TouchDream: 3D Object Completion through Imagined Touch

Supplementary Material

In this supplementary material, we present additional experimental results, including guidance comparison details, a complete-process visual analysis, and a discussion of limitations, along with further visual comparisons on the ShapeNet-55/34 dataset.

Combination of Touch and Symmetry Guidance

In Section 4.3 of this paper, to compare the effects of touch guidance and the symmetry guidance proposed by SymmCompletion [34] on point cloud completion, we experiment with three configurations: using only touch guidance, only the symmetry guidance from SymmCompletion, or a combination of both to guide the point refinement. The architectures of the combination of touch guidance and symmetry guidance is illustrated in Figure 8. In the symmetry guidance module from SymmCompletion, an LSTNet is used to extract the input point cloud feature F_{input} and the symmetry feature F_{symm}. The coarse feature F_{coarse} is then processed through two separate branches, where it is fused with F_{input} and F_{symm} via cross-attention and self-attention, respectively. The resulting features are finally concatenated. In the combined guidance approach, an additional feature, derived from tactile features F_{touch}, is concatenated.

Based on the results shown in Table 5 and Figure 7 in the main body of this paper, touch guidance leads to better completion outcomes, particularly in the recovery of fine details. Notably, the touch guidance module achieves performance comparable to that of the combined symmetry and touch guidance.

Visualization Analysis of TouchDream Process

Figure 9 presents the generated tactile point clouds and their corresponding complete ground truth point clouds, which can be observed to align with the ground truth. Figure 10 illustrates the standard completion pipeline of TouchDream. The process consists of three stages: First, a pretrained coarse completion network (specifically, the LSTNet from SymmCompletion [34]) generates an initial coarse point cloud that captures the global structure of the target object. Subsequently, this coarse point cloud is utilized to sample touch poses, from which corresponding touch latent codes are generated through a conditioned diffusion model and decoded into tactile point cloud. Finally, the coarse point cloud is refined with the generated tactile points, producing a final output that visual results show is both more complete and richer in local details.

We also visualize the touch generation process in Figure 11. The interaction action space is defined by uniformly sampling 50 points on a sphere centered at the centroid of the scaled coarse point cloud. The valid poses are then determined by identifying the intersection points between the approach rays and the object's convex hull. The simulator output all the valid poses for touch generation. For each valid pose, we employ a diffusion model to generate the tactile latent representation, which is subsequently decoded into local 3D points. These local points are then transformed to align with the ground truth surfaces using the predicted transformation matrix. After merging and resampling all the transformed tactile points, we rescale the resulting point cloud back to its original dimensions to obtain the final tactile points for touch-guided refinement.

Limitations Analysis

Although our model achieves state-of-the-art performance, its effectiveness is contingent upon the initial coarse point cloud. When input data is limited, leading to a coarse prediction that significantly deviates from the ground truth, the subsequent tactile point cloud may fail to capture the object's true geometry, thereby compromising the final output. Failure cases are illustrated in Figure 12. It is important to note that this represents a common challenge across point cloud completion methods. Nevertheless, our approach still produces more plausible results than prior methods. A promising solution to this limitation lies in a unified multimodal framework that integrates tactile and visual generation for multimodel-guided completion.

A further limitation is the computational cost. Since our approach generates a tactile latent code for every sampled touch pose and then performs decoding, transformation, and sampling to obtain the local geometry, the process is inherently more time-consuming. Notably, a full evaluation on the 1,200 shapes from the PCN dataset required 6.7 hours on a single NVIDIA GTX 4090 GPU. To mitigate this in the future, one could explore a unified model that leverages a diffusion model to predict a complete global tactile point cloud at once, conditioned on all valid interactions. However, this approach must address the fundamental challenge that tactile perception is inherently local and tied to curvature, resulting in significant coordinate variations for points generated from different poses.

Visualized Results on ShapeNet-55/34

We also present additional qualitative results on the ShapeNet-55/34 datasets in Figure 13. To better highlight the local refinement capability of our approach, we zoom in on selected regions (marked with red boxes), where our method shows clear advantages in reconstructing fine-grained details.

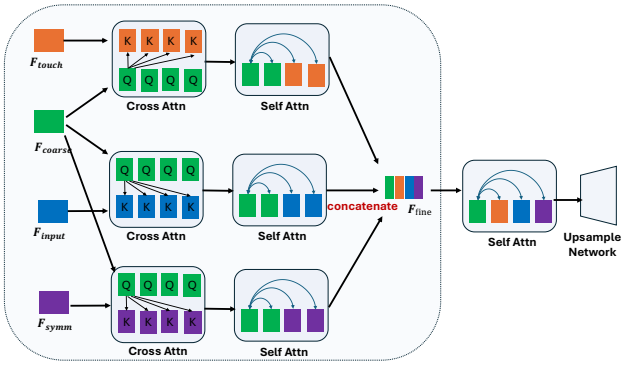

Figure 8. Combination of touch guidance and symmetry guidance.

Figure 9. Generated touch points (256 sampled points) and their corresponding ground-truth point clouds.

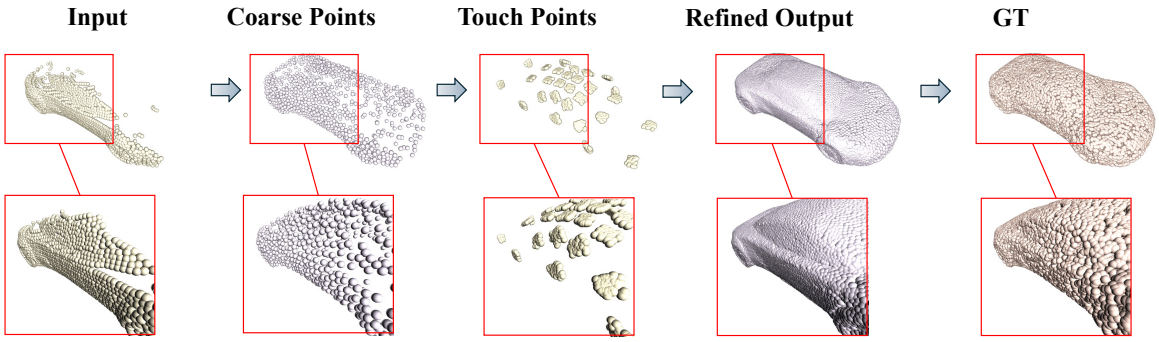

Figure 10. Visualized completion process of our TouchDream.

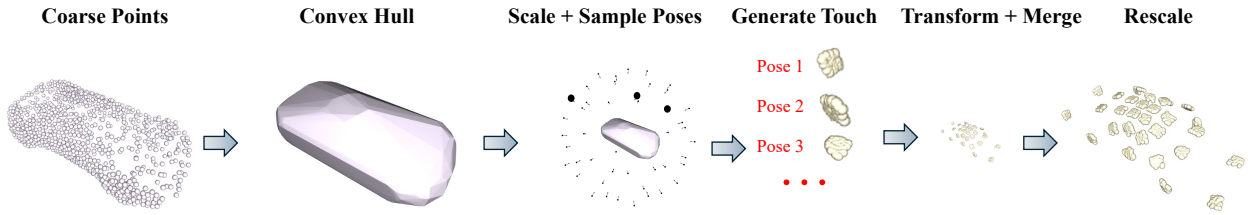

Figure 11. Visualized touch generation process.

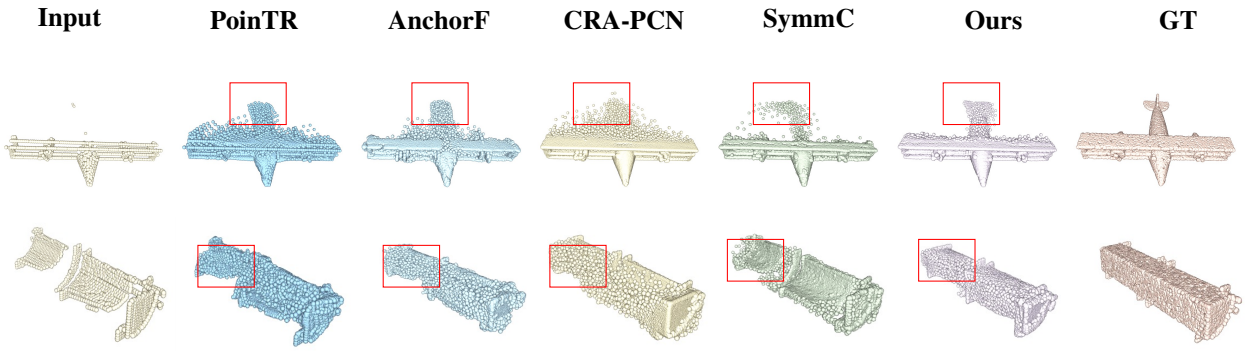

Figure 12. Examples of failure cases on PCN dataset.

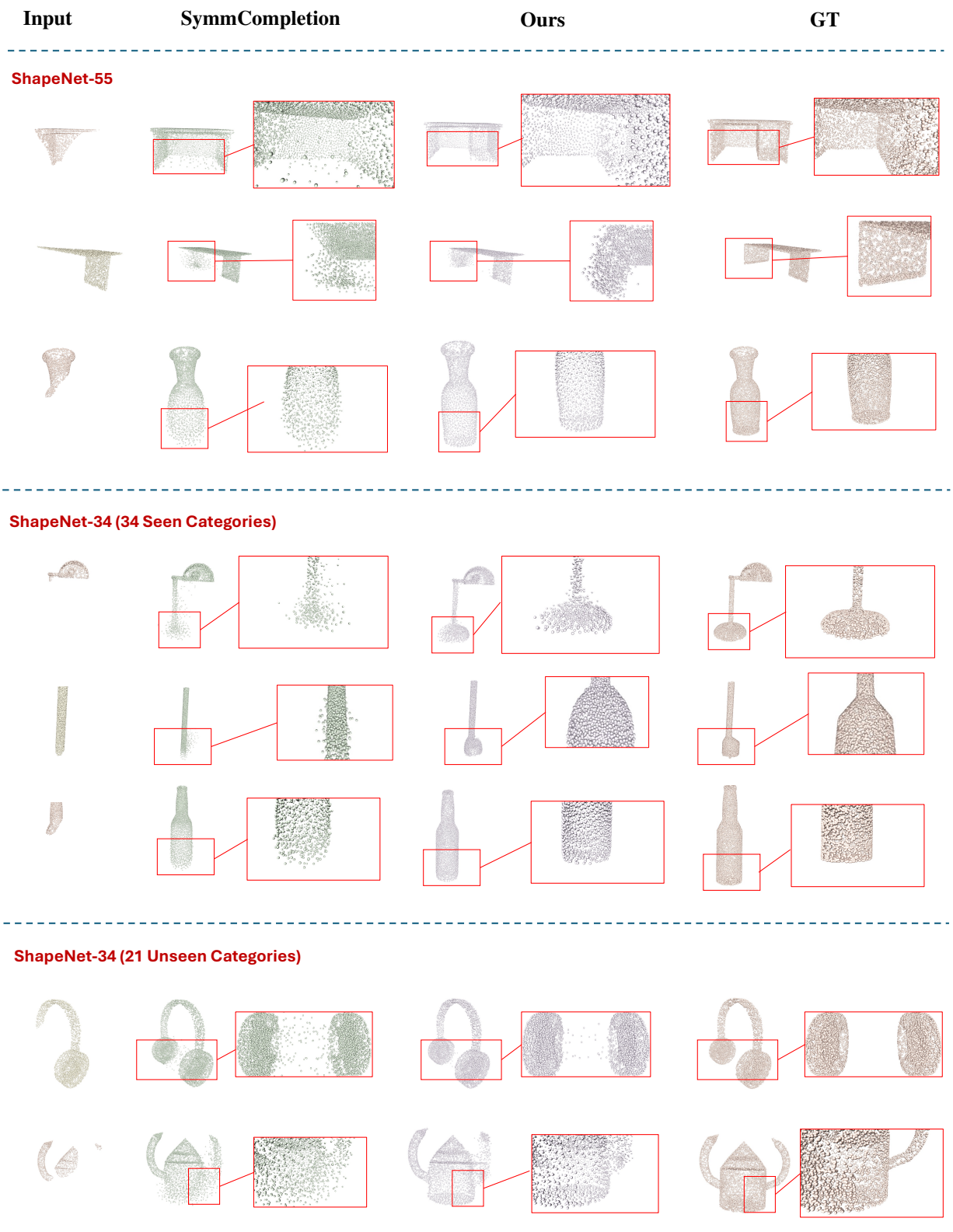

Figure 13. Qualitative Comparison with SymmCompletion on the ShapeNet-55/34 dataset.