Supplementary Materials of "VIPNet: Combining Viewpoint Information and Shape Priors for Instant Multi-View 3D Reconstruction"

Weining Ye[1], Zhixuan Li[2], and Tingting Jiang[3]

[1] School of Computer Science, Peking University, Beijing 100871, China
[2] College of Computing and Data Science, Nanyang Technological University, Singapore
[3] National Engineering Research Center of Visual Technology, National Key Laboratory for Multimedia Information Processing, School of Computer Science, National Biomedical Imaging Center, Peking University, Beijing 100871, China

1 Per-Category results of Multi-View 3D Reconstruction

We compare our VIPNet and VIPNet+ with other CNN-based methods testing on each category of ShapeNet. We show both the IoU results and F-Score results when inputting 24 views. As shown in Table 1, our VIPNet+ performs better than other methods for nearly all categories. For the display category, our VIPNet+ outperforms other methods by about 6% on IoU and about 0.06 on F-Score@1%. For the lamp category, our VIPNet+ outperforms other methods by about 4% on IoU and about 0.03 on F-Score@1%.

Table 1: Per-category results testing on ShapeNet when inputting 24 views. The results are IoU / F-Score@1%

Category	Pix2Vox++ [3]	GARNet [4]	GARNet+	VIPNet	VIPNet+
airplane	0.729 / 0.614	0.724 / 0.606	**0.739** / **0.628**	0.720 / 0.601	**0.739** / 0.624
bench	0.686 / 0.522	0.698 / 0.536	0.707 / 0.551	0.694 / 0.535	**0.711** / **0.554**
cabinet	0.829 / 0.456	0.841 / 0.473	0.840 / 0.505	0.860 / 0.503	**0.863** / **0.506**
car	0.883 / 0.598	0.888 / 0.608	0.894 / 0.623	0.891 / 0.614	**0.896** / **0.625**
chair	0.647 / 0.341	0.674 / 0.369	0.683 / 0.384	0.685 / 0.383	**0.699** / **0.396**
display	0.613 / 0.335	0.668 / 0.386	0.665 / 0.396	0.721 / 0.445	**0.730** / **0.452**
lamp	0.493 / 0.351	0.516 / 0.366	0.513 / 0.369	0.537 / 0.382	**0.556** / **0.400**
speaker	0.762 / 0.326	0.773 / 0.338	0.772 / 0.346	0.798 / 0.374	**0.803** / **0.375**
rifle	0.686 / 0.624	0.697 / 0.634	0.709 / 0.647	0.708 / 0.647	**0.733** / **0.674**
sofa	0.782 / 0.454	0.807 / 0.489	0.810 / 0.500	0.817 / 0.503	**0.823** / **0.513**
table	0.666 / 0.419	0.693 / 0.449	0.692 / 0.452	0.707 / 0.462	**0.710** / **0.464**
telephone	0.849 / 0.666	0.871 / 0.698	0.879 / 0.716	0.890 / 0.724	**0.895** / **0.732**
watercraft	0.668 / 0.460	0.693 / 0.494	0.696 / 0.504	0.700 / 0.501	**0.709** / **0.516**
overall	0.720 / 0.473	0.737 / 0.493	0.742 / 0.505	0.748 / 0.506	**0.758** / **0.518**

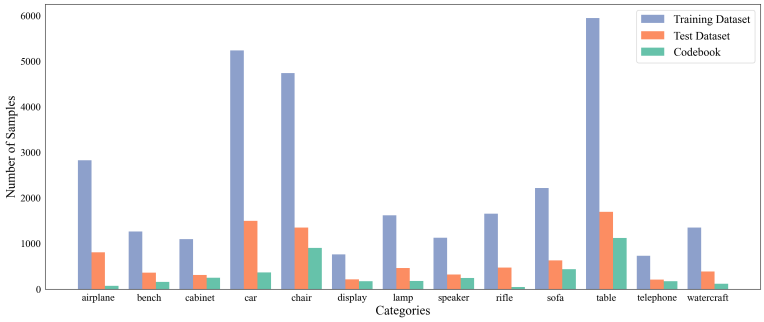

Fig. 1: The number of samples for each category in the training dataset, test dataset, and the codebook.

2 Shape Distribution in the Codebook

Figure 1 shows the number of shapes in each category in the codebook. The total number of shapes in the codebook is 4000. We also demonstrate the number of samples for each category in the training and test dataset. The figure indicates that chairs, sofas, and tables are relatively large in number in the codebook due to the large number in the training dataset and their variety of shapes. In contrast, although airplanes and cars have many samples in the training set, they are not numerous in the codebook because of their high shape similarity and small variance.

3 More Visualization Results

Figure 2 and Figure 3 illustrate more visualization results on ShapeNet [1, 2] when taking 5, 10, 15, and 20 viewpoints as inputs. We compare our method with 3D-R2N2 [1], Pix2Vox++ [3], GARNet [4] and GARNet++. With the help of the viewpoint information and shape priors, our method is able to reconstruct more accurate shapes.

4 Limitations and Failure Cases

Figure 4 shows some failure cases of our VIPNet and VIPNet+ on ShapeNet when taking 5, 10, 15, and 20 viewpoints as inputs. When the 3D shape is irregular and unusual, such as the first example of Figure 4, our VIPNet and VIPNet+ may give undesirable results. Besides, when some parts of the object are very thin, our model may not be able to reconstruct these parts, such as the lamp and the table in Figure 4.

The issues with uncommon shapes arise because our method relies on training data and lacks strong zero-shot capabilities. Fine details are challenging due

to the limited feature extraction capacity of our CNN-based model. Improving the model's structure and using larger and more complex training datasets can help mitigate these problems. Additionally, our current voxel resolution is 32, which limits 3D object representation. Increasing the resolution to 64 or 128 can enhance detail capture.

References

1. Choy, C.B., Xu, D., Gwak, J., Chen, K., Savarese, S.: 3D-R2N2: A Unified Approach for Single and Multi-View 3D Object Reconstruction. In: Proceedings of the European Conference on Computer Vision. pp. 628–644. Springer (2016) 2
2. Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., Xiao, J.: 3D ShapeNets: A Deep Representation for Volumetric Shapes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1912–1920 (2015) 2
3. Xie, H., Yao, H., Zhang, S., Zhou, S., Sun, W.: Pix2Vox++: Multi-Scale Context-Aware 3D Object Reconstruction from Single and Multiple images. International Journal of Computer Vision **128**(12), 2919–2935 (2020) 1, 2
4. Zhu, Z., Yang, L., Lin, X., Yang, L., Liang, Y.: GARNet: Global-aware Multi-View 3D Reconstruction Network and the Cost-Performance Tradeoff. Pattern Recognition **142**, 109674 (2023) 1, 2

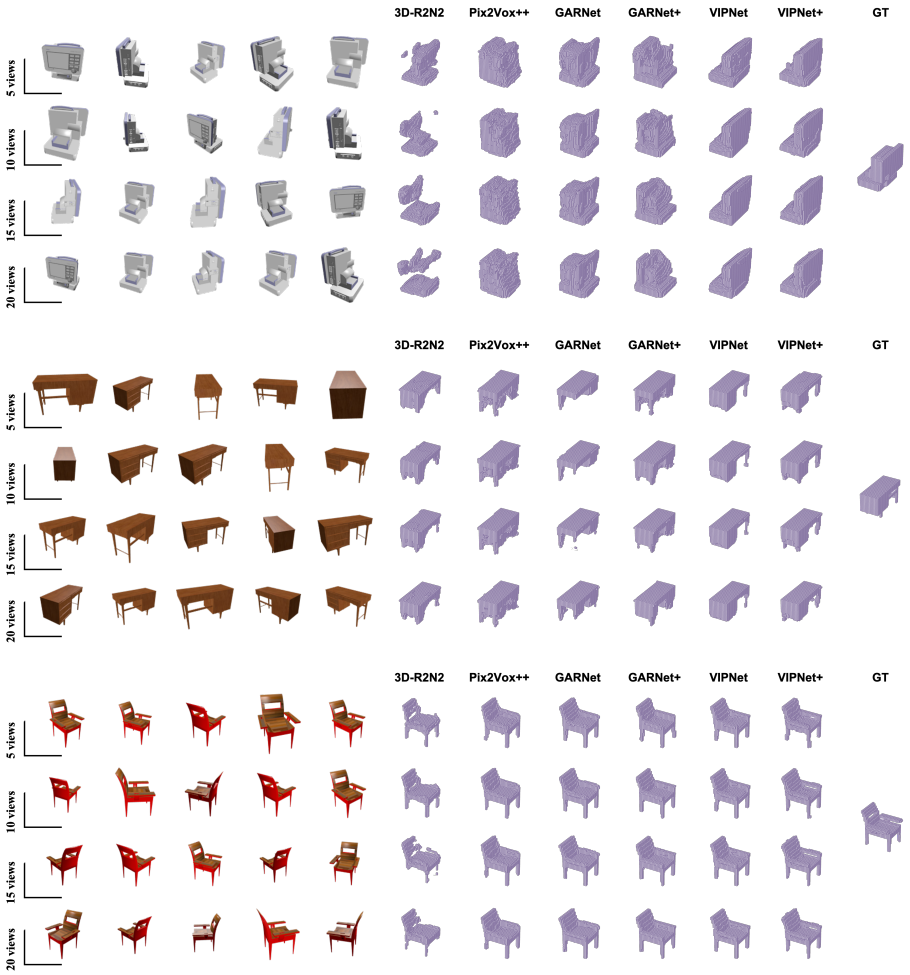

Fig. 2: Visualization of Multi-View 3D reconstruction with other CNN-based methods on ShapeNet when taking 5, 10, 15, and 20 viewpoints as inputs.

Fig. 3: Visualization of Multi-View 3D reconstruction with other CNN-based methods on ShapeNet when taking 5, 10, 15, and 20 viewpoints as inputs.

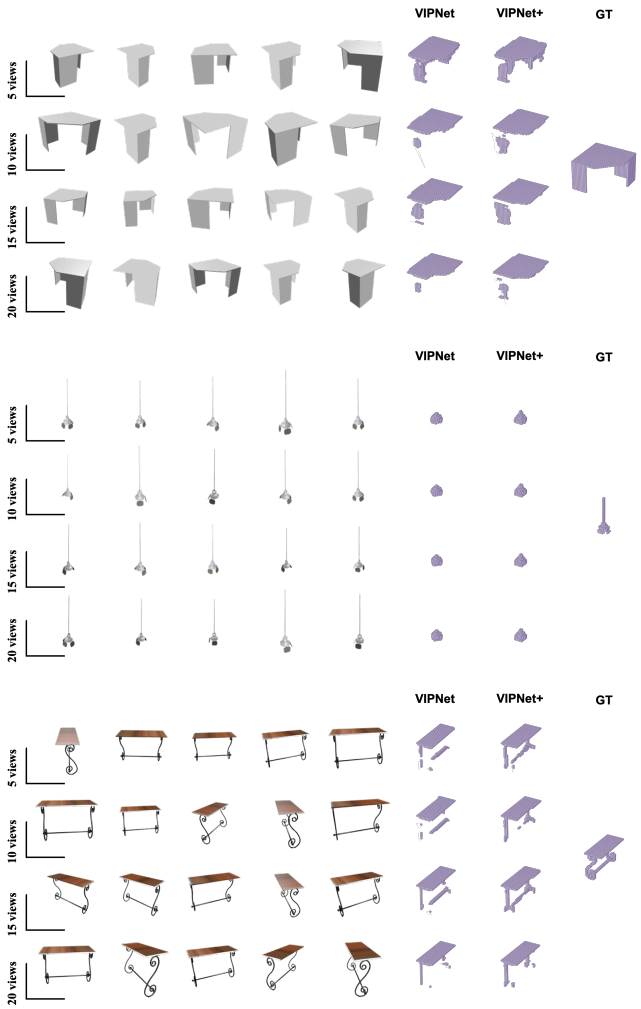

Fig. 4: Failure cases of our VIPNet and VIPNet+ on ShapeNet when taking 5, 10, 15, and 20 viewpoints as inputs.