SuperPrimitive: Scene Reconstruction at a Primitive Level

Supplementary Material

Method	ATE
Constant Depth	0.257
Constant Normals	0.208
Ours	**0.153**

Table 3. **Ablation study on TUM.** We ablate our odometry by flattening out SuperPrimitives. See the main text for the details.

Method	MAE	RMSE	iMAE	iRMSE
Ours With Depth From DPT-Hybrid	128.03	213.32	57.86	91.82
Ours	**109.0**	**204.15**	**47.32**	**83.40**

Table 4. **Ablation Study on VOID.** The quality of the unscaled depth obtained from method via surface normal integration is quantitatively compared against direct monocular depth estimation with DPT-Hybrid.

6. Ablation Study

6.1. Surface Normal vs Depth Prior

While a monocular depth prior has been widely used in 3D reconstruction [2, 8], in this work we use a surface normal prior instead. This choice is driven by the fact that surface normals have stronger generalisation abilities and capture better scene geometry. This observation is confirmed with an ablation study on the depth completion experiment described in Sec. 4.1.

In this ablation experiment, we replaced our unscaled depth estimates $\mathfrak{D}$ (obtained by integrating surface normals) with the outputs of DPT-Hybrid [37], a state-of-the-art depth estimator. All depth scales are then optimised in the same way as in the original method.

Our approach marginally outperforms its DPT counterpart on the VOID depth completion benchmark on all metrics (Tab. 4). Qualitatively (Fig. 7), our method is better at preserving structural properties of the scene, such as walls.

6.2. Surface Normal Quality Impact

We investigate how the surface normal quality affects the performance of our method, especially for pose estimation. There are two ablation levels we performed in this case. The first one just assumes constant depth $z = const$ within each SuperPrimitive. This makes our method akin to Multiplane images (MPI) [44, 59]. To emulate planar but possibly slanted segments, we replaced surface normal vectors with their averaged value within each SuperPrimitive independently. Our full odometry system performs significantly better than its two ablated counterparts, see Tab. 3.

7. Implementation Details

Segments Post Processing. Segments Ω_i extracted from the segmentation model may not be connected *a priori*. We perform a simple mask connectivity check and split a primitive into two in the case of detected mask discontinuities.

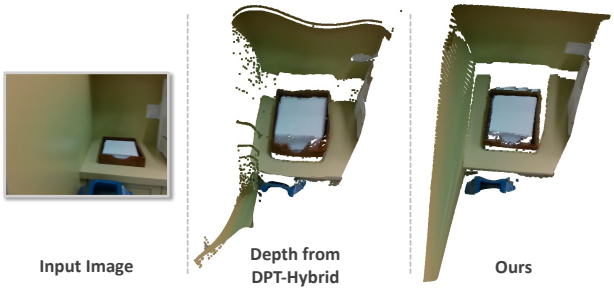

Figure 7. **Qualitative Ablation on VOID.**

Depth scales. Our depth scaling is implemented via storing a point p_i within the segment Ω_i. We represent depth scales s_i as the log-depth value at p_i. This point p_i is the same as the query point provided to the segmentation model as an input. This depth scale parametrisation allows converting partially available depth maps into a set of depth scaled SuperPrimitives.

7.1. Few-View Structure-From-Motion

We initialised depth scales uniformly to 1.0 for each primitive. All supplementary poses are initialised at identity. We solve SfM in a coarse-to-fine fashion to ensure the segments would not stuck in a local minima.

A small penalty on depth scale was added with the weight $w = 1\mathrm{e}{-5}$ to constraint segments that may not have photometric information from other views.

7.2. Depth Reinitialisation in MonoVO

Given a new keyframe $\mathrm{I}_{\mathrm{kf}}^{\mathrm{next}}$ with an estimated pose $\mathrm{T}_{\mathrm{kf}}^{\mathrm{next}}$, we scale depth for each new SuperPrimitve by using the geometry estimates of the previous keyframe $\mathrm{I}_{\mathrm{kf}}^{\mathrm{prev}}$.

More precisely, we transform the point cloud $\mathcal{G}_{\mathrm{prev}}$ of the previous keyframe into the coordinate system of the new keyframe and then render a partial depth map. Then, the depth scales s_i of the new keyframe are estimated as in the depth completion experiments in Sec. 4.1.

8. Experimental details

8.1. VOID Dataset

For depth completion evaluation on the VOID dataset, we follow the protocol of [55]. The ground truth depth is considered to be valid between 0.2 and 5.0 meters. The test set consists of 800 images. The dataset also provides sparse depth measurements obtained by an external visual-inertial

Metric	Units	Definition
MAE	*mm*	$\frac{1}{\lvert\Omega\rvert}\sum_{\boldsymbol{u}\in\Omega}\lvert\hat{z}(\boldsymbol{u})-z_{\text{gt}}(\boldsymbol{u})\rvert$
RMSE	*mm*	$\left(\frac{1}{\lvert\Omega\rvert}\sum_{\boldsymbol{u}\in\Omega}\lvert\hat{z}(\boldsymbol{u})-z_{\text{gt}}(\boldsymbol{u})\rvert^2\right)^{1/2}$
iMAE	*1/km*	$\frac{1}{\lvert\Omega\rvert}\sum_{\boldsymbol{u}\in\Omega}\lvert 1/\hat{z}(\boldsymbol{u})-1/z_{\text{gt}}(\boldsymbol{u})\rvert$
iRMSE	*1/km*	$\left(\frac{1}{\lvert\Omega\rvert}\sum_{\boldsymbol{u}\in\Omega}\lvert 1/\hat{z}(\boldsymbol{u})-1/z_{\text{gt}}(\boldsymbol{u})\rvert^2\right)^{1/2}$

Table 5. **Error metrics.** The definition of the error metrics used in depth quality valuation. Here, $\hat{z}$ and z_{gt} are predicted and ground truth depth values respectively.

odometry. We choose the setting with least sparse depth measurements available, "150 points" — where the SLAM system was configured to estimate depth of around 150 feature points (which constitutes 0.05% of the full image size).

8.2. TUM RGB-D Dataset

For evaluation of the estimated trajectory we used ATE RMSE metric [46].

9. Hardware Details

All of our experiments were conducted in the following hardware setup: Intel Core i7 3.60GHz processor, 32 GB RAM, and NVIDIA GeForce RTX 4090 with 24GB VRAM. Our method is implemented in PyTorch [35] and CuPy [33] libraries.

References

[1] Gwangbin Bae, Ignas Budvytis, and Roberto Cipolla. Estimating and exploiting the aleatoric uncertainty in surface normal estimation. In *International Conference on Computer Vision (ICCV)*, 2021. 3, 4

[2] M. Bloesch, J. Czarnowski, R. Clark, S. Leutenegger, and A. J. Davison. CodeSLAM — learning a compact, optimisable representation for dense visual SLAM. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 2, 12

[3] Aljaž Božič, Pablo Palafox, Justus Thies, Angela Dai, and Matthias Nießner. Transformerfusion: Monocular rgb scene reconstruction using transformers. *Proc. Neural Information Processing Systems (NeurIPS)*, 2021. 2

[4] Carlos Campos, Richard Elvira, Juan J. Gomez, Jose M. M. Montiel, and Juan D. Tardos. ORB-SLAM3: An accurate open-source library for visual, visual-inertial and multi-map SLAM. *IEEE Transactions on Robotics*, 2021. 8

[5] Xu Cao, Hiroaki Santo, Boxin Shi, Fumio Okura, and Yasuyuki Matsushita. Bilateral normal integration. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 4

[6] Alejo Concha and Javier Civera. Using superpixels in monocular SLAM. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2014. 2

[7] G. Cross and A. Zisserman. Quadric Reconstruction from Dual-Space Geometry. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 1998. 2

[8] J. Czarnowski, T. Laidlow, R. Clark, and A. J. Davison. Deepfactors: Real-time probabilistic dense monocular SLAM. In *IEEE Robotics and Automation Letters*, 2020. 8, 12

[9] A. J. Davison, N. D. Molton, I. Reid, and O. Stasse. MonoSLAM: Real-Time Single Camera SLAM. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2007. 2

[10] Eric Dexheimer and Andrew J. Davison. Learning a depth covariance function. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 8

[11] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Neural Information Processing Systems (NeurIPS)*, 2014. 1

[12] David Eigen, Christian Puhrsch, and Rob Fergus. Depth Map Prediction from a Single Image using a Multi-Scale Deep Network. In *Neural Information Processing Systems (NeurIPS)*, 2014. 2

[13] Jakob Engel, Thomas Schoeps, and Daniel Cremers. LSD-SLAM: Large-scale direct monocular SLAM. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014. 3

[14] J. Engel, V. Koltun, and D. Cremers. Direct sparse odometry. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2017. 3, 5

[15] Jiahui Fu, Yilun Du, Kurran Singh, Joshua B Tenenbaum, and John J Leonard. Neuse: Neural se (3)-equivariant embedding for consistent spatial understanding with objects. In *Proceedings of Robotics: Science and Systems (RSS)*, 2023. 2

[16] Y. Furukawa, B. Curless, S. M. Seitz, and R. Szeliski. Towards Internet-scale multi-view stereo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010. 2

[17] D. Gallup, J.-M. Frahm, and M. Pollefeys. Piecewise planar and non-planar stereo for urban scene reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010. 2

[18] Magnus R. Hestenes and Eduard Stiefel. Methods of conjugate gradients for solving linear systems. *Journal of research of the National Bureau of Standards*, 1952. 4

[19] M Kaess. Simultaneous localization and mapping with infinite planes. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2015. 2

[20] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015. 5

[21] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023. 3, 4

[22] G. Klein and D. W. Murray. Parallel Tracking and Mapping for Small AR Workspaces. In *Proceedings of the International Symposium on Mixed and Augmented Reality (ISMAR)*, 2007. 2, 5

[23] T. Laidlow and A. J. Davison. Simultaneous localisation and mapping with quadric surfaces. In *International Conference on 3D Vision (3DV)*, 2022. 2

[24] Stefan Leutenegger, Simon Lynen, Michael Bosse, Roland Siegwart, and Paul Furgale. Keyframe-based visual-inertial odometry using nonlinear optimization. *International Journal of Robotics Research (IJRR)*, 2014. 6

[25] Chen Liu, Jimei Yang, Duygu Ceylan, Ersin Yumer, and Yasutaka Furukawa. Planenet: Piece-wise planar reconstruction from a single rgb image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[26] Chen Liu, Kihwan Kim, Jinwei Gu, Yasutaka Furukawa, and Jan Kautz. Planercnn: 3d plane detection and reconstruction from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[27] Tianlin Liu, Parth T. Agrawal, Allison Chen, Byung-Woo Hong, and A. Wong. Monitored distillation for positive congruent depth completion. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 7

[28] B. D. Lucas and T. Kanade. An Iterative Image Registration Technique with an Application to Stereo Vision. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 1981. 6

[29] J. McCormac, R. Clark, M. Bloesch, A. J. Davison, and S. Leutenegger. Fusion++:volumetric object-level slam. In *Proceedings of the International Conference on 3D Vision (3DV)*, 2018. 2

[30] R. A. Newcombe, S. Lovegrove, and A. J. Davison. DTAM: Dense Tracking and Mapping in Real-Time. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2011. 1, 2, 5

[31] Lachlan Nicholson, Michael Milford, and Niko Sünderhauf. Quadricslam: Dual quadrics from object detections as landmarks in object-oriented slam. *IEEE Robotics and Automation Letters*, 2019. 2

[32] Michael Niemeyer, Jonathan T. Barron, Ben Mildenhall, Mehdi S. M. Sajjadi, Andreas Geiger, and Noha Radwan. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[33] Ryosuke Okuta, Yuya Unno, Daisuke Nishino, Shohei Hido, and Crissman Loomis. Cupy: A numpy-compatible library for nvidia gpu calculations. In *Proceedings of Workshop on Machine Learning Systems (LearningSys) in The Thirty-first Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2017. 13

[34] Jinsun Park, Kyungdon Joo, Zhe Hu, Chi-Kuei Liu, and In So Kweon. Non-local spatial propagation network for depth completion. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 7

[35] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. PyTorch: An imperative style, high-performance deep learning library. In *Neural Information Processing Systems (NeurIPS)*, 2019. 13

[36] Albert Pumarola, Alexander Vakhitov, Antonio Agudo, Alberto Sanfeliu, and Francese Moreno-Noguer. Pl-slam: Real-time monocular visual slam with points and lines. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2017. 2

[37] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 2, 6, 12

[38] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2022. 2, 6

[39] Xiaofeng Ren and Jitendra Malik. Learning a classification model for segmentation. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2003. 3

[40] Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M. Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 7

[41] R. F. Salas-Moreno, R. A. Newcombe, H. Strasdat, P. H. J. Kelly, and A. J. Davison. SLAM++: Simultaneous Localisation and Mapping at the Level of Objects. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013. 2

[42] Mohamed Sayed, John Gibson, Jamie Watson, Victor Prisacariu, Michael Firman, and Clément Godard. Simplerecon: 3d reconstruction without 3d convolutions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 2

[43] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[44] Jonathan Shade, Steven Gortler, Li-wei He, and Richard Szeliski. Layered depth images. In *Proceedings of the 25th annual conference on Computer graphics and interactive techniques*, 1998. 12

[45] Jingjia Shi, Shuaifeng Zhi, and Kai Xu. Planerectr: Unified query learning for 3d plane recovery from a single view. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2023. 2

[46] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers. A Benchmark for the Evaluation of RGB-D SLAM Systems. In *Proceedings of the IEEE/RSJ Conference on Intelligent Robots and Systems (IROS)*, 2012. 8, 13

[47] E. Sucar, K. Wada, and A. J. Davison. NodeSLAM: Neural object descriptors for multi-view shape reconstruction. In *Proceedings of the International Conference on 3D Vision (3DV)*, 2020. 2

[48] 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[49] Zachary Teed and Jia Deng. Deepv2d: Video to depth with differentiable structure from motion. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020. 7, 8

[50] Alexander Vakhitov, Jan Funke, and Francesc Moreno-Noguer. Accurate and linear time pose estimation from points and lines. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016. 2

[51] Tom van Dijk and Guido C.H.E. de Croon. How do neural networks see depth in single images? In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2019. 2

[52] Dor Verbin, Peter Hedman, Ben Mildenhall, Todd Zickler, Jonathan T. Barron, and Pratul P. Srinivasan. Ref-NeRF: Structured view-dependent appearance for neural radiance fields. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1

[53] Wenshan Wang, Yaoyu Hu, and Sebastian Scherer. TartanVO: A Generalizable Learning-based VO. In *Conference on Robot Learning (CoRL)*, 2020. 8

[54] C. S. Weerasekera, Y. Latif, R. Garg, and I. Reid. Dense monocular reconstruction using surface normals. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2017. 2

[55] Wofk, Diana and Ranftl, René and Müller, Matthias and Koltun, Vladlen. Monocular Visual-Inertial Depth Estimation. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2023. 6, 7, 12

[56] Alex Wong and Stefano Soatto. Unsupervised depth completion with calibrated backprojection layers. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 7

[57] Alex Wong, Xiaohan Fei, Stephanie Tsuei, and Stefano Soatto. Unsupervised depth completion from visual inertial odometry. *IEEE Robotics and Automation Letters*, 2020. 6, 7

[58] Alex Wong, Safa Cicek, and Stefano Soatto. Learning topology from synthetic data for unsupervised depth completion. *IEEE Robotics and Automation Letters*, 2021. 7

[59] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. In *Proceedings of SIGGRAPH*, 2018. 12