, pages 1912–1920, 2015.

[27] Haozhe Xie, Hongxun Yao, Shangchen Zhou, Jiageng Mao, Shengping Zhang, and Wenxiu Sun. Grnet: Gridding residual network for dense point cloud completion. *arXiv preprint arXiv:2006.03761*, 2020.

[28] 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*, pages 206–215, 2018.

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

[30] Wenxiao Zhang, Qingan Yan, and Chunxia Xiao. Detail preserved point cloud completion via separated feature aggregation. *arXiv preprint arXiv:2007.02374*, 2020.

[31] Hengshuang Zhao, Jiaya Jia, and Vladlen Koltun. Exploring self-attention for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10076–10085, 2020.

[32] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. Pluralistic image completion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1438–1447, 2019.