ProtoComp: Diverse Point Cloud Completion with Controllable Prototype

Anonymous ECCV 2024 Submission

Paper ID #6685

A. Implementation Details

Basic HyperParameters: We follow the foundational framework of vanilla Point-E, leveraging lightweight 40M pretrained weights designed specifically for text-only prompting. For both the frozen branch and the control branch, we deploy 12 layers of self-attention and integrate the output from the final 6 layers of the control branch into corresponding layers within the frozen backbone. Our approach incorporates a multi-layer graph-based convolution to extract high-level semantics while downsampling the input to N = 256 central points and we set k of kNN operation to K = 16. The detailed configuration for multi-layer graph-based convolution is: $\texttt{Linear}(C_{in} = 3, C_{out} = 8) \rightarrow \texttt{Graph}\phi(C_{in} = 8, C_{out} = 32, K = 16, N_{out} = 2048) \rightarrow \texttt{Graph}\phi(C_{in} = 32, C_{out} = 64, K = 16, N_{out} = 512) \rightarrow \texttt{Graph}\phi(C_{in} = 64, C_{out} = 64, K = 16, N_{out} = 512) \rightarrow \texttt{Graph}\phi(C_{in} = 64, C_{out} = 128, K = 16, N_{out} = 256)$. Here, C_{in} and C_{out} represent the input and output channels, respectively, while N_{out} denotes the number of points after downsampling. Regarding the Geometric block, we use the same encoder architecture mentioned earlier, which is a same multi-layer graph-based convolution. We set the momentum for the momentum encoder to 0.999. The features encoded from prototypes and partials are projected into 384 channels, and we employ a two-layer cross-attention with 384 channels to merge these features. The cross-attention has 6 heads. We train our model with a batch size of 48 across 100 epochs on PCN dataset. For the ablation experiment, we train our model on a reduced subset of ShapeNet55. This subset encompasses 8 categories, all of which are also present in the PCN dataset. Specifically, we intercept the initial 100 samples for each category, yielding a total of 800 samples. We train on this subset with a batch size of 48 and across only about 20000 iterations. For all experiments, we employ the AdamW optimizer with an initial learning rate set at 0.0001 and a weight decay of 0.001. During inference stage, we set the guidance scale to 1.3 for empty text prompt, 3.0 for non-empty text prompt, and we always use the simple text, devoid of directional information, as the input prompt.

Augmentation: During the training stage, we augment the partial input by introducing random rotations at degrees of 0°, 90°, 180°, or 270°. Simultaneously, we enhance the contextual information associated with the text prompt by appending the phrases "Facing West," "Facing North," "Facing East," and "Facing South" to both the beginning and end of the simple text input in accordance

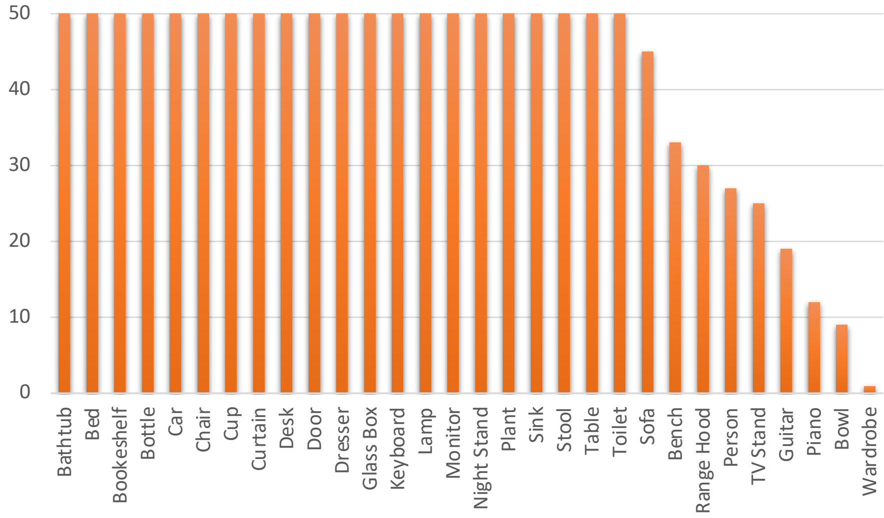

Fig. 1: The category distribution for Real-Sensors Benchmark. It encompasses a data distribution that includes 30 categories, with approximately 1200 samples available for evaluating real-world scan completion.

with the corresponding rotation angle. To illustrate, if a chair undergoes a 90° rotation and the text prompt is not randomly set to be empty, the comprehensive prompt for it will become "Facing North. A chair. Facing North."

B. Real-Sensors Benchmark

In order to assess the model's capabilities in real-world scenarios, we have developed the Real-Sensors Benchmark. This benchmark comprises 30 distinct object categories, totaling 1251 samples. Among them, 29 categories are sourced from indoor scenes in the ScanNet200 dataset, while the remaining single category pertains to the 'cars' from the KITTI dataset. We set an upper limit of 50 for each category to limit the number of common objects. Importantly, all these categories coexist in the ModelNet40 dataset, enabling the utilization of PointNet++ pre-trained on this dataset for classification purpose. The detailed distribution of sample categories can be observed in Figure. 1

C. More Experiments

To further evaluate the impact of varying degrees of incompleteness on our model, we conduct experiments on Real-Sensors with different mask ratios and report the results in Table. 1. We follow the same cropping masking scheme proposed in [AdaPoinTr, TPAMI 2023]. It is observed that as the proportion of masked ratio increases, the model's GD declines. However, it is noteworthy

Table 1: Different Mask Ratios on Real-Sensors Benchmark.

Mask Ratio	GD_C	GD_I	Mask Ratio	GD_C	GD_I	Mask Ratio	GD_C	GD_I
0%	40.8	43.6	20%	40.4	43.3	40%	39.8	42.0
10%	40.5	43.5	30%	39.8	42.2	50%	39.4	41.8

that even under mild masking conditions (10%, 20%), the model still achieves favorable results, indicating a certain level of robustness to the degree of input incompleteness.

D. More Visualization Results

In Figure. 2, we present additional visualization results showcasing the efficacy of our model at the object level within the Scannet200 dataset. Notably, our model, trained on the same PCN dataset, exhibits superior performance in object completion tasks within indoor open scenes, giving more semantic consistency and realism than other state-of-the-art point cloud completion models. Figure. 3 extends our analysis to encompass more complex and realistic scenarios, offering more comprehensive presentation of whole scene recovery ability within open environments. Rows (a)–(b) and (c)–(d) specifically show the scene recovery capabilities on Scannet200 and KITTI datasets, respectively. In addition, employing the settings of Model E in our ablation experiments, we conducted experiments with varying proportions of masking on four different categories of objects (where trash bin and mailbox are not included in the training set), as illustrated in Figure. 4. It is observed that even under very high masking proportions (80%, 90%), our model is capable of imaginatively and reasonably completing the incomplete point clouds. As the masking proportion decreases, the incomplete point clouds exert a greater influence on the model, leading to completion results that are closer to the ground truth labels.

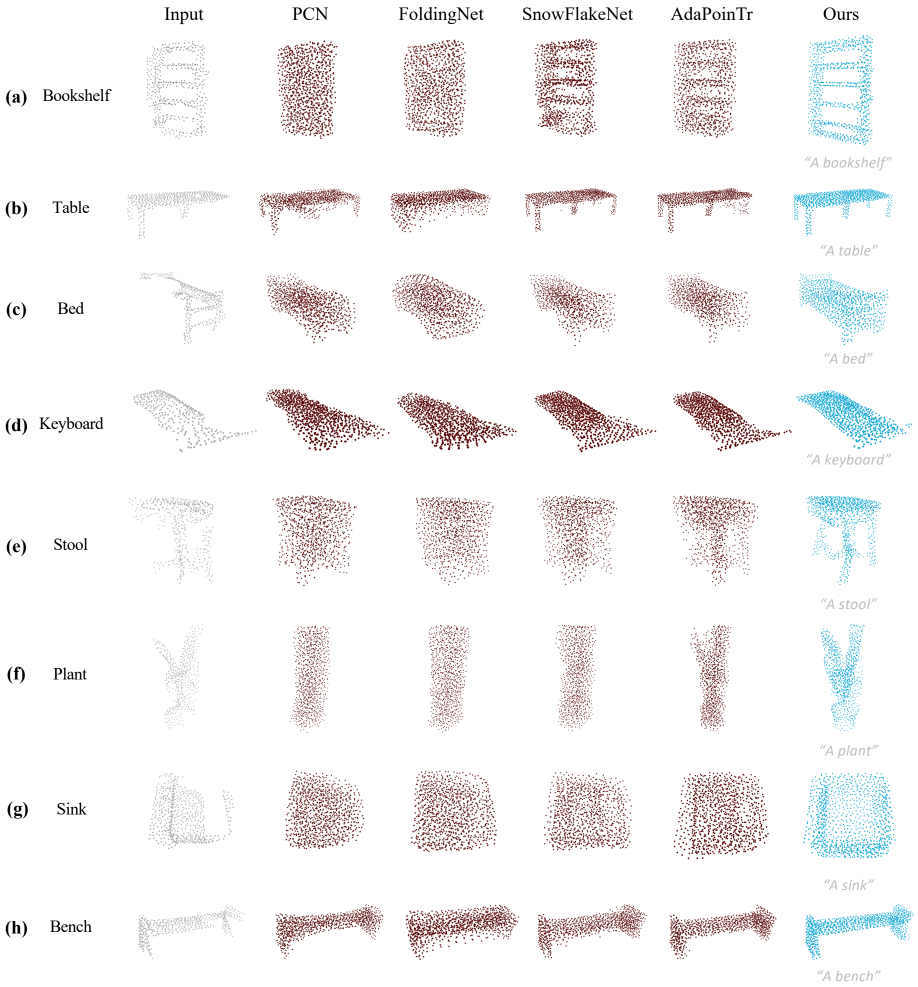

Fig. 2: More visualization results. We show the completion results with identity objects cropped from ScanNet200.

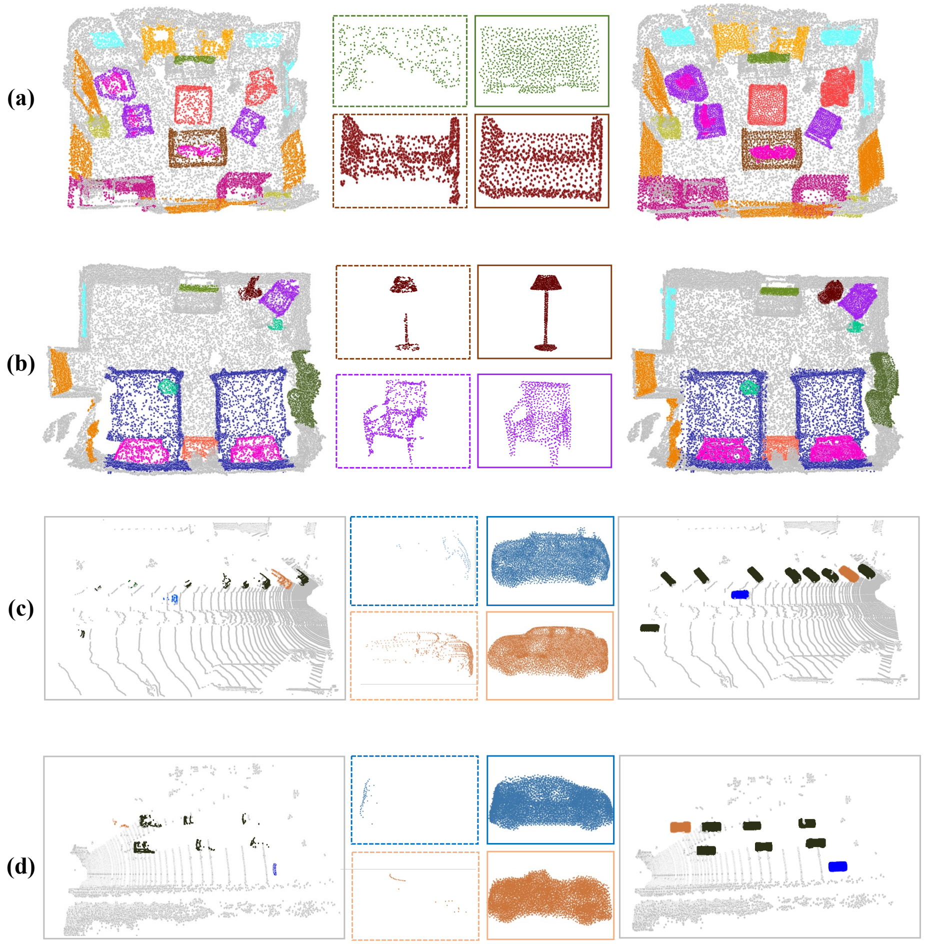

Fig. 3: More visualization results. We showcase the completion results using entire scenes from KITTI and ScanNet200.

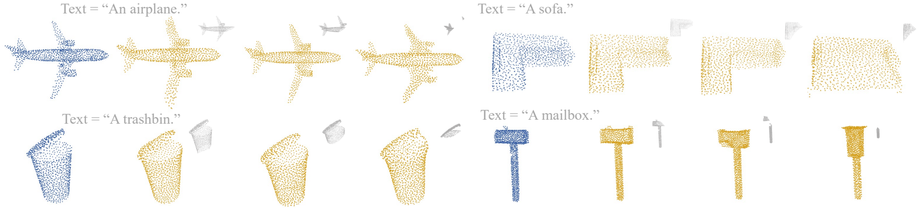

Fig. 4: More visualization results.We present visualization results showcasing the completion outcomes generated by our model across various mask ratios. The visual representations depict three distinct components: the ground truth object depicted in blue, our model's prediction illustrated in yellow, and the partial input displayed in grey. The partial input exhibits different mask ratios, specifically 20%, 50%, and 80%, arranged from left to right in ascending order.