Energy-Based Residual Latent Transport for Unsupervised Point Cloud Completion

Ruikai Cui[1]
ruikai.cui@anu.edu.au
Shi Qiu[1,2]
shi.qiu@anu.edu.au
Saeed Anwar[1,2]
saeed.anwar@csiro.au
Jing Zhang[1]
jing.zhang@anu.edu.au
Nick Barnes[1]
nick.barnes@anu.edu.au

[1] Australian National University, Canberra, ACT, AU
[2] Data61, CSIRO Canberra, ACT, AU

Abstract

Unsupervised point cloud completion aims to infer the whole geometry of a partial object observation without requiring partial-complete correspondence. Differing from existing deterministic approaches, we advocate generative modeling based unsupervised point cloud completion to explore the missing correspondence. Specifically, we propose a novel framework that performs completion by transforming a partial shape encoding into a complete one using a latent transport module, and it is designed as a latent-space energy-based model (EBM) in an encoder-decoder architecture, aiming to learn a probability distribution conditioned on the partial shape encoding. To train the latent code transport module and the encoder-decoder network jointly, we introduce a residual sampling strategy, where the residual captures the domain gap between partial and complete shape latent spaces. As a generative model-based framework, our method can produce uncertainty maps consistent with human perception, leading to explainable unsupervised point cloud completion. We experimentally show that the proposed method produces high-fidelity completion results, outperforming state-of-the-art models by a significant margin.

1 Introduction

Point cloud data is a fundamental representation of 3D geometry, contributing to numerous applications in robotics, auto-navigation, augmented reality, *etc*. Limited by viewing angle, occlusion, and acquisition resolution, raw point clouds are generally sparse and incomplete. We argue that the completion of partial scans is not only essential for a better understanding of 3D scenes, but also beneficial for many downstream 3D computer vision tasks, including classification [6, 29], segmentation [14, 27], and detection [25, 28]. To this end, increasing attention has been dedicated to point cloud completion.

Pioneered by PCN [40], supervised methods [23, 31, 33, 39] have achieved impressive completion results. However, they usually rely on large-scale datasets with both partial and corresponding ground truth complete shapes, where the latter is hard to collect. To tackle the difficulty in data collection, unsupervised point cloud completion has recently gained popularity due to its capability of utilizing both synthetic and real-world datasets. In the unsupervised setting, only unpaired samples from the partial observation domain and complete shape domain are provided, so a model needs to infer the partial-to-complete relationship. Existing methods [2, 7, 32] address the problem mainly by mapping a partial shape to a latent code that can be decoded as a valid complete shape. Nevertheless, a deterministic one-to-one mapping is assumed in existing methods, which can be biased because a partial shape can correspond to multiple complete shapes due to the missing geometries.

We introduce a generative model-based strategy to explore the one-to-many mapping inherent in point cloud completion; thus, we can capture the prediction uncertainties, representing our ignorance about the complete shape space. Specifically, we leverage an encoder-decoder architecture, where parameters of the encoder and decoder for complete and incomplete point clouds, respectively, are shared. For an incomplete shape latent code, we assume it locates somewhere near its corresponding complete shape codes in the latent space. Consequently, we design an energy-based model (EBM) [1] and deploy it in the latent space of the encoder-decoder architecture. The latent-space EBM aims to learn a conditional distribution of complete shape code given a partial shape code. By sampling with the gradient-based Markov chain Monte Carlo (MCMC) (*e.g.* Langevin dynamics [10]) initialized by a partial shape code, a latent code corresponding to a valid complete shape can be generated.

The vanilla Langevin dynamics [10] do not facilitate back-propagation since a computationally expensive second-order gradient is needed when back-propagating through the Langevin iterations. Therefore, existing EBMs are either integrated into a deep neural network as a prior model [24, 41] or utilized with a pre-trained network [24]. Unlike existing techniques, we propose a residual sampling strategy that, for the first time, enables joint training of a latent-space EBM and a task-related generator. Particularly, instead of sampling the conditional latent code of a complete shape directly, we generate a residual that captures the gap for transporting a code from one domain (partial shape space) to the other (complete shape space). By adding the residual back to the partial latent code, we achieve both complete latent code sampling, and gradient back-propagation with parameter updating for the encoder-decoder.

Our main contributions are summarized below:

- We propose a novel energy-based latent transport mechanism, enabling generative modeling of the unsupervised point cloud completion task for the first time.
- We present a residual sampling strategy that allows joint training of a latent-space EBM and an encoder-decoder.
- Experimental results indicate that our model not only achieves state-of-the-art performance on synthetic (ShapeNet [5]) and real-world (KITTI [11], ScanNet [8], MatterPort3D [4]) datasets, but is also capable of generating explainable uncertainty maps.

2 Related Works

Unsupervised Point Cloud Completion. As a pioneering work for unsupervised point cloud completion [2, 3, 7, 32, 42], Pcl2Pcl [7] proposed an adversarial learning-based

approach, where they firstly train two autoencoders for incomplete and complete point clouds, respectively, then a generator [12] was used to transform latent code of the incomplete shape to that of the complete shape. Following [7], Cycle4Completion [32] introduced two cycle transformations between the latent space of the complete and partial shapes, achieving dual-direction transformation of the two shape codes. However, it imposes a strong "one-to-one deterministic mapping" constraint [45], neglecting the uncertainty in the partial and complete shape domains. To address this issue, ShapeInversion [42] searched for a latent code in the latent space of a pre-trained GAN [12] with a partial-complete matching loss. Nevertheless, the search, as a minimization program, is highly non-convex, where the optimization can be easily stuck in a local minimum, leading to poor completion quality. More recently, Cai *et al.* [2] encoded a series of related partial point clouds into a unified latent space as a complete shape code and multiple occlusion codes, but a deterministic one-to-one mapping is assumed to perform completion, while a partial shape can have multiple valid, complete predictions due to occlusion. Different from existing methods, we propose a novel energy-based latent transport module, with which we aim to model the distribution gap between the partial and the complete shape codes. Further, as a generative model, our method can produce uncertainty maps representing ignorance of the model towards its prediction.

Energy-Based Models. Learning a data distribution for new sample generation is an important task in machine learning. Recent works have shown that energy-based models (EBMs) [1] have a strong capability in modeling high dimensional data, such as images [10], videos [37], and point clouds [36]. Yet, the capability of EBMs in point cloud completion remains unclear, which makes our method a research pioneer on the topic. Despite leveraging EBMs in data space, Pang *et al.* [24] proposed to jointly learn a latent space energy-based prior with a top-down generator network. Our method differs from [24] as our EBM takes an output from a learnable module. Therefore, gradient back-propagation is necessary to train the network while [24] does not support it. LETIT [44] is the most related art to our method, where an EBM is deployed in the latent space of a pre-trained variational autoencoder (VAE) [18] to transform a latent code from one domain to another domain. Despite that, the VAE and the EBM are trained in two stages since sampling from an EBM blocks gradient back-propagation to the encoder. To the best of our knowledge, we achieve simultaneous training of a latent space EBM for the first time, thanks to the proposed residual sampling strategy.

3 Method

Let $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ be samples from the partial and complete point cloud domain, respectively. We are provided with samples from two marginal distributions $p(x)$ and $p(y)$ instead of the joint distribution $p(x,y)$ in the unsupervised point cloud completion task. The goal of this task is to estimate the conditional distribution $p(y|x)$ with a point cloud completion model $p(y_{\mathcal{X}\to\mathcal{Y}}|x)$, where $y_{\mathcal{X}\to\mathcal{Y}}$ is a sample produced by translating a sample from the partial point cloud domain $\mathcal{X}$ to the complete point cloud domain $\mathcal{Y}$. Given the one-to-many mapping attribute inherent in the point cloud completion task, we present a generative unsupervised point cloud completion model, where a latent variable is introduced to explore the missing correspondence knowledge from the partial observation. Our model consists of four main parts, namely **1)** an encoder ($\mathcal{E}_\alpha$) for point cloud code extraction, **2)** a latent-space energy-based model with energy function E_θ to fill the code gap between the partial point cloud and its corresponding completion, **3)** a decoder ($\mathcal{D}_\beta$) for point cloud reconstruction and **4)** a point domain discriminator (D_γ) to achieve adversarial learning (see Figure 1).

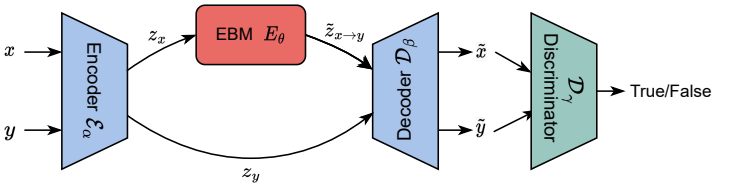

Figure 1: Model overview. Our model consists of 1) an encoder $\mathcal{E}_\alpha$ and a decoder $\mathcal{D}_\beta$ for both domain $\mathcal{X}$ and $\mathcal{Y}$, 2) an EBM (E_θ) that transports a partial shape latent code z_x to a complete shape latent code $\tilde{z}_{x\to y}$, 3) and a point domain discriminator D_γ.

Notably, the encoder projects a point cloud into the latent space, leading to a partial point cloud code z_x or a complete point cloud code z_y. Since z_y is the representation of a shape with full geometry, we feed it directly to the decoder for point cloud reconstruction. However, the decoder cannot effectively reconstruct the corresponding complete point cloud with the latent code z_x of the incomplete shape due to information loss caused by partial observation. The latent-space EBM is then introduced to transport z_x to its corresponding complete shape latent code $\tilde{z}_{x\to y}$. In this way, the same decoder can complete a partial observation with the inferred latent code $\tilde{z}_{x\to y}$. Finally, adversarial training is presented to guarantee the quality of the inferred complete shape $\tilde{x}$ for effective unsupervised point cloud completion.

In the following sections, first, we introduce the energy-based transport module via residual sampling (Section 3.1). Then, we discuss our learning pipeline to achieve joint training of our encoder-decoder framework (Section 3.2). Finally, we present our objective functions (Section 3.3) for the proposed generative unsupervised point cloud completion task.

3.1 Energy-based Residual Latent Transport

EBM for supervised point cloud completion - A preliminary: Following Pang *et al.* [24], we define the conditional distribution of the complete shape latent code as: $p_\theta(z_y|z_x) = \frac{p_\theta(z_y,z_x)}{\int_\theta p_\theta(z_y,z_x)dz_y} = \frac{\exp[-E_\theta(z_y,z_x)]}{Z(z_x;\theta)}$, where $E_\theta(z_y,z_x)$ is the energy function parameterized by a deep neural network, mapping the code pair (z_y,z_x) to a scalar that measures their compatibility. $Z(z_x;\theta) = \int \exp[-E_\theta(z_y,z_x)]dz_y$ is the normalizing constant. For supervised point cloud completion, $E_\theta(z_y,z_x)$ can be easily modeled as we have paired training samples. In this case, given z_x, z_y can then be achieved by Langevin dynamics [10] following:

$$z_y^{t+1} = z_y^t - \frac{\delta^2}{2}\frac{\partial}{\partial z_y}E_\theta(z_y^t,z_x) + \delta\varepsilon^t, \tag{1}$$

where t represents Langevin steps, δ is the step size, $\varepsilon^t \sim \mathcal{N}(0,\mathbf{I})$ is the Gaussian random noise. The prediction process via Langevin sampling in Eq. 1 can be considered as finding z_y to minimize the cost $E_\theta(z_y,z_x)$ given z_x.

Challenges and solution of applying EBM for unsupervised point cloud completion: In the unsupervised setting, we have no access to paired partial-complete point cloud samples, thus, we are incapable of directly modeling the compatibility of z_x and z_y through the energy function $E_\theta(z_y,z_x)$. Alternatively, we model the conditional distribution of the residual r_{xy}

instead, representing the distribution gap between the latent space of the two domains as:

$$p_\theta(r_{xy}|z_x) = \frac{p_\theta(r_{xy}, z_x)}{\int_\theta p_\theta(r_{xy}, z_x) dz_x} = \frac{\exp[-E_\theta(r_{xy}, z_x)]}{Z(r_{xy}; \theta)}. \tag{2}$$

Different from E_θ for the conditional distribution of complete code in the supervised setting, the energy function $E_\theta(r_{xy}, z_x)$ in Eq. 2 is designed to build the bridge between z_x and z_y, which we term "latent code transport". The benefits of learning the residual instead of the complete code mainly lie in (see Section 3.2): **1)** No paired partial-complete point cloud data is needed for the residual-based EBM, leading to unsupervised point cloud completion. Particularly, we can define the initial state of r_{xy} as 0, representing no domain gap between partial-complete point cloud codes. The model is trained to gradually refine r_{xy}. **2)** It leads to joint training of the EBM and the encoder-decoder framework, which is significantly different from existing techniques, where the two modules are updated separately [24, 41].

3.2 Joint Training of EBM and the Encoder-Decoder Framework

For the conditional distribution $p_\theta(r_{xy}|z_x)$ modeled with an energy-based model, we can sample the residual r_{xy} given a partial shape code z_x, by running several Langevin steps similar to Eq. 1. With multiple (K) Langevin steps, it has been proved that the final $r_{xy} = r_{xy}^K$ is sampled from the true conditional residual distribution $q(r_{xy}|z_x)$ [24]. In this way, a complete shape code $\tilde{z}_{x\to y}$ can be obtained as:

$$\tilde{z}_{x\to y} = z_x + \Omega(r_{xy}) \tag{3}$$

Here, we apply the stop gradient operation ($\Omega(\cdot)$) [10] to avoid unfolding the Langevin dynamics iteration and involving the second-order gradient of r_{xy} in future gradient computation, which is computationally expensive as shown in our supplementary material. Specifically, we define the parameter set of the encoder as α and the loss function as $\mathcal{L}$. The encoder can be trained with standard gradient descent via: $\frac{\partial \mathcal{L}}{\partial \alpha} = \frac{\partial \mathcal{L}}{\partial \tilde{z}_{x\to y}} \frac{\partial \tilde{z}_{x\to y}}{\partial z_x} \frac{\partial z_x}{\partial \alpha}$. With the stop gradient operator $\Omega(\cdot)$, the gradient can be simplified as $\frac{\partial \mathcal{L}}{\partial \alpha} = \frac{\partial \mathcal{L}}{\partial z_x} \frac{\partial z_x}{\partial \alpha}$, which can be supported by automatic differentiation.

The EBM module can be trained via maximizing the log-likelihood [22, 35] as: $L_{\text{ebm}} = \frac{1}{N}\sum_{i=1}^{N} \log p_\theta(r_{xy}^i|z_x^i)$, where N is the number of the partial point clouds. The gradients can be obtained as: $\Delta\theta = \frac{1}{N}\sum_{i=1}^{N}\{\mathbb{E}_{p_\theta(r_{xy}|z_x^i)}\left[\frac{\partial}{\partial\theta}E_\theta(r_{xy}, z_x^i)\right] - \frac{\partial}{\partial\theta}E_\theta(r_{xy}^i, z_x^i)\}$. In practice, we feed $\tilde{z}_{x\to y}$ to E_θ, leading to: $\Delta\theta = \frac{1}{N}\sum_{i=1}^{N}\{\mathbb{E}_{\tilde{z}_{x\to y}\sim p_\theta}\left[\frac{\partial}{\partial\theta}E_\theta(\tilde{z}_{x\to y}^i)\right] - \frac{\partial}{\partial\theta}E_\theta(r_{xy}^i, z_x^i)\}$.

The first term of $\Delta\theta$ is tractable. However, we cannot directly estimate $\frac{\partial}{\partial\theta}E_\theta(r_{xy}^i, z_x^i)\}$ as the paired residual and partial code (r_{xy}, z_x) is unavailable. We find that $E_\theta(r_{xy}^i, z_x^i)$ is minimized when r_{xy} can indeed represent the true partial-complete code gap. We then claim that the minimization of $E_\theta(r_{xy}^i, z_x^i)$ is equivalent to finding z_y of the complete point cloud code that minimizes the same energy function E_θ. In this way, we approximate $E_\theta(r_{xy}, z_x)$ of the partial point cloud with $E_\theta(z_y)$ of the complete point cloud with:

$$\Delta\theta = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{\tilde{z}_{x\to y}\sim p_\theta}[\frac{\partial}{\partial\theta}E_\theta(\tilde{z}_{x\to y}^i)] - \mathbb{E}_{z_y\sim p_{z_y}}[\frac{\partial}{\partial\theta}E_\theta(z_y)] + C, \tag{4}$$

where C is a constant which can be ignored during training. We use the proposed residual sampling strategy to sample a $\tilde{z}_{x\to y}$ from p_θ, and Adam [17] with $\Delta\theta$ to update θ. Given

the inferred latent code for both the partial and complete observation, the decoder $\mathcal{D}_\beta$ parameterized with β can then be updated with stochastic gradient descent.

3.3 Training Objectives

Reconstruction loss: We adopt Chamfer Distance (CD) as the reconstruction loss. Given two sets of point clouds S_1 and S_2, CD measures the point cloud similarity via:

$$\mathrm{CD}(S_1,S_2) = \frac{1}{|S_1|}\sum_{p_1\in S_1}\min_{p_2\in S_2}\|p_1-p_2\|_2 + \frac{1}{|S_2|}\sum_{p_2\in S_2}\min_{p_1\in S_1}\|p_2-p_1\|_2, \tag{5}$$

where p_1 and p_2 are points from S_1 and S_2, respectively. Given the reconstructed complete point cloud $\tilde{y}$ and the ground truth y, the reconstruction loss is defined as: $\mathcal{L}_{\text{recon}} = CD(y,\tilde{y})$
Fidelity loss: "Fidelity loss" is used to measure how much geometrical details of a partial observation are preserved in the complete prediction. Specifically, the fidelity loss is defined as a Unidirectional Chamfer Distance (UCD) as shown in Eq. 6:

$$\mathcal{L}_{\text{fidelity}} = \frac{1}{|x|}\sum_{p_1\in x}\min_{p_2\in\tilde{x}}\|p_1-p_2\|_2, \tag{6}$$

where $\tilde{x}$ is the reconstructed complete point cloud from the inferred complete code $\tilde{z}_{x\to y}$, and x is a partial shape input.
Adversarial loss: We introduce adversarial training to our framework with a discriminator parameterized by a deep neural network to achieve regularizing point predictions to be complete shapes. Specifically, we define adversarial loss following [13, 20] as:

$$\mathcal{L}_{\text{adv}} = \mathbb{E}[\min(0,-1+D_\gamma(\tilde{y}))] + \mathbb{E}[\min(0,-1-D_\gamma(\tilde{x}))] \tag{7}$$

where $\mathcal{D}_\gamma$ is the discriminator, $\tilde{x}$ and $\tilde{y}$ are point clouds reconstructed given partial input and complete input, respectively.
EBM loss: The EBM parameters θ can be updated with $\Delta\theta$ in Eq. 4. Additionally, we follow [10] to add a weak ℓ_2 normalization term on the energy magnitude for numerical stability. Therefore, the loss function for the latent transport module is defined as:

$$\mathcal{L}_{\text{ebm}} = -L_{\text{ebm}} + \lambda\left(E_\theta(z_y)^2 + E_\theta(\tilde{z}_{x\to y})^2\right) \tag{8}$$

where λ is the weight for the regularization term, and we define $\lambda = 0.1$ empirically.

Given the definition of the above four types of loss functions, we train the proposed model using the following steps: **1)** For partial shape x and complete shape y, the encoder generates their corresponding latent code z_x and z_y; **2)** The energy-based transport translates z_x to its corresponding complete shape latent code $\tilde{z}_{x\to y}$, where the Langevin step size $\delta^2 = 0.05$; **3)** The decoder reconstructs $\tilde{x}$ and $\tilde{y}$ from $\tilde{z}_{x\to y}$ and z_y, representing the reconstructed complete shape of x and that of y; **4)** The encoder-decoder framework with parameters $\{\alpha,\beta\}$ is updated with the loss function $\mathcal{L}_{\text{ed}} = \mathcal{L}_{\text{recon}} + \lambda_1\mathcal{L}_{\text{fidelity}} - \lambda_2\mathbb{E}[\mathcal{D}_\gamma(\tilde{x})]$ where the last term $(-\mathbb{E}[\mathcal{D}_\gamma(\tilde{x})])$ is adopted from Miyato *et al.* [21] to regularize the distribution of predicted point clouds in the complete shape domain. Empirically, we set $\lambda_1 = 2$ and $\lambda_2 = 1$; **5)** The latent transport module is updated with $\mathcal{L}_{\text{ebm}}$; **6)** The discriminator is updated with $\mathcal{L}_{\text{adv}}$.

During testing, given the partial point cloud x, we first obtain its latent code $z_x = \mathcal{E}_\alpha(x)$. Then we feed it to the latent transport module with an initial state of $r^0_{xy} = 0$. We run $K = 8$ Langevin steps to obtain the residual $r_{xy} = r^K_{xy}$, which is used to generate the complete shape code $\tilde{z}_{x\to y}$ via Eq. 3. Finally, we generate the complete shape $\tilde{x} = \mathcal{D}_\beta(\tilde{z}_{x\to y})$.

Table 1: Shape Completion results of supervised (upper three rows) and unsupervised (lower five rows) methods on the 3D-EPN dataset. The numbers shown are CD ↓ scaled by 10^4.

Method	Avg.	Plane	Cabinet	Car	Chair	Lamp	Sofa	Table	Boat
3D-EPN [9]	29.1	60.0	27.0	24.0	16.0	38.0	45.0	14.0	9.0
FoldingNet [38]	9.2	2.4	8.5	7.2	10.3	14.1	9.1	13.6	8.8
PCN [40]	7.6	2.0	8.0	5.0	9.0	13.0	8.0	10.0	6.0
Pcl2Pcl [7]	17.4	4.0	19.0	10.0	20.0	23.0	26.0	26.0	11.0
C4C. [32]	14.3	3.7	12.6	8.1	14.6	18.2	26.2	22.5	8.7
ShapeInv. [42]	23.6	4.3	20.7	11.9	20.6	25.9	54.8	38.0	12.8
Cai *et al.* [7]	13.6	3.5	12.2	9.0	12.1	17.6	26.0	19.8	8.6
Ours	**9.4**	**2.3**	**12.2**	**5.8**	**12.0**	**12.8**	**10.3**	**13.8**	**5.7**

4 Experiments

4.1 Implementation Details

We implement the encoder with the PointNet++ [26] set abstraction operation and point transformers [43]. There are three set abstraction layers in our encoder; following each is a point transformer block. The decoder consists of a latent code projection layer, three self-attention blocks [30], and a multi-layer perceptron [15] with one hidden layer. The discriminator is implemented in the same way as PU-GAN [19].

4.2 Performance on ShapeNet Dataset

Dataset. In the footstep of [7, 32], we evaluate our method on the 3D-EPN dataset [9], which is a point cloud completion benchmark derived from the ShapeNet [5] dataset. Eight incomplete shapes are produced for each 3D shape by projecting a 3D shape to a 2.5D depth image with one of eight fixed viewpoints and then back-projected to 3D coordinates, where 2048 points are uniformly sampled from object surfaces.

Quantitative and qualitative evaluation. Table 1 shows performance comparison on the 3D-EPN dataset, where we adopt Chamfer Distance (CD) as the metric, and compare with both unsupervised methods [7, 32, 42] and supervised methods [9, 38, 40]. Table 1 explains that our method outperforms existing unsupervised methods with a large margin. Specifically, our method achieves the best result in all categories and surpasses Cai *et al.* [7] by 4.2 on the average CD metric. Figure 2 illustrates the qualitative results of our method compared with the competitive techniques. In general, there are many outliers in the prediction of Cycle4Completion [32] while ShapeInversion [42] struggles to yield shapes with uniformly distributed points. It can be clearly observed that our method can achieve better quality with more uniform point distribution and complete geometries.

4.3 Performance on Real-World Datasets

To test the generalization ability of our model on real-world data, we extract partial objects from MatterPort3D [4], ScanNet [8], and KITTI [11]. Our model, as well as other competitive state-of-the-art unsupervised methods, are trained on the 3D-EPN dataset without further fine-tuning. As Table 2 indicates, our model can outperform Cycle4Completion [32] by a significant margin on all datasets. However, since ShapeInversion [42] directly minimizes

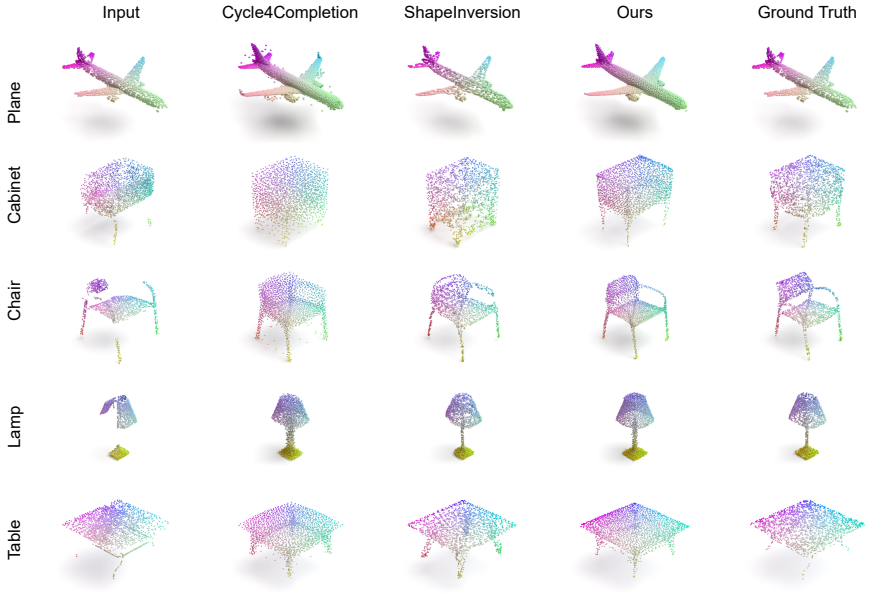

Figure 2: Qualitative Result on the 3D-EPN dataset. From left to right by column: input incomplete point clouds, results from Cycle4Completion [32] ShapeInversion [42], ours, and the ground truth.

the UCD metric between partial and prediction pairs, the results are comparable to ours on real-world datasets.

4.4 Model Analysis

Uncertainty Estimation: As the first generative unsupervised completion model, our framework can provide meaningful uncertainty maps that summarize the stochasticity of the inference process. Specifically, an uncertainty map is generated by performing multiple times of inferences, and we define the variance of the multiple reconstructed complete point clouds as uncertainty following [16]. Figure 3 visualizes sample uncertainty maps by rendering the per-point mean as a point and colorizing points with a heatmap converted from the magnitude of variance. The general trend is as follows: 1) an area with low uncertainty indicates that

Table 2: Shape completion on real-world scans. The numbers shown are Unidirectional Chamfer Distance (UCD)↓ scaled by 10^4. The *sup* indicates supervised or unsupervised method.

Method	sup.	ScanNet		MatterPort3D		KITTI
		Chair	Table	Chair	Table	Car
GRNet [34]	yes	1.6	1.6	1.6	1.5	2.2
PoinTr [39]	yes	1.7	1.5	1.8	1.3	1.9
C4C. [32]	no	12.0	31.0	12.0	34.0	13.7
ShapeInv. [42]	no	5.0	3.0	5.0	3.0	6.0
Ours	no	4.0	3.0	4.0	3.0	7.3

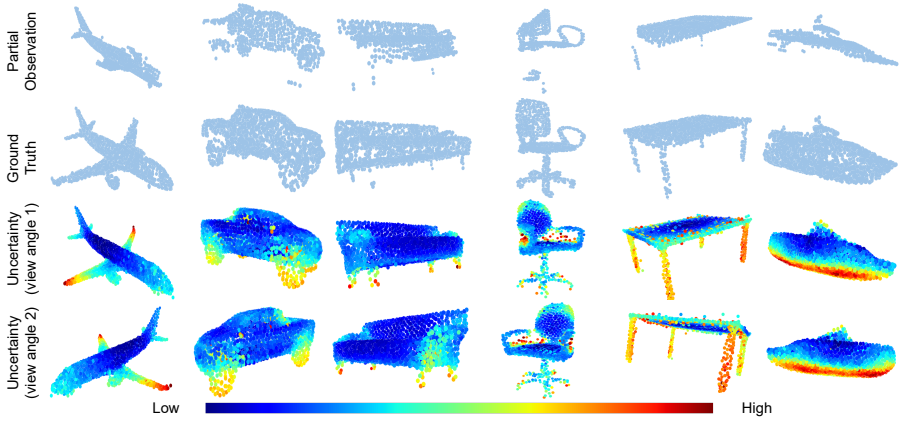

Figure 3: Uncertainty maps. The first and second rows show partial observations and their complete ground truth, and the third and fourth rows show two views of uncertainty maps.

Table 3: Effect of different modules. CD↓ scaled by 10^4 are reported here

Method	Chair	Lamp	Sofa	Table	Avg.
w/o EB transport	13.6	14.3	11.9	14.7	13.6
w/o residual sampling	14.4	14.6	12.2	16.6	14.5
w/o adversarial loss	16.6	19.5	17.7	25.5	19.8
Full Model	**12.0**	**12.8**	**10.3**	**13.8**	**12.2**

this region or its symmetric correspondence has been observed in the partial shape so that our model is confident with its prediction; 2) unobserved regions tend to have higher uncertainty since there can be multiple variants to form a valid completion. For example, the plane wing is not observed in the partial shape, so our model assigns higher uncertainty scores when approaching the end of the wing. Note that our model produces high uncertainty on outliers, such as the abnormal points in the table or chair sample.

Effect of Main Modules: To comprehensively analyse the proposed framework, we provide ablation studies to show the effectiveness of our proposed methods. There are mainly three components that we deployed to the encoder-decoder backbone, *i.e.*, energy-based latent transport, residual sampling, and adversarial loss. We study their effect by removing one of them from the full model. Table 3 shows the full model's performance and performance of models with one of the above modules removed on the 3D-EPN [9] dataset.

For each of the models with one module removed, we observe deteriorated performance on the four test categories, demonstrating the effectiveness of each module. Note that the first row of Table 3 indicates the capability of the encoder-decoder backbone. The backbone is capable of predicting complete shapes with designed losses as well as the adversarial training framework. The energy-based latent transport module boosts the performance by 1.4 in terms of the average CD. However, suppose the model is not trained with the residual sampling strategy, in that case, the latent-space EBM can even harm the performance, as shown in the second row of Table 3. The model without adversarial loss performs the worst since the adversarial loss regularizes predictions to be the complete shapes, and if it is removed, the model tends to reconstruct the input (partial shapes) instead of completing the geometries.

5 Conclusions

We present the first generative framework for unsupervised point cloud completion that is capable of generating prediction uncertainty maps. Specifically, we introduce a latent code transport module based on a conditional EBM formulation for partial-to-complete transform. A novel residual sampling strategy is proposed to facilitate end-to-end training for a latent-space EBM. Our experiments show that the proposed method outperforms existing state-of-the-art models on both synthetic and real-world datasets.

References

[1] David H Ackley, Geoffrey E Hinton, and Terrence J Sejnowski. A learning algorithm for boltzmann machines. *Cognitive science*, 9(1):147–169, 1985.

[2] Yingjie Cai, Kwan-Yee Lin, Chao Zhang, Qiang Wang, Xiaogang Wang, and Hongsheng Li. Learning a structured latent space for unsupervised point cloud completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5543–5553, June 2022.

[3] Zhen Cao, Wenxiao Zhang, Xin Wen, Zhen Dong, Yu-Shen Liu, and Bisheng Yang. Mfm-net: Unpaired shape completion network with multi-stage feature matching. *CoRR*, abs/2111.11976, 2021. URL https://arxiv.org/abs/2111.11976.

[4] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *International Conference on 3D Vision (3DV)*, 2017.

[5] Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. ShapeNet: An Information-Rich 3D Model Repository. Technical Report arXiv:1512.03012 [cs.GR], Stanford University — Princeton University — Toyota Technological Institute at Chicago, 2015.

[6] R. Qi Charles, Hao Su, Mo Kaichun, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 77–85, 2017. doi: 10.1109/CVPR.2017.16.

[7] Xuelin Chen, Baoquan Chen, and Niloy J Mitra. Unpaired point cloud completion on real scans using adversarial training. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020.

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

[9] Angela Dai, Charles Ruizhongtai Qi, and Matthias Nießner. Shape completion using 3d-encoder-predictor cnns and shape synthesis. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017.

[10] Yilun Du and Igor Mordatch. Implicit generation and modeling with energy based models. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[11] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research (IJRR)*, 2013.

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

[13] Ishaan Gulrajani, Faruk Ahmed, Martín Arjovsky, Vincent Dumoulin, and Aaron C. Courville. Improved training of wasserstein gans. *CoRR*, abs/1704.00028, 2017. URL http://arxiv.org/abs/1704.00028.

[14] 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*, 43(12):4338–4364, 2021. doi: 10.1109/TPAMI.2020.3005434.

[15] Simon Haykin. *Neural networks: a comprehensive foundation*. Prentice Hall PTR, 1994.

[16] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems*, 2017.

[17] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

[18] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In Yoshua Bengio and Yann LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

[19] Ruihui Li, Xianzhi Li, Chi-Wing Fu, Daniel Cohen-Or, and Pheng-Ann Heng. Pu-gan: a point cloud upsampling adversarial network. In *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[20] Jae Hyun Lim and Jong Chul Ye. Geometric GAN. *CoRR*, abs/1705.02894, 2017. URL http://arxiv.org/abs/1705.02894.

[21] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018.

[22] Jiquan Ngiam, Zhenghao Chen, Pang Wei Koh, and Andrew Y. Ng. Learning deep energy models. In Lise Getoor and Tobias Scheffer, editors, *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011*, pages 1105–1112. Omnipress, 2011.

[23] 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 (CVPR)*, pages 8524–8533, June 2021.

[24] Bo Pang, Tian Han, Erik Nijkamp, Song-Chun Zhu, and Ying Nian Wu. Learning latent space energy-based prior model. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

[25] Charles R. Qi, Or Litany, Kaiming He, and Leonidas J. Guibas. Deep hough voting for 3d object detection in point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[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 I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[27] Shi Qiu, Saeed Anwar, and Nick Barnes. Semantic segmentation for real point cloud scenes via bilateral augmentation and adaptive fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1757–1767, 2021.

[28] Shi Qiu, Yunfan Wu, Saeed Anwar, and Chongyi Li. Investigating attention mechanism in 3d point cloud object detection. In *International Conference on 3D Vision (3DV)*. IEEE, 2021.

[29] Shi Qiu, Saeed Anwar, and Nick Barnes. Geometric back-projection network for point cloud classification. *IEEE Transactions on Multimedia*, 24:1943–1955, 2022.

[30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[31] X. Wen, T. Li, Z. Han, and Y. Liu. Point cloud completion by skip-attention network with hierarchical folding. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1936–1945, Los Alamitos, CA, USA, jun 2020. IEEE Computer Society. doi: 10.1109/CVPR42600.2020.00201. URL https://doi.ieeecomputersociety.org/10.1109/CVPR42600.2020.00201.

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

[33] Peng Xiang, Xin Wen, Yu-Shen Liu, Yan-Pei Cao, Pengfei Wan, Wen Zheng, and Zhizhong Han. Snowflakenet: Point cloud completion by snowflake point deconvolution

with skip-transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5499–5509, October 2021.

[34] 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.

[35] Jianwen Xie, Yang Lu, Song-Chun Zhu, and Ying Nian Wu. A theory of generative convnet. In Maria-Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 2635–2644. JMLR.org, 2016. URL http://proceedings.mlr.press/v48/xiec16.html.

[36] Jianwen Xie, Yifei Xu, Zilong Zheng, Song-Chun Zhu, and Ying Nian Wu. Generative pointnet: Deep energy-based learning on unordered point sets for 3d generation, reconstruction and classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14976–14985, June 2021.

[37] Jianwen Xie, Song-Chun Zhu, and Ying Nian Wu. Learning energy-based spatial-temporal generative convnets for dynamic patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(2):516–531, 2021. doi: 10.1109/TPAMI.2019.2934852.

[38] Yaoqing Yang, Chen Feng, Yiru Shen, and Dong Tian. Foldingnet: Point cloud auto-encoder via deep grid deformation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 206–215, 2018. doi: 10.1109/CVPR.2018.00029.

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

[40] 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, 2018. doi: 10.1109/3DV.2018.00088.

[41] Jing Zhang, Jianwen Xie, Nick Barnes, and Ping Li. Learning generative vision transformer with energy-based latent space for saliency prediction. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=LoUdcqLuPej.

[42] Junzhe Zhang, Xinyi Chen, Zhongang Cai, Liang Pan, Haiyu Zhao, Shuai Yi, Chai Kiat Yeo, Bo Dai, and Chen Change Loy. Unsupervised 3d shape completion through gan inversion. In *CVPR*, 2021.

[43] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip H.S. Torr, and Vladlen Koltun. Point transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16259–16268, October 2021.

[44] Yang Zhao and Changyou Chen. Unpaired image-to-image translation via latent energy transport. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16418–16427, June 2021.

[45] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.