Semantic Iterative Closest Point through Expectation-Maximization

Steven A. Parkison
sparki@umich.edu
Lu Gan
ganlu@umich.edu
Maani Ghaffari Jadidi
maanigj@umich.edu
Ryan M. Eustice
eustice@umich.edu

Perceptual Robotics Laboratory
University of Michigan
Ann Arbor, Michigan, USA

Abstract

In this paper, we develop a novel point cloud registration algorithm that directly incorporates pixelated semantic measurements into the estimation of the relative transformation between two point clouds. The algorithm uses an Iterative Closest Point (ICP)-like scheme and performs joint semantic and geometric inference using the Expectation-Maximization technique in which semantic labels and point associations between two point clouds are treated as latent random variables. The minimization of the expected cost on the three-dimensional special Euclidean group, i.e., SE(3), yields the rigid body transformation between two point clouds. The evaluation on publicly available RGBD benchmarks shows that, in comparison with both the standard Generalized ICP (GICP) available in the Point Cloud Library and GICP on SE(3), the registration error is reduced.

1 Introduction

Point cloud registration, the task of finding the rigid body transformation between two point clouds, is an integral part of geometric inference in many modern perception systems. The most successful algorithm is known as the Iterative Closest Point (ICP) [5, 12] algorithm. ICP was further developed to the probabilistic framework known as Generalized ICP (GICP) [39].

Semantic inference on images and point clouds has shown increasing value in vision-based applications. Early algorithms relied on classifiers trained on a set of hand-crafted features [8, 41]. However, their computational efficiency limits their application in real-time scenarios. Advances in Convolutional Neural Networks (CNNs) have improved the computational efficiency of semantic segmentation techniques with superior performance in both indoor and outdoor benchmarks [23, 26, 33, 34, 45]. Together with pose estimation techniques, multiple scenes can be segmented and combined to perform semantic mapping [28]; nevertheless, most semantic mapping research has focused on combining geometry and semantics into a map representation, and not on how semantics can improve pose estimation.

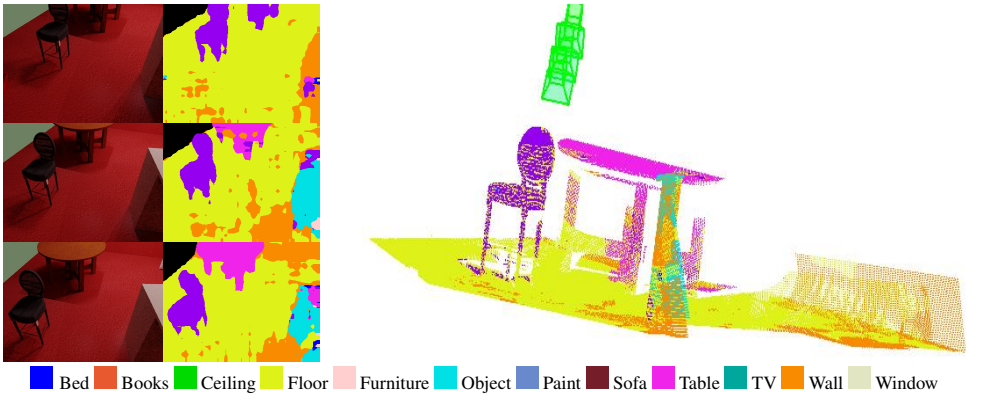

Figure 1: An example of three point clouds from the SceneNet RGBD dataset [27, 29] aligned using Semantic ICP. The left column shows the source images. The middle column shows the inferred semantic classes, labeled by the most likely class. The right figure shows three point clouds projected into a common reference frame by the estimated transformations, with the camera positions marked in green. Class labels are indicated below the image. The crisp objects are a sign of good alignment. Moreover, data association is addressed through performing a joint semantic and geometric inference by using the EM technique in which semantic labels and point associations between two point clouds are treated as latent random variables. Despite inaccuracies in the inference, semantics improve point cloud registration results.

In this paper, we develop the Semantic ICP algorithm that directly incorporates pixel semantics into the registration problem between two overlapping point clouds. The primary motivation is aiding tasks that rely on joint semantic segmentation and relative pose estimation, such as semantic mapping and object tracking. Figure 1 illustrates this concept where semantic labels in an indoor scene aid the alignment. In particular, this work has the following contributions:

1. Development of the Semantic ICP algorithm which uses joint semantic and geometric probabilities for finding the associations in the GICP-SE(3) algorithm, where GICP-SE(3) algorithm solves the point cloud registration problem with respect to the motion group manifold structure.

2. The open source implementation of the proposed algorithms as well as code to reproduce the provided results [1].

3. We provide experimental evaluations using publicly available benchmarks, KITTI [20] and SceneNet RGBD [27, 29] datasets, that show improved registration performance over current methods.

2 Related Work

Point cloud registration is generally formulated as an optimization problem over the rigid body transformation that minimizes some residual between points in the source cloud to

[1] https://bitbucket.org/saparkison/semantic-icp

points in the target cloud. The ICP algorithm [5] defines the objective function as the Euclidean distance between points in the source cloud, to an associated point in the target cloud. That association is rarely known and is unobserved by the sensor, thus the approach taken by Besl and McKay [5] is to alternate between finding the Euclidean Nearest Neighbor (NN) association between points in each cloud and optimizing the point-to-point distance function over the transformation variables. This approach is generalized slightly using point-to-line [10] and point-to-plane [12] objective functions that have been shown to improve convergence speed and accuracy [32, 35].

There has also been work on defining probability distributions over points in the point clouds. Biber and Straßer [6] define normal distributions using points in the target point cloud that fall into voxels of the environment. The objective function is defined as the probability that a point in the source point cloud is within the distributions of the target point cloud. Generalized ICP [39] also defines a Gaussian distribution over the source and target point clouds, but computes these distributions by calculating the sample covariance of neighboring points, where neighbors are those points that are the closest in the Euclidean distance. These probabilistic formulations of ICP and the iterative nature of the algorithm have led to Expectation Maximization (EM) approaches to the point cloud registration problem [21]. Lee and Lee [24] use an EM approach to align sensor measurements to 3D models of objects while also learning the covariance of the observation to improve the alignment. Their method performs well on the model alignment task, but they treat the distribution on model and observation points differently, which does not generalize to point cloud registration between two observations. Gabriel Agamennoni and Sorrenti [18] also formulate point cloud registration as an EM problem. They model points in the source cloud drawn from a t-distribution centered at points in the target cloud. None of these methods include semantics. Our approach, which includes semantics, improves upon these methods' registration accuracy, as we show in the evaluation.

We conduct joint geometric and semantic inference to improve the registration task. Semantics have been combined with geometry in a variety of ways. Object level classification has been used on premade maps [9, 15]. Bao and Savarese [3] use an object detector in the structure-from-motion setup to jointly estimate camera parameters, 3D points, and object instances and poses. Their method, unlike ours, requires parametrizing objects, and only estimates sparse object poses and not dense point labels. These approaches are developed to improve scene estimation by providing more geometric constraints. Conversely, Pillai and Leonard [31] use monocular SLAM to aggregate multiple views of a single object to provide more evidence to the object detector. This approach treats pose estimation and semantic classification as independent, solving the first to improve the second.

Dense 3D priors of objects have also been used for scene estimation and mapping. Salas-Moreno et al. [37] align 3D mesh model priors of objects to the RGBD frame. The technique treats objects as landmarks and each alignment as a *factor* in the *graphical* SLAM framework. Choudhary et al. [14] also use objects as landmarks, but instead of having a dense 3D prior over every object, the objects are discovered via segmentation and modeled during the mapping process. Bowman et al. [7] approach data association in SLAM using EM, though at the sparse object level as opposed to the dense point level. The iterative nature of ICP is closer to that of an EM framework then SLAM is, which requires finding the solution to the full SLAM problem multiple times at each step to converge on an association. Yu et al. [46] use semantics in city-wide mapping by extracting semantic features from point clouds and then matching and aligning the features. This contrasts with our approach in that we propose to densely align points through joint geometric and semantic inference.

Sevilmis and Kimia [40] leveraged shapes of objects to improve optical flow matching using richer representations such as SIFT and CNN features. It was found that the greater the visual variation between the images, the more their approach was aided by shape correspondences. There has been other work, which while it does not directly use semantic class labels, that uses object and feature geometry to improve association in the registration problem. Gressin et al. [22] use feature computed on the local geometry around a point to both select good points to use and to improve association search. Similarly, Weinmann and Jutzi [43] use the local geometry of a point to assess the quality, which improves the number of inliers for their RANSAC based registration method, while also improving the convergence rate. Zaganidis et al. [47] propose an approach that adds semantics to the Normal Distribution Transform. In the latter work, their definition of semantics is geometric edges and planes. Instead, our definition is object and class labels. These last approaches also differ in that they strictly enforce NNs; whereas we treat semantics as noisy measurements that assist in modeling the probability of association.

3 Problem Statement and Formulation

We wish to find the 3D rigid body transformation that aligns two semantically labeled point clouds. We will be using $\mathcal{X} \subset \mathbb{R}^3$ to represent a set of spatial coordinates collected by a range/depth sensor. The following definitions are used throughout the paper.

Definition 1 (Target point cloud)**.** *The point cloud $\mathcal{X}_t$ which is considered to be in a fixed reference frame is called the target point cloud.*

Definition 2 (Source point cloud)**.** *The point cloud $\mathcal{X}_s$ which $\mathbf{T} \in \mathrm{SE}(3)$ acts on is called the source point cloud.*

The action of $\mathbf{T}$ on any point $\mathbf{x}_i \in \mathcal{X}$ is $\mathbf{T}(\mathbf{x}_i) = \mathbf{R}\mathbf{x}_i + \mathbf{p}$, where $\mathbf{R} \in \mathrm{SO}(3)$ and $\mathbf{p} \in \mathbb{R}^3$. The likelihood function for aligning two point clouds sampled from the same environment depends on data association between them. We define the association variable $\mathcal{I} \triangleq \{i_k, j_k\}_{k=1}^n \in \mathbb{I}$ where i_k, j_k indicate $\mathbf{x}_k^t \triangleq \mathbf{x}_{i_k}^t \in \mathcal{X}_t$ is a measurement of the same point as $\mathbf{x}_k^s \triangleq \mathbf{x}_{j_k}^s \in \mathcal{X}_s$, and $\mathbb{I}$ is the set of all possible associations (permutations). In short, the association set $\mathcal{I}$ gives the indices of points in the target and source cloud which are independent measurements of the same point. We also introduce a new random variable, $\mathcal{R} \triangleq \{\mathbf{r}_k\}_{k=1}^n$, to represent the residual where $\mathbf{r}_k \triangleq \mathbf{x}_k^t - \mathbf{T}(\mathbf{x}_k^s)$. To emphasize that the likelihood term includes the action of $\mathbf{T} \in \mathrm{SE}(3)$ on $\mathcal{X}_s$, we shall write the log-likelihood function as $f(\mathbf{T}; \mathcal{R}|\mathcal{X}_t, \mathcal{X}_s, \mathcal{I}) \triangleq \log p(\mathcal{R}|\mathcal{X}_t, \mathcal{X}_s, \mathcal{I}; \mathbf{T})$. However, for simplicity, we use $p(\mathcal{R}|\mathcal{X}_t, \mathcal{X}_s, \mathcal{I})$ whenever $\mathbf{T}$ is irrelevant.

Most ICP-based approaches follow an iterative two-step procedure for solving the point cloud registration problem. **Step 1**: Determine the association $\mathcal{I}$. **Step 2**: Minimize the cost defined using the residual, $\mathcal{R}$, over the parameter $\mathbf{T}$.

Thus, the geometric point cloud registration problem in **Step 2**, give a fixed set of associations $\mathcal{I}$, is expressed as follows.

Problem 1 (Point cloud registration)**.** *Let $\mathcal{X}_t$ and $\mathcal{X}_s$ be two geometric point clouds. Given correspondences between target and source point clouds, $\mathcal{I}$, the optimal transformation that aligns the source to the target can be computed by solving the following maximum likelihood estimation (MLE) problem:*

$$\max_{\mathbf{T} \in \mathrm{SE}(3)} \; f(\mathbf{T}; \mathcal{R}|\mathcal{X}_t, \mathcal{X}_s, \mathcal{I}) \tag{1}$$

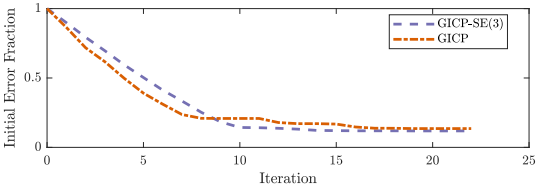

Figure 2: Convergence evaluation for GICP and GICP-SE(3). Median Fraction (taken from 50 alignments) of initial error, measure as $\mathrm{d}_{\mathrm{SE}(3)}(\cdot,\cdot)$, over outer loop iterations. While GICP initially converges faster, GICP-SE(3) reaches steady state in fewer iterations.

Point clouds $\mathcal{X}_t$ and $\mathcal{X}_s$ observe the geometry of the environment; however, through the inclusion of semantic knowledge more information can be inferred. Let $\mathcal{C}$ be the set of semantic class labels. Define $\mathcal{S} \triangleq \{s_k\}_{k=1}^n$, where $s_k \in \mathcal{C}$. $\mathcal{S}$ represents the semantic class labels of points in the environment. Now $\mathcal{I}$ also encodes the association of a pair of points, one in $\mathcal{X}_s$ and one in $\mathcal{X}_t$, to a semantic label $s_k \in \mathcal{S}$. The joint distribution of the residuals $\mathcal{R}$, semantics $\mathcal{S}$, and association $\mathcal{I}$, conditioned on the source and target point clouds is $f(\mathbf{T};\mathcal{R},\mathcal{S},\mathcal{I}|\mathcal{X}_t,\mathcal{X}_s) \triangleq \log p(\mathcal{R},\mathcal{S},\mathcal{I}|\mathcal{X}_t,\mathcal{X}_s)$. Thus, if the assumption of know associations is removed, the semantic point cloud registration is an optimization over the log-likelihood as follows.

Problem 2 (Semantic point cloud registration)**.** *Let $\mathcal{X}_t$ and $\mathcal{X}_s$ be two independent, overlapping point clouds. Let $\mathcal{S}$ be the semantic labels of the environment observed by the point clouds. The optimal transformation that aligns the source to the target can be computed by solving the following MLE problem:*

$$\max_{\mathbf{T}\in\mathrm{SE}(3)} \; f(\mathbf{T};\mathcal{R},\mathcal{S},\mathcal{I}|\mathcal{X}_t,\mathcal{X}_s) \tag{2}$$

4 Generalized ICP on SE(3)

Segal et al. [39] modeled measurements in the target and source clouds as being drawn from Gaussian distributions, i.e., $\mathbf{x}_k^t \sim \mathcal{N}(\hat{\mathbf{x}}_k^t, \Sigma_k^t)$, and $\mathbf{x}_k^s \sim \mathcal{N}(\mathbf{T}(\hat{\mathbf{x}}_k^s), \Sigma_k^s)$, respectively. Therefore, the residual log-likelihood, excluding the normalization constant, becomes $f_{\texttt{GICP}}(\mathbf{T};\mathcal{R}|\mathcal{X}_t,\mathcal{X}_s,\mathcal{I}) \triangleq \sum_{k=1}^n ||\mathbf{x}_k^t - \mathbf{T}(\mathbf{x}_k^s)||^2_{\mathbf{C}_k}$ where $\mathbf{C}_k \triangleq \Sigma_k^t + \mathbf{R}\Sigma_k^s\mathbf{R}^\top$. The analytical gradient of this cost function in the ambient Euclidean space is $\frac{\partial f_{\texttt{GICP}}}{\partial \mathbf{p}} = \sum_{k=1}^n -2\mathbf{C}_k^{-1}\mathbf{r}_k$ with respect to the translation and $\frac{\partial f_{\texttt{GICP}}}{\partial \mathbf{R}} = \sum_{k=1}^n -2\mathbf{C}_k^{-1}\mathbf{r}_k(\mathbf{x}_k^{s\top} + \mathbf{r}_k^\top\mathbf{C}_k^{-1}\mathbf{R}\Sigma_k^s)$ with respect to rotation. This sets up Problem 1 as an optimization over the SE(3) manifold. While it does not change the formulation, optimizing over SE(3) is a more efficient way to parametrize the optimization problem with respect to the nature of the rigid body transformation, and by itself shows improvements over the Euler angle parametrization used in the original implementation of GICP [38]. Our approach follows that of Absil et al. [1], i.e., lifting the problem on to the tangent space of the Lie group, solving the reparametrized problem, then retracting it back to the manifold. For more details see Appendix A.

The original GICP algorithm removes residuals larger than a certain value to ensure that any point in the source cloud which does not have a counterpart will not affect the solution. To avoid having a hard threshold, we replace this step with a robust estimator using the

Algorithm 1 GICP-SE(3)

Require: Initial transformation $\mathbf{T}^{\text{init}}$, target point cloud $\mathcal{X}_t$, source point cloud $\mathcal{X}_s$;
1: $\mathbf{T}^* \leftarrow \mathbf{T}^{\text{init}}$
2: **while** not converged **do**
3: $\quad\mathbf{T}^{\text{old}} \leftarrow \mathbf{T}^*$
4: $\quad\mathcal{I} \leftarrow \texttt{nnsearch}(\mathcal{X}_s, \mathcal{X}_t, \mathbf{T}^{\text{old}})$ // Find Association
5: $\quad\mathbf{T}^* \leftarrow \arg\max_{\mathbf{T}\in \text{SE}(3)} f_{\texttt{GICP}}(\mathbf{T}; \mathcal{R}|\mathcal{X}_t, \mathcal{X}_s, \mathcal{I})$ // Optimize over SE(3)
6: $\quad$**if** $\text{d}_{\text{SE}(3)}(\mathbf{T}^{\text{old}}, \mathbf{T}^*) < \varepsilon$ **then** // Check convergence using distance threshold ε
7: $\quad\quad$converged $\leftarrow$ true
8: $\quad$**end if**
9: **end while**
10: **return** $\mathbf{T}^*$

Cauchy loss function, $\rho_\alpha(x) = \alpha^2 \ln(1 + \frac{x}{\alpha^2})$, where α is a parameter that controls where the loss begins to scale sublinearly. Similar to the approach in GICP, the robust estimator diminishes the effect of outliers while avoiding removal of potential inliers. Consequently, our cost function becomes

$$f_{\texttt{GICP}}(\mathbf{T}; \mathcal{R}|\mathcal{X}_t, \mathcal{X}_s, \mathcal{I}) = \sum_k^n \rho_\alpha(||\mathbf{x}_k^t - \mathbf{T}(\mathbf{x}_k^s)||^2_{\mathbf{C}_k}) \tag{3}$$

and the effect of the loss function on the gradient is trivial to derive using the chain rule. The algorithmic implementation of GICP-SE(3) is shown in Algorithm 1. For **Step 1** in the ICP framework, finding the association is done using a NN search in line 4. In **Step 2** (3) is solved by the lift-solve-retract scheme over the SE(3) Lie group as described in Appendix A. The inner loops are stopped once the change in $\mathbf{T}^*$ is less than a distance threshold ε. We use a distance metric on SE(3) defined as $\text{d}_{\text{SE}(3)}(\mathbf{T}_1, \mathbf{T}_2) \triangleq \|\log(\mathbf{T}_1\mathbf{T}_2^{-1})^\vee\|$ where $\log(\cdot)$ computes matrix logarithm. We will also use this metric in the evaluations. Figure 2 shows the convergence of GICP-SE(3) compared to that of GICP. The parametrization over SE(3) leads to convergence in fewer iterations versus Euler angles.

5 Semantic Iterative Closest Point

Problem 2 frames the semantic point cloud registration problem as an optimization of the joint log-likelihood. However, both semantic class, $\mathcal{S}$, and association, $\mathcal{I}$, are in fact latent variables. The domain of the semantic random variables are small in this work and we can directly marginalize them; unfortunately, the same does not hold for associations. Following an EM approach, the joint likelihood is $p(\mathcal{R}, \mathcal{S}, \mathcal{I}|\mathcal{X}_t, \mathcal{X}_s)$. We now assume, given the semantic class and association, the points $\mathbf{x}_k^t$ and $\mathbf{x}_k^s$ have independent noise, i.e., they are independent measurements. Together with applying Bayes' rule, it is easy to show that

$$p(\mathcal{R}, \mathcal{S}, \mathcal{I}|\mathcal{X}_t, \mathcal{X}_s) = p(\mathcal{R}|\mathcal{S}, \mathcal{I}, \mathcal{X}_t, \mathcal{X}_s) p(\mathcal{S}|\mathcal{I}, \mathcal{X}_t) p(\mathcal{S}|\mathcal{I}, \mathcal{X}_s) \frac{p(\mathcal{I}|\mathcal{X}_t, \mathcal{X}_s)}{p(\mathcal{S}|\mathcal{I})} \tag{4}$$

where $p(\mathcal{S}|\mathcal{I})$ can be seen as an uninformative prior term.

Similar to McCormac et al. [28], given the point cloud and association, we use a CNN to model the per point semantic observation term, $p(\mathcal{S}|\mathcal{I}, \mathcal{X})$. We model $p(\mathcal{R}|\mathcal{S}, \mathcal{I}, \mathcal{X}_t, \mathcal{X}_s)$ using the same Gaussian distribution as in GICP [39]. Since the residual is a function of the

$\mathbf{x}_k^t$, $\mathbf{x}_k^s$, and the association i_k, it is independent of the semantic class given those variables, and we can simplify $p(\mathcal{R}|\mathcal{S},\mathcal{I},\mathcal{X}_t,\mathcal{X}_s)$ to $p(\mathcal{R}|\mathcal{I},\mathcal{X}_t,\mathcal{X}_s)$. Consequently, (4) simplifies to

$$p(\mathcal{R},\mathcal{S},\mathcal{I}|\mathcal{X}_t,\mathcal{X}_s) \propto \underbrace{p(\mathcal{R}|\mathcal{I},\mathcal{X}_t,\mathcal{X}_s)}_{\text{residual}} \underbrace{p(\mathcal{S}|\mathcal{I},\mathcal{X}_t)}_{\text{target semantic}} \underbrace{p(\mathcal{S}|\mathcal{I},\mathcal{X}_s)}_{\text{source semantic}} \underbrace{p(\mathcal{I}|\mathcal{X}_t,\mathcal{X}_s)}_{\text{geometric association}} \tag{5}$$

Our approach in **Step 2** differs from Section 4 in how we handle the latent variable that represents the associations. The standard nearest neighbor approach can be seen as a heuristic of picking the geometrically closest point as a *hard association*. In contrast, we defined the geometric association, $p(\mathcal{I}|\mathcal{X}_t,\mathcal{X}_s) = \prod_{k=1}^{n} p(i_k|\mathbf{x}_k^t,\mathbf{x}_k^s)$, as

$$p(i_k|\mathbf{x}_k^t,\mathbf{x}_k^s) \triangleq \begin{cases} \frac{1}{N} & \text{if } \mathbf{x}_k^t \text{ is } N \text{ nearest neighbors of } \mathbf{x}_k^s \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

The EM approach to infer the latent variables and the optimal transformation, $\mathbf{T}$, can be expressed as follows:

– *Expectation:* We wish to compute the expected value of the log-likelihood function with respect to the probability of the association given the current transformation and point clouds, or the $Q(\cdot,\cdot)$ function.

$$\begin{aligned} Q(\mathbf{T},\mathbf{T}^{\text{old}}) &= \mathbb{E}_{p(\mathcal{I}|\mathcal{R},\mathcal{S},\mathcal{X}_t,\mathcal{X}_s;\mathbf{T}^{\text{old}})}[\log p(\mathcal{R},\mathcal{S},\mathcal{I}|\mathcal{X}_t,\mathcal{X}_s;\mathbf{T})] \\ &= \sum_{\mathcal{I}\in\mathbb{I}} p(\mathcal{I}|\mathcal{R},\mathcal{S},\mathcal{X}_t,\mathcal{X}_s;\mathbf{T}^{\text{old}}) \log p(\mathcal{R}|\mathcal{I},\mathcal{X}_t,\mathcal{X}_s;\mathbf{T}) + \text{const.} \end{aligned} \tag{7}$$

– *Maximization:* We wish to maximize $Q(\cdot,\cdot)$ over the transformation variable $\mathbf{T}$

$$\mathbf{T}^* = \underset{\mathbf{T}\in\text{SE}(3)}{\arg\max}\, Q(\mathbf{T},\mathbf{T}^{\text{old}}) \tag{8}$$

We can see that (7) is log-likelihood of (4). The probability of the latent variable given data and the current transformation estimate, $p(\mathcal{I}|\mathcal{R},\mathcal{S},\mathcal{X}_t,\mathcal{X}_s;\mathbf{T}^{\text{old}})$, can be expanded as

$$\begin{aligned} p(\mathcal{I}|\mathcal{R},\mathcal{S},\mathcal{X}_t,\mathcal{X}_s;\mathbf{T}^{\text{old}}) &= \frac{p(\mathcal{R},\mathcal{S},|\mathcal{I},\mathcal{X}_t,\mathcal{X}_s,;\mathbf{T}^{\text{old}})p(\mathcal{I}|\mathcal{X}_t,\mathcal{X}_s)}{\sum_{\tilde{\mathcal{I}}\in\mathbb{I}} p(\mathcal{R},\mathcal{S}|\tilde{\mathcal{I}},\mathcal{X}_t,\mathcal{X}_s,;\mathbf{T}^{\text{old}})p(\tilde{\mathcal{I}}|\mathcal{X}_t,\mathcal{X}_s)} \\ &\triangleq \eta\; p(\mathcal{R},\mathcal{S}|\mathcal{I},\mathcal{X}_t,\mathcal{X}_s,;\mathbf{T}^{\text{old}})p(\mathcal{I}|\mathcal{X}_t,\mathcal{X}_s) \end{aligned} \tag{9}$$

where η is constant with respect to $\mathcal{I}$. Using (5), we calculate a weight based on the conditional probability of every possible association, denoted by $i_k \in \mathbb{I}$, excluding the normalization constant η and uninformative priors, as follows.

$$w_k \triangleq \sum_{s_k\in\mathcal{C}} p(\mathbf{r}_k|\mathbf{x}_k^t,\mathbf{x}_k^s,i_k;\mathbf{T}^{\text{old}})p(s_k|\mathcal{X}_t,i_k)p(s_k|\mathcal{X}_s,i_k)p(i_k|\mathbf{x}_k^t,\mathbf{x}_k^s) \tag{10}$$

We combine the weights from (10) into a weight array, $\mathbf{w} = \text{vec}(w_1,\ldots,w_{n\times N})$, that is $n\times N$ counting non-zero weights. Subsequently, the maximization step becomes

$$\mathbf{T}^* = \underset{\mathbf{T}\in\text{SE}(3)}{\arg\max}\, f_{\texttt{SICP}}(\mathbf{T},\mathbf{w};\mathcal{R}|\mathcal{X}_t,\mathcal{X}_s,\mathcal{I}) \triangleq \underset{\mathbf{T}\in\text{SE}(3)}{\arg\max} \sum_{k=1}^{n\times N} \rho_\alpha(w_k||\mathbf{x}_k^t - \mathbf{T}(\mathbf{x}_k^s)||^2_{\mathbf{C}_k}) \tag{11}$$

The algorithmic implementation of semantic ICP is shown in Algorithm 2. The steps are similar to the presented GICP-SE(3) and the main difference is in line 4 where the weights are calculated, turning 5 into a weighted non-linear least squares problem over SE(3).

Algorithm 2 Semantic ICP

Require: Initial transformation $\mathbf{T}^{\text{init}}$, target point cloud $\mathcal{X}_t$, source point cloud $\mathcal{X}_s$, semantic labels;
$\mathbf{T}^* \leftarrow \mathbf{T}^{\text{init}}$
while not converged **do**
  $\mathbf{T}^{\text{old}} \leftarrow \mathbf{T}^*$
  $\mathbf{w} \leftarrow$ Compute weights using (10) // Expectation
  $\mathbf{T}^* \leftarrow \arg\max_{\mathbf{T}\in \text{SE}(3)} f_{\texttt{SICP}}(\mathbf{T}, \mathbf{w}; \mathcal{R}|\mathcal{X}_t, \mathcal{X}_s, \mathcal{I})$ // Maximization: Optimize over SE(3)
  if $\text{d}_{\text{SE}(3)}(\mathbf{T}^{\text{old}}, \mathbf{T}^*) < \varepsilon$ **then** // Check convergence using distance threshold ε
    converged $\leftarrow$ true
  end if
end while
return $\mathbf{T}^*$

6 Evaluation

We evaluate the performance of our proposed method by comparing the relative transformation error of the algorithms on two open source datasets, namely the KITTI Vision benchmark dataset [20], and SceneNet RGBD Dataset [29]. We use two different CNNs for semantic inference on these datasets as one shows good performance on outdoor images while the other on the indoor scenes in our experiments. We show how using semantics in association and optimizing over SE(3) benefit the point cloud registration task by comparing the following algorithms: Generalized ICP [39] available in the Point Cloud Library [36], GICP-SE(3) described in Section 4 and Algorithm 1, Semantic ICP as described in Section 5 and Algorithm 2, and finally, the EM approach in Iterative Probabilistic Data Association (IPDA) [18].

6.1 KITTI Visual Odometry Dataset

We use the KITTI visual odometry dataset [20] to evaluate the accuracy of the estimated transformations on the stereo data available as part of the dataset. The semantic inference is performed using Dilation CNN [45]. Further explanation of data preprocessing can be found in Appendix C. We use the distance metric $\text{d}_{\text{SE}(3)}(\cdot,\cdot)$ to compare estimated transformations to the ground truth trajectory provided in the dataset. We also provide $\text{d}_{\text{SO}(3)}(\cdot,\cdot)$ and $\text{d}_{\mathbb{R}^3}(\cdot,\cdot)$ distances in Table 1. We define those distances as $\text{d}_{\text{SO}(3)}(\mathbf{T}_1,\mathbf{T}_2) = \|\log(\mathbf{R}_1\mathbf{R}_2^\top)^\vee\|$ and $\text{d}_{\mathbb{R}^3}(\mathbf{T}_1,\mathbf{T}_2) = \|\mathbf{t}_1 - \mathbf{R}_1\mathbf{R}_2^\top\mathbf{t}_2\|$ which are consistent with the $\text{d}_{\text{SE}(3)}(\cdot,\cdot)$ definition.

We were only able to run IPDA on a subset of the dataset due to its high processing time. Table 1 summarizes the quantitative results on that subset. It shows that changing the parametrization of the optimization improved the results. In addition, adding semantics to aid the association problem further improved the results. The IPDA algorithm performed similarly to GICP, and it would potentially perform better when used to align two point clouds of varying density. The table also includes average runtimes, which are slower for the EM-based approaches. This is expected because soft associations add a factor more terms (equal to the number of neighbors of each point considered) to the cost function summation.

In Figure 3, to show the distribution of errors, we also plot the cumulative distribution function and box plots of the $\text{d}_{\text{SE}(3)}(\cdot,\cdot)$ for each algorithm. The plots show that both Semantic ICP and GICP-SE(3) outperform GICP regarding their best two quartiles. We can also observe that Semantic ICP starts to outperform GICP-SE(3) in its final quartile, which

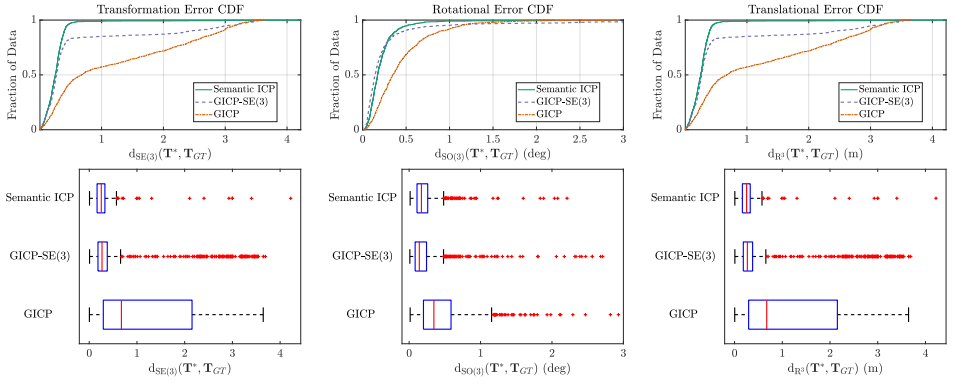

Figure 3: Error CDF and box plots of the proposed algorithms compared with GICP computed using KITTI sequence 05 dataset [20]. The metrics used for comparison are $d_{SE(3)}(\cdot,\cdot)$, $d_{SO(3)}(\cdot,\cdot)$, $d_{\mathbb{R}^3}(\cdot,\cdot)$. The proposed algorithms, Semantic ICP and GICP-SE(3), show better performance by exploiting the structure of SE(3).

Table 1: KITTI results with the distance metrics and runtime. Best results for each column are in **bold**.

Algorithm	**Tranfomation Error** $d_{SE(3)}(\mathbf{T}^*, \mathbf{T}_{GT})$		**Rotational Error** $d_{SO(3)}(\mathbf{T}^*, \mathbf{T}_{GT})$ (deg)		**Translation Error** $d_{\mathbb{R}^3}(\mathbf{T}^*, \mathbf{T}_{GT})$ (m)		**Runtime** (s)	
	Mean	Median	Mean	Median	Mean	Median	Mean	Median
Semantic ICP	**0.2619**	**0.2078**	**0.2041**	0.1561	**0.2618**	**0.2078**	109.4	101.5
GICP-SE(3)	0.4923	0.2295	0.2467	**0.1308**	0.4922	0.2295	38.6	36.5
GICP	0.9095	0.4860	0.4200	0.3264	0.9094	0.4869	**12.5**	**12.0**
IPDA	1.1808	0.8732	1.2830	0.8341	1.1798	0.8731	2672.0	2555.0

accounts for its better mean value. Since all algorithms use gradient-based optimizers, the initialization affects the accuracy of results. To explore how our approach influences the basin of convergence, we plot the initial offset versus the final offset in Figure 4. We find that using SE(3) parametrization and incorporating semantic information improve convergence.

6.2 SceneNet RGBD Dataset

The SceneNet RGBD dataset is a synthetic rendered dataset that provides pixel level semantics and ground truth depth and camera trajectory [27, 29]. The dataset was made by randomly generating indoor scenes and placing models of household objects in rooms. Random trajectories are then sampled and synthetic images are rendered. The dataset gives the ability to evaluate the compared registration algorithms to a known ground truth trajectory. For semantic inference, we trained DeepLab-ResNet [11] on the SceneNet RGBD training data. The training procedure is described in Appendix C. Each algorithm was used to align consecutive point clouds in the dataset and $d_{SE(3)}(\cdot)$, $d_{SO(3)}(\cdot)$, and $d_{\mathbb{R}^3}(\cdot)$ were collected with respect to the provided ground truth. The results are summarized in Table 2. The mean values are tighter than those of the KITTI visual odometry dataset, and larger than the medians, indicating a significant tail of errors. This is most likely caused by strong geometric features, such as perpendicular wall and ceiling, either being correctly associated or, in some

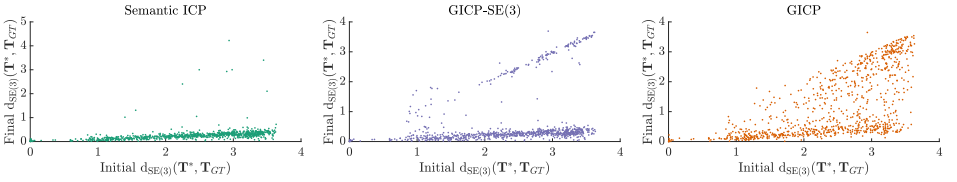

Figure 4: Scatter plots of the initial alignment vs. final alignment using $d_{SE(3)}(\cdot, \cdot)$ for each algorithm on the KITTI visual odometry dataset. We can see that GICP is less likely to converge as the initial offset gets larger, while Semantic ICP and GICP-SE(3) are more of a bimodal distribution, either staying near the initial transformation, or converging.

Table 2: SceneNet RGBD results with the distance metrics and runtime. Best result for each column is in **bold**.

Algorithm	**Tranfomation Error** $d_{SE(3)}(\mathbf{T}^*, \mathbf{T}_{GT})$		**Rotational Error** $d_{SO(3)}(\mathbf{T}^*, \mathbf{T}_{GT})$ (deg)		**Translation Error** $d_{\mathbb{R}^3}(\mathbf{T}^*, \mathbf{T}_{GT})$ (m)		**Runtime** (s)	
	Mean	Median	Mean	Median	Mean	Median	Mean	Median
Semantic ICP	**0.4430**	**0.0377**	**9.98**	**0.5339**	**0.3778**	**0.0349**	54.0	32.0
GICP-SE(3)	0.4602	0.0443	10.70	0.6874	0.3878	0.0425	15.2	9.0
GICP	0.4582	0.0629	10.29	1.04	0.3915	0.0598	**2.8**	**2.0**

outlier cases, completely miss-associated. Nevertheless, the Semantic ICP algorithm shows a quantitative improvement over GICP-SE(3). Further results are shown in Appendix E.

7 Conclusion

In this paper, we proposed a novel algorithm for the point cloud registration problem that is based on the joint semantic and geometric inference. Our proposed Semantic ICP algorithm treats point associations as latent random variables leading to an EM-style solution. We showed semantic labels together with EM data associations improves the algorithm's performance in comparison with standard GICP and our GICP-SE(3). This evaluation was performed on two publicly available datasets. The extension of this work to a framework for semantic SLAM or an odometry system for real-time applications is an interesting future direction [42, 44]. Extensions to optimizing over multiple rigid body transformations to compensate for dynamic objects in the scene is also an interesting direction.

Acknowledgements

This work was partially supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE1256260, and by the Toyota Research Institute (TRI), partly under award number N021515, however, this article solely reflects the opinions and conclusions of its authors and not TRI or any other Toyota entity.

References

[1] P-A Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2009.

[2] Sameer Agarwal, Keir Mierle, and Others. Ceres solver. http://ceres-solver.org.

[3] Sid Yingze Bao and Silvio Savarese. Semantic structure from motion. In *Proc. IEEE Int. Conf. Computer Vision and Pattern Recog.*, pages 2025–2032. IEEE, 2011.

[4] Timothy D Barfoot and Paul T Furgale. Associating uncertainty with three-dimensional poses for use in estimation problems. *IEEE Trans. Robot.*, 30(3):679–693, 2014.

[5] Paul J Besl and Neil D McKay. Method for registration of 3-d shapes. In *Sensor Fusion IV: Control Paradigms and Data Structures*, volume 1611, pages 586–607. International Society for Optics and Photonics, 1992.

[6] Peter Biber and Wolfgang Straßer. The normal distributions transform: A new approach to laser scan matching. In *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, volume 3, pages 2743–2748. IEEE, 2003.

[7] Sean L Bowman, Nikolay Atanasov, Kostas Daniilidis, and George J Pappas. Probabilistic data association for semantic SLAM. In *Proc. IEEE Int. Conf. Robot. Automat.*, pages 1722–1729, 2017.

[8] Gabriel Brostow, Jamie Shotton, Julien Fauqueur, and Roberto Cipolla. Segmentation and recognition using structure from motion point clouds. *European Conf. on Computer Vision*, pages 44–57, 2008.

[9] Robert O Castle, Darren J Gawley, Georg Klein, and David W Murray. Towards simultaneous recognition, localization and mapping for hand-held and wearable cameras. In *Proc. IEEE Int. Conf. Robot. Automat.*, pages 4102–4107, 2007.

[10] Andrea Censi. An ICP variant using a point-to-line metric. In *Proc. IEEE Int. Conf. Robot. Automat.*, pages 19–25. IEEE, 2008.

[11] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. DeepLab: Semantic image segmentation with deep convolutional nets, Atrous convolution, and fully connected CRFs. 2016.

[12] Yang Chen and Gérard Medioni. Object modeling by registration of multiple range images. In *Proc. IEEE Int. Conf. Robot. Automat.*, pages 2724–2729. IEEE, 1991.

[13] Gregory S Chirikjian. *Stochastic Models, Information Theory, and Lie Groups, Volume 2: Analytic Methods and Modern Applications*. Springer Science & Business Media, 2011.

[14] Siddharth Choudhary, Alexander JB Trevor, Henrik I Christensen, and Frank Dellaert. SLAM with object discovery, modeling and mapping. In *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, pages 1018–1025, 2014.

[15] Javier Civera, Dorian Gálvez-López, Luis Riazuelo, Juan D Tardós, and JMM Montiel. Towards semantic SLAM using a monocular camera. In *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, pages 1277–1284, 2011.

[16] Camille Couprie, Clément Farabet, Laurent Najman, and Yann LeCun. Indoor semantic segmentation using depth information. *arXiv preprint arXiv:1301.3572*, 2013.

[17] Simone Fontana, Timo Hinzmann, and Gabriel Agamennoni. Iterative Probabilistic Data Association. https://github.com/ethz-asl/robust_point_cloud_registration, 2016. [Online; accessed 30-January-2018].

[18] Roland Y. Siegwart Gabriel Agamennoni, Simone Fontana and Domenico G. Sorrenti. Point clouds registration with probabilistic data association. In *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, pages 4092–4098. IEEE, 2016.

[19] Andreas Geiger, Martin Roser, and Raquel Urtasun. Efficient large-scale stereo matching. In *Asian Conf. Computer Vision*, 2010.

[20] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *Proc. IEEE Int. Conf. Computer Vision and Pattern Recog.*, 2012.

[21] Sébastien Granger, Xavier Pennec, and Alexis Roche. Rigid point-surface registration using an em variant of icp for computer guided oral implantology. In Wiro J. Niessen and Max A. Viergever, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2001*, pages 752–761, Berlin, Heidelberg, 2001. Springer Berlin Heidelberg.

[22] Adrien Gressin, Clément Mallet, Jérôme Demantké, and Nicolas David. Towards 3d lidar point cloud registration improvement using optimal neighborhood knowledge. *ISPRS journal of photogrammetry and remote sensing*, 79:240–251, 2013.

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. IEEE Int. Conf. Computer Vision and Pattern Recog.*, pages 770–778, 2016.

[24] Bhoram Lee and Daniel D Lee. Learning anisotropic icp (la-icp) for robust and efficient 3d registration. In *Proc. IEEE Int. Conf. Robot. Automat.*, pages 5040–5045. IEEE, 2016.

[25] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conf. on Computer Vision*, pages 740–755. Springer, 2014.

[26] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proc. IEEE Int. Conf. Computer Vision and Pattern Recog.*, pages 3431–3440, 2015.

[27] John McCormac, Ankur Handa, Stefan Leutenegger, and Andrew J Davison. Scenenet RGB-D: 5M photorealistic images of synthetic indoor trajectories with ground truth. 2016.

[28] John McCormac, Ankur Handa, Andrew J Davison, and Stefan Leutenegger. Semanticfusion: Dense 3d semantic mapping with convolutional neural networks. In *Proc. IEEE Int. Conf. Robot. Automat.*, pages 4628–4635, May 2017.

[29] 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. IEEE Int. Conf. Computer Vision*, 2017.

[30] Richard M Murray, Zexiang Li, S Shankar Sastry, and S Shankara Sastry. *A mathematical introduction to robotic manipulation*. CRC press, 1994.

[31] Sudeep Pillai and John Leonard. Monocular SLAM supported object recognition. In *Robotics: Science and Systems*, Rome, Italy, July 2015.

[32] François Pomerleau, Francis Colas, Roland Siegwart, and Stéphane Magnenat. Comparing ICP variants on real-world data sets. *Auton. Robot*, 34(3):133–148, 2013.

[33] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3D classification and segmentation. *arXiv preprint arXiv:1612.00593*, 2016.

[34] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in Neural Information Processing Systems 30*, pages 5105–5114. 2017.

[35] Szymon Rusinkiewicz and Marc Levoy. Efficient variants of the ICP algorithm. In *3-D Digital Imaging and Modeling, 2001. Proceedings. Third International Conference on*, pages 145–152. IEEE, 2001.

[36] Radu Bogdan Rusu and Steve Cousins. 3D is here: Point Cloud Library (PCL). In *Proc. IEEE Int. Conf. Robot. Automat.*, Shanghai, China, May 9-13 2011.

[37] Renato F Salas-Moreno, Richard A Newcombe, Hauke Strasdat, Paul HJ Kelly, and Andrew J Davison. SLAM++: Simultaneous localisation and mapping at the level of objects. In *Proc. IEEE Int. Conf. Computer Vision and Pattern Recog.*, pages 1352–1359, 2013.

[38] Aleksandr Segal. Generalized-ICP. https://github.com/avsegal/gicp, 2009. [Online; accessed 30-January-2018].

[39] Aleksandr Segal, Dirk Haehnel, and Sebastian Thrun. Generalized-ICP. In *Robotics: Science and Systems*, volume 2, 2009.

[40] Berk Sevilmis and Benjamin Kimia. Shape-based image correspondence. In Edwin R. Hancock Richard C. Wilson and William A. P. Smith, editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 66.1–66.12. BMVA Press, September 2016. ISBN 1-901725-59-6. doi: 10.5244/C.30.66. URL https://dx.doi.org/10.5244/C.30.66.

[41] Jamie Shotton, Matthew Johnson, and Roberto Cipolla. Semantic texton forests for image categorization and segmentation. In *IEEE Conf. on Computer vision and pattern recognition*, pages 1–8. IEEE, 2008.

[42] Rafael Valencia, Ernesto H Teniente, Eduard Trulls, and Juan Andrade-Cetto. 3D mapping for urban service robots. In *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, pages 3076–3081, 2009.

[43] Martin Weinmann and Boris Jutzi. Geometric point quality assessment for the automated, markerless and robust registration of unordered tls point clouds. *ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences*, 2, 2015.

[44] Ryan W. Wolcott and Ryan M. Eustice. Robust LIDAR localization using multiresolution Gaussian mixture maps for autonomous driving. *The Int. J. Robot. Res.*, 36: 292–319, 3 2017.

[45] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. In *ICLR*, 2016.

[46] Fisher Yu, Jianxiong Xiao, and Thomas Funkhouser. Semantic alignment of LiDAR data at city scale. In *Proc. IEEE Int. Conf. Computer Vision and Pattern Recog.*, pages 1722–1731, 2015.

[47] Anestis Zaganidis, Martin Magnusson, Tom Duckett, and Grzegorz Cielniak. Semantic-assisted 3D normal distributions transform for scan registration in environments with limited structure. In *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, September 2017.