VisFusion: Visibility-aware Online 3D Scene Reconstruction from Videos ——Supplementary Material——

Huiyu Gao, Wei Mao, Miaomiao Liu
Australian National University
{huiyu.gao, wei.mao, miaomiao.liu}@anu.edu.au

In this supplementary material, we first introduce more preprocessing details about datasets in Sec. 1. In Sec. 2, we present the definitions of evaluation metrics for 3D geometry and standard 2D depth, separately. In Sec. 3, we illustrate the surface update in the incremental reconstruction process. In Sec. 4, we provide additional quantitative results in 2D depth metrics following [14] and 3D geometry metrics following the evaluation protocol defined in [1]. In addition, more qualitative reconstructions on the 7-Scenes dataset are shown in Sec. 4. In Sec. 5, we provide more visualizations of similarity maps and learned visibility weights for both surface voxels and empty voxels. In Sec. 6, we discuss the limitations of our approach.

1. Details about Datasets

In the ScanNet dataset [3], each RGB image shares the same camera pose parameters with its corresponding ground truth depth map. All scenes are located in the first quadrant of the world coordinate system with the floors orthogonal to the z-axis. However, in the 7-Scenes dataset [12], the RGB and depth cameras have not been calibrated. We follow the preprocessing method in [10] to align the RGB image and depth maps. For each scene, we manually find the normal of the floor and calculate a transformation matrix that transforms the scene to the first quadrant of the world coordinate system with its floor orthogonal to the z-axis. All poses for each scene are updated by multiplying with this transformation matrix. The ground truth 3D mesh for each scene is generated from the calibrated ground truth depth maps using the standard TSDF fusion algorithm [2] with a voxel size of 1cm.

2. Definitions of Metrics

Following [8, 14], we evaluate our method on both 3D geometry metrics presented in [8] and standard 2D depth metrics defined in [5]. The definitions of these metrics are shown in Tab. 1.

2D Depth Metrics		3D Geometry Metrics	
Abs Rel	$\frac{1}{n}\sum \lvert d-d^*\rvert/d^*$	Acc	$\text{mean}_{p\in P}(\min_{p^*\in P^*}\lVert p-p^*\rVert)$
Abs Diff	$\frac{1}{n}\sum \lvert d-d^*\rvert$	Comp	$\text{mean}_{p^*\in P^*}(\min_{p\in P}\lVert p-p^*\rVert)$
Sq Rel	$\frac{1}{n}\sum \lvert d-d^*\rvert^2/d^*$	Chamfer	$\frac{1}{2}(\text{Acc}+\text{Comp})$
RMSE	$\sqrt{\frac{1}{n}\sum \lvert d-d^*\rvert^2}$	Prec	$\text{mean}_{p\in P}(\min_{p^*\in P^*}\lVert p-p^*\rVert<0.05)$
$\delta<1.25^i$	$\frac{1}{n}\sum(\max(\frac{d}{d^*},\frac{d^*}{d})<1.25^i)$	Recall	$\text{mean}_{p^*\in P^*}(\min_{p\in P}\lVert p-p^*\rVert<0.05)$
Comp	% valid predictions	F-score	$\frac{2\times\text{Prec}\times\text{Recal}}{\text{Prec}+\text{Recal}}$

Table 1. **Definitions of Metrics.** n is the number of pixels with both valid ground truth and predictions. d and d^* denote the predicted and ground truth depths. t and t^* denote the predicted and ground truth TSDFs. p and p^* denote the predicted and ground truth point clouds, respectively.

3. Global Update and Reconstruction

Different from depth-based methods [6, 7, 10, 15] that predict depth map for each input keyframe images *independently* and fuse it into a TSDF volume, our global feature fusion module makes the current-fragment reconstruction conditional on previous reconstructions and updates the surface geometry *globally*. An illustration of the surface update in the incremental reconstruction process is shown in Fig. 2. Thanks to our global feature fusion, the inaccurate surface reconstructed by previous fragments will be corrected with the new sequentially online input segment.

4. More Reconstruction Results

We report the experimental results evaluated by 2D depth metrics on ScanNet and 7-Scenes datasets in Tab. 2 and Tab. 3, respectively. Our method outperforms existing online feature fusion methods [1, 14] in most metrics. We also evaluate our method on ScanNet following the evaluation protocol defined in [1] and show the results in Tab. 4. Compared to the existing online feature fusion methods [1, 14], we achieve the best performance in the *Chamfer* distance metric. The qualitative comparison on 7-Scenes is shown in Fig. 3. Similar to results on ScanNet, our method is able to reconstruct more complete scenes compared to NeuralRecon [14] and more coherent scenes compared to SimpleRecon [10]. These results further demonstrate the generalization ability of the proposed method.

		Method	Abs Rel ↓	Abs Diff ↓	Sq Rel ↓	RMSE ↓	$\delta < 1.25$ ↑	Comp ↑
Depth Fusion		MVDNet [15]	0.121	0.193	0.088	0.327	0.870	0.945
		GPMVS [6]	0.094	0.153	0.069	0.282	0.902	0.948
		DPSNet [7]	0.109	0.177	0.080	0.306	0.882	0.948
		DeepVMVS [4]	0.067	0.112	0.040	0.216	0.936	0.945
		SimRec [10]	**0.046**	**0.083**	**0.022**	**0.173**	**0.954**	**0.944**
Feature Fusion	Offline	Atlas [8]	0.065	0.123	0.045	0.251	0.936	**0.999**
		3DVNet [9]	0.062	0.107	0.042	0.214	**0.941**	0.984
		VoRTX [13]	**0.058**	**0.092**	**0.036**	**0.199**	0.938	0.950
	Online	NeuRec [14]	0.065	0.106	0.031	0.195	**0.948**	0.909
		TF [1]	0.065	0.099	0.042	0.205	0.934	0.905
		Ours	**0.055**	**0.088**	**0.030**	**0.183**	0.942	**0.923**

Table 2. **Quantitative results of 2D metrics on ScanNet.** We show the results of two-stage depth fusion methods (top) and those for end-to-end feature fusion works (bottom) following the evaluation protocol in [14]. These depth scores are calculated by comparing the rendered depths from a fused mesh to the ground truth. We highlight the best results for *Depth Fusion*, *Feature Fusion Offline* and *Feature Fusion Online* methods in blue, teal, and violet, respectively. Offline methods assume to observe the whole video sequence. Our method performs the best among all online and offline feature fusion methods for most metrics.

Method	Abs Rel ↓	Abs Diff ↓	Sq Rel ↓	RMSE ↓	$\delta < 1.25$ ↑	Comp ↑
SimRec [10]	0.212	0.109	0.448	0.207	0.952	0.983
NeuRec [14]	**0.194**	0.125	**0.322**	0.231	0.932	0.871
Ours	0.215	**0.109**	0.454	**0.220**	**0.942**	**0.949**

Table 3. **Quantitative results of 2D metrics on 7-Scenes.** We evaluate our method on the official test split of 7-scenes using the same 2D metrics and evaluation protocol as in [14]. All methods are trained on ScanNet and for baseline methods, we use their released pre-trained models.

		Method	Acc ↓	Comp ↓	Chamfer ↓	Prec ↑	Recall ↑	F-score ↑
Depth Fusion		COLMAP [11]	10.22	11.88	11.05	0.509	0.474	0.489
		MVDNet [15]	12.94	8.34	10.64	0.443	0.487	0.460
		GPMVS [6]	12.90	8.02	10.46	0.453	0.510	0.477
		DPSNet [7]	11.94	7.58	9.77	0.474	0.519	0.492
		DeepVMVS [4]	10.68	6.90	8.79	0.541	0.592	0.563
		SimRec [10]	**5.53**	**6.09**	**5.81**	**0.686**	**0.658**	**0.671**
Feature Fusion	Offline	Atlas [8]	7.16	7.61	7.38	0.675	0.605	0.636
		3DVNet [9]	7.72	**6.73**	7.22	0.655	0.596	0.621
		VoRTX [13]	**4.31**	7.23	**5.77**	**0.767**	**0.651**	**0.703**
	Online	NeuRec [14]	5.09	9.13	7.11	0.630	**0.612**	0.619
		TF [1]	5.52	**8.27**	6.89	0.728	0.600	**0.655**
		Ours	**4.17**	9.05	**6.61**	**0.751**	0.580	0.653

Table 4. **Quantitative results of 3D metrics on ScanNet following the evaluation protocol in [1].** We show the results of two-stage depth fusion methods (top) and those for end-to-end feature fusion works (bottom). We highlight the best results for *Depth Fusion*, *Feature Fusion Offline* and *Feature Fusion Online* methods in blue, teal, and violet, respectively. All baseline results are taken from previous papers [1, 10]. Offline methods assume to observe the whole video sequence. Compared to all existing online feature fusion methods [1, 14], our method achieves the best performance in the *Chamfer* distance metric.

5. More Visualizations of Visibility

In Fig. 4, we visualize more similarity maps and learned visibility fusion weights for surface voxels and empty voxels in different local scene segments. These figures demonstrate that our local feature fusion module is able to distinguish the relevant views from the irrelevant ones for different kinds of voxels via the visibility weights.

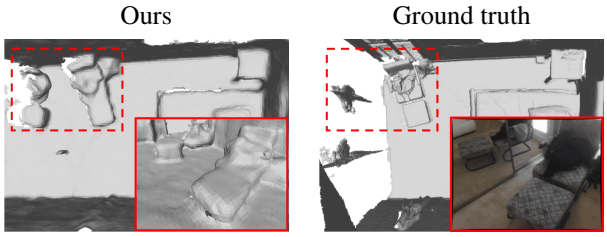
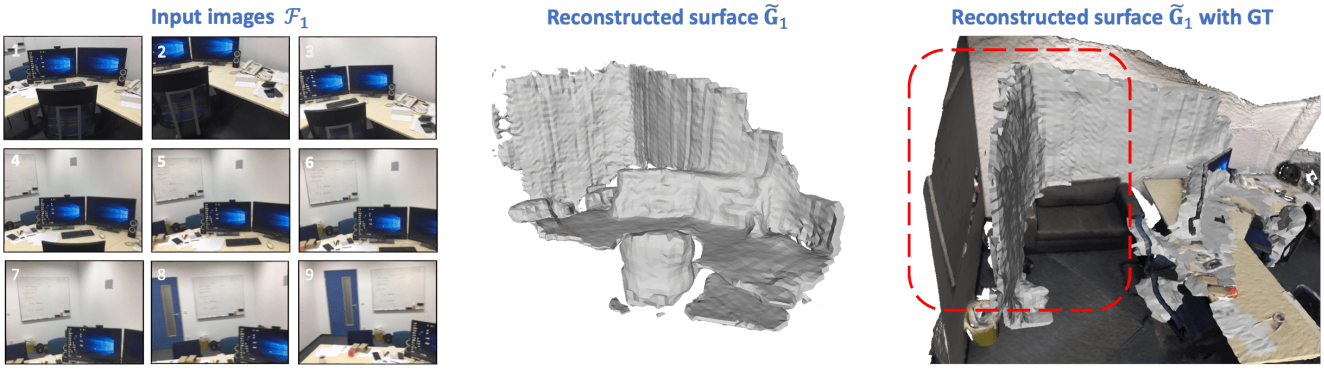


Figure 1. **Failure case.** Our approach tends to recover the 3D structure of both the real and virtual scenes caused by the mirror.

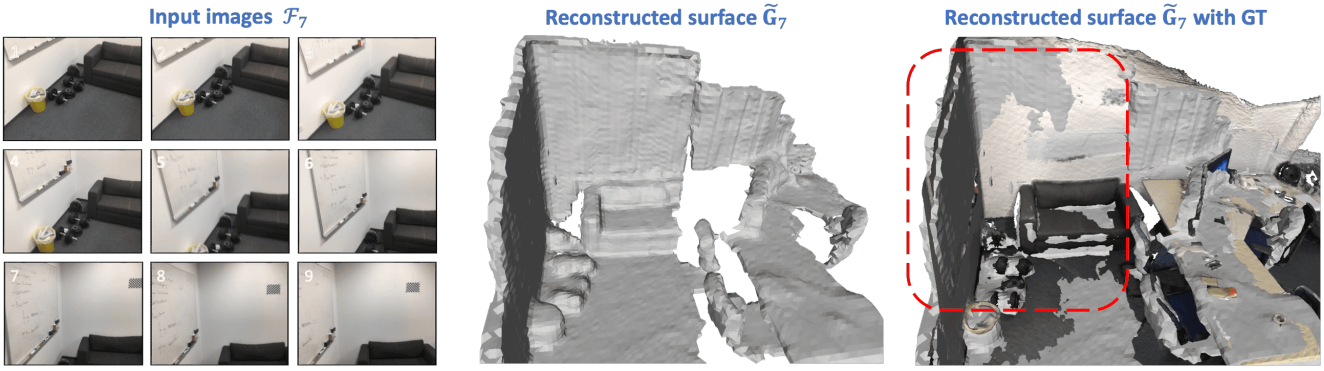
6. Limitation

First, our method cannot distinguish whether there is a mirror in the scene. One failure case is shown in Fig. 1. Instead of purely reconstructing the real scene surface, our approach tends to recover the 3D structure of both the real and virtual scenes caused by the mirror.

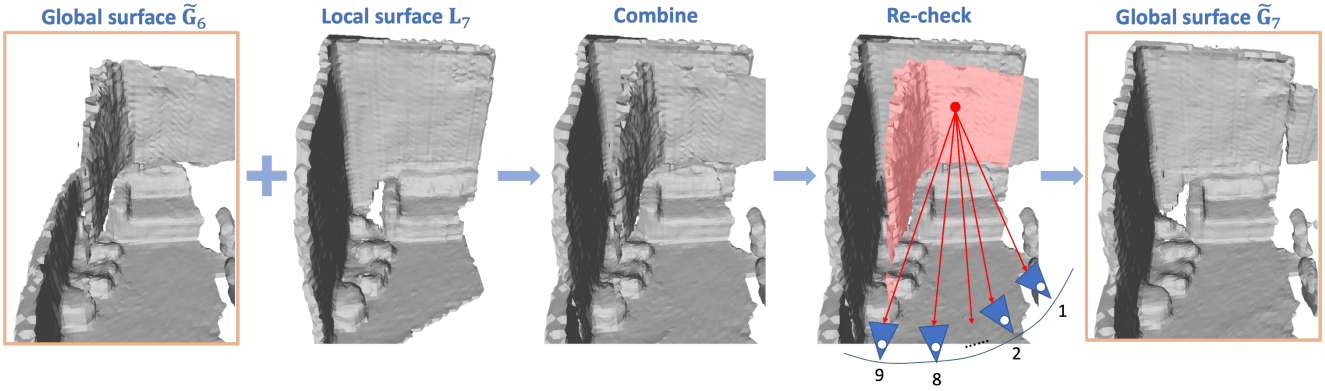
Second, although depth-based methods struggle to generate coherent surfaces, the detailed structures reconstructed by the depth-based methods are better than that reconstructed by feature fusion based methods which are limited by the resolution of the feature volume. We include more qualitative comparisons in Fig. 5. GPMVS [6] and DeepVideoMVS [4] are two real-time depth-fusion baselines with consistent video depth estimation. Since the temporal information is encoded in the latent space for depth estimation without directly considering the geometrical consistency of the scene, their results still have artifacts especially for textureless regions like walls. However, as highlighted in red boxes, GPMVS and DeepVideoMVS could reconstruct more detailed structures than VoRTX [13], which is the state-of-the-art *offline* feature fusion based method and our approach. Due to the global observation of the whole scene, VoRTX can achieve a more complete surface than our results or even the ground truth (highlighted in the blue boxes).



(a) Given the input fragment $\mathcal{F}_1$ consisting of 9 consecutive keyframes (left), the network outputs the surface geometry (middle) based on current observations, where the reconstruction of the distant wall is inaccurate (right).



(b) Given the input fragment $\mathcal{F}_7$ consisting of 9 consecutive keyframes (left), the network outputs the surface geometry based on current observations as well as previous reconstructions stored in the global map. In the updated global map (middle), the reconstruction of the wall falling in the current fragment bounding volume (FBV) is accurate (right).



(c) Given the global surface $\tilde{\mathbf{G}}_6$, reconstructed by previous fragments $\{\mathcal{F}_i\}_{i=1}^{6}$, and the local surface $\mathbf{L}_7$, reconstructed by the current fragment $\mathcal{F}_7$, our global feature fusion module fuses the local features into the global map. Specifically, in addition to the voxels reserved after local sparsification, the voxels that store global features and lie within the current camera views are also updated. By projecting these voxels into the input images of $\mathcal{F}_7$ and leveraging the photometric consistency between features from different views, the inaccurate surface reconstructed by previous fragments will be removed due to the low occupancy probabilities.

Figure 2. **Illustration of the surface update in the incremental reconstruction process.** Our global feature fusion module makes the current-fragment reconstruction conditional on previous reconstructions and updates the surface geometry *globally*.

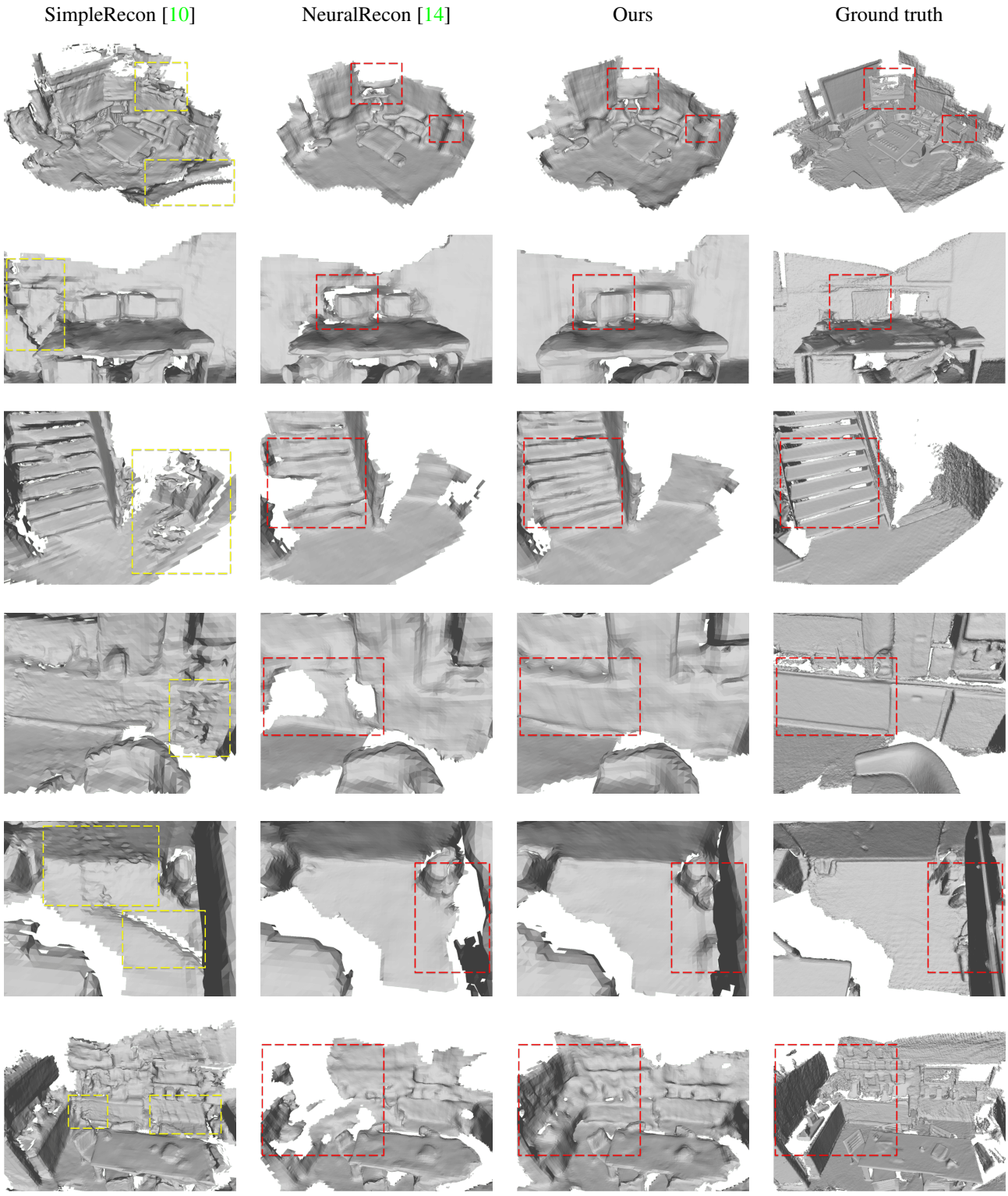
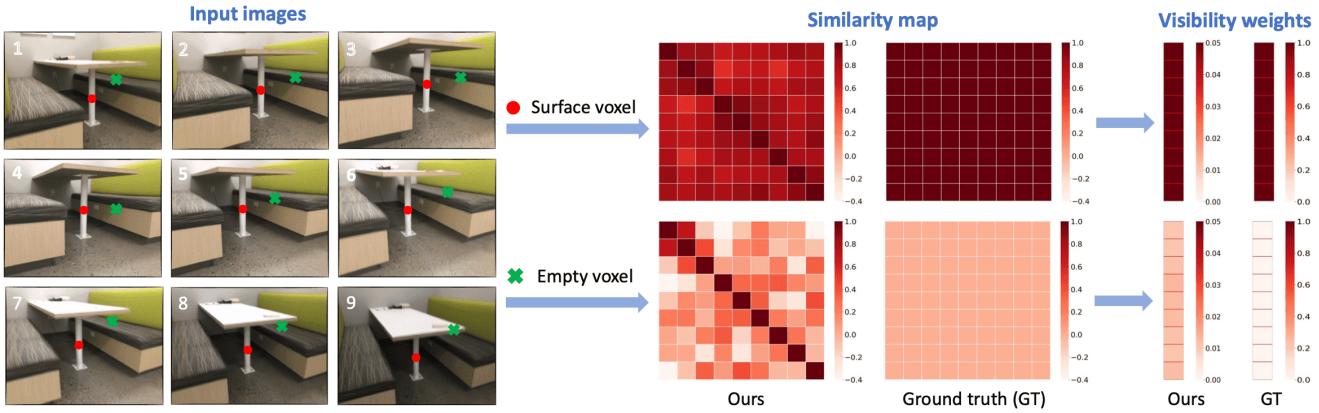
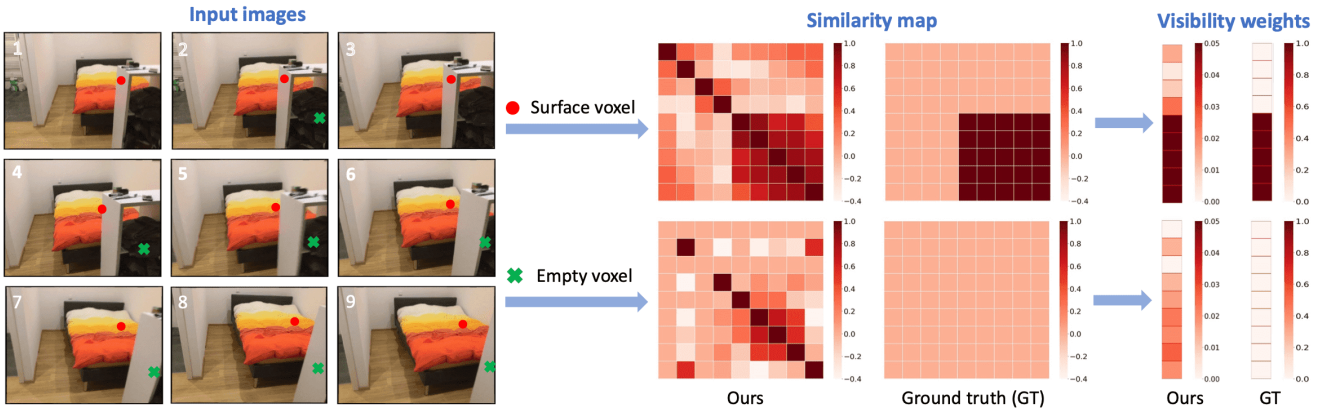


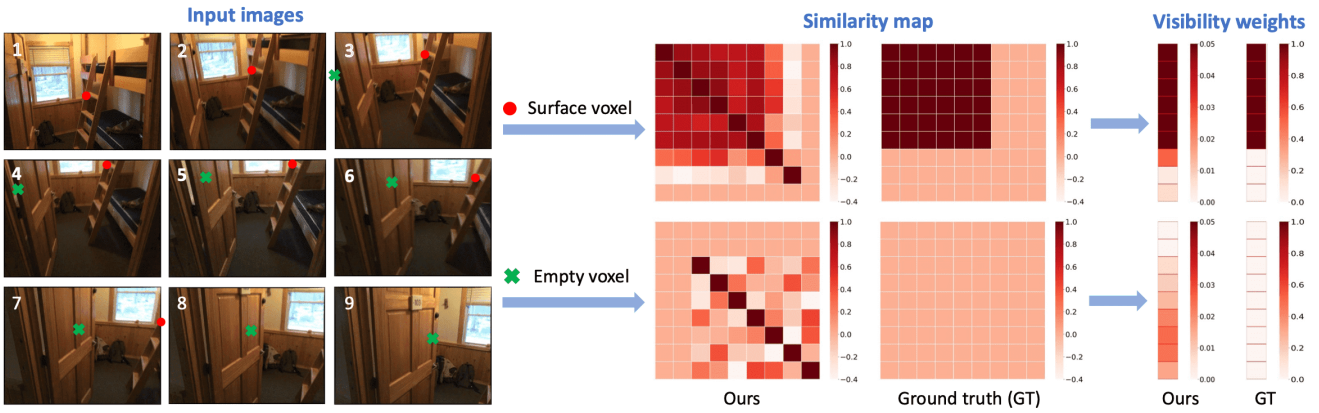
Figure 3. **Qualitative comparison on 7-Scenes.** Compared to NeuralRecon [14], our reconstruction results are more complete and contain more details (highlighted in the red boxes). Since SimpleRecon [10] is a two-stage depth-based method, it generates many artifacts and is not coherent (highlighted in the yellow boxes).



(a) The surface voxel (red dot) occupied by the leg of the table is visible in all views.



(b) The surface voxel (red dot) occupied by the quilt is visible in the last 5 views while occluded in the first 4 views.



(c) The surface voxel (red dot) occupied by the ladder is visible in the first 6 views, projected to the image edge in 7^{th} view, and to the outside of the image in the 8^{th} and 9^{th} views.

Figure 4. **More visualizations of the relations between the similarity map and the visibility weights.** We illustrate this relationship using two kinds of voxels. For the surface voxel (red dot), the features extracted from visible views have higher similarity, resulting in higher visibility weights. For the empty voxel (green cross), the features from different images are different leading to lower visibility weights.

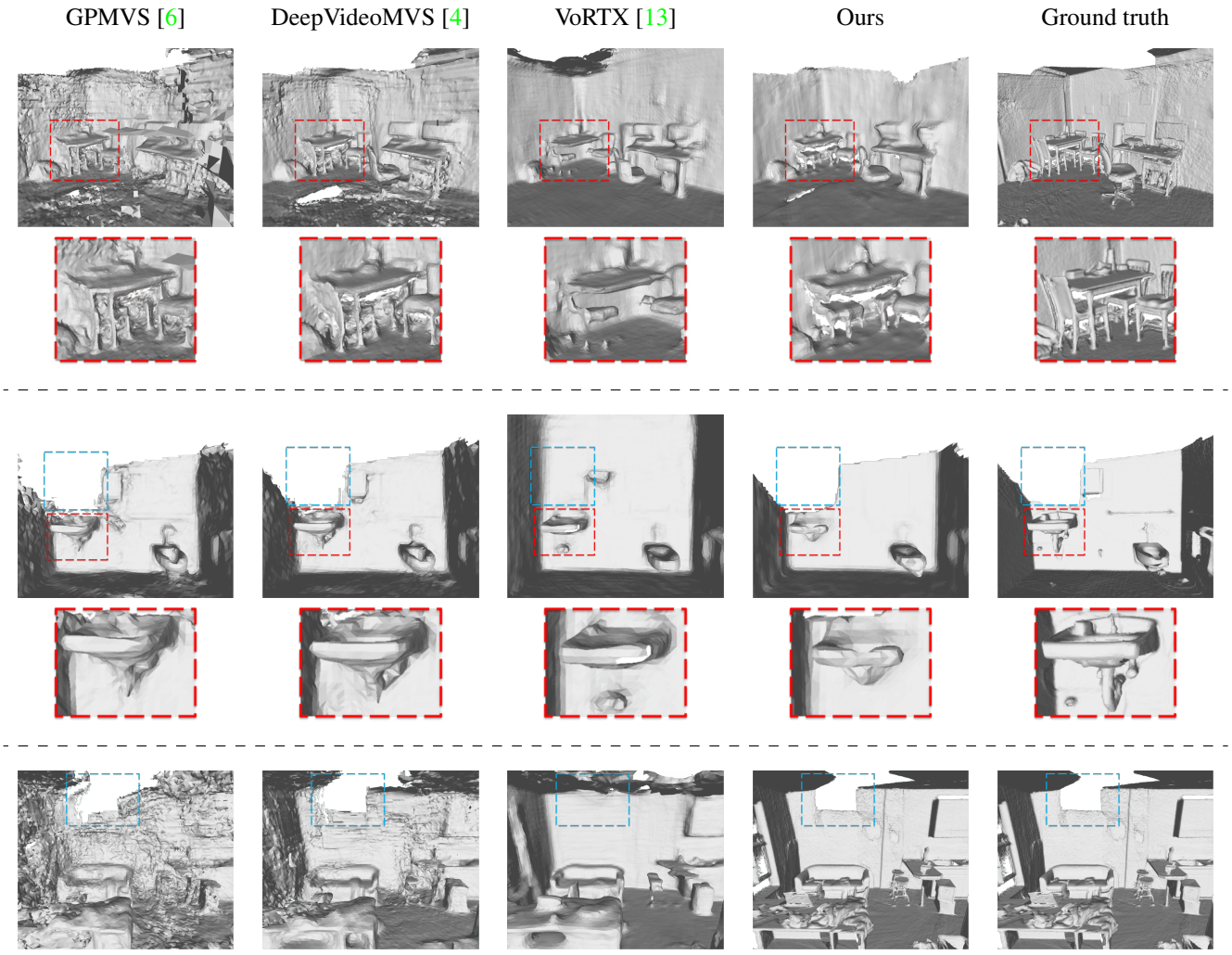


Figure 5. **Limitation of feature fusion based methods.** Compared to depth fusion based methods [4, 6], feature fusion based methods (including VoRTX [13], which is the state-of-the-art *offline* volumetric approach, and ours) are limited by feature volume resolution in the reconstruction of detailed structures (highlighted in the red boxes).

References

[1] Aljaz Bozic, Pablo Palafox, Justus Thies, Angela Dai, and Matthias Nießner. Transformerfusion: Monocular rgb scene reconstruction using transformers. *Advances in Neural Information Processing Systems*, 34:1403–1414, 2021. 1, 2

[2] Brian Curless and Marc Levoy. A volumetric method for building complex models from range images. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 303–312, 1996. 1

[3] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 1

[4] Arda Duzceker, Silvano Galliani, Christoph Vogel, Pablo Speciale, Mihai Dusmanu, and Marc Pollefeys. Deepvideomvs: Multi-view stereo on video with recurrent spatio-temporal fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15324–15333, 2021. 2, 6

[5] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*, 27, 2014. 1

[6] Yuxin Hou, Juho Kannala, and Arno Solin. Multi-view stereo by temporal nonparametric fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2651–2660, 2019. 1, 2, 6

[7] Sunghoon Im, Hae-Gon Jeon, Stephen Lin, and In So Kweon. Dpsnet: End-to-end deep plane sweep stereo. *arXiv preprint arXiv:1905.00538*, 2019. 1, 2

[8] Zak Murez, Tarrence van As, James Bartolozzi, Ayan Sinha, Vijay Badrinarayanan, and Andrew Rabinovich. Atlas: End-

to-end 3d scene reconstruction from posed images. In *European conference on computer vision*, pages 414–431. Springer, 2020. 1, 2

[9] Alexander Rich, Noah Stier, Pradeep Sen, and Tobias Höllerer. 3dvnet: Multi-view depth prediction and volumetric refinement. In *2021 International Conference on 3D Vision (3DV)*, pages 700–709. IEEE, 2021. 2

[10] Mohamed Sayed, John Gibson, Jamie Watson, Victor Prisacariu, Michael Firman, and Clément Godard. Simplerecon: 3d reconstruction without 3d convolutions. In *European Conference on Computer Vision*, pages 1–19. Springer, 2022. 1, 2, 4

[11] Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *European conference on computer vision*, pages 501–518. Springer, 2016. 2

[12] Jamie Shotton, Ben Glocker, Christopher Zach, Shahram Izadi, Antonio Criminisi, and Andrew Fitzgibbon. Scene coordinate regression forests for camera relocalization in rgb-d images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2930–2937, 2013. 1

[13] Noah Stier, Alexander Rich, Pradeep Sen, and Tobias Höllerer. Vortx: Volumetric 3d reconstruction with transformers for voxelwise view selection and fusion. In *2021 International Conference on 3D Vision (3DV)*, pages 320–330. IEEE, 2021. 2, 6

[14] Jiaming Sun, Yiming Xie, Linghao Chen, Xiaowei Zhou, and Hujun Bao. Neuralrecon: Real-time coherent 3d reconstruction from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15598–15607, 2021. 1, 2, 4

[15] Kaixuan Wang and Shaojie Shen. Mvdepthnet: Real-time multiview depth estimation neural network. In *2018 International conference on 3d vision (3DV)*, pages 248–257. IEEE, 2018. 1, 2