

Cascaded Refinement Network for Point Cloud Completion

Xiaogang Wang Marcelo H Ang Jr Gim Hee Lee
National University of Singapore

xiaogangw@u.nus.edu {mpeangh, gimhee.lee}@nus.edu.sg

Abstract

Point clouds are often sparse and incomplete. Existing shape completion methods are incapable of generating details of objects or learning the complex point distributions. To this end, we propose a cascaded refinement network together with a coarse-to-fine strategy to synthesize the detailed object shapes. Considering the local details of partial input with the global shape information together, we can preserve the existing details in the incomplete point set and generate the missing parts with high fidelity. We also design a patch discriminator that guarantees every local area has the same pattern with the ground truth to learn the complicated point distribution. Quantitative and qualitative experiments on different datasets show that our method achieves superior results compared to existing state-of-the-art approaches on the 3D point cloud completion task. Our source code is available at <https://github.com/xiaogangw/cascaded-point-completion.git>.

1. Introduction

Despite the significant progress on image generation and translation [41, 16], synthesizing and generating 3D point clouds remains as a very challenging task due to the sparseness, incompleteness and irregularity of the points. More specifically, the inabilities of learning accurate point features and various point distributions make it difficult to obtain a complete and dense object shape. In this work, we focus on the point cloud completion [52, 36] task, which completes missing parts of the occluded object. 3D shape completion has wide applications such as robotic navigation [7, 26], scene understanding [17, 5] and augmented reality [2, 44].

Existing methods [52, 36, 3, 6, 22, 34] have shown promising results on shape completion for different inputs: distance fields, meshes, voxel grids and point clouds. Voxel representations are a direct generalization of pixels to the 3D case. However, generating 3D shape with voxel format suffers from memory inefficiency, hence it is difficult to obtain high-resolution results. Although data-driven methods

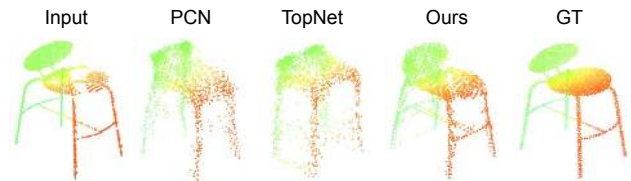


Figure 1. Our method can generate complete point clouds with finer details compared to existing state-of-the-art methods.

on mesh representations [38, 12, 42] are able to generate complicated surfaces, they are limited to the fixed vertex connection patterns. As a result, it is difficult to change the topology during the training process. In contrast, it is easy to add new points for point clouds and several studies have shown promising results. The pioneer work [52] proposes an encoder-decoder based pipeline on both the synthetic dataset ShapeNet [4] and the real scene dataset KITTI [9]. A following work TopNet [36] proposes a hierarchical rooted tree structure decoder to generate object shapes. Even though they have achieved impressive performances on shape completion, they are both unable to generate the detailed geometric structure of 3D objects and produce unsatisfactory coarse object outputs. Several approaches [24, 28, 25] propose to learn 3D structures in a function space, and they achieve impressive results for various input formats. However, these methods require post-processing to refine the outputs.

We propose to synthesize the dense and complete objects shapes in a cascaded refinement manner, and jointly optimize the reconstruction loss and an adversarial loss end-to-end. Our framework is designed to keep the object details in the partial inputs, and to produce realistic reconstructions of the missing parts. Figure 1 shows an example between our method and existing approaches [52, 36]. Although the legs of a chair are clearly present in the input, existing works are incapable of keeping this structural details in the outputs. On the contrary, our approach successfully captures this fine-grained details. To this end, we make a skip connection between the incomplete points and coarse outputs. However, simple concatenation between inputs and our coarse outputs give rise to unevenly distributed points. Consequently, we design an iterative refinement de-

coder together with a feature contraction and expansion unit to refine the point positions. We adopt an adversarial loss that penalizes inaccurate points from the ground truth to learn the complex point distributions and further improve the performance. Instead of classifying the whole object by predicting a single confidence value like conventional generative adversarial networks (GANs) [21, 46], we design a patch-based discriminator to explicitly force every local patch of generated point clouds to have the same pattern with real complete point clouds inspired by [16, 45]. We show state-of-the-art quantitative and qualitative results on different datasets by various experiments. Our key contributions are as follows:

- We propose a novel point cloud completion network which is able to preserve object details from partial points and generate missing parts with fine details at the same time;
- Our cascaded refinement strategy together with the coarse-to-fine pipeline can refine the points positions locally and globally;
- Experiments on different datasets show that our framework achieves superior results to existing methods on the 3D point cloud completion task.

2. Related work

In this section, we review existing works on point generation, upsampling and shape completion that are related to our task.

3D Generation. The pioneering work PointNet [29] proposed a method on point cloud analysis and inspired a large amount of works on point cloud generation. Early works [1, 37, 33] have proposed generative models by using GAN or variational auto-encoder (VAE) on 3D generation. Achlioptas et al. [1] proposed r-GAN for 3D point clouds generation, in which both generator and discriminator are fully connected layers. Valsesia et al. [37] proposed a graph neural network to synthesize object shapes. They calculated the adjacency matrix by the feature vectors from each vertex in each graph convolution layer. Despite their superior results, the calculation of the adjacency matrix requires quadratic computation complexity and consumes a lot of memory. The above methods successfully synthesize object shapes from noise. However, simple GANs or VAE can only generate small scale (1024 or 2048) point sets due to the complex point distribution and the notoriously difficult training of GANs. Although improved methods [47, 33, 48] show superior performance on synthesizing 3D objects, they are limited to synthesizing the general shapes of objects and are not suitable for shape completion.

3D Upsampling. Similar to point cloud completion, several works [51, 50, 49, 21, 45] aim at generating dense and uniform point clouds given sparse and non-uniform point sets. PU-Net [51] adopted the PointNet++ [30] as a backbone to extract point features and expand feature dimensions by a series of convolutions. Following PU-Net, EC-Net [50] generated sharp edges by penalizing the distance between points and edge labels. While they show exciting results, both methods are limited to upsampling point clouds by a small ratio (e.g. 4×). To alleviate this problem, Yifan et al. [49] introduced a hierarchical point feature extraction and multi-stage generation network and achieved 16× upsampling, but the training process consumes more computation memory. More importantly, they are all limited to upsampling the sparse points and are not applicable for completion tasks.

3D Completion. 3D shape completion plays an important role in robotics and perception, and has obtained significant development in recent years. Existing methods have shown impressive performance on various formats: voxel grids, meshes and point clouds. Inspired by 2D CNN operations, earlier works [6, 14, 35, 20] focus on the voxel and distance fields formats generation with 3D convolution. Several approaches [6, 35] have proposed a 3D encoder-decoder based network for shape completion and shown promising performance. However, voxel-based methods consume a large amount of memory and are unable to generate high-resolution outputs. To increase the resolution, several works [39, 40] have proposed to use the octree structure to gradually voxelize specific areas. However, due to the quantization effect of the voxelization operation, recent works gradually discard the voxel format and focus on the mesh reconstruction. Existing mesh representations [12, 38] are based on deforming a template mesh to a target mesh and hence not flexible to any typologies. In comparison to voxels and meshes, point clouds are easy to add new points during the training procedure. Yuan et al. [52] proposed the pioneering work PCN on point cloud completion, which was a simple encoder-decoder network to reconstruct dense and complete point set from an incomplete point cloud. They adopted the folding mechanism [48] to generate high resolution outputs (16,384). TopNet [36] proposed a hierarchical tree-structure network to generate point cloud without assuming any specific topology for the input point set. However, both PCN and TopNet are unable to synthesize the fine-grained details of 3D objects.

3. Our Method

3.1. Overview

Our objective is to produce complete and high-resolution 3D objects from corrupted and low resolution point clouds. Specifically, given the sparse incomplete point sets $P =$

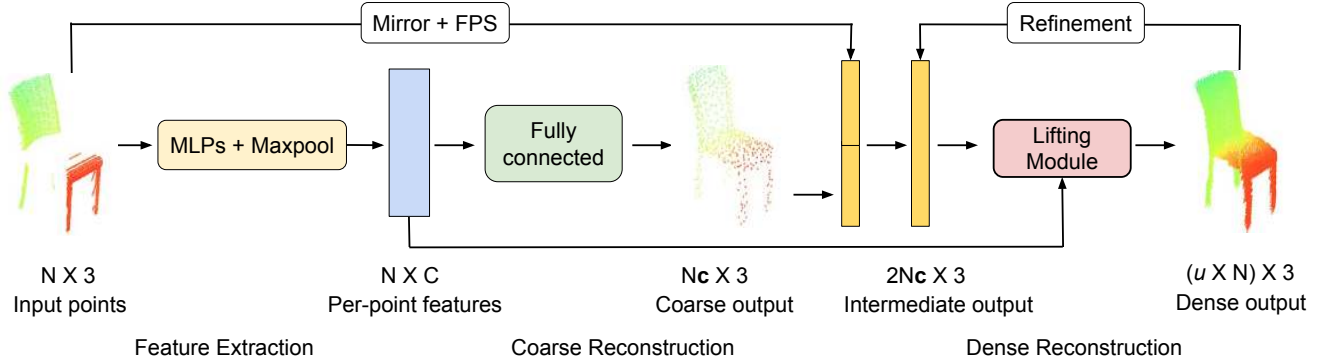


Figure 2. An illustration of our generator network. The generator includes three sub-networks: feature extraction, coarse reconstruction and dense reconstruction. The feature extractor consists of two MLPs and max-pooling layers. The coarse reconstruction comprises several fully-connected layers. The dense reconstruction is a cascaded refinement sub-network with a lifting module in each step. We generate dense and complete point clouds given partial and sparse inputs. μ is the upsampling factor.

$\{p_i\}_{i=1}^N$ of N points, we aim to generate a dense and complete point set $Q = \{q_i\}_{i=1}^{u \times N}$ of $u \times N$ points, where u is the upsampling scalar. We expect our method to fulfill three requirements: (1) preserve the fine details of the input point cloud P , (2) inpaint the missing parts with detailed geometric structures, and (3) generate evenly distributed points on object surfaces.

Our point cloud completion architecture is shown in Figure 2. Traditional GANs [11, 1, 37] map a noise distribution z to the data space, we extend the general GAN framework by modelling the generator G (Section 3.2) as a feature extraction encoder and a conditional coarse-to-fine decoder. The discriminator D (Section 3.5) aims to distinguish between the generated fake output and the ground truth.

3.2. Generator

Our generator G consists of three components: (1) feature extraction h , (2) coarse reconstruction g_1 and (3) dense reconstruction g_2 .

Feature Extraction. Same with PCN [52], we use two stacked PointNet feature extraction architecture with max-pooling operation to extract the global point features f . Specifically, the feature extractor h can be modelled by the composition of two functions expressed as:

$$f = h(P | w_h), \quad h = h_1 \circ h_2, \quad (1)$$

where w_h denotes the parameters of h_1 and h_2 , h_1 and h_2 represent the two extraction sub-networks, respectively.

Coarse Reconstruction. g_1 consists of several fully-connected layers, which maps the latent embedding f to the coarse point cloud. We denote the size of P_{coarse} as $N_c \times 3$. From Figure 2, we can observe that the coarse output roughly capture the complete object shape but loses fine details, which we aim to recover in the second stage.

Dense Reconstruction. Our second stage g_2 is a conditional iterative refinement sub-network. The synthesis begins at generating low resolution points (2048×3), and points with higher resolutions are then progressively refined. Following TopNet [36], our outputs have four resolutions: $N = \{2048, 4096, 8192, 16, 384\}$, for which the numbers of iterations are 1, 2, 3 and 4, respectively. Parameters are shared among each iteration.

Existing methods [52, 36, 33] exploit either folding based operations or tree structure to generate dense and complete objects. Although they have achieved impressive qualitative results, the fine details of the objects are often lost. As can be seen in Figure 1, both PCN [52] and TopNet [36] fail to generate the details of 3D objects (e.g. the legs of the chair). The reason is that the latent embedding f is obtained by the last max-pooling layer of the encoder, and it only represents the rough global shape, hence it is difficult to recover the detailed object structures. We propose to preserve the object shape details in the partial inputs and exploit the global shape information from P_{coarse} at the same time. Inspired by the skip-connection from U-Net [31], we concatenate the partial inputs with the global shape P_{coarse} to synthesize the dense points. However, direct concatenation resulted in a poor visual quality because of the serious uneven distributed points. To alleviate this problem, we propose to dynamically subsample $N_c \times 3$ points from the partial inputs P before concatenating with the coarse output P_{coarse} . We denote the combined point sets as P_S with the size of $2N_c \times 3$, which are fed into the lifting module (Section 3.3) to obtain a higher resolution points P_i . We use the efficient farthest point sampling (FPS) algorithm [30] to subsample points. We also design a feature contraction-expansion unit (Section 3.3) to refine the point positions gradually. We progressively refine the point positions and upsample the point size by a factor of two by the lifting module. For the subsequent iterates, the input for the lifting module is the intermediate output P_i from last step.

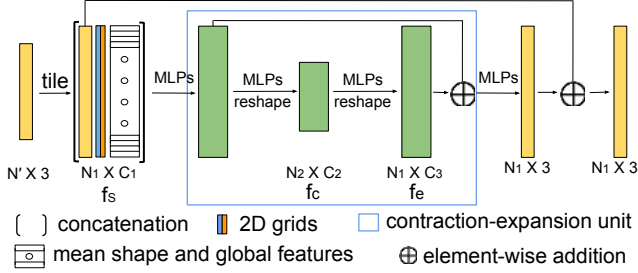


Figure 3. The architecture of the lifting module. The input is $N' \times 3$, and we upsample it by a factor of 2 to obtain the output of $N_1 \times 3$. The feature contraction and expansion unit predicts residual point features instead of final results.

3.3. Lifting Module

We design a lifting module to upsample the point size by a factor of two, and concurrently refine the point positions by the feature contraction and expansion unit. To upsample the point set P_S , we first tile the points P_S two times to obtain a new point set P'_S . Then we sample a unique 2D grid vector and append it after each point coordinates to increase the variations among the duplicated points [48]. We also utilize the mean shape prior f_m (Section 3.4) in our iterative refinement [18] to alleviate the domain gap of point features between the incomplete and complete point clouds. We concatenate the point P'_S , mean shape vectors f_m , global feature f and the sampled 2D grids to obtain a new feature f_s . We aim to predict per-vertex displacements $\{d_x, d_y, d_z\}$ for each point P'_S given the point feature f_s .

Feature Contraction-expansion Unit. Inspired by the hourglass network [27], we consolidate the local and global information by a bottom-up and top-down fashion to refine points positions and make them evenly distributed on object surfaces. However, it is not straightforward to subsample and upsample features between different scales for point clouds. Although some operations are introduced in PointNet++ [30] and graph convolution [53], they consume a large amount of memory and computation time, especially for high-resolution points. Consequently, we use shared multilayer perceptrons (MLPs) [51] to make feature contraction and expansion. Specifically, we assume the dimension of f_s to be $N_1 \times C_1$, and sizes of outputs features f_c and f_e are $N_2 \times C_2$ and $N_1 \times C_3$, respectively. The two operations are represented as $f_c = \mathcal{RS}(\mathcal{C}_C(f_s))$ and $f_e = \mathcal{RS}(\mathcal{C}_E(f_c))$, where $\mathcal{RS}(\cdot)$ is a reshaping operation. $\mathcal{C}_C(\cdot)$ and $\mathcal{C}_E(\cdot)$ are MLPs for contraction and expansion, respectively. Our lifting module predicts point feature residuals rather than the final output since deep neural networks are better at predicting residuals [38]. Our lifting module is shown in Figure 3.

Overall, in one-step refinement, the output point set P_1 is

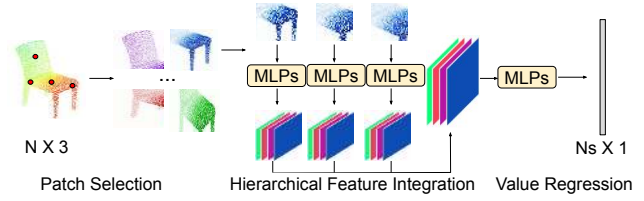


Figure 4. The discriminator architecture sub-network. It includes the patch selection, hierarchical feature integration and confidence value regression.

represented as:

$$P_i = F(P'_S) + P'_S, \quad (2)$$

where $F(\cdot)$ predicts per-vertex offsets by the lifting module for the input point P'_S .

3.4. Shape Priors

For each object class, we take mean values of latent embeddings from all the instances within that category as our mean shape vectors. The calculation is represented as:

$$f_m^i = \frac{1}{N^i} \sum_{i=1}^{i=N^i} f_i, \quad (3)$$

where N^i is the total number of objects from category i . The latent embeddings are obtained by a pre-trained PointNet auto-encoder¹ on eight object categories following [18].

Following 3DN [42], we mirror the partial input with respect to the xy -plane as we assume the reflection symmetry plane (xy -plane) of objects to be known since many man-made models show global reflection symmetry. Then we subsample N_c points from the mirrored points to obtain the point set P_C . Note that not all training objects are symmetric, and 40 of 1200 testing data are asymmetric. Our mirror operation can be seen as an initialization for the missing points and reasonable point positions are generated by the whole optimization.

3.5. Discriminator

To generate various realistic dense and complete point clouds, we adopt the adversarial training and jointly optimize the reconstruction loss and the adversarial loss end-to-end. Instead of only considering the global shape by regressing one single confidence value as conventional GANs [21], we design a patch discriminator to further guarantee that every local area is realistic. We employ the hierarchical point set feature learning in PointNet++ [30] with different radii to consider multi-scale local patches. Specifically, we first uniformly sample N_s point seeds by FPS, and then set three radii $\{0.1, 0.2, 0.4\}$ around the seeds to extract a set of local patch. Finally, we obtain N_s scores from

¹<https://github.com/charlesq34/pointnet-autoencoder>

the discriminator instead of calculating one single value for binary classification. Our discriminator consists of patch selection, hierarchical feature integration and value regression. The discriminator sub-network is shown in Figure 4.

3.6. Optimization

Our training loss comprises two components, a reconstruction loss to encourage the completed point cloud to be the same as the ground truth, and an adversarial loss to penalize the unrealistic outputs.

Reconstruction Loss. We adopt the Chamfer Distance (CD) [8] as our reconstruction loss, i.e.,

$$\begin{aligned} \text{CD}(X, Y) &= \mathcal{L}_{X,Y} + \mathcal{L}_{Y,X}, \text{ where} \\ \mathcal{L}_{X,Y} &= \frac{1}{|X|} \sum_{x \in X} \min_{y \in Y} \|x - y\|_2, \text{ and} \\ \mathcal{L}_{Y,X} &= \frac{1}{|Y|} \sum_{y \in Y} \min_{x \in X} \|x - y\|_2, \end{aligned} \quad (4)$$

which calculates the average closest point distance between two point clouds X and Y . There are two variants for CD which we denote as CD-P and CD-T. Specifically, CD-P = $(\sqrt{\mathcal{L}_{X,Y}} + \sqrt{\mathcal{L}_{Y,X}})/2$ and CD-T = $\mathcal{L}_{X,Y} + \mathcal{L}_{Y,X}$. We show different results with these two variants, and we adopt CD-P in all our experiments during training. Hence, our reconstruction loss can be expressed as:

$$\mathcal{L}_{\text{rec}} = \text{CD}(P_{\text{coarse}}, Q) + \lambda_f \text{CD}(P_{\text{fine}}, Q), \quad (5)$$

where P_{coarse} and P_{fine} correspond to the coarse output and fine output, respectively, and λ_f is the weight for the reconstruction loss of P_{fine} .

Adversarial Loss. We adopt the stable and efficient objective function of LS-GAN [23] for our adversarial losses. Specifically, the adversarial losses for the generator and discriminator are:

$$\mathcal{L}_{\text{GAN}}(G) = \frac{1}{2} [D(\tilde{x}) - 1]^2, \quad (6)$$

$$\mathcal{L}_{\text{GAN}}(D) = \frac{1}{2} [D(\tilde{x})^2 + (D(x) - 1)^2], \quad (7)$$

where \tilde{x} and x are the generated fake result and the target ground truth, respectively.

Overall Loss. Our overall loss function is the weighted sum of the reconstruction loss and the adversarial losses:

$$\mathcal{L} = \lambda \mathcal{L}_{\text{GAN}} + \beta \mathcal{L}_{\text{rec}}, \quad (8)$$

where λ and β are the weights for GAN loss and the reconstruction loss, respectively. During training, G and D are optimized alternatively.

4. Experiments

4.1. Evaluation Metrics

We compare our method with several existing methods 3D-EPN [6], PCN [52] and TopNet [36]. We use two evaluation metrics to evaluate results quantitatively. The first metric is the Chamfer Distance (CD) following [52, 36]. More specifically, we use CD-P for experiments in Section 4.4 and use CD-T in the remaining experiments for fair comparison. The other metric is Fréchet Point Cloud Distance (FPD) adopted from [33]. FPD calculates the 2-Wasserstein distance between the real and fake Gaussian measures in the feature spaces of the point sets:

$$\text{FPD}(X, Y) = \|m_X - m_Y\|_2^2 + \text{Tr}(\Sigma_X + \Sigma_Y - 2(\Sigma_X \Sigma_Y)^{\frac{1}{2}}), \quad (9)$$

where m and Σ represent the mean vector and covariance matrix of the points, respectively. $\text{Tr}(A)$ is the sum of the diagonal elements from matrix A . More evaluation details are shown in supplementary material.

4.2. Datasets

For a fair comparison, we evaluate our method on the datasets of PCN [52] and TopNet [36]. Partial inputs are obtained by back-projecting 2.5D depth images into 3D. 30,974 objects from eight categories are selected: airplane, cabinet, car, chair, lamp, sofa, table and vessel. We also create our smaller training dataset to measure the generalization ability on fewer training data. We only render the partial scans with one random virtual view instead of eight random views like PCN, hence the number of our training data is one eighth of PCN, but we keep the testing data the same with PCN. The resolutions of the partial and complete point clouds are 2048 in our created dataset following TopNet [36]. We use our testing data for evaluation when training on the dataset of TopNet.

4.3. Implementation Details

All our models are trained using the Adam [19] optimizer. We adopt the two time-scale update rule (TTUR) [15] and set learning rates for the generator and discriminator as 0.0001 and 0.00005, respectively. The learning rates are decayed by 0.7 after around every 40 epochs, and clipped by 10^{-6} . λ and β are set to 1 and 200, respectively. λ_f increases from 0.01 to 1 within the first 50,000 iterations. N_s in discriminator is 256. The size N_c of coarse output is 512. We train one single network for all eight categories of data.

4.4. Point Completion on the Dataset of PCN

Quantitative and qualitative results are shown in Table 1 and Figure 6. Point resolutions for the output and the

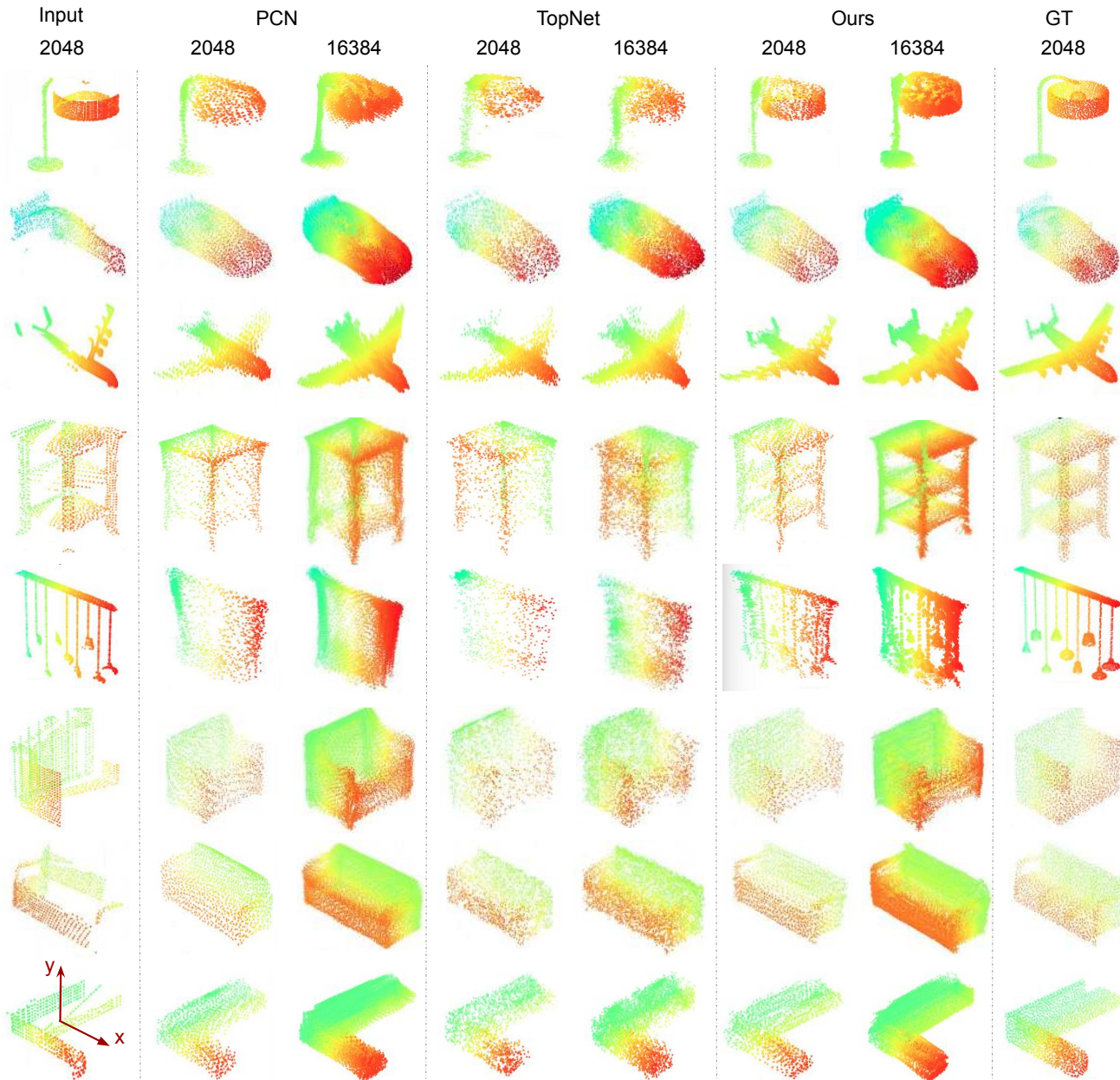


Figure 5. Qualitative comparison on our created ShapeNet dataset. The resolution for both partial and ground truth are 2048. We show the generated results of size 2048 and 16,384 from different methods.

ground truth are 16,384. The quantitative results in Table 1 show that we obtain the best performance on all categories of objects compared to other methods. We obtain 11.74% relative improvement on the average value compared to the second best method PCN. The results indicate that we achieve better performance with more accurate global shape and finer local structures. From Figure 6 we can observe that PCN and TopNet fail to recover the fine details such as legs of a chair and aircraft tails, while our method successfully generates such structures.

4.5. Point Completion on the Dataset of TopNet

In this experiment, we train our model on the training data from TopNet² and then test on our created testing data. Since we observe that object scales of the training data are larger than scales of the testing data, we adopt random scaling augmentation technique [36] during training for all methods and the scale values are uniformly sampled between $[1/1.5, 1]$. We can see that we achieve better quantitative results for all resolutions in Table 2.

²<https://github.com/lynetcha/completion3d>

Methods	Mean Chamfer Distance per point (10^{-3})								
	Avg	Airplane	Cabinet	Car	Chair	Lamp	Sofa	Table	Vessel
3D-EPN[6]	20.147	13.161	21.803	20.306	18.813	25.746	21.089	21.716	18.543
PCN-FC[52]	9.799	5.698	11.023	8.775	10.969	11.131	11.756	9.320	9.720
PCN[52]	9.636	5.502	10.625	8.696	10.998	11.339	11.676	8.590	9.665
TopNet [36]	9.890	6.235	11.628	9.833	11.498	9.366	12.347	9.362	8.851
Ours	8.505	4.794	9.968	8.311	9.492	8.940	10.685	7.805	8.045

Table 1. Quantitative comparison for point cloud completion on eight categories objects of ShapeNet.

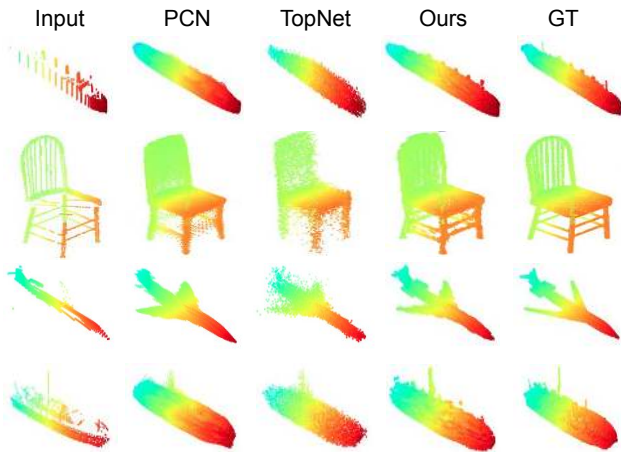


Figure 6. Qualitative comparison on the dataset of PCN. Point resolutions for the output and ground truth are 16,384.

Methods	Resolution			
	2048	4096	8192	16384
PCN [52]	9.36	8.17	7.28	6.28
TopNet [36]	10.23	8.85	7.47	6.64
Ours	7.61	6.57	5.72	5.21

Table 2. Quantitative comparison on the training data of TopNet.

4.6. Point Completion on Our Training Data

In this section, we show the results on our smaller training data. Quantitative and qualitative results are shown in Table 3 and Figure 5, respectively. As shown in Table 3, our method outperforms both PCN and TopNet on all resolutions. The relative improvements of our method compared to PCN are 16.08%, 12.97%, 15.36% and 15.56% for all resolutions on our smaller training data. The improvements on our smaller training data verify the robustness and generality of our method. We also generate 2048, 4096, 8192 and 16,384 resolution objects by training one single model on 16,384 points, and compare the results with that obtained from independent training of PCN and TopNet. We still achieve lower CD errors, which verifies the accuracy of our method.

We get three conclusions from the qualitative results in Figure 5: (1) Our method is able to generate the details not only included in the partial scan, but also for the missing parts, on both high and low resolutions. For example, the

Resolution	Methods			
	PCN [52]	TopNet [36]	Ours*	Ours
2048	9.02	9.88	8.03	7.57
4096	7.71	8.52	6.78	6.71
8192	6.90	7.56	5.98	5.84
16,384	6.17	6.60	5.21	5.21

Table 3. Quantitative results on our smaller training data. We take CD (10^{-4}) as evaluation. Ours* represents the results obtained by using the single model trained on the 16,384 resolution output.

lampshade (Row 1), the empennage of the car (Row 2), and the engines of the airplane (Row 3). While both PCN and TopNet miss the detailed structure and only obtain the general object shapes. (2) Our generated points are more evenly distributed. From the results of desk and chandelier (Row 4 and 5), we can see more points are located on the top surface of the desk and in the top left corner of the chandelier from PCN, while ours are evenly distributed on the object surface. (3) Although we mirror the partial input with respect to the xy -plane, our method does not memorize the mirrored points. As the results shown in the last row of Figure 5, our generated object is not symmetric with respect to xy -plane. This verifies that mirroring operation provides a initialization for the missing points, and accurate point deformations are estimated by our whole network. More results are shown in our supplementary material.

4.7. Robustness to Occlusion

To further test the robustness of the models, we manually occlude the partial inputs from testing dataset by p percent of points following PCN [52], and p ranges from 20% to 70% with a step of 10%. The quantitative results are shown in Table 4. Our method achieves the best performance, although the error increases gradually as more regions are occluded. This shows that our method is more robust to noise data. More qualitative results are shown in our supplementary material.

4.8. Point Completion for Classification

Following [32], we also measure the completion quality by calculating the classification accuracy on the synthesized complete point clouds. Specifically, we train one classification model by PointNet [29]. The upper bound (UP) is cal-

culated on the complete points from the testing data and the lower bound (LP) is calculated on the partial points from the testing data. The remaining values are obtained by evaluating the classification model on the generated outputs from different methods. The quantitative results are shown in Table 5. Clearly, the complete outputs provide higher accuracy because of the defects in the partial data. Our generated results improve the accuracy by 1.59% compared to PCN and TopNet, which demonstrates that our outputs are more realistic and our results preserve more accurate semantic information.

Methods	Occlusion ratios					
	20%	30%	40%	50%	60%	70%
PCN [52]	7.69	8.84	10.63	13.30	17.20	23.60
TopNet [36]	8.46	9.57	11.30	13.60	17.60	23.20
Ours	5.52	6.72	8.46	11.36	15.26	21.27

Table 4. Quantitative comparison for occluded point clouds under different occlusion rates. The evaluation metric is mean CD per point (10^{-4}).

Methods	LB	UB	PCN [52]	TopNet [36]	Ours
Acc. (%)	70.50	97.33	92.58	92.58	94.17

Table 5. Comparison of classification results among different methods. The upper bound (UB) represents the result tested on the complete points (ground truth) of the testing data. The lower bound (LB) represents the result tested on the partial points of the testing data. The remaining results are obtained by the synthesized objects.

4.9. Ablation Study

We evaluate different components in our network, including the adversarial training, mean shape, contraction-expansion unit, mirror operation and different Chamfer Distance calculations during training. We denote our method without discriminator as the baseline (BS). We use CD-P as the evaluation metric and the quantitative comparison are shown in Table 6. All experiments are done on the 2048 resolution points. We can see that our full pipeline performs the best. Removing any component decreases the performance, which verifies that each component contributes.

Training loss	Methods				
	w/o MS	w/o CE	w/o Mir	BS	w/ Dis
CD-P*	7.78	7.83	7.67	7.67	7.61
CD-P \diamond	7.80	7.73	7.71	7.68	7.57
CD-T*	7.93	7.90	7.76	7.75	7.68
CD-T \diamond	8.00	8.01	7.95	7.75	7.62

Table 6. Quantitative comparisons for the ablation study. Dis represents the discriminator, MS represents the mean shape features, CE represents the contraction-expansion unit, Mir represents the mirror operation for partial points. * and \diamond represent the TopNet training data and our training data, respectively.

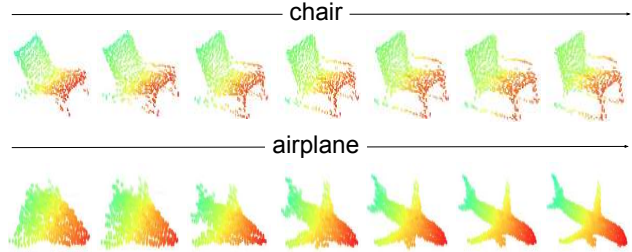


Figure 7. Shape interpolation results for chair and airplane.

4.10. Shape Arithmetic for Feature Learning

Following previous GAN methods [46, 10, 41, 13, 43], we show shape transformation by interpolating latent vectors from the encoder. Qualitative results are shown in Figure 7. The smooth transitions indicate that our learned features preserve critical geometric information. The synthesized reasonable object shapes verify the effectiveness of our cascaded refinement strategy.

4.11. Model Size Comparison

We evaluate the model size in Table 7 from two aspects for the resolution of 16,384 points: the number of parameters and the size of the trained models. We can see that our model has fewer parameters and smaller size compared to PCN and TopNet, since we share the parameters in each cascaded refinement step.

Methods	PCN [52]	TopNet [36]	Ours
#Paras	6.85M	9.96M	5.14M
Size of Model	82.30M	79.80M	61.90M

Table 7. Quantitative comparisons for model size.

5. Conclusion

In this work, we propose a novel point completion network to generate complete points given the partial inputs. The generator is a cascaded refinement network, which exploits the existing details of the partial input points and synthesize the missing parts with high quality. We design a patch discriminator that leverages on adversarial training to learn the accurate point distribution and penalize the generated objects from infidelity to the ground truth. We evaluate our proposed method on the completion datasets. Various experiments show that our method achieves state-of-the-art performances.

Acknowledgments. This research was supported in part by the Singapore Ministry of Education (MOE) Tier 1 grant R-252-000-A65-114, National University of Singapore Scholarship Funds and the National Research Foundation, Prime Ministers Office, Singapore, under its CREATE programme, Singapore-MIT Alliance for Research and Technology (SMART) Future Urban Mobility (FM) IRG.

References

- [1] Panos Achlioptas, Olga Diamanti, Ioannis Mitliagkas, and Leonidas Guibas. Learning representations and generative models for 3D point clouds. In *Proceedings of the 35th International Conference on Machine Learning*. PMLR, 2018.
- [2] Andrew C Boud, David J Haniff, Chris Baber, and SJ Steiner. Virtual reality and augmented reality as a training tool for assembly tasks. In *1999 IEEE International Conference on Information Visualization (Cat. No. PR00210)*, pages 32–36. IEEE, 1999.
- [3] Andrew Brock, Theodore Lim, James M Ritchie, and Nick Weston. Generative and discriminative voxel modeling with convolutional neural networks. *arXiv preprint arXiv:1608.04236*, 2016.
- [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] Angela Dai, Daniel Ritchie, Martin Bokeloh, Scott Reed, Jürgen Sturm, and Matthias Nießner. Scancomplete: Large-scale scene completion and semantic segmentation for 3d scans. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2018.
- [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] Jakob Engel, Thomas Schöps, and Daniel Cremers. Lsdslam: Large-scale direct monocular slam. In *European conference on computer vision*, pages 834–849. Springer, 2014.
- [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] 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*, pages 484–499. Springer, 2016.
- [11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.
- [12] Thibault Groueix, Matthew Fisher, Vladimir G Kim, Bryan C Russell, and Mathieu Aubry. Atlasnet: A paperm^{ach}e approach to learning 3d surface generation. *arXiv preprint arXiv:1802.05384*, 2018.
- [13] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems*, pages 5767–5777, 2017.
- [14] 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*, pages 85–93, 2017.
- [15] 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.
- [16] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.
- [17] Hou Ji, Angela Dai, and Matthias Nießner. 3d-sis: 3d semantic instance segmentation of rgb-d scans. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2019.
- [18] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7122–7131, 2018.
- [19] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [20] Truc Le and Ye Duan. Pointgrid: A deep network for 3d shape understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9204–9214, 2018.
- [21] Ruihui Li, Xianzhi Li, Chi-Wing Fu, Daniel Cohen-Or, and Pheng-Ann Heng. Pu-gan: a point cloud upsampling adversarial network. *arXiv preprint arXiv:1907.10844*, 2019.
- [22] Or Litany, Alex Bronstein, Michael Bronstein, and Ameesh Makadia. Deformable shape completion with graph convolutional autoencoders. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1886–1895, 2018.
- [23] 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.
- [24] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4460–4470, 2019.
- [25] Mateusz Michalkiewicz, Jhony K Pontes, Dominic Jack, Mahsa Baktashmotlagh, and Anders Eriksson. Deep level sets: Implicit surface representations for 3d shape inference. *arXiv preprint arXiv:1901.06802*, 2019.
- [26] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. Orb-slam: a versatile and accurate monocular slam system. *IEEE transactions on robotics*, 31(5):1147–1163, 2015.

- [27] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hour-glass networks for human pose estimation. In *European conference on computer vision*, pages 483–499. Springer, 2016.
- [28] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. DeepSDF: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 165–174, 2019.
- [29] 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.
- [30] 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.
- [31] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.
- [32] Muhammad Sarmad, Hyunjoo Jenny Lee, and Young Min Kim. RL-gan-net: A reinforcement learning agent controlled gan network for real-time point cloud shape completion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5898–5907, 2019.
- [33] Dong Wook Shu, Sung Woo Park, and Junseok Kwon. 3d point cloud generative adversarial network based on tree structured graph convolutions. *arXiv preprint arXiv:1905.06292*, 2019.
- [34] Ayan Sinha, Asim Unmesh, Qixing Huang, and Karthik Ramani. Surfnet: Generating 3d shape surfaces using deep residual networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6040–6049, 2017.
- [35] David Stutz and Andreas Geiger. Learning 3d shape completion from laser scan data with weak supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1955–1964, 2018.
- [36] Lyne P Tchammi, Vineet Kosaraju, S. Hamid Rezatofighi, Ian Reid, and Silvio Savarese. Topnet: Structural point cloud decoder. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [37] Diego Valsesia, Giulia Fracastoro, and Enrico Magli. Learning localized generative models for 3d point clouds via graph convolution. 2018.
- [38] Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2mesh: Generating 3d mesh models from single rgb images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 52–67, 2018.
- [39] 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):72, 2017.
- [40] Peng-Shuai Wang, Chun-Yu Sun, Yang Liu, and Xin Tong. Adaptive o-cnn: a patch-based deep representation of 3d shapes. In *SIGGRAPH Asia 2018 Technical Papers*, page 217. ACM, 2018.
- [41] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.
- [42] Weiyue Wang, Duygu Ceylan, Radomir Mech, and Ulrich Neumann. 3dn: 3d deformation network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1038–1046, 2019.
- [43] Weiyue Wang, Qiangui Huang, Suya You, Chao Yang, and Ulrich Neumann. Shape inpainting using 3d generative adversarial network and recurrent convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2298–2306, 2017.
- [44] Anthony Webster, Steven Feiner, Blair MacIntyre, William Massie, and Theodore Krueger. Augmented reality in architectural construction, inspection and renovation. In *Proc. ASCE Third Congress on Computing in Civil Engineering*, volume 1, page 996, 1996.
- [45] Huikai Wu, Junge Zhang, and Kaiqi Huang. Point cloud super resolution with adversarial residual graph networks. *arXiv preprint arXiv:1908.02111*, 2019.
- [46] Jiajun Wu, Chengkai Zhang, Tianfan Xue, Bill Freeman, and Josh Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In *Advances in neural information processing systems*, pages 82–90, 2016.
- [47] Guandao Yang, Xun Huang, Zekun Hao, Ming-Yu Liu, Serge Belongie, and Bharath Hariharan. Pointflow: 3d point cloud generation with continuous normalizing flows. *arXiv preprint arXiv:1906.12320*, 2019.
- [48] 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.
- [49] Wang Yifan, Shihao Wu, Hui Huang, Daniel Cohen-Or, and Olga Sorkine-Hornung. Patch-based progressive 3d point set upsampling. In *The IEEE Conference on Computer Vision and Pattern Recognition*, June 2019.
- [50] Lequan Yu, Xianzhi Li, Chi-Wing Fu, Daniel Cohen-Or, and Pheng-Ann Heng. Ec-net: an edge-aware point set consolidation network. In *Proceedings of the European Conference on Computer Vision*, pages 386–402, 2018.
- [51] Lequan Yu, Xianzhi Li, Chi-Wing Fu, Daniel Cohen-Or, and Pheng-Ann Heng. Pu-net: Point cloud upsampling network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2790–2799, 2018.
- [52] Wentao Yuan, Tejas Khot, David Held, Christoph Mertz, and Martial Hebert. Pcn: Point completion network. In *2018 International Conference on 3D Vision*, pages 728–737. IEEE, 2018.
- [53] Ziwei Liu Sanjay E. Sarma Michael M. Bronstein Justin M. Solomon Yue Wang, Yongbin Sun. Dynamic graph cnn for learning on point clouds. *arXiv preprint arXiv:1801.07829*, 2018.