2019 International Conference on Robotics and Automation (ICRA)
Palais des congres de Montreal, Montreal, Canada, May 20-24, 2019

Probabilistic Projective Association and Semantic Guided Relocalization for Dense Reconstruction

Sheng Yang[1], Zheng-Fei Kuang[1], Yan-Pei Cao[1], Yu-Kun Lai[2], and Shi-Min Hu[1]

Abstract— **We present a real-time dense mapping system which uses the predicted 2D semantic labels for optimizing the geometric quality of reconstruction. With a combination of Convolutional Neural Networks (CNNs) for 2D labeling and a Simultaneous Localization and Mapping (SLAM) system for camera trajectory estimation, recent approaches have succeeded in incrementally fusing and labeling 3D scenes. However, the geometric quality of the reconstruction can be further improved by incorporating such semantic prediction results, which is not sufficiently exploited by existing methods. In this paper, we propose to use semantic information to improve two crucial modules in the reconstruction pipeline, namely tracking and loop detection, for obtaining mutual benefits in geometric reconstruction and semantic recognition. Specifically for tracking, we use a novel probabilistic projective association approach to efficiently pick out candidate correspondences, where the confidence of these correspondences is quantified concerning similarities on all available short-term invariant features. For the loop detection, we incorporate these semantic labels into the original encoding through Randomized Ferns to generate a more comprehensive representation for retrieving candidate loop frames. Evaluations on a publicly available synthetic dataset have shown the effectiveness of our approach that considers such semantic hints as a reliable feature for achieving higher geometric quality.**

I. Introduction

A dense 3D representation of scenes with precise geometry and reliable semantic information is the basis for intelligent motion planning and modern VR/AR applications. As the most prevalent dense reconstruction pipeline, the Kinect-Fusion [1] and its subsequent variants for enhancing the capacity [2], [3] and quality [4], [5] have recently been integrated with modern CNN techniques [6], [7] for 2D semantic labeling to develop incremental dense semantic mapping systems such as SemanticFusion [8]. Soon afterwards, by exploring the role of semantics, MaskFusion [9] and DynSLAM [10] propose to use such 2D labeling for pre-segmenting moveable objects (e.g., persons and vehicles), leading to a divide-and-conquer reconstruction strategy for dynamic scenes. Fusion++ [11] also maintains such a list of objects and further formulates an object-level pose graph for a globally consistent map of objects. We argue that semantic hints have more profound effects on SLAM algorithms, by contributing to the two most influential modules in a reconstruction scenario, namely the tracking module and the loop refinement module. In this section, we first review related work on these two modules and then discuss our proposed methods for refinement.

The tracking module is used for estimating sensor poses via frame registration so that measurements and predictions can be projected and fused into the map. The most prevalent solution for dense reconstruction systems [4], [5] is to minimize a joint cost involving both geometric and photometric terms [12], where dense pixel-wise correspondences are efficiently established through projective data association [1]. Recent methods that exploit additional semantic hints can be classified into two major categories: (1) The first category suggests using semantic hints to form higher-level representative entities for the factor graph optimization [13], [14]. Starting from primitive-level geometric entities such as lines [15], [16] and planes [17], [18], coarse [19], [20] and fine [21] representations for instances extracted by recent object detection [22] and segmentation [7] approaches are also introduced nowadays for establishing a richer representation of the involved landmarks. (2) The latter category takes a different view by reformulating the associations between frames and entities into a probabilistic form, where semantic hints are regarded as invariant features for estimating their likelihood. For this type of approaches, the expectation-maximization (EM) [23] algorithm is broadly applied: Bowman et al. [24] use two interconnected stages, as an estimation of discrete data association and continuous optimization over the metric states, for optimizing the constructed factor graphs containing probabilistic associations. Lianos et al. [25] also use such a scheme to determine and minimize semantic reprojection errors for visual odometry. Parkison et al. [26] also follow the strategy to register unorganized point clouds by updating and minimizing weighted residuals from a group of established probabilistic correspondences. Unlike existing methods, we propose to register organized frames, where the efficiency of the original projective association method [1] can be maintained in its probabilistic form to support the real-time application.

The loop refinement module is used for reducing cumulative drifts brought in by sequential tracking. It often consists of two sub-modules, i.e. detection and optimization. The loop detection module finds reliable events of place revisiting for constructing trajectory constraints. These constraints are then used by the latter module for refining the trajectory [5] or deforming the scene [4]. Although the behavior of loop detection is similar to the relocalization problem in the computer vision community, many end-to-end deep learning techniques such as PoseNet [27] and Dual-Stream CNN [28] that predict recovered poses cannot

[1] These authors are with the Department of Computer Science and Technology, Tsinghua University, China. Shi-Min Hu is the corresponding author, shimin@tsinghua.edu.cn

[2] Yu-Kun Lai is with the School of Computer Science & Informatics, Cardiff University, UK. Yukun.Lai@cs.cardiff.ac.uk

978-1-5386-6026-3/19/$31.00 ©2019 IEEE

be directly applied in such an online scenario, because the estimated poses of those continuously recorded frames need to be optimized on the fly, i.e., the trained model should be efficiently evolved during the scanning process. Hence, loop detection methods for such scenarios are mainly based on efficiently retrieving visually similar frames among a database consisting of keyframes. Although some approaches introduce deep neural networks to predict feature descriptors for representing these keyframes [29], [30], the most widely used methods for acquiring such a compact representation are still based on low-level descriptors such as Bag-of-Words (BoW) for monocular images [31] and Randomized Ferns for RGBD frames [32]. By introducing additional semantic hints, such representations can be evolved into a more comprehensive form, leading to more accurate retrieval.

In this paper, we advance those key modules with predicted semantic labels, and form a novel system in which geometric and semantic information can achieve mutual benefits. For the tracking module, we revise the original projective data association [1] into a probabilistic form. Although such a change often comes at the expense of computational efficiency as reported by previous methods [26], we propose a reliable criterion for quickly determining target regions containing most correspondences within a constant time complexity. Furthermore, the probability of such an association is assigned through a joint likelihood considering all available short-term invariant features (geometric, semantic, and photometric). On the other hand, for the loop detection module, we incorporate the predicted semantic labels with the original low-level raw measurements to obtain a compound encoding, which can better represent their corresponding keyframe in the frame retrieval problem. Generally, these improvements can be applied to other dense reconstruction systems [5], and we choose to work on ElasticFusion [4], which is also used as an underlying system by SemanticFusion [8], to test our effectiveness. Experiments on a high-fidelity synthetic dataset [33] with ground-truth trajectory and labeling for quantitative comparisons demonstrate the effectiveness of our improved modules and system.

II. Pipeline Overview

Fig. 1 describes the pipeline of our system based on ElasticFusion [4]. The input to our system is a stream of RGB-D (color and depth) frames. It uses two different processing frequencies (illustrated with green and black arrows) for integrating the computationally intensive 2D semantic labeling task into the real-time processing. Instead of densely predicting semantic labels on each input frame, we use ray-casted labels from the map for sequential tracking.

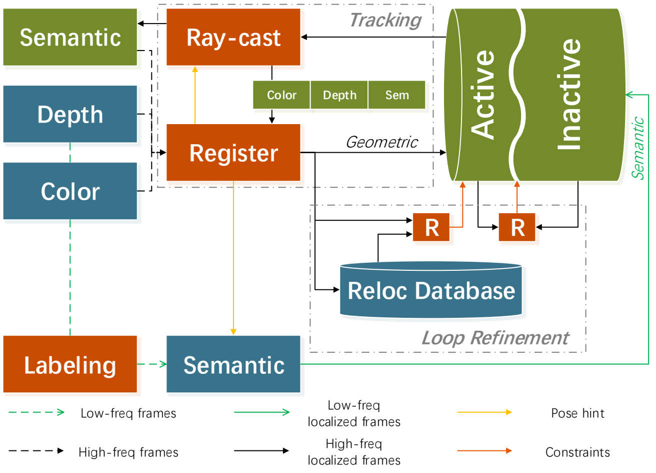


Fig. 1. The data flow of our proposed system. Processing modules, frames, and the frames ray-casted through the map are colored in red, blue and green, respectively. 2D semantic labeling is performed sparsely and fused into the map in a low frequency (green arrows), and the ray-casted semantic labels (top-left) are used for tracking sequential frames and constructing the relocalization database. The red R's stand for registration attempts, and they generate constraints for deforming the reconstructed map.

Map Structure. The reconstructed map is an unordered list of surfels $\mathcal{M}$, where each surfel $\mathcal{M}_k$ stores various types of attributes. In addition to the original attributes such as location p_k, normal n_k, color c_k, and radius r_k, we also store the uncertainty of its semantic and geometric information: The semantic information of $\mathcal{M}_k$ is stored as a normalized histogram of possibilities $L_k = \{l_k^1, \ldots, l_k^M\}$ and updated with a recursive Bayesian rule [8], where M is the number of classes and l_k^m denotes the possibility of $\mathcal{M}_k$ belonging to class m. L_k has 2 bytes for each class along with a maximum likelihood cache, so we allocate 14×2 bytes in total for a surfel in the target dataset [33] containing $M = 13$ classes. Surfels are obtained by mapping pixels in RGB-D frames, so for simplicity, we also use $\mathcal{M}_k$ to represent the corresponding RGB-D pixel. A potential choice for storing more categories is to use the Least-Recently-Used (LRU) strategy due to the sparsity of the possible categories through such predictions. For the geometric uncertainty, a 3×3 covariance matrix $\mathbf{\Sigma}_k$ is used for describing the confidence of position according to multiple noisy depth measurements. For each pixel $\mathcal{M}_z$ on a raw depth frame, its back-projection $p_z \triangleq f(p_z^{uvd}) = \mathbf{K}^{-1} \cdot p_z^{uvd}$ with regard to the camera intrinsics $\mathbf{K}$ maps the point in the image space $p_z^{uvd} = d_z[u_z\ v_z\ 1]^\top$ to the camera space p_z, where (u_z, v_z) and d_z are the coordinates and depth of the pixel in the image. Under the assumption that there is no correlation of noise between different axis directions, the covariance in the image space can be set in a diagonal form $\mathbf{\Sigma}_z^{uvd} = diag(\sigma_u^2, \sigma_v^2, \sigma_d^2)$, so that such uncertainty in the camera space is calculated as:

$$\mathbf{\Sigma}_z = \mathbf{J}_f(p_z^{uvd}) \cdot \mathbf{\Sigma}_z^{uvd} \cdot \mathbf{J}_f^\top(p_z^{uvd}), \qquad (1)$$

where $\mathbf{J}_f$ is the Jacobian matrix of the back-projection function $f(p_z^{uvd})$. For rasterized depth pixels on a frame, we assume $\sigma_u = \sigma_v = 0.5$ and σ_d as concluded by Handa et al. [34]. Then, once a new observation $\mathcal{M}_z$ is added to $\mathcal{M}_k$ with an estimated pose $\mathbf{T}_z = [\mathbf{R}_z|\mathbf{t}_z]$, we use $\mathbf{\Sigma}_k{}^{-1} \leftarrow \mathbf{\Sigma}_k{}^{-1} + \mathbf{R}_z\mathbf{\Sigma}_z{}^{-1}\mathbf{R}_z^\top$ to update its uncertainty according to the Gaussian mixture model [35]. In total, we use 112 bytes (with 64 extra bytes) for each surfel. Through ray-casting, these surfels are rendered as small plain disks with regard to their radius and normal for subsequent registration.

Tracking module. By ray-casting the reconstructed dense map into organized images, all these three types of registration (frame-to-model, frame-to-frame, and model-to-model) can be performed in a unified correspondence search manner. Instead of the original projective association [1], we propose to use a novel probabilistic projective association (Sec. III) strategy for constructing correspondences and registering images. Specifically, in the tracking module when performing frame-to-model registration, the semantic labels for the input frame are iteratively refined with the estimated pose and ultimately finalized after pose estimation before being sent to the loop refinement module.

Loop Refinement module. The other two types of registration, i.e., frame-to-frame and model-to-model in the original ElasticFusion [4], are used for verifying candidate global and local loops, respectively. For detecting global loops, the proposed system maintains a database consisting of historical keyframes, and a new frame will trigger the verification when a similar keyframe is found. For detecting local loops, the surfels in the map are split into two types as either 'active' or 'inactive' by their last update time, and the verification is performed between these two ray-casted frames to detect and recover possible misalignments caused by cumulative drifts of sequential tracking. We incorporate higher-level semantic features into the original encoding through Randomized Ferns [32] to obtain a representative code of each keyframe for more effective retrieval (Sec. IV). In addition, these verifications are based on the joint likelihood considering all available short-term invariant features rather than the original ICP residual that only assesses geometric convergence (Sec. V). For how to apply these established constraints for scene deformation, we refer readers to the original ElasticFusion [4] for their implementation details.

III. Probabilistic Projective Association

Given a source frame and a target frame, the original projective data association [1] locates at most one corresponding pixel $\mathcal{M}_j$ for each source pixel $\mathcal{M}_i$ as the nearest neighbor of its reprojection on the target image domain, where the reprojection $p'_i = d'_i[u'_i\ v'_i\ 1]^\top$ is calculated as:

$$p'_i = f'(\mathbf{T} \cdot f(p_i^{uvd})), \tag{2}$$

where $f' \triangleq f^{-1}$ and $\mathbf{T} = [\mathbf{R}|\mathbf{t}]$ is the estimated relative transformation for transforming points in the source camera space to the target camera space. If such a candidate pixel $\mathcal{M}_j$ on the target frame p_j^{uvd} is picked, a post verification based on the difference of their depth values d'_i and d_j by a hand-tuned threshold is required to reject unsuitable correspondences.

The main advantage of this reprojection strategy is that its time complexity is independent of the total number of pixels, which outperforms other correspondence searching data structures such as Kd-trees, but it also suffers from noise of depth measurements and may result in erroneous correspondences, especially when commodity depth sensors are used. To address this, we instead formulate correspondence search into a probabilistic form taking into account possible sensor noise. As shown in Fig. 2, we replace the original projective data association [1] (red) with our new probabilistic projective association (green), where the probabilities of such correspondences are quantified through the joint likelihood based on available pixel features. We first present the definition of the probability of such associations, and then derive the criterion for the candidate region (orange) for quickly locating possible candidates.

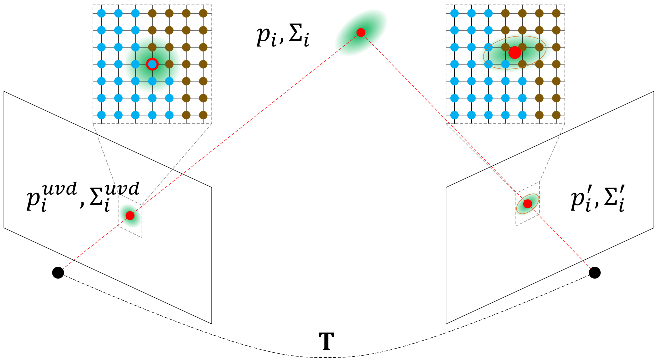


Fig. 2. Schematic of probabilistic projective association in our tracking module. The original projective association [1] is shown in red, while our proposed probabilistic association shown in green uses a region considering possible sensor noise to find multiple candidate correspondences, with the confidence of each correspondence assigned according to the similarity of the appearance between pixels (cyan and brown represent two different semantic labels).

Under the assumption that geometric, semantic and color measurements are independent, such a joint probability $P_{i,j}$ between the source pixel $\mathcal{M}_i$ and its candidate corresponding pixel $\mathcal{M}_j$ in the target frame is defined as:

$$P_{i,j} = P_{geo}(p_i, p_j; \mathbf{\Sigma}_{i,j}) \cdot P_{sem}(L_i, L_j) \cdot P_{clr}(c_i, c_j; \lambda_c^2), \tag{3}$$

where P_{geo}, P_{sem}, and P_{clr} are similarities according to the different types of measurements. The geometric similarity P_{geo} is calculated as:

$$P_{geo}(p_i, p_j; \mathbf{\Sigma}_{i,j}) \propto \exp(-\frac{1}{2}\|p_i - p_j\|^2_{\mathbf{\Sigma}_{i,j}}) \triangleq w_{geo}, \tag{4}$$

where $\|\mathbf{x}\|^2_{\mathbf{\Sigma}} \triangleq \mathbf{x}^\top \mathbf{\Sigma}^{-1} \mathbf{x}$ is the squared Mahalanobis distance with covariance matrix $\mathbf{\Sigma}$, and $\mathbf{\Sigma}_{i,j}^{-1} = \mathbf{\Sigma}_i^{-1} + \mathbf{R}\mathbf{\Sigma}_j^{-1}\mathbf{R}^\top$ again through the Gaussian mixture model [35]. The semantic labeling similarity P_{sem} is defined as the similarity of their normalized histograms:

$$P_{sem}(L_i, L_j) \propto \sum_{x \in \{1 \ldots M\}} l_i^x \cdot l_j^x \triangleq w_{sem}, \tag{5}$$

and the color similarity P_{clr} is similarly defined by the difference of colors under the assumption of normally distributed uncertainty:

$$P_{clr}(c_i, c_j; \lambda_c^2) \propto \exp(-\frac{1}{2}\|c_i - c_j\|^2_{\lambda_c^2 \mathbf{I}}) \triangleq w_{clr}, \tag{6}$$

where a parameter λ_c is introduced to set the confidence for the consistency of color measurements, and $\lambda_c = 0.1$ is used in our experiments.

Based on the joint probability defined above (Equ. 3), for fast rejecting those less possible correspondences, we suggest ignoring those correspondences if their joint probability is relatively small. According to the following relations, we choose to use the geometric similarity for such a fast rejection:

$$\begin{aligned} P_{i,j} &< P_{geo}(p_i, p_j; \mathbf{\Sigma}_{i,j}) \\ &= P_{geo}(p_i, p_j; \mathbf{\Sigma}_i) \cdot P_{geo}(\mathbf{R}^\top p_i, \mathbf{R}^\top p_j; \mathbf{\Sigma}_j) \\ &< P_{geo}(p_i, p_j; \mathbf{\Sigma}_i) \\ &= P_{geo}(p'_i, p_j^{uvd}; \mathbf{\Sigma}'_i), \end{aligned} \tag{7}$$

where $\mathbf{\Sigma}'_i$ is the reprojected covariance on the target frame shown in Fig. 2 and we use its first order approximation $\mathbf{\Sigma}'_i \approx \mathbf{J}_{f'} \cdot \mathbf{R}^\top \mathbf{\Sigma}_i \mathbf{R} \cdot \mathbf{J}_{f'}^\top$ for computing. Since such an approximation only relies on the geometry information from the source frame, we thus can define a condition as $w_{geo}(p'_i, p_j^{uvd}; \mathbf{\Sigma}'_i) < \lambda_p$ for fast rejecting less possible correspondences (with $\lambda_p = 0.1$ in our implementation). Then, the rejection region $\Omega_i \subset \mathbb{R}^2$ can be solved by the following problem with the condition represented in the logarithm form:

$$\begin{aligned} &\text{find } \Omega_i \text{ that } \forall (u_j, v_j) \in \Omega_i, \forall d_j \in (\lambda_n, \lambda_f), \\ &\text{s.t. } \|p'_i - d_j [u_j \ v_j \ 1]^\top\|^2_{\mathbf{\Sigma}'_i} > -2 \log \lambda_p, \end{aligned} \tag{8}$$

where the condition can be treated as a quadratic function of d_j for solving for the image domain Ω_i. $\lambda_n = 0.0, \lambda_f = 4.0$ are assigned as the possible scanning range of the target frame.

Finally, we traverse each source pixel $\mathcal{M}_i$ and its candidate corresponding pixels $\mathcal{M}_j$ that lies within such an estimated range Ω_i to compute the weighted residual and solve the relative transformation $\mathbf{T}$ as:

$$\underset{\mathbf{T}}{\operatorname{argmin}} \sum_i \sum_{j \in \Omega_i} w_{i,j} \|p_j - \mathbf{T} \cdot p_i\|^2_{\mathbf{\Sigma}_{i,j}}, \tag{9}$$

where these joint weights $w_{i,j}$ are calculated as the product of w_{geo}, w_{sem}, and w_{clr} according to Equ. 3. These weights are held constant during each inner iteration.

IV. Semantic Encoding for Keyframe Retrieval

Reliable and compact codes for keyframes are essential for efficiently retrieving similar poses. As used in various reconstruction systems, the original encoding strategy based on the Randomized Ferns [32] defines 4-channel binary tests at randomized but fixed image locations (N in total) to generate a compact code for each keyframe X. Intuitively by integrating higher-level labeling results with low-level color and depth tests, a richer and more effective code can be acquired for better measuring similarities between keyframes. We expand the original code $b_X = [b_{X_1} \dots b_{X_N}] \in \mathbb{B}^{4N}$ by adding the maximum likelihood of labeling denoted as q_{X_*} of each location, and thus form a revised code $b'_X = [(b_{X_1}, q_{X_1}) \dots (b_{X_N}, q_{X_N})] \in \mathbb{B}^{(4 + \lceil \log_2 M \rceil)N}$. Thus, the dissimilarity measured by both pixel and semantic differences of two encodings b'_I and b'_J can be calculated as:

$$Dis(b'_I, b'_J) = \frac{1}{N} \sum_{y=1}^{N} (b_{I_y} \equiv b_{J_y} \wedge q_{I_y} \equiv q_{J_y}), \tag{10}$$

where the equivalent operator $\equiv$ returns 0 if two blocks are identical and 1 otherwise. We follow the original ID look-up table as an efficient structure for searching and maintaining the keyframe database.

V. Pose Evaluation through Joint-Likelihood

Registration attempts may fail or result in unexpected convergence. The original ElasticFusion [4] does not explicitly address such problems and chooses to use a hand-tuned threshold on its ICP residual reflecting the geometric convergence for verification. However, a better strategy concerning all available features can be applied in our scenario, based on the observation that a reasonable relative transformation should have geometric, semantic, and photometric features converged simultaneously. Hence, we choose to verify the independent pixel-wise maximum joint-confidence rather than only rely on the geometric residual for evaluating estimated poses, and the score $Scr(I, J; \mathbf{T})$ of a candidate relative transformation $\mathbf{T}$ for the source frame I and the target frame J is defined as:

$$Scr(I, J; \mathbf{T}) = (\prod_{i \in I} \max_{j \in \Omega_i} \ w_{i,j})^{1/|I|}, \tag{11}$$

where $|I|$ stands for the total number of those valid pixels $\mathcal{M}_i$ on the source frame I, and the total score is calculated as the geometric mean of those individual scores for each $\mathcal{M}_i$. Empirically, we use a threshold λ_v for accepting the estimated pose if $Scr(I, J, \mathbf{T}) > \lambda_v$. This parameter and its performance in comparison to the original threshold are further discussed in Sec. VI-F.

VI. Experiments and Evaluations

A. Evaluation Dataset

We perform experiments on a publicly available dataset: SceneNetRGBD [33]. This dataset contains a rendering engine and randomized scenes, where random trajectories can be automatically generated for rendering photorealistic videos as well as ground truth depth measurements and instance labels. Also, it provides a collected dataset for training 2D semantic labeling models. We generated 53,787 frames of 42 trajectories from 7 scenes (6 trajectories each), where the lighting and textures of each scene remain unchanged in its trajectories for testing relocalization algorithms. We sorted these scenes ascendingly through their density of objects and denote them from S-1 to S-7. For rendering color images, we keep its original configuration to synthesize visual artifacts such as motion blur, while for depth images, we follow the noise model proposed by Handa et al. [34] to synthesize realistic noisy depth scans. These ground-truth labelings of instances are converted into 13 classes.

B. System Implementation

Network Training. We uniformly sampled a subset (168,650 images as 1/30) of the publicly available training set for our training. We apply the DeepLabv3 [7] for training a 2D semantic labeling model, and the network was trained using the ‘poly’ learning rate policy with the base learning

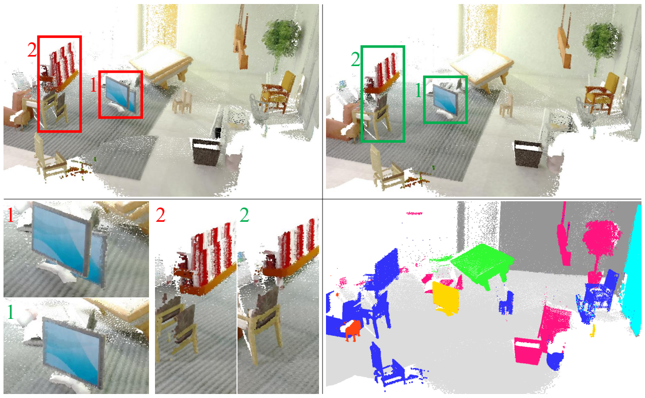


Fig. 3. An example of the dense reconstruction of ours (Top-Right) in comparison to the ElasticFusion (Top-Left) with close views (Bottom-Left). We also show the 3D labeling result according to the incremental semantic fusion on the Bottom-Right.

rate set to 0.007 and power to 0.9. Momentum and weight decay were set to 0.9 and 0.0002 respectively. We used a mini-batch size of 10 and trained the network for a total of 300K iterations for about three days on an NVIDIA Titan Xp. The performance of the network is evaluated with mean Intersection over Union (IoU) of 44.52% and pixel accuracy of 80.19% on its publicly available validation set.

Parameters. For registering frames, (λ_n, λ_f) are used for picking reliable depth measurements due to the reported accuracy of depth cameras [34]. Decreasing λ_p may accept more less-important correspondences in the iteration, but we found that the chosen value is suitable for maintaining the quality of registration while reducing the computational cost. λ_c can be assigned according to the quality of color frames, and it is necessary to increase the value to reduce the sensitivity of color features when scanning scenes with changing lighting conditions. For the threshold λ_v used for pose evaluation, we demonstrate its effectiveness in comparison to the original threshold on ICP residual in Sec. VI-F.

Hardware Configurations. All experiments were performed on a desktop PC with an i7-6850K CPU, 32 GB RAM, and two NVIDIA Titan Xp GPUs. We use one GPU for continuously predicting poses and the other for the main system. The 2D labeling task takes 104ms on average for each frame (~10Hz), and the main system requires 3.42GB graphics memory.

C. Reconstruction Quality

For estimating the geometric reconstruction quality, we calculate the RMSE (root mean square error) of distances for every surfel to its nearest ground-truth surface. For comparison, we use two publicly available reconstruction systems ElasticFusion [4] and the state-of-the-art BundleFusion [5]. Since SemanticFusion [8] directly applies ElasticFusion [4] for its geometric reconstruction, they will obtain the same geometric quality under the same configuration. We use the default parameters as suggested by their paper, and $\lambda_f = 4.0$ to cut off far noisy measurements. Results of geometric reconstruction are listed in Tab. I, where the RMSE of each scene is averaged over all the trajectories. Fig. 3 and our supplementary video also show visualized examples. According to our results, the original projective data association (Equ. 2) used by ElasticFusion [4] and BundleFusion [5] has generated false correspondences and thus influence the final reconstruction quality. Although BundleFusion adds sparse features during the registration, it failed to precisely register frames in some textureless scenes. On the other hand, the effect of such semantic hints on scenes with complicated object arrangements is more evident according to the difference between ElasticFusion and ours.

	S-1	S-2	S-3	S-4	S-5	S-6	S-7
ElasticFusion	10.5	**15.7**	17.9	19.6	23.7	24.2	28.2
BundleFusion	12.5	21.9	21.2	18.9	22.8	20.0	16.3
Ours	**10.0**	15.9	**16.6**	**16.7**	**22.0**	**19.4**	**16.2**

TABLE I

STATISTICS OF GEOMETRIC QUALITY FOR DIFFERENT METHODS EVALUATED IN RMSE (MILLIMETERS).

D. Performance of Registration

To further test the performance of our proposed registration algorithm (Sec. III), we randomly select 20,000 pairs whose ground-truth relative transformation between its two frames is less than 0.05m and 5° (since these registration methods are mainly used for fine-level local registration) to construct the test set. Our registration method is performed in comparison with the original point-to-plane ICP used in KinectFusion [1] and the RGBD odometry [12] used in ElasticFusion [4]. As a reference, if the estimated relative transformation has less than both 1cm translational and 1° rotational error, we consider the registration as successful. The results of all these methods are shown in Tab. II. We compare 3 semantic labeling sources, namely RGBD-CNN pre-trained by SemanticFusion [8], our trained model based on DeepLabv3 [7], and the ground truth generated from rendering, for assessing the influence of the quality of the input labels. Although the semantic quality will affect the accuracy of tracking, using currently available labeling strategies yields better results than the classical tracking algorithms. DeepLabv3 works better than RGBD-CNN and the performance is reasonably close to using ground-truth labels. Meanwhile, our probabilistic association method considers more corresponding pixels but can still maintain real-time efficiency (less than 33ms for 30Hz RGB-D streams).

	Err-T.(mm)	Err-R.$(^\circ)$	Suc.(%)	Avt.(ms)
Point-to-Plane ICP	19.6	0.673	45.6	**0.96**
RGBD Odometry	5.66	0.246	81.9	3.13
Ours (RGBD-CNN)	5.45	0.236	84.8	12.93
Ours (DeepLabv3)	5.42	0.229	85.4	12.93
Ours (Ground-Truth)	**5.36**	**0.203**	**87.2**	12.92

TABLE II

STATISTICS OF THE REGISTRATION PERFORMANCE FOR DIFFERENT METHODS. ERR-T./R. - AVERAGE TRANSLATIONAL/ROTATIONAL ERROR. SUC. - SUCCESS RATE. AVT. - AVERAGE RUNNING TIME.

E. Performance of Keyframe Encoding

Since the purpose of the encoding strategies discussed in Sec. IV for loop detection is to find spatially close frames based on visual appearance, we use 4 trajectories of each scene to construct their keyframe database and the other 2 trajectories are used for retrieval attempts. As the most common usage of such loop detectors, we retrieve the best candidate keyframe, and record its (1) initial translational/rotational difference after retrieval, and (2) final translational/rotational error after frame-to-frame registration through the RGBD Odometry [12]. Also, if a recovery within $\langle$1cm,1°$\rangle$ difference to its ground-truth pose is obtained, we treat it as successful. The detailed comparison results are listed in Tab. III for relocalization. Our method performs both higher success rate and lower average difference of pose in vast majority of cases, demonstrating the effectiveness of such incorporation. It is worth mentioning that the reconstructed databases of keyframes are also enlarged, since more dissimilarity, i.e., diversity, is recorded in consideration of such a higher-level feature.

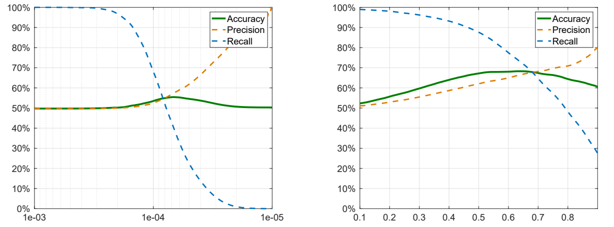
	Method	Size	Suc.	Init-T.(mm)	Init-R.(°)	Final-T.(mm)	Final-R.(°)
S-1	R-Ferns	586	102	**32.6**	2.10	10.5	0.65
	Ours	636	102	33.2	**1.97**	**9.0**	**0.60**
S-2	R-Fern	1026	108	43.2	1.10	7.1	0.46
	Ours	1162	**114**	**35.6**	**0.96**	**7.0**	**0.45**
S-3	R-Fern	508	51	47.3	1.27	4.8	0.21
	Ours	590	**57**	**42.6**	**1.18**	**4.6**	**0.18**
S-4	R-Fern	438	95	44.1	1.22	**14.6**	0.74
	Ours	614	**123**	**35.4**	**0.95**	15.6	**0.58**
S-5	R-Fern	356	61	43.4	2.36	9.0	0.67
	Ours	424	**74**	**35.4**	**2.00**	**8.8**	**0.61**
S-6	R-Fern	444	98	41.7	1.97	**15.7**	0.81
	Ours	546	**108**	**35.1**	**1.71**	16.1	**0.77**
S-7	R-Fern	404	49	42.8	1.23	8.4	**0.43**
	Ours	480	**59**	**38.2**	**1.08**	8.4	0.45

TABLE III

STATISTICS OF THE RELOCALIZATION PERFORMANCE. SIZE - TOTAL NUMBER OF KEYFRAMES. SUC. - SUCCESSFUL RECOVERY. INIT-T./R. - INITIAL TRANSLATIONAL/ROTATIONAL DIFFERENCE. FINAL-T./R. - FINAL TRANSLATIONAL/ROTATIONAL DIFFERENCE.

F. Performance of Pose Evaluation

For our proposed pose evaluation criterion (Sec. V) based on the joint-likelihood, we compare the performance of the threshold λ_c for the joint-likelihood and the original threshold for ICP residual. based on the previously constructed 20,000 pairs of frames, and the $\langle$1cm,1°$\rangle$ criterion, we use the 6-DOF normal distribution to randomly generate positive and negative relative poses with approximately equal numbers (nearly 50,000 for each). Then, we test the performance of different configurations of parameters, and record their accuracy, precision and recall as shown in Fig. 4. As shown in Fig. 4(a), the best accuracy is around 55% obtained with the geometric threshold set to 10^{-4}, which is also chosen by many reconstruction approaches [4] for constructing loop constraints. However, the performance of our proposed likelihood evaluation through the threshold λ_v (Fig. 4(b)) can reach almost 69% accuracy, demonstrating that the criterion based on the consistency tracking is a suitable choice for assessing estimated poses.



(a) By thresholding residuals [4]. (b) By thresholding likelihood λ_v.

Fig. 4. Performance (y-axis) of different pose evaluation criteria w.r.t. the chosen threshold (x-axis).

G. Limitations

Our proposed method suffers from the following limitations. First, our system does not construct distributions of color measurements. Although it can be inferred similarly using a Gaussian mixture model, the uncertainty of raw measurements is hard to be quantified and such models are not applicable to specular and highlight surface regions. Second, although our probabilistic form of associations has considered different types of features, we only use the geometric information for computing residuals during registration, where a better choice would be jointly considering all available residuals together.

VII. Conclusion

In this paper, we presented a real-time dense mapping system which uses the predicted 2D semantic labeling results for enhancing the geometric reconstruction quality. For registering frames and models, we propose a probabilistic projective data association approach that constructs possible correspondences between pixels, where the confidence of an association is quantified by the joint likelihood considering geometric, semantic and photometric information together. Such joint likelihood is also used for evaluating the validity of an estimated transformation. For detecting candidate loops through a database of keyframes, we incorporate these predicted labels into the original coding to obtain a more effective representation that better retrieves candidates for loop refinement.

Acknowledgment

We would like to thank the reviewers for their constructive comments and the authors of the SceneNetRGBD dataset for their kind suggestions of its usage. This work was supported by the National Key Technology R&D Program (Project Number 2017YFB1002604), the Joint NSFC-DFG Research Program (project number 61761136018), the Natural Science Foundation of China (Project Number 61521002), and Research Grant of Tsinghua-Tencent Joint Laboratory for Internet Innovation Technology.

References

[1] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohi, J. Shotton, S. Hodges, and A. Fitzgibbon, "KinectFusion: Real-time dense surface mapping and tracking," in *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, 2011, pp. 127–136.

[2] T. Whelan, M. Kaess, M. Fallon, H. Johannsson, J. Leonard, and J. McDonald, "Kintinuous: Spatially extended KinectFusion," in *RSS Workshop on RGB-D: Advanced Reasoning with Depth Cameras*, 2012.

[3] M. Nießner, M. Zollhöfer, S. Izadi, and M. Stamminger, "Real-time 3D reconstruction at scale using voxel hashing," *ACM Transactions on Graphics (TOG)*, vol. 32, no. 6, pp. 169:1–169:11, 2013.

[4] T. Whelan, S. Leutenegger, R. F. Salas-Moreno, B. Glocker, and A. J. Davison, "ElasticFusion: Dense SLAM without a pose graph," in *Robotics: Science and Systems*, 2015.

[5] A. Dai, M. Nießner, M. Zollhöfer, S. Izadi, and C. Theobalt, "BundleFusion: Real-time globally consistent 3D reconstruction using on-the-fly surface reintegration," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 3, pp. 24:1–24:18, 2017.

[6] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1520–1528.

[7] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 40, no. 4, pp. 834–848, 2018.

[8] J. McCormac, A. Handa, A. Davison, and S. Leutenegger, "SemanticFusion: Dense 3D semantic mapping with convolutional neural networks," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2017, pp. 4628–4635.

[9] M. Rünz and L. Agapito, "MaskFusion: Real-time recognition, tracking and reconstruction of multiple moving objects," *arXiv preprint arXiv:1804.09194*, 2018.

[10] I. A. Bârsan, P. Liu, M. Pollefeys, and A. Geiger, "Robust dense mapping for large-scale dynamic environments," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2018.

[11] J. McCormac, R. Clark, M. Bloesch, A. Davison, and S. Leutenegger, "Fusion++: Volumetric object-level slam," in *International Conference on 3D Vision (3DV)*. IEEE, 2018, pp. 32–41.

[12] T. Whelan, M. Kaess, H. Johannsson, M. Fallon, J. J. Leonard, and J. Mcdonald, "Real-time large-scale dense RGB-D SLAM with volumetric fusion," *The International Journal of Robotics Research (IJRR)*, vol. 34, no. 4–5, pp. 598–626, 2015.

[13] G. Grisetti, R. Kummerle, C. Stachniss, and W. Burgard, "A tutorial on graph-based SLAM," *IEEE Intelligent Transportation Systems Magazine*, vol. 2, no. 4, pp. 31–43, 2010.

[14] R. Kümmerle, G. Grisetti, H. Strasdat, K. Konolige, and W. Burgard, "g2o: A general framework for graph optimization," in *International Conference on Robotics and Automation (ICRA)*, 2011, pp. 3607–3613.

[15] A. Pumarola, A. Vakhitov, A. Agudo, A. Sanfeliu, and F. Moreno-Noguer, "PL-SLAM: Real-time monocular visual SLAM with points and lines," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2017, pp. 4503–4508.

[16] D. G. Kottas and S. I. Roumeliotis, "Efficient and consistent vision-aided inertial navigation using line observations," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2013, pp. 1540–1547.

[17] L. Ma, C. Kerl, J. Stückler, and D. Cremers, "CPA-SLAM: Consistent plane-model alignment for direct RGB-D SLAM," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2016, pp. 1285–1291.

[18] Y. Shi, K. Xu, M. Niessner, S. Rusinkiewicz, and T. Funkhouser, "PlaneMatch: Patch coplanarity prediction for robust RGB-D reconstruction," *European Conference on Computer Vision (ECCV)*, 2018.

[19] S. Y. Bao, M. Bagra, Y.-W. Chao, and S. Savarese, "Semantic structure from motion with points, regions, and objects," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 2703–2710.

[20] P. Gay, V. Bansal, C. Rubino, and A. Del Bue, "Probabilistic structure from motion with objects (PSFMO)," in *IEEE International Conference on Computer Vision (ICCV)*, vol. 2, 2017.

[21] N. Fioraio and L. Di Stefano, "Joint detection, tracking and mapping by semantic bundle adjustment," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 1538–1545.

[22] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2980–2988.

[23] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the royal statistical society. Series B (methodological)*, pp. 1–38, 1977.

[24] S. L. Bowman, N. Atanasov, K. Daniilidis, and G. J. Pappas, "Probabilistic data association for semantic SLAM," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2017, pp. 1722–1729.

[25] K.-N. Lianos, J. L. Schönberger, M. Pollefeys, and T. Sattler, "VSO: Visual semantic odometry," 2018.

[26] S. A. Parkison, L. Gan, M. G. Jadidi, and R. M. Eustice, "Semantic iterative closest point through expectation-maximization," 2018.

[27] A. Kendall, M. Grimes, and R. Cipolla, "PoseNet: A convolutional network for real-time 6-DOF camera relocalization," in *IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 2938–2946.

[28] R. Li, Q. Liu, J. Gui, D. Gu, and H. Hu, "Indoor relocalization in challenging environments with dual-stream convolutional neural networks," *IEEE Transactions on Automation Science and Engineering*, vol. 15, no. 2, pp. 651–662, 2018.

[29] Y. Xia, J. Li, L. Qi, and H. Fan, "Loop closure detection for visual SLAM using PCANet features," in *IEEE International Joint Conference on Neural Networks*, 2016, pp. 2274–2281.

[30] X. Gao and T. Zhang, "Unsupervised learning to detect loops using deep neural networks for visual SLAM system," *Autonomous robots*, vol. 41, no. 1, pp. 1–18, 2017.

[31] D. Gálvez-López and J. D. Tardos, "Bags of binary words for fast place recognition in image sequences," *IEEE Transactions on Robotics*, vol. 28, no. 5, pp. 1188–1197, 2012.

[32] B. Glocker, J. Shotton, A. Criminisi, and S. Izadi, "Real-time RGB-D camera relocalization via randomized ferns for keyframe encoding," *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, vol. 21, no. 5, pp. 571–583, 2015.

[33] J. McCormac, A. Handa, S. Leutenegger, and A. J. Davison, "SceneNet RGB-D: Can 5m synthetic images beat generic ImageNet pre-training on indoor segmentation," in *IEEE International Conference on Computer Vision (ICCV)*, vol. 4, 2017.

[34] A. Handa, T. Whelan, J. McDonald, and A. J. Davison, "A benchmark for RGB-D visual odometry, 3D reconstruction and SLAM," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2014, pp. 1524–1531.

[35] B. G. Lindsay, "Mixture models: theory, geometry and applications," in *NSF-CBMS regional conference series in probability and statistics*, 1995, pp. i–163.