Supplementary Material for NeuralRecon: Real-Time Coherent 3D Reconstruction from Monocular Video

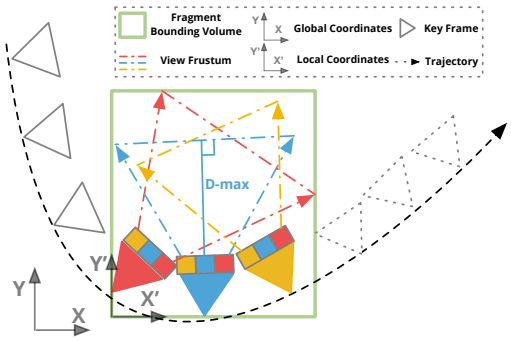

Figure 1. **Illustration for the definitions of local and global coordinate and FBV.**

1. Evaluation Metrics

The definitions of the evaluation metrics are detailed in Tab. 1.

2. Relationship with Atlas [2]

We discuss the relationship between Atlas and the proposed method NeuralRecon as follows: 1) We share the same idea of predicting the TSDF volume directly from the 3D feature volume, 2) One of the major focuses of Atlas is the scene completion (i.e. TSDF refinement) capability of this design. Atlas achieves impressive reconstruction completeness that sometimes even surpass the ground-truth. However, the major focus of NeuralRecon is on the computation efficiency of this design compared to depth-based methods, 3) Being a real-time method, NeuralRecon incrementally reconstructs local fragment geometry, instead of aggregating the image feature of the entire sequence and then predicts the global geometry as in Atlas and 4) The local fragment separation in our method is also acting as a view-selection mechanism that avoids irrelevant image features to be fused into the 3D volume.

3. The Single- and Double-Layered Mesh

Due to the TSDF Fusion implementation used in Atlas, the output mesh has two layers. The evaluation in Atlas is conducted using double-layered predictions and single-layered ground-truths. In the meantime, the results of other baselines are single-layered mesh. In order to make a fair comparison with these baselines, we report the evaluation results using both double-layered mesh and single-layered mesh. The single-layered mesh and double-layered mesh are visualized in Fig. 2.

4. Visualizations for Different Settings in the Ablation Study

The different settings in the ablation study are illustrated visually in Fig. 3.

5. More Qualitative Results

We provide more qualitative results in Fig. 4. When image textures are applied to the reconstructed mesh, the reconstruction quality is suitable for most applications.

6. Discussion on Depth Filtering and Fusion Methods

We experimented with several methods for depth filtering and fusion for depth-based baseline methods and discuss results here.

There are two main-stream methods for depth map fusion, namely TSDF fusion and point cloud fusion. As illustrated in the main paper, TSDF fusion is usually applied in real-time reconstruction pipelines. For point cloud fusion, depth maps from multiple views are back-projected to 3D and aggregated to a single point cloud according to the distance between each point. This is mostly done in offline MVS since the quality of the depth map is higher compared to real-time methods. Poisson surface reconstruction [1] is usually applied to reconstruct the 3D surface.

Point cloud filtering is often done in conjunction with point cloud fusion, through which outliers of depth maps are filtered according to the depth consistency across multiple depth maps. We found that off-the-shelf depth filtering technique in COLMAP [3] does not work well in TSDF fusion. As shown in Fig.5 (v), although the filtered point cloud is more accurate after the depth filtering (ii), the re-

maining points tend to cancel each other during TSDF fusion, resulting in almost no actual surface left.

Since the proposed method focuses on real-time 3D reconstruction, we opt to use TSDF fusion without depth filtering (instead of Poisson reconstruction) for surface reconstruction in most depth-based baseline methods.

7. Supplementary Video

In the supplementary video, we demonstrate the incremental reconstruction process of NeuralRecon in real-time applications. We also visually compare the reconstructions with other depth-based and volume-based baselines. We finally demonstrate the scalability of NeuralRecon by constructing a $30 \times 10m^2$ space.

2D		3D	
Abs Rel	$\frac{1}{n}\sum \lvert d - d^* \rvert / d^*$	Acc	$\text{mean}_{p \in P}(\min_{p^* \in P^*} \lvert\lvert p - p^* \rvert\rvert)$
Abs Diff	$\frac{1}{n}\sum \lvert d - d^* \rvert$	Comp	$\text{mean}_{p^* \in P^*}(\min_{p \in P} \lvert\lvert p - p^* \rvert\rvert)$
Sq Rel	$\frac{1}{n}\sum \lvert d - d^* \rvert^2 / d^*$	Prec	$\text{mean}_{p \in P}(\min_{p^* \in P^*} \lvert\lvert p - p^* \rvert\rvert < .05)$
RMSE	$\sqrt{\frac{1}{n}\sum \lvert d - d^* \rvert^2}$	Recal	$\text{mean}_{p^* \in P^*}(\min_{p \in P} \lvert\lvert p - p^* \rvert\rvert < .05)$
$\delta < 1.25$	$\frac{1}{n}\sum (\max(\frac{d}{d^*}, \frac{d^*}{d}) < 1.25)$	F-score	$\frac{2 \times \text{Perc} \times \text{Recal}}{\text{Prec} + \text{Recal}}$
Comp	% valid predictions		
RMSE log	$\sqrt{\frac{1}{n}\sum \lvert \log(d) - \log(d^*) \rvert^2}$		
Sc Inv	$(\frac{1}{n}\sum_i z_i^2 - \frac{1}{n^2}(\sum_i z_i)^2)^{1/2}$		

Table 1. **Metric definitions.** n is the number of pixels with both valid ground truth and prediction, d and d^* are the predicted and ground truth depth. t and t^* are the predicted and ground truth TSDFs, p and p^* are the predicted and ground truth point clouds, $z_i = \log d_i - \log d_i^*$.

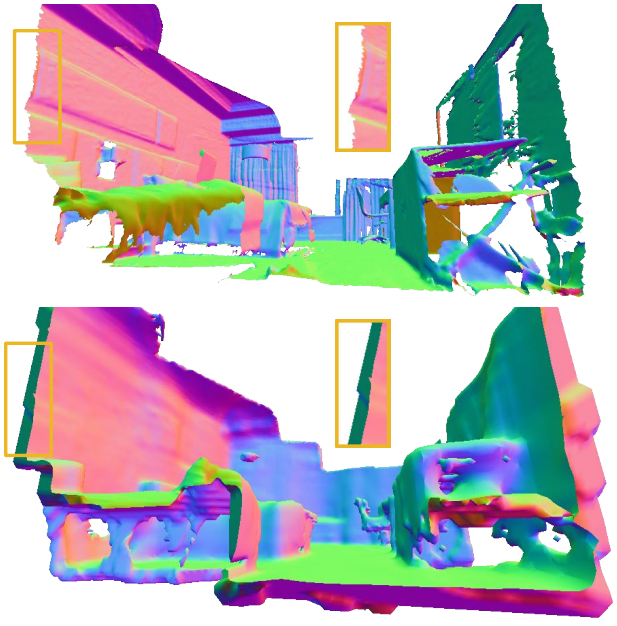

Figure 2. **Single-layered mesh (top) and double-layered mesh (bottom).**

References

[1] Michael Kazhdan and Hugues Hoppe. Screened Poisson Surface Reconstruction. *ACM Transactions on Graphics*, 32(3):1–13, 2013.

[2] Zak Murez, Tarrence van As, James Bartolozzi, Ayan Sinha, Vijay Badrinarayanan, and Andrew Rabinovich. Atlas: End-to-End 3D Scene Reconstruction from Posed Images. *ECCV*, 2020.

[3] Johannes L. Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise View Selection for Unstructured Multi-View Stereo. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, volume 9907, pages 501–518. Springer International Publishing, Cham, 2016.

[4] Zachary Teed and Jia Deng. DeepV2D: Video to Depth with Differentiable Structure from Motion. *ICLR*, 2020.

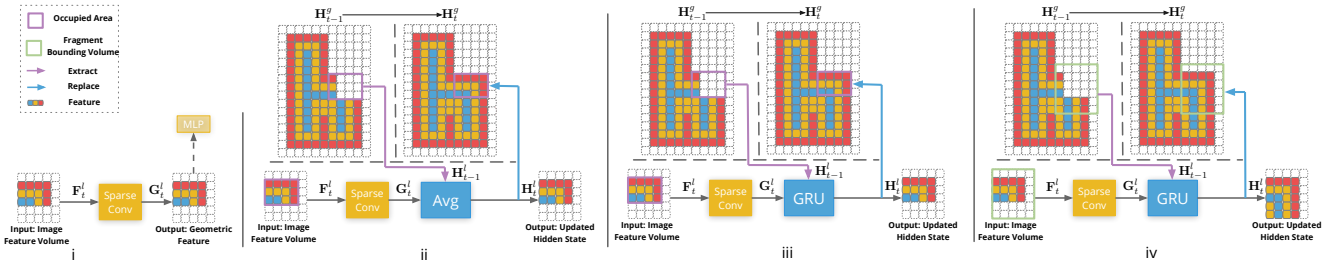

Figure 3. **2D toy examples to illustrate the different ablation settings in main paper.** The indications of Roman numerals are in Tab. 4 in main paper.

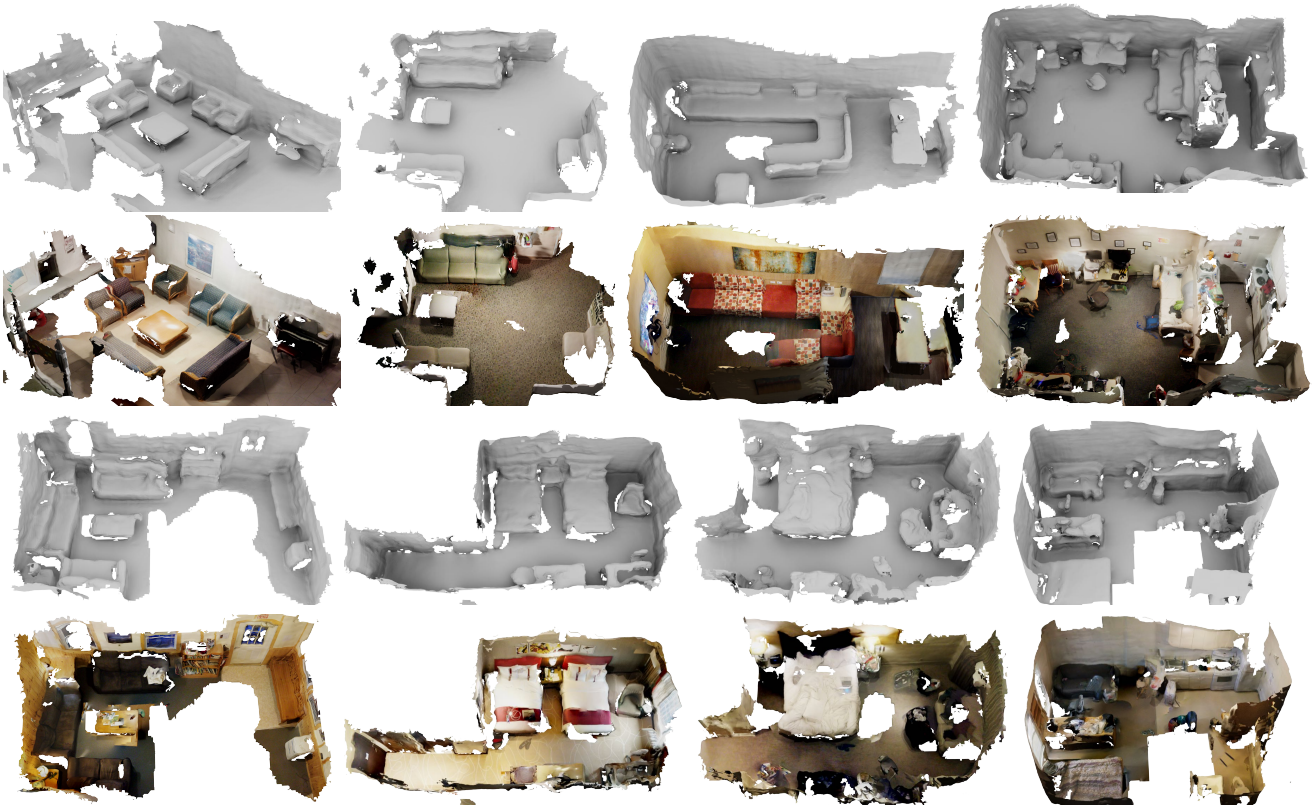

Figure 4. **More qualitative results** without and with image textures.

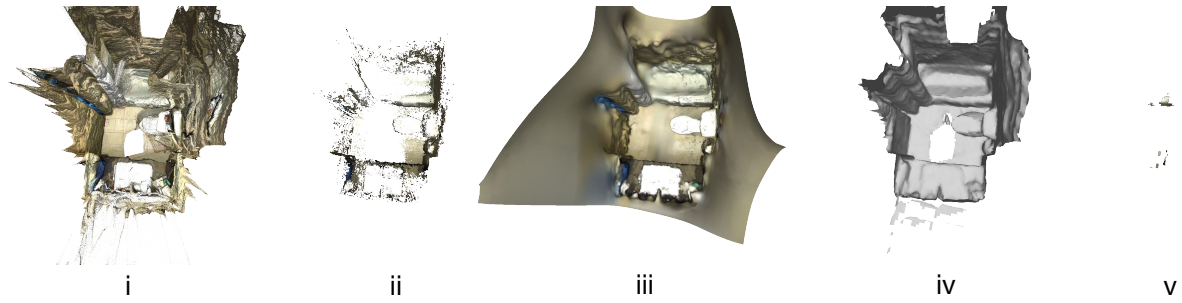

Figure 5. **Comparison on different depth filtering and meshing methods.** i. point cloud obtained from directly aligning multiple depth maps. ii. point cloud fused with multi-view depth consistency checking as depth filtering. iii. meshing through Poisson reconstruction with point cloud from (ii). iv. meshing through TSDF fusion with point cloud from (i). v. meshing through TSDF fusion with point cloud from (ii). The depth maps are obtained from DeepV2D [4] given ground-truth camera pose. (v) is achieved by projecting the filtered point cloud back to depth maps and fuse them with standard TSDF fusion. Zoom in for details.