FLOW-BASED POINT CLOUD COMPLETION NETWORK WITH ADVERSARIAL REFINEMENT

Rong Bao[1,3], *Yurui Ren*[1,3], *Ge Li**[1,3], *Wei Gao*[1,3], *Shan Liu*[2]

[1]School of Electronic and Computer Engineering, Peking University Shenzhen Graduate School
[2]Media Lab, Tencent [3]Peng Cheng Laboratory

ABSTRACT

Point cloud completion is the task of estimating the complete point cloud from the partial observation. Most of the existing methods tend to recover global shapes of 3D objects and usually lack local details. These methods rely merely on distance metrics between point sets as loss functions, which have the insufficient capability of supervising fine structures. In this work, we propose a coarse-to-fine approach to complete the partial point cloud with two stages: **1) Flow-based Completion Network**, a principled probabilistic model that built on continuous normalizing flow to generate coarse completions conditioned on partial inputs. **2) Adversarial Refinement Network**, a hierarchical refinement network constrained by the proposed patch discriminator to refine local details based on coarse completions. Experimental results show that our method can progressively complete 3D point clouds with fine details. Compared with other competitive methods, our method achieves better results on both quantitative and qualitative evaluations.

***Index Terms*—** Point Cloud Completion, Continuous Normalizing Flow, Adversarial Training

1. INTRODUCTION

Point clouds serve as efficient representations of 3D objects and scenes. However, raw point clouds directly obtained by 3D scanning devices are usually incomplete and sparse due to occlusions, noises, and limited sensor resolution. Therefore, it is essential to recover the complete point cloud from its partial observation to benefit a wide range of downstream applications such as autonomous driving, scene understanding, and augmented reality.

With the development of point-based networks [1, 2, 3], several point cloud completion networks have been proposed [4, 5, 6, 7, 8, 9, 10, 11]. However, most of these methods tend to recover shapes of 3D objects globally, which prevents them from predicting local details. Moreover, they merely adopt *Chamfer Distance* (CD) as loss function, which is not capable of supervising fine structures effectively [4, 9].

To this end, we propose a coarse-to-fine approach, which imposes more constraints on the completion process in order to progressively generate detailed completions from incomplete point clouds by two sub-networks. In the first stage, we propose a Flow-based Completion Network (FCNet), which is a principled probabilistic model built on a continuous normalizing flow to generate overall structures. The normalizing flow [12, 13, 14] is a sequence of invertible and differentiable mappings that transforms a simple probability distribution into a more complicated one, showing great potentials in point cloud generation [15, 16, 17]. While some previous works have limited modeling capability and are restricted to a fixed size, FCNet can construct complex distributions and generate point clouds with arbitrary resolutions by modeling an invertible transformation of points from a prior distribution. In the second stage, we propose an Adversarial Refinement Network (ARNet) with self-attention modules [11] to refine local details based on generated coarse completions. Considering that CD is not capable of supervising fine structures effectively [9], we propose an additional patch discriminator with self-attention kernels, which guides the distribution of the generated patches to simulate that of the ground truth patches. The adversarial loss is helpful for the network to reduce invalid points and generate visually-pleasing completions free from the corruption by the noise points. Experiments on datasets of different resolutions demonstrate that our method outperforms other competitive methods quantitatively and qualitatively.

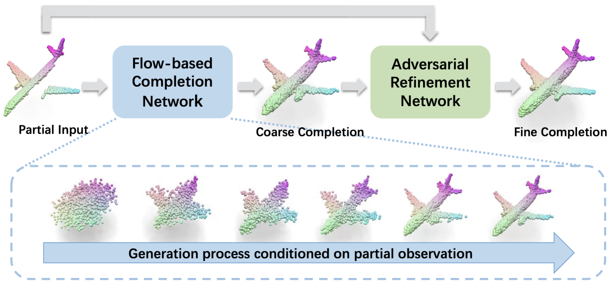
Flow-based Completion Network
Adversarial Refinement Network
Partial Input
Coarse Completion
Fine Completion
Generation process conditioned on partial observation

Fig. 1. The overall architecture of our method. FCNet generates complete point clouds conditioned on partial observations. ARNet refines local details based on coarse completions.

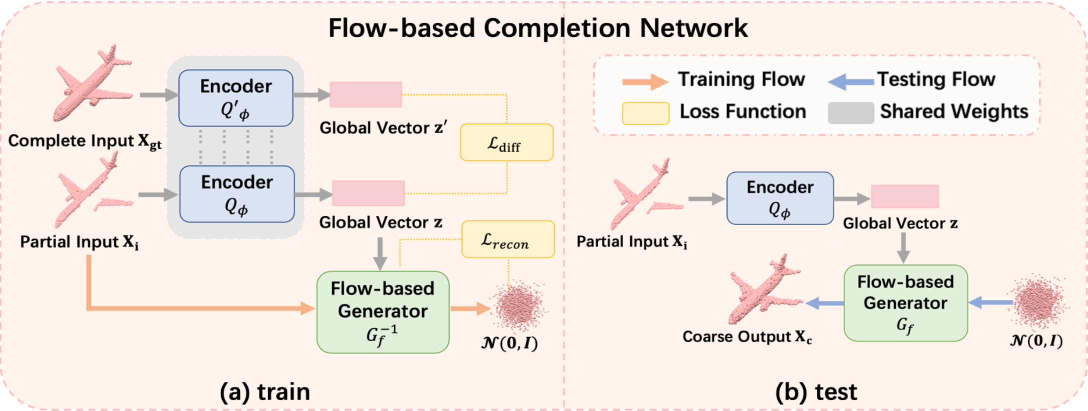
Flow-based Completion Network
Training Flow
Testing Flow
Loss Function
Shared Weights
Complete Input $\mathbf{X}_{gt}$
Encoder Q'_ϕ
Global Vector $\mathbf{z}'$
$\mathcal{L}_{diff}$
Partial Input $\mathbf{X}_i$
Encoder Q_ϕ
Global Vector $\mathbf{z}$
$\mathcal{L}_{recon}$
Flow-based Generator G_f^{-1}
$\mathcal{N}(0,I)$
(a) train
Partial Input $\mathbf{X}_i$
Encoder Q_ϕ
Global Vector $\mathbf{z}$
Flow-based Generator G_f
Coarse Output $\mathbf{X}_c$
$\mathcal{N}(0,I)$
(b) test

Fig. 2. The architecture of FCNet.

2. OUR METHOD

The architecture of our method is shown in Fig.1. Given an incomplete point cloud $\mathbf{X_i}$ as the input, we first generate a coarse completion $\mathbf{X_c}$ by using our flow-based completion network. Subsequently, we refine the local details of the fine completion $\mathbf{X_f}$ based on the coarse output by using our adversarial refinement network.

2.1. Flow-Based Completion Network

We propose the flow-based completion network to predict the coarse completion, as shown in Fig.2. The encoder is designed to embed the partial input $\mathbf{X_i}$ into the shape representation $\mathbf{z}$. The continuous normalizing flow is employed to build the generator G_f, which aims at generating the coarse completion $\mathbf{X_c}$ given a shape $\mathbf{z}$.

Specifically, a point $\mathbf{x}$ of the point cloud $\mathbf{X}$ is generated by transforming a point $\mathbf{y}(t_0)$ of the prior distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$ conditioned on the shape representation $\mathbf{z}$:

$$\mathbf{x} = G_f(\mathbf{y}(t_0); \mathbf{z}) = \mathbf{y}(t_0) + \int_{t_0}^{t_1} g_f(\mathbf{y}(t), t, \mathbf{z})dt, \quad (1)$$

where t_o and t_1 denote the starting and the ending time of continuous transformation respectively. $\mathbf{y}(t_0) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\mathbf{y}(t_1) = \mathbf{x}$.

Due to the invertibility of the continuous normalizing flow, the flow-based model can compute the exact likelihood by applying the *instantaneous-change-of-variables* formula [13, 14]:

$$logP(\mathbf{x}|\mathbf{z}) = logP(G_f^{-1}(\mathbf{x}; \mathbf{z})) - \int_{t_0}^{t_1} Tr(\frac{\partial g_f}{\partial \mathbf{y}(t)})dt, \quad (2)$$

where $logP(G_f^{-1}(\mathbf{x}; \mathbf{z}))$ can be easily computed since we choose standard Gaussian as the prior distribution. Therefore, we can train the network by directly minimizing the negative log-likelihood:

$$\mathcal{L}_{recon} = \mathbb{E}_{Q_\phi(\mathbf{z}|\mathbf{X_i})}[logP(\mathbf{X}|\mathbf{z})]$$
$$= \mathbb{E}_{Q_\phi(\mathbf{z}|\mathbf{X_i})}[\sum_{\mathbf{x}\in\mathbf{X}}(logP(G_f^{-1}(\mathbf{x}; \mathbf{z})) - \int_{t_0}^{t_1} Tr(\frac{\partial g_f}{\partial \mathbf{y}(t)})dt)]. \quad (3)$$

Additionally, in order to obtain a reasonable shape representation, we design a Siamese encoder that encodes the partial and the complete point clouds separately at training time. We constrain that the embedded shape representation of incomplete point cloud is consistent with the complete one by minimizing the difference between them, which is defined as: $\mathcal{L}_{diff} = |\mathbf{z} - \mathbf{z}'|$. The total loss of FCNet can be formulated as: $\mathcal{L}_{FCNet} = -\lambda_{recon}\mathcal{L}_{recon} + \lambda_{diff}\mathcal{L}_{diff}$.

2.2. Adversarial Refinement Network

After obtaining the coarse completion, we propose the adversarial refinement network to refine local details, as shown in Fig.3. The generator G_a based on U-Net architecture takes the incomplete point cloud and the coarse completion as inputs and generates the fine completion. Inspired by the success of self-attention modules [11] in learning and fusing multi-scale point features, we adopt a self-attention residual block as the basic building block of G_a.

The point self-attention kernel (PSA) can adaptively aggregate local neighboring point features:

$$\mathbf{y}_i = \sum_{j\in N_k(i)} \gamma([\sigma(\mathbf{x}_i), [\xi(\mathbf{x}_j)]_{\forall j\in N_k(i)}]) \odot \beta(\mathbf{x}_j), \quad (4)$$

where $N_k(i)$ is the K-Nearest Neighboring points of point i, x_j is the corresponding point feature of point $j \in N_k(i)$. The multilayer perceptron (MLP) is represented by γ, σ, ξ and β. The relation operation $[*, *]$ concatenates point features, while the map operation γ and β transform features into vectors of the same dimension. The element-wise product is represented by $\odot$.

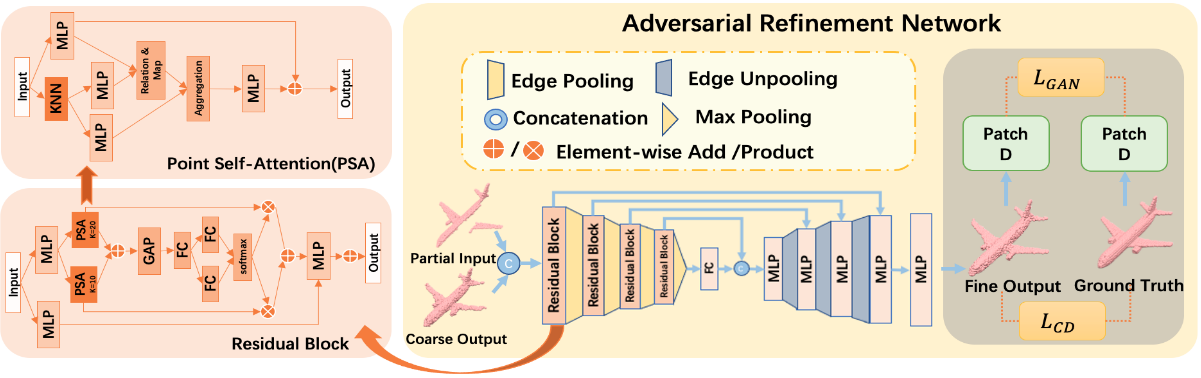

Fig. 3. The architecture of ARNet.

The residual block can adaptively fuse multi-scale features extracted by point self-attention kernel:

$$\mathbf{V} = [\mathbf{V}_1, ..., \mathbf{V}_C], \mathbf{V}_c = \sum_{i=1}^{M} \mathbf{U}_c^i \frac{exp(\mathbf{W}_c^i \mathbf{z})}{\sum_{j=1}^{M} exp(\mathbf{W}_c^j \mathbf{z})} \forall c \in C,$$

$$\mathbf{z} = f(\mathbf{W}\mathbf{s}), \mathbf{s}_c = \frac{1}{N} \sum_{i=1}^{N} \mathbf{U}_c(i), \mathbf{U} = \sum_{i=1}^{M} \mathbf{U}^i, \tag{5}$$

where $\mathbf{V}$ is the fused feature, $\mathbf{U}^i$ is the point feature obtained by PSA of different selected scale, $\mathbf{z}$ is the compact feature which guides adaptive selections, and $\mathbf{s}$ is a global feature obtained by global average pooling over all points. $\mathbf{W}^i \in \mathbb{R}^{C \times d}$ and $\mathbf{W} \in \mathbb{R}^{d \times C}$ are learnable weights, where d is a reduced feature size. M denotes the number of PSA branches. The fully-connected layer (FC) is represented by f.

We adopt CD as our training loss to regress the ground-truth: $\mathcal{L}_{CD} = \text{CD}(\mathbf{X_f}, \mathbf{X_{gt}})$, where $\mathbf{X_f}$ and $\mathbf{X_{gt}}$ represent the fine completion and the ground-truth respectively.

CD usually causes situations where points over-populate in common areas (e.g. seats of chairs) with the details of the changeable parts (e.g. legs of chairs) blurred [9, 4]. Therefore, an additional adversarial loss [18, 10] is proposed to help supervise fine details. We design a patch discriminator D_a with self-attention kernels to guide the network in generating fine completions with higher perceptual quality. Specifically, n points are randomly selected as centers of local patches by farthest point sampling [2]. The self-attention residual block is employed to extract and fuse multi-scale features of each local patch. Finally, scores are taken at each local patches to ensure the consistency between the generated and the real point clouds. The adversarial losses can be formulated as:

$$\mathcal{L}_{GAN}(G_a) = \frac{1}{2} \mathbb{E}_{\widetilde{\mathbf{x}} \sim \mathbf{X_f}} (D(\widetilde{\mathbf{x}}) - 1)^2,$$

$$\mathcal{L}_{GAN}(D_a) = \frac{1}{2} \mathbb{E}_{\widetilde{\mathbf{x}} \sim \mathbf{X_f}} D(\widetilde{\mathbf{x}})^2 + \frac{1}{2} \mathbb{E}_{\mathbf{x} \sim \mathbf{X_{gt}}} (D(\mathbf{x}) - 1)^2. \tag{6}$$

The total loss of ARNet can be formulated as: $\mathcal{L}_{ARNet} = \lambda_{GAN} \mathcal{L}_{GAN} + \lambda_{CD} \mathcal{L}_{CD}$.

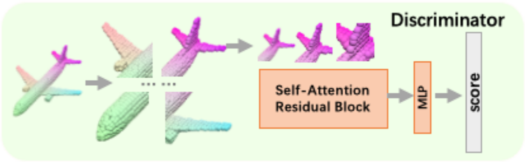

Fig. 4. The architecture of patch discriminator.

3. EXPERIMENTS

3.1. Datasets and Evaluation Metrics

Datasets. We evaluate our proposed method on the commonly used datasets for point cloud completion: the MVP dataset [11] and the Shapenet Completion3D dataset [8]. The MVP dataset consists of over 100,000 partial and complete point clouds pairs with different resolutions. We follow the same split and test on overall 16 categories. The ShapeNet Completion3D dataset consists of 28,974 samples for training and 800 samples for validation derived from ShapeNet [19]. We evaluate all methods on our created testing set, which consists of 800 samples and is generated by back-projecting 2.5D depth images into 3D. Different from the MVP dataset, there are only 2,048 points in the partial and complete point clouds.

Evaluation Metrics. To align with previous methods, we evaluate the completion results in terms of Chamfer Distance between the completion and the ground-truth:

$$\text{CD}(\mathbf{X}, \mathbf{Y}) = \frac{1}{|\mathbf{X}|} \sum_{\mathbf{x} \in \mathbf{X}} \min_{\mathbf{y} \in \mathbf{Y}} \|\mathbf{x} - \mathbf{y}\|_2^2 + \frac{1}{|\mathbf{Y}|} \sum_{\mathbf{y} \in \mathbf{Y}} \min_{\mathbf{x} \in \mathbf{X}} \|\mathbf{x} - \mathbf{y}\|_2^2, \tag{7}$$

where $\mathbf{X}$ and $\mathbf{Y}$ denote the generated completion and the ground-truth respectively. In addition, we use the F-Score@1% [20] as an extra metric to evaluate the distance between object surfaces, which is defined as the harmonic mean between precision and recall.

Method	2048pts										16384pts									
	CD									F1	CD									F1
	plane	cabinet	car	chair	lamp	sofa	table	vessel	avg.	avg.	plane	cabinet	car	chair	lamp	sofa	table	vessel	avg.	avg.
TopNet	6.05	10.92	7.89	15.66	23.24	13.58	14.58	9.50	10.11	0.31	2.75	4.25	3.40	7.95	17.01	6.04	7.42	6.04	6.36	0.60
PCN	4.11	8.86	6.48	12.27	18.06	9.65	12.26	8.58	9.77	0.32	2.95	4.13	3.04	7.07	14.93	5.56	7.06	6.08	6.02	0.64
MSN	2.99	10.87	7.04	12.20	14.78	9.97	12.06	8.07	7.90	0.43	2.07	3.82	2.76	6.21	12.72	4.74	5.32	4.80	4.90	0.71
CRN	2.55	9.01	6.17	9.22	11.46	8.86	9.94	6.54	7.25	0.43	1.59	3.64	2.60	5.24	9.02	4.42	5.45	4.26	4.30	0.74
VRCNet	2.17	**7.83**	5.53	**7.31**	8.29	7.42	**7.07**	5.15	5.95	0.50	1.15	**3.20**	2.14	**3.58**	5.57	3.58	**4.17**	2.47	3.06	0.80
Ours	**1.94**	7.85	**5.32**	7.50	**8.04**	**7.00**	7.56	**5.14**	**6.29**	**0.52**	**1.13**	3.35	**2.08**	3.67	**5.51**	**3.43**	4.28	**2.37**	**2.76**	**0.81**

Table 1. Completion comparison on MVP dataset in terms of CD $\times 10^4$ (lower is better) and F-score@1% (higher is better). Note that avg. indicates the average of the overall 16 categories in dataset.

Method	plane	cabinet	car	chair	lamp	sofa	table	vessel	avg.
PCN	6.82	26.20	15.29	22.30	21.22	17.96	23.08	16.03	18.61
MSN	5.96	25.64	15.52	21.43	20.93	18.83	24.76	15.06	18.52
CRN	5.87	25.02	12.91	21.53	18.62	18.09	22.82	11.05	16.99
VRCNet	4.86	23.14	13.38	**17.84**	**18.44**	14.90	19.25	14.42	15.78
Ours-A	5.83	26.57	11.60	22.06	25.54	20.54	24.37	16.43	19.12
Ours-B	4.05	20.6	9.43	21.50	23.60	15.11	18.75	11.06	15.51
Ours	**3.94**	**19.17**	**9.09**	21.12	20.35	**14.20**	**18.55**	**9.98**	**14.55**

Table 2. Completion comparison and ablation study on ShapeNet in terms of CD $\times 10^4$. Ours-A denotes FCNet. Ours-B denotes FCNet and ARNet without patch discriminator. Ours denotes our full method.

3.2. Comparison Experiments

We compare our method with the following competitive approaches: 1) PCN [7] is a pioneering work for point cloud completion that folds 2D patches into the 3D surface. 2) TopNet [8] proposes a tree-structured decoder. 3) MSN [9] generates coarse completions with a folding-based decoder and refines local details with a residual network. 4) CRN [10] proposes an adversarial coarse-to-fine cascaded refinement network. 5) VRCNet [11] is a very recent method that generates detailed completions by learning structural relations.

The quantitative comparison results are presented in Table 1 and 2. Our method outperforms existing methods in terms of CD and F-score on both datasets. Moreover, we evaluate our method on two different resolutions (2048 or 16384 points), demonstrating the superiority of our method in the multi-resolution completion.

The qualitative comparison results are shown in Fig. 5. Our method can generate neat completions that are visually pleasing while other methods tend to generate blurry results (as indicated by red arrows). In particular, our method can preserve and predict delicate local details such as the fine structures of the chair, the mast of the vessel.

3.3. Ablation Study

The ablation studies are presented in Table 2. We denote FCNet as the baseline. The addition of proposed modules will lead to better completion results. Moreover, we can clearly observe that our method progressively consolidates the fine structures in Fig. 6, which demonstrates the effectiveness of each proposed component.

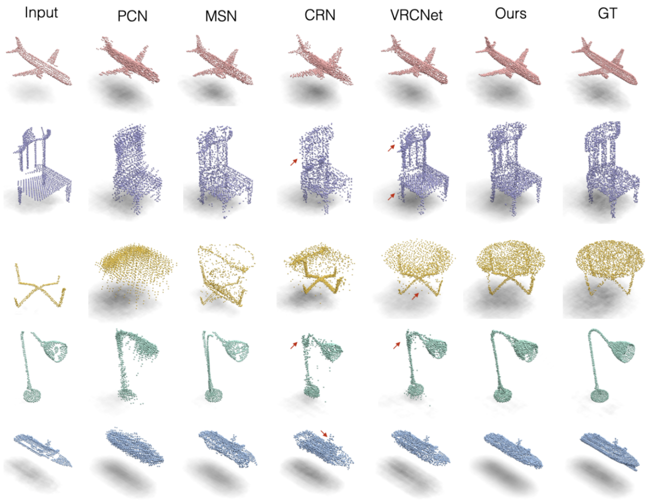

Fig. 5. Visualized completion comparison on ShapeNet.

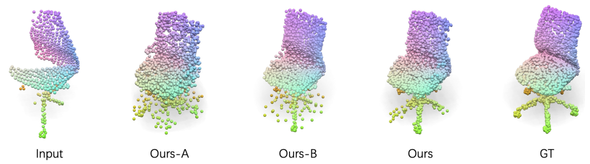

Fig. 6. Visualized ablation study.

4. CONCLUSION

In this paper, we propose a novel coarse-to-fine approach for point cloud completion, which consists of the flow-based completion network and the adversarial refinement network. Various experiments show that our method can progressively consolidate fine structures and outperform other methods quantitatively and qualitatively.

Acknowledgements. This research received support from the National Key R&D Program of China (No. 2020AAA0103501).

5. REFERENCES

[1] 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*, 2017, pp. 652–660.

[2] Charles R Qi, Li Yi, Hao Su, and Leonidas J Guibas, “Pointnet++: Deep hierarchical feature learning on point sets in a metric space,” *arXiv preprint arXiv:1706.02413*, 2017.

[3] 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.

[4] Panos Achlioptas, Olga Diamanti, Ioannis Mitliagkas, and Leonidas Guibas, “Learning representations and generative models for 3d point clouds,” in *International conference on machine learning*. PMLR, 2018, pp. 40–49.

[5] 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*, 2018, pp. 206–215.

[6] Maria Vakalopoulou, Guillaume Chassagnon, Norbert Bus, Rafael Marini, Evangelia I Zacharaki, M-P Revel, and Nikos Paragios, “Atlasnet: multi-atlas non-linear deep networks for medical image segmentation,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 658–666.

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

[8] 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*, 2019, pp. 383–392.

[9] Minghua Liu, Lu Sheng, Sheng Yang, Jing Shao, and Shi-Min Hu, “Morphing and sampling network for dense point cloud completion,” in *Proceedings of the AAAI conference on artificial intelligence*, 2020, vol. 34, pp. 11596–11603.

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

[11] Liang Pan, Xinyi Chen, Zhongang Cai, Junzhe Zhang, Haiyu Zhao, Shuai Yi, and Ziwei Liu, “Variational relational point completion network,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8524–8533.

[12] Danilo Rezende and Shakir Mohamed, “Variational inference with normalizing flows,” in *International conference on machine learning*. PMLR, 2015, pp. 1530–1538.

[13] Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud, “Neural ordinary differential equations,” *arXiv preprint arXiv:1806.07366*, 2018.

[14] Will Grathwohl, Ricky TQ Chen, Jesse Bettencourt, Ilya Sutskever, and David Duvenaud, “Ffjord: Free-form continuous dynamics for scalable reversible generative models,” *arXiv preprint arXiv:1810.01367*, 2018.

[15] Guandao Yang, Xun Huang, Zekun Hao, Ming-Yu Liu, Serge Belongie, and Bharath Hariharan, “Pointflow: 3d point cloud generation with continuous normalizing flows,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4541–4550.

[16] Albert Pumarola, Stefan Popov, Francesc Moreno-Noguer, and Vittorio Ferrari, “C-flow: Conditional generative flow models for images and 3d point clouds,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7949–7958.

[17] Roman Klokov, Edmond Boyer, and Jakob Verbeek, “Discrete point flow networks for efficient point cloud generation,” in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII 16*. Springer, 2020, pp. 694–710.

[18] 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*, 2017, pp. 2794–2802.

[19] 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.

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