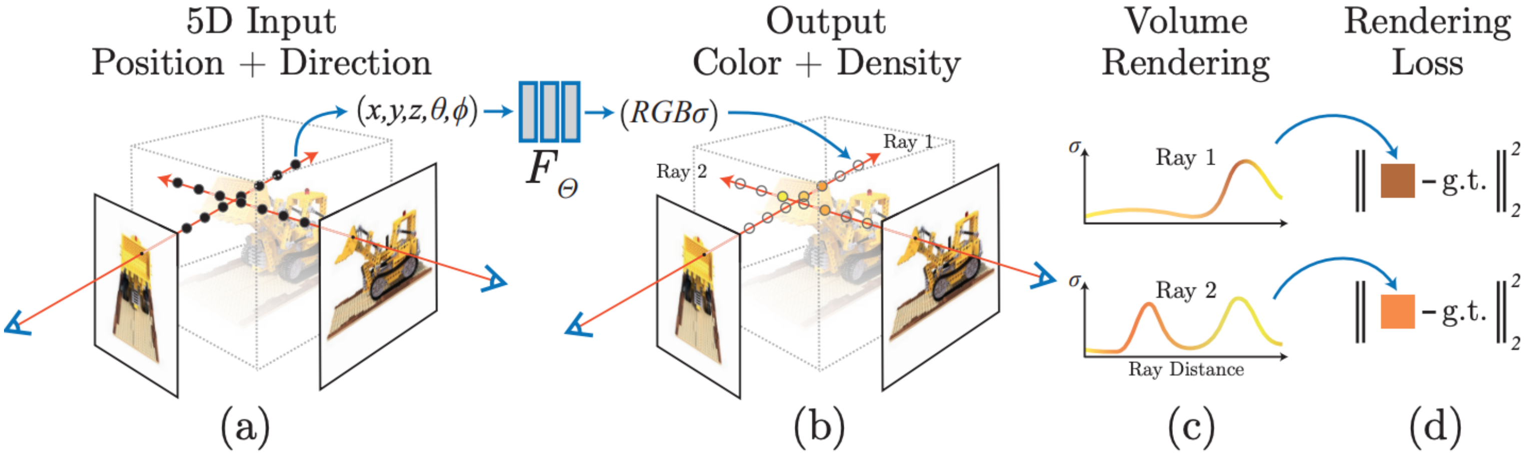
pixelNeRF: Neural Radiance Fields from One or Few Images

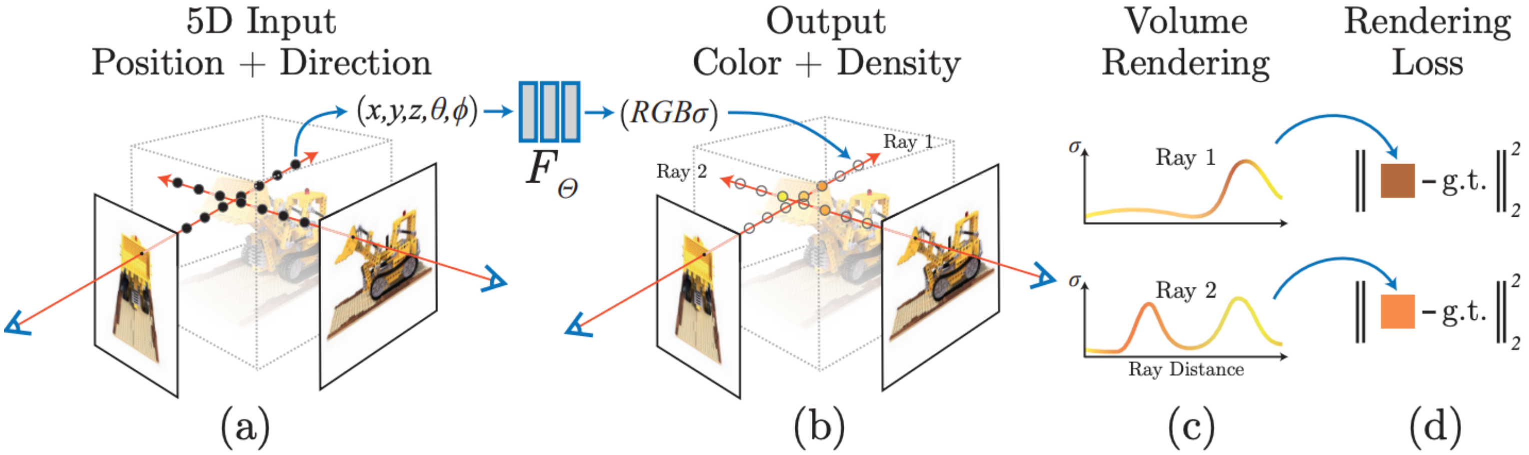
Yu et al.

Background Knowledge

1. NeRF: (Neural Radiance Field)



1. **NeRF: (Neural Radiance Field)**
   a. Map (x,y,z, angle) to pixel intensity and color
   b. Use classical differentiable rendering to compute loss
   c. Coarse + Fine network for importance sampling

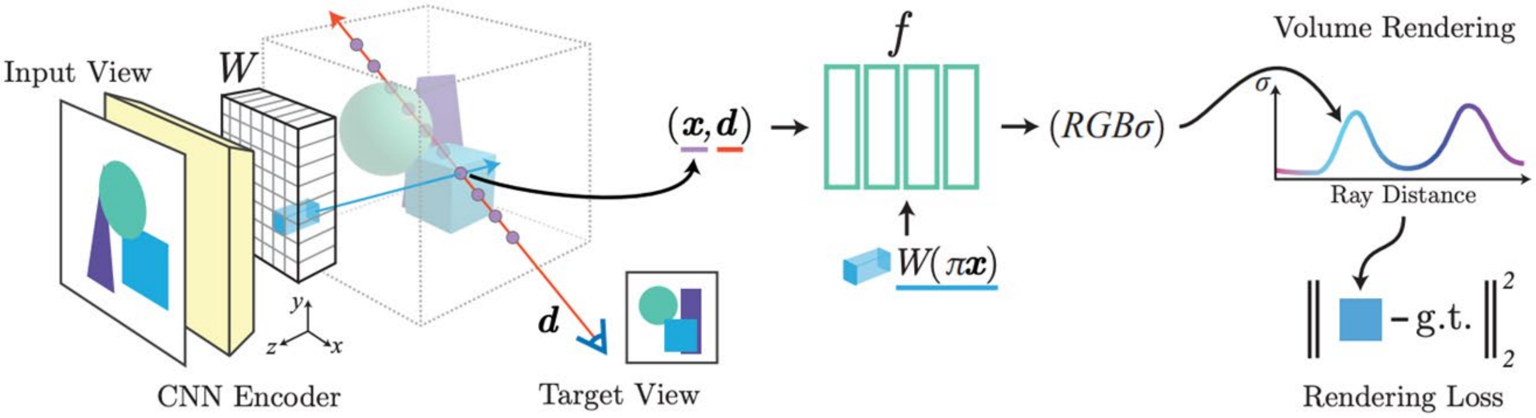


New Problems

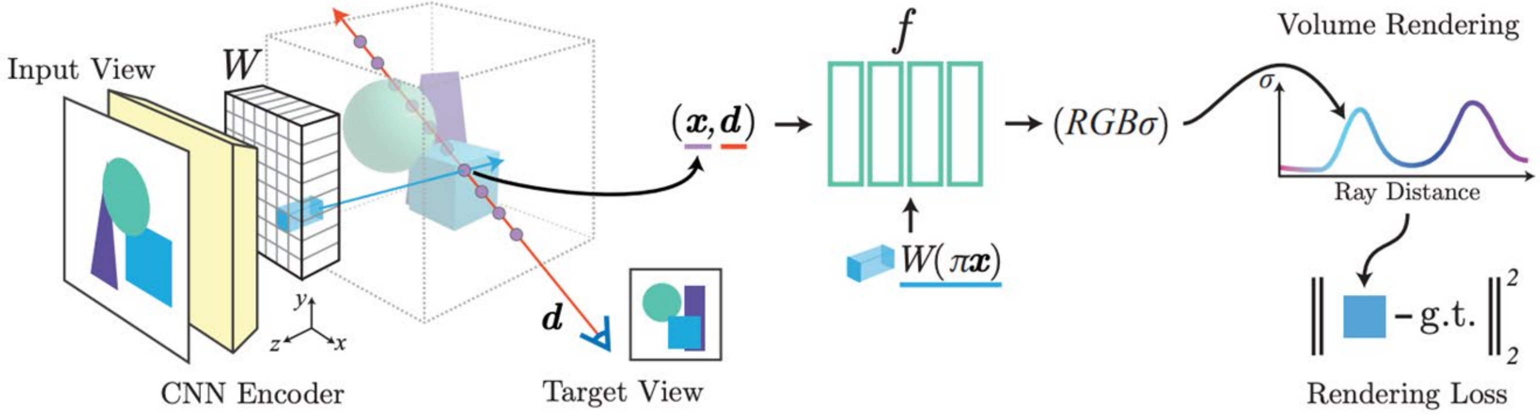
What if the number of projections is very small (i.e. fewer than 5), can we still construct the 3D representation as in NeRF using only those projections?

If we have a large number of projections not related to the current objects that we are interested in, can we utilize those projections?

1. **NeRF**
2. **PixelNeRF: (Neural Radiance Field)**
    a. Add an additional prior image embedding input to NeRF, keeping everything else the same



1. **NeRF**
2. **PixelNeRF: (Neural Radiance Field)**
    1. Add an additional prior image embedding input
    2. Project the embedding from camera plane to canonical plane



PixelNeRF Details

1. Instead of use that embedding as input, add this as a residual to the output of the initial output
2. Prior image embedding taken from pretrained networks, i.e. ImageNet (not trained simultaneously).
3. If multiple prior images
    a. Construct multiple NeRF networks and share weights
    b. Take the average from the outputs of multiple networks

PixelNeRF Details

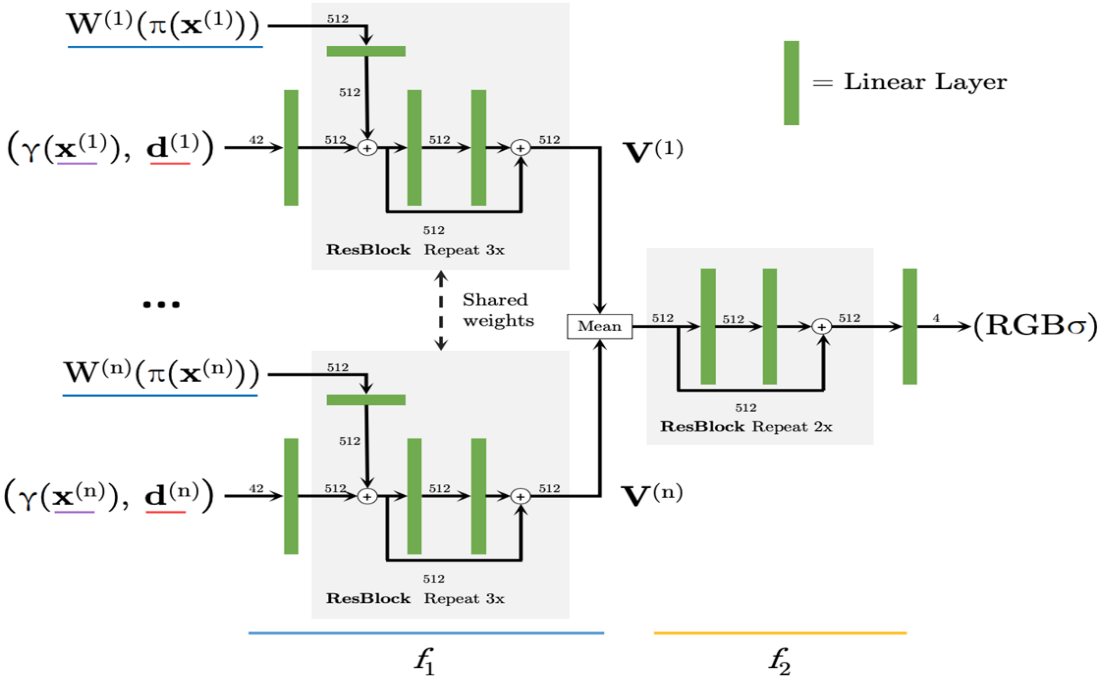
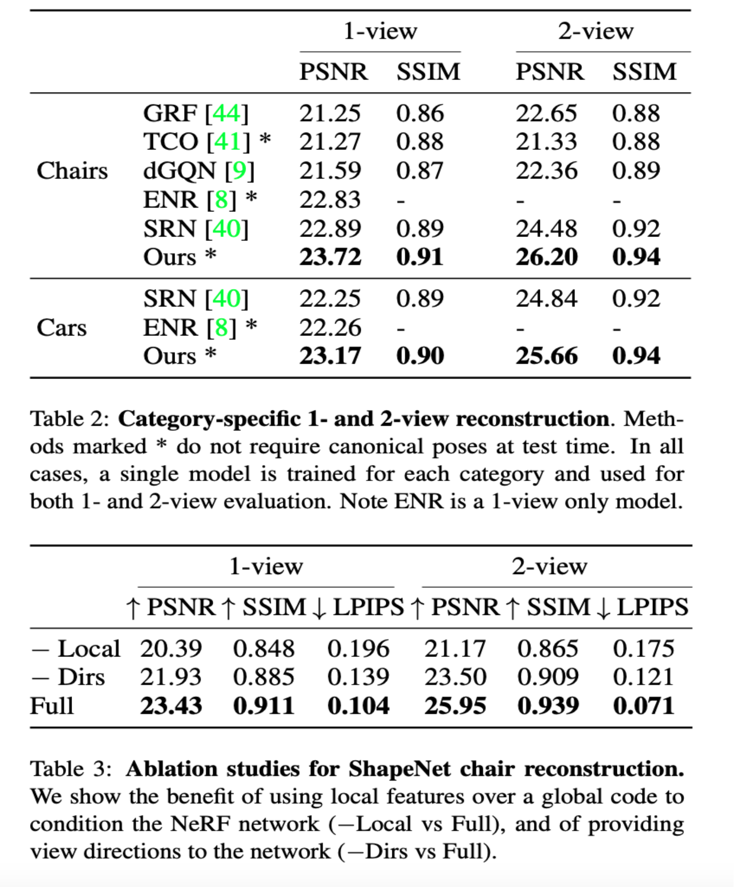


Figure 18: **Multi-view NeRF Network Architecture.** We use notation established in § 5.1.2 of the main paper, where γ denotes a positional encoding with 6 exponentially increasing frequencies. Each linear layer is followed by a ReLU activation. Note that in the single-view case, f_1 and f_2 can be considered a single ResNet $f = f_2 \circ f_1$.

Results



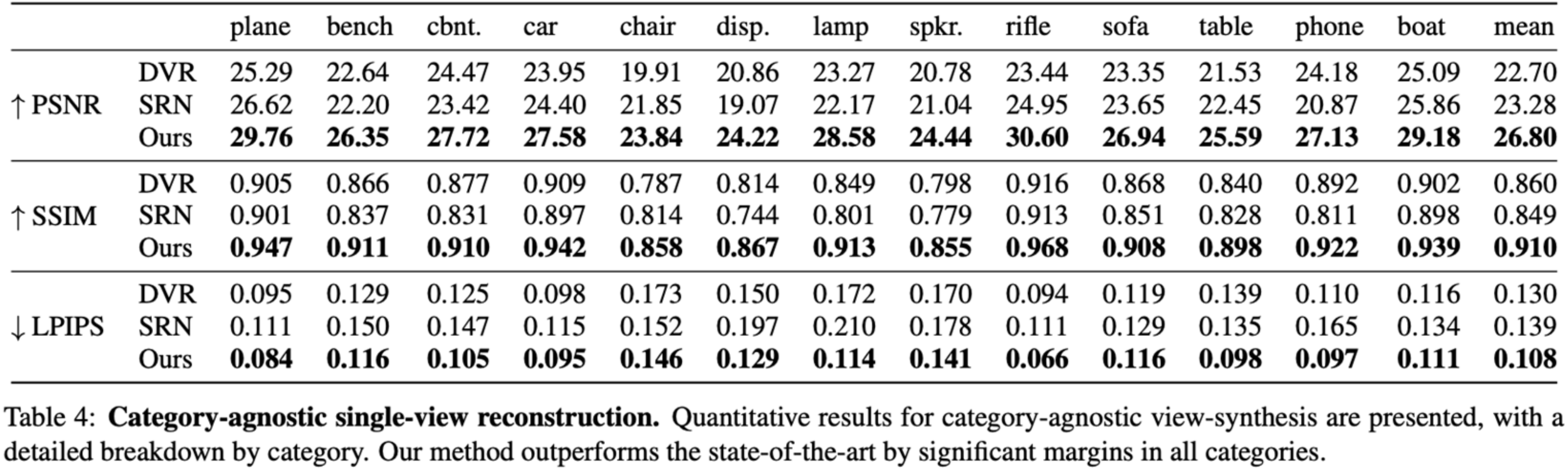
		1-view		2-view	
		PSNR	SSIM	PSNR	SSIM
Chairs	GRF [44]	21.25	0.86	22.65	0.88
	TCO [41] *	21.27	0.88	21.33	0.88
	dGQN [9]	21.59	0.87	22.36	0.89
	ENR [8] *	22.83	-	-	-
	SRN [40]	22.89	0.89	24.48	0.92
	Ours *	**23.72**	**0.91**	**26.20**	**0.94**
Cars	SRN [40]	22.25	0.89	24.84	0.92
	ENR [8] *	22.26	-	-	-
	Ours *	**23.17**	**0.90**	**25.66**	**0.94**

Table 2: **Category-specific 1- and 2-view reconstruction**. Methods marked * do not require canonical poses at test time. In all cases, a single model is trained for each category and used for both 1- and 2-view evaluation. Note ENR is a 1-view only model.

	1-view			2-view		
	↑ PSNR	↑ SSIM	↓ LPIPS	↑ PSNR	↑ SSIM	↓ LPIPS
− Local	20.39	0.848	0.196	21.17	0.865	0.175
− Dirs	21.93	0.885	0.139	23.50	0.909	0.121
Full	**23.43**	**0.911**	**0.104**	**25.95**	**0.939**	**0.071**

Table 3: **Ablation studies for ShapeNet chair reconstruction.** We show the benefit of using local features over a global code to condition the NeRF network (−Local vs Full), and of providing view directions to the network (−Dirs vs Full).

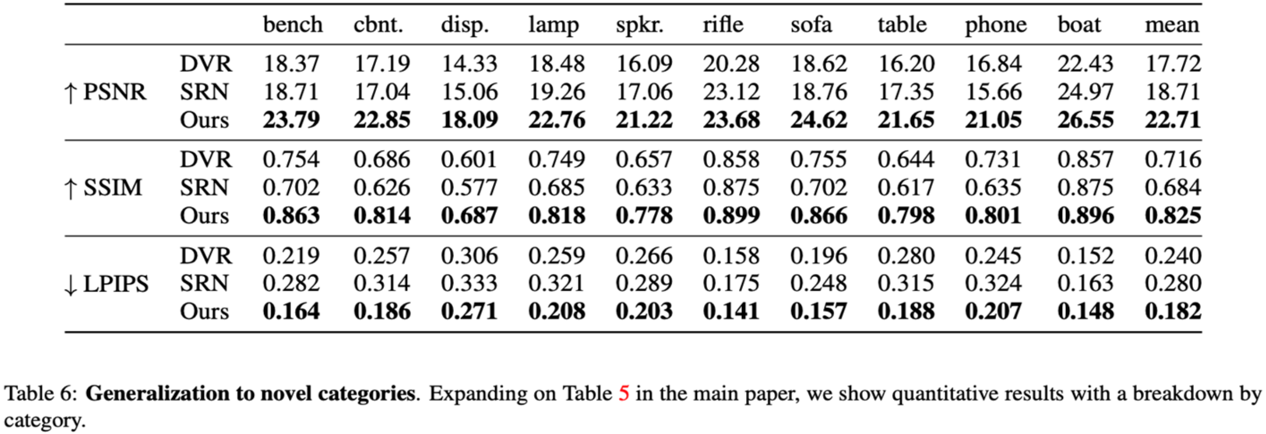
Results



		plane	bench	cbnt.	car	chair	disp.	lamp	spkr.	rifle	sofa	table	phone	boat	mean
↑ PSNR	DVR	25.29	22.64	24.47	23.95	19.91	20.86	23.27	20.78	23.44	23.35	21.53	24.18	25.09	22.70
	SRN	26.62	22.20	23.42	24.40	21.85	19.07	22.17	21.04	24.95	23.65	22.45	20.87	25.86	23.28
	Ours	**29.76**	**26.35**	**27.72**	**27.58**	**23.84**	**24.22**	**28.58**	**24.44**	**30.60**	**26.94**	**25.59**	**27.13**	**29.18**	**26.80**
↑ SSIM	DVR	0.905	0.866	0.877	0.909	0.787	0.814	0.849	0.798	0.916	0.868	0.840	0.892	0.902	0.860
	SRN	0.901	0.837	0.831	0.897	0.814	0.744	0.801	0.779	0.913	0.851	0.828	0.811	0.898	0.849
	Ours	**0.947**	**0.911**	**0.910**	**0.942**	**0.858**	**0.867**	**0.913**	**0.855**	**0.968**	**0.908**	**0.898**	**0.922**	**0.939**	**0.910**
↓ LPIPS	DVR	0.095	0.129	0.125	0.098	0.173	0.150	0.172	0.170	0.094	0.119	0.139	0.110	0.116	0.130
	SRN	0.111	0.150	0.147	0.115	0.152	0.197	0.210	0.178	0.111	0.129	0.135	0.165	0.134	0.139
	Ours	**0.084**	**0.116**	**0.105**	**0.095**	**0.146**	**0.129**	**0.114**	**0.141**	**0.066**	**0.116**	**0.098**	**0.097**	**0.111**	**0.108**

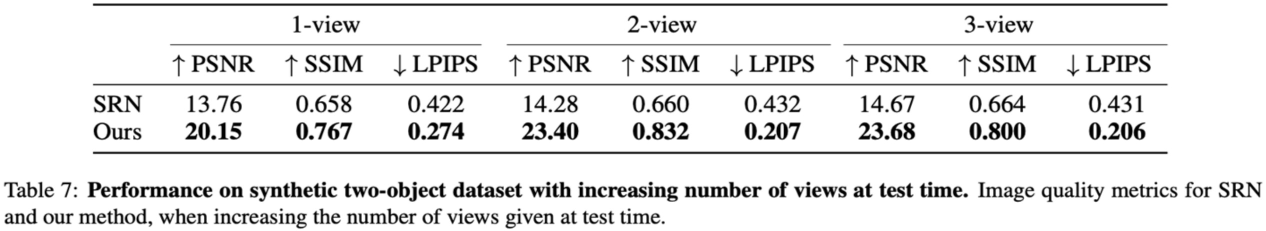
Table 4: **Category-agnostic single-view reconstruction.** Quantitative results for category-agnostic view-synthesis are presented, with a detailed breakdown by category. Our method outperforms the state-of-the-art by significant margins in all categories.

Results



		bench	cbnt.	disp.	lamp	spkr.	rifle	sofa	table	phone	boat	mean
↑ PSNR	DVR	18.37	17.19	14.33	18.48	16.09	20.28	18.62	16.20	16.84	22.43	17.72
	SRN	18.71	17.04	15.06	19.26	17.06	23.12	18.76	17.35	15.66	24.97	18.71
	Ours	**23.79**	**22.85**	**18.09**	**22.76**	**21.22**	**23.68**	**24.62**	**21.65**	**21.05**	**26.55**	**22.71**
↑ SSIM	DVR	0.754	0.686	0.601	0.749	0.657	0.858	0.755	0.644	0.731	0.857	0.716
	SRN	0.702	0.626	0.577	0.685	0.633	0.875	0.702	0.617	0.635	0.875	0.684
	Ours	**0.863**	**0.814**	**0.687**	**0.818**	**0.778**	**0.899**	**0.866**	**0.798**	**0.801**	**0.896**	**0.825**
↓ LPIPS	DVR	0.219	0.257	0.306	0.259	0.266	0.158	0.196	0.280	0.245	0.152	0.240
	SRN	0.282	0.314	0.333	0.321	0.289	0.175	0.248	0.315	0.324	0.163	0.280
	Ours	**0.164**	**0.186**	**0.271**	**0.208**	**0.203**	**0.141**	**0.157**	**0.188**	**0.207**	**0.148**	**0.182**

Table 6: **Generalization to novel categories**. Expanding on Table 5 in the main paper, we show quantitative results with a breakdown by category.



	1-view			2-view			3-view		
	↑ PSNR	↑ SSIM	↓ LPIPS	↑ PSNR	↑ SSIM	↓ LPIPS	↑ PSNR	↑ SSIM	↓ LPIPS
SRN	13.76	0.658	0.422	14.28	0.660	0.432	14.67	0.664	0.431
Ours	**20.15**	**0.767**	**0.274**	**23.40**	**0.832**	**0.207**	**23.68**	**0.800**	**0.206**

Table 7: **Performance on synthetic two-object dataset with increasing number of views at test time.** Image quality metrics for SRN and our method, when increasing the number of views given at test time.

Results

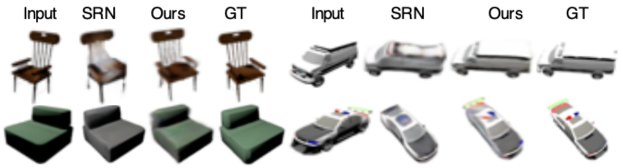


Figure 3: **Category-specific single-view reconstruction benchmark**. We train a separate model for cars and chairs and compare to SRN. The corresponding numbers may be found in Table 2.

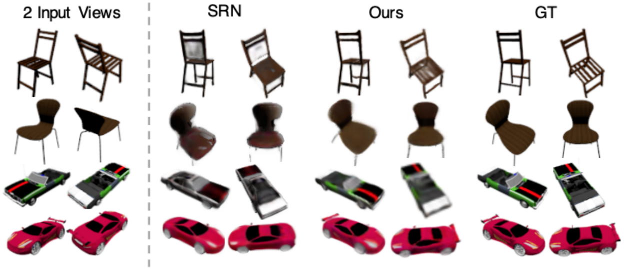


Figure 4: **Category-specific 2-view reconstruction benchmark**. We provide two views (left) to each model, and show two novel view renderings in each case (right). Please also refer to Table 2.

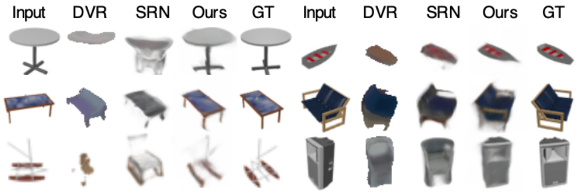


Figure 6: **Generalization to unseen categories.** We evaluate a model trained on planes, cars, and chairs on 10 unseen ShapeNet categories. We find that the model is able to synthesize reasonable views even in this difficult case.

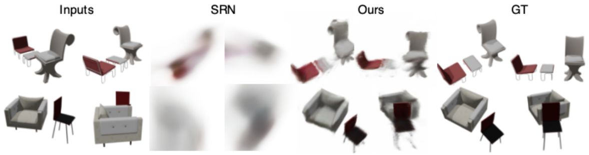


Figure 7: **360° view prediction with multiple objects.** We show qualitative results of our method compared with SRN on scenes composed of multiple ShapeNet chairs. We are easily able to handle this setting, because our prediction is done in view space; in contrast, SRN predicts in canonical space, and struggles with scenes that cannot be aligned in such a way.

References:

1. Mildenhall, Ben, et al. "Nerf: Representing scenes as neural radiance fields for view synthesis." *European conference on computer vision*. Springer, Cham, 2020.
2. Yu, Alex, et al. "pixelnerf: Neural radiance fields from one or few images." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.