P2M2-Net: Part-Aware Prompt-Guided Multimodal Point Cloud Completion

Linlian Jiang¹, Pan Chen¹, Ye Wang¹, Tieru Wu^{1,2}, and Rui Ma^{1,2,*}

¹ Jilin University

 $^2\,$ Engineering Research Center of Knowledge-Driven Human-Machine Intelligence,

MOE

{jiangll21, chenpan21, yewang22}@mails.jlu.edu.cn
 {wutr, ruim}@jlu.edu.cn

Abstract. Inferring missing regions from severely occluded point clouds is highly challenging. Especially for 3D shapes with rich geometry and structure details, inherent ambiguities of the unknown parts are existing. Existing approaches either learn a one-to-one mapping in a supervised manner or train a generative model to synthesize the missing points for the completion of 3D point cloud shapes. These methods, however, lack the controllability for the completion process and the results are either deterministic or exhibiting uncontrolled diversity. Inspired by the prompt-driven data generation and editing, we propose a novel prompt-guided point cloud completion framework, coined P2M2-Net, to enable more controllable and more diverse shape completion. Given an input partial point cloud and a text prompt describing the part-aware information such as semantics and structure of the missing region, our Transformer-based completion network can efficiently fuse the multimodal features and generate diverse results following the prompt guidance. We train the P2M2-Net on a new large-scale PartNet-Prompt dataset and conduct extensive experiments on two challenging shape completion benchmarks. Quantitative and qualitative results show the efficacy of incorporating prompts for more controllable part-aware point cloud completion and generation. Code and data are available at https://github.com/JLU-ICL/P2M2-Net.

Keywords: Multimodal · Point Cloud Completion.

1 Introduction

Point cloud is one of the most commonly used 3D shape representations, which requires less memory to store detailed geometry and structural information about a 3D shape. Nowadays, point clouds can easily be obtained through depth cameras or other 3D scanning devices. However, due to the resolution of the scanning devices, occlusion or the limitation of accessible scanning regions, raw point clouds are often sparse or incomplete. Point cloud completion aims to completing

^{*} Rui Ma is the corresponding author.





Fig. 1. Given an input point cloud with a large missing region (left column), our P2M2-Net can use different text prompts to guide the shape completion and generate diverse outputs in a controllable manner.

the geometry and structure of the missing region given the partial point cloud as input. When the missing region is significantly large, inherent *ambiguities* may exist when performing the completion, i.e., there may be multiple options for the missing parts. How to obtain the expected completion result is a challenging task for the point cloud completion.

Existing point completion methods [47,48,46,40,41,53,43] usually take the incomplete point cloud of a 3D shape as input and train an encoder-decoder network to map the input to a complete shape. However, it is difficult to learn the mapping due to the sparsity and limited information from the input. One way to provide more information for point cloud completion is using images as the guidance [51,10,16,18,39]. Though ambiguities can be resolved and improved performance can be obtained by considering more image constraints during the completion, the images and point clouds need to be matched so that features from the image modality can be used to complete the missing features for the corresponding point clouds. Furthermore, as no explicit semantic and structure information is considered in these methods, their completion mainly focus on the global geometry level.

On the other hand, the point cloud completion can also be regarded as a conditional generation problem, in which the complete shape is generated based on the input partial point cloud. For example, some methods [42,4,54,1,49] work

on the multimodal shape completion with the goal of generating diverse 3D shapes from a single partial point cloud. In this way, the point cloud completion is modeled as a one-to-many mapping which allows multiple outputs as long as the completed shapes respect to the input and their geometry and structure are plausible. Although diverse results can be obtained from these generative approaches, it is hard to generate a specific complete shape that meets the expectation or requirement of the user. How to allow the user to control or guide the completion in an intuitive and efficient manner is worth to investigate.

In this paper, we aim for controllable point cloud completion that can accept a simple form of guidance (e.g., text prompt) to generate plausible outputs that satisfy the user's specification. Also, in addition to the global geometry, we attempt to explicitly focus on the part semantics and structures when performing the completion. Such part-aware modeling can allow more fine-grained control on the completion process. To this end, we propose P2M2-Net, a novel part-aware prompt-guided multimodal point cloud completion framework, to enable more controllable and diverse shape completion. With a text prompt describing the part-aware information such as semantics and structure of the missing region, the P2M2-Net can efficiently fuse the features from two modalities, i.e., 3D point clouds and text prompts, and predict a complete shape that matches to the text prompt. The word *multimodal* in our paper has two kinds of meanings: one indicates the completion is based on features from multiple data modalities; the other one represents that multiple different shapes can be generated when different text prompts are used as guidance for the same input point cloud (see Figure 1).

To enable the joint learning between text prompts and the 3D point data, we construct a novel large-scale dataset, named PartNet-Prompt, which contains part-level text prompt annotations for three representative shape categories (chair, table and lamp) from the PartNet dataset [22]. Each prompt is a short text phrase that describes the geometry or structure of the corresponding part, such as *inclined back*, *straight legs* etc. With the paired data from our PartNet-Prompt dataset, we first perform a cross-modal contrastive pre-training to align the part-level features of the text and 3D point. For the prompt-guided completion network, we adopt a Transformer-based network PoinTr [46] and adapt it to perform point cloud completion using multimodal features. A new multimodal feature encoder is proposed to extract the feature of each modality and then fuse them together using a attention-based feature fusion module. Next, the fused multimodal feature is passed to the multimodal query generator and the multimodal-based point cloud decoder to predict the complete shape in a coarse-to-fine manner.

To evaluate the performance of our P2M2-Net, we conduct extensive experiments on two challenging PartNet-based point cloud completion benchmarks. Quantitative and qualitative comparisons with the state-of-the-art methods show the the efficacy of incorporating prompts for the guided completion. We also perform ablation studies on the cross-modal pre-training and the attention-based multimodal feature fusion and the results verify the effectiveness of each module. 4 Linlian Jiang, Pan Chen, Ye Wang, Tieru Wu, and Rui Ma

In summary, our contributions are as follows:

- 1. We build the PartNet-Prompt, a novel large-scale dataset with part-level text prompt annotations. With the paired cross-modal data on the semantics and structures of shape parts, various applications such as part-aware point cloud completion and generation as well as fine-grained shape understanding and retrieval can be supported.
- 2. We propose P2M2-Net, a novel part-aware prompt-guided framework which can achieve the point cloud completion in a more controllable manner. A contrastive pre-training and a multimodal feature encoder are proposed to better align and fuse the cross-modal features. Once trained, when guided by different text prompts, the P2M2-Net can generate diverse results from a single input.
- 3. Extensive experiments on two challenging PartNet-based shape completion benchmarks demonstrate the superiority of P2M2-Net comparing to the state-of-the-art point cloud completion methods. Also, our prompt-guided completion can also be regarded as cross-modal compositional modeling and the diverse results show its potential in generating novel shapes.

2 Related Work

2.1 3D Point Cloud Shape Completion

Point cloud completion for 3D shapes has been widely studied in recent years. The task is usually modeled as a one-to-one mapping which outputs a deterministic complete shape from a given input. To learn the mapping, the input partial point cloud is often encoded into a feature vector using conventional point-based encoders such as [29,30,27]. With the encoded point feature, PCN [47] designs a multi-stage decoder which first predict a coarse complete shape and then employs the FoldingNet [45] to refine the initial result. Furthermore, Transformer-based encoder-decoder [46,53] which can learn more comprehensive relationships among the points has also been investigated. Comparing to these methods, since we learn transferrable features via the cross-modal pre-training, we can enable one-to-many mapping by using different text prompt as guidance.

Meanwhile, some methods [42,4,54,1,49] formulate the point cloud completion as a shape generation problem and employ generative models to obtain diverse completion results. These methods can inherently learn a one-to-many mapping, but their completion is not controllable. For example, the multimodal point cloud completion (or MPC) [42] develops the first shape completion framework which can generate multimodal (i.e., diverse) results based on the conditional generative modeling, but it is difficult to incorporate user's constraints into the completion process. In contrast, our method allows intuitive user control in the form of text prompt and we can also generate diverse outputs when different text prompts are used.

2.2 Multimodal-based Point Cloud Completion

Since there are inherent ambiguities when completing the partial point cloud, information from other data modalities may be used to guide the completion process. In ViPC [51], a single-view image that matches to the target shape is used to provide the information about missing region. The additional image information has also been explored in the completion of RGB-D scenes which contain severe missing data due to the occlusion. With aligned RGB and depth images, some approaches [10,16,18,39] propose different schemes to fuse the information from the multimodal input data. Our method also takes the advantage of using multimodal input to alleviate the ambiguities for the completion. Instead of the image, we utilize the text prompt which is more flexible to provide the shape completion guidance. To resolve the domain gap between the text prompt and the 3D point cloud, we adopt the part-level cross-modal pre-training to obtain aligned and transferable features for each data modality. Meanwhile, our multimodal Transformer can also efficiently fuse the text and point features and generate the completion result based on the multimodal input data.

2.3 Prompt-Driven Multimodal Learning

Recently, the prompt-driven multimodal learning has attracted great attention for its applications in zero-shot learning [31,13], 2D/3D visual perception [11,14], content generation [32,33,28,25,20,5] and editing [15,21,34]. With text or other types of the prompt [50,24], impressive 2D images and 3D models can be generated in a controllable manner. One key for the success of prompt-driven learning is the large-scale pre-trained model such as CLIP [31] which learns aligned features from paired data of two modalities. To enable the multimodal learning between the text prompt and 3D shapes, we construct PartNet-Prompt dataset by manually annotating the 3D parts of representative PartNet shapes with text descriptions about their geometry and structure. With such part-level annotation, we can achieve more fine-grained control in prompt-guided part-aware completion and generation.

2.4 Cross-Modal Contrastive Pre-Training

Due to the difference between the data modalities, there is a large domain gap between the features of point clouds and text prompts. To facilitate the multimodal feature fusion, the 3D point cloud and text features need to be aligned into the same embedding space. Contrastive pre-training methods [37,7,23,19,31,13,17,44,3,12] has been widely used to learn aligned and transferrable features for data of different modalities, e.g., images and natural language. For contrastive pre-training that involves 3D point clouds, CrossPoint [3] jointly learns the aligned representations of 3D point cloud shapes and their corresponding images, while the learned representations are used for point cloud understanding tasks such as 3D classification and segmentation. However, to the best our knowledge, contrastive pre-training has not been sufficiently explored for joint learning of point cloud and text features, nor the prompt-guided point cloud completion task.



Fig. 2. Overview of P2M2-Net. (a) Illustration of the cross-modal contrastive pretraining. Embeddings of the same part sare pulled closer; (b) Pipeline of the multimodal Transformer for prompt-guided completion.

 Table 1. The statistics of PartNet-Prompt dataset.

Category		C	Chair		Tabl	e	Lamp			
#Shape		8	,176		9,90	3,408				
# Part		22	2,175		17,83	9,201				
Part-type	Back	Seat	Armrest	Leg	Tabletop	Leg	Head	Post	Base	
#Prompt-type	12	7	9	22	6	16	8	6	6	
#Annotation	7,553	$5,\!384$	2,506	$4,\!664$	9,272	8,558	3,431	$2,\!460$	$3,\!310$	

3 Method

6

In this section, we introduce the PartNet-Prompt dataset and present the details of the P2M2-Net. Figure 2 illustrates an overview of our pipeline which contains two stages of training: contrastive pre-training and multimodal Transformer for feature fusion and point cloud decoding.

3.1 PartNet-Prompt Dataset

To enable part-level joint learning on the 3D shape and text prompt, we propose a large-scale PartNet-Prompt dataset which contains paired data of text prompt and their corresponding parts. To construct the dataset, we manually annotate short text prompts for the pre-segmented parts of the Table, Chair, and Lamp categories based on their semantic segmentation from the original PartNet [22] dataset. A pre-defined set of vocabularies is combined with the semantic label of a part to describe part-level geometry, structure and semantics, e.g., curved back, single-rod leg, cylindrical head etc. In total, the PartNet-Prompt contains part-level text prompt annotations for 8,176 chairs, 9,906 tables and 3,408 lamps, and the numbers of annotated parts are 22,175, 17,830 and 9,201, respectively. The dataset is further split to train, validation and test set with the ratio 7:1:2 following the PartNet. The statistics of the PartNet-Prompt dataset is shown in Table 1.

3.2 Contrastive Pre-Training for Prompt-Guided Completion

Following the general idea of CrossPoint [3], we design a contrastive pre-training module that is specifically aiming for part-aware prompt-guided point cloud completion. As shown in Figure 2, this module contains a point cloud encoder and a text prompt encoder which can be some established networks, as well as two MLP-based projection heads to further map each modality into the target embedding space. In the following, we introduce the details of the contrastive pre-training module.

Point cloud encoder. We design an encoder f_P that combines DGCNN [27] and Point Transformer [52] to extract the point cloud feature. Given a point cloud w.r.t. a particular shape part, DGCNN is used to extract the local features for a set of points sampled by farthest point sampling (FPS). Then, the global point cloud feature is obtained using Point Transformer in which self-attention is applied to mine the relationships among the local features.

Text prompt encoder. Given a text prompt that describes the geometry, structure and semantics of the corresponding part, we apply BERT [8] as the text prompt encoder f_T to extract a 768-dim vector as the initial feature for the text prompt.

Contrastive pre-training of point cloud and text prompt. With the initial features extracted by point cloud encoder f_P and text prompt encoder f_T , the point cloud projection head g_P and text projection head g_T further map the features into a unified embedding space R^{256} . For a shape part represented by point cloud p_i and text prompt t_i , we denote the final feature embedding of each modality as $z_P^i = g_P^i(f_P^i(p_i))$ and $z_T^i = g_T^i(f_T^i(t_i))$. To align the point cloud and text prompt features, contrastive pre-training adopts the InfoNCE loss [26] to fine-tune the encoders f_P , f_T and projection heads g_P and g_T . Formally, the InfoNCE loss for contrastive pre-training is defined as:

$$L_{InfoNCE} = (L_{P \to T} + L_{T \to P})/2, \quad where \tag{1}$$

$$L_{P \to T} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp(s(z_P^i, z_T^i)/\tau)}{\sum_{k=1}^{N} \exp(s(z_P^i, z_T^k)/\tau)},$$
(2)

and $L_{T \to P}$ is defined similarly as $L_{P \to T}$. Here, $s(\cdot, \cdot)$ is the cosine similarity between features of cross-modal pairs and τ is a temperature parameter which can adjust the feature learning. By performing the cross-modal pre-training using $L_{InfoNCE}$, the final embeddings of the same shape part, e.g., z_P^i and z_T^i are pulled closer, and the embeddings of the different parts, e.g., z_P^i and z_T^i , (when $i \neq k$), are repelled far away from each other. Similar to CrossPoint [3], after pre-training, the features extracted by the encoders f_P and f_T are passed to the downstream point cloud completion task.

3.3 Multimodal Transformer for Point Cloud Completion

To achieve the prompt-guided point cloud completion, we extend the PoinTr [46] to work with multimodal features.

Multimodal feature encoder. With the features extracted by the point cloud and text prompt encoders, we design a multimodal feature fusion module which contains self-attention and cross-attention layers as in [38] to fuse the features into a 1024-dim multimodal feature f_M . Moreover, a coarse point cloud which can be used to guide the following completion process is also generated by applying two additional linear layers to f_M and reshaping the output.

Multimodal query generator. The query generator, which is composed of three Conv1D layers, takes the fused multimodal feature f_M and coarse point cloud as input and generate a sequence of multimodal query proxies which can be used to query the related features from f_M . Similar to PointTr [46], the coarse point cloud is utilized to provide the spatial coordinates when generating the query proxies.

Multimodal-based point cloud decoder. With the multimodal query proxies and the fused multimodal feature f_M as input, the point cloud decoder first predicts a sequence of predicted proxies which contain the multimodal information. When performing the querying, a kNN model is used to query the multimodal features around each point in the coarse point cloud. Then, the FoldingNet [45] is employed to recover detailed local shapes centered around the generated proxies. Note that we only predict the missing part that matches to the text prompt and concatenate the output to the input point cloud to obtain a complete shape.

Training Details. The P2M2-Net is trained in two stages using the annotated data from the PartNet-Prompt dataset. Each part is uniformly sampled into 1024 points similar to MPC [42]. Separate encoders or models are trained during the contrastive pre-training stage and the prompt-guided completion stage, using the parts from the training set of Chair, Table and Lamp, respectively. For the contrastive pre-training stage, the InfoNCE loss $L_{InfoNCE}$ is used and we train the point cloud encoder from scratch and the text prompt encoder based on the pre-trained BERT. The pre-training is running for 2000 epoches. For the prompt-guided completion stage, we use the Chamfer Distance (CD) loss and train the multimodal Transformer-based network for 500 epoches.

4 Experiments

We conduct quantitative and qualitative evaluations of P2M2-Net on the prompteguided completion. We also perform ablation studies to verify the effectiveness of key modules in P2M2-Net.

4.1 Evaluation Metrics and Benchmarks

Following previous works [47,42,46], we adopt the following metrics for quantitative evaluation of the completion or diverse generation performance:

Chamfer Distance (CD) [9]: the average Chamfer Distance which measures the set-wise distance between the completed and the ground truth point clouds is used as a metric for evaluating the prompt-guided completion.

F1 Score (F-Score) [35]: F1 score is defined as the harmonic mean of precision and recall when performing 3D classification, reconstruction or completion. It explicitly evaluates the distance between the predicted points and the GT points and intuitively measures the percentage of points that is predicted correctly. We set the distance threshold d = 0.01 when computing the F1 score.

Total Mutual Difference (TMD) [42] : given k completion results for an input, the mutual difference d_i is defined as the average Chamfer Distance between the i-th shape to the other k-1 shapes. Then, the Total Mutual Difference is define as $\sum_{i=1}^{k} d_i$ to measure the **diversity** of the completion results.

Minimal Matching Distance (MMD) [2]: for each input, we calculate the minimal matching distance between 10 predicted shapes w.r.t the corresponding ground truth shape to measure the *quality* of the completion results.

Unidirectional Hausdorff Distance (UHD) [42]: we compute the average Hausdorff distance from the input partial point cloud to each of the completed shapes to measure the *fidelity* of the completion results.

Benchmarks for Point Completion. We follow the settings in MPC [42] and use their released code to generate two challenging benchmarks to evaluate the P2M2-Net for part-ware completion. The *PartNet* benchmark is obtained by removing points of randomly selected parts from the shapes in the PartNet dataset [22]. The *PartNet-Scan* benchmark is obtained by first randomly removing the parts and then virtually scanning the remaining parts. These two benchmarks both focus on part-level incompleteness and exhibit high ambiguities for the missing region.

4.2 Evaluation for One-to-One Completion

We first conduct comparisons following the conventional setting in previous methods, i.e., predicting a complete shape given a partial input. For our method, since the text prompt is needed to provide the guidance for completion, we use GT text prompt of the missing part as an additional input. At the first glance, using the GT text prompt to provide additional information seems to be unfair for other methods when conducting the comparison. Meanwhile, since our method can be regarded as a multimodal version of PoinTr [46], the comparison still can show how much improvement can be achieved by incorporating the prompt guidance comparing the the PoinTr baseline and other state-of-the-art methods.

Quantitative comparison. For each benchmark, we compare our results with representative and state-of-the-art point cloud completion methods, i.e., PCN [47], PFNet [48], FoldingNet [45], MPC [42], TopNet [36], PoinTr [46],

Mathad	C	hair	Τa	able	Lamp		
Method	CD	F-Score	CD	F-Score	CD	F-Score	
PCN [47]	2.098	0.152	3.560	0.131	9.133	0.110	
PFNet [48]	3.734	0.087	6.282	0.120	14.652	0.075	
FoldingNet [45]	2.733	0.082	5.194	0.193	12.466	0.116	
MPC [42]	2.081	0.132	4.132	0.236	10.465	0.088	
TopNet [36]	1.480	0.171	3.069	0.174	7.388	0.081	
PoinTr [46]	1.292	0.364	2.682	0.356	6.017	0.354	
PMP-Net++ [41]	1.236	0.385	2.427	0.369	5.987	0.326	
P2M2-Net	1.351	0.333	2.320	0.373	5.675	0.372	

Table 2. Quantitative comparison on the PartNet benchmark. $CD-L2(\times 10^3)$ and F-Score@0.01 are used to compare with other methods.

Table 3. Quantitative comparison on the PartNet-Scan benchmark. $CD-L2(\times 10^3)$ and F-Score@0.01 are used to compare with other methods.

Mathad	C	hair	T	able	Lamp		
Method	CD	F-Score	CD	F-Score	CD	F-Score	
PCN [47]	3.421	0.097	4.662	0.104	10.019	0.091	
PFNet [48]	4.571	0.065	7.031	0.098	15.351	0.073	
FoldingNet [45]	3.843	0.071	5.972	0.120	13.544	0.111	
MPC [42]	3.464	0.074	4.694	0.213	11.096	0.076	
TopNet [36]	2.016	0.117	3.473	0.120	6.972	0.079	
PoinTr [46]	1.325	0.359	2.990	0.343	7.623	0.301	
PMP-Net++[41]	1.294	0.375	2.641	0.357	6.132	0.318	
P2M2-Net	1.421	0.356	2.423	0.361	6.307	0.305	

and PMP-Net++ [41]. For each compared method, we use their released code and train their models on the training set of our benchmark. Table 2 and 3 show the results of quantitative comparison measured by CD and F-Score. It can be observed that for both benckmarks, our P2M2-Net outperforms most of the baseline methods and just underperforms in a few cases comparing to PoinTr and PMP-Net++. The reasons may be as follows: 1) by fusing the text prompt feature to the point cloud feature, some noise may actually be introduced since the pre-training is not perfect due to the amount of data and the ambiguity between different parts; 2) since the same text prompt may be used to annotate parts with minorly different geometry, the learned text feature may be related to an average shape of the described part. Nevertheless, by incorporating with the text prompt, our method can guide the completion with the expected geometry and structure, while makes the completion more controllable.

Qualitative comparison. Figure 3 shows the qualitative results of methods which achieve relative high performance in the quantitative comparison, i.e., TopNet[36], PoinTr[46] and PMP-Net++[41], on two benchmark datasets, namely PartNet (dashed line on the left) and PartNet-scan (dashed line on the



Fig. 3. Qualitative comparisons on PartNet and PartNet-Scan benchmarks.

right). Among the compared methods, PMP-Net++ achieves the best performance in terms of the quality of the results and the similarity to the GT. However, all methods including PMP-Net++ cannot predict the expected shape for certain cases, such as the chair leg in the second column and the back with rounded corners in the fifth column of Figure 3. In contrast, our method can generate the expected shape when the text prompt is used as guidance.

4.3 Evaluation for Multimodal (Diverse) Completion

With different text prompts as input, our P2M2-Net can also generate multimodal (or diverse) completion results. Table 4 shows the quantitative comparisons with MPC [42] and the baselines used in their paper. Since we use the same benckmarks *PartNet* and *PartNet-Scan* as MPC, we directly compare with the metrics reported in their paper. For other methods, we randomly select one text prompt from the set of prompt annotations for the corresponding part type and generate the result with the prompt-guidance. From the results, our method achieves the best MMD (quality), TMD (diversity) and UHD (fidelity) metrics. Figure 4 shows qualitative comparisons with MPC and our results achieve both superior diversity and quality. In Figure 5, more qualitative results are provided. It can be seen our method can generate diverse, plausible or even novel shapes with different prompts.





Fig. 4. Qualitative comparison with MPC [42] for shapes in the PartNet benchmark.



Fig. 5. Qualitative results on PartNet and PartNet-Scan. We visualize the generated results from different prompts. P2M2-Net not only preserves the originally observed structure but also achieves diverse generated results that comply with the prompt.

4.4 Ablation Studies

We conduct ablation studies to evaluate the key modules of our framework: cross-modal pre-training and attention-based feature fusion. We implement two variations of methods that have the corresponding module disabled. The baseline model A directly uses features from pre-trained models of DGCNN [27] and BERT, and performs the feature fusion by simple concatenation. Such model A can be regarded as a simple multimodal extension of PoinTr [46]. The baseline model B performs the cross-modal pre-training but still uses the simple

Table 4. Quantitative comparison for multimodal completion on PartNet benchmark. Note that MMD (quality), TMD (diversity) and UHD (fidelity) are multiplied by 10^3 , 10^2 and 10^2 , respectively.

PartNet	MMD↓				TMD↑		UHD↓			
Method	Chair	Table	Lamp	Chair	Table	Lamp	Chair	Table	Lamp	
pcl2pcl [6]	1.90	1.90	2.50	0.00	0.00	0.00	4.88	4.64	4.78	
KNN-latent	1.39	1.30	1.72	2.28	2.36	4.18	8.58	7.61	8.47	
MPC [42]	1.52	1.46	1.97	2.75	3.30	3.31	6.89	5.56	5.72	
P2M2-Net	1.35	1.39	1.62	3.07	3.51	3.41	2.65	2.53	3.24	

Table 5. Quantitative ablation study on the PartNet benchmark. The effectiveness of cross-modal pre-training (Pre-train) and attention-based feature fusion (Attention) are evaluated. We investigate different designs including Pre-train Module (Pre-train) and Multi-modal Fusion Module (Attention)

Model	Pro train	Attention	Chair			Table			Lamp		
Model	I le-train A		CD	TMD	UHD	CD	TMD	UHD	CD	TMD	UHD
А			1.669	0.26	4.58	2.714	0.19	4.37	6.208	0.22	4.63
В	 ✓ 		1.425	1.32	3.42	2.503	1.29	2.98	5.864	1.80	4.05
P2M2-Net		~	1.365	3.07	2.65	2.320	3.51	2.53	5.675	3.41	3.24

concatenation-based feature fusion. Table 5 and Figure 6 show the quantitative and qualitative results of ablation studies. It can be seen both the two modules are important to achieve diverse results while respecting to the input prompts.

To further examine the effectiveness of the cross-modal pre-training, we use t-SNE to visualize the embedding space of different features for 150 chairs with 394 parts in Figure 7. It can be seen the initial point cloud features of different parts are mixed together. This is because different parts may have similar shapes. If simply concatenating the point cloud feature with the text prompt feature without pre-training, the fused features are still not representative to corresponding parts. After performing pre-training, since the point cloud and text prompt are aligned into the same space, even if we just simply concatenate them together, the resulting features can better represent the parts as the parts with similar geometry and semantics are closer in the space.

5 Conclusion

In this paper, we propose P2M2-Net, a novel part-aware prompt-guided framework for multimodal point cloud completion. With the guidance of the text prompt, our P2M2-Net can resolve the ambiguities for the large missing region and enable more controllable completion process. Moreover, with different text prompts as input, we can also generate diverse completion results for the same partial point cloud. To enable the joint learning of point cloud and text prompts, we construct a novel large-scale dataset PartNet-Prompt which has the poten-



14 Linlian Jiang, Pan Chen, Ye Wang, Tieru Wu, and Rui Ma

Fig. 6. Qualitative ablation study on cross-modal pre-training and attention-based feature fusion modules.



Fig. 7. t-SNE visualization of features obtained with different schemes (see the main text for more details).

tial to support more part-level multimodal learning tasks such as prompt-guided generation and editing. Currently, our P2M2-Net is based on supervised learning with the paired data from PartNet-Prompt and our completion is still deterministic when the same text prompt is used to the same partial point cloud. In the future, we would like to introduce generative models such as GAN or diffusion models to achieve more diverse results for controllable part-aware shape completion and generation.

References

- Achlioptas, P., Diamanti, O., Mitliagkas, I., Guibas, L.: Learning representations and generative models for 3d point clouds. In: International conference on machine learning. pp. 40–49. PMLR (2018)
- 2. Achlioptas, P., Diamanti, O., Mitliagkas, I., Guibas, L.: Learning representations and generative models for 3d point clouds (Jul 2017)
- Afham, M., Dissanayake, I., Dissanayake, D., Dharmasiri, A., Thilakarathna, K., Rodrigo, R.: Crosspoint: Self-supervised cross-modal contrastive learning for 3d point cloud understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9902–9912 (2022)
- Arora, H., Mishra, S., Peng, S., Li, K., Mahdavi-Amiri, A.: Multimodal shape completion via implicit maximum likelihood estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2958– 2967 (2022)
- Bahmani, S., Park, J.J., Paschalidou, D., Yan, X., Wetzstein, G., Guibas, L., Tagliasacchi, A.: Cc3d: Layout-conditioned generation of compositional 3d scenes. arXiv preprint arXiv:2303.12074 (2023)
- 6. Chen, X., Chen, B., Mitra, N.: Unpaired point cloud completion on real scans using adversarial training (Apr 2020)
- Desai, K., Johnson, J.: Virtex: Learning visual representations from textual annotations. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11162–11173 (2021)
- Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
- Fan, H., Su, H., Guibas, L.J.: 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. pp. 605–613 (2017)
- Garbade, M., Chen, Y.T., Sawatzky, J., Gall, J.: Two stream 3d semantic scene completion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 0–0 (2019)
- 11. Hegde, D., Valanarasu, J.M.J., Patel, V.M.: Clip goes 3d: Leveraging prompt tuning for language grounded 3d recognition. arXiv preprint arXiv:2303.11313 (2023)
- Huang, S., Chen, Y., Jia, J., Wang, L.: Multi-view transformer for 3d visual grounding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15524–15533 (2022)
- Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International Conference on Machine Learning. pp. 4904–4916. PMLR (2021)
- Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. arXiv preprint arXiv:2304.02643 (2023)
- Li, G., Zheng, H., Wang, C., Li, C., Zheng, C., Tao, D.: 3ddesigner: Towards photorealistic 3d object generation and editing with text-guided diffusion models. arXiv preprint arXiv:2211.14108 (2022)
- Li, J., Liu, Y., Gong, D., Shi, Q., Yuan, X., Zhao, C., Reid, I.: Rgbd based dimensional decomposition residual network for 3d semantic scene completion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7693–7702 (2019)

- 16 Linlian Jiang, Pan Chen, Ye Wang, Tieru Wu, and Rui Ma
- Li, J., Selvaraju, R.R., Gotmare, A., Joty, S., Xiong, C., Hoi, S.C.H.: Align before fuse: Vision and language representation learning with momentum distillation (2021)
- Liu, Y., Li, J., Yan, Q., Yuan, X., Zhao, C., Reid, I., Cadena, C.: 3d gated recurrent fusion for semantic scene completion. arXiv preprint arXiv:2002.07269 (2020)
- Liu, Y., Fan, Q., Zhang, S., Dong, H., Funkhouser, T., Yi, L.: Contrastive multimodal fusion with tupleinfonce. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 754–763 (2021)
- Liu, Z., Wang, Y., Qi, X., Fu, C.W.: Towards implicit text-guided 3d shape generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17896–17906 (2022)
- Mikaeili, A., Perel, O., Cohen-Or, D., Mahdavi-Amiri, A.: Sked: Sketch-guided text-based 3d editing. arXiv preprint arXiv:2303.10735 (2023)
- 22. Mo, K., Zhu, S., Chang, A.X., Yi, L., Tripathi, S., Guibas, L.J., Su, H.: Partnet: A large-scale benchmark for fine-grained and hierarchical part-level 3d object understanding. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 909–918 (2019)
- Morgado, P., Vasconcelos, N., Misra, I.: Audio-visual instance discrimination with cross-modal agreement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12475–12486 (2021)
- Mou, C., Wang, X., Xie, L., Zhang, J., Qi, Z., Shan, Y., Qie, X.: T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. arXiv preprint arXiv:2302.08453 (2023)
- Nichol, A., Jun, H., Dhariwal, P., Mishkin, P., Chen, M.: Point-e: A system for generating 3d point clouds from complex prompts. arXiv preprint arXiv:2212.08751 (2022)
- van den Oord, A., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv: Learning (2018)
- Phan, A.V., Le Nguyen, M., Nguyen, Y.L.H., Bui, L.T.: Dgcnn: A convolutional neural network over large-scale labeled graphs. Neural Networks 108, 533–543 (2018)
- Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. arXiv preprint arXiv:2209.14988 (2022)
- Qi, C.R., Su, H., Mo, K., Guibas, L.J.: Pointnet: Deep learning on point sets for 3d classification and segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 652–660 (2017)
- Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. Advances in neural information processing systems **30** (2017)
- Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
- Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical textconditional image generation with clip latents. arXiv preprint arXiv:2204.06125 (2022)
- Sanghi, A., Chu, H., Lambourne, J.G., Wang, Y., Cheng, C.Y., Fumero, M., Malekshan, K.R.: Clip-forge: Towards zero-shot text-to-shape generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18603–18613 (2022)

 Sella, E., Fiebelman, G., Hedman, P., Averbuch-Elor, H.: Vox-e: Text-guided voxel editing of 3d objects. arXiv preprint arXiv:2303.12048 (2023)

17

- Tatarchenko, M., Richter, S.R., Ranftl, R., Li, Z., Koltun, V., Brox, T.: What do single-view 3d reconstruction networks learn? In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3405–3414 (2019)
- Tchapmi, L.P., Kosaraju, V., Rezatofighi, H., Reid, I., Savarese, S.: Topnet: Structural point cloud decoder. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 383–392 (2019)
- Tian, Y., Krishnan, D., Isola, P.: Contrastive multiview coding. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16. pp. 776–794. Springer (2020)
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
- Wang, X., Lin, D., Wan, L.: Ffnet: Frequency fusion network for semantic scene completion. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 2550–2557 (2022)
- 40. Wen, X., Xiang, P., Han, Z., Cao, Y.P., Wan, P., Zheng, W., Liu, Y.S.: Pmp-net: Point cloud completion by learning multi-step point moving paths. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7443–7452 (2021)
- Wen, X., Xiang, P., Han, Z., Cao, Y.P., Wan, P., Zheng, W., Liu, Y.S.: Pmp-net++: Point cloud completion by transformer-enhanced multi-step point moving paths. IEEE Transactions on Pattern Analysis and Machine Intelligence 45(1), 852–867 (2022)
- Wu, R., Chen, X., Zhuang, Y., Chen, B.: Multimodal shape completion via conditional generative adversarial networks. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16. pp. 281–296. Springer (2020)
- 43. Xiang, P., Wen, X., Liu, Y.S., Cao, Y.P., Wan, P., Zheng, W., Han, Z.: Snowflakenet: Point cloud completion by snowflake point deconvolution with skiptransformer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5499–5509 (2021)
- 44. Yang, J., Duan, J., Tran, S., Xu, Y., Chanda, S., Chen, L., Zeng, B., Chilimbi, T., Huang, J.: Vision-language pre-training with triple contrastive learning. pp. 15671–15680 (2022)
- Yang, Y., Feng, C., Shen, Y., Tian, D.: Foldingnet: Point cloud auto-encoder via deep grid deformation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 206–215 (2018)
- 46. Yu, X., Rao, Y., Wang, Z., Liu, Z., Lu, J., Zhou, J.: Pointr: Diverse point cloud completion with geometry-aware transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 12498–12507 (2021)
- 47. Yuan, W., Khot, T., Held, D., Mertz, C., Hebert, M.: Pcn: Point completion network. In: 2018 international conference on 3D vision (3DV). pp. 728–737. IEEE (2018)
- Zhang, J., Shao, J., Chen, J., Yang, D., Liang, B., Liang, R.: Pfnet: an unsupervised deep network for polarization image fusion. Optics letters 45(6), 1507–1510 (2020)
- Zhang, J., Chen, X., Cai, Z., Pan, L., Zhao, H., Yi, S., Yeo, C.K., Dai, B., Loy, C.C.: Unsupervised 3d shape completion through gan inversion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1768–1777 (2021)

- 18 Linlian Jiang, Pan Chen, Ye Wang, Tieru Wu, and Rui Ma
- Zhang, L., Agrawala, M.: Adding conditional control to text-to-image diffusion models. arXiv preprint arXiv:2302.05543 (2023)
- Zhang, X., Feng, Y., Li, S., Zou, C., Wan, H., Zhao, X., Guo, Y., Gao, Y.: Viewguided point cloud completion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15890–15899 (2021)
- Zhao, H., Jiang, L., Jia, J., Torr, P.H., Koltun, V.: Point transformer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 16259– 16268 (2021)
- 53. Zhou, H., Cao, Y., Chu, W., Zhu, J., Lu, T., Tai, Y., Wang, C.: Seedformer: Patch seeds based point cloud completion with upsample transformer. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part III. pp. 416–432. Springer (2022)
- Zhou, L., Du, Y., Wu, J.: 3d shape generation and completion through point-voxel diffusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5826–5835 (2021)