Supplementary Materials of 3D-C2FT: Coarse-to-fine Transformer for Multi-view 3D Reconstruction

Leslie Ching Ow Tiong[1,*] [0000−0003−3786−2117], Dick Sigmund[2,*] [0000−0002−6207−5804], and Andrew Beng Jin Teoh[3,†] [0000−0001−5063−9484]

[1] Computational Science Research Center, Korea Institute of Science and Technology, 5, Hwarang-ro 14-gil, Seongbuk-gu, Seoul 02792, Republic of Korea
tiongleslie@kist.re.kr
[2] AIDOT Inc., 128, Beobwon-ro, Songpa-gu, Seoul 05854, Republic of Korea
dsigmund@aidot.ai
[3] School of Electrical and Electronic Engineering, Yonsei University, Seoul 120-749, Republic of Korea
bjteoh@yonsei.ac.kr

6 Appendix

6.1 Additional Ablation Study

In this section, all the experiments are conducted with the ShapeNet dataset described in main content Section 4.1.

Training with Different Number of C2F Patch Attention Block. In supplementary Fig. 1, we visualize the performances on the ShapeNet validation dataset with respect to the number of C2F patch attention blocks, $J = 3, 4, 5$ used in the proposed network. We observe that 3D-C2FT with J=3 achieves the smallest loss. In the main text, this justifies $J = 3$ is set for 3D-C2FT.

Training with Different View Counts. This ablation investigates the object reconstruction performance with respect to the number of input view images used in 3D-C2FT training. Here, we fix the view input numbers at 4, 8, and 12. In addition, the input views are randomly sampled at every training iteration. From supplementary Table 1, it is interesting to see that the best number of input views for training is 8, but not the larger view count such as 12 by intuition. This phenomenon could be associated with the limitation of the decoder to aggregate coarse to fine-grained features with similar orientation views.

*The authors have contributed equally to this work.
†Corresponding author.

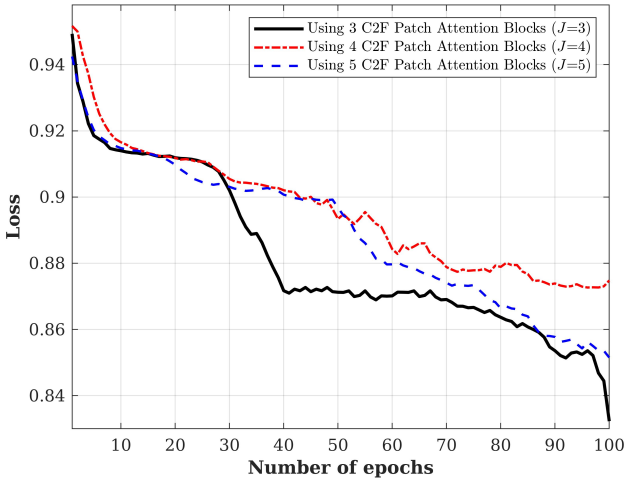

Fig. 1. Performance comparisons on different C2F patch attention blocks in the proposed network.

Table 1. Performance comparisons of single and multi-view 3D reconstruction on 3D-C2FT using specific view counts for training. The best score for each view is written in bold.

Training view counts	**Testing view counts**								
	1	2	3	4	5	8	12	18	20
Metric: IoU									
4 views	**0.629**	0.672	0.688	0.694	0.699	0.707	0.711	0.713	0.714
8 views	**0.629**	**0.678**	**0.695**	**0.702**	**0.708**	**0.716**	**0.720**	**0.723**	**0.725**
12 views	0.628	**0.678**	**0.695**	**0.702**	**0.708**	**0.716**	**0.720**	**0.723**	**0.725**
Metric: F-score									
4 views	**0.374**	0.421	0.438	0.446	0.451	0.460	0.466	0.469	0.470
8 views	0.371	**0.424**	**0.443**	**0.452**	**0.458**	**0.468**	**0.476**	**0.477**	**0.479**
12 views	0.370	0.423	0.442	**0.452**	**0.458**	**0.468**	**0.476**	**0.477**	**0.479**

Training with Different Backbone Networks. This study demonstrates the performance of 3D-C2FT using different backbone networks as image embedding modules, namely VGG16, ResNet50, and DenseNet121, which is tabulated in Supplementary Table 2. As can be seen in this table, the result indicates that using DenseNet121 as a backbone network achieves better 3D reconstruction results, which can assist the proposed model in obtaining better feature representations.

Table 2. Performance comparisons of single and multi-view 3D reconstruction on 3D-C2FT using different backbone networks for training. The best score for each view is written in bold.

Backbone	**Number of views**								
Network	1	2	3	4	5	8	12	18	20
Metric: IoU									
VGG16	0.587	0.633	0.650	0.658	0.665	0.674	0.678	0.682	0.683
ResNet50	0.593	0.648	0.661	0.670	0.679	0.682	0.690	0.696	0.697
DenseNet121	**0.629**	**0.678**	**0.695**	**0.702**	**0.708**	**0.716**	**0.720**	**0.723**	**0.725**
Metric: F-score									
VGG16	0.334	0.381	0.400	0.409	0.417	0.427	0.433	0.437	0.438
ResNet50	0.341	0.392	0.409	0.421	0.426	0.431	0.440	0.447	0.449
DenseNet121	**0.371**	**0.424**	**0.443**	**0.452**	**0.458**	**0.468**	**0.476**	**0.477**	**0.479**

6.2 Occlusion Box

In this study, we introduce the occlusion box to the odd number ordered lists of 2D images, which impedes the important parts of the images randomly. Supplementary Fig. 2 illustrates several test images from the ShapeNet dataset [2] with different sizes of occlusion boxes. In this experiment, all the models are tested on same experimental settings.

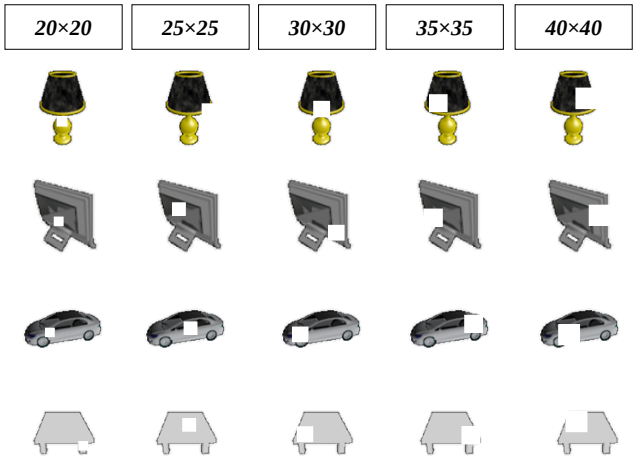

Fig. 2. Demonstrations of occlusion images on ShapeNet with several sizes of occlusion box: 20×20, 25×25, 30×30, 35×35, and 40×40.

As an illustration, we demonstrate several reconstruction results on different sizes of occlusion boxes, as shown in Supplementary Fig. 3 and 4. Among the CNN-based models [3], 3D-RETR [1], and LegoFormer [4], 3D-C2FT performs the best over various occlusion boxes.

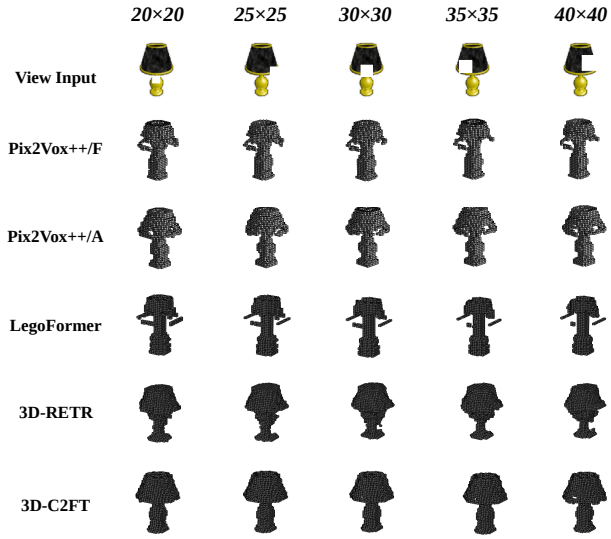

Fig. 3. Demonstration of 3D object reconstruction results for *lamp* category with several sizes of occlusion box. The experiments were conducted using 12 views (only one is shown).

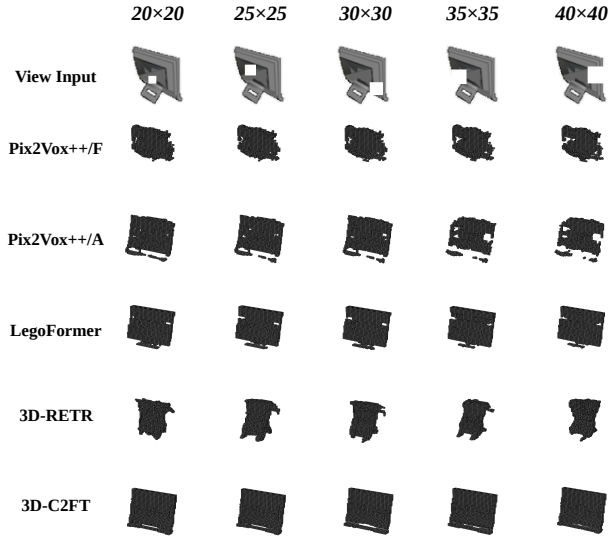

Fig. 4. Demonstration of 3D object reconstruction results for *display* category with several sizes of occlusion box. The experiments were conducted using 12 views (only one is shown).

6.3 Experiments on Multi-view Real-life Dataset

This section conducts additional qualitative experiments for single and multi-view 3D reconstruction with a Multi-view Real-life dataset.

Supplementary Fig. 5 demonstrates the performance comparisons on *Case I*. Our proposed model performs substantially better than the benchmark models in terms of the refined 3D volume and surface quality. In particular, LegoFormer performs very poorly for all categories, except *cabinet*, which is comparable to the 3D-C2FT.

Fig. 5. 3D object reconstruction results for *Case I*. Each test sample mainly contains 1 to 5 views.

In addition, Supplementary Fig. 6 illustrates the performance comparisons in *Case II*. 3D-C2FT remains significantly better than the benchmark models. Pix2Vox++/A and LegoFormer were slightly improved in this case compared to Case I. Although both models can barely reconstruct the 3D objects, the fineness is not on par with 3D-C2FT.

For *Case III*, Supplementary Fig. 7 illustrates several qualitative examples of 3D reconstruction by using more than 12 views. The reconstructed objects by 3D-C2FT are more well structured with all fine-grained details. However, surprisingly, most competing models perform poorly under *Case III*. For instance, the 3D volumes are not appropriately reconstructed for most categories, or the surfaces of 3D objects are not generated correctly.

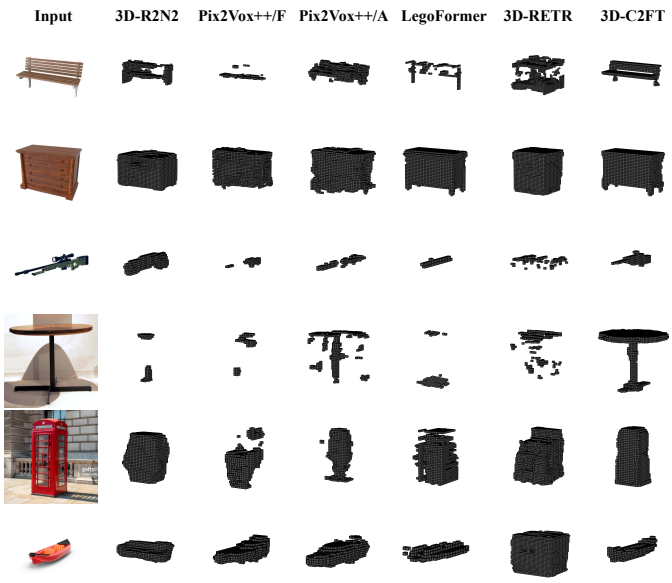

Fig. 6. 3D object reconstruction results for *Case II*. Each test sample mainly contains at least 6 to 11 views.

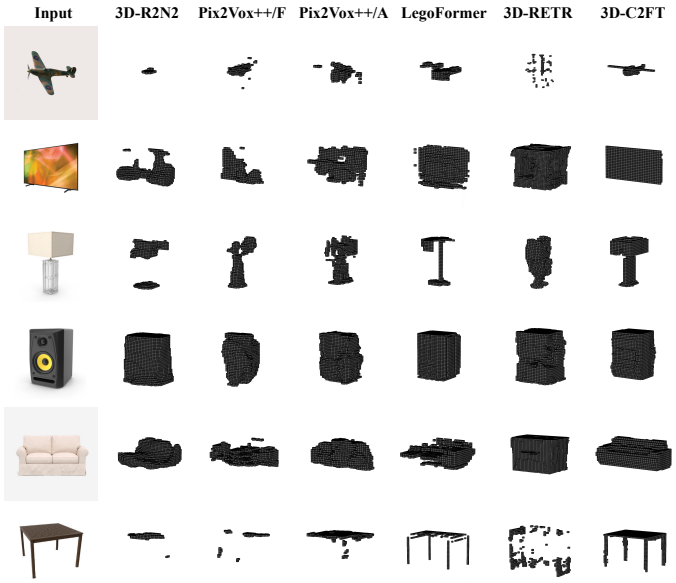

Fig. 7. 3D object reconstruction results for *Case III*. Each test sample mainly contains at least 12 views or more.

In summary, from the analysis of Supplementary Fig. 5–7, we find all the competing models cannot reconstruct the real-life 3D objects decently. Notably, the proposed model performs pretty well even in *Case I*, which is considered the most challenging one. This observation again vindicates the effectiveness of the C2F attention mechanism applied in the 3D-C2FT. We additionally demonstrate several sample results in Video02.mp4.

References

1. Shi, Z., Meng, Z., Xing, Y., Ma, Y., Wattenhofer, R.: 3D-RETR: End-to-end single and multi-view 3D reconstruction with transformers. In: British Machine Vision Conference (BMVC). pp. 1–14 (2021)
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 Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1912–1920 (2015)
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)
4. Yagubbayli, F., Tonioni, A., Tombari, F.: LegoFormer: Transformers for block-by-block multi-view 3D reconstruction. arXiv e-prints (2021), http://arxiv.org/abs/2106.12102