Fast Monocular Scene Reconstruction with Global-Sparse Local-Dense Grids
Supplementary Material

Wei Dong* Chris Choy† Charles Loop† Or Litany† Yuke Zhu† Anima Anandkumar†

1. Depth Scaler Optimization

Our system adopts monocular depth map predictions from off-the-shelf networks [2] using the DPT backbone [11]. However, these depth priors are not metric and the scale of each depth prediction is independent of others. Thus, we define the unary and binary (pairwise) constraints to estimate consistent metric scales.

1.1. Unary Constraints

Our pipeline relies on COLMAP's [12] sparse reconstruction for unary constraints. COLMAP supports sparse reconstruction with or without poses. Both modes start with SIFT [6] feature extraction and matching. The *with pose* mode then runs triangulation, while the *without pose* mode runs bundle adjustment to also estimate poses. *With pose* mode usually runs within 1 min, while the *without pose* mode often finishes around 5 mins for a sequence with several hundred frames. While our system integrates both modes, for fair comparison on the benchmark datasets, we adopt the *with pose* mode in quantitative experiments where ground truth poses from RGB-D SLAM are given. Fig. 1 shows the sparse reconstructions from the *with pose* mode.

1.2. Binary Constraints

Once we have camera poses and the sparse reconstruction, we can define which triangulated feature points are visible to which cameras (covisible). Thus, we can create pairwise reprojection constraints between frames, similar to loop closures in the monocular SLAM context [8]. We directly retrieve the feature matches obtained by COLMAP, and setup such frame-to-frame covsibility constraints. Fig. 1 shows the covisibility matrices, where entry (i, j) indicates the number of covisible features between frame i and j. They are used to establish binary constraints between frames for refining monocular depth scales.

*CMU RI. Work done during the internship at NVIDIA.
†NVIDIA Research

2. Volumetric Fusion

Eq. 9 in the main paper shows the least squares to initialize voxel-wise SDF. The more detailed implementation follows KinectFusion [9], where a truncation function ψ is used to reject associations.

$$\theta_d(\mathbf{v}) = \arg\min_d \sum_i \|d - \psi\big(d^o, \mu\big)\|^2, \tag{1}$$

$$d^o = d_{\mathbf{v}\to i} - \mathcal{D}_i(\mathbf{p}_{\mathbf{v}\to i})\phi_i(\mathbf{p}_{\mathbf{v}\to i}), \tag{2}$$

$$\psi(x, \mu) = \min(x, \mu), \tag{3}$$

where μ is the truncation distance. μ is associated with the *Dilate* operation and voxel block resolution in Eq. 7-8 in the main paper. Formally, we define

$$\text{Dilate}_R(\mathbf{x}) = \left\{\mathbf{x}_i \mid \left\|\mathbf{x}_i - \left\lfloor\frac{\mathbf{x}}{L}\right\rfloor\right\|_0 \leq R\right\}, \tag{4}$$

where L is the voxel block size, $\mathbf{x}_i$ are quantized grid points around, and R is the dilation radius. We use $R = 2$ (corresponding to two 8^3 voxel blocks) to account for the uncertainty around surfaces from the monocular depth prediction. Correspondingly, we use $\mu = L \cdot R$ to truncate the SDF.

The volumetric fusion runs at 50 Hz with RGB and SDF fusion, and at 30 Hz when additional semantic labels are also fused, hence serves as a fast initializer.

3. Hyper Parameters

We followed [17]'s hyperparameter choices and used $\lambda_d = 0.1, \lambda_n = 0.05$ for the rendering loss.

For regularizors, we obtained from hyper param sweeps from the 0084 scene of ScanNet that $\lambda_{\text{eik}} = 0.1$ for the Eikonal loss, and $\lambda_{\text{color}} = 10^{-3}, \lambda_{\text{label}} = 0.1, \lambda_{\text{normal}} = 1$ for the CRF loss.

In Gaussian kernels, we fix $\sigma_{\text{sdf}} = 1.0$ and $\sigma_{\text{color}} = 0.1$.

4. Evaluation

4.1. Metrics

We follow the evaluation protocols defined by ManhattanSDF [4], where the metrics between predicted point set

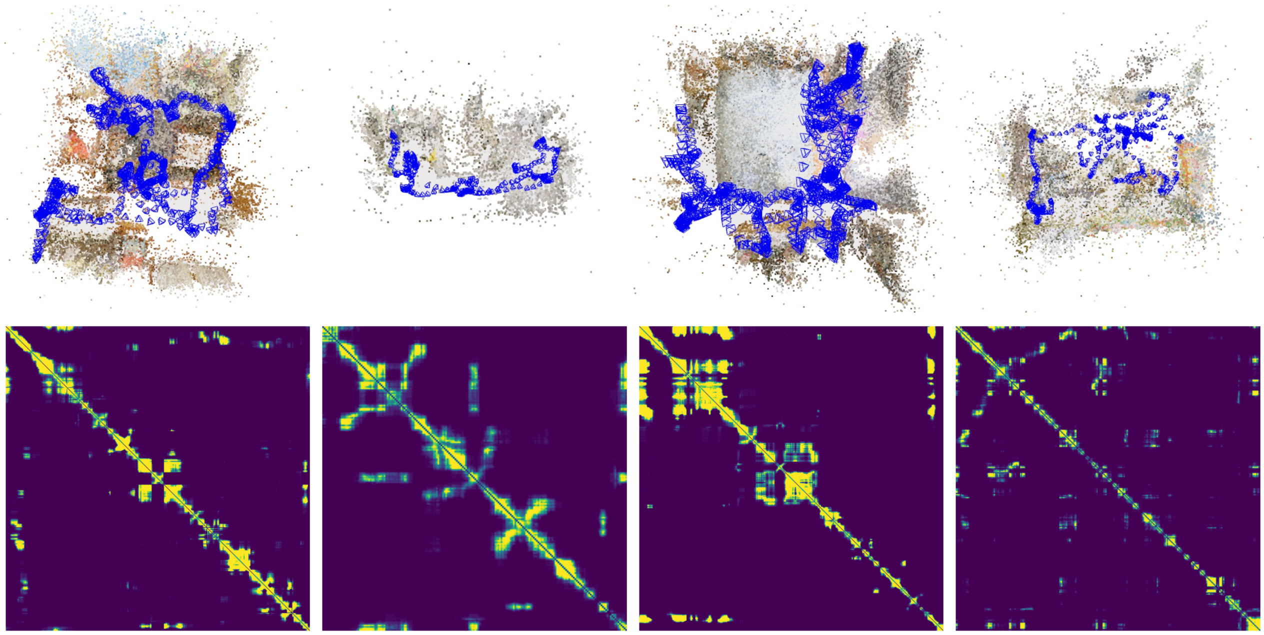


Figure 1. Sparse reconstruction and covisibility matrix of ScanNet scenes selected by ManhattanSDF [4].

P and ground truth point set P^* are

$$D(p, p^*) = \|p - p^*\|, \tag{5}$$
$$\mathrm{D_{Acc}}(P, P^*) = \underset{p \in P}{\mathrm{mean}} \min_{p^* \in P^*} D(p, p^*), \tag{6}$$
$$\mathrm{D_{Comp}}(P, P^*) = \underset{p^* \in P^*}{\mathrm{mean}} \min_{p \in P} D(p, p^*), \tag{7}$$
$$\mathrm{Prec}(P, P^*) = \underset{p \in P}{\mathrm{mean}} \Big(\big(\min_{p^* \in P^*} D(p, p^*) \big) < T \Big), \tag{8}$$
$$\mathrm{Recall}(P, P^*) = \underset{p^* \in P^*}{\mathrm{mean}} \Big(\big(\min_{p \in P} D(p, p^*) \big) < T \Big), \tag{9}$$
$$\text{F-score}(P, P^*) = \frac{2 \cdot \mathrm{Prec} \cdot \mathrm{Recall}}{\mathrm{Prec} + \mathrm{Recall}}, \tag{10}$$

where $T = 5$cm.

4.2. Generation of P and P^*

We follow previous works [4, 17] that applied TSDF refusion to generate P for evaluation: use Marching Cubes [5] to generate a global mesh; render depth map from mesh at selected viewpoints to crop points out of viewports; apply TSDF fusion [18] to obtain the final mesh and point cloud P. For fairness, we render depth at the resolution 480×640 for all approaches to be consistent with input (in contrast to MonoSDF that uses 968×1296 in their released evaluation code), and conduct refusion to a voxel grid at the resolution of 1cm.

To ensure the same surface coverage, we generate ground truth P^* at the same viewpoints with the same image and voxel resolution, only replacing rendered depth with ground truth depth obtained by an RGB-D sensor.

5. Additional Experimental Results

5.1. Ablation of scale optimization

To further illustrate the necessity of per-frame scale optimization, we show quantitative reconstruction results without scale optimization in Table 1. Here, volumetric fusion is conducted on an estimated single scale factor across all frames between monocular depth and SfM, resulting in poor initial reconstruction.

Table 1. Initial reconstruction results without per-frame scale optimization (*c. f.* Ours (Init) in Table 2-3.)

	Acc ↓	Comp ↓	Prec ↑	Recall ↑	F-score ↑
ScanNet	0.42	0.19	0.13	0.28	0.17
7-Scenes	0.36	0.12	0.19	0.43	0.26

5.2. Fusion and Refinement

Please see video supplementary for the incremental fusion from scaled depth, and the refinement stage that converges to general shapes within several hundred steps.

5.3. Scene-wise statistics on ScanNet

We use reconstructed mesh provided by ManhattanSDF [4], and report scene-wise statistics in Table 2. Reconstructions and corresponding ground truths are shown in Fig. 2.

It is observable that our reconstructions have low error at fine details with rich textures (*e.g.* 0050, furniture in 0580), but problems exist at texture-less regions (*e.g.* walls in 0580 and 0616, floor in 0084) due to the inaccurate scale estimate from sparse reconstructions. We plan to improve these by learning-based sparse or semi-dense reconstruction, *e.g.* [13, 14].

5.4. Scene-wise statistics on 7-scenes

The reconstructed mesh and scene-wise statistics are not provided by ManhattanSDF [4] for COLMAP, NeRF, UNISURF, NeuS, VolSDF, and ManhattanSDF. Therefore, we reuse their reported averages as a reference in the main paper. Here we report scene-wise numbers in Table 3 for the state-of-the-art MonoSDF [17] and our method. Reconstructions and ground truths are in Fig. 3.

7-scenes have challenging camera motion patterns and complex scenes, thus the overlaps between viewpoints are small, leading to reduced accuracy for all the approaches. Although our approach produces less accurate floor and walls with fewer features, it achieves fine reconstruction of desktop objects in general.

References

[1] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, pages 5828–5839, 2017. 4

[2] Ainaz Eftekhar, Alexander Sax, Jitendra Malik, and Amir Zamir. Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans. In *ICCV*, pages 10786–10796, 2021. 1

[3] Ben Glocker, Shahram Izadi, Jamie Shotton, and Antonio Criminisi. Real-time rgb-d camera relocalization. In *International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, October 2013. 5

[4] Haoyu Guo, Sida Peng, Haotong Lin, Qianqian Wang, Guofeng Zhang, Hujun Bao, and Xiaowei Zhou. Neural 3d scene reconstruction with the manhattan-world assumption. In *CVPR*, pages 5511–5520, 2022. 1, 2, 3, 4

[5] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. *ACM siggraph computer graphics*, 21(4):163–169, 1987. 2

[6] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004. 1

[7] 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. *Communications of the ACM*, 65(1):99–106, 2021. 4

[8] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. Orb-slam: a versatile and accurate monocular slam system. *IEEE Trans. Robotics*, 31(5):1147–1163, 2015. 1

[9] Richard A Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J Davison, Pushmeet Kohi, Jamie Shotton, Steve Hodges, and Andrew Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *2011 10th IEEE international symposium on mixed and augmented reality*, pages 127–136. Ieee, 2011. 1

[10] Michael Oechsle, Songyou Peng, and Andreas Geiger. Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *ICCV*, pages 5589–5599, 2021. 4

[11] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *ICCV*, pages 12179–12188, 2021. 1

[12] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, pages 4104–4113, 2016. 1, 4

[13] Zachary Teed and Jia Deng. Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. *NeurIPS*, 34:16558–16569, 2021. 3

[14] Zachary Teed, Lahav Lipson, and Jia Deng. Deep patch visual odometry. *arXiv preprint arXiv:2208.04726*, 2022. 3

[15] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021. 4

[16] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. *NeurIPS*, 34:4805–4815, 2021. 4

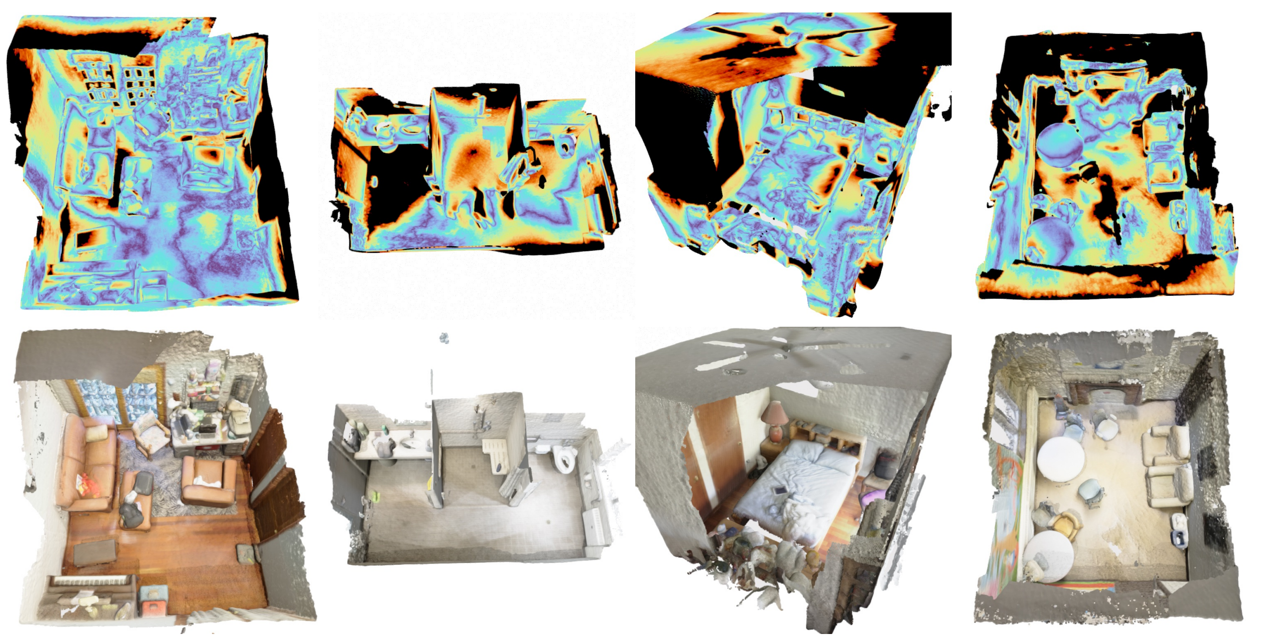
[17] Zehao Yu, Songyou Peng, Michael Niemeyer, Torsten Sattler, and Andreas Geiger. Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction. 2022. 1, 2, 3, 4, 5

[18] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Open3d: A modern library for 3d data processing. *arXiv preprint arXiv:1801.09847*, 2018. 2

Table 2. Scene-wise quantitative results on ScanNet.

Method	0050					0084				
	Acc ↓	Comp ↓	Prec ↑	Recall ↑	F-score ↑	Acc ↓	Comp ↓	Prec ↑	Recall ↑	F-score ↑
COLMAP [12]	0.049	0.129	0.707	0.531	0.607	0.032	0.121	0.807	0.577	0.673
NeRF [7]	0.704	0.081	0.215	0.517	0.304	0.733	0.248	0.157	0.213	0.181
UNISURF [10]	0.432	0.087	0.309	0.482	0.376	0.594	0.242	0.218	0.339	0.266
NeuS [15]	0.091	0.103	0.528	0.455	0.489	0.231	0.365	0.159	0.090	0.115
VolSDF [16]	0.071	0.071	0.600	0.599	0.599	0.507	0.165	0.163	0.247	0.196
ManhattanSDF [4]	0.032	0.050	0.849	0.755	0.800	**0.029**	**0.041**	**0.822**	**0.784**	**0.802**
MonoSDF (MLP) [17]	**0.025**	0.054	0.865	0.713	0.781	0.036	0.048	0.700	0.646	0.672
MonoSDF (Grid) [17]	0.027	0.045	0.854	0.764	0.807	0.035	0.043	0.796	0.774	0.785
Ours (Init)	0.034	0.051	0.775	0.684	0.727	0.047	0.048	0.705	0.725	0.715
Ours (+Rendering)	0.026	**0.044**	0.875	0.780	0.825	0.038	0.046	0.762	0.748	0.755
Ours (+CRF)	0.026	**0.044**	**0.880**	**0.788**	**0.832**	0.043	0.043	0.750	0.780	0.765

Method	0580					0616				
	Acc ↓	Comp ↓	Prec ↑	Recall ↑	F-score ↑	Acc ↓	Comp ↓	Prec ↑	Recall ↑	F-score ↑
COLMAP [12]	0.169	0.300	0.204	0.112	0.145	0.045	0.406	0.689	0.230	0.344
NeRF [7]	0.402	0.186	0.125	0.216	0.159	0.582	0.196	0.249	0.263	0.256
UNISURF [10]	0.392	0.192	0.131	0.188	0.155	0.571	0.148	0.237	0.300	0.265
NeuS [15]	0.206	0.275	0.167	0.114	0.135	0.137	0.140	0.330	0.289	0.308
VolSDF [16]	0.197	0.183	0.197	0.189	0.193	0.736	0.129	0.176	0.284	0.217
ManhattanSDF [4]	0.205	0.240	0.149	0.124	0.135	0.058	0.066	0.684	0.513	0.586
MonoSDF (MLP) [17]	**0.025**	**0.040**	**0.867**	**0.759**	**0.809**	0.039	0.087	0.702	0.488	0.576
MonoSDF (Grid) [17]	0.039	0.048	0.718	0.661	0.688	**0.033**	**0.048**	**0.815**	**0.646**	**0.721**
Ours (Init)	0.076	0.059	0.574	0.582	0.578	0.076	0.097	0.566	0.427	0.487
Ours (+Rendering)	0.070	0.080	0.760	0.636	0.692	0.046	0.070	0.699	0.504	0.586
Ours (+CRF)	0.046	0.050	0.707	0.682	0.694	0.057	0.080	0.659	0.504	0.571



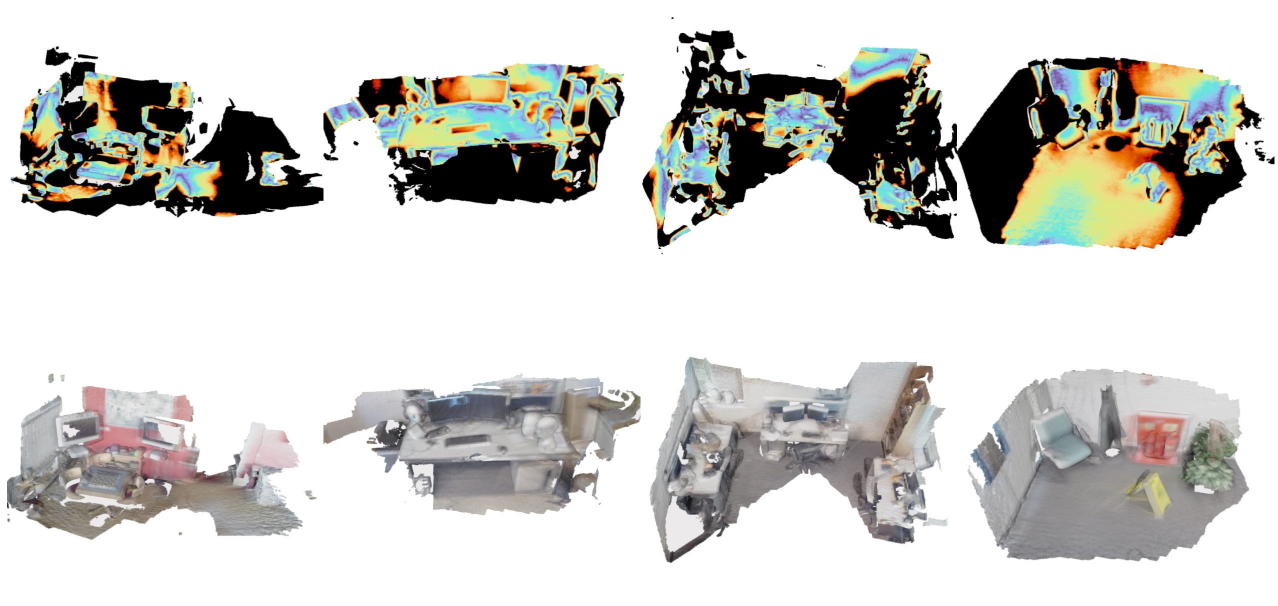
0050 0084 0580 0616

Figure 2. Error heatmap from our reconstruction (first row) to groundtruth (second row) for each scene in ScanNet [1]. Points are colorized by distance error ranging from 0 (blue) to 5cm (red) to its nearest neighbor in ground truth. Points with error larger than 5cm are regarded as outliers and colored in black.

Table 3. Scene-wise quantitative results on 7-Scenes.

Method	chess					heads				
	Acc ↓	Comp ↓	Prec ↑	Recall ↑	F-score ↑	Acc ↓	Comp ↓	Prec ↑	Recall ↑	F-score ↑
MonoSDF (MLP) [17]	0.160	0.390	0.250	0.132	0.173	**0.068**	0.188	**0.586**	0.353	0.440
MonoSDF (Grid) [17]	**0.113**	0.143	0.324	0.267	0.293	0.133	0.099	0.305	0.327	0.315
Ours (Init)	0.164	0.108	0.278	0.350	0.310	0.186	0.083	0.288	0.401	0.335
Ours (+Rendering)	0.147	0.111	0.367	0.389	0.378	0.074	0.062	0.543	0.568	0.555
Ours (+CRF)	0.147	**0.107**	**0.368**	**0.391**	**0.379**	0.071	**0.057**	0.559	**0.626**	**0.591**

Method	office					fire				
	Acc ↓	Comp ↓	Prec ↑	Recall ↑	F-score ↑	Acc ↓	Comp ↓	Prec ↑	Recall ↑	F-score ↑
MonoSDF (MLP) [17]	**0.087**	0.128	0.338	0.236	0.278	0.075	0.064	**0.592**	0.522	**0.555**
MonoSDF (Grid) [17]	0.147	0.077	**0.539**	0.471	**0.503**	**0.061**	0.081	0.564	0.504	0.533
Ours (Init)	0.168	**0.068**	0.398	**0.483**	0.436	0.087	**0.058**	0.503	**0.616**	0.554
Ours (+Rendering)	0.180	0.081	0.330	0.400	0.362	0.160	0.072	0.426	0.445	0.435
Ours (+CRF)	0.164	0.080	0.340	0.400	0.367	0.162	0.068	0.474	0.490	0.482



chess heads office fire

Figure 3. Error heatmap from our reconstruction (first row) to groundtruth (second row) for each scene in 7-Scenes [3]. The colorization is the same as Fig. 2.