

# Style-based Point Generator with Adversarial Rendering for Point Cloud Completion

Chulin Xie<sup>1\*</sup> Chuxin Wang<sup>2\*</sup> Bo Zhang<sup>3</sup> Hao Yang<sup>3</sup> Dong Chen<sup>3</sup> Fang Wen<sup>3</sup>

<sup>1</sup>University of Illinois at Urbana-Champaign <sup>2</sup>University of Science and Technology of China

<sup>3</sup>Microsoft Research Asia

chulinx2@illinois.edu

wcx0602@mail.ustc.edu.cn

{zhanbo, haya, doch, fangwen}@microsoft.com

## Abstract

*In this paper, we proposed a novel Style-based Point Generator with Adversarial Rendering (SpareNet) for point cloud completion. Firstly, we present the channel-attentive EdgeConv to fully exploit the local structures as well as the global shape in point features. Secondly, we observe that the concatenation manner used by vanilla foldings limits its potential of generating a complex and faithful shape. Enlightened by the success of StyleGAN, we regard the shape feature as style code that modulates the normalization layers during the folding, which considerably enhances its capability. Thirdly, we realize that existing point supervisions, e.g., Chamfer Distance or Earth Mover’s Distance, cannot faithfully reflect the perceptual quality of the reconstructed points. To address this, we propose to project the completed points to depth maps with a differentiable renderer and apply adversarial training to advocate the perceptual realism under different viewpoints. Comprehensive experiments on ShapeNet and KITTI prove the effectiveness of our method, which achieves state-of-the-art quantitative performance while offering superior visual quality.*

## 1. Introduction

As the 3D scanning devices such as depth camera and LiDAR become ubiquitous, point clouds get easier to acquire and have recently attracted a surge of research interest in the vision and robotics community. However, raw points directly captured by those devices are usually sparse and incomplete due to the limited sensor resolution and occlusions. Hence, it is essential to infer the complete shape from the partial observation so as to facilitate various downstream tasks [12] such as classification and shape manipulation as required in real-world applications.

\*Equal contribution. Authors did this work during the internship at Microsoft Research Asia.

Due to the irregularity and unorderedness of point clouds, one workaround is to leverage intermediate representations, e.g., depth map [15] or voxels [35], that are more amenable to neural networks. However, the representation transform may result in information loss, so the detailed structures can not be well preserved. With the emergence of point-based networks [25, 26, 29, 32, 12], predominant methods [37, 39, 11, 33, 21, 28, 5, 31, 17, 41] nowadays digest the partial inputs directly and estimate the complete point clouds in an end-to-end manner. These methods typically follow the encoder-decoder paradigm and adopt permutation invariant losses [8], e.g., Chamfer Distance or Earth Mover’s Distance, for regressing the groundtruth.

Ideally, the point completion network should simultaneously meet the following needs: 1) The output is desired to faithfully preserve the detailed structures of the partial input; 2) The network has a strong imaginative power to infer the global shape from the partial clue; 3) the local structure should be sharp, accurate, and free from the corruption by the noise points. Nonetheless, existing methods fail to achieve the above goals because of the neglecting of the global context during the feature extraction, the insufficient capability of modeling the structural details, and the lack of perceptual metrics for measuring the visual quality.

In this paper, we propose *Style-based Point generator with Adversarial Rendering*, i.e., the SpareNet, to circumvent the above issues. We have made improvements from encoder, generator, and loss function, and proposed 3 new modules: channel-attentive EdgeConv, Style-based Point Generator, and Adversarial Point Rendering. Firstly, while previous works employ PointNet or PointNet++ to learn point-wise or local features, we propose channel-attentive EdgeConv (Section 3.1), which not only considers the local information within the  $k$ -nearest neighbors but also wisely leverages the global context by aggregating the global features and weighting the feature channel attention for each point accordingly. The fusion of local and global context

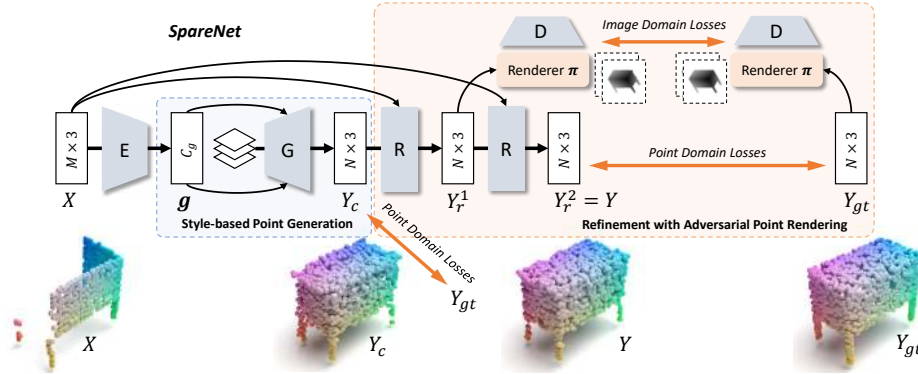


Figure 1: The architecture of SpareNet. An encoder E encodes the partial points  $X$  into a shape code  $g$ , leveraged by a style-based generator  $G$  to synthesize a coarse completion  $Y_c$ , which is recurrently improved with refiner  $R$  into the final result  $Y$ . Adversarial point rendering is applied to advocate the perceptual realism of completed points under different views.

enriches the learnt representation, so the network is more powerful to characterize fine structures of the input.

Further, we claim that the vanilla folding module [37] in conventional methods, which outputs the 3D shapes by morphing the 2D surfaces through multilayer perceptrons (MLP), has limited modeling capability due to the improper usage of the features, *i.e.*, the features are tiled and concatenated to each location of the 2D lattice. Drawn by the success of StyleGAN [19] in image synthesis, we boost the folding capability by regarding the learnt features as style codes, which can be used to modulate the feature normalization within the folding MLPs. The resulting style-based generator, as elaborated in Section 3.2, shows considerably improved capability of modeling structural details.

Last but not least, in order to generate visually-pleasing results, we propose to project the generated point clouds to view images (Section 3.3), whose realism is further examined by adversarial discriminators. Since the renderer we use is differentiable, the gradient from the discriminators will guide the network to learn the completion with high perceptual quality when viewed at different angles. We conduct extensive experiments on ShapeNet [4] and KITTI [9] datasets, and our SpareNet performs favorably over state-of-the-art methods both quantitatively and qualitatively.

## 2. Related Work

**Point cloud processing.** Pioneer works [25, 40] propose max-pooling to aggregate the features of individual points to ensure permutation invariance. Such aggregation, however, neglects the contextual relationships among different points, resulting in the representation incapable to characterize the fine structures. To amend this, [26] hierarchically groups the points and extracts features for the local context. Inspired by the tremendous success of 2D convolution, a surge of conv-based methods [16, 36, 10, 3, 20, 29, 32] has recently emerged, which generalizes the convolution to ir-

regular coordinate space. Meanwhile, graph-based methods [32, 30] regard the point clouds as graph structures where points are treated as nodes, and their local connectivity is denoted by edges. In [32], the EdgeConv is proposed to process the  $k$ -nearest neighbor graph and dynamically models the locality according to not only the coordinate distance but also the semantic affinity. While these networks are powerful to characterize the local structure, they fail to simultaneously consider the global shape, *i.e.*, the local feature extraction is unaware of the global information.

**Point cloud reconstruction.** To hallucinate the complete 3D coordinates, previous works design point decoders in various forms: multilayer perceptrons (MLPs) [5], hierarchical structures like a tree [28] or a multi-level pyramid [17], or through iterative refinement [31].

While some works resort to proxy representations other than points for shape completion, like volumetric grids [6, 35] or depth maps [15], FoldingNet [37] uses 2D manifolds to represent 3D point clouds, which also inspires others to model the target point clouds as non-linear foldings of 2D grids. PCN [39] predicts the folding of a single 2D patch, whereas AtlasNet [11] and MSN [21] generate the output with multiple patches. Recently, SA-Net [33] proposes a hierarchical folding that progressively hallucinates the detailed structures. Instead of improving the folding mechanism, SFA-Net [41] addresses the information loss during the global feature extraction. Albeit effective, these folding-based methods feed the partial features to folding modules via concatenation, but we claim that such a concatenation manner would impair the capacity of the foldings. In comparison, our style-based point generator greatly enhances the capacity for modeling structural details.

## 3. The SpareNet

The overview architecture of SpareNet is exhibited in Figure 1. Given a partial and low-res point cloud  $X$  as in-

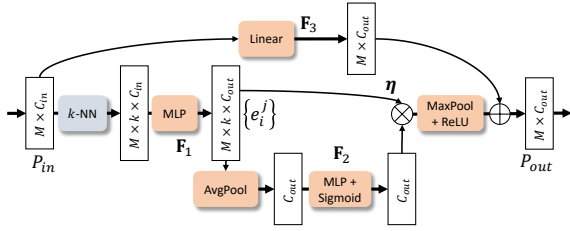


Figure 2: The structure of the Channel-Attentive EdgeConv.

put, SpareNet first completes  $X$  with a coarse point cloud  $Y_c$  through an encoder and a generator: the encoder  $E$  embeds  $X$  into a shape code  $\mathbf{g}$ , the style-based generator  $G$  exploits the shape code  $\mathbf{g}$  and synthesizes the coarse output  $Y_c$ . In addition, SpareNet adopts a refinement part that employs adversarial point rendering to further refine the coarse points  $Y_c$  and outputs a final complete and high-res point cloud  $Y$  with improved visual quality.

### 3.1. Channel-Attentive EdgeConv

We devise the *Channel-Attentive EdgeConv* (CAE) to simultaneously integrate both local and global context from point features. It is inspired by the EdgeConv [32] that captures a local context and the Squeeze-and-Excitation blocks [14] for capturing a global context. Our point encoder  $E$  heavily relies on the CAE blocks.

Let  $P_{in}$  be the input of a CAE block. Suppose it has  $M$  points with feature dimension  $C_{in}$ . For each point  $p_i \in P_{in}$ , we first find its  $k$ -nearest neighbors in  $P_{in}$  with respect to the Euclidean distance defined in the  $C_{in}$ -dimensional feature space ( $k = 8$  in experiments). Denote these neighbors as  $\{q_i^j, 1 \leq j \leq k\}$ , we have  $k$  directional edges on  $p_i$ , with each edge represented as  $(p_i, q_i^j - p_i)$ . Then we use a multilayer perceptron (MLP) (denoted as  $\mathbf{F}_1$ ) to compute a new feature  $e_i^j = \mathbf{F}_1(p_i, q_i^j - p_i)$  from each edge.

In order to leverage a global context from the  $k$ -NN graph, we feed the global average of all edge features  $\{e_i^j, 1 \leq i \leq M, 1 \leq j \leq k\}$  into a second MLP (denoted as  $\mathbf{F}_2$ ) to calculate a gating vector  $\eta$ :

$$\eta = \sigma \circ \mathbf{F}_2 \left[ \frac{1}{kM} \times \sum_{i,j} e_i^j \right], \quad (1)$$

where  $\sigma$  represents *sigmoid*. We re-calibrate every edge feature  $e_i^j$  by multiplying it with  $\eta$ . Finally, for each point, we reduce its  $k$  edge features into a new point feature through maximum pooling and ReLU activation. We also add an additional linear layer  $\mathbf{F}_3$  that skip-connects the output with the input to make the block residual.

Unlike the T-net presented in [25] that predicts an affine transformation to apply on point coordinates, we predict a gating vector that applies on edge features, in order to leverage both the local and global context for feature activation.

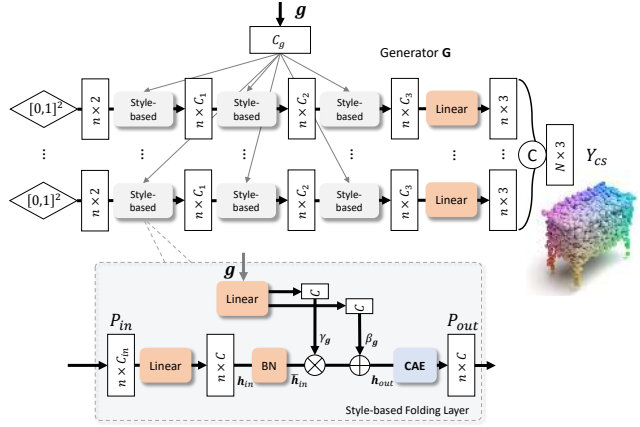


Figure 3: The styled-based folding and the generator  $G$ .

Our point encoder  $E$  is built with four sequential CAE blocks. The input  $X$  is fed into the first one and passes through all the rest, resulting in four point features with different dimensions and different receptive fields. We concatenate them together and compress the feature dimension through a final MLP. The output shape code  $\mathbf{g}$  is derived as a concatenation of two global poolings of the above result: a maximum pooling and an average pooling. It has a dimension of  $C_g$ , which is set to 4,096 in our experiments.

### 3.2. Style-based Point Generator

We present a *style-based point generator*  $G$  to generate a completed point cloud from the shape code  $\mathbf{g}$  through novel style-based folding layers. In previous folding methods [37, 39, 11, 21], the shape code is tiled and concatenated with  $N$  2D coordinates all sampled from a unit square  $[0, 1]^2$ . They learn a mapping from such combination into the 3D space using MLP, to emulate the morphing of a 2D grid into a 3D surface. Under such foldings, the shape code determines the morphing through the input concatenations ahead but hardly affects all the layers behind in an effective way. Such concatenation-based folding induces bottleneck that limits its capacity to represent different 3D surfaces.

Enlightened by the success of StyleGAN [19] in image generation, we propose style-based folding to circumvent these disadvantages in point cloud generation. Figure 3 exhibits the structure of our generator  $G$ . Instead of combining the shape code  $\mathbf{g}$  with grid coordinates as an input to vanilla folding MLPs, we directly inject  $\mathbf{g}$  into the generator  $G$ 's internal layers, to ensure a more extensive information aggregation in point cloud synthesis. We design novel style-based folding layers to accomplish such injection.

A style-based folding layer transforms the input point features  $P_{in}$  into new point features  $P_{out}$  under the modulation of the shape code  $\mathbf{g}$ . In specific, let  $\mathbf{h}_{in} \in \mathbb{R}^{B \times M \times C}$  be a mini-batch of point activations that are linearly transformed from  $P_{in}$ , with  $B$  being the batch size,  $M$  the num-

ber of points and  $C$  the dimension of point features. We first normalize  $\mathbf{h}_{in}$  to be  $\bar{\mathbf{h}}_{in} \in \mathbb{R}^{B \times M \times C}$  batch-wisely:

$$\bar{\mathbf{h}}_{in} = \frac{\mathbf{h}_{in} - \mu_{\mathbf{h}_{in}}}{\sigma_{\mathbf{h}_{in}}}, \quad (2)$$

with  $\mu_{\mathbf{h}_{in}}, \sigma_{\mathbf{h}_{in}} \in \mathbb{R}^{1 \times 1 \times C}$  being the means and standard deviations of  $\mathbf{h}_{in}$ 's channel-wise activations. In order to integrate the shape code, we compute new activation  $\mathbf{h}_{out}$  by denormalizing the normalized  $\bar{\mathbf{h}}_{in}$  according to the shape code  $\mathbf{g}$ , with the formulation

$$\mathbf{h}_{out} = \gamma_{\mathbf{g}} \otimes \bar{\mathbf{h}}_{in} + \beta_{\mathbf{g}}, \quad (3)$$

where  $\gamma_{\mathbf{g}}$  and  $\beta_{\mathbf{g}}$  are two modulation parameters both transformed from  $\mathbf{g}$  through linear layers. Finally we append a CAE block to  $\mathbf{h}_{out}$  to compute the output  $P_{out}$ .

Same with [11, 21], our generator  $G$  employs  $K$  (32 in experiments) surface elements to form a complex shape, as depicted in Figure 3. For each surface element, the generator learns a mapping from a unit square  $[0, 1]^2$  to a 3D surface through three sequential style-based folding layers and one linear layer. We sample  $n = N/K$  points for each surface element. Finally, the  $K$  3D surfaces (each surface represented as  $n = N/K$  points) are directly merged together forming the coarse output point cloud  $Y_c$  with  $N$  points.

### 3.3. Adversarial Point Rendering

After generating a coarse point cloud  $Y_c$ , we additionally refine it to acquire a final result  $Y$  with improved quality. But unlike some previous works [21, 31], our refinement guarantees an advantageous visual quality of our final point cloud  $Y$  by employing a novel *adversarial point rendering*.

By *point rendering*, we mean a fully differentiable point renderer that enables end-to-end rendering from 3D point cloud to 2D depth maps. The renderer makes it possible that we can supervise the training not only in the point domain but also in the image domain. By *adversarial*, we mean to adversarially improve the point cloud quality with a discriminator  $D$ , which not directly discriminates the 3D point clouds, but their rendered 2D depth maps. Observing the success of image-based convolutional networks, we believe that an image-based convolutional discriminator, combined with our differentiable point renderer, can better capture geometric details in point clouds than the point-based discriminators used by [31, 5].

**Renderer** Let  $P$  be a point cloud and  $v$  be a camera pose, the point renderer aims to generate a 2D depth map  $I^v(P)$  whose pixels reflect  $P$ 's geometry that is visible in  $v$ .

As shown in Figure 4, we start the rendering pipeline by projecting every 3D point  $p = (p_x, p_y, p_z) \in P$  onto a projection plane, retrieving a 2D pixel location  $(\hat{p}_x^v, \hat{p}_y^v)$  and a depth  $\hat{p}_z^v$  (a.k.a. distance from the projection). We calculate such projection with a projective transformation  $T_v$ , which is derived from camera pose  $v$  by combining both its extrinsic and intrinsic parameters.

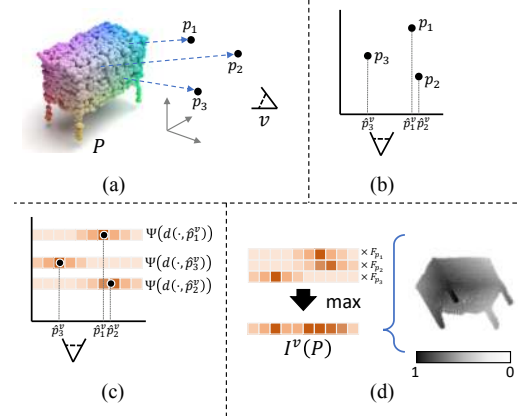


Figure 4: Pipeline of our differentiable point rendering. (a) Given 3D points  $P$  and a camera view  $v$ , (b) the 3D points  $p_i \in P$  are firstly projected as 2D points  $\hat{p}_i^v$  with depths according to  $v$ . (c) We regard each 2D point as a smooth density function modeled by kernel  $\Psi$ . (d) A depth map  $I^v(P)$  is generated through a pixel-wise maximum reduction of negated point depths  $F$  weighted by point densities.

We then rasterize these 2D points  $\hat{p}^v = (\hat{p}_x^v, \hat{p}_y^v)$  with depths  $\hat{p}_z^v$  to generate a rendered image. To guarantee differentiability, we regard each point not as a hard image pixel, but like a density function that spread smooth influence around its center. Let  $F_p$  be the point feature used for rendering, the rasterization can be formulated as

$$I_{x,y}^v(P) = \max_{p \in P} \{ \Psi(\|(x, y), \hat{p}^v\|_2) \times F_p, 0 \}, \quad (4)$$

where  $\Psi(x) = \exp(-x^2/2\rho^2)$  is a Gaussian-shape kernel that models the density function of a point, with  $\rho$  being a hyper-parameter that controls the radius of its influence. In order to render  $I^v(P)$  as a depth map, we define  $F_p = 1 - (\hat{p}_z^v - \min_{p \in P} \hat{p}_z^v) / (\max_{p \in P} \hat{p}_z^v - \min_{p \in P} \hat{p}_z^v)$  as a negative point depth normalized within  $[0, 1]$ . Choosing a different  $F$  for rendering is also supported, our implementation is able to render any point features without limits.

Comparing with previous differentiable point renderers [18, 38, 2, 34, 7], ours is much simpler yet effective for rendering depth maps. Unlike [7], our renderer does not attach additional parameters for training. We don't need to perform z-buffer sorting like [34, 18], since the maximum reduction among negative depths can automatically locate the nearest point. More importantly, unlike [2] where point coordinates are fixed, our renderer is fully differentiable: it supports gradients to be back-propagated not only to the feature  $F_p$ , but also to the 2D coordinates  $(\hat{p}_x^v, \hat{p}_y^v)$ .

To reduce information loss in rendering, we further propose the *multi-view point renderer*  $\pi$ , which utilizes our differentiable point renderer to simultaneously render a point cloud  $P$  into eight depth maps, each observed from a different viewpoint. The resulting  $\pi(P)$  with shape  $H \times W \times 8$  is an ordered concatenation of all eight depth maps in the

channel dimension. The size  $H \times W$  is set to  $256 \times 256$  in experiments; the eight viewpoints are set as the eight corners of a cube:  $[\pm 1, \pm 1, \pm 1]$ , to cover a wide angle of observation. Unlike the multi-view depth maps used by [15], our multi-view depth maps are rendered in a differentiable way: gradients can be back-propagated from depths maps on to the rendered points. Hence, our renderer enables end-to-end training with both image and point supervisions, which considerably promotes the perceptual quality of results.

**Refiner** Our refiner R shares similar structure with [21]; they both consist of a minimum density sampling and a residual network that resembles PointNet [25]. But different from [21], we add CAE blocks to the residual network for enhanced capability. Moreover, we recurrently deploy the refiner twice upon the coarse result  $Y_c$  to get a first and a second refining result,  $Y_r^1$  and  $Y_r^2$ , with  $Y = Y_r^2$  being the final output, as illustrated by Figure 1.

**Discriminator** We render three point clouds during training: the partial input  $X$ , the groundtruth  $Y_{gt}$  and the first refined result  $Y_r^1$ . Our discriminator D utilizes a cGAN [23] strategy: the real sample is a concatenation of  $\pi(Y_{gt})$  and  $\pi(X)$  in channel dimension; the fake sample is a concatenation of  $\pi(Y_r^1)$  and  $\pi(X)$ . This makes the adversarial update of  $Y_r^1$  to be conditioned on the input  $X$ . We implement D as a sequence of 2D convolution layers with spectral normalizations [24] and LeakyReLU activations.

### 3.4. Training Losses

During training, the *reconstruction loss*  $\mathcal{L}_{rec}$  is required to match the output point cloud to the ground-truth. Even though the Chamfer Distance (CD) is very popular among existing works due to its efficiency in computation, we follow [21] and implement  $\mathcal{L}_{rec}$  with the Earth Mover’s Distance (EMD) instead, which is more faithful to visual quality as verified by [8, 1, 21]. The  $\mathcal{L}_{rec}$  supervises both the coarse output  $Y_c$  and the final output  $Y$  with

$$\mathcal{L}_{rec} = d_{EMD}(Y_c, Y_{gt}) + d_{EMD}(Y, Y_{gt}). \quad (5)$$

The *fidelity loss*  $\mathcal{L}_{fd}$  is employed to preserve structures of the input  $X$  within the output  $Y$  as

$$\mathcal{L}_{fd} = \frac{1}{|X|} \sum_{p \in X} \min_{q \in Y} \|p - q\|_2^2. \quad (6)$$

We also introduce losses in the image domain. We impose the *depth map matching loss*  $\mathcal{L}_{depth}$  as a  $\mathcal{L}$ -1 distance on the multi-view depth maps. We also adopt the *feature matching loss*  $\mathcal{L}_{fea}$  as a  $\mathcal{L}$ -2 distance on the discriminator features. These two losses are formulated as

$$\mathcal{L}_{depth} = \frac{1}{8HW} \|\pi(Y_r^1), \pi(Y_{gt})\|_1, \quad (7)$$

$$\mathcal{L}_{fea} = \sum_i \frac{\alpha_i}{H_i W_i D_i} \|D_i[\pi(Y_r^1)], D_i[\pi(Y_{gt})]\|_2^2, \quad (8)$$

with  $D_{1 \leq i \leq 4}$  being features extracted from intermediate layers of discriminator D,  $H_i W_i D_i$  the feature shape,  $\alpha_i =$

Method	plane	cabinet	car	chair	lamp	sofa	table	vessel	avg
PointFCAE	1.554	2.631	2.132	2.954	4.067	2.997	2.899	2.619	2.732
FoldingNet	1.682	2.576	2.183	2.847	3.062	3.003	2.500	2.357	2.526
AtlasNet	1.324	2.582	2.085	2.442	2.718	2.829	2.160	2.114	2.282
PCN*	2.426	<b>1.888</b>	2.744	2.200	2.383	<b>2.062</b>	<b>1.242</b>	2.208	2.144
MSN	1.334	2.251	2.062	2.346	2.449	2.712	1.977	2.001	2.142
GRNet	1.376	2.128	1.918	2.127	2.150	2.468	1.852	1.876	1.987
<i>Ours</i>	<b>1.131</b>	2.014	<b>1.783</b>	<b>2.050</b>	<b>2.063</b>	2.333	1.729	<b>1.790</b>	<b>1.862</b>

Table 1: Completion comparison on ShapeNet in terms of EMD  $\times 10^3$  (lower is better).

Method	plane	cabinet	car	chair	lamp	sofa	table	vessel	avg
PointFCAE	1.424	3.878	0.519	2.404	6.989	2.594	2.673	8.998	3.683
FoldingNet	1.593	5.918	0.649	1.355	4.344	2.400	2.243	5.508	3.001
AtlasNet	0.512	2.536	0.706	1.181	2.295	2.460	1.810	2.475	1.747
PCN*	0.484	1.221	0.200	1.417	0.947	1.680	<b>0.566</b>	0.926	0.930
MSN	0.324	1.858	0.412	<b>0.968</b>	1.652	2.409	0.917	0.744	1.161
GRNet	4.649	1.002	1.152	1.712	2.977	2.717	1.713	5.528	2.681
<i>Ours</i>	<b>0.307</b>	<b>0.691</b>	<b>0.142</b>	1.113	<b>0.774</b>	<b>0.945</b>	0.668	<b>0.523</b>	<b>0.645</b>

Table 2: Completion comparison on ShapeNet in terms of FPD  $\times 0.1$  (lower is better).

$D_i / \sum_i^4 D_i$  a re-weighting factor of each feature. The final loss for the end-to-end SpareNet training combines all individual losses as

$$\mathcal{L} = w_{rec}\mathcal{L}_{rec} + w_{fd}\mathcal{L}_{fd} + w_{depth}\mathcal{L}_{depth} + w_{fea}\mathcal{L}_{fea} + w_{adv}\mathcal{L}_{adv} + w_{exp}\mathcal{L}_{exp}, \quad (9)$$

with  $\mathcal{L}_{adv}$  the adversarial loss from the discriminator,  $\mathcal{L}_{exp}$  an expansion loss also used by [21]. We follow [22] and implement  $\mathcal{L}_{adv}$  as mean square in GAN training. The loss weights are set as  $w_{rec} = 200, w_{fd} = 0.5, w_{adv} = 0.1, w_{depth} = w_{fea} = 1, w_{exp} = 0.1$ .

## 4. Experiments

During training, we use the Adam optimizer with  $\beta_1 = 0$  and  $\beta_2 = 0.9$ . Following TTUR [13], we set imbalanced initial learning rates,  $1 \times 10^{-4}$  for the generator and  $4 \times 10^{-4}$  for the discriminator. We train the network for 200 epochs, with the learning rates decayed by 0.1 at 100 and 150 epochs. With the batch size 32, it takes 5 days for training on the ShapeNet dataset with 4 Tesla V100 GPUs.

### 4.1. Datasets

We conduct experiments on two datasets commonly used for point cloud completion: ShapeNet [4] and KITTI [9].

\*The PCN scores are calculated from a released PCN[39] model that is trained on a fully sized ShapeNet, which is 8 times larger than the training set of all the other models reported in Tables 1, 2, 3. Even though, our model demonstrates superiority over it in terms of both EMD and FPD.

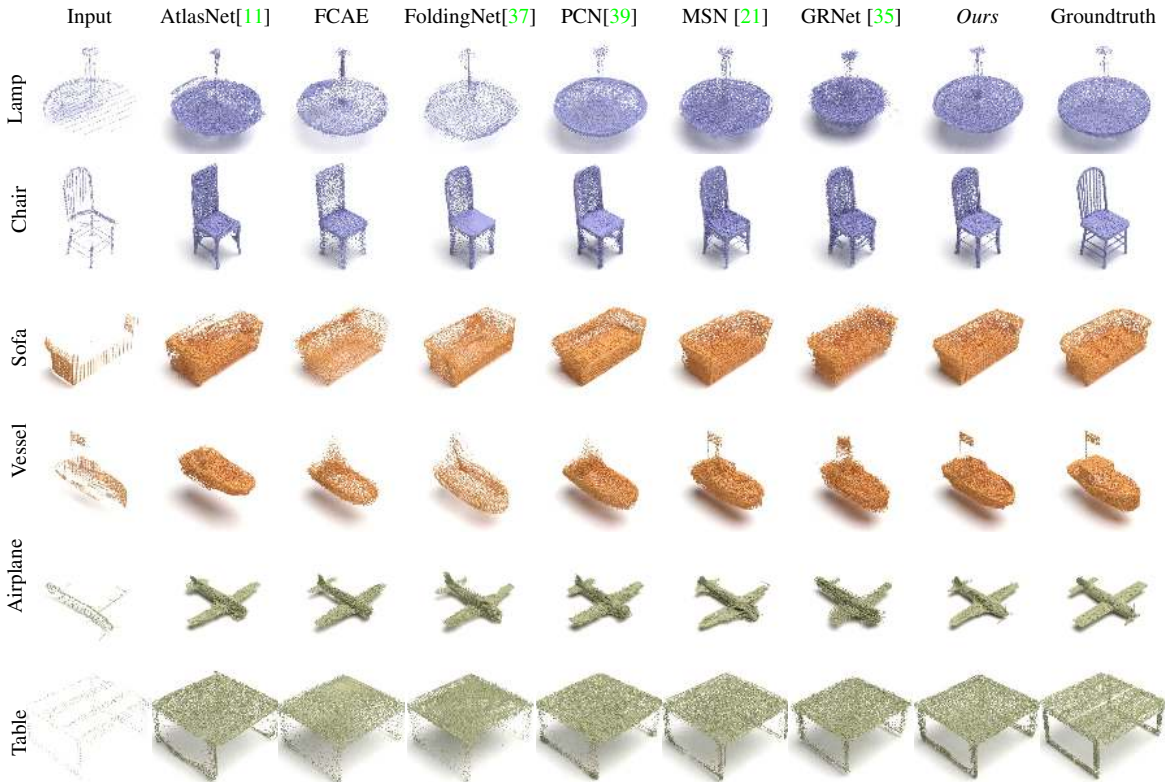


Figure 5: Visualized completion comparison on ShapeNet.

Method	plane	cabinet	car	chair	lamp	sofa	table	vessel	avg
PointFCAE	0.340	1.170	0.480	1.240	2.294	1.312	1.433	0.893	1.145
FoldingNet	0.622	1.608	0.619	1.553	2.025	1.543	1.534	0.910	1.302
AtlasNet	0.301	0.967	0.440	0.858	1.126	1.174	0.813	0.639	0.790
PCN*	0.559	<b>0.389</b>	0.581	<b>0.466</b>	0.684	<b>0.263</b>	<b>0.156</b>	0.395	<b>0.437</b>
MSN	0.252	0.974	0.445	0.770	0.933	1.152	0.669	0.491	0.711
GRNet	0.293	0.560	0.363	0.583	0.690	0.935	0.532	0.389	0.543
Ours	<b>0.176</b>	0.664	<b>0.362</b>	0.616	<b>0.631</b>	0.789	0.498	<b>0.384</b>	0.515

Table 3: Completion comparison on ShapeNet in terms of  $CD \times 10^3$  (lower is better).

**ShapeNet.** The ShapeNet dataset derived from [39] consists of 30,974 3D models that belong to 8 categories: airplane, cabinet, car, chair, lamp, sofa, table, and vessel. Each groundtruth point cloud comprises 16,384 points that are uniformly sampled from the corresponding 3D model. The partial point clouds are constructed by projecting the 2.5D depth maps of the model back into 3D. This leads to 8 different partial point clouds for each groundtruth in the training set. Following [35, 31], we randomly select only one partial point cloud from the 8 to build our training pairs. This results in a training set only 1/8 the size of the one used by PCN [39]. For a fair comparison, we use the same train/val/test splits as [35, 31].

**KITTI.** The KITTI dataset comprises a sequence of real LiDAR scans. For each frame, the cars are extracted accord-

ing to the labeled bounding box, resulting in 2,401 point partial inputs. Since no groundtruth exist in this dataset, we cannot rely on paired metrics for evaluation.

## 4.2. Comparison on ShapeNet

**Metrics.** We employ three evaluation metrics to measure completion accuracy on ShapeNet: the *Chamfer Distance* (CD), the *Earth Mover’s Distance* (EMD) and the *Fréchet Point cloud Distance* (FPD). Among these distances, CD is a widely used one due to its computation efficiency. EMD is more discriminative to the local details and density distribution. Previous works [21, 8, 1] have claimed EMD to be a more reliable measure for visual quality than CD. The FPD [27], inspired by the FID [13], is a metric that computes the Fréchet distance between Gaussian fitted distributions. It evaluates not the accuracy of an individual point cloud, but the overall perceptual quality of all predicted point clouds.

**Completion evaluation.** We compare SpareNet with the following state-of-the-art approaches: 1) PointFCAE, a simple autoencoder adopting PointNet as encoder and MLPs as decoder; 2) AtlasNet [11], which generates points by sampling from the parametric surface elements; 3) FoldingNet [37], which proposes folding-based decoder to directly generate points; 4) PCN [39] that uses stacked PointNet for feature extraction; 5) MSN [21], the baseline of this work; 6) GRNet [35], a recent leading approach that oper-

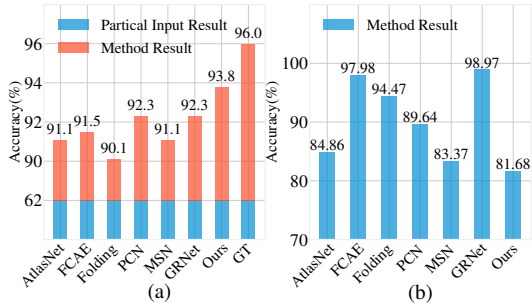


Figure 6: (a) Comparing informativeness. Higher is better. (b) Comparing fake detection accuracies. Lower is better.

Method	Atlasnet	PCN	Folding	TopNet	MSN	GRNet	Ours
Consistency↓	0.700	1.557	1.053	0.568	1.951	0.313	<b>0.249</b>
Fidelity↓	1.759	2.235	7.467	5.354	<b>0.434</b>	0.816	1.461
MMD↓	2.108	1.366	0.537	0.636	2.259	0.568	<b>0.368</b>

Table 4: Quantitative comparison on KITTI dataset in terms of consistency, fidelity and minimum matching distance (MMD). The best results are highlighted in bold.

ates on volumetric grids. For a fair comparison, we train all the models using the EMD loss with their released codes. All models utilize the same training set except PCN\*.

Tables 1 and 3 show that our method outperforms prior works in terms of both EMD and CD metric. In particular, our method is more advantageous in terms of the FPD score as shown in Table 2, reducing the FPD from the second-best 0.930 to 0.645. The qualitative results in Figure 5 further corroborate the perceptual advantage of our method. In comparison, our method is able to generate fine structures with shapes whereas other methods are prone to give blurry results. Thin structures, such as the mast of the vessel, can also be faithfully generated. Overall, our completed result appears much less noisy and visually pleasing.

**Comparison via classification.** To better quantify the completion quality, we propose two novel metrics for comparison. First, we employ a pretrained PointNet model and test its classification accuracy on the completed point clouds by different methods. A pretrained classifier should perform better when the input is more resemblance to the real point clouds. In Figure 6 (a), the PointNet achieves the highest classification accuracy when using our completed results as input, showing that our method can well preserve the semantics of the partial input and is able to complete the points as the real ones. Besides, we compare the perceptual quality of various methods by examining their rendered view images. As we do not have the labeling of quality scores, we train an image classifier (ResNet-18) to differentiate the fake or real data based on the complete results of all seven methods, and then make use of this fake detector to test each method. Here we assume that different meth-

Ablation		EMD	CD	FPD
[A]	w/o $\eta$ in CAE	2.060	0.606	1.012
[B]	w/o EdgeConv in CAE	1.972	0.542	0.692
[C]	w/o Style-based Folding	3.274	1.779	3.111
[D]	w/o Recurrent Refine	2.184	0.708	1.856
<i>Ours Full</i>		<b>1.862</b>	<b>0.515</b>	<b>0.645</b>

Table 5: Ablation results of SpareNet.

ods share similar types of artifacts and the fake detection trained on one method is generalizable to other approaches. Figure 6 (b) shows that our method can better mislead the fake detector by giving a lower detection accuracy, proving that our methods suffer from less noticeable artifacts when observed at different viewpoints.

### 4.3. Comparison on KITTI

This dataset contains the LiDAR scans from auto-driving scenes, and the objects are much sparser than the ShapeNet dataset. For better transferred performance, we finetune all the models using the ShapeNetCars subset that only contains the cars from the ShapeNet so that the prior knowledge for this category can be better leveraged.

Since the groundtruth point clouds are not available in this real-world dataset, we cannot apply the full-reference metrics (CD, EMD and FPD) for evaluation. As such, we follow the practice in [39], and use the following metrics: 1) the *Temporal Consistency* that measures the Chamfer distance for the consecutive frames; 2) the *Fidelity*, a single-directional Chamfer distance that measures how the input structures are preserved in the output; 3) the *Minimum Matching Distance* (MMD), the CD between the output and the point cloud in ShapeNetCars that is closest to the output in terms of CD. This metric makes sure that the output resembles a car model in ShapeNet. Table 4 shows the quantitative results. Our method shows superior consistency and MMD, indicating that our method can complete the real point points with high stability and quality. Yet, MSN and GRNet outperform in fidelity, possibly because our adversarial rendering that enforces high perceptual quality brings a slight sacrifice of the fidelity relative to the input.

### 4.4. Ablation Study

We report the results of extensive ablation experiments in Table 5, with each row corresponding to one ablation setting. We observe that the ablation settings all cause the drop of performance (*i.e.* raise of the three distance numbers) comparing to our full model shown at the bottom. It in turn exhibits the effectiveness of each ablated component we propose. For example, in [A], the structure that computes and applies the channel attention vector  $\eta$  (Equation 1) is removed from CAE blocks. The raise of metric numbers reflects the importance of our proposed channel-

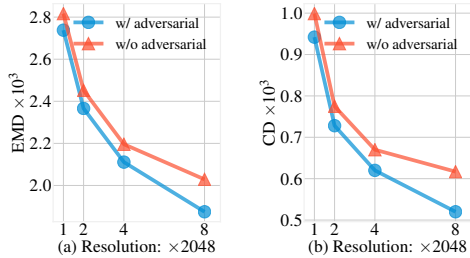


Figure 7: Ablation of adversarial rendering under different point resolutions. Our adversarial rendering demonstrates larger improvement when predicting denser point clouds.

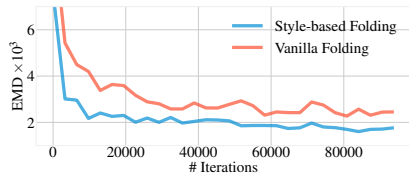


Figure 8: Loss curves of training two models: one uses the vanilla folding (red); the other applies style-based folding (blue). The latter converges much faster than the former.

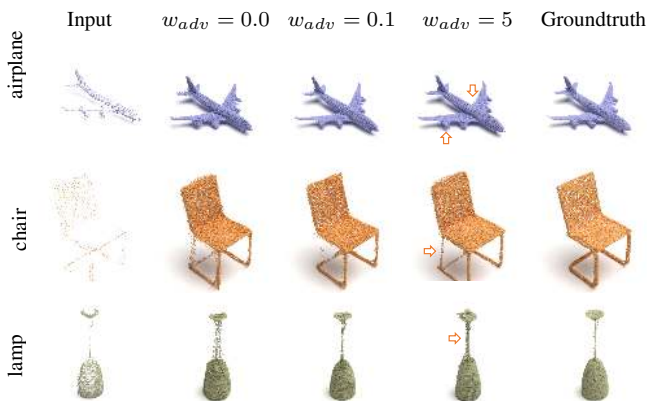


Figure 9: Results with different adversarial loss weights.

attentive functionality. On the other hand, the advantage of edge features over point features can be observed by comparing the full model with [B] where its CAE blocks abandon the EdgeConv structure but preserve the channel attention  $\eta$  on point features. We also ablate on the style-based generator in [C], where the generator  $G$  adopts the vanilla concatenation-based folding instead of our style-based folding. The large-margin performance drop indicates the criticalness of our style-based folding in point cloud completion. It is also verified by Figure 8, where the style-base folding leads to much faster convergence of training loss than the vanilla folding. In addition, we use a single-step refinement (with renderings applied on the final output) instead of the recurrent refinement and report results in [D], the performance drop indicates a necessity of recurrent refinement in our framework.

$\mathcal{Y}_{adv}$	EMD	CD	FPD	$w_{adv}$	EMD	CD	FPD
$\emptyset$	2.029	0.617	1.308	0	2.029	0.617	1.308
$\{Y_r^1\}$	<b>1.862</b>	<b>0.515</b>	<b>0.645</b>	0.1	<b>1.862</b>	<b>0.515</b>	<b>0.645</b>
$\{Y_r^2\}$	1.937	0.601	0.874	0.2	1.898	0.535	0.654
$\{Y_r^1, Y_r^2\}$	1.898	0.535	0.736	5	1.980	0.593	1.013

Table 6: Ablation of adversarial training settings. By default we apply discriminator only on  $Y_r^1$  with  $w_{adv} = 0.1$ .

More experiments are conducted to study the adversarial training settings. We compare the accuracies with or without adversarial training in terms of both EMD and CD in Figure 7. It shows that 1) adversarial rendering brings significant improvement under all point resolutions, from 2,048 to 16,384; 2) a denser predicted point cloud leads to a larger improvement. We believe it is due to the rendering quality: a denser point cloud renders smoother depth maps with less point scattering effects, aiding the discriminator to better capture the intrinsic geometry of point clouds.

Let  $\mathcal{Y}_{adv}$  be the point clouds which we apply adversarial rendering on; let  $w_{adv}$  be the weight of the adversarial loss. Different settings of  $\mathcal{Y}_{adv}$  and  $w_{adv}$  are exploited in Table 6 showing that: 1) when adversarial rendering involves the final output, *i.e.* when  $Y_r^2 \in \mathcal{Y}_{adv}$ , the completion accuracy is slightly impaired; 2) a large adversarial loss weight ( $w_{adv} = 5$ ) also hinders the completion accuracy. We deem this a conflict between the adversarial loss  $\mathcal{L}_{adv}$  and the reconstruction loss  $\mathcal{L}_{rec}$  (defined by Equation 5) when they are applied on the same output, or when their loss weights are comparable:  $\mathcal{L}_{rec}$  encourages the output to match the single corresponding groundtruth, whereas the  $\mathcal{L}_{adv}$  promotes an improved visual quality through the discriminator, which is learned from a larger distribution instead of a single groundtruth. This conflict is also visualized by Figure 9, where a larger adversarial loss weight  $w_{adv}$ , though helps forge some structures (*e.g.* the airplane engine, the chair leg) reasonable for the category, nevertheless often distracts the resulting shape from its true groundtruth.

## 5. Conclusion

This paper presents a novel framework, SpareNet, for point cloud completion. It comprises channel-attentive Edge-Convs for the fusion of local and global context in point feature extraction. It also performs style-based folding for an enhanced capability in point cloud synthesis. In addition, adversarial point rendering is adopted: by leveraging a fully differentiable point renderer, an image-based discriminator is utilized for capturing geometric details in point clouds. Extensive experiments on ShapeNet and KITTI verify the state-of-the-art performance of SpareNet, as well as the effectiveness of each proposed component.



## References

- [1] Panos Achlioptas, Olga Diamanti, Ioannis Mitliagkas, and Leonidas Guibas. Learning representations and generative models for 3d point clouds. In *International conference on machine learning*, pages 40–49. PMLR, 2018.
- [2] Kara-Ali Aliev, Artem Sevastopolsky, Maria Kolos, Dmitry Ulyanov, and Victor Lempitsky. Neural point-based graphics. *arXiv preprint arXiv:1906.08240*, 2019.
- [3] Matan Atzmon, Haggai Maron, and Yaron Lipman. Point convolutional neural networks by extension operators. *arXiv preprint arXiv:1803.10091*, 2018.
- [4] 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] Xuelin Chen, Baoquan Chen, and Niloy J Mitra. Unpaired point cloud completion on real scans using adversarial training. *arXiv preprint arXiv:1904.00069*, 2019.
- [6] 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*, pages 5868–5877, 2017.
- [7] Peng Dai, Yinda Zhang, Zhuwen Li, Shuaicheng Liu, and Bing Zeng. Neural point cloud rendering via multi-plane projection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7830–7839, 2020.
- [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*, pages 605–613, 2017.
- [9] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013.
- [10] Fabian Groh, Patrick Wieschollek, and Hendrik PA Lensch. Flex-convolution. In *Asian Conference on Computer Vision*, pages 105–122. Springer, 2018.
- [11] Thibault Groueix, Matthew Fisher, Vladimir G. Kim, Bryan Russell, and Mathieu Aubry. AtlasNet: A Papier-Mâché Approach to Learning 3D Surface Generation. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [12] 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*, 2020.
- [13] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in neural information processing systems*, pages 6626–6637, 2017.
- [14] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.
- [15] Tao Hu, Zhizhong Han, Abhinav Shrivastava, and Matthias Zwicker. Render4completion: Synthesizing multi-view depth maps for 3d shape completion. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019.
- [16] Binh-Son Hua, Minh-Khoi Tran, and Sai-Kit Yeung. Pointwise convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 984–993, 2018.
- [17] 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/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
- [18] Eldar Insafutdinov and Alexey Dosovitskiy. Unsupervised learning of shape and pose with differentiable point clouds. In *Advances in neural information processing systems*, pages 2802–2812, 2018.
- [19] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4401–4410, 2019.
- [20] Yangyan Li, Rui Bu, Mingchao Sun, Wei Wu, Xinhan Di, and Baoquan Chen. Pointcnn: Convolution on x-transformed points. In *Advances in neural information processing systems*, pages 820–830, 2018.
- [21] Minghua Liu, Lu Sheng, Sheng Yang, Jing Shao, and Shi-Min Hu. Morphing and sampling network for dense point cloud completion. *AAAI 2020*, 2019.
- [22] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802, 2017.
- [23] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.
- [24] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.
- [25] 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*, pages 652–660, 2017.
- [26] 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.
- [27] Dong Wook Shu, Sung Woo Park, and Junseok Kwon. 3d point cloud generative adversarial network based on tree structured graph convolutions. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3859–3868, 2019.
- [28] Lyne P. Tchapmi, Vineet Kosaraju, Hamid Rezatofighi, Ian Reid, and Silvio Savarese. Topnet: Structural point cloud decoder. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [29] 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.
- [30] Chu Wang, Babak Samari, and Kaleem Siddiqi. Local spectral graph convolution for point set feature learning. In *Proceedings of the European conference on computer vision (ECCV)*, pages 52–66, 2018.
- [31] Xiaogang Wang, Marcelo H. Ang Jr. , and Gim Hee Lee. Cascaded refinement network for point cloud completion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
- [32] 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)*, 2019.
- [33] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 1939–1948, 2020.
- [34] Olivia Wiles, Georgia Gkioxari, Richard Szeliski, and Justin Johnson. Synsin: End-to-end view synthesis from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7467–7477, 2020.
- [35] Haozhe Xie, Hongxun Yao, Shangchen Zhou, Jiageng Mao, Shengping Zhang, and Wenxiu Sun. Grnet: Gridding residual network for dense point cloud completion. In *ECCV*, 2020.
- [36] Yifan Xu, Tianqi Fan, Mingye Xu, Long Zeng, and Yu Qiao. Spidercnn: Deep learning on point sets with parameterized convolutional filters. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 87–102, 2018.
- [37] 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)*, June 2018.
- [38] Wang Yifan, Felice Serena, Shihao Wu, Cengiz Öztireli, and Olga Sorkine-Hornung. Differentiable surface splatting for point-based geometry processing. *ACM Transactions on Graphics (TOG)*, 38(6):1–14, 2019.
- [39] 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.
- [40] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander J Smola. Deep sets. In *Advances in neural information processing systems*, pages 3391–3401, 2017.
- [41] Wenxiao Zhang, Qingan Yan, and Chunxia Xiao. Detail preserved point cloud completion via separated feature aggregation, 2020.