From Image Collections to Point Clouds with Self-supervised Shape and Pose Networks: Supplementary Material

K L Navaneet[1] Ansu Mathew[1] Shashank Kashyap[1] Wei-Chih Hung[2]
Varun Jampani[3] R. Venkatesh Babu[1]

[1]Indian Institute of Science [2]University of California, Merced [3]Google Research

Along with the supplementary document, we provide a supplementary video for easier understanding of the proposed approach and better visualization of the 3D reconstruction results. The supplementary document is arranged as follows: We provide training schedule and additional implementation details in the initial sections. We add more detailed ablations on the role of individual components of the proposed cycle consistency based losses. Subsequently we present experimental results on the effect of number of nearest neighbours, consistency and symmetry losses, dense point correspondence, inference stage optimization and colored point cloud reconstruction. We provide qualitative results on failure modes of our approach. Lastly, we provide the architectural details of our reconstruction and pose prediction networks.[1]

1. Training Schedule

We train our networks for 400000 iterations using Adam optimizer with a learning rate of 0.0005. For training our approach, we observe that the pose prediction network converges at a much earlier stage compared to the reconstruction network. At the half-way stage (200000 iterations), we freeze the pose network and train the reconstruction network with just image and mask losses, similar to the DIFFER baseline. We observe that this helps in obtaining better 3D shape reconstructions and eliminate outlier points in predictions.

2. Additional Implementation Details

We choose the optimal hyperparameter values based on the reconstruction performance on the validation set. The weight for geometric consistency loss, β is set to 10000 and pose consistency loss, ρ is set to 1. The weight for nearest neighbours consistency loss κ is set to be same as that for mask loss α. During the second half of the training schedule, the weights for consistency losses β and ρ are set to 0 and that of image and mask losses α is reduced to 10. In the experiments on nearest neighbours, we consider five nearest neighbours for every input among which n images are sampled randomly. The effect of the number of neighbours chosen, n, is presented in Fig. 1 and Table 2. In inference stage optimization experiments, the weights for regularization and symmetry loss, λ and κ are both set to 500.

3. Role of Cycle Consistency Losses

We present quantitative ablation on the role of individual components of our proposed cycle consistency loss in Table 1. We present qualitative comparison of reconstruction with and without these losses in a self-supervised setting in Fig. 2. The network fails to learn meaningful 3D shapes in the absence of the proposed losses, while the reconstructions closely match the input when the losses are utilized. We also observe that each of the individual losses help improve the reconstructions and the best performance is obtained when all the losses are combined. Fig. 3, displays the qualitative results on the effect of geometric consistency loss on the pose supervised ULSP_Sup approach. We observe a significant improvement in the reconstruction quality, suggesting the portable nature of the proposed loss.

4. Effect of Nearest Neighbours

Fig. 3 and Tables 1 and 3 in the main submission demonstrate the efficacy of the nearest neighbours consistency loss. Here, we analyze the effect of the number of chosen nearest neighbours for each image. Table 2 presents quantitative comparison of reconstruction performance for different number of neighbours. We observe a significant improvement when just a single image is utilized. The performance improves or remains nearly same as more number of images are considered. When more than 3 images are

[1] Code is available at https://github.com/val-iisc/ssl_3d_recon

Geometric CC	Pose CC	Nearest		Chamfer			EMD	
		Neighbor CC	Car	Chair	Aero	Car	Chair	Aero
✗	✗	✗	10.33	21.84	15.06	18.32	23.40	16.12
✓	✗	✗	5.78	27.89	10.77	7.07	26.9	15.76
✗	✓	✗	11.31	11.46	12.47	11.59	14.97	15.26
✓	✓	✗	6.39	13.58	8.66	6.42	16.46	12.53
✓	✓	✓	**5.48**	**10.91**	**7.11**	**4.95**	**14.93**	**11.07**

Table 1: **Effect of Consistency Loss:** We evaluate the effect of the proposed consistency losses on reconstruction metrics. The network fails to train in the absence of the consistency losses in the self-supervised setting. Each of the proposed losses is necessary to obtain the optimal performance.

used in loss calculation, we observe a drop in performance. This behaviour is consistent with our expectations, since the farther nearest neighbours have lower geometric similarity with the input image.

5. Effect of Symmetry Loss

Symmetry loss (Section 3.4 of main paper) was proposed as an additional regularization to obtain meaningful 3D reconstructions and to align the reconstructions to a predefined canonical pose. Here, we present quantitative results (Table 3) for reconstruction performance with and without symmetry loss. We use both consistency and nearest neighbor losses for both the methods. We observe that symmetry loss is crucial in getting reasonable reconstructions for the airplane category. It does not affect the performance for the car category while it has a negative impact on chair reconstructions. Similar trends were observed on the validation set too. Based on these observations, we choose the best combination for Ours-NN model. We use the symmetry loss only for the airplane category in Ours-NN.

6. Results on Point Correspondence

We observe that the reconstructed point clouds have dense point-wise correspondence. That is, points with similar indices in the regressed list of points are present in semantically simillar regions. To visualize this, we use a colored UV map to obtain point correspondences on the point cloud. Fig. 4 depicts the UV mapped point clouds. We observe that points with similar color are grouped together and have correspondence across different samples.

7. Results on Inference Stage Optimization

Fig. 4 of the paper demonstrates that ISO results in significant improvement in correspondence of the reconstructions to the input image. We present the corresponding quantitative results in Table 4. The metrics are consistent with our observations that the point cloud structure remains intact in occluded regions while closely matching the input image in the visible regions.

8. Results on Color Prediction

Since our networks predict colored point clouds, we present qualitative and quantitative results on it in Fig 5 and Table 5. Due to the absence of good ground-truth for evaluation of color prediction on point clouds, we project our reconstructions from 10 randomly sampled view-points and perform comparison in the 2D domain. We observe a greater correspondence to the input image in our projections compared to those of the pose supervised DIFFER approach, particularly in the case of car category. Since the color metrics are dependent on the quality of our reconstructions, DIFFER has improved performance in the chair category, while we outperform it in the car category.

9. Failure Cases

Fig. 6 presents a few failure cases. Some reconstructions have high density clusters leaving very few points to model the thinner structures (Fig. 6(a)). Clusters in airplane category lead to reconstructions with thin structures. However, we note that such failure modes are also observed in earlier point cloud reconstruction literature [1] and addressing these forms an important future work. Our approach also fails to accurately model certain structures like the spoilers in cars and complex leg and handle structures in chairs (Fig. 6(b)). Training with larger number of such examples might help alleviate the problem.

10. Network Architecture

Details of our reconstruction and pose network architectures are provided in Tables 6 and 7. We use a dual branch reconstruction network similar to DIFFER [1] for reconstructing point locations and color values. The structure branch of the reconstruction network and the pose network have similar architecture except for the output layer. We use the output of the D_{s1} (6) layer our reconstruction network as the embedding to obtain the nearest neighbours in our experiments.

Neighbours	Car		Chair		Aero	
	Chamfer	EMD	Chamfer	EMD	Chamfer	EMD
0	6.39	6.42	13.58	16.46	8.66	12.53
1	5.47	4.93	10.91	14.93	8.35	12.3
2	5.51	5.29	10.65	15.46	7.1	11.07
3	5.57	5.16	10.90	14.84	8.99	14.02
4	5.54	5.24	11.93	16.64	8.65	13.35

Table 2: **Effect of Nearest Neighbours:** We examine the effect of number of images of nearest neighbours on the reconstruction metrics. The performance improves or remains nearly same as more number of images are considered. When more than 3 images are used in loss calculation, we observe a drop in reconstruction performance due to the increased disparity between neighbours and input image.

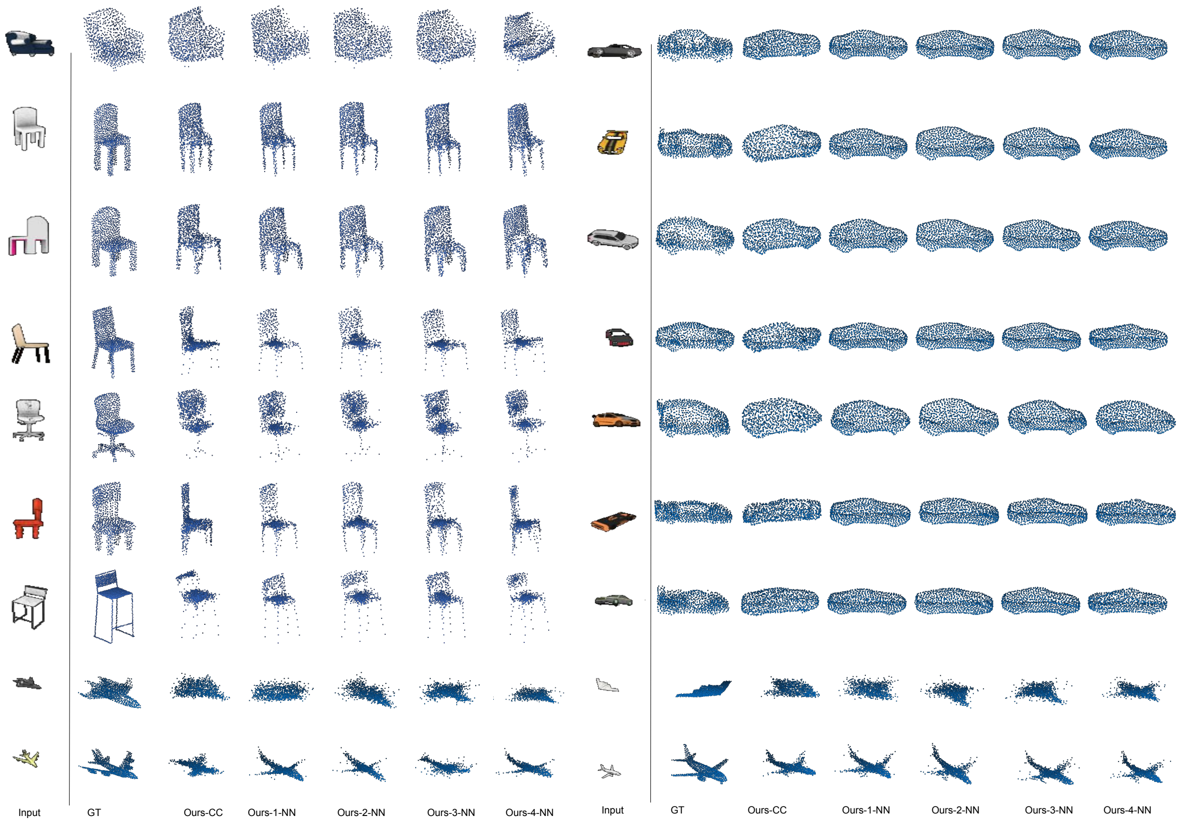

Figure 1: **Effect of Nearest Neighbours:** We examine the effect of number of images of nearest neighbours on the reconstruction metrics. The performance improves or remains nearly same as more number of images are considered. The best performance is achieved when one or two images are used while reconstructions suffer when more than 3 images are utilized.

References

[1] K L Navaneet, Priyanka Mandikal, Varun Jampani, and R Venkatesh Babu. DIFFER: Moving beyond 3d reconstruction with differentiable feature rendering. In *CVPR Workshops*, 2019. 2, 7

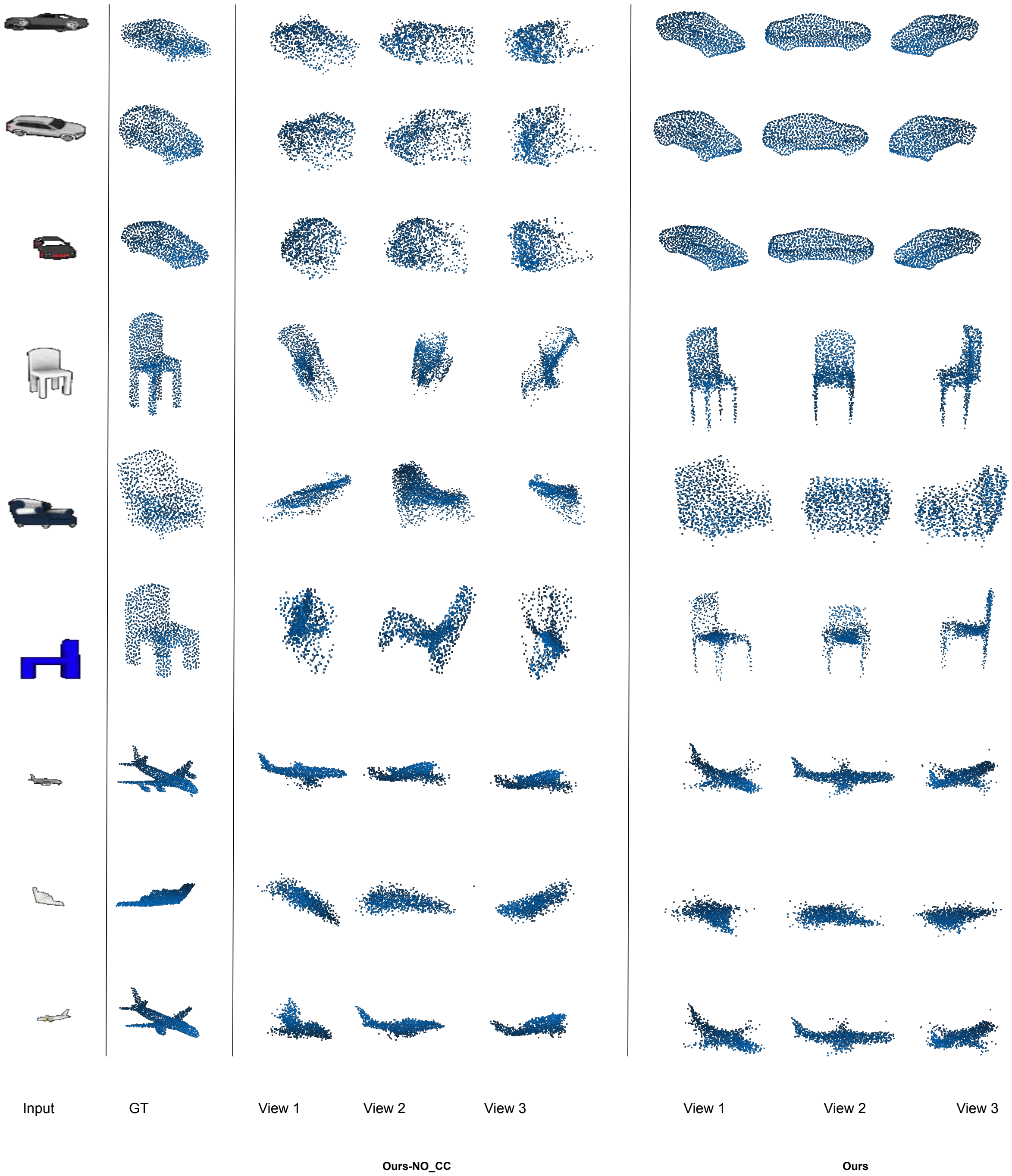

Figure 2: **Effect of Cycle Consistency Loss** The network fails to learn meaningful 3D shapes in the absence of the proposed geometric and pose cycle consistency losses. The reconstructions closely match the input when the losses are utilized.

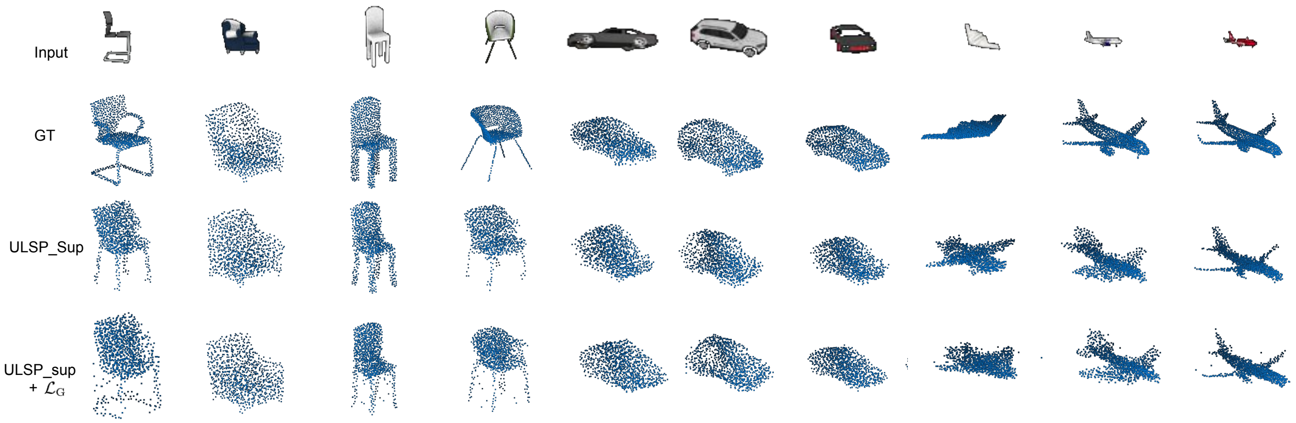

Figure 3: **Portability of Proposed Loss** We employ the proposed geometric cycle consistency loss atop the pose-supervised ULSP approach. We observe a significant improvement in the reconstruction quality, suggesting the portable nature of the proposed loss.

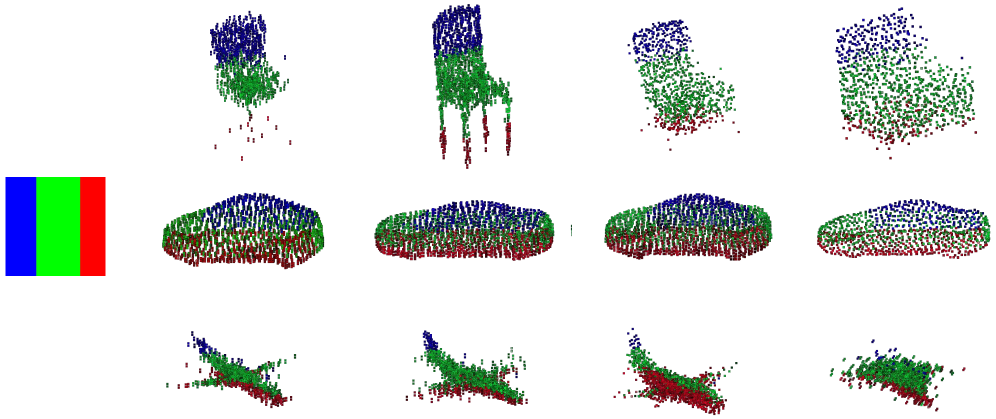

Figure 4: **Point Correspondence:** Similar indices in the point cloud are visualized with the same color. The reconstructions exhibit high point correspondence across models.

Method	Chamfer			EMD		
	Car	Chair	Aero	Car	Chair	Aero
Ours-No-Sym	5.48	10.91	7.91	4.95	14.93	13.98
Ours-Sym	5.72	12.34	7.11	5.24	16.67	11.07

Table 3: **Effect of Symmetry Loss:** Symmetry loss is crucial for effective reconstructions on airplane category. We choose the best settings from the ablation for each category in Ours-NN model.

Categ.	Method	Chamfer	EMD
Car	Ours-NN	5.47	4.93
	Ours-NN post ISO	5.49	5.01
Chair	Ours-NN	10.91	14.93
	Ours-NN post ISO	15.32	17.79
Aero	Ours-NN	7.1	11.07
	Ours-NN post ISO	7.62	11.09

Table 4: **Quantitative Analysis of ISO:** Chamfer and EMD metrics before and after inference stage optimization are comparable. This indicates that the point cloud structures are not degraded in occluded regions due to ISO.

Method	Car	Chair	Aero
DIFFER	8.59	12.81	4.69
Ours-CC	8.58	14.19	4.8
Ours-NN	8.09	13.51	4.77

Table 5: **Color Metrics:** We present the $\mathcal{L}_2$ distance between predicted projections and ground-truth images to evaluate color prediction. We either outperform or perform comparably to the pose supervised DIFFER approach.

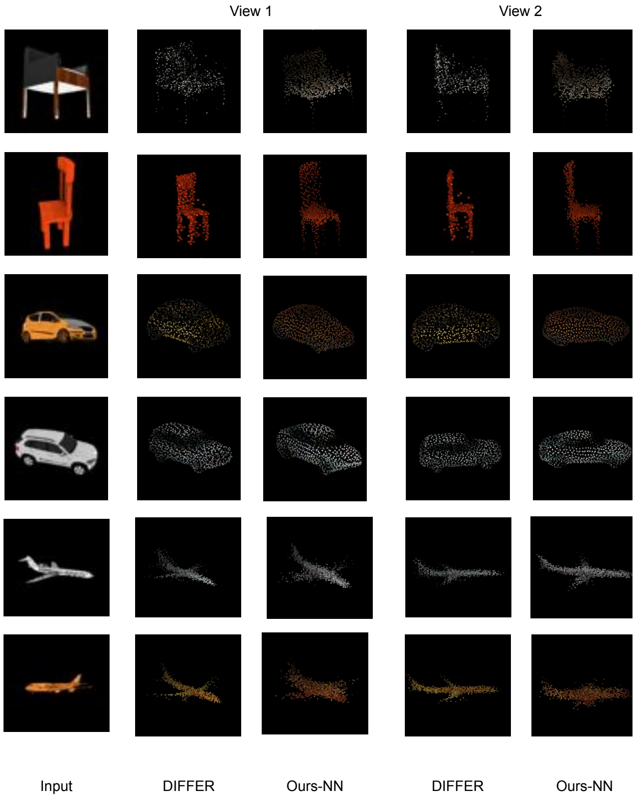

Figure 5: **Colored Point Cloud Reconstruction** We compare the colored point cloud reconstructions of DIFFER and our approach. We achieve higher correspondence in color to the input image compared to DIFFER.

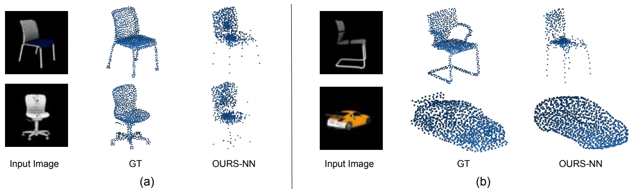

Figure 6: **Failure cases:** (a) Points are clustered with very few points being used for thin structures like the legs of the chair. (b) Details like car spoilers and complex chair legs/handles are not accurately reconstructed.

S.No.	Layer	Filter Size/ Stride	Output Size
Structure Branch			
E$_{s1}$	conv	3x3/2	32x32x32
E$_{s2}$	conv	3x3/2	16x16x64
E$_{s3}$	conv	3x3/2	8x8x128
E$_{s4}$	conv	3x3/2	4x4x256
D$_{s1}$	linear	-	128
D$_{s2}$	linear	-	128
D$_{s3}$	linear	-	128
D$_{s4}$	linear	-	1024*3
Color Branch			
E$_{c1}$	conv	3x3/2	32x32x32
E$_{c2}$	conv	3x3/2	16x16x64
D$_{c1}$	linear	-	128
D$_{c2}$	linear	-	128
D$_{c3}$	linear	-	128
D$_{c3}$	concat(D$_{s3}, D_{c3}$)	-	256
D$_{c4}$	linear	-	128
D$_{c4}$	linear	-	1024*3

Table 6: **Reconstruction Network Architecture:** We use dual branch network architecture for regressing point locations and color as it is shown to be highly effective [1]

S.No.	Layer	Filter Size/ Stride	Output Size
E$_{s1}$	conv	3x3/2	32x32x32
E$_{s2}$	conv	3x3/2	16x16x64
E$_{s3}$	conv	3x3/2	8x8x128
E$_{s4}$	conv	3x3/2	4x4x256
D$_{s1}$	linear	-	128
D$_{s2}$	linear	-	128
D$_{s3}$	linear	-	128
D$_{s4}$	linear	-	2

Table 7: **Pose Network Architecture:** We use an architecture similar to reconstruction network except for the output layer. In the pose prediction network, two values corresponding to azimuth and elevation parameters of the camera are regressed.