Supplementary Material:
Neural 3D Scene Reconstruction with the Manhattan-world Assumption

Haoyu Guo[1*] Sida Peng[1*] Haotong Lin[1] Qianqian Wang[2]
Guofeng Zhang[1] Hujun Bao[1] Xiaowei Zhou[1†]
[1] Zhejiang University [2] Cornell University

1. Network architecture

The architecture of our network is illustrated in Figure 1.

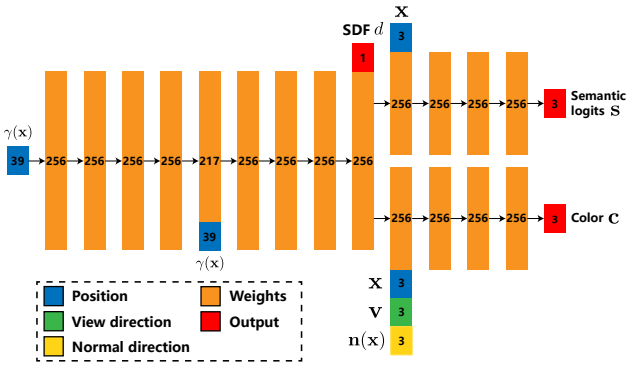

Figure 1. **Network architecture.** Our network takes as inputs the spatial point $\mathbf{x}$, the view direction $\mathbf{v}$, and outputs SDF d, color $\mathbf{c}$, semantic logits $\mathbf{s}$.

2. Evaluation metrics

The definitions of 3D reconstruction metrics are shown in Table 1.

Metric	Definition
Acc	$\text{mean}_{p\in P}(\min_{p^*\in P^*} \|\|p-p^*\|\|)$
Comp	$\text{mean}_{p^*\in P^*}(\min_{p\in P} \|\|p-p^*\|\|)$
Prec	$\text{mean}_{p\in P}(\min_{p^*\in P^*} \|\|p-p^*\|\| < .05)$
Recal	$\text{mean}_{p^*\in P^*}(\min_{p\in P} \|\|p-p^*\|\| < .05)$
F-score	$\frac{2\times\text{Perc}\times\text{Recal}}{\text{Prec}+\text{Recal}}$

Table 1. **Metric definitions.** P and P^* are the point clouds sampled from predicted and ground truth mesh.

3. Quantitative results on individual scenes

We conduct reconstruction experiments on 8 randomly select scenes from ScanNet and 7-Scenes datasets, and compare our method with state-of-the-art MVS and implicit neural representations based methods. We show quantitative results on each individual scene in Table 2.

4. Details of scene parameters

As mentioned in the main paper, we set $\mathbf{n}_f$ as $\langle 0,0,1\rangle$ and fix it during training. In practice, capturing RGB images using mobile platforms with a gravity sensor (e.g. Apple ARKit and Android ARCore) can easily ensure that the estimated camera poses are aligned in z-axis with real world coordinate. However, it is difficult to ensure the normal directions in wall regions to be aligned with x/y-axis, that is the reason why we optimize $\mathbf{n}_w$ together with network parameters during training.

In our experiments, we found that $\mathbf{n}_w$ can converge well. Since GT mesh with semantic label can be obtained on ScanNet dataset, we cluster the normal directions of wall regions using Mean shift algorithm and get 4 clustering centers $\{\mathbf{n}_i\}, i \in \{1,2,3,4\}$ for each scene. To evaluate how well $\mathbf{n}_w$ converges, we define cost function as:

$$L_{\text{normal}}(\mathbf{n}_w) = \frac{1}{4}\sum_{i=1}^{4} \min_{j\in\{-1,0,1\}} |j - \mathbf{n}_w \cdot \mathbf{n}_i| \,. \quad (1)$$

The curves of cost function $L_{\text{normal}}(\mathbf{n}_w)$ are ploted in Figure 2.

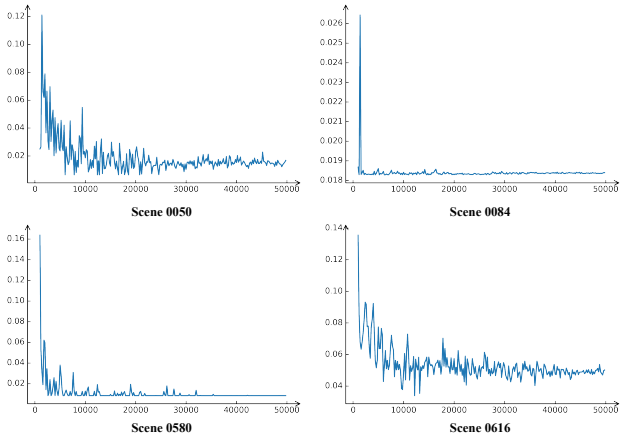

Figure 2. **Convergence of $\mathbf{n}_w$.** x-axis represents training iteration number, y-axis represents the value of cost function in Equation (1).

	Method	scene 0050					scene 0084				
		Acc↓	Comp↓	Prec↑	Recall↑	**F-score↑**	Acc↓	Comp↓	Prec↑	Recall↑	**F-score↑**
ScanNet	COLMAP	0.059	0.174	0.659	0.491	0.563	0.042	0.134	0.736	0.552	0.631
	COLMAP*	0.511	0.070	0.222	0.587	0.322	0.239	0.052	0.430	0.702	0.533
	ACMP	0.123	0.097	0.560	0.594	0.577	0.122	0.060	0.567	0.652	0.607
	NeRF	0.855	0.089	0.146	0.476	0.224	0.908	0.249	0.107	0.197	0.139
	UNISURF	0.485	0.102	0.258	0.432	0.323	0.638	0.247	0.189	0.326	0.239
	NeuS	0.130	0.115	0.441	0.406	0.423	0.255	0.360	0.128	0.084	0.101
	VolSDF	0.092	0.079	0.512	0.544	0.527	0.551	0.162	0.127	0.232	0.164
	Ours	0.058	0.059	0.707	0.642	0.673	0.055	0.053	0.639	0.621	0.630
	Method	scene 0580					scene 0616				
		Acc↓	Comp↓	Prec↑	Recall↑	**F-score↑**	Acc↓	Comp↓	Prec↑	Recall↑	**F-score↑**
	COLMAP	0.034	0.176	0.809	0.465	0.590	0.054	0.457	0.638	0.256	0.365
	COLMAP*	0.479	0.054	0.223	0.661	0.333	0.354	0.147	0.211	0.429	0.283
	ACMP	0.088	0.070	0.557	0.594	0.575	0.138	0.095	0.438	0.486	0.460
	NeRF	0.457	0.138	0.099	0.252	0.142	0.718	0.233	0.172	0.237	0.199
	UNISURF	0.376	0.116	0.218	0.399	0.282	0.716	0.193	0.183	0.293	0.225
	NeuS	0.161	0.215	0.413	0.327	0.365	0.171	0.142	0.269	0.284	0.276
	VolSDF	0.091	0.088	0.529	0.540	0.534	0.922	0.150	0.115	0.259	0.160
	Ours	0.104	0.062	0.616	0.650	0.632	0.072	0.098	0.521	0.431	0.472

	Method	Heads					Office				
		Acc↓	Comp↓	Prec↑	Recall↑	**F-score↑**	Acc↓	Comp↓	Prec↑	Recall↑	**F-score↑**
7-Scenes	COLMAP	0.036	0.284	0.815	0.206	0.329	0.080	0.247	0.408	0.195	0.264
	COLMAP*	0.921	0.299	0.110	0.137	0.122	0.769	0.143	0.125	0.251	0.167
	ACMP	0.099	0.220	0.427	0.259	0.322	0.261	0.180	0.259	0.203	0.228
	NeRF	0.144	0.323	0.294	0.047	0.081	0.669	0.341	0.185	0.079	0.111
	UNISURF	0.129	0.117	0.354	0.419	0.384	0.534	0.145	0.127	0.248	0.168
	NeuS	0.091	0.204	0.520	0.328	0.402	0.174	0.242	0.203	0.135	0.162
	VolSDF	0.136	0.097	0.396	0.417	0.406	0.412	0.147	0.139	0.195	0.162
	Ours	0.120	0.086	0.407	0.428	0.417	0.116	0.165	0.254	0.172	0.205
	Method	Chess					Fire				
		Acc↓	Comp↓	Prec↑	Recall↑	**F-score↑**	Acc↓	Comp↓	Prec↑	Recall↑	**F-score↑**
	COLMAP	0.112	0.772	0.271	0.119	0.165	0.047	0.365	0.652	0.287	0.399
	COLMAP*	0.373	0.197	0.117	0.180	0.142	0.616	0.220	0.114	0.293	0.164
	ACMP	0.747	0.194	0.158	0.211	0.181	0.064	0.181	0.555	0.402	0.466
	NeRF	0.502	0.405	0.118	0.061	0.081	0.979	0.216	0.038	0.153	0.061
	UNISURF	0.285	0.160	0.208	0.281	0.239	0.682	0.121	0.092	0.256	0.136
	NeuS	0.206	0.404	0.199	0.136	0.162	0.134	0.139	0.330	0.316	0.323
	VolSDF	0.364	0.180	0.148	0.232	0.181	0.229	0.135	0.197	0.296	0.236
	Ours	0.129	0.214	0.289	0.244	0.265	0.083	0.066	0.455	0.460	0.458

Table 2. **3D reconstruction metrics of individual scenes on ScanNet and 7-Scenes.**

5. Novel View synthesis results

To evaluate the performance of our method on novel view synthesis, we manually set novel camera poses by adding 0.2 to training camera poses along the z-axis. Figure 3 presents the qualitative comparisons on ScanNet. Please refer to the supplementary video for more results.

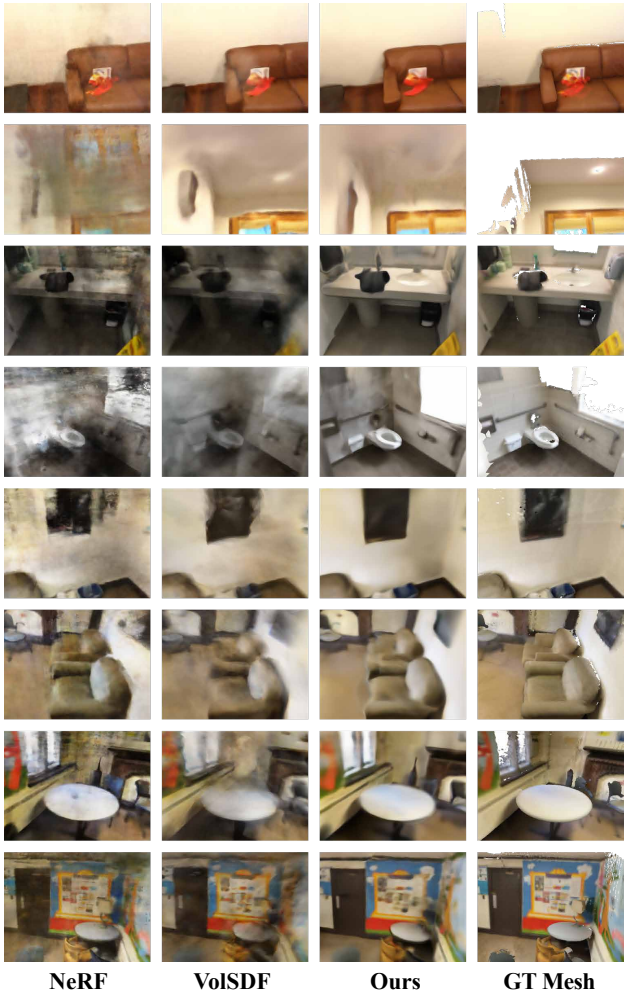

Figure 3. **Novel view synthesis results on ScanNet.** Due to the lack of ground truth images in novel views, we render ground truth mesh in these views for reference.

6. Performance on non-Manhattan scenes.

Our method is not suitable to non-Manhattan scenes due to there may exist sloped ground and walls maybe not vertical to each other, making our geometric constraints cannot work well.

7. Zoom-in visualization of non-planar regions.

We provide zoom-in visualization of geometric details of non-planar regions in Fig. 4.

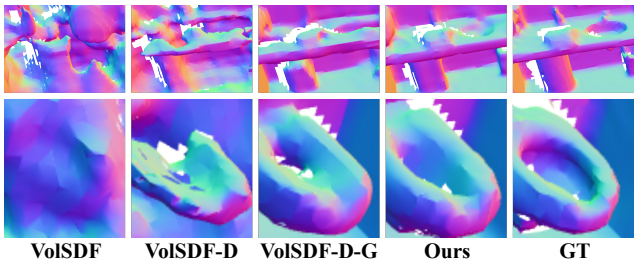

Figure 4. **Zoom-in visualization of non-planar regions.**

8. Effect of inaccurate segmentation.

If non-planar regions are mis-classified as planar, the reconstruction may be misled by the wrong regularization as shown in Fig. 5 (a). Our joint optimization can correct it as shown in Fig. 5 (b).

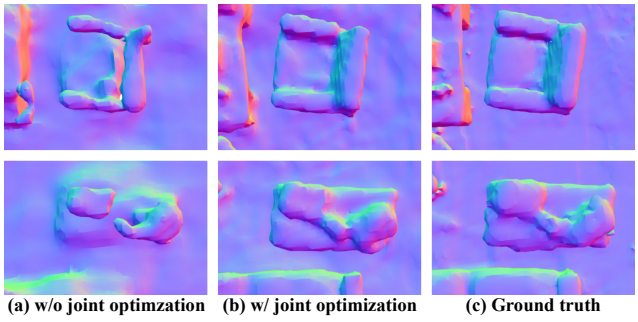

Figure 5. **Effect of inaccurate segmentation.**

9. Why non-planar regions are improved?

As the whole scene is represented by a single SDF network with the Eikonal regularization, the convergence and reconstruction quality of planar regions (dominant area) will also influence those in non-planar regions.

10. Comparison with supervised method.

We provide results of DeepVideoMVS in Tab. 3.

	Acc↓	Comp↓	Prec↑	Recall↑	F-score↑
DeepVideoMVS	0.206	**0.032**	0.286	**0.827**	0.424
Ours	**0.072**	0.068	**0.621**	0.586	**0.602**

Table 3. **Comparison with DeepVideoMVS on ScanNet.**