We thank the reviewers for their time and insightful comments. We were able to address most of them, which helped improve the paper. We focus on the major comments below, and minor ones will be addressed directly in the paper.

Direct supervision between the two source images [R1, 2, 4]. This is a viable approach; however, it does not generalize as well as our network, as our method reconstructs the source images from unseen target viewpoints, ensuring visual coherence for these views. To support this claim, we designed the proposed approach and retrained the network. Our approach outperforms direct supervision by over 20% (Table/Fig. 1), confirming our approach's superiority.

Methods	Car			Chair		
	$L_1(\downarrow)$	SSIM (↑)	LPIPS (↓)	$L_1(\downarrow)$	SSIM (↑)	LPIPS (↓)
Supervised	0.0368	0.8481	0.1283	0.0786	0.72	0.2501
Ours	**0.0338**	**0.8632**	**0.1029**	**0.0663**	**0.7568**	**0.1766**

Table 1: Direct supervision vs. Ours

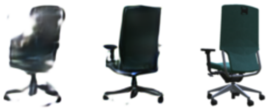

Figure 1: Supervised (left), ours (middle), ground truth (right)

Claim of using no 2D supervision is misleading [R1, 2]. Unsupervised in the paper refers to no target view supervision. Our method does not use any supervision for target viewpoints, only self-supervision on the two source views. We have clarified this point in the paper.

Perceptual evaluation metric [R2]. This is a great point, and we have added the suggested metric to our comparison (Table 1). Full results will be added to the paper.

Model has access to 108 training images during training [R1]. We select two images per object *before* training and only use these for training (LN 187-188). We rewrote the section to avoid confusion.

Evaluation on other datasets than ShapeNet [R1, 3]. This is a great suggestion and we show preliminary results on the real capture dataset (Fig. 2). Additional details and results will be added to the paper.

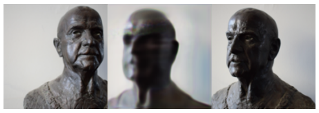

Figure 2: Source (left), prediction (middle), target (right)

Comparison to SynSin [R4]. SynSin allows rotations of limited range (±20°) since it uses a depth point cloud for the 3D features. In our method we generate novel views on the entire viewing hemisphere and therefore did not include this comparison.

Neural renderer description [R3]. We agree with the reviewer and added additional details of the rendering function in the supplementary material.

Choice of neural rendering function [R3]. This is a very active research area, without a clear consensus on which rendering functions are preferable (see the recent summary report "State of the art in neural rendering"). We chose a rasterization based rendering function as it is fast, differentiable, and the splatting of features to the image plane allows for good gradient flow.

Refinement network and adaptive sampling [R3]. The refinement network fills in gaps and improves rendering quality of the rasterization based neural renderer. Adaptive sampling is an interesting suggestion, and our network would likely benefit from such a step. We will keep this as future work.

Target poses T_G specification [R2]. Poses are represented with spherical coordinates represented by azimuth and elevation. The target pose is selected randomly and given to the network as an input (Fig. 2 in the paper).

More realistic scenes [R3, 4]. This is a research direction that we would like to investigate in future work (LN 254). We plan to investigate model generalization to natural images with background and scenes containing multiple objects.

Bilinear sampling vs. Attentive Neural Processes [R1]. Using "Attentive Neural Processes" looks like a viable option, but we have not investigated this as an alternative and do not know if this would improve performance.

Single image cyclic consistency [R4]. This is technically correct. However, the network collapses to a trivial solution (the identity function) when using a single image. We have added this to the paper.

Supervising the occupancy using binary cross-entropy [R4]. For a single instance, the visual hull can be outside of the object's 3D footprint. However, we find that the multi-view consistency enforces the predicted occupancy features to be coherent with the object's actual 3D structure. We also have preliminary results for extracting the 3D model using the occupancy features.

Performance saturation [R4]. We think that using multiple views will improve performance, but we have not investigated this claim nor the issue of performance saturation. Our approach shows that novel view synthesis is feasible, even if only two images per object are available.

Segmentation network [R1]. We do not use any (pre-trained or not) segmentation networks in our method.