A METRIC FOR EVALUATING 3D RECONSTRUCTION AND MAPPING PERFORMANCE WITH NO GROUND TRUTHING

Guoxiang Zhang and YangQuan Chen

Mechatronics, Embedded Systems and Automation Lab,
University of California, Merced, Merced, USA

ABSTRACT

It is not easy when evaluating 3D mapping performance because existing metrics require ground truth data that can only be collected with special instruments. In this paper, we propose a metric, dense map posterior (DMP), for this evaluation. It can work without any ground truth data. Instead, it calculates a comparable value, reflecting a map posterior probability, from dense point cloud observations. In our experiments, the proposed DMP is benchmarked against ground truth-based metrics. Results show that DMP can provide a similar evaluation capability. The proposed metric makes evaluating different methods more flexible and opens many new possibilities, such as self-supervised methods and more available datasets.

Index Terms— Metric, SLAM, 3D reconstruction, Evaluation, Benchmark

1. INTRODUCTION

Background and motivation. Simultaneous localization and mapping (SLAM) [1, 2] and 3D reconstruction [3, 4, 5] methods have achieved dramatic advancements recently due to the increasing need for autonomous vehicles and consumer robots. But, it is still not an easy task to collect data for performance evaluation. Existing metrics require additional ground truth data, such as ground truth camera trajectories or ground truth 3D models, that need to be collected by special instruments, such as motion capture systems, RTK-GPS, and LiDAR. These instruments are not only expensive. They make it impossible to add ground truth to existing data. As a result, there are only a limited number of datasets with ground truth information available [6].

To address this problem, some researchers turn to generate synthetic data from graphical rendering on virtual scenes [7, 8]. This makes accessing ground truth data easy, but it introduces new challenges to create realistic virtual environments and camera trajectories. Some datasets [9, 10, 11] choose to use pseudo-ground truth data which are generated by estimation algorithms, while some others resort to external markers [12, 13].

Key contributions. In this work, we propose a metric, dense map posterior (DMP), for 3D reconstruction and mapping performance evaluation that can work without any ground truth data. Instead, it calculates a comparable value, reflecting a map posterior probability, from dense point cloud observations. In our experiments, the proposed metric is benchmarked against ground truth-based metrics. Results show that the proposed DMP can provide a similar evaluation capability.

Significance. The proposed metric makes a broad impact to SLAM and 3D mapping. We list a few major ones here:

1. It makes 3D mapping evaluation more flexible and easily accessible.
2. It provides a supervisory figure-of-merit signal for robust loop closure optimization.
3. It helps build large real-world SLAM datasets with minimal effort.
4. It makes it possible to introduce self-supervised machine learning algorithms to 3D reconstruction methods.

2. RELATED WORK

When evaluating 3D mapping results, the common error metrics are the absolute trajectory error (ATE) [2, 14, 15], the relative position error (RPE) [14], and surface distance error [3, 4, 16, 15]. The ATE represents the difference between the ground truth and an estimated trajectory in a common world frame. The RPE calculates the relative relation differences between estimated and ground truth trajectories. The surface distance error specifies the mean and median of distances between a reconstructed surface and a ground-truth surface. All these metrics require ground truth data that is not trivial to collect.

The work closely related to ours is [17]. Olson and Kaess discussed using sparse map posteriors for map evaluation, but they concluded that the posterior subjects to over-fitting during the optimization process. Thus it is not reliable. In their formulation, the sparse map posterior is calculated on a map of sparse landmarks. We agree that, with sparse data only, it is not as reliable as ground truth-based metrics. However, in our method, we consider the posterior of a dense 3D map given

geometry-related observations. The over-fitting is no longer a problem.

3. DENSE MAP POSTERIOR (DMP)

3D reconstruction and mapping methods usually take a set of sensor reading and output both an estimated 3D model (*e.g.* mesh or surfels, *etc.*) $\mathcal{M}$ and an estimated camera or LiDAR trajectory $\mathbb{T}$. For simplicity, we denote $M = \{\mathcal{M}, \mathbb{T}\}$ to include both outputs. For the input sensor readings, we denote geometry related readings as $\mathbb{Z}$, where $\mathbb{Z}$ can be a set of depth images from RGB-D cameras or a set of point clouds from LiDARs.

The idea behinds the proposed metric, dense map posterior (DMP), is to calculate a value reflecting a posterior probability $p(M|\mathbb{Z})$. For this calculation, DMP takes both M and $\mathbb{Z}$ as evaluation input. Then, performance can be evaluated by comparing metric values of different 3D reconstruction estimates $\{M_i\}$.

Because there is no direct way to calculate $p(M|\mathbb{Z})$, derivations are made to get a computable form. First, the Bayes' rule is applied to get

$$p(M|\mathbb{Z}) = \frac{p(\mathbb{Z}|M)p(M)}{p(\mathbb{Z})}. \tag{1}$$

Then, for the same set of observations $\mathbb{Z}$, if there are two different 3D reconstruction estimates M_1 and M_2, we can get a ratio

$$\frac{p(M_1|\mathbb{Z})}{p(M_2|\mathbb{Z})} = \frac{\frac{p(\mathbb{Z}|M_1)p(M_1)}{p(\mathbb{Z})}}{\frac{p(\mathbb{Z}|M_2)p(M_2)}{p(\mathbb{Z})}}. \tag{2}$$

If we assume $p(M_1) = p(M_2)$, we can simplify (2) to be

$$\frac{p(M_1|\mathbb{Z})}{p(M_2|\mathbb{Z})} = \frac{p(\mathbb{Z}|M_1)}{p(\mathbb{Z}|M_2)}, \tag{3}$$

where

$$p(\mathbb{Z}|M) = \prod_{Z \in \mathbb{Z}} p(Z|M), \tag{4}$$

where Z is a depth image or a single point cloud scan of LiDAR. Note that Z consists of many independent observations (*i.e.*, depth image pixels or 3D points). We denote each of them as $z \in Z$, then

$$p(Z|M) = \prod_{z \in Z} p(z|M). \tag{5}$$

Assume that z follows a Gaussian distribution:

$$p(z|M) \sim \mathcal{N}(z', \sigma^2) \tag{6}$$

where z' and σ are the mean and covariance of the Gaussian distribution. z' is produced by applying sensor model $\mathcal{S}(\cdot)$ on M with the same observing configuration as z:

$$z' = \mathcal{S}_z(M) = \mathcal{S}_z(\mathcal{M}, T_z), \tag{7}$$

where $\mathcal{S}(\cdot)$ is a camera projection model for depth sensors and a range-bearing model for LiDARs. T_z is the estimated sensor observing perspective for z. One can easily get T_z from the estimated trajectory $\mathbb{T}$. For numerical stability, we take log on both sides of (3). We get a log-likelihood ratio (LLR)

$$LLR = \log p(\mathbb{Z}|M_1) - \log p(\mathbb{Z}|M_2). \tag{8}$$

We define

$$r(M, \mathbb{Z}) = -\log p(\mathbb{Z}|M). \tag{9}$$

When $LLR > 0$, which equivalents to $r(M_1, \mathbb{Z}) < r(M_2, \mathbb{Z})$, M_1 is more likely to be better than M_2 in terms of dense map posterior probability. This means that we can compare different $r(M_i, \mathbb{Z})$ directly. The lower, the better. To further simplify (9), one can plug (4), (5), and (6) into (9) to get

$$-\log p(\mathbb{Z}|M) = -\log \left[\prod_{Z \in \mathbb{Z}} \prod_{z \in Z} \frac{1}{\sigma\sqrt{2\pi}} e^{-(z-z')^2/2\sigma^2} \right]$$
$$= \sum_{Z \in \mathbb{Z}} \sum_{z \in Z} \frac{(z-z')^2}{2\sigma^2} + L_0, \tag{10}$$

where L_0 is a constant. Since L_0 is the same for the same set of $\mathbb{Z}$, then (9) can be simplified to be

$$r(M, \mathbb{Z}) \sim \overline{r}(M, \mathbb{Z}) = \sum_{Z \in \mathbb{Z}} \sum_{z \in Z} (z-z')^2, \tag{11}$$

where $\overline{r}(M, \mathbb{Z})$ is the proposed metric DMP, the lower, the better. To calculate its value, one needs to provide a 3D reconstruction estimate M, sensor readings $\mathbb{Z}$, and a sensor model $\mathcal{S}(\cdot)$.

4. EVALUATION OF DMP – THE PROPOSED METRIC

To evaluate the proposed metric DMP, we conduct experiments to compare the results of different metrics on multiple datasets: TUM RGB-D dataset [14] and Augmented ICL-NUIM [4]. In all of our experiments, 3D models are fused using Surfels implemented by ElasticFusion [15].

4.1. TUM RGB-D dataset

The TUM RGB-D dataset [14] is widely used for evaluating SLAM systems. The dataset has RGB-D sequences with ground truth camera trajectories available. We run experiments on a subset of sequences that are commonly used for SLAM evaluation [2]. Following [14], we adapt the absolute trajectory error (ATE) as the baseline metric.

For each data sequence, four 3D reconstruction estimates, M_1, M_2, M_3, and M_4 are used as benchmarking data. These estimates are from SLAM systems with different characteristics. The results are reported in Fig. 1. For better visualization, M_1 to M_4 are sorted with the ATE metric.
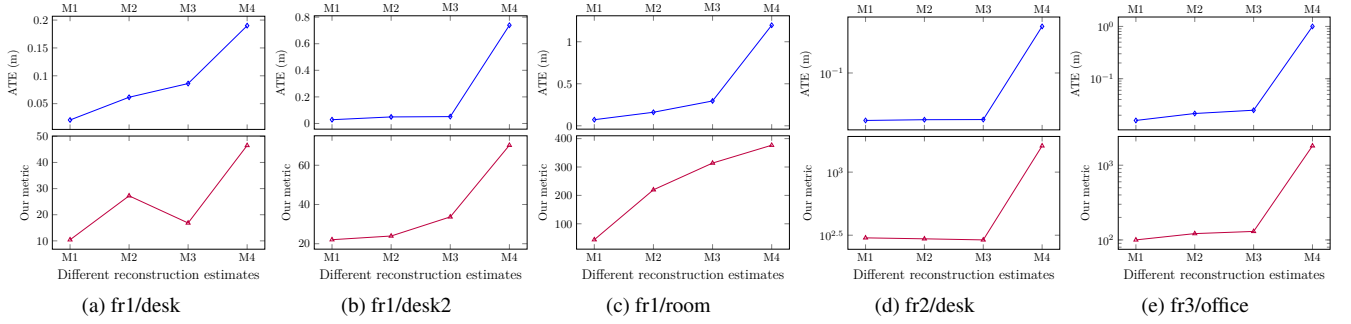
Fig. 1. Comparison of ATE and our metric DMP. For each sequence, curves of two metrics follow almost the same ascending order, which means metrics have very similar comparison results. The outlier case in (a) is discussed in section 4.3.

Key observation. In the results, the proposed metric DMP can report the same ascending order as the ATE metric in most cases. This means our metric can provide similar evaluation results as the ATE.

4.2. Augmented ICL-NUIM dataset

The augmented ICL-NUIM (AIN) dataset [4] is a synthetic dataset with ground-truth surface models and camera trajectories. Experiments are performed to quantitatively compare performance of the proposed metric DMP with ground truth-based metrics: trajectory RMSE and surface mean distance (SMD).

Results are reported in Fig. 2. Reconstruction estimates data M_1-M_5 for evaluation are from SLAM methods and ranked with the SMD metric. **Key observations:** Our metric DMP positively correlates to SMD and Traj. RMSE, which means our metric can provide almost the same ranking results as ground truth-based metrics. To further visualize our metric, we include screen-shots of different 3D models and corresponding DMP values in Fig. 3.

4.3. Ineffective cases

Among all the evaluations on the two datasets, there are rare cases where the proposed metric DMP gives different results from ground truth-based metrics: M_3 in `TUM/frl/desk`, M_2 in `AIN/office 1`, and M_2 in `AIN/office 2`. We believe that it is because these data sequences have most of the space scanned only once without loopy coverage. That is the case where our metric cannot handle perfectly. However, it is simple to avoid this problem by adding overlapping scans.

5. AN APPLICATION OF DMP: LOOP EVALUATION

One important use case of the proposed metric DMP is to evaluate and filter loop detections for SLAM and 3D reconstruction systems. Because most loop detection systems do not have a perfect precision performance, which means false
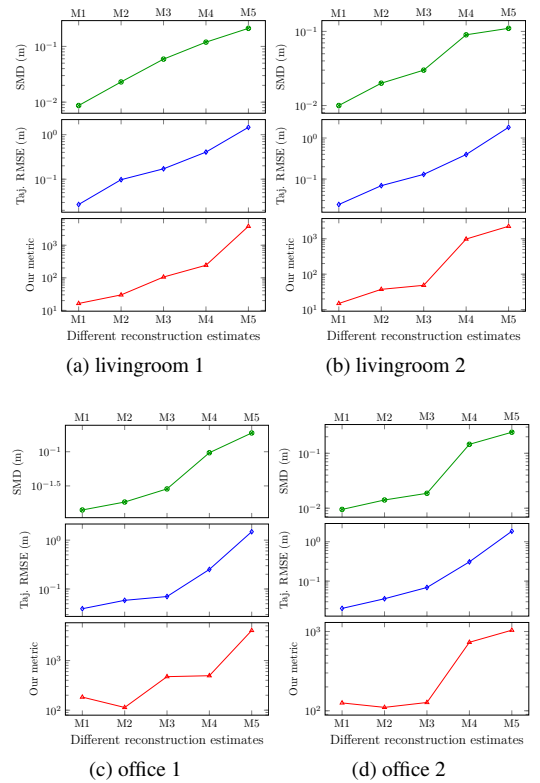

Fig. 2. Comparison of surface mean distance (SMD), trajectory RMSE (Traj. RMSE), and the proposed metric DMP. Our metric DMP positively correlates to SMD and Traj. RMSE, which means our metric can provide almost the same ranking results as ground truth-based metrics. The two outliers can be avoided by overlapping scans.

positive loop detections are inevitable. Meanwhile, SLAM and 3D reconstruction systems tend to be very sensitive to false-positive loops, which are likely to dramatically decrease the 3D model accuracy, as happened in Fig. 3 (e). With the proposed metric, it becomes possible to rank different loops
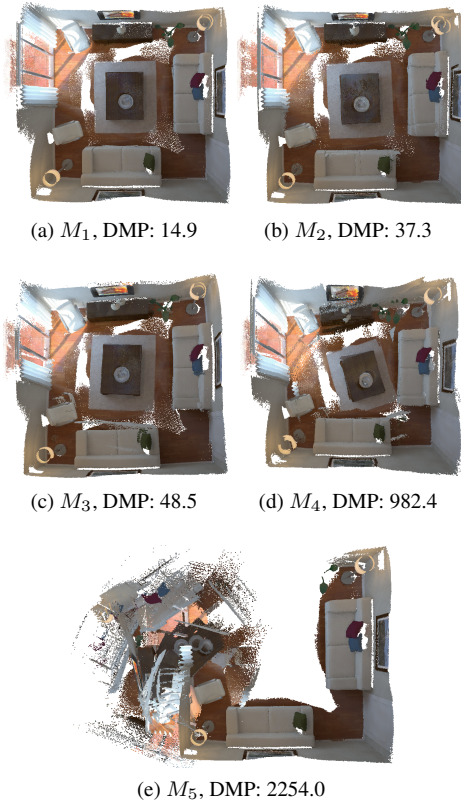
(a) M_1, DMP: 14.9

(b) M_2, DMP: 37.3

(c) M_3, DMP: 48.5

(d) M_4, DMP: 982.4

(e) M_5, DMP: 2254.0

Fig. 3. Visualization of estimated 3D models of the livingroom 2 sequence and their corresponding DMP values. In (a), M_1 leads to the best 3D mapping quality among the five estimates. There are not any noticeable mismatches. In (b), there are mismatches at the upper right-hand corner. The lamp and sofa are not well reconstructed. In (c), the sofa at the bottom of the image is not perfectly matched between frames. In (d), the coffee table at the center is off. (e) shows a collapsed 3D model optimized with a false positive loop.

and filter out the false positive ones.

Again, we run experiments on the Aug ICL-NUIM dataset [4], which also contains a point cloud registration benchmark operating on point cloud scene fragments. [4] also provides a set of point cloud registration results with high recall but low precision. These registration results can be used as a source of loop detections. A subset of the loops that are not connected by ORB-SLAM2 [2] co-visibility graph are selected for our experiments.

In our evaluation, we calculate a DMP value for each of the loops by running the following steps:

1. Perform pose graph optimization from ORB-SALM2 with a single loop added.
2. Fuse a 3D surfels model out of the optimization result.
3. Evaluate DMP.

Finally, the loops are sorted by their DMP values. Precision-Recall curves, shown in Fig. 4, are generated by varying threshold position.

Summary of observations. In the results, we can see the curves all starts from 100% precision. Then the curves keep on high precision values when recall increases. There are a few drop points, which means false positive loops. But the number is limited. The majority of the false-positive loops are at the end of the list reflected by the sharp drops when recall reaches 100%. These mean that the ranking has excellent performance, which proves that the proposed DMP is effective.
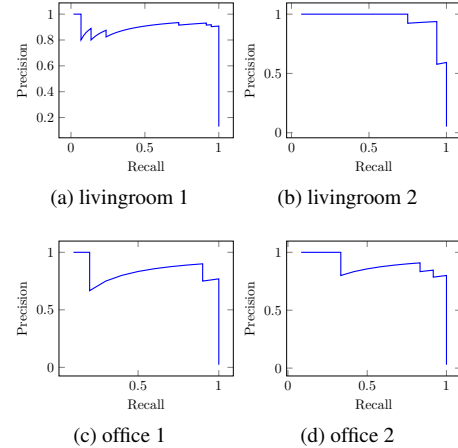

(a) livingroom 1

(b) livingroom 2

(c) office 1

(d) office 2

Fig. 4. Precision and recall curves for ranking results based on the proposed metric DMP. The curves show a very good ranking performance.

6. COMPUTATIONAL PERFORMANCE

Our metric can be efficiently parallelized using OpenGL and CUDA. In our implementation, on an NVIDIA Titan X Pascal, the average evaluation time is $2.7\ s$ for a TUM model and $4.2\ s$ for an Augmented ICL-NUIM model. The speed can be further improved if only a sampled subset of data frames are used for evaluation.

7. CONCLUSION

In this work, we propose a metric, which can work without any ground truth data, for evaluating 3D reconstruction and mapping performance. In our experiments, the proposed metric DMP is benchmarked against ground truth based metrics. Results show that DMP can provide a similar evaluation capability.

The proposed metric not only makes 3D mapping evaluation simpler, but also opens many new opportunities. Our experiments show that it can evaluate loop detection results and lead to good precision-recall performance. We envision that more can be done with this metric, such as self-supervised methods and more available datasets.

8. REFERENCES

[1] Cesar Cadena, Luca Carlone, Henry Carrillo, Yasir Latif, Davide Scaramuzza, Jose Neira, Ian Reid, and John J. Leonard, "Past, Present, and Future of Simultaneous Localization and Mapping: Toward the Robust-Perception Age," *IEEE Transactions on Robotics*, vol. 32, no. 6, pp. 1309–1332, Dec. 2016. 1

[2] Raul Mur-Artal and Juan D. Tardos, "ORB-SLAM2: An Open-Source SLAM System for Monocular, Stereo, and RGB-D Cameras," *IEEE Transactions on Robotics*, vol. 33, no. 5, pp. 1255–1262, Oct. 2017. 1, 2, 4

[3] Guoxiang Zhang and YangQuan Chen, "LoopSmart: Smart Visual SLAM Through Surface Loop Closure," Jan. 2018, arXiv:1801.01572. 1

[4] Sungjoon Choi, Qian-Yi Zhou, and Vladlen Koltun, "Robust reconstruction of indoor scenes," in *Proc. of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2015, IEEE. 1, 2, 3, 4

[5] Adrian Garcea, Jiazhen Zhu, Dominik Van Opdenbosch, and Eckehard Steinbach, "Robust map alignment for cooperative visual SLAM," in *Proc. of the 2018 25th IEEE International Conference on Image Processing (ICIP)*. Oct. 2018, IEEE. 1

[6] Michael Firman, "RGBD datasets: Past, present and future," in *Proc. of the 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. June 2016, IEEE. 1

[7] John McCormac, Ankur Handa, Stefan Leutenegger, and Andrew J. Davison, "SceneNet RGB-D: Can 5M Synthetic Images Beat Generic ImageNet Pre-training on Indoor Segmentation?," in *Proc. of the 2017 IEEE International Conference on Computer Vision (ICCV)*. Oct. 2017, IEEE. 1

[8] Shuran Song, Fisher Yu, Andy Zeng, Angel X. Chang, Manolis Savva, and Thomas Funkhouser, "Semantic scene completion from a single depth image," in *Proc. of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. July 2017, IEEE. 1

[9] Jan Bayer, Petr Čížek, and Jan Faigl, "On construction of a reliable ground truth for evaluation of visual SLAM algorithms," in *Proc. of the 2016 Conference on Planning in Artificial Intelligence and Robotics*, Nov. 2016, vol. 6, pp. 1–5. 1

[10] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias NieBner, "ScanNet: Richly-Annotated 3D Reconstructions of Indoor Scenes," in *Proc. of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. July 2017, IEEE. 1

[11] Binh-Son Hua, Quang-Hieu Pham, Duc Thanh Nguyen, Minh-Khoi Tran, Lap-Fai Yu, and Sai-Kit Yeung, "SceneNN: A scene meshes dataset with aNNotations," in *Proc. of the 2016 Fourth International Conference on 3D Vision (3DV)*. Oct. 2016, IEEE. 1

[12] Harsha Kikkeri, Gershon Parent, Mihai Jalobeanu, and Stan Birchfield, "An inexpensive method for evaluating the localization performance of a mobile robot navigation system," in *Proc. of the 2014 IEEE International Conference on Robotics and Automation (ICRA)*. May 2014, IEEE. 1

[13] John G. Rogers, Jason M. Gregory, Jonathan Fink, and Ethan Stump, "Test your SLAM! the SubT-tunnel dataset and metric for mapping," in *Proc. of the 2020 IEEE International Conference on Robotics and Automation (ICRA)*. May 2020, IEEE. 1

[14] Jrgen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers, "A benchmark for the evaluation of RGB-D SLAM systems," in *Proc. of the 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*. Oct. 2012, IEEE. 1, 2

[15] Thomas Whelan, Stefan Leutenegger, Renato Salas Moreno, Ben Glocker, and Andrew Davison, "Elastic-Fusion: Dense SLAM Without A Pose Graph," in *Proc. of the Robotics: Science and Systems XI*. July 2015, Robotics: Science and Systems Foundation. 1, 2

[16] Ankur Handa, Thomas Whelan, John McDonald, and Andrew J. Davison, "A benchmark for RGB-D visual odometry, 3D reconstruction and SLAM," in *Proc. of the 2014 IEEE International Conference on Robotics and Automation (ICRA)*. May 2014, IEEE. 1

[17] Edwin Olson and Michael Kaess, "Evaluating the performance of map optimization algorithms," in *Proc. of the RSS Workshop on Good Experimental Methodology in Robotics*, 2009, vol. 15. 1