2021 IEEE International Conference on Robotics and Automation (ICRA 2021)
May 31 - June 4, 2021, Xi'an, China

Point Set Registration With Semantic Region Association Using Cascaded Expectation Maximization

Lan Hu, Jiaxin Wei, Zhanpeng Ouyang and Laurent Kneip
Mobile Perception Laboratory, ShanghaiTech University
{hulan, weijx, ouyangzhp, lkneip}@shanghaitech.edu.cn

***Abstract*— We introduce a new solution to point set registration, a fundamental geometric problem occurring in many computer vision and robotics applications. We consider the specific case in which the point sets are segmented into semantically annotated parts. Such information may for example come from object detection or instance-level semantic segmentation in a registered RGB image. Existing methods incorporate the additional information to restrict or re-weight the point-pair associations occurring throughout the registration process. We introduce a novel hierarchical association framework for a simultaneous inference of semantic region association likelihoods. The formulation is elegantly solved using cascaded expectation-maximization. We conclude by demonstrating a substantial improvement over existing alternatives on open RGBD datasets.**

I. Introduction

The alignment of two point sets is a fundamental geometric problem that occurs in many computer vision and robotics applications. In computer vision, the technique is used to stitch together partial 3D reconstructions in order to form a more complete model of an object or environment [34]. In robotics, point set registration is an essential ingredient to simultaneous localization and mapping with affordable consumer depth cameras [31] or 3D Lidars [8].

The dominant solution to 2D and 3D rigid registration is the ICP algorithm [3], [43] and its variants. ICP is conceptually simple, easy to use, and has good performance. ICP algorithm does not require initial correspondences, its variants will be found as part of the registration process which alternates between geometric correspondence establishment (e.g. by simple nearest neighbour search) and Procrustes alignment. The convergence of the iterative process depends on a sufficiently accurate initial guess about the relative transformation. Conversely, without a good initialisation, ICP may easily get trapped in a local minimum, thus producing erroneous estimates in the absence of reliable fault detection.

Gaussian Mixture Alignment (GMA) [12], [29], [38] was introduced to address these shortcomings. GMA involves representing discrete point sets by Gaussian Mixture Models (GMMs), and reformulates the point set registration problem as a minimization of a statistical discrepancy measure between the two corresponding mixture models. The ICP algorithm can be reinterpreted as minimizing an approximation of the KL divergence between Gaussian mixtures, which explains its lack of robustness against missing correspondences caused by partial overlap or occlusions. GMA mitigates the problem of local minima by widening the basin of convergence [22].

The present work aims at a further improvement of the robustness and convergence basin of the point set registration method by utilizing semantic information inferred from images, a technique that has recently shown increasing value in vision-based applications such as for example the localization or mapping part of a SLAM pipeline. Semantic detections can be used as landmarks for localization [16], [15], [37]. Together with pose estimation techniques, multiple registered scans can also be segmented and combined to form a semantic map [28], [27], [33], [30], thereby giving robots a better understanding of their environment.

Our work is inspired by the work of Bowmann et al. [6], who use GMMs to model associations between semantic 3D landmarks and image detections. We propose the addition of a semantic region association layer to the point set alignment problem, again modelled by a GMM. Semantic regions are for example obtained from an instance-level segmentation of RGB images generated by a neural network. The regions are defined by a mask and a label. To summarize, we perform point set registration by a hierarchical model with semantic region association on the outside, and point-pair association on the inside.

Our contributions are summarized as follows:

- We exploit a GMM for semantic region association, and solve the registration problem with hierarchical association using cascaded expectation maximization.
- We analyse the performance of the algorithm for four different point set distance metrics.
- We test on many open RGBD datasets, and demonstrate a clear improvement in point set registration performance compared to existing both traditional and semantically augmented alignment methods. We also compare against more modern deep learning [40] as well as feature-based [35], global registration alternatives.

II. Related Work

In general terms, point set registration consists of an optimization problem over a rigid body transformation that minimizes certain residuals between a source point set and

L. Kneip is also with the Shanghai Engineering Research Center of Intelligent Vision and Imaging. The authors would like to thank the funding sponsoredby Natural Science Foundation of Shanghai (grant number: 19ZR1434000).

a target point set. The traditional ICP algorithm [9], [3], [43] does not depend on a prior derivation of point-to-point correspondences, and simply aligns the two sets by iteratively alternating between the two steps of finding nearest neighbours (e.g. by minimising point-to-point distances), and computing the alignment (e.g. using Arun's method [2]). GICP [36] generalizes ICP by modelling the source and target point clouds with Gaussian distributions derived from sample covariances of neighboring points.

Probabilistic formulations of ICP and the iterative nature of the algorithm have led to Expectation Maximization (EM) approaches to the point cloud registration problem [19]. To improve robustness of the algorithm against occlusions and reduced overlap, the method has been extended by outlier rejection [43], [18] or data trimming [10] techniques. Recently, there are some deep learning based solutions [40], [26] showing promising results on some restricted scenarios.

An entire family of alternative approaches relies on the idea of expressing the point sets by GMMs and aligning the latter using GMA. Notable GMA-based techniques for rigid and non-rigid registration are given by the robust point matching algorithm by Chui and Rangarajan [12], the coherent point drift strategy by Myronenko and Song [29], kernel correlation by Tsin and Kanade [38], and the GMMReg algorithm by Jian and Vemuri [22]. GMA is advertised by improved robustness against poor initialisations, noise, and outliers. On the down-side, GMA is typically much slower than ICP.

More recently, there have been lots of results from the deep learning community that enable high quality, instance-level semantic segmentation of a single image [39], [25], [20], [21], [5], [24], [4]. Owing to the availability of large-scale open training datasets with extensive semantic labeling, the number of detectable classes can be in the order of 1000. Semantic information has already been widely incorporated into SLAM frameworks [16], [15], [37], [28], [27], [33], [30]. Recent techniques have discovered the merits of performing simultaneous camera pose refinement and data association between semantic 3D landmarks and image measurements (e.g. object detections, or instance-level segmentations) [16], [6], [23], [15]. However, the problem is commonly addressed in a sequential manner in which we first solve the purely geometric alignment problem, followed by semantic association. Most methods perform joint geometric registration and semantic association only in the concluding refinement step.

There are only few exceptions which directly utilize semantic information during the geometric registration step. [32], [37], [42], [41]. [42] utilizes semantic information as a hard constraint, and only finds nearest neighbors within subsets of identical semantic labels. However, both [42] and [32] still work at the pure point-pair association level. [37] approximates semantic regions by 3D ellipsoids and maximizes the number of matched regions for robust re-localization. [41] uses semantic information of pixels as a guidance for re-localization. [32] is most related to our algorithm. It employs semantic information in EM-ICP [18] as a prior to calculate the weight of each residual term. However, it still limits the inference of associations to the point-pair level.

In contrast to point-level semantic association, our work introduces semantic regions as additional elements for data association. The likelihood of the semantic region associations is again modelled by a GMM, thus leading to a hierarchical, two-layer association model. The semantic region associations can be regarded as latent variables able to improve registration convergence and accuracy.

III. Preliminaries

Notations and problem definition: The inputs to the registration problem consist of two point sets $A = \{a_1, a_2, \cdots a_n\}$ and $B = \{b_1, b_2, \cdots, b_m\}$, where n and m are the number of points A and B, respectively. Let $B' = f(B, \mathbf{x})$ be the point set of B after applying the transformation $T(\mathbf{x})$, where $\mathbf{x}$ are the transformation parameters. The distance $d(a_i, B')$ between the point a_i and the point set B' is then given by

$$d(a_i, B') = \min_{b_j \in B'} d(a_i, b_j), \tag{1}$$

and $d(a_i, b_j)$ furthermore denotes the Euclidean distance between points a_i and b_j.

ICP as a log-likelihood maximization problem: With the above definition of the point-set distance, the objective of the classical ICP algorithm can be written as

$$\arg\min_{\mathbf{x}} \sum_{i=1}^{n} d^2(a_i, f(B, \mathbf{x})). \tag{2}$$

The main idea behind the ICP algorithm consists of alternating between finding the nearest points in the target point set and minimizing the objective function. Using the common assumption of Gaussian noise distribution of the residual terms, the minimization problem can be easily transformed into the log-likelihood maximization problem[18].

$$\begin{aligned} & \arg\max_{\mathbf{x}} \log \prod_i \frac{1}{\sqrt{2\pi}\sigma} \exp \frac{\|d(a_i, f(B,\mathbf{x})) - 0\|^2}{2\sigma^2} \\ = \; & \arg\max_{\mathbf{x}} \log \mathbf{p}(A, \mathcal{I} | B, \mathbf{x}), \end{aligned} \tag{3}$$

where σ represents the standard deviation of the residual error distribution, and $\mathcal{I}$ the association set which indicates the correspondences (defined further in Section IV-B).

Alternative point set distance metrics: Our work will introduce semantic region association into the objective, the likelihood of which will be formulated as a function of a geometric distance metric between two regions. A complete exposition and comparison of point set distance metrics can be found in the literature [1]. Here we only list the following alternatives as they are used in our work:

- Mean Distance Function:

$$d_m(A, B) = d\Big(\frac{1}{n}\sum_i^n a_i, \frac{1}{m}\sum_j^m b_j\Big) \tag{4}$$

- Average Distance Function:

$$d_{\text{avg}}(A, B) = \frac{1}{nm}\sum_i^n \sum_j^m d(a_i, b_j) \tag{5}$$

- Sum of Minimums Distance Function:

$$d_{\text{som}}(A,B) = \frac{1}{2}\big(\frac{1}{n}\sum_{a_i\in A} d(a_i,B) + \frac{1}{m}\sum_{b_i\in B} d(b_j,A)\big) \quad (6)$$

- Hausdorff Distance Function:

$$d_{\text{haus}}(A,B) = \max\big\{\max_{a_i\in A}\{d(a_i,B)\} + \max_{b_j\in B}\{d(b_j,A)\}\big\} \quad (7)$$

IV. Registration with Semantic Region Association

Assuming that the point set is for example measured by a depth camera, it can be divided into semantically labelled regions by applying instance-level segmentation (or object detection) in the RGB image. Let $\mathcal{A} = \{\mathbf{A}^i\}|_{\{i=1,\cdots,N\}} = \{(A^i, s_i)\}|_{\{i=1,\cdots,N, s_i\in\mathcal{S}\}}$ and $\mathcal{B} = \{\mathbf{B}^j\}|_{\{j=1,\cdots,M\}} = \{(B^j, s_j)\}|_{\{j=1,\cdots,M, s_j\in\mathcal{S}\}}$ be the sets of semantic regions in the source data and target data. N and M are the number of regions, $\mathbf{A}^i$ contains the 3D points A^i and the corresponding semantic label s_i of the i-th region, $\mathbf{B}^j$ is defined analogously, and $\mathcal{S}$ is the set of possible semantic labels. $A^i \in R^{N_i\times 3}$ and $B^j \in R^{N_j\times 3}$, so N_i and N_j denote the number of 3D points in A^i and B^j, respectively.

Our core idea consists of performing hierarchical data association and additionally infer the unknown correspondences between the regions of $\mathcal{A}$ and $\mathcal{B}$. Let $\mathcal{D} = \{D_k\}|_{k=1,\cdots,K} = \{(\alpha_k,\beta_k)\}|_{k=1,\cdots,K}$ denote the set of all possible region associations, K is the number of all associations. Each element $D_k = (\alpha_k, \beta_k)$ denotes a potential association between the α_k-th semantic region in the source data and the β_k-th semantic region in the target data. We consider all possible pair-wise region associations for which the semantic labels $s \in \mathcal{S}$ are identical, and perform soft data association [6].

Our goal remains to infer the pose which transforms the source points to the target points. However, the maximum-likelihood objective is now changed in that it also involves the simultaneous estimation of the likelihood of each possible association D_k, i.e.

$$\mathbf{x}^* = \arg\max_{\mathbf{x}} \mathbf{p}(\mathcal{A}, \mathcal{D}|\mathcal{B}, \mathbf{x}). \quad (8)$$

It is difficult to solve this problem directly, as the likelihood of the associations are latent variables for which no prior is given. Instead, the probabilities of the possible semantic region associations need to be estimated alongside $\mathbf{x}$.

The common way to solve such problems is given by the EM framework. The idea consists of iteratively estimating the transformation parameters $\mathbf{x}$ where—in each iteration—the joint probability distribution is changed by taking the expectation value over the likelihoods of $\mathcal{D}$, followed by a maximization of the remaining objective over $\mathbf{x}$. Formally, we have

$$\begin{aligned}\mathbf{x}^{i+1} &= \arg\max_{\mathbf{x}} \mathbb{E}_{\mathcal{D}|\mathcal{A},\mathcal{B},\mathbf{x}^i}[\log \mathbf{p}(\mathcal{A},\mathcal{D}|\mathcal{B},\mathbf{x})] \\ &= \arg\max_{\mathbf{x}} \sum_{k=1}^{K} \mathbf{p}(D_k|\mathbf{A}^{\alpha_k},\mathbf{B}^{\beta_k},\mathbf{x}^i)\log \mathbf{p}(\mathbf{A}^{\alpha_k}|\mathbf{B}^{\beta_k},\mathbf{x}).\end{aligned} \quad (9)$$

Expectation-maximization executes two steps in each iteration:

- E step: Given $\mathbf{x}^i$, we estimate

$$h(D_k|\mathbf{x}^i) = \mathbf{p}(D_k|\mathbf{A}^{\alpha_k},\mathbf{B}^{\alpha_k},\mathbf{x}^i). \quad (10)$$

- M step: Using $w_k = h(D_k|\mathbf{x}^i)$, obtain $\mathbf{x}^{i+1}$ from the objective

$$\mathbf{x}^{i+1} = \arg\max_{\mathbf{x}}\{\sum_{k=1}^{K} w_k \log \mathbf{p}(\mathbf{A}^{\alpha_k}|\mathbf{B}^{\beta_k},\mathbf{x})\}. \quad (11)$$

  w_k denotes the association probability between $\mathbf{A}^{\alpha_k}$ and $\mathbf{B}^{\beta_k}$. It is independent of the optimization variables $\mathbf{x}$, and quantifies our soft semantic region association.

As can be observed, the objective considers multiple possible associations between semantic regions commensurate with their respective probabilities. Picturing the regions as nodes in a graph, the associations represent the edges in the graph whereas the association probabilities denote the edge weights. The formulation can be considered as implicitly adding semantic graph constraints to the registration.

A. Expectation Step

The weights $w_k = h(D_k|\mathbf{x}^i)$ indicating the likelihood of a certain association are decomposed as in

$$\begin{aligned}h(D_k|\mathbf{x}^i) &= \mathbf{p}(D_k|\mathbf{A}^{\alpha_k},\mathbf{B}^{\beta_k},\mathbf{x}^i) \\ &= \mathbf{p}(s|\mathbf{A}^{\alpha_k}) * \mathbf{p}(s|\mathbf{B}^{\beta_k}) * \mathbf{p}(e|\mathbf{A}^{\alpha_k},\mathbf{B}^{\beta_k},\mathbf{x}^i).\end{aligned} \quad (12)$$

$\mathbf{p}(s|\mathbf{A}^{\alpha_k})$ and $\mathbf{p}(s|\mathbf{B}^{\beta_k})$ are the semantic detection confidences (e.g. object detection confidences) provided by the neural network. The density $\mathbf{p}(e|\mathbf{A}^{\alpha_k},\mathbf{B}^{\beta_k},\mathbf{x}^i)$ then expresses the geometric distance between the regions. In our work, we use the four classical distance metrics introduced in Section III, given by the *mean point distance* (4), *average distance* (5), the *sum of minimums distance* (6), and the *Hausdorff distance* (7). The influence of the choice of distance metric is analysed in detail in our experimental results section.

Some special cases for semantic region association exist, which we summarize as follows:

(i) The instance segmentation generates only a single segment, in which case our algorithm would naturally transition to the classical ICP algorithm, i.e. $|\mathcal{D}|$=1.

(ii) Certain regions have no potential association to any other regions. We would simply redefine such regions to become part of the background. The three primary reasons for this case are: 1) a new region entered the view due to a large view-point change, 2) an old region left the view due to a large view-point change, or 3) a wrong or missing detection in one of the views.

(iii) A region has received possible associations with other regions, but none of the associations is the correct one. The reasons for this case are essentially the same than for the previous point.

If we perform point-set registration without semantic region association [32], [42], situation (iii) would likely decrease registration quality. However—in our algorithm—the relevant association weights would likely be low, thus alleviating the influence of the false set of potential associations.

B. Maximization Step: Cascaded EM or ICP

The registration probability in (11) between the point subsets of two semantic regions is given by

$$\mathbf{p}(\mathbf{A}^{\alpha_k}|\mathbf{B}^{\beta_k},\mathbf{x}) = \Pi_{n=1}^{N_{\alpha_k}}\mathbf{p}(a_n^{\alpha_k}|B^{\beta_k},\mathbf{x}). \tag{13}$$

Let $\mathcal{I} = \{\mathbf{I}^k|k = 1,\cdots,K\}$ be the region-wise set of point-pair association sets, where $\mathbf{I}^k = \{I_n^k\}|_{n=1,\cdots,N_{\alpha_k}}$ is the set of associations for all the points in A^{α_k}. More specifically, $I_n^k = \{(n,m)\}|_{m\in M_{\beta_k}}$ contains all point-level correspondences between the n-th point $a_n^{\alpha_k}$ from the source pointset A^{α_k} and any of the points $b_m^{\beta_k}$ from the target data B^{β_k}. Adding the likelihoods over the unobserved, latent point-pair associations gives

$$\mathbf{x}^{i+1} = \arg\max_{\mathbf{x}}\{\sum_{k=1}^{K} w_{D_k}\sum_{n=1}^{N_{\alpha_k}}\log\mathbf{p}(a_n^{\alpha_k},I_n^k|B^{\beta_k},\mathbf{x}). \tag{14}$$

Different registration methods have different realizations of this maximization step. Here we use GMA and ICP as two examples. Given the weights w_{D_k} of the semantic region association, the maximization is in any case solved iteratively. We therefore initialize $\mathbf{x}^{i,0} = \mathbf{x}^i$ for the maximization step. Later, we would update $\mathbf{x}^{i+1} = \mathbf{x}^{i,j}$, where j is the total number of iterations executed during the maximization step.

Gaussian Mixture Alignment: Using GMA again means that we need to construct residual terms for each possible point-pair (all-to-all correspondences), which is why the association layer $\mathcal{I}$ can also be interpreted as a soft association. However, compared against classical GMA, our algorithm adds the semantic region association weights to each point-pair residual term. The derivation again leads to the EM framework and proceeds by taking the expectation over $\mathcal{I}$ in each iteration, followed by an inner maximization step. Given $\mathbf{x}^{i,j}$, each EM-iteration therefore becomes

$$\begin{aligned}\mathbf{x}^{i,j+1} = \arg\max_{\mathbf{x}}\{&\sum_{k=1}^{K} w_{D_k}\\ &\sum_{n}^{N_{\alpha_k}}(I_n^k|a_n^{\alpha_k},B^{\beta_k},\mathbf{x}^{i,j})\log\mathbf{p}(a_n^{\alpha_k}|B^{\beta_k},I_n^k,\mathbf{x})\}.\end{aligned} \tag{15}$$

Similar to the region association, the two steps in each iteration are given by:

- E step:

$$h(I_n^k|\mathbf{x}^{i,j}) = \mathbf{p}(I_n^k|a_n^{\alpha_k},B^{\beta_k},\mathbf{x}^{i,j}). \tag{16}$$

- M step: Given $w_{D_k} = h(D_k|\mathbf{x}^i)$ and $w_{I_n^k} = h(I_n^k|\mathbf{x}^{i,j})$, obtain $\mathbf{x}^{i,j+1}$ by solving the problem.

$$\mathbf{x}^{i,j+1} = \arg\max_{\mathbf{x}}\{\sum_{k=1}^{K} w_{D_k}\sum_{n}^{N_{\alpha_k}} w_{I_n^k}\log\mathbf{p}(a_n^{\alpha_k}|B^{\beta_k},I_n^k,\mathbf{x})\}. \tag{17}$$

Again, note that I_n^k here is the association set that associates $a_n^{\alpha_k}$ to all the points in B^{β_k}. Furthermore note that semantic region associations hold the potential of rejecting lots of point-pairs and could therefore decrease the computational cost of the optimization (i.e. all-to-all point-pairs only happen within associated regions).

ICP registration: The Iterative Closest Point (ICP) algorithm aligns the two sets by iteratively alternating between the two steps of finding nearest neighbours, and computing the alignment. Compared against GMA, the I_n^k within ICP only encodes a single point-pair correspondence and it is fixed before each update of $\mathbf{x}$. The two alternating maximisation steps are given by

$$\mathcal{I} = \arg\max_{\mathcal{I}}\{\sum_{k=1}^{K} w_{D_k}\sum_{n=1}^{N_{\alpha_k}}\log\mathbf{p}(a_n^{\alpha_k},I_n^k|B^{\beta_k},\mathbf{x}^{i,j})\}, \tag{18}$$

$$\mathbf{x}^{i,j+1} = \arg\max_{\mathbf{x}}\{\sum_{k=1}^{K} w_{D_k}\sum_{n=1}^{N_{\alpha_k}} w_{I_n^k}\log\mathbf{p}(a_n^{\alpha_k}|B^{\beta_k},I_n^k,\mathbf{x})\}. \tag{19}$$

The algorithm is similar to the EM-ICP framework presented in [18]. Note however that—since we consider all possible region associations—the same point could potentially be asked to find a nearest neighbor more than once, and there could be more than one residual term for each point in the source data. This is a stark difference to existing ICP algorithms, which all perform point-pair associations only (including more modern ICP variants that make use of semantic information). An abstract is given in Alg.1.

V. Experiments

We test our algorithm on several open datasets, including ScanNet [13], BundleFusion [14], RGB-D Scenes v2 [45] and 7-Scenes [17]. Our non-semantic baseline implementation is given by EM-ICP [18], which performs better than classical ICP [43]. We furthermore compare against the closely related semantic EM-ICP [32], which incorporates semantic information into the point-level associations. We denote this work [32] SICP-NDA, as it performs no data associations at the level of semantic regions. Conversely, our own algorithm is denoted SICP-DA. We use BlendMask [7] for object instance segmentation. The background (or missed objects) are set with the additional semantic label 0. Regarding implicit parameter σ for Guassian distribution in (10) for region association, we set $\frac{1}{\sigma^2} = 15$ and estimate it

Algorithm 1 Main Algorithm: Semantic ICP with region-wise data association(SICP-DA)

Input: Source data: $\mathcal{A}$, target data: $\mathcal{B}$.
Output: Optimal $\mathbf{x}^*$.
Initialization $\mathbf{x}^0, \mathcal{D}^0, \mathcal{I}^0$
loop
  E step: compute the semantic region association weights (10) for each pair of semantic regions;
  M step: given weights w_{D_k} and $\mathbf{x}^{i,0} = \mathbf{x}^i$.
  loop
    (16) $\rightarrow$ (17) (or (18) $\rightarrow$ (19));
  end loop
  Update $\mathbf{x}^{i+1} = \mathbf{x}^{i,j}$.
end loop

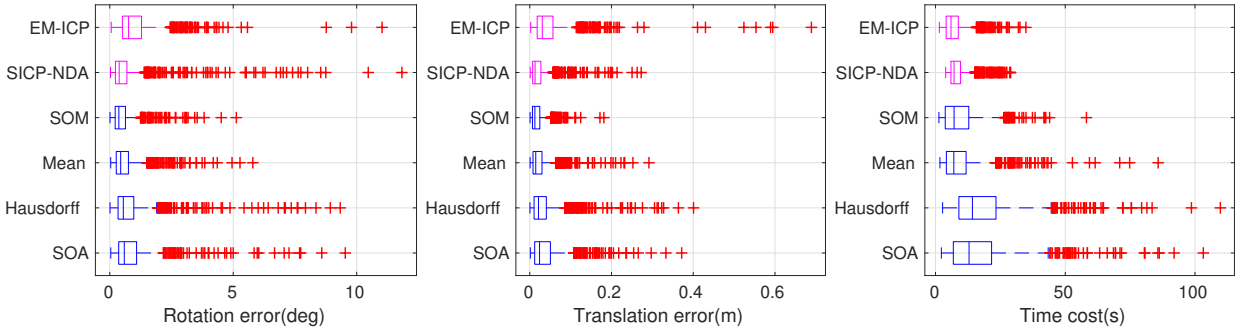

Fig. 1. Error box plots of the proposed algorithms compared against EM-ICP and SICP-NDA on ScanNet *scene0015_00*. The left two plots are the transformation errors, and the right one indicates the time cost for each algorithm. The Blue box plots are our proposed algorithm with different distance metrics for region association estimation.

using an annealing strategy. We implement all the ICP-based algorithms in Matlab on a computer with 3.6 GHz CPU and 32 GB RAM.

Registration algorithms typically ignore associations for which the geometric distance exceeds a certain threshold. This ensures that points in the source point set that have no counterpart in the target set will not affect the solution. We avoid a hard threshold and replace the trimming step by a robust estimator using the Huber loss. The addition of a robust kernel is equally performed for all algorithms, thus ensuring a fair comparison. There are many further variations of ICP and GMA to improve robustness and speed, which could again be equally applied to all algorithms. They are however omitted to ensure a simple, fair evaluation of the impact of semantic associations at the point or region level.

Each experiment consists of aligning two views from a varying initial relative transformation (i.e. view-point change). Since all algorithms rely on a local optimizer, the amount of initial transformation will impact on the achievable accuracy. To analyze the ability of each method to deal with the initial misalignment (i.e. the size of the convergence basin), sets of experiments with varying initial transformation are created by varying the frame gap between the chosen pair of views from 1 to 11. Rotation errors are given in degrees, and translation errors are reported in meters (m).

A. Maximization using ICP

In this section, the internal maximisation step of our method is performed using the ICP variant. We use the ScanNet sequence "scene0015_00" to compare our method against the listed alternatives. We further compare different implementations of our algorithm to analyse the influence of the point set distance metrics ("Mean"(4),"SOA"(5), "SOM"(6), and "Hausdorff"(7)). The result is shown in Figure 1.

As can be observed, the "Hausdorff" and "SOA" metrics have higher mean error while their median error remains low. We trace this back to their low resilience against segmentation errors, and deem them unsuitable for the purpose of semantic region association. The "Mean" and "SOM" metrics in turn produce much more stable results, and "SOM" performs best. In the following experiments, we use "SOM"

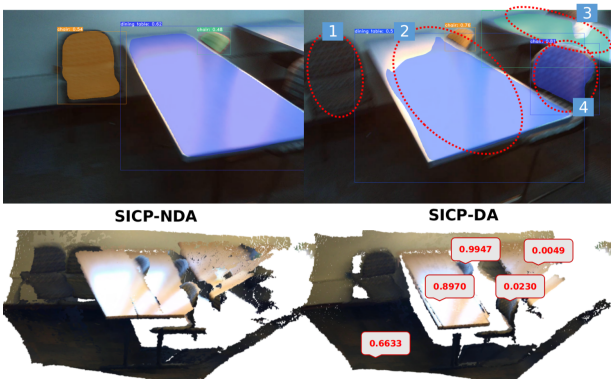

Fig. 2. Illustration of the influence of missing object detections. Top row: 2 key frames with instance segmentation masks. Bottom row: Alignement obtained by SICP-NDA and SICP-DA. Numbers indicate semantic region association likelihoods.

for the region association. EM-ICP has higher transformation errors than semantic ICP alternatives. SICP-NDA has similar average errors than SICP-DA with "SOM", but our algorithm performs more stable and has higher convergence radius. On the down side, our algorithm is slightly less computationally efficient than EM-ICP or SICP-NDA, as the semantic data association simply adds more residual terms to the estimation (cf. third column in Figure 1).

As outlined in Section (IV-A), SICP-NDA is easily influenced by missing and false positive object detections. An example is given in Figure 2, where (1) the segmentation in the source frame misses one chair for which points will be assigned to the background, (2) the segmentation of a table in the source frame is much worse than in the target frame, (3) the target frame misses one table, and (4) the target frame misses one chair. As indicated in the Figure, SICP-DA significantly outperforms SICP-NDA with the help of region associations. Owing to the large view-point change, the association likelihood of the background is 0.6633. Regions with missing counter-part have very low association weights (i.e. 0.0049 and 0.0230, respectively). Comparing the remaining regions, the more consistent segmentation has a higher association weight (i.e. 0.9947 for the chair versus 0.8970 for the table).

We test the algorithms on more open datasets. Figure 3

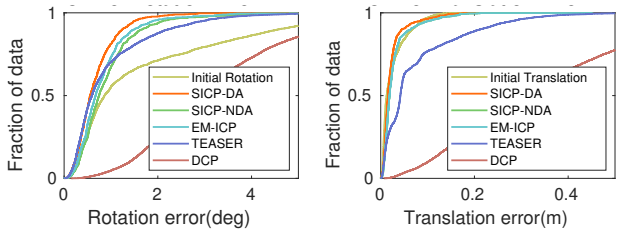

Fig. 3. Transformation error CDF plots on BundleFusion *apt2*.

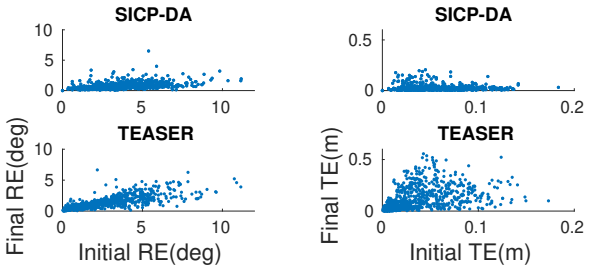

Fig. 4. Scatter plots of the initial alignment vs. final alignment on the BundleFusion *apt2*.

plots the CDF error on *apt2* of the BundleFusion dataset. We add two more comparison methods: DCP [40] (re-trained on ScanNet) and TEASER [35] (we use the python package TEASER 0.4.4 and the 3D feature FPFH[46]). As can be observed, DCP produces low accuracy relative pose estimates. TEASER on the other hand is a very fast feature-based global registration algorithm that shows high robustness against noisy point sets and limited overlap ratios. A detailed opposition of SICP-DA and TEASER is given in scatter diagram in Figure 4, showing that TEASER is more sensitive to the initial transformation for relative pose estimation compared with SICP-DA. To conclude, Table V-A summarizes the average rotation (RE) and translation (TE) errors on *redkitchen* from 7-Scenes, *scene_10* from RGBD Scenes V2, and *scene0025_01* from ScanNet. We present the mean and median errors as well as the standard deviation. As can be observed, SICP-DA generally performs best.

TABLE I

AVERAGE TRANSFORMATION ERROR.

Data	Error	DCP	TEASER	EM-ICP	SICP-NDA	SICP-DA
7 Scene	RE (deg)	3.297	0.574	0.695	0.488	**0.409**
mean		2.915	0.415	0.632	0.368	**0.313**
median		1.281	0.460	0.351	0.255	**0.251**
	TE (m)	0.354	0.041	0.025	0.018	**0.013**
mean		0.309	0.027	0.018	0.009	**0.008**
median		0.145	0.053	0.244	0.127	**0.010**
Scene V2	RE (deg)	1.071	0.365	0.846	0.312	**0.316**
mean		1.014	0.321	0.681	**0.297**	0.301
median		0.484	0.235	0.670	0.162	**0.162**
	TE (m)	0.023	0.015	0.021	**0.006**	**0.006**
mean		0.021	0.012	0.017	**0.005**	**0.005**
median		0.012	0.007	0.014	**0.003**	**0.003**
Scan Net	RE (deg)	1.934	0.782	1.039	0.695	**0.611**
mean		1.467	0.553	0.762	**0.382**	0.388
median		1.641	1.502	0.886	1.058	**0.631**
	TE (m)	0.072	0.033	0.047	0.033	**0.025**
mean		0.070	0.021	0.031	**0.015**	0.016
median		0.042	0.085	0.055	0.035	**0.023**

B. Maximization using GMA

To conclude, we also compare the GMA variants of all registration methods. Registration without semantic information is denoted GMA, with semantics and point-level associations SGMA-NDA, and with semantic region associations SGMA-DA. We only utilize the "SOM" point distance metric. Figure 5 illustrates the obtained rotation errors and time cost on ScanNet "scene0015_00", demonstrating that SGMA-DA clearly outperforms the alternative methods.

The computation time of GMA-based methods increases quickly with the number of points. We accelerate convergence by incorporating the strategy in [18] to decrease the number of potential associations. Nonetheless, it should be noted that SGMA-DA is faster than SICP-NDA and GMA. While the non-uniqueness of point-level associations make SICP-DA slower than EM-ICP and SICP-NDA, SGMA-DA performs "all-to-all" point associations within associated regions, only, and thus significantly outperforms SGMA-NDA and GMA in terms of computational efficiency.

Figure 6 shows example progressions of the RMS of the registration errors for GMA, SGMA-NDA, and SGMA-DA (we use groundtruth to define the overlap region of the two point sets, and consider only points within the overlap). As can be observed, SGMA-DA generally converges fastest and achieves the lowest RMS errors.

VI. CONCLUSION

We propose the addition of semantic region associations as an implicit supervision signal within point set registration. Our results demonstrate that this novel, hierarchical registration provides better, more stable results than pure point-level association frameworks. Our proposed local registration method furthermore outperforms recent state-of-the-art global registration methods in terms of accuracy.

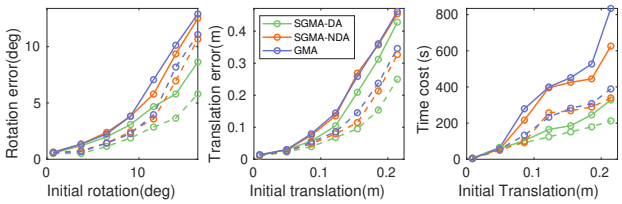

Fig. 5. Comparison of GMA-based implementations on ScanNet *scene0122_00*. The solid lines denote the mean error and the dotted lines are the median error.

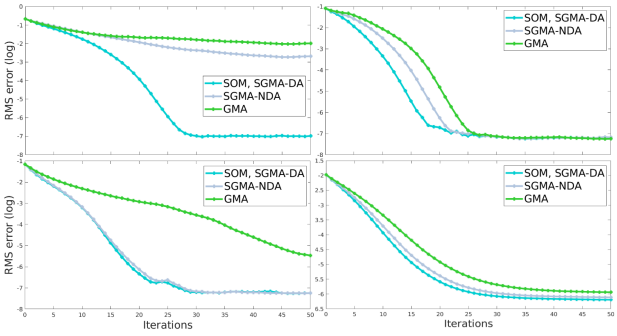

Fig. 6. Examples of RMS registration error progressions of GMA-based methods.

REFERENCES

[1] M. Ahmed Sherif and A.-C. Ngonga Ngomo. A systematic survey of point set distance measures for link discovery. *Semantic Web*, 9(5):589–604, 2018.

[2] K. Arun, T. Huang, and S. Blostein. Least-Squares Fitting of Two 3-D Point Sets. *Transations of Pattern Analysis and Machine Intelligence*, 9(5):698–700, 1987.

[3] P. J. Besl and N. D. McKay. Method for registration of 3-d shapes. In *Sensor fusion IV: control paradigms and data structures*, volume 1611, pages 586–606. International Society for Optics and Photonics, 1992.

[4] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee. Yolact++: Better real-time instance segmentation. *arXiv preprint arXiv:1912.06218*, 2019.

[5] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee. Yolact: real-time instance segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9157–9166, 2019.

[6] S. L. Bowman, N. Atanasov, K. Daniilidis, and G. J. Pappas. Probabilistic data association for semantic slam. In *2017 IEEE international conference on robotics and automation (ICRA)*, pages 1722–1729. IEEE, 2017.

[7] H. Chen, K. Sun, Z. Tian, C. Shen, Y. Huang, and Y. Yan. Blendmask: Top-down meets bottom-up for instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8573–8581, 2020.

[8] J. Chen, X. Wu, M. Y. Wang, and X. Li. 3d shape modeling using a self-developed hand-held 3d laser scanner and an efficient ht-icp point cloud registration algorithm. *Optics & Laser Technology*, 45:414–423, 2013.

[9] Y. Chen and G. Medioni. Object modelling by registration of multiple range images. *Image and vision computing*, 10(3):145–155, 1992.

[10] D. Chetverikov, D. Stepanov, and P. Krsek. Robust euclidean alignment of 3d point sets: the trimmed iterative closest point algorithm. *Image and vision computing*, 23(3):299–309, 2005.

[11] C. Choy, W. Dong, and V. Koltun. Deep global registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2514–2523, 2020.

[12] H. Chui and A. Rangarajan. A new point matching algorithm for non-rigid registration. *Computer Vision and Image Understanding*, 89(2-3):114–141, 2003.

[13] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. 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.

[14] 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)*, 36(4):1, 2017.

[15] K. Doherty, D. Baxter, E. Schneeweiss, and J. Leonard. Probabilistic data association via mixture models for robust semantic slam. *arXiv preprint arXiv:1909.11213*, 2019.

[16] K. Doherty, D. Fourie, and J. Leonard. Multimodal semantic slam with probabilistic data association. In *2019 international conference on robotics and automation (ICRA)*, pages 2419–2425. IEEE, 2019.

[17] B. Glocker, S. Izadi, J. Shotton, and A. Criminisi. Real-time rgb-d camera relocalization. In *2013 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 173–179. IEEE, 2013.

[18] S. Granger and X. Pennec. Multi-scale em-icp: A fast and robust approach for surface registration. In *European Conference on Computer Vision*, pages 418–432. Springer, 2002.

[19] S. Granger, X. Pennec, and A. Roche. Rigid point-surface registration using an em variant of icp for computer guided oral implantology. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 752–761. Springer, 2001.

[20] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[21] H. Hu, S. Lan, Y. Jiang, Z. Cao, and F. Sha. Fastmask: Segment multi-scale object candidates in one shot. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 991–999, 2017.

[22] B. Jian and B. C. Vemuri. Robust point set registration using gaussian mixture models. *IEEE transactions on pattern analysis and machine intelligence*, 33(8):1633–1645, 2010.

[23] X. Kang and S. Yuan. Robust data association for object-level semantic slam. *arXiv preprint arXiv:1909.13493*, 2019.

[24] Y. Lee and J. Park. Centermask: Real-time anchor-free instance segmentation. *arXiv preprint arXiv:1911.06667*, 2019.

[25] Y. Li, H. Qi, J. Dai, X. Ji, and Y. Wei. Fully convolutional instance-aware semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2359–2367, 2017.

[26] W. Lu, G. Wan, Y. Zhou, X. Fu, P. Yuan, and S. Song. Deepicp: An end-to-end deep neural network for 3d point cloud registration. *arXiv preprint arXiv:1905.04153*, 2019.

[27] J. McCormac, R. Clark, M. Bloesch, A. Davison, and S. Leutenegger. Fusion++: Volumetric object-level slam. In *2018 international conference on 3D vision (3DV)*, pages 32–41. IEEE, 2018.

[28] J. McCormac, A. Handa, A. Davison, and S. Leutenegger. Semanticfusion: Dense 3d semantic mapping with convolutional neural networks. In *2017 IEEE International Conference on Robotics and automation (ICRA)*, pages 4628–4635. IEEE, 2017.

[29] A. Myronenko and X. Song. Point set registration: Coherent point drift. *IEEE transactions on pattern analysis and machine intelligence*, 32(12):2262–2275, 2010.

[30] Y. Nakajima and H. Saito. Efficient object-oriented semantic mapping with object detector. *IEEE Access*, 7:3206–3213, 2018.

[31] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohli, J. Shotton, S. Hodges, and A. Fitzgibbon. KinectFusion: Real-time dense surface mapping and tracking. In *International Symposium on Mixed and Augmented Reality*, 2011.

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

[33] Q.-H. Pham, B.-S. Hua, T. Nguyen, and S.-K. Yeung. Real-time progressive 3d semantic segmentation for indoor scenes. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1089–1098. IEEE, 2019.

[34] K. Pulli. Multiview registration for large data sets. In *Second International Conference on 3-D Digital Imaging and Modeling (Cat. No. PR00062)*, pages 160–168. IEEE, 1999.

[35] H. Yang, J. Shi, and L. Carlone. Teaser: Fast and certifiable point cloud registration. *arXiv preprint arXiv:2001.07715*, 2020.

[36] A. Segal, D. Haehnel, and S. Thrun. Generalized-icp. In *Robotics: science and systems*, volume 2, page 435. Seattle, WA, 2009.

[37] P. Speciale, D. P. Paudel, M. R. Oswald, H. Riemenschneider, L. Van Gool, and M. Pollefeys. Consensus maximization for semantic region correspondences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7317–7326, 2018.

[38] Y. Tsin and T. Kanade. A correlation-based approach to robust point set registration. In *European conference on computer vision*, pages 558–569. Springer, 2004.

[39] X. Wang, T. Kong, C. Shen, Y. Jiang, and L. Li. Solo: Segmenting objects by locations. *arXiv preprint arXiv:1912.04488*, 2019.

[40] Y. Wang and J. M. Solomon. Deep closest point: Learning representations for point cloud registration. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3523–3532, 2019.

[41] S. Yang, Z.-F. Kuang, Y.-P. Cao, Y.-K. Lai, and S.-M. Hu. Probabilistic projective association and semantic guided relocalization for dense reconstruction. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 7130–7136. IEEE, 2019.

[42] A. Zaganidis, L. Sun, T. Duckett, and G. Cielniak. Integrating deep semantic segmentation into 3-d point cloud registration. *IEEE Robotics and Automation Letters*, 3(4):2942–2949, 2018.

[43] Z. Zhang. Iterative point matching for registration of free-form curves and surfaces. *International journal of computer vision*, 13(2):119–152, 1994.

[44] Q.-Y. Zhou, J. Park, and V. Koltun. Fast global registration. In *European Conference on Computer Vision*, pages 766–782. Springer, 2016.

[45] K. Lai, L. Bo, and D. Fox. Unsupervised feature learning for 3d scene labeling. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3050–3057. IEEE, 2014.

[46] R. B. Rusu, N. Blodow, and M. Beetz. Fast point feature histograms (fpfh) for 3d registration. In *2009 IEEE international conference on robotics and automation*, pages 3212–3217. IEEE, 2009.