Supplementary Material

1. Voxel Representation

In this paper, our method employs voxel-based 3D representation. It is not necessary to use any extra information for reconstruction. However, methods using implicit representation require camera parameters and methods with mesh representation need an assistive tool.

2. Evaluation Metrics

In Section 4.1 of the main paper, we mention that Intersection over Union (IoU) and F-Score@1% are used as the evaluation metrics to measure the performance of methods. It is emphasized that these two metrics are commonly used and recognized in works related to voxel-based 3D reconstruction. Their details are also kept the same as previous research works and will be elaborated as follow:

Intersection over Union. The predicted probabilities p should be binarized according to a preset threshold and then compare the voxel grids with the ground truth gt, which is defined as:

$$\mathrm{IoU} = \frac{\sum_{(i,j,k)} \mathrm{I}(p_{(i,j,k)} > t)\mathrm{I}(gt_{(i,j,k)})}{\sum_{i,j,k} \mathrm{I}[\mathrm{I}(p_{(i,j,k)} > t) + \mathrm{I}(gt_{(i,j,k)})]}, \tag{1}$$

where $\mathrm{I}(\cdot)$ is an indicator function and the subscript (i, j, k) denotes the occupancy probability of the grid located on the corresponding position.

F-Score@1%. [8] firstly proposes F-Score and [10] introduces it as an extra metric to 3D reconstruction task. F-Score is defined as:

$$\text{F-Score }(d) = \frac{2\mathrm{P}(d)\mathrm{R}(d)}{\mathrm{P}(d) + \mathrm{R}(d)}, \tag{2}$$

where $P(d)$ and $R(d)$ indicate the precision and recall while the distance threshold is d. The precision and recall can be calculated as:

$$\mathrm{P}(d) = \frac{1}{|\mathcal{R}|} \sum_{\mathbf{r}\in\mathcal{R}} \left[min_{\mathbf{g}\in\mathcal{G}} \|\mathbf{r} - \mathbf{g}\| < d\right], \tag{3}$$

$$\mathrm{R}(d) = \frac{1}{|\mathcal{R}|} \sum_{\mathbf{r}\in\mathcal{G}} \left[min_{\mathbf{r}\in\mathcal{R}} \|\mathbf{g} - \mathbf{r}\| < d\right], \tag{4}$$

where $\mathcal{R}$ and $\mathcal{G}$ are respectively the point clouds of the reconstruction object and ground truth. The surface of voxel objects is generated by the marching cubes algorithm [6]. The point clouds with 8192 points sampled from the surface are utilized to calculate F-Score. F-Score@1% means the F-Score value when $d = 1\%$. The above settings are completely consistent with [10].

3. Setup Details of Experiments in Table 5

The experiments as shown in Table 5 of our paper attempt to verify that the inter-view-decoupled strategy based on mining the correlations between similar tokens is most suitable for handling unordered multiple images in the multi-view 3D reconstruction task. Consequently, the control groups adapt different encoding strategies and end up with attention-based fusion [12], the most advanced aggregate approach before our work, to connect to the shape reconstruction stage. In addition, all methods are based on ViT [4] with 12 transformer blocks to manipulate the variables. The setup details of them are as follows:

- **Independent Branches.** As the baseline, ViT [4] extracts the feature from view images in parallel while there is no communication between each branch until the fusion module.
- **Video Transformer with Joint Attention** [13]. All tokens from various views are processed uniformly in each attention layers without decoupled encoding.
- **Video Transformer.** The three methods establish intra-view-decoupled encoding utilizing ViT and insert inter-view-decoupled encoding based on the temporally-coherence property in different ways.
  - **Factorised Transformer Block** [5]. The 12 transformer blocks are divided into 2 types: intra-view-decoupled transformer blocks and inter-view-decoupled transformer blocks. The former independently processes each image. Whereas, the latter processes on different images at the same spatial regions and the regions are defined to a size of 7 × 7 tokens without overlap. These two types of blocks are executed alternately.
  - **Factorised Attention** [2]. An extra multi-head attention (MHA) layer initialised with zero for

its all weights is inserted between the original MHA layer and feed-forward network in each transformer block. The original MHA layer computes self-attention on the intra-view-dimension and the inserted MHA layer does on the inter-view-dimension.

  - **Factorised Dot-Product** [1]. In each MHA, intra-view-decoupled attention operation and inter-view-decoupled attention operation are factorised to compute using different heads in parallel. This method has the same number of parameters as ViT.

- **Ours.** This experiment adapts the UMIFormer model.

Obviously, the inter-view-decoupled strategies used in video transformer networks do not conform to the nature of unordered multiple images, while the strategy based on mining the correlations between similar tokens can establish a relatively reasonable connection between views. Therefore, our proposed inter-view-decoupled strategy is more suitable for multi-view 3D reconstruction, as demonstrated by the experiment results.

4. Number of View Input during Training

For multi-view reconstruction algorithms, the performance to process a heavy amount of input can be better by increasing the view number during training. All SOTA methods we compared in Table 1 of the main paper employ not less than 3 views as input for training. Among them, the training view number of AttSets [12] even attain 24. Therefore, the effectiveness of UMIFormer can be fully verified through the great multi-view reconstruction performance when only adopting 3 views during training.

5. Supplementary Experiments

5.1. 24-View Reconstruction Results

As a work on multi-view 3D reconstruction, it is necessary to pay attention to the performance when facing a large number of view inputs. As shown in Table 1, we take 24 view inputs as an example. UMIFormer and UMIFormer+ have significant advantages over the other SOTA methods in terms of reconstructing each category.

5.2. More Reconstruction Examples

Figure 1, Figure 2 and Figure 3 supplement more reconstruction examples on the test set of ShapeNet using various methods, including UMIFormer, UMIFormer+ and [3, 12, 10, 11, 7, 14] when facing 5 views, 10 views, 15 views and 20 views as input.

5.3. Various Decoders

In Table 2, we provide extra ablation experiments that use other decoder architectures (from EVolT [9] and LegoFormer [11]) to supplement the experiment results in Table 2 of the main paper. It verifies that IVDB and STM hold effective under various decoder networks. Among them, the performance when using the decoder of EVolT is even better than our proposed model shown in the main paper which uses the decoder of 3D-RETR [7]. However, EVolT lacks some implementation details in the public information, hence we are worried whether our reproduction follows the original work exactly. Therefore, we do not use this better-performance version in this paper.

5.4. Scheme of Inserting IVDB

Our proposed UMIformer model uses multiple IVDBs. In order to prove the necessity of this scheme, Table 3 compares the performance when using IVDB once and using it repeatedly. Obviously, using IVDB once is also effective while weaker than using it repeatedly. It makes us realize that the number of IVDBs used is actually a tradeoff between performance and efficiency.

5.5. Pre-Training of ViT

In Section 4.3 of the main paper, we mention that inserting IVDB needs to preserve the pre-training advantages of ViT. Because the model performance relies heavily on it. If intra-view modules are without pre-learned parameters, the performance will be extremely poor as shown in Table 4. There is even an abnormal result that performance seriously degrades when the view increases.

References

[1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6836–6846, 2021.

[2] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, volume 2, page 4, 2021.

[3] Christopher B Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *European conference on computer vision*, pages 628–644. Springer, 2016.

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

	24-view IoU						24-view F-Score@1%					
Category	**Pix2Vox++ [10]**	**EVolT[9]**	**GARNet [14]**	**GARNet+**	**UMIFormer**	**UMIFormer+**	**Pix2Vox++**	**EVolT**	**GARNet**	**GARNet+**	**UMIFormer**	**UMIFormer+**
airplane	0.729	0.741	0.724	0.739	0.769	**0.789**	0.614	0.636	0.606	0.628	0.667	**0.691**
bench	0.686	0.707	0.698	0.707	0.738	**0.761**	0.522	0.548	0.536	0.551	0.498	**0.600**
cabinet	0.829	0.832	0.841	0.840	0.861	**0.877**	0.456	0.464	0.473	0.473	0.498	**0.515**
car	0.883	0.894	0.888	0.894	0.895	**0.903**	0.598	0.624	0.608	0.623	0.622	**0.641**
chair	0.647	0.681	0.674	0.683	0.713	**0.735**	0.341	0.373	0.369	0.384	0.399	**0.419**
display	0.613	0.674	0.668	0.665	0.742	**0.768**	0.335	0.403	0.386	0.396	0.454	**0.485**
lamp	0.493	0.520	0.516	0.513	0.570	**0.610**	0.351	0.366	0.366	0.369	0.410	**0.451**
speaker	0.762	0.772	0.773	0.772	0.820	**0.840**	0.326	0.339	0.338	0.346	0.392	**0.418**
rifle	0.686	0.711	0.697	0.709	0.760	**0.784**	0.624	0.653	0.634	0.647	0.707	**0.736**
sofa	0.782	0.800	0.807	0.810	0.825	**0.840**	0.454	0.478	0.489	0.500	0.505	**0.528**
table	0.666	0.675	0.693	0.692	0.726	**0.744**	0.419	0.431	0.449	0.452	0.467	**0.481**
telephone	0.849	0.867	0.871	0.879	0.887	**0.904**	0.666	0.687	0.698	0.716	0.709	**0.736**
watercraft	0.668	0.693	0.693	0.696	0.723	**0.745**	0.460	0.494	0.494	0.504	0.534	**0.567**
Overall	0.720	0.738	0.737	0.742	0.771	**0.790**	0.473	0.497	0.493	0.505	0.525	**0.548**

Table 1: Evaluation and comparison of the performance for 24-view reconstruction on the test set of ShapeNet using IoU / F-Score@1%.

IVDB	**Fusion**	**Decoder of EVolT [9]**						**Decoder of LegoFormer [11]**					
		3 views	**5 views**	**8 views**	**12 views**	**16 views**	**20 views**	**3 views**	**5 views**	**8 views**	**12 views**	**16 views**	**20 views**
✗	**ABM**	0.7399	0.7484	0.7523	0.7546	0.7561	0.7567	0.7325	0.7401	0.7443	0.7465	0.7480	0.7485
✗	**STM**	0.7504	0.7594	0.7631	0.7649	0.7655	0.7655	0.7453	0.7532	0.7581	0.7596	0.7610	0.7615
✓	**ABM**	0.7419	0.7509	0.7555	0.7577	0.7592	0.7598	0.7425	0.7512	0.7560	0.7588	0.7605	0.7609
✓	**STM**	**0.7536**	**0.7633**	**0.7676**	**0.7699**	**0.7710**	**0.7714**	**0.7490**	**0.7589**	**0.7642**	**0.7664**	**0.7681**	**0.7682**

Table 2: Supplementary ablation experiments about various decoders.

Scheme	**3 views**	**5 views**	**8 views**	**12 views**	**16 views**	**20 views**
w/o IVDB	0.7477	0.7557	0.7587	0.7598	0.7606	0.7606
only IVDB once	0.7476	0.7564	0.7606	0.7624	0.7633	0.7636
IVDB repeatedly	**0.7518**	**0.7612**	**0.7661**	**0.7682**	**0.7696**	**0.7702**

Table 3: Comparison of the performance when using different schemes of inserting IVDB.

Pre-training	**3 views**	**5 views**	**8 views**	**12 views**	**16 views**	**20 views**
✗	0.6075	0.5903	0.5505	0.5247	0.5144	0.5082
✓	**0.7518**	**0.7612**	**0.7661**	**0.7682**	**0.7696**	**0.7702**

Table 4: Comparison of the performance of whether the intra-view-decoupled transformer blocks are initialized by the pre-trained ViT.

[5] Rui Liu, Hanming Deng, Yangyi Huang, Xiaoyu Shi, Lewei Lu, Wenxiu Sun, Xiaogang Wang, Jifeng Dai, and Hongsheng Li. Decoupled spatial-temporal transformer for video inpainting. *arXiv preprint arXiv:2104.06637*, 2021.

[6] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. *ACM siggraph computer graphics*, 21(4):163–169, 1987.

[7] Zai Shi, Zhao Meng, Yiran Xing, Yunpu Ma, and Roger Wattenhofer. 3d-retr: End-to-end single and multi-view 3d reconstruction with transformers. In *British Machine Vision Conference*, 2021.

[8] Maxim Tatarchenko, Stephan R Richter, René Ranftl, Zhuwen Li, Vladlen Koltun, and Thomas Brox. What do single-view 3d reconstruction networks learn? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3405–3414, 2019.

[9] Dan Wang, Xinrui Cui, Xun Chen, Zhengxia Zou, Tianyang Shi, Septimiu Salcudean, Z Jane Wang, and Rabab Ward. Multi-view 3d reconstruction with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5722–5731, 2021.

[10] Haozhe Xie, Hongxun Yao, Shengping Zhang, Shangchen Zhou, and Wenxiu Sun. Pix2vox++: Multi-scale context-aware 3d object reconstruction from single and multiple images. *International Journal of Computer Vision*, 128(12):2919–2935, 2020.

[11] Farid Yagubbayli, Alessio Tonioni, and Federico Tombari. Legoformer: Transformers for block-by-block multi-view 3d reconstruction. *arXiv preprint arXiv:2106.12102*, 2021.

[12] Bo Yang, Sen Wang, Andrew Markham, and Niki Trigoni. Robust attentional aggregation of deep feature sets for multi-view 3d reconstruction. *International Journal of Computer Vision*, 128(1):53–73, 2020.

[13] Yanhong Zeng, Jianlong Fu, and Hongyang Chao. Learning joint spatial-temporal transformations for video inpainting. In *European Conference on Computer Vision*, pages 528–543. Springer, 2020.

[14] Zhenwei Zhu, Liying Yang, Xuxin Lin, Lin Yang, and Yanyan Liang. Garnet: Global-aware multi-view 3d reconstruction network and the cost-performance tradeoff. *Pattern Recognition*, page 109674, 2023.

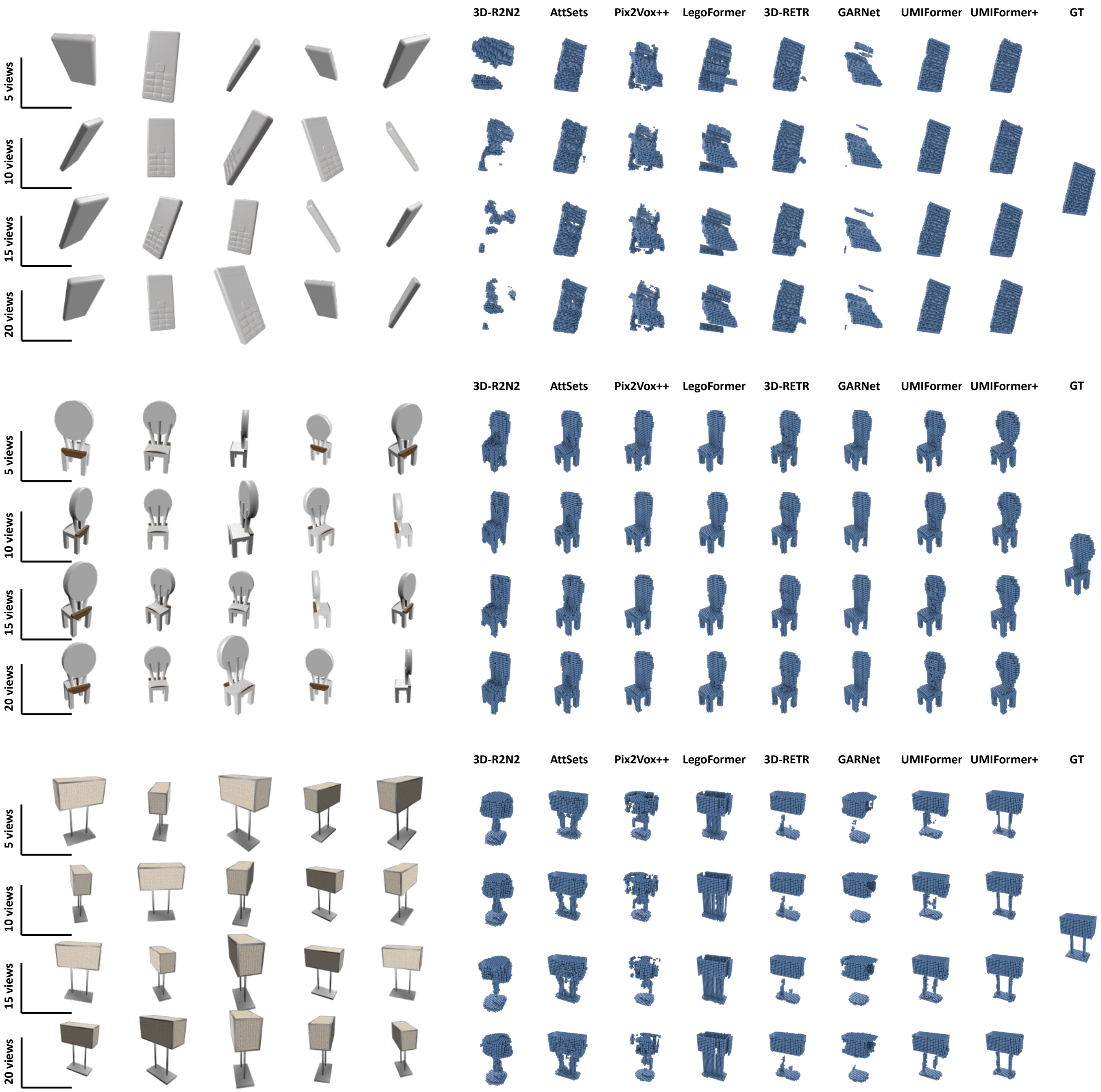

Figure 1: Qualitative reconstruction results when facing 5 views, 10 views, 15 views and 20 views as input for telephone, chair and lamp.

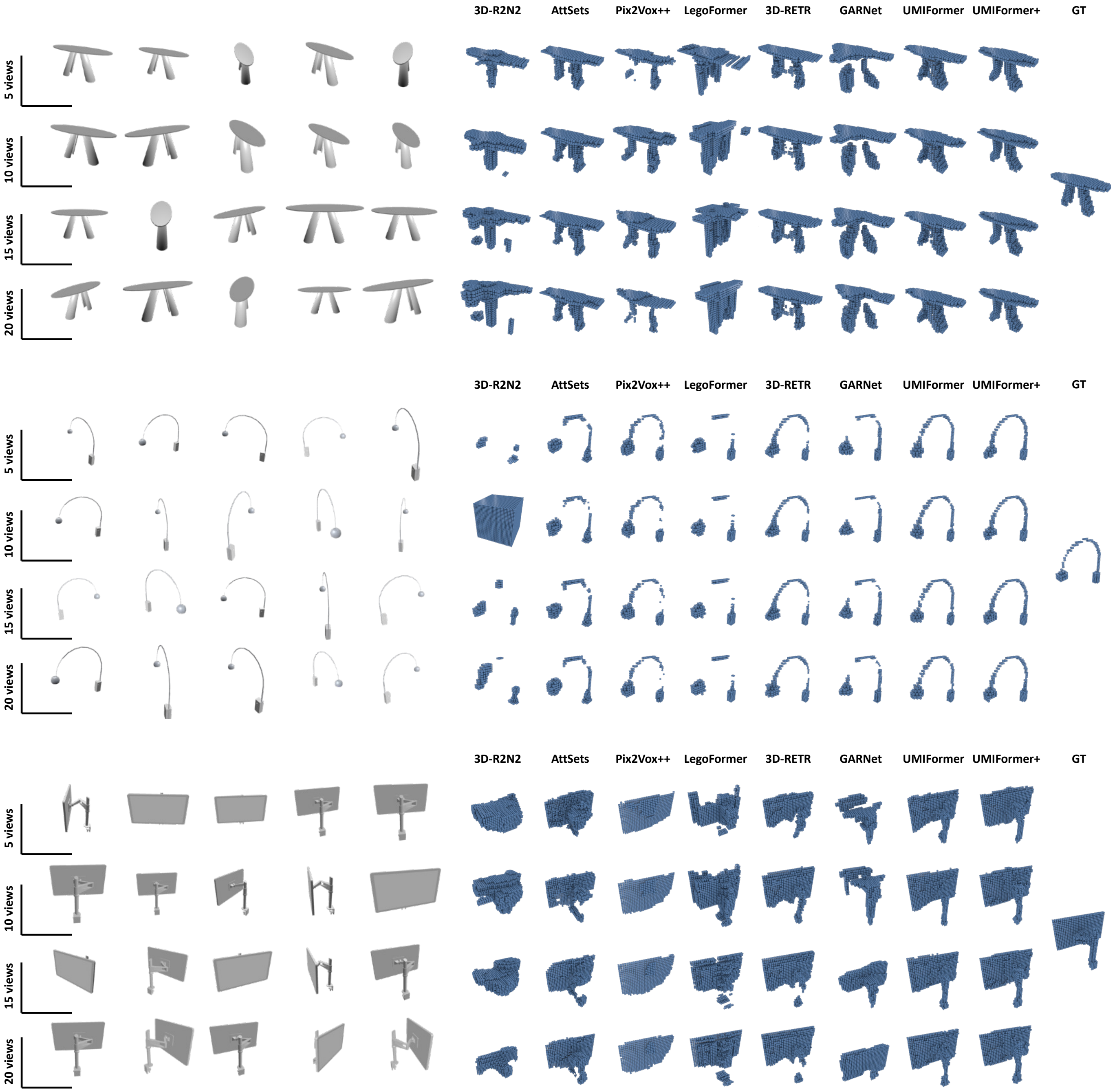

Figure 2: Qualitative reconstruction results when facing 5 views, 10 views, 15 views and 20 views as input for table, lamp and display.

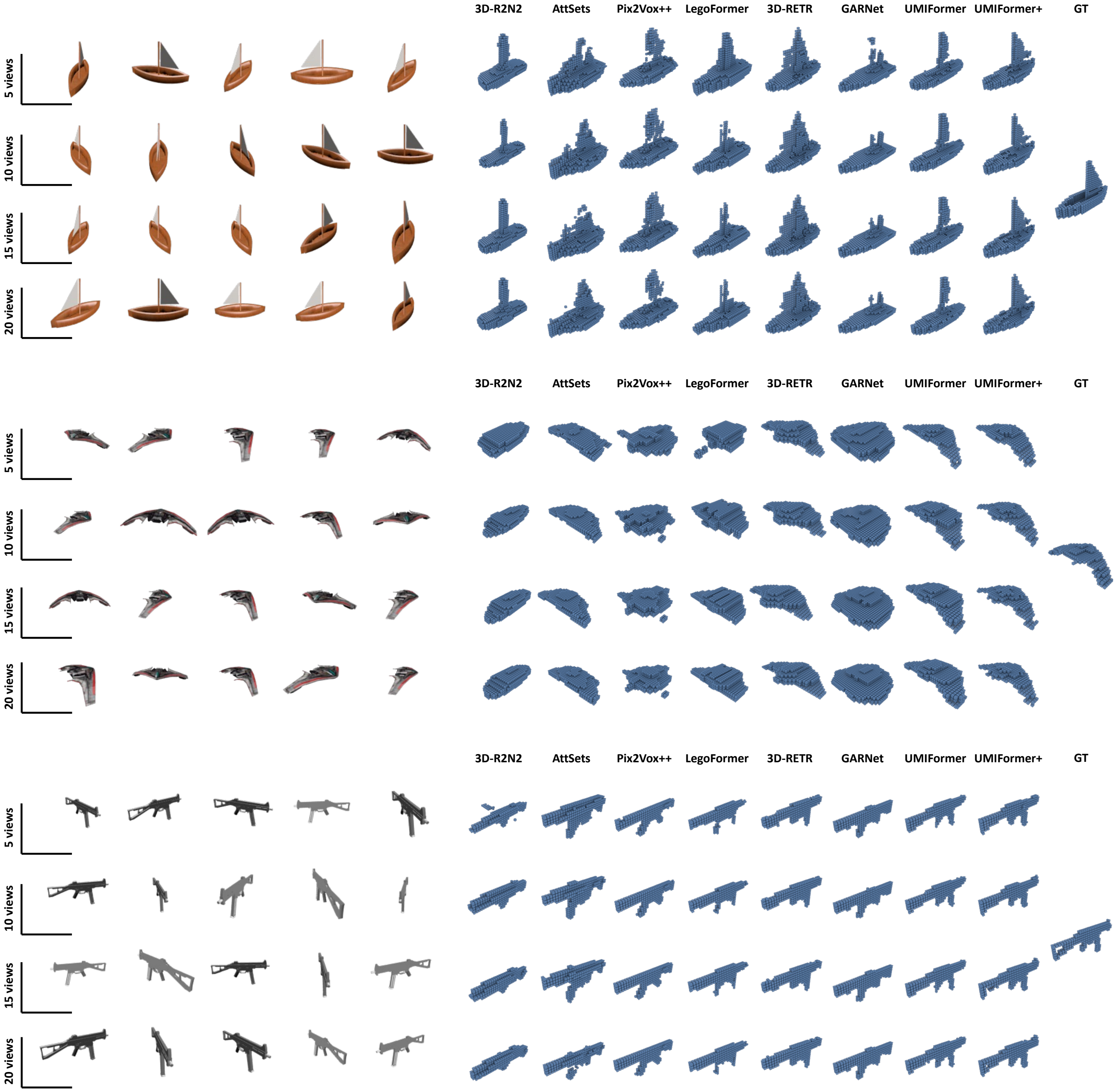

Figure 3: Qualitative reconstruction results when facing 5 views, 10 views, 15 views and 20 views as input for watercraft, airplane and rifle.