Coordinates are not lonely - Codebook Prior Helps Implicit Neural 3D Representations Supplementary Materials

Anonymous Author(s)
Affiliation
Address
email

1 Novel View Synthesis Videos

We visualize some of our experiments in a video, including spares views and few views setting on the DTU [1], BlendedMVS [5] and H3DS [4] dataset (see CoCo-INR_Sparse_Views_Setting.mp4 and CoCo-INR_Few_Views_Setting.mp4). Specifically, two opposite views (i.e., front and behind) are selected as the start and end view, then an arc-shaped sequence of novel views with even adjacent view degree interval are then synthesized between the start and end views, constituting a total number of 60 views covering 180 degrees variation. Note that the DTU dataset is not selected from 360-degree surrounding views so it is not suitable for the above view generation method. For DTU dataset, we chose the leftmost and rightmost views to generate a arc-shaped view sequence between them and use masks to reduce the background ambiguity caused by the non-surround setting.

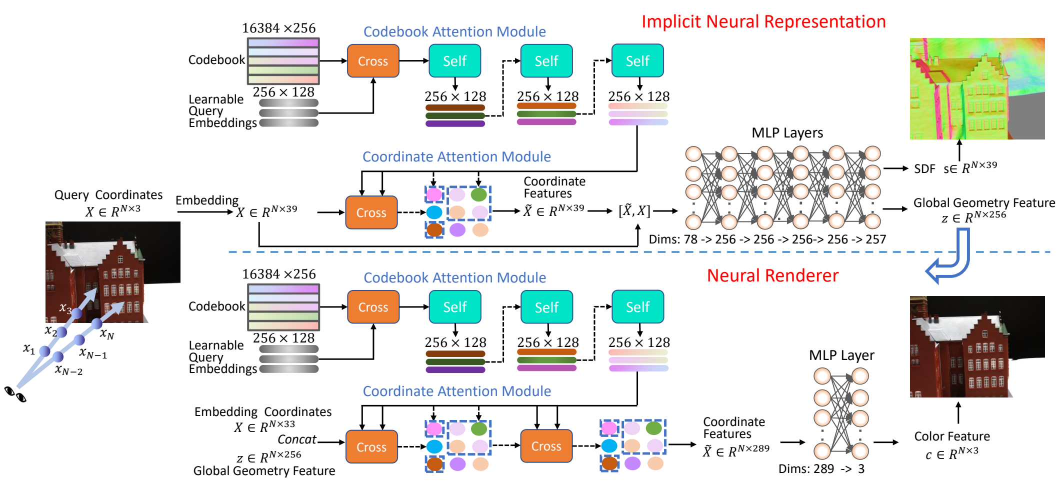

Figure 1: Pipeline of our method.

2 Details of Our CoCo-INR

We now present the pipeline and details of our proposed CoCo-INR. As shown in Fig. 1, it contains two modules: implicit neural representation and neural renderer, each of which consists of several codebook attention and coordinate attention modules and MLP layers. The structure of the codebook

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

attention and coordinate attention module has been described in our paper. Here is the workflow of our method.

3 Additional Experiments

In our paper, we have comprehensively evaluated the performance of our method and other state-of-the-art methods on DTU [1], BlendedMVS [5], and H3DS [4] datasets under different numbers of training views. In this section, we will present detailed data for qualitative comparisons, challenge fewer training views, and analyze limitations.

3.1 Detailed Quantitative Results

Qualitative comparisons with NeRF [2], UNISURF [3], and VolSDF [6] have been shown in paper. Our CoCo-based method outperforms others in terms of PSNR, SSIM, and LPIPS whether sparse views or few views are available. This section further presents the evaluation results for each scan on DTU [1], BlendedMVS [5], and H3DS [4] datasets.

For sparse views setting (16-32 in total), Table 1 shows the comparison of each scan on the DTU dataset, and Table 2 shows the comparison of each scan on the BlendedMVS dataset, and Table 3 shows the comparison of each scan on the H3DS dataset. It can be seen that our method outperforms other methods in most scans, which means that our method has good performance and robustness under sparse views.

Table 1: The results of different methods on the DTU dataset for each scan with sparse views (16-32) and without object masks. ↑ means the higher, the better.

Scan	PSNR↑				SSIM↑				LPIPS↓			
	NeRF	UNISURF	VolSDF	Ours	NeRF	UNISURF	VolSDF	Ours	NeRF	UNISURF	VolSDF	Ours
24	10.720	20.060	23.100	**23.896**	0.478	0.704	0.676	**0.791**	0.377	0.277	0.246	**0.227**
37	19.789	20.065	**22.576**	22.022	0.628	0.650	0.607	**0.703**	**0.227**	0.277	0.254	0.230
40	9.424	12.185	21.705	**22.803**	0.364	0.512	**0.593**	0.569	0.338	0.313	0.234	**0.231**
55	20.323	22.264	**24.364**	23.804	0.734	0.783	**0.835**	0.821	0.329	0.379	0.333	**0.318**
63	24.492	24.539	25.378	**26.568**	**0.826**	0.811	0.699	0.705	**0.184**	0.229	0.233	0.200
65	18.161	23.859	**27.904**	26.121	0.780	0.837	**0.870**	0.850	**0.262**	0.310	0.271	0.282
69	23.478	24.427	25.434	**25.935**	0.867	0.882	0.910	**0.915**	0.302	0.339	**0.274**	0.283
83	25.733	25.808	26.165	**27.975**	0.922	0.905	0.925	**0.944**	**0.270**	0.370	0.360	0.391
97	22.119	21.833	23.765	**24.029**	0.854	0.845	0.887	**0.896**	**0.297**	0.392	0.308	0.339
105	24.746	24.240	27.257	**27.577**	0.895	0.888	**0.926**	0.924	**0.268**	0.372	0.363	0.300
106	27.629	26.919	30.394	**30.813**	0.904	0.907	0.932	**0.940**	0.381	0.389	0.345	**0.338**
110	25.353	26.046	28.084	**30.190**	0.916	0.910	0.924	**0.939**	**0.369**	0.426	0.392	**0.369**
114	24.297	24.394	28.039	**28.707**	0.863	0.875	**0.907**	0.902	0.348	0.387	0.340	**0.338**
118	29.082	28.484	**32.218**	31.358	0.919	0.921	**0.948**	0.943	**0.312**	0.377	0.341	0.329
122	27.091	28.125	**32.765**	29.270	0.901	0.921	**0.952**	0.942	0.334	0.387	0.350	**0.294**
Mean	22.162	23.549	26.609	**26.738**	0.790	0.823	0.839	**0.852**	0.306	0.348	0.309	**0.298**

For few views setting (5-8 in total), Table 4 shows the comparison of each scan on the DTU dataset, and Table 5 shows the comparison of each scan on the BlendedMVS dataset, and Table 6 shows the comparison of each scan on the H3DS dataset. VolSDF [6] achieves better performance than other state-of-the-art methods in the sparse views setting, so we only compare with VolSDF in this more difficult setting. It can be seen that in each scan, using fewer training views will inevitably lead to performance degradation compared with sparse views, but our method still maintains better robustness and performance for the majority of scans.

Table 2: The results of different methods on the BlendedMVS dataset for each scan with sparse views (16-32) and without object masks. ↑ means the higher, the better.

Scan	PSNR↑				SSIM↑				LPIPS↓			
	NeRF	UNISURF	VolSDF	Ours	NeRF	UNISURF	VolSDF	Ours	NeRF	UNISURF	VolSDF	Ours
1	19.179	15.586	20.004	**20.038**	0.675	0.626	**0.725**	**0.725**	0.256	0.295	0.215	**0.211**
2	14.220	14.848	19.297	**20.951**	0.630	0.651	0.757	**0.810**	0.320	0.351	0.234	**0.203**
3	16.684	14.717	16.808	**17.263**	0.647	0.600	0.671	**0.682**	0.293	0.336	0.268	**0.265**
4	**23.684**	15.200	23.214	23.007	0.838	0.755	0.847	**0.854**	0.170	0.258	0.154	**0.148**
5	17.160	14.361	16.744	**17.528**	0.745	0.700	0.758	**0.784**	0.244	0.304	0.215	**0.207**
6	15.575	13.739	19.775	**21.741**	0.670	0.630	0.788	**0.819**	0.282	0.304	0.175	**0.148**
7	8.722	12.086	17.972	**17.997**	0.528	0.610	**0.721**	**0.721**	0.349	0.333	0.237	**0.223**
8	17.851	17.187	18.924	**19.827**	0.743	0.737	0.821	**0.837**	0.183	0.171	0.140	**0.128**
9	15.634	15.019	17.747	**17.997**	0.535	0.540	0.639	**0.640**	0.356	0.346	0.284	**0.279**
Mean	16.523	14.749	18.942	**19.594**	0.667	0.649	0.747	**0.764**	0.272	0.299	0.213	**0.201**

Table 3: The results of different methods on the H3DS dataset for each scan with sparse views (16-32) and without object masks. ↑ means the higher, the better.

Scan	PSNR↑				SSIM↑				LPIPS↓			
	NeRF	UNISURF	VolSDF	Ours	NeRF	UNISURF	VolSDF	Ours	NeRF	UNISURF	VolSDF	Ours
*a287	21.826	14.314	25.283	**26.070**	0.896	0.801	**0.941**	**0.941**	0.121	0.198	0.082	**0.076**
*42a8	19.783	20.170	22.437	**23.785**	0.825	0.832	0.863	**0.887**	0.179	0.202	0.138	**0.121**
*1d54	23.779	18.028	25.420	**25.933**	0.878	0.810	0.893	**0.904**	0.130	0.184	0.114	**0.107**
*0854	20.302	20.236	22.819	**24.274**	0.840	0.835	0.865	**0.875**	0.151	0.151	0.127	**0.124**
*0c89	22.241	21.371	23.658	**25.116**	0.911	0.898	0.932	**0.942**	0.117	0.122	0.087	**0.084**
*0226	22.207	12.104	25.582	**25.742**	0.861	0.762	**0.907**	**0.907**	0.130	0.246	**0.088**	**0.088**
*4baa	21.458	15.378	24.949	**26.755**	0.884	0.820	0.909	**0.931**	0.139	0.187	0.106	**0.094**
*512f	21.009	20.250	24.217	**25.141**	0.830	0.826	0.878	**0.895**	0.183	0.223	0.141	**0.127**
*3924	21.108	18.979	23.774	**26.090**	0.865	0.832	0.897	**0.918**	0.133	0.201	0.112	**0.091**
*e5bc	23.274	12.605	24.051	**24.395**	0.908	0.768	0.918	**0.926**	0.099	0.210	0.084	**0.079**
*ee0b	21.128	12.636	22.629	**24.136**	0.886	0.763	0.911	**0.921**	0.131	0.217	0.100	**0.091**
*e187	24.039	20.294	25.256	**26.738**	0.893	0.846	0.913	**0.923**	0.114	0.146	0.092	**0.084**
*d436	19.579	18.488	21.114	**21.356**	0.817	0.799	0.847	**0.851**	0.168	0.237	0.160	**0.136**
*b9c0	21.514	20.944	23.069	**27.221**	0.854	0.855	0.877	**0.916**	0.153	0.182	0.129	**0.101**
*9c4e	21.905	17.540	25.714	**27.557**	0.911	0.881	0.943	**0.950**	0.110	0.144	0.080	**0.073**
*7be3	20.288	14.227	20.190	**22.983**	0.835	0.787	0.843	**0.866**	0.138	0.198	0.138	**0.106**
*fd85	22.270	9.058	24.448	**25.250**	0.873	0.658	0.900	**0.910**	0.140	0.296	0.114	**0.102**
*c1e6	21.110	15.816	25.507	**26.204**	0.889	0.832	0.932	**0.938**	0.132	0.193	0.090	**0.082**
*4b87	21.769	22.298	24.960	**25.244**	0.897	0.901	0.931	**0.934**	0.126	0.116	0.094	**0.092**
*2a8f	11.296	15.006	24.133	**25.439**	0.697	0.821	0.928	**0.931**	0.285	0.180	0.094	**0.093**
*244e	22.966	14.668	23.207	**25.295**	0.833	0.758	0.853	**0.875**	0.155	0.226	0.123	**0.106**
*deb0	22.809	20.577	25.772	**28.006**	0.898	0.862	0.931	**0.942**	0.139	0.141	0.114	**0.097**
*2091	19.607	18.199	22.038	**22.685**	0.839	0.838	0.864	**0.871**	0.159	0.176	0.134	**0.122**
Mean	21.185	17.095	23.922	**25.279**	0.861	0.816	0.898	**0.911**	0.144	0.190	0.110	**0.098**

Table 4: The results of different methods on the DTU dataset for each scan with few views (5-8) and without object masks. ↑ means the higher, the better.

Scan	PSNR↑		SSIM↑		LPIPS↓	
	VolSDF	**Ours**	VolSDF	**Ours**	VolSDF	**Ours**
24	17.476	**17.867**	0.544	**0.617**	**0.274**	0.285
37	10.783	**15.372**	0.554	**0.615**	0.310	**0.271**
40	**14.489**	13.627	**0.574**	0.496	**0.286**	0.294
55	14.840	**18.014**	0.641	**0.720**	0.393	**0.382**
63	7.661	**18.391**	0.482	**0.572**	0.361	**0.226**
65	19.173	**19.265**	0.769	**0.771**	0.319	**0.307**
69	**16.769**	16.268	**0.720**	0.705	0.387	**0.379**
83	**18.746**	17.119	**0.866**	0.864	**0.375**	0.391
97	18.321	**18.821**	**0.827**	**0.827**	0.382	**0.378**
105	**18.914**	18.165	**0.873**	0.861	0.360	**0.344**
106	**23.132**	23.103	**0.845**	0.836	**0.377**	0.407
110	21.763	**25.954**	0.843	**0.900**	0.437	**0.436**
114	**24.510**	22.408	**0.861**	0.834	0.372	**0.352**
118	**28.324**	26.602	**0.913**	0.905	0.397	**0.368**
122	**25.488**	23.385	**0.901**	0.836	0.400	**0.372**
Mean	18.693	**19.624**	0.748	**0.757**	0.362	**0.346**

Table 5: The results of different methods on the BlendedMVS dataset for each scan with few views (5-8) and without object masks. ↑ means the higher, the better.

Scan	PSNR↑		SSIM↑		LPIPS↓	
	VolSDF	**Ours**	VolSDF	**Ours**	VolSDF	**Ours**
1	13.313	**16.597**	0.562	**0.630**	0.369	**0.265**
2	15.037	**15.277**	0.635	**0.637**	0.311	**0.303**
3	**12.312**	10.965	**0.556**	0.490	**0.358**	0.398
4	14.132	**14.614**	0.722	**0.735**	0.278	**0.254**
5	13.781	**14.358**	0.675	**0.699**	0.299	**0.268**
6	11.662	**12.114**	**0.579**	0.575	0.375	**0.366**
7	14.708	**15.955**	0.647	**0.659**	0.281	**0.269**
8	**17.438**	16.384	**0.760**	0.759	**0.171**	0.178
9	12.876	**13.491**	0.478	**0.487**	0.415	**0.386**
Mean	13.918	**14.417**	0.624	**0.630**	0.317	**0.299**

Table 6: The results of different methods on the H3DS dataset for each scan with few views (5-8) and without object masks. ↑ means the higher, the better.

Scan	PSNR↑		SSIM↑		LPIPS↓	
	VolSDF	**Ours**	VolSDF	**Ours**	VolSDF	**Ours**
*a287	18.014	**18.804**	0.864	**0.878**	0.152	**0.136**
*42a8	18.207	**18.538**	**0.811**	**0.811**	0.202	**0.195**
*1d54	**22.271**	16.674	**0.872**	0.761	**0.139**	0.202
*0854	18.837	**19.950**	0.809	**0.827**	0.187	**0.171**
*0c89	**20.037**	14.813	**0.892**	0.779	**0.129**	0.226
*0226	20.999	**21.706**	0.863	**0.869**	**0.132**	0.134
*4baa	18.800	**20.792**	0.847	**0.875**	0.174	**0.148**
*512f	19.587	**19.780**	0.829	**0.832**	**0.190**	0.200
*3924	**19.950**	19.362	**0.857**	0.851	**0.155**	0.157
*e5bc	**22.159**	20.907	**0.913**	0.900	**0.113**	0.120
*ee0b	15.451	**16.959**	0.817	**0.841**	0.209	**0.187**
*e187	20.125	**21.774**	0.858	**0.891**	0.154	**0.128**
*d436	16.528	**17.806**	0.780	**0.794**	0.211	**0.193**
*b9c0	14.505	**19.288**	0.783	**0.843**	0.230	**0.188**
*9c4e	19.285	**20.137**	**0.894**	**0.894**	**0.121**	0.123
*7be3	19.439	**20.038**	0.825	**0.864**	**0.151**	**0.151**
*fd85	**19.874**	19.391	**0.858**	0.850	**0.168**	**0.168**
*c1e6	20.311	**20.470**	0.891	**0.894**	0.135	**0.122**
*4b87	**20.192**	19.527	**0.890**	0.880	0.136	**0.135**
*2a8f	**17.871**	16.654	**0.859**	0.831	**0.173**	0.196
*244e	18.218	**18.508**	0.797	**0.825**	**0.179**	0.182
*deb0	**20.181**	19.616	**0.865**	**0.865**	**0.152**	0.171
*2091	14.767	**17.450**	0.789	**0.816**	0.220	**0.170**
Mean	18.939	**19.085**	0.846	**0.847**	0.166	**0.165**

3.2 Setting with Fewer Views

In this section we attempt a challenge: only use three training views to train our CoCo-based network. Specifically, we select three representative views for each scan as training views (usually left, right, and top views), and the remaining views are used as test views to evaluate the performance of our method under extremely limited views. In this challenge, three training views mean that it is difficult for the network to learn background information, so we use masks to make the network more focus on the foreground. We conduct experiments on the DTU dataset and compare with mask-based VolSDF [6].

As shown in Table 7, our method has significantly improved over VolSDF with a higher mean PSNR, SSIM, and LPIPS respectively. Especially for the LPIPS, VolSDF is nearly 25% higher than our method, which means that the new view images generated by our method perform better at the semantic level. We visualize part of the experimental results in Fig. 2, and it can be seen that our CoCo-INR can still render satisfactory RGB images and surface features under the extremely limited 3 training views.

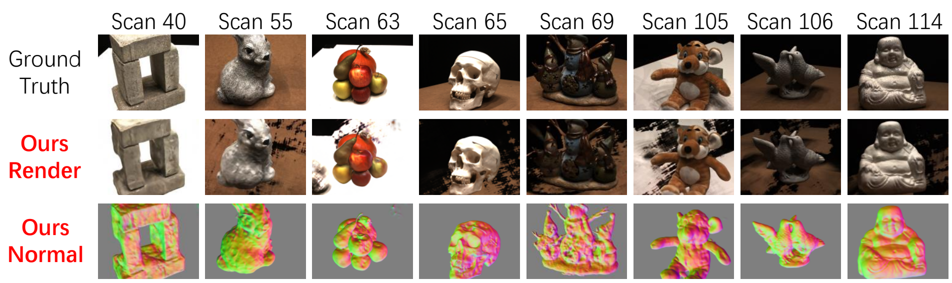

Figure 2: Qualitative visualization results (zoom-in for the best of views) on the DTU dataset with 3 extremely limited views.

Table 7: The results of different methods on the DTU dataset for each scene with 3 training views and objects masks. ↑ means the higher, the better.

Scan	PSNR↑		SSIM↑		LPIPS↓	
	VolSDF	**Ours**	VolSDF	**Ours**	VolSDF	**Ours**
24	**12.118**	11.266	**0.518**	0.443	0.352	**0.332**
37	9.840	**10.413**	**0.464**	0.387	0.337	**0.334**
40	**10.743**	9.963	**0.443**	0.372	**0.309**	0.317
55	14.709	**14.743**	**0.630**	0.619	**0.401**	0.414
63	**11.516**	11.179	**0.428**	0.374	0.294	**0.276**
65	15.411	**15.762**	0.674	**0.689**	0.337	**0.324**
69	**17.798**	17.472	**0.771**	0.761	**0.364**	0.375
83	10.316	**10.681**	0.675	**0.715**	0.430	**0.405**
97	11.936	**12.302**	0.707	**0.714**	0.419	**0.412**
105	**11.296**	11.261	0.716	**0.727**	0.410	**0.405**
106	17.311	**19.219**	0.513	**0.746**	0.713	**0.427**
110	**16.703**	16.367	**0.497**	**0.497**	0.769	**0.730**
114	16.763	**19.308**	0.524	**0.754**	0.672	**0.400**
118	15.928	**20.354**	0.281	**0.770**	0.780	**0.396**
122	16.549	**21.122**	0.445	**0.799**	0.789	**0.407**
Mean	13.929	**14.761**	0.552	**0.624**	0.491	**0.396**

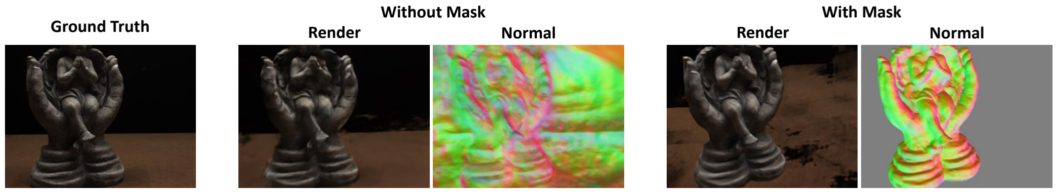

Figure 3: A failure case of our method with extremely limited views.

3.3 Limitations

For settings where the number of training views is extremely limited, it is possible that the limited views can not cover the entire background. So it is difficult to distinguish objects and background when they have similar color, as shown in Fig. 3. To improve the robustness of our method, for the setting of three training views, we have to use extra masks to reduce the background interference. In fact, we can fuse unsupervised methods and implicit neural representations to learn masks to reduce the manual annotation. Therefore, we will study self-supervised multi-task assisted CoCo-based methods to further improve the effectiveness under restricted conditions. Moreover, if we replace the per-scene learnable query embeddings in codebook attention modules with the image/pose-dependent feature tokens, our framework could be extended to a generalizable network across scenes rather than the current per-scene optimization.

References

[1] Rasmus Jensen, Anders Dahl, George Vogiatzis, Engin Tola, and Henrik Aanæs. Large scale multi-view stereopsis evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 406–413, 2014.

[2] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European conference on computer vision*, pages 405–421. Springer, 2020.

[3] Michael Oechsle, Songyou Peng, and Andreas Geiger. Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5589–5599, 2021.

[4] Eduard Ramon, Gil Triginer, Janna Escur, Albert Pumarola, Jaime Garcia, Xavier Giro-i Nieto, and Francesc Moreno-Noguer. H3d-net: Few-shot high-fidelity 3d head reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5620–5629, 2021.

[5] Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1790–1799, 2020.

[6] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. *Advances in Neural Information Processing Systems*, 34, 2021.