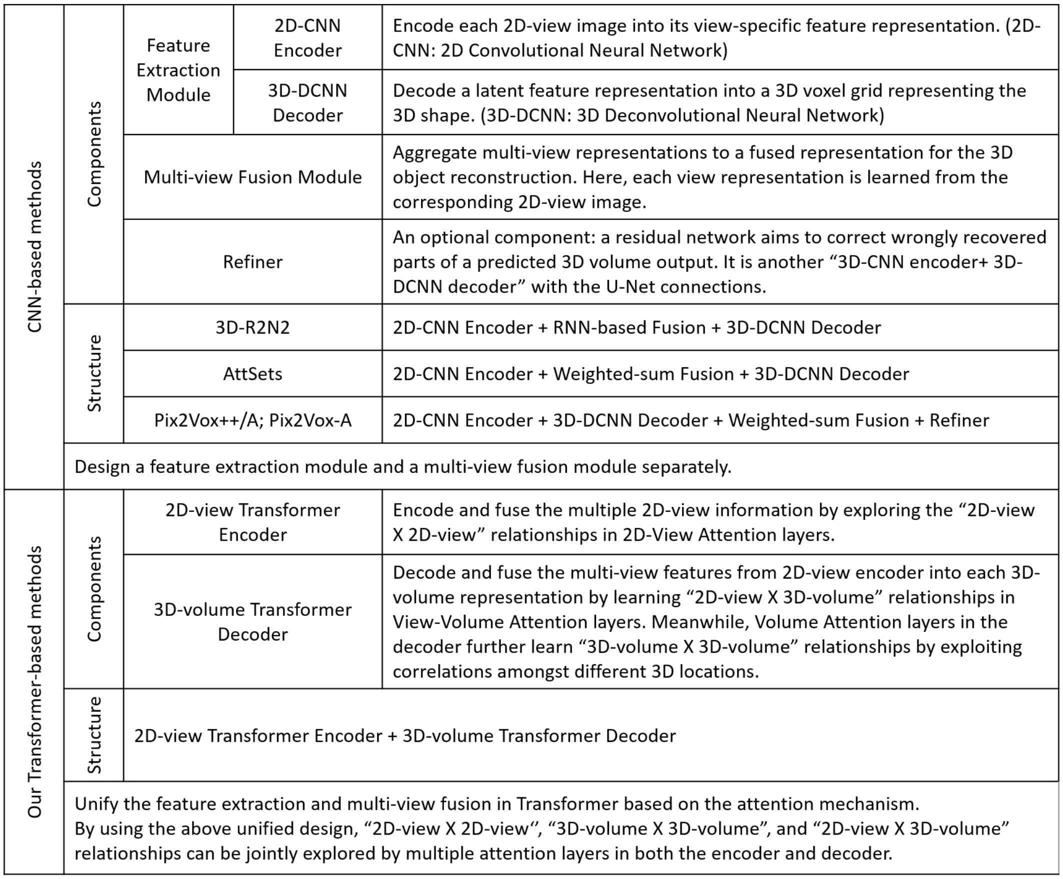
Supplementary Material:
Multi-view 3D Reconstruction with Transformers

Dan Wang[1] Xinrui Cui[2*] Xun Chen[2]
Zhengxia Zou[3] Tianyang Shi[4] Septimiu Salcudean[1] Z. Jane Wang[1] Rabab Ward[1]
[1] University of British Columbia [2] University of Science and Technology of China
[3] University of Michigan, Ann Arbor [4] NetEase Fuxi AI Lab

1. Comparison of Structures in Competing Methods

Figure 1 gives a comparison between the proposed transformer-based 3D reconstruction methods and existing CNN-based methods from the perspective of components and structure.



CNN-based methods	Components	Feature Extraction Module	2D-CNN Encoder	Encode each 2D-view image into its view-specific feature representation. (2D-CNN: 2D Convolutional Neural Network)
			3D-DCNN Decoder	Decode a latent feature representation into a 3D voxel grid representing the 3D shape. (3D-DCNN: 3D Deconvolutional Neural Network)
		Multi-view Fusion Module		Aggregate multi-view representations to a fused representation for the 3D object reconstruction. Here, each view representation is learned from the corresponding 2D-view image.
		Refiner		An optional component: a residual network aims to correct wrongly recovered parts of a predicted 3D volume output. It is another "3D-CNN encoder+ 3D-DCNN decoder" with the U-Net connections.
	Structure	3D-R2N2		2D-CNN Encoder + RNN-based Fusion + 3D-DCNN Decoder
		AttSets		2D-CNN Encoder + Weighted-sum Fusion + 3D-DCNN Decoder
		Pix2Vox++/A; Pix2Vox-A		2D-CNN Encoder + 3D-DCNN Decoder + Weighted-sum Fusion + Refiner
	Design a feature extraction module and a multi-view fusion module separately.			
Our Transformer-based methods	Components	2D-view Transformer Encoder		Encode and fuse the multiple 2D-view information by exploring the "2D-view X 2D-view" relationships in 2D-View Attention layers.
		3D-volume Transformer Decoder		Decode and fuse the multi-view features from 2D-view encoder into each 3D-volume representation by learning "2D-view X 3D-volume" relationships in View-Volume Attention layers. Meanwhile, Volume Attention layers in the decoder further learn "3D-volume X 3D-volume" relationships by exploiting correlations amongst different 3D locations.
	Structure	2D-view Transformer Encoder + 3D-volume Transformer Decoder		
	Unify the feature extraction and multi-view fusion in Transformer based on the attention mechanism. By using the above unified design, "2D-view X 2D-view", "3D-volume X 3D-volume", and "2D-view X 3D-volume" relationships can be jointly explored by multiple attention layers in both the encoder and decoder.			

Figure 1. Comparison of CNN-based methods (3D-R2N2 [1], AttSets [4], Pix2Vox/A [2], Pix2Vox++/A [3]) and the proposed Transformer-based methods.

[*] Corresponding Author. Email:xinruic@ece.ubc.ca

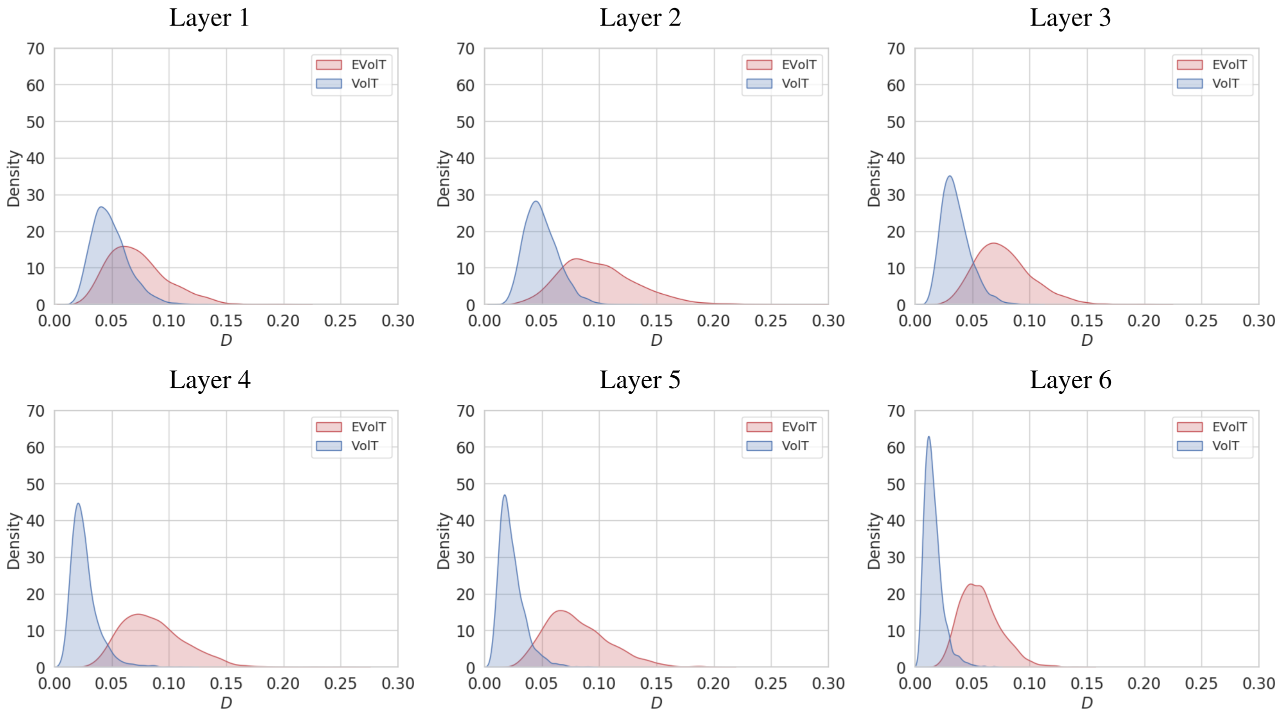


Figure 2. Kernel density estimation of D value in different attention layers for VolT and EVolT.

2. Additional Results

2.1. View Divergence

In Figure 2, we plot the estimated probability density of the D value at different attention layers for VolT and EVolT. We use kernel density estimation (KDE) to compute the probability density and explore the convergence of multi-view representations in different attention layers. A small D means a more considerable convergence of multi-view representations.

In each view attention layer, the probability density function $\hat{p}(D)$ of D is estimated as

$$\begin{aligned} \hat{p}(D) &= \frac{1}{N_{object}N_{view}h}\sum_{i}^{N_{object}}\sum_{m}^{N_{view}} K(\frac{D_m^i - D}{h}), \\ \text{where}\quad D_m^i &= \|\mathbf{s}_m^i - \frac{1}{N_{view}}\sum_{m}^{N_{view}} \mathbf{s}_m^i\|_2. \end{aligned} \tag{1}$$

where $\mathbf{s}_m^i$ is the attention vector of the m-th view for the i-th object. The number of random objects is set to $N_{object} = 100$. The input view number is set to $N_{view} = 24$. Here, we used the Gaussian kernel $K(x) = \frac{1}{\sqrt{2\pi}}\exp(-\frac{x^2}{2})$. h is computed by the rule of thumb of Scott.

It is shown in Figure 2 that the density of EVolT has a much larger variance than that of the VolT. Also, as the attention layers go deeper, the D value of the VolT gradually moves closer to 0 while the EVolT can still cover a larger range of D values. This indicates that the divergence enhancement function in EVolT can effectively slow down the convergence degradation of multi-views in deeper layers.

2.2. Qualitative Results

We provide more object reconstruction results of competing methods, as shown in Figure 3, 4, 5, and 6. In each object sample, we provide object reconstruction results from different numbers of input views, i.e., 12 views, 18 views, and 24 views. The first two rows on the left part of Figure 3 show the 12 input views of an object, and the corresponding reconstruction results of competing methods are shown at the second row on the right. Similarly, the first three rows on the left are the 18 input views corresponding to the results on the right. The qualitative comparison suggests the superiority of the proposed method in terms of the reconstruction topology and details.

References

[1] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016. 1

[2] Haozhe Xie, Hongxun Yao, Xiaoshuai Sun, Shangchen Zhou, and Shengping Zhang. Pix2Vox: Context-aware 3D reconstruction from single and multi-view images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2690–2698, 2019. 1

[3] 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. 1

[4] 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. 1

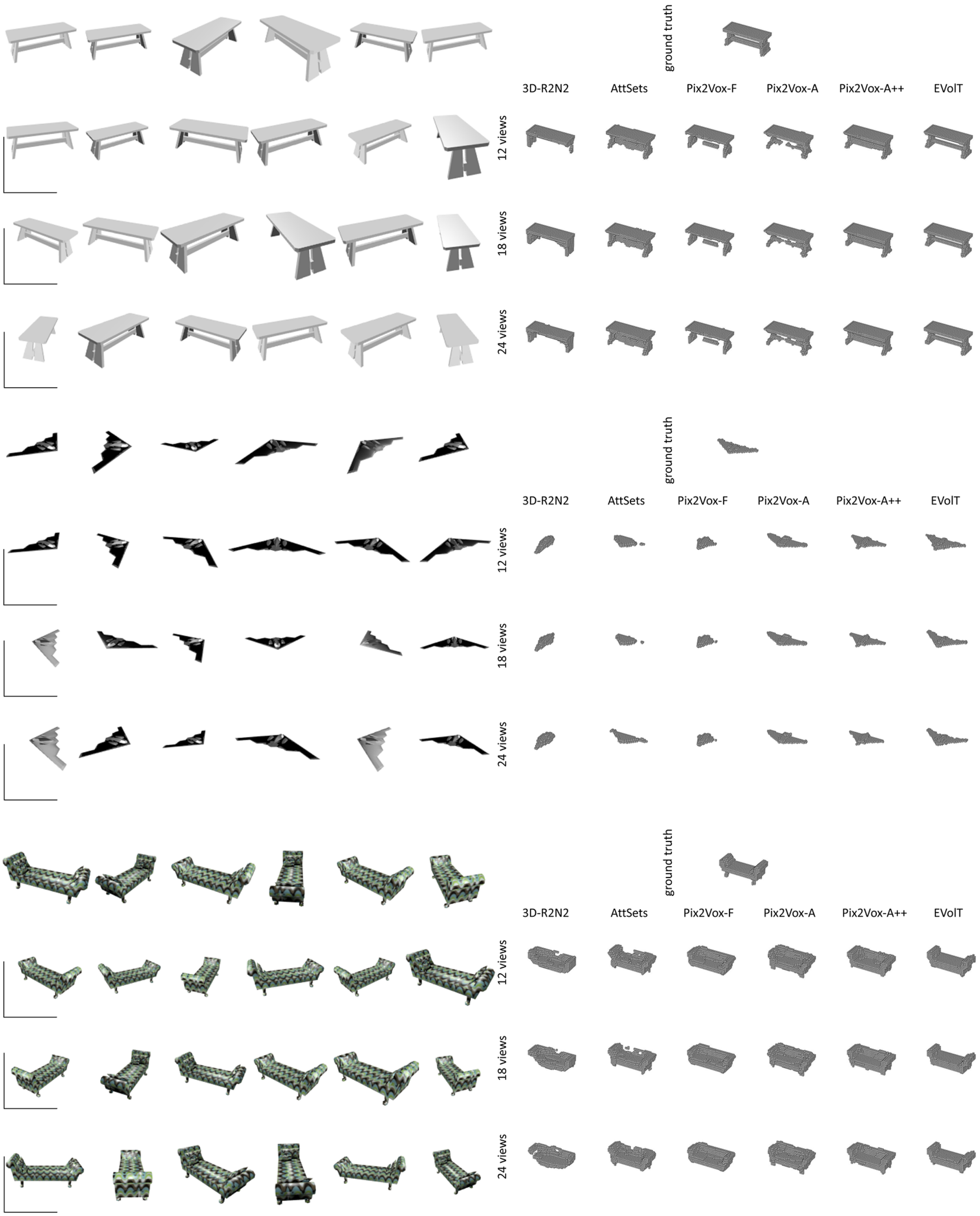


Figure 3. Qualitative reconstruction results of competing methods for bench (top), aeroplane (middle), and sofa (bottom).

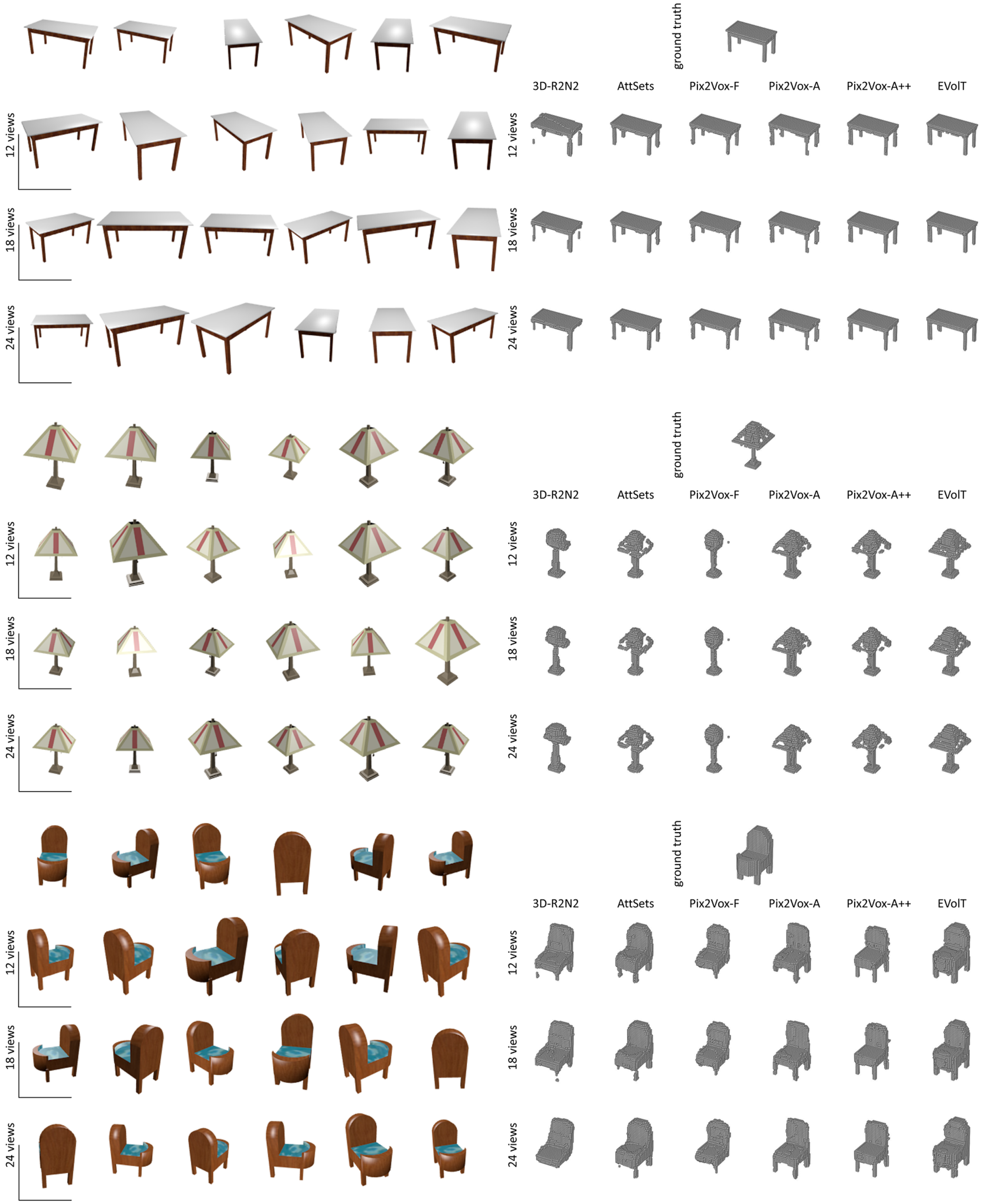


Figure 4. Qualitative reconstruction results of competing methods for table (top), lamp (middle), and chair (bottom).

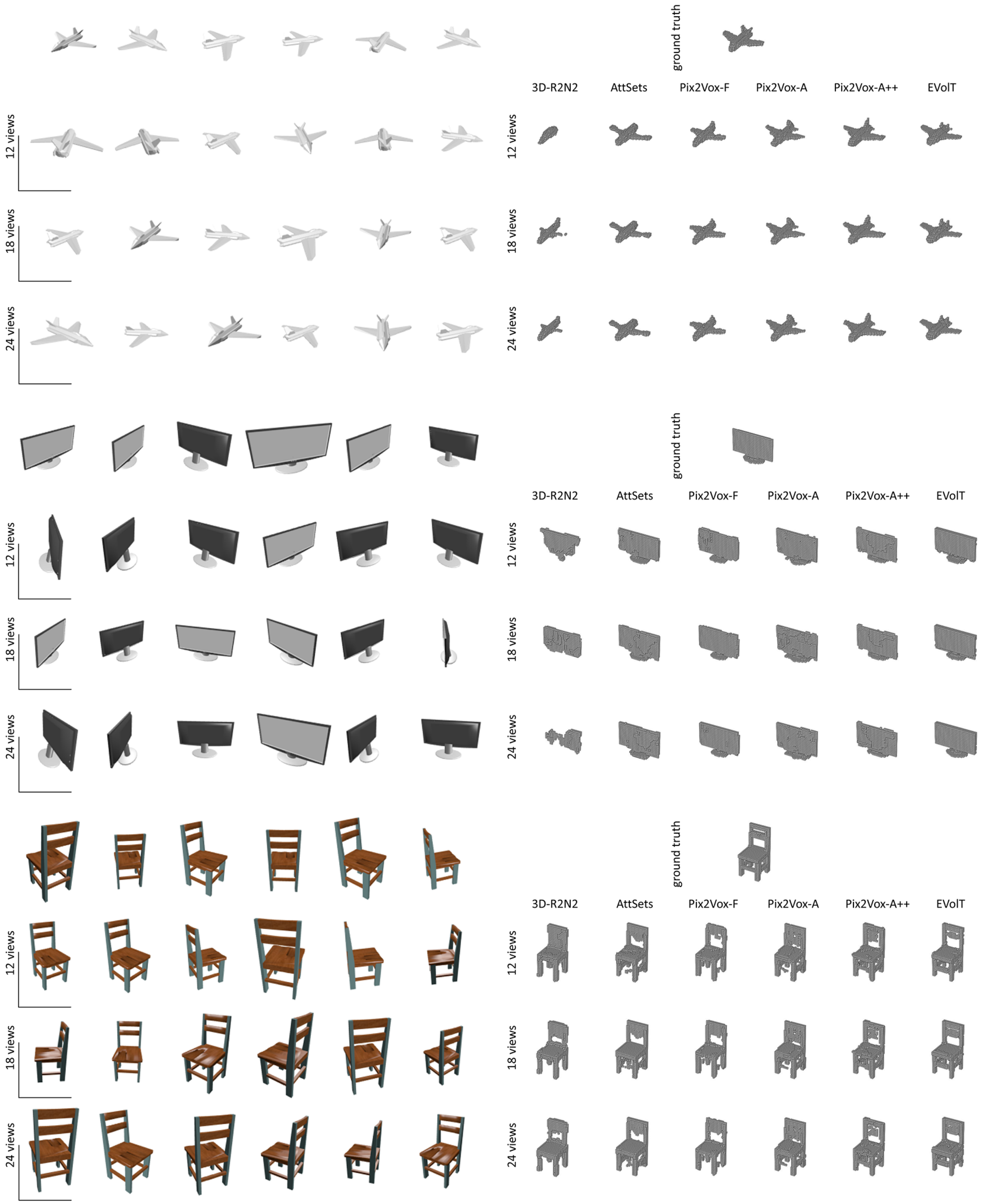


Figure 5. Qualitative reconstruction results of competing methods for aeroplane (top), display (middle), and chair (bottom).

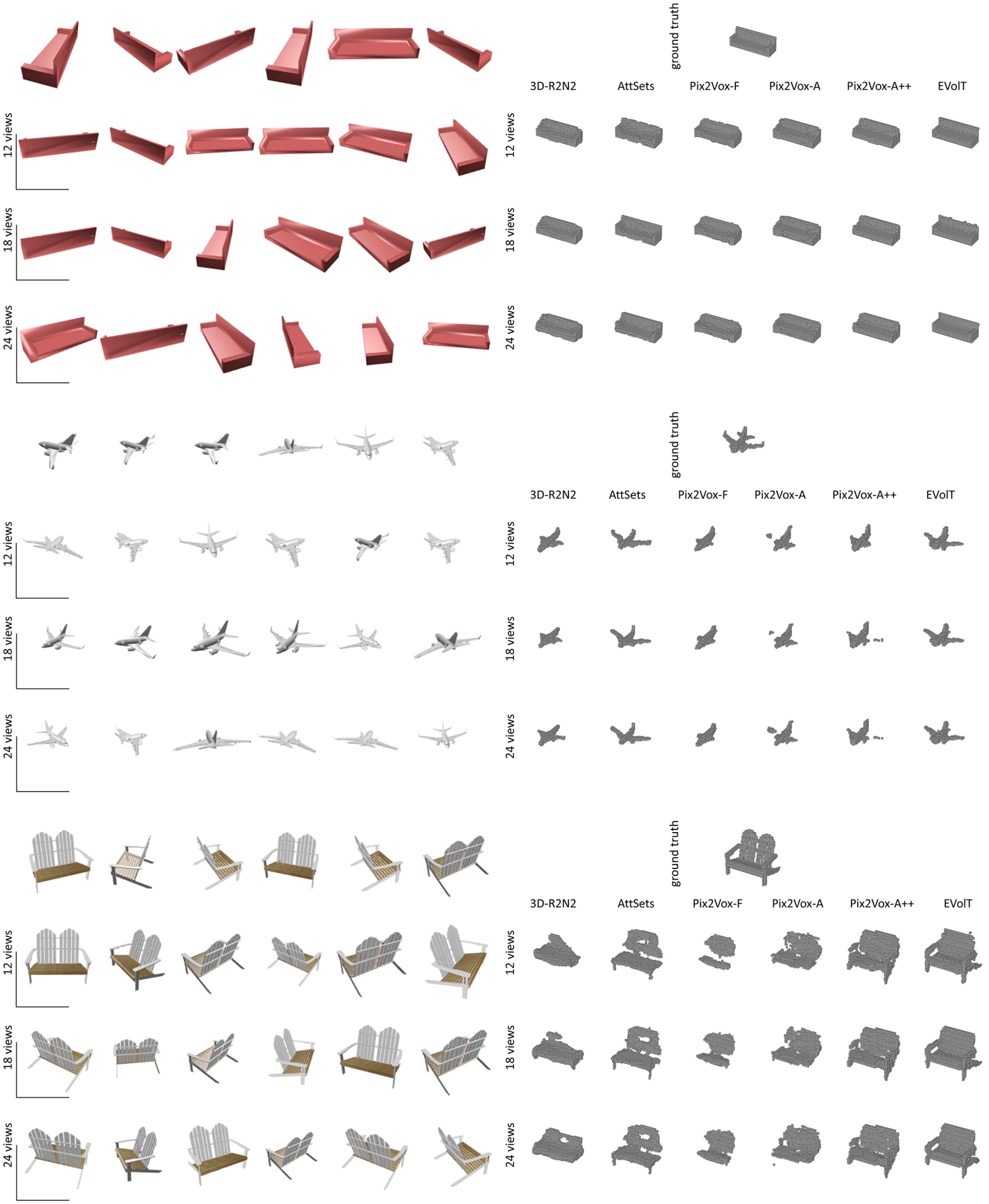


Figure 6. Qualitative reconstruction results of competing methods for sofa (top), aeroplane (middle), and bench (bottom).