Few-shot Single-view 3D Reconstruction with Memory Prior Contrastive Network(Appendix)

Zhen Xing[1], Yijiang Chen[1], Zhixin Ling[1],
Xiangdong Zhou[1], and Yu Xiang[2]

[1] School of Computer Science, Fudan University, Shanghai, China, 200433
{zxing20,chenyj20,2021201005,xdzhou}@fudan.edu.cn
[2] The University of Texas at Dallas {yu.xiang}@utdallas.edu

This supplementary Appendix contains the following.

- Section 1:Details of experiment setup.
- Section 2:Implementation details of different prior module analysis.
- Section 3:Additional qualitative examples from different datasets.
- Section 4:Limitations and Future work.

1 Experiment setup

Implementation Details MPCN is implemented using PyTorch 1.5.0 on a single GTX TITAN X GPU on Ubuntu16.04 system with 64G memory. The CPU is i7-6800K. The GPU memory is 12.2GB.

We use the datasets of ShapeNet [1] provided by 3D-R2N2 [3] with image size of 137 × 137. We adopt the same data augmentation strategy as Pix2vox [8] by random crop, random color jittering and random noise. Finally, the input image of MPCN is center cropped to 224 × 224.

Dataset Details Pascal3D+ [6] has 12 categories: aeroplane, bicycle, boat, bottle, bus, car, chair, diningtable, motorbike, sofa, train and tvmonitor. For the experiments on real-world dataset, we firstly pre-train the MPCN on all 13 categories in ShapeNet [1] dataset. We put objects in front of random backgrounds from the SUN database [7]. Then we finetune the model on 8 base categories in Pascal3D+ [6] dataset: aeroplane, bicycle, boat, bus, car, chair, diningtable, sofa, tvmonitor. Finally we test on 4 novel categories in Pascal3D+: bicycle, train, motorbike, bottle.

The Comparison of Efficiency (1) PADMix [2] mentioned that it requires 8 Tesla V100 with 256GB GPU memory to train the model, while our method only need a single NVIDIA TITAN X with 12 GB GPU memory, and the training time takes about 80 hours on single GPU, which shows the lightness of our method. (2) Besides, our 3D-aware contrastive loss could train together with reconstruction loss while PADMix needs an additional pre-training stage to solve the pose-invariance issue. (3) The cross-attention computational cost is quite small because that the length of the sequence is only 5, and the training

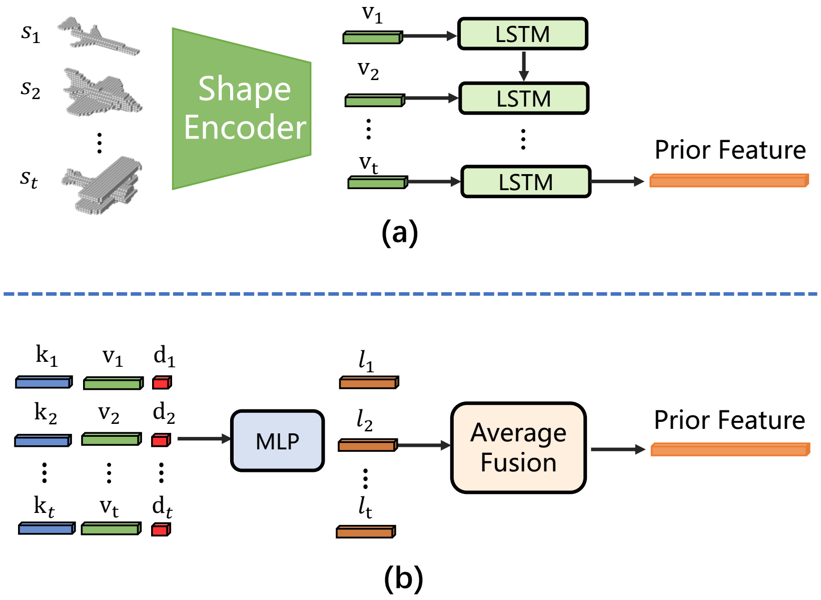

Fig. 1. (a) The overview of MPCN-LSTM. (b) The overview of MPCN-average-fusion.

computational costs of Wallace [5] and CGCE [4] are similar to ours, but the performance of these two methods lag far behind ours (*i.e.*, up to 10% mIoU on ShapeNet [1] with 10-shot and 25-shot settings). (4) For the inference time, it only takes about 0.02s for a single reconstruction inference on a single TITAN X for our method.

2 Prior Module Analysis Details

In the Ablation Study of the main submission, we have proved that the memory network provides a well guidance shape prior for 3D reconstruction. In order to prove the effectiveness of our attention-based prior module, we also use two other prior fusion baseline modules as comparison.

2.1 MPCN-LSTM

We use a standard LSTM module instead of our attention-based prior module to compare the advantages of our method. As shown in Fig. 1 (a), firstly we extract voxel feature with shape encoder, and at each time step we feed in the neighbor voxel feature embedding. The LSTM is rolled out for k time steps, where k is the number of neighbors. Its final output state is taken as prior feature.

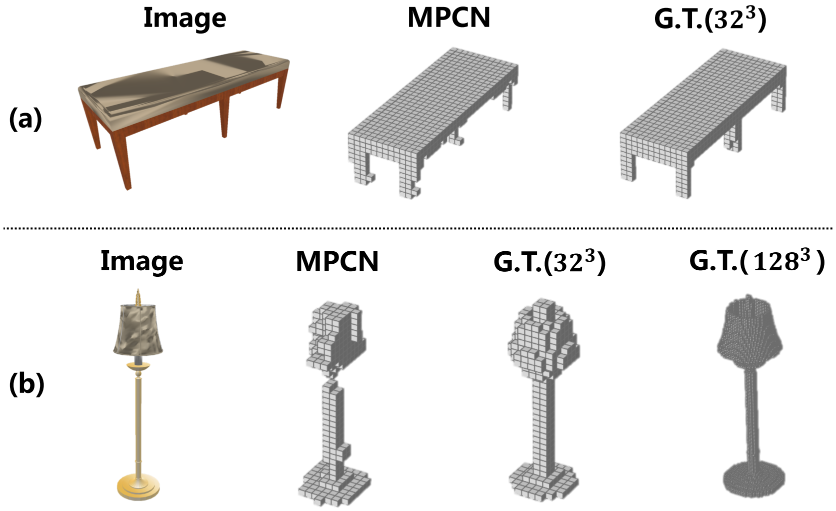

Fig. 2. Examples of limitation and future work.

2.2 MPCN-average-fusion

As shown in Fig. 1 (b), the front part of the network structure is completely consistent with MPCN-Attention. In the fusion part of k neighbors of l, we calculate the average of these k vectors, which is used as a lower baseline for comparison. The number of k is 5. The final average result is defined as prior feature.

3 Qualitative Examples

In this section, we show our MPCN 3D reconstructions compared to baseline method in visualizations. We provide more visualizations of our reconstructions as compared to baseline which demonstrate higher quality predictions obtained using our proposed method in Fig. 3. Besides, Fig. 4 shows the results on Pascal3D+ [6] dataset with the shot number of 10-shot. We show two different views for each 3D volumes in Fig. 4 .

4 Limitation and Future Work

Previous baseline methods [5,4,2] based on voxel adopt 32^3 resolution to represent the shape of 3D objects. To compare with them fairly, we also follow the volume resolution. It works well for squared 3D shape, as shown in Fig. 2 (a). However, as shown in Fig. 2 (b), for this complex 3D object with curved shape, it is difficult to accurately represent the shape of the voxel with low resolution. But the resolution of 128^3 voxel can better reflect the shape of the object, so our future work will be devoted to the research of higher resolution 3D reconstruction based on voxel.

References

1. Chang, A.X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., et al.: Shapenet: An information-rich 3d model repository. arXiv preprint arXiv:1512.03012 (2015)
2. Cheng, T.Y., Yang, H.R., Trigoni, N., Chen, H.T., Liu, T.L.: Pose adaptive dual mixup for few-shot single-view 3d reconstruction. In: AAAI (2022)
3. 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: ECCV (2016)
4. Michalkiewicz, M., Parisot, S., Tsogkas, S., Baktashmotlagh, M., Eriksson, A., Belilovsky, E.: Few-shot single-view 3-d object reconstruction with compositional priors. In: ECCV (2020)
5. Wallace, B., Hariharan, B.: Few-shot generalization for single-image 3d reconstruction via priors. In: ICCV (2019)
6. Xiang, Y., Mottaghi, R., Savarese, S.: Beyond pascal: A benchmark for 3d object detection in the wild. In: WACV (2014)
7. Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A.: Sun database: Large-scale scene recognition from abbey to zoo. In: CVPR (2010)
8. Xie, H., Yao, H., Sun, X., Zhou, S., Zhang, S.: Pix2vox: Context-aware 3d reconstruction from single and multi-view images. In: ICCV (2019)

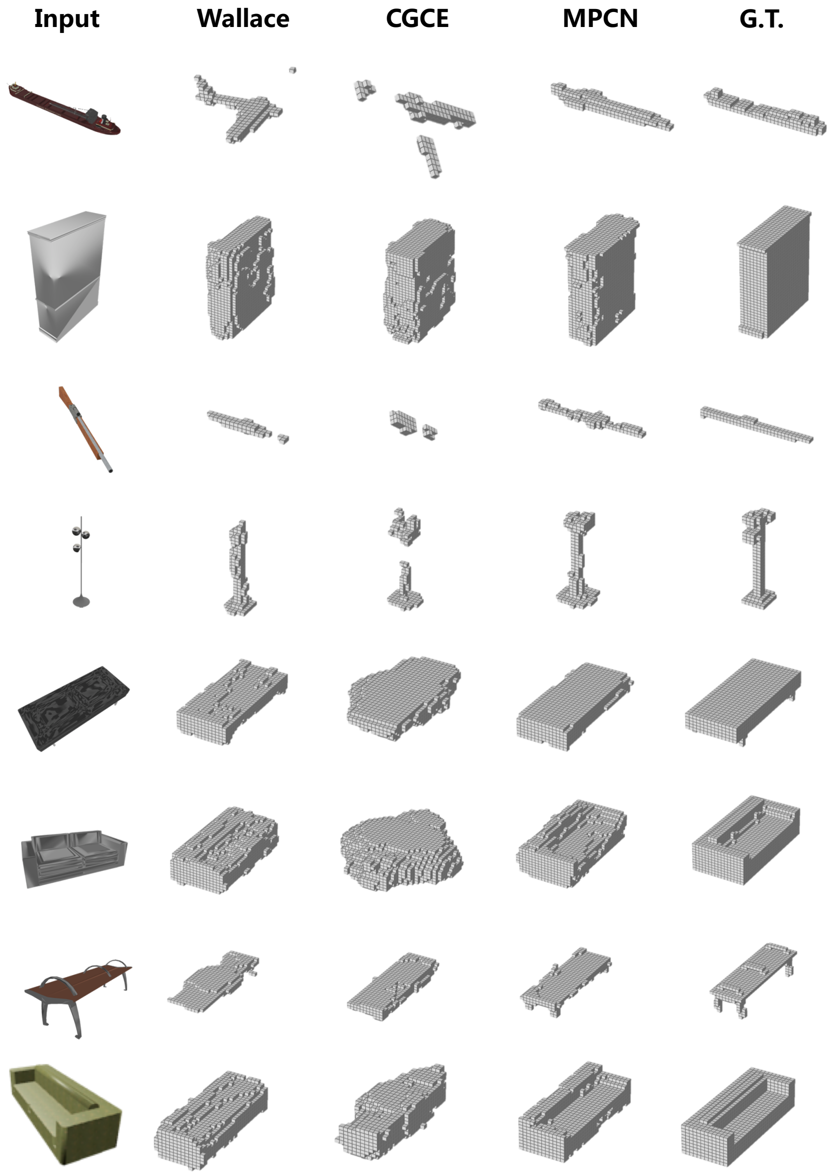

Fig. 3. Qualitative results of single image 3D reconstruction on the ShapeNet dataset with our method and baseline method.

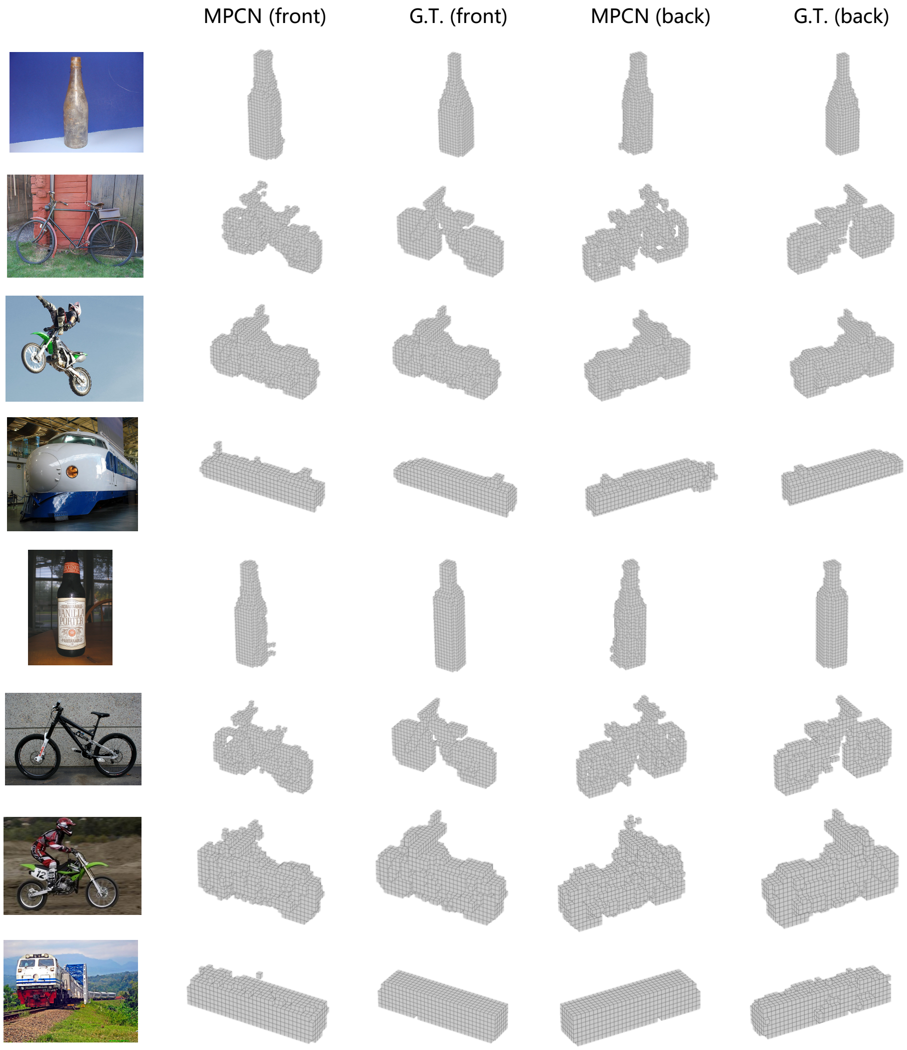

Fig. 4. Qualitative examples of shape inference obtained by proposed MPCN approach on Pascal3D+ with 10-shot. We show two different views (front and back) for each 3D representation.