

StickyPillars: Robust and Efficient Feature Matching on Point Clouds using Graph Neural Networks

Kai Fischer¹, Martin Simon¹, Florian Ölsner², Stefan Milz^{2,4}, Horst-Michael Groß³, Patrick Mäder⁴

¹ Valeo Schalter und Sensoren GmbH, Kronach, Germany

² Spleenlab GmbH, Saalburg-Ebersdorf, Germany

³ Neuroinformatics and Cognitive Robotics Lab, Ilmenau University of Technology, Germany

⁴ Software Engineering for Safety-Critical Systems, Ilmenau University of Technology, Germany

{kai.fischer, martin.simon}@valeo.com

Abstract

Robust point cloud registration in real-time is an important prerequisite for many mapping and localization algorithms. Traditional methods like ICP tend to fail without good initialization, insufficient overlap or in the presence of dynamic objects. Modern deep learning based registration approaches present much better results, but suffer from a heavy runtime. We overcome these drawbacks by introducing StickyPillars, a fast, accurate and extremely robust deep middle-end 3D feature matching method on point clouds. It uses graph neural networks and performs context aggregation on sparse 3D key-points with the aid of transformer based multi-head self and cross-attention. The network output is used as the cost for an optimal transport problem whose solution yields the final matching probabilities. The system does not rely on hand crafted feature descriptors or heuristic matching strategies. We present state-of-the-art accuracy results on the registration problem demonstrated on the KITTI dataset while being four times faster than leading deep methods. Furthermore, we integrate our matching system into a LiDAR odometry pipeline yielding most accurate results on the KITTI odometry dataset. Finally, we demonstrate robustness on KITTI odometry. Our method remains stable in accuracy where state-of-the-art procedures fail on frame drops and higher speeds.

1. Introduction

Point cloud registration, the process of finding a spatial transformation aligning two point clouds, is an essential computer vision problem and a precondition for a wide range of tasks in the domain of real-time scene understanding or applied robotics, such as odometry, mapping, re-localization or SLAM. New generations of 3D sensors, like depth cam-

eras or LiDARs (light detection and ranging), as well as multi-sensor setups provide substantially more fine-grained and reliable data enabling dense range perception at a large field of view. These sensors substantially increase the expectations on point cloud registration and an exact matching of feature correspondences.

State-of-the-art 3D point cloud registration employs locally describable features in a global optimization step [43, 57, 21]. Most methods do not rely on modern machine learning algorithms, although they are part of the best performing approaches on odometry challenges like KITTI [14]. In contrast, recent research for point cloud processing, e.g. classification and segmentation [34, 35, 18, 60], relies on neural networks and promises substantial improvements for registration, mapping and odometry [12, 20]. The limitation of all none neural network-based odometry and mapping methods is that they perform odometry estimation using a global rigid body operation. Those approaches assume many static objects within the environment and proper viewpoints. However, real world measurements are generally unstable under challenging situations, e.g. many dynamic objects or widely varying viewpoints and small overlapping areas. Hence, the mapping quality itself is suffering from artifacts (blurring) and is often not evaluated qualitatively. To overcome these limitations, we propose StickyPillars a novel registration approach for point clouds utilizing graph neural networks based on pillar shaped point descriptors. Inspired by [9, 52], our approach computes feature correspondence rather than end-to-end odometry estimations. By reducing the registration problem to an inter-cloud correspondence search on a sparse subset of selected key-points and utilizing a ground truth match matrix in the training process, we are able to predict poses with a very low computation time while being robust against the influence of dynamic objects. We demonstrate StickyPillars's robust real-time registration capabilities (see Fig. 1) and its confidence under challeng-

ing conditions, such as dynamic environments, challenging viewpoints and small overlapping areas. We evaluate our technique on the KITTI odometry benchmark [14] and significantly outperform state-of-the-art frame to frame matching approaches e.g. the one used in LOAM[57]. Those improvements enable more precise odometry estimation for applied robotics.

2. Related Work

Deep Learning on point clouds is a novel field of research [7, 45, 44]. The fact that points are typically stored in unordered sets and the need for viewpoint invariance prohibits the direct use of classical CNN architectures. Existing solutions tackle this problem by converting the point cloud into a voxel grid [60], by projecting it to a sphere [26, 20] or by directly operating on the set using well designed network architectures [34, 35].

Point cloud registration aims to find the relative rigid 3D transformation that aligns two sets of points representing the same 3D structure. Most established methods try to find feature correspondences in both point clouds. The transformation is found by minimizing a distance metric using standard numerical solvers. A simple but powerful approach called iterative closest point (ICP) was exhaustively investigated by [4, 59, 38] and adapted in a wide range of applications [48, 43, 57, 21]. Here each point is a feature which is matched to its closest neighbor in the other point cloud in an iterative process. ICP’s convergence and runtime highly depends on the initial relative pose, the matching accuracy and the overlap [38]. This problem can be avoided by using global feature matching approaches [39, 25, 54, 1, 19] in combination with solvers like RANSAC [13, 36] that are robust to outlier correspondences. Alternatively the transformation can be directly estimated using a neural network as proposed by [20]. Scene flow based methods pursue a different approach by estimating point-wise translational motion vectors instead of a single rigid transformation [10, 50, 22, 35, 16, 23]. These methods naturally handle dynamic and non-rigid movements but are computationally quite demanding.

Lidar odometry estimation is a typical task that involves point cloud registration in real-time. Many of the aforementioned methods [1, 19, 15, 2, 25], especially those using deep neural networks, are not fast enough to deal e.g. with lidar sensors that run at 10-15 Hz. LOAM [57], the currently best performing approach on the KITTI odometry benchmark [14] uses a variant of ICP that operates on a sparse set of feature points located on edge and surface patches. The feature extraction exploits the special scan structure of an rotating lidar sensor. After a rough pairwise matching the point clouds are accurately aligned to a local map. There exist several variants and improvements [21, 43]. In the work by [20] it is shown that the pairwise matching

step can be achieved by an appropriately sized CNN, but their method wasn’t able to reduce the error any further.

Feature based matching is more widely used in the domain of image processing with prominent approaches, such as FLANN [28] and SIFT [24]. The fundamental approach consists of several steps, point detection, feature calculation and matching. Such models based on neighborhood consensus were evaluated by [5, 41, 49, 6] or in a more robust way combined with a solver called RANSAC [13, 36]. Recently, deep learning based approaches, i.e. convolutional neural networks (CNNs), were proposed to learn local descriptors and sparse correspondences [11, 31, 37, 56]. However, these approaches operate on sets of matches and ignore the assignment structure. In contrast, [40] focuses on bundling aggregation, matching and filtering based on novel *Graph Neural Networks*.

3. The StickyPillars Architecture

The idea behind StickyPillars is the development of a robust-point cloud registration and matching algorithm to replace standard methods (e.g. ICP) as most common matcher in applied robotics and computer vision algorithms like odometry, mapping or SLAM. The 3D point cloud features (*pillars*) are flexible and fully composed by learnable parameters. [18] and [60] have proposed a 3D feature learning mechanism for perception tasks. We transform the concept to feature learning within a matching pipeline, but only using sparse sets of key points to ensure real-time capability and leanness. We propose an architecture using graph neural networks to learn geometrical context aggregation of two point sets in an end-to-end manner. The overall architecture is composed by three important layers: 1. *Pillar Layer*, 2. *Graph Neural Network layer* and 3. *Optimal Transport layer* (see Fig. 1).

Problem description Let \mathcal{P}^K and \mathcal{P}^L be two point clouds to be registered. The key-points of those point clouds will be denoted as π_i^K and π_j^L with $\{\pi_0^K, \dots, \pi_n^K\} \subset \mathcal{P}^K$ and $\{\pi_0^L, \dots, \pi_m^L\} \subset \mathcal{P}^L$, while other points will be defined as $\mathbf{x}_k^K \in \mathcal{P}^K$ and $\mathbf{x}_l^L \in \mathcal{P}^L$. Each key-point with index i is associated to a *point pillar*, which can be pictured as a cylinder with an endless height, having a centroid position π_i^K and a center of gravity $\hat{\pi}_i^K$. All points (\mathcal{P}_i^K) within a pillar i are associated with a pillar feature stack $\mathbf{f}_i^K \in \mathbb{R}^D$, with D as pillar encoder input depth. The same applies for π_j^L . $\mathbf{c}_{i,j}$ and $\mathbf{f}_{i,j}$ compose the input for the graph. The overall goal is to find partial assignments $\langle \pi_i^K, \pi_j^L \rangle$ for the optimal re-projection $\tilde{\mathcal{P}}$ with $\tilde{\mathcal{P}} := f_{\pi_j^L \rightarrow \pi_i^K}(\mathcal{P}^L) \approx \mathcal{P}^K$.

3.1. Pillar Layer

Key-Point Selection is the initial part of the pillar layer with the aim to describe a dense point set with a sparse significant subset of key-points to ensure real-time capability.

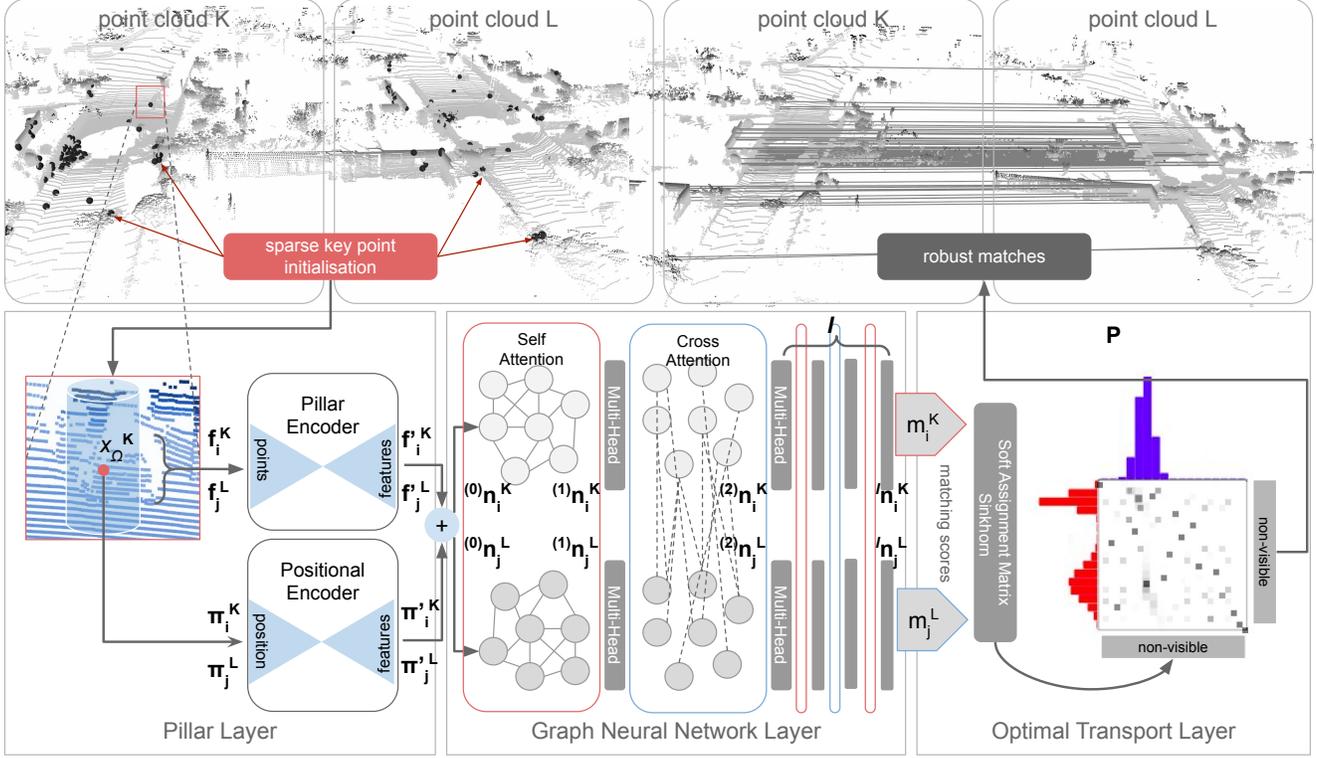


Figure 1: **StickyPillars Architecture** is composed by three layers: 1. *Pillar Layer*, 2. *Graph Neural Network layer* and 3. *Optimal Transport layer*. With the aid of 1, we learn 3D features (*pillar encoder*) and spatial clues (*positional encoder*) directly. In 2 *Self-* and *Cross* Multi-Head Attention is performed in a graph architecture for contextual aggregation. The resulting matching scores are used in 3 to generate an assignment matrix for key-point correspondences via numerical optimal transport.

Most common 3D sensors deliver dense point clouds having more than 120k points. Similar to [57], we place the centroid pillar coordinates on sharp edges and planar surface patches as areas of interest. A smoothness term c identifies smooth or sharp areas. For a point cloud \mathcal{P}^K the smoothness term c^K is defined by:

$$c^K = \frac{1}{|\mathcal{S}| \cdot \|\mathbf{x}_k^K\|} \cdot \left\| \sum_{k' \in \mathcal{S}, k' \neq k} (\mathbf{x}_k^K - \mathbf{x}_{k'}^K) \right\| \quad (1)$$

where k and k' being point indices within the point cloud \mathcal{P}^K having coordinates $\mathbf{x}_k^K, \mathbf{x}_{k'}^K \in \mathbb{R}^3$. \mathcal{S} is a set of neighboring points of k and $|\mathcal{S}|$ is the cardinality of \mathcal{S} . With the aid of the sorted smoothness factors in \mathcal{P}^K , we define two thresholds c_{\min}^K and c_{\max}^K to pick a fixed number n of key-points π_i^K in sharp $c_k^K > c_{\max}^K$ and planar regions $c_k^K < c_{\min}^K$. This is also repeated for the target point-set with c^L on \mathcal{P}^L selecting m points with index j .

Pillar Encoder is designed to learn features in 3D inspired by [34, 18]. Any selected key-point, π_i^K and π_j^L , is associated with a *point pillar* i and j describing a set of specific points \mathcal{P}_i^K and \mathcal{P}_j^L . We sample points into a pillar using

an 2D euclidean distance function (x, y plane) assuming a pillar alignment along the z coordinate (vertical direction) using a projection function $g \rightarrow [x, y, z] = [x, y]$:

$$\mathcal{P}_i^K := \{\mathbf{x}_0^K, \mathbf{x}_\Omega^K, \dots, \mathbf{x}_z^K\} \quad \left\| g(\pi_i^K) - g(\mathbf{x}_\Omega^K) \right\| < d \quad (2)$$

Similar equations apply for \mathcal{P}_j^L . Due to a fixed input size of the pillar encoder, we draw a maximum of z points per pillar, where $z = 100$ is used in our experiments. d is the distance threshold defining the pillar radius (e.g. 50 cm). To enable efficient computation, we organized point clouds within a k -d tree [3]. Based on π_i^K the z closest samples \mathbf{x}_Ω^K are drawn into the pillar \mathcal{P}_i^K , whereas points with a projection distance greater d were rejected.

To compose a sufficient feature input stack for the *pillar encoder* $\mathbf{f}_i^K \in \mathbb{R}^D$, we stack for each sampled point \mathbf{x}_Ω^K with $\Omega \in \{1, \dots, z\}$ in the style of [18]:

$$\mathbf{f}_i^K = \left\{ \left[\mathbf{x}_\Omega^K, i_\Omega^K, (\mathbf{x}_\Omega^K - \hat{\pi}_i^K), \|\mathbf{x}_\Omega^K\|_2, (\mathbf{x}_\Omega^K - \pi_i^K) \right], \dots \right\} \quad (3)$$

$\mathbf{x}_\Omega^K \in \mathbb{R}^3$ denotes sample points' coordinates $(x, y, z)^T$. $i_\Omega^K \in \mathbb{R}$ is a scalar and represents the intensity value (e.g.

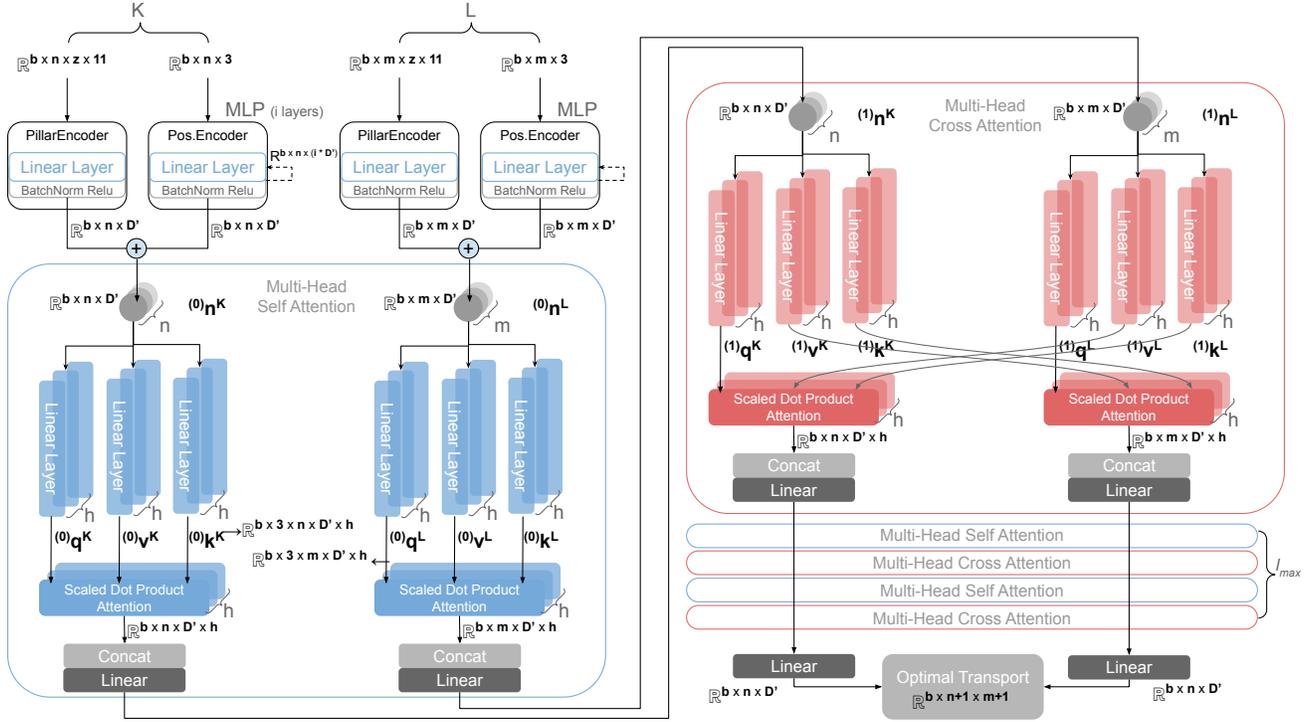


Figure 2: **The StickyPillars Tensor Graph** identifies the data flow throughout the network architecture especially during *self*- and *cross* attention, where b describes the batch-size, n and m the number of pillars, h is the number of heads and l_{\max} the maximum layer depth. D' is the feature depth per node. The result is an assignment matrix \mathbf{P} with an extra column and row for invisible pillars.

LiDAR reflectance), $(\mathbf{x}_\Omega^K - \hat{\pi}_i^K) \in \mathbb{R}^3$ being the difference to the pillar's center of gravity and $(\mathbf{x}_\Omega^K - \pi_i^K) \in \mathbb{R}^3$ is the difference to the pillar's key-point. $\|\mathbf{x}_\Omega^K\|_2 \in \mathbb{R}$ is the L2 norm of the point itself. This leads to an overall input depth $D = z \times 11$. The pillar encoder is a single linear projection layer with shared weights across all pillars and frames followed by a batchnorm and a ReLU layer with an output depth of D' (e.g. 32 in our experiments) and $\mathbf{f}'_i^K, \mathbf{f}'_i^L \in \mathbb{R}^{D'}$:

$$\begin{aligned} \mathbf{f}'_i^K &= W_f \cdot \mathbf{f}_i^K \quad \forall i \in \{1, \dots, n\} \\ \mathbf{f}'_j^L &= W_f \cdot \mathbf{f}_j^L \quad \forall j \in \{1, \dots, m\} \end{aligned} \quad (4)$$

The aim of the **Positional Encoder** is learning geometrical aggregation using a context without applying pooling operations. The positional encoder is inspired by [34] and utilizes a single multi-layer-perceptron (MLP) shared across \mathcal{P}^L and \mathcal{P}^K such as all pillars including batchnorm and ReLU. From the centroid coordinates $\pi_i^K, \pi_j^L \in \mathbb{R}^3$, we calculate positional features via MLP with depth of D' and $\pi'_{iK}, \pi'_{jL} \in \mathbb{R}^{D'}$:

$$\begin{aligned} \pi'_{iK} &= \text{MLP}_\pi(\pi_i^K) \quad \forall i \in \{1, \dots, n\} \\ \pi'_{jL} &= \text{MLP}_\pi(\pi_j^L) \quad \forall j \in \{1, \dots, m\} \end{aligned} \quad (5)$$

3.2. Graph Neural Network Layer

The **Graph Architecture** relies on two complete graphs \mathcal{G}^L and \mathcal{G}^K , whose nodes are related and equivalent to the pillars quantity. The initial $(0)\mathbf{n}_i^K, (0)\mathbf{n}_j^L$ node conditions are denoted as:

$$\begin{aligned} (0)\mathbf{n}_i^K &= \pi'_{iK} + \mathbf{f}'_i^K \\ (0)\mathbf{n}_j^L &= \pi'_{jL} + \mathbf{f}'_j^L \quad (0)\mathbf{n}_i^K, (0)\mathbf{n}_j^L \in \mathbb{R}^{D'} \end{aligned} \quad (6)$$

The overall composed graph $(\mathcal{G}^L, \mathcal{G}^K)$ is a *multiplex* graph inspired by [27, 30]. It is composed by intra-frame edges, i.e. *self* edges connecting each key-point within \mathcal{G}^L and each key-point within \mathcal{G}^K respectively. Additionally, to perform global matching using context aggregation inter-frame edges are introduced, i.e. *cross* edges that connect all nodes of \mathcal{G}^K with \mathcal{G}^L and vice versa.

Multi-Head Self- and Cross-Attention allows us to integrate contextual cues intuitively and increase its distinctiveness considering its spatial and 3D relationship with other co-visible pillars, such as those that are salient, self-similar or statistically co-occurring [40]. An attention function \mathcal{A} [52] is a mapping function of a query and a set of key-point

pairs to an output, with query q , keys k , and values v being vectors. We define attention as:

$$\mathcal{A}(q, k, v) = \text{softmax} \left(\frac{q^T \cdot k}{\sqrt{D'}} \right) \cdot v \quad (7)$$

where D' describes the feature depth analogous to the depth of every node. We apply the multi-head attention function to each node ${}^l \mathbf{n}_i^K, {}^l \mathbf{n}_j^L$ at state l calculating its next condition $l+1$. The node's conditions $l \in \{0, 1, \dots, l_{\max}\}$ are represented as network layers to propagate information to the graph:

$$\begin{aligned} (l+1) \mathbf{n}_i^K &= (l) \mathbf{n}_i^K + (l) \mathcal{M}^K(q_i^K, v_\alpha^\Omega, k_\alpha^\Omega) \\ (l+1) \mathbf{n}_j^L &= (l) \mathbf{n}_j^L + (l) \mathcal{M}^L(q_j^L, v_\beta^\Omega, k_\beta^\Omega) \end{aligned} \quad (8)$$

We alternate the indices for α and β to perform *self* and *cross* attention alternately with increasing depth of l through the network, where the following applies $\Omega \in \{K, L\}$:

$$\alpha, \beta := \begin{cases} i, j & \text{if } l \equiv \text{even} \\ j, i & \text{if } l \equiv \text{odd} \end{cases} \quad (9)$$

The multi-head attention function is defined as:

$$(l) \mathcal{M}^K(q_i^K, v_\alpha^\Omega, k_\alpha^\Omega) = (l) W_0 \cdot (l) (\text{head}_1^K \| \dots \| \text{head}_h^K) \quad (10)$$

with $\|$ being the concatenation operator. A single head is composed by the attention function:

$$\begin{aligned} (l) \text{head}_h^K &= (l) \mathcal{A}(q_i^K, v_\alpha^\Omega, k_\alpha^\Omega) \\ &= (l) \mathcal{A}(W_{1h} \cdot \mathbf{n}_i^K, W_{2h} \cdot \mathbf{n}_\alpha^\Omega, W_{3h} \cdot \mathbf{n}_\alpha^\Omega) \end{aligned} \quad (11)$$

The multi-head attention function is also defined for $(l) \mathcal{M}^L$. All weights $(l) W_0, (l) W_{11} \dots (l) W_{3h}$ are shared throughout all pillars and both graphs $(\mathcal{G}^L, \mathcal{G}^K)$ within a single layer l .

Final predictions are computed by the last layer within the Graph Neural Network and designed as single linear projection with shared weights across both graphs $(\mathcal{G}^L, \mathcal{G}^K)$ and pillars:

$$\begin{aligned} \mathbf{m}_i^K &= W_m \cdot (l_{\max}) \mathbf{n}_i^K \\ \mathbf{m}_j^L &= W_m \cdot (l_{\max}) \mathbf{n}_j^L \\ \mathbf{m}_j^L, \mathbf{m}_i^K &\in \mathbb{R}^{D'} \end{aligned} \quad (12)$$

3.3. Optimal Transport Layer

Following the approach by [40] the final matching is performed in two steps. First a score matrix $\mathbf{M} \in \mathbb{R}^{n \times m}$ is constructed by computing the unnormalized cosine similarity between each pair of features:

$$\begin{aligned} \mathbf{M} &= (\mathbf{m}^K)^T \cdot \mathbf{m}^L, \\ \mathbf{m}^K &= [\mathbf{m}_1^K, \dots, \mathbf{m}_n^K], \mathbf{m}^L = [\mathbf{m}_1^L, \dots, \mathbf{m}_m^L] \end{aligned} \quad (13)$$

In the second step a soft-assignment matrix $\mathbf{P} \in \mathbb{R}^{(n+1) \times (m+1)}$ is computed that contains matching probabilities for each pair of features. Each row and column of \mathbf{P} corresponds to a key-point in \mathcal{P}^K and \mathcal{P}^L respectively. The last column and the last row represent an auxiliary *dustbin* point to account for unmatched features. Accordingly \mathbf{M} is extended to a matrix $\bar{\mathbf{M}} \in \mathbb{R}^{(n+1) \times (m+1)}$ with all new elements initialized using a learnable parameter W_v . Finding the optimal assignment then corresponds to maximizing the sum $\sum_{i,j} \bar{\mathbf{M}}_{i,j} \mathbf{P}_{i,j}$ subject to the following constraints:

$$\begin{aligned} \sum_{i=1}^{n+1} \mathbf{P}_{i,j} &= \begin{cases} 1 & \text{for } 1 \leq j \leq m \\ n & \text{for } j = m+1 \end{cases} \\ \sum_{j=1}^{m+1} \mathbf{P}_{i,j} &= \begin{cases} 1 & \text{for } 1 \leq i \leq n \\ m & \text{for } i = n+1 \end{cases} \end{aligned} \quad (14)$$

This represents an optimal transport problem [46, 51, 8] which can be solved in a fast and differentiable way using a slightly modified version of the Sinkhorn algorithm [46, 8]. Let \mathbf{r}_i and \mathbf{c}_j denote the i^{th} row and j^{th} column of $\bar{\mathbf{M}}$ respectively. A single iteration of the algorithm consists of 2 steps:

1. $(t+1) \mathbf{r}_i \leftarrow (t) \mathbf{r}_i - \log \sum_j e^{\mathbf{r}_i \cdot \mathbf{c}_j - \alpha}$, with $\alpha = \log m$ for $i = n$ and $\alpha = 0$ otherwise
2. $(t+1) \mathbf{c}_j \leftarrow (t) \mathbf{c}_j - \log \sum_i e^{\mathbf{r}_i \cdot \mathbf{c}_j - \beta}$, with $\beta = \log n$ for $j = m$ and $\beta = 0$ otherwise

After $T = 100$ iterations we obtain $\mathbf{P} = \exp \left((T) \bar{\mathbf{M}} \right)$. The overall tensor graph is shown in Fig 2 including architectural details from the pillar layer to the optimal transport layer.

3.4. Loss

The overall architecture with its three layer types: *Pillar Layer*, *Graph Neural Network Layer* and *Optimal Transport Layer* is fully differentiable. Hence, the network is trained in a supervised manner. The ground truth being the set \mathcal{GT} including all index tuples (i, j) with pillar correspondences in our datasets accompanied by unmatched labels (\bar{n}, j) and (i, \bar{m}) , with (\bar{n}, \bar{m}) being redundant. We consider a negative log-likelihood loss $\mathcal{L}_{\text{NLL}} = - \sum_{i,j \in \mathcal{GT}} \log \mathbf{P}_{ij}$. By utilizing the ground truth matrix \mathcal{GT} characterizing fix inter-cloud point correspondences, the network is implicitly trained to predict close point matches and simultaneously disregarding distant ones, making it robust against occlusions e.g. caused by dynamic objects.

4. Experiments

We separated the experiments into two subsections. First we validate the quality of StickyPillars in a point cloud registration task where we compare our results to geometry

and DNN based state-of-the-art approaches. Subsequently we show how StickyPillars can be deployed as middle-end inside LiDAR odometry and mapping approaches by replacing the standard odometry estimation step to reduce drift and instabilities. We compare the performance on the KITTI Odometry benchmark to state-of-the-art methods like LOAM [58] and LO-Net [20]. Finally we demonstrate the robustness of StickyPillars simulating high speed scenarios by skipping certain amounts of frames on scenes from the KITTI odometry dataset.

Model configuration: For key-point extraction, we used variable c_{\min} and c_{\max} to achieve $n = m = 500$ key-points π_i as inputs for the *pillar layer*. Each *point pillar* is sampled with up to $z = 128$ points x_{Ω} using an Euclidean distance threshold of $d = 0.5$ m. Our implemented feature depth is $D' = 32$. The key-point encoder has five layers with the dimensions set to 32, 64, 128, 256 channels respectively. The graph is composed of $l_{\max} = 6$ *self* and *cross* attention layers with $h = 8$ heads each. Overall, this results in 33 linear layers. Our model is implemented in *PyTorch* [32] v1.6 with *Python* 3.7.

Training procedure: We process KITTI’s [14] odometry training sequences 00 to 10, using our key-point selection strategy (cp. Sec. 3.1) by computing the proposed smoothness function (Eq. 1). Ground truth correspondences and unmatched sets are generated using KITTI’s odometry ground truth. Ground truth correspondences are either key-point pairs with a nearest neighbor distance smaller than 0.1 m or invisible matches, i.e. all pairs with distances greater 0.5 m remain unmatched. We ignore all associations with a distance in range 0.1 m to 0.5 m ensuring variances in resulting features. We trained our model for 200 epochs using Adam [17] with a constant learning rate of 10^{-4} and a batch size of 16. For the point cloud registration experiment we chose sequence 00 – 05 for training and sequence 08 – 10 for evaluation as stated in [25] and [2] with frame differences between source and target frame of 1 – 10 (frame $\Delta = [1, 10]$). For the LiDAR odometry estimation we used sequence 00 – 06 for training and 07 – 10 for evaluation similar to [20] with frame $\Delta = [1, 5]$.

4.1. Point Cloud Registration

Validation metrics: For point cloud registration we adopted the experiments as stated in [25] where we sampled the test data sequences in 30 frame intervals and select all frames within a 5m radius as registration targets. We calculate the transformation error compared to the ground truth poses provided by the KITTI dataset for each frame pair based on Euclidean distance for translation and $\Theta = 2\sin^{-1}\left(\frac{\|R - \bar{R}\|_F}{\sqrt{8}}\right)$ for rotation, with $\|R - \bar{R}\|_F$ being the Frobenius norm of the chordal distance between estimation and ground truth rotation matrix. Finally we also present the average runtime for registering one frame pair

for each approach.

Comparison to state-of-the-art methods: We filter matches below a confidence threshold (e.g. 0.6) and subsequently apply singular value decomposition to determine the pose. Figure 3 shows predicted correspondences by the network on an unseen pair of frames for different temporal distances of the target frame. We validate the performance of StickyPillars comparing it to state-of-the-art geometric approaches like ICP [4], G-ICP [42], AA-ICP [33], NDT-P2D [47] and CPD [29] and the DNN based methods 3DFeat-Net [55], DeepVCP [25] and D3Feat [2] based on the correspondences predicted by the network. We adopted the results presented in [25] and extended the table by running the experiments on D3Feat and StickyPillars. For D3Feat we changed the number of key-points to 500 instead of 250 since [2] reported better performance using this configuration but leaving all other parameters at default. Furthermore according to [25] the first 500 frame of sequence 08 involve large errors in the ground truth poses and were therefore neglected during the experiments for point cloud registration. The final results for all methods considered are listed in Table 1. We are reaching comparable results to all state-of-the-art methods regarding the considered metrics. Moreover we achieve lowest mean angular and second lowest mean translational error for the deep learning based methods without the necessity of an initial pose estimate unlike DeepVCP. In consideration of a mean processing time of 15ms for feature extraction on CPU and an average inference time on a *Nvidia Geforce GTX 1080 Ti* of 101ms for correspondence finding, StickyPillars outperforms all methods regarding the total average runtime.

METHOD	ANGULAR ERR(°)		TRANSL. ERR(M)		T(S)
	MEAN	MAX	MEAN	MAX	
ICP-PO2PO [4]	0.139	1.176	0.089	2.017	8.17
ICP-PO2PL [4]	0.084	1.693	0.065	2.050	2.92
GICP [42]	0.067	0.375	0.065	2.045	6.92
AA-ICP [33]	0.145	1.406	0.088	2.020	5.24
NDT-P2D [47]	0.101	4.369	0.071	2.000	8.73
CPD [29]	0.461	5.076	0.804	7.301	3241
3DFEAT-NET [55]	0.199	2.428	0.116	4.972	15.02
DEEPCVP [25]	0.164	1.212	0.071	0.482	2.3
D3FEAT [2]	0.110	1.819	0.087	0.734	0.43
OURS	0.109	1.439	0.073	1.451	0.12

Table 1: Point cloud registration results: Our method shows comparable results to state-of-the-art methods and depicting a much lower average runtime.

4.2. LiDAR Odometry

Validation metrics: For validation on LiDAR Odometry, we are using the KITTI Odometry dataset provided with ground truth poses calculating the average translational RMSE t_{rel} (%) and rotational RMSE r_{rel} (°/100m)

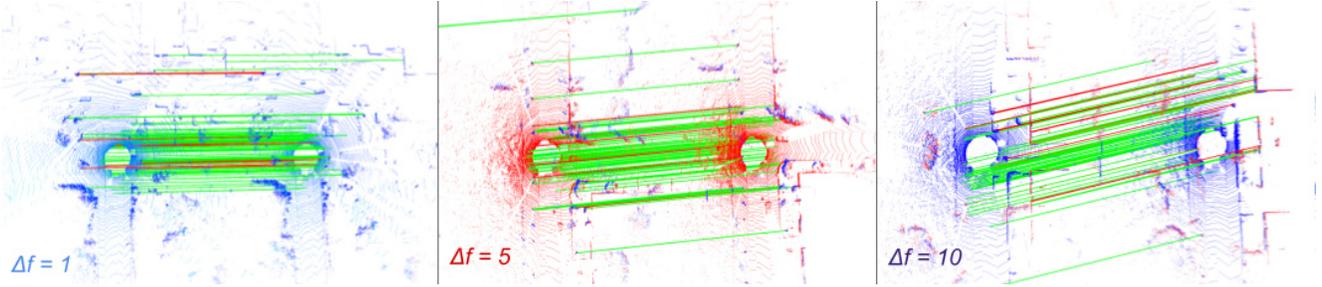


Figure 3: **Qualitative Results** from two point clouds with increasing frame Δ , i.e., increasing difficulty, of $\Delta = 1$ (blue - top row), $\Delta = 5$ (red - middle row), and $\Delta = 10$ (purple - bottom row) frames. The figure shows samples of the validation sets unseen during training. Green lines highlight correct matches, while red lines highlight incorrect ones.

on lengths of 100m-800m errors per scene according to [14].

Comparison to state-of-the-art methods: We evaluate the performance of StickyPillars in combination with a subsequent mapping step. For this purpose we utilize the A-LOAM¹ algorithm which is an advanced version of [58] and exchange the simple point cloud registration step prior to the mapping with StickyPillars. For our experiments we changed the voxel grid size of the surface features in the mapping step to 1.0m but leaving all other parameters at their default value. In order to achieve real-time capability for LiDAR Odometry, we infer StickyPillars on a *Nvidia Geforce RTX Titan* resulting in a mean runtime of 50ms per frame. For all following experiments A-LOAM was processed sequentially, neglecting all kinds of parallel implementations by ROS to ensure a reliable baseline for benchmark comparisons. We compare our results based on the KITTI odometry benchmark to different versions of the ICP algorithm [4] [42], CLS [53] and LOAM [58] which is widely considered as baseline in terms of point cloud based odometry estimation. Furthermore we validate against LO-Net [20] which is using a similar hybrid approach consisting of a Deep Learning method for point cloud registration and subsequent geometry based mapping. We adopted the values stated in [20] and extended the table with our results as shown in Table 2. We outperform the considered methods in the majority of sequences regarding t_{rel} and in almost every scene with respect to r_{rel} , leading to best results for average translational error on par with LO-Net and lowest average rotational error among all compared approaches.

In order to demonstrate the robustness of our method we also compared the standard A-LOAM implementation to our approach where we replaced the point cloud registration module with StickyPillars in the context of simulating higher speed scenarios and frame drops respectively. This is done by skipping a certain amount of frames of the particular sequence e.g. $\Delta = 3$ means providing every third consecutive frame to the algorithm. For evaluation we again use

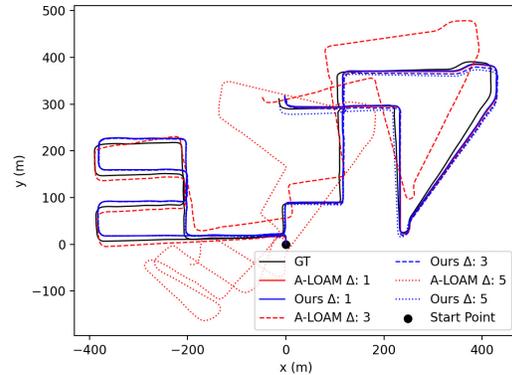


Figure 4: **Trajectory plots** for A-LOAM and A-LOAM + StickyPillars (Ours) for different frame Δ with ground truth.

the relative translational and rotational errors. For $\Delta = 1$, which equals to common processing of a sequence, standard A-LOAM provides some minor improvements to t_{rel} and also r_{rel} in selected scenes where average speed levels are lower, thus leading to large cloud overlap. In such cases the ordinary frame matching algorithm shows good performance. For other sequences like 02 with a more dynamic environment, the standard implementation fails and the robust transformations provided by StickyPillars help to correct the induced drift leading to much lower average transformation errors. The robustness of our approach in the context of varying vehicle velocities can also be observed by taking a look at the results for higher frame Δ where, with one exception, our approach outperforms standard A-LOAM in all considered scenes and also depicting comparable average errors to the ones of $\Delta = 1$. Furthermore we observed partially better results for larger frame Δ compared to the smaller ones on certain scenes (e.g. sequence 03) for our approach which probably is related to a reduction of drift effects caused by close frame to frame matchings. Figure 4 shows qualitative results of the estimated trajectories for the

¹<https://github.com/HKUST-Aerial-Robotics/A-LOAM>

SEQ.	ICP-PO2PO [4]		ICP-PO2PL [4]		GICP [42]		CLS [53]		LOAM [58] ¹		LO-NET+MAP [20]		OURS	
	t_{rel}	r_{rel}	t_{rel}	r_{rel}	t_{rel}	r_{rel}	t_{rel}	r_{rel}	t_{rel}	r_{rel}	t_{rel}	r_{rel}	t_{rel}	r_{rel}
00 [†]	6.88	2.99	3.80	1.73	1.29	0.64	2.11	0.95	1.10 (0.78)	0.53	0.78	0.42	0.65	0.26
01 [†]	11.21	2.58	13.53	2.58	4.39	0.91	4.22	1.05	2.79 (1.43)	0.55	1.42	0.40	1.82	0.45
02 [†]	8.21	3.39	9.00	2.74	2.53	0.77	2.29	0.86	1.54 (0.92)	0.55	1.01	0.45	1.00	0.34
03 [†]	11.07	5.05	2.72	1.63	1.68	1.08	1.63	1.09	1.13 (0.86)	0.65	0.73	0.59	0.91	0.45
04 [†]	6.64	4.02	2.96	2.58	3.76	1.07	1.59	0.71	1.45 (0.71)	0.50	0.56	0.54	0.53	0.17
05 [†]	3.97	1.93	2.29	1.08	1.02	0.54	1.98	0.92	0.75 (0.57)	0.38	0.62	0.35	0.46	0.23
06 [†]	1.95	1.59	1.77	1.00	0.92	0.46	0.92	0.46	0.72 (0.65)	0.39	0.55	0.33	0.56	0.25
07 [*]	5.17	3.35	1.55	1.42	0.64	0.45	1.04	0.73	0.69 (0.63)	0.50	0.56	0.45	0.43	0.24
08 [*]	10.04	4.93	4.42	2.14	1.58	0.75	2.14	1.05	1.18 (1.12)	0.44	1.08	0.43	1.02	0.29
09 [*]	6.93	2.89	3.95	1.71	1.97	0.77	1.95	0.92	1.20 (0.77)	0.48	0.77	0.38	0.67	0.24
10 [*]	8.91	4.74	6.13	2.60	1.31	0.62	3.46	1.28	1.51 (0.79)	0.57	0.92	0.41	1.00	0.41
MEAN	7.36	3.41	4.74	1.93	1.92	0.73	2.12	0.91	1.28 (0.84)	0.51	0.82	0.43	0.82	0.30

¹: The results on KITTI dataset outside the brackets are obtained by running the code, and those in the brackets are taken from [58].

[†]: KITTI Odometry dataset sequences used for training

^{*}: KITTI Odometry dataset sequences used for testing

Table 2: LiDAR Odometry results on the KITTI Odometry dataset. We get comparable results regarding t_{rel} and outperform state-of-the-art methods with respect to r_{rel} .

SEQ.	A-LOAM						A-LOAM+STICKYPILLARS					
	$\Delta = 1$		$\Delta = 3$		$\Delta = 5$		$\Delta = 1$		$\Delta = 3$		$\Delta = 5$	
	t_{rel}	r_{rel}	t_{rel}	r_{rel}	t_{rel}	r_{rel}	t_{rel}	r_{rel}	t_{rel}	r_{rel}	t_{rel}	r_{rel}
00 [†]	0.70	0.27	0.97	0.38	31.16	12.10	0.65	0.26	0.79	0.31	1.29	0.48
01 [†]	1.86	0.46	4.30	0.96	96.04	10.36	1.82	0.45	2.14	0.59	2.55	0.56
02 [†]	4.58	1.43	5.29	1.61	26.74	8.72	1.00	0.34	0.91	0.37	1.04	0.42
03 [†]	0.95	0.47	1.36	0.50	15.04	4.39	0.91	0.45	0.83	0.47	0.77	0.51
04 [†]	0.54	0.18	89.48	0.27	102.35	0.21	0.53	0.17	0.63	0.25	0.65	0.23
05 [†]	0.47	0.24	0.53	0.24	12.47	4.21	0.46	0.23	0.52	0.24	0.57	0.27
06 [†]	0.55	0.24	1.75	0.24	42.18	12.32	0.56	0.25	0.57	0.26	0.58	0.28
07 [*]	0.42	0.26	0.45	0.28	11.04	3.38	0.43	0.24	0.43	0.24	0.49	0.35
08 [*]	0.96	0.30	1.85	0.62	31.41	12.39	1.02	0.29	0.92	0.30	1.16	0.38
09 [*]	0.66	0.24	0.78	0.30	36.67	12.03	0.67	0.24	0.72	0.28	0.69	0.31
10 [*]	0.87	0.34	1.34	0.46	29.10	12.16	1.00	0.41	0.89	0.38	1.41	0.57
MEAN	1.14	0.40	9.83	0.53	39.47	8.39	0.82	0.30	0.85	0.34	1.02	0.40

Table 3: Extensive experiments demonstrating performance under simulated higher speed / frame drop scenarios with various frame Δ . Our approach shows very high robustness in terms of large environment changes compared to the standard point cloud registration used in A-LOAM.

two methods for different frame Δ on sequence 08 which was not seen during the training process of StickyPillars. For $\Delta > 1$ there are large odometry drifts for the standard implementation of A-LOAM whereas the trajectories for the extended version by StickyPillars are almost identical.

5. Conclusion

We present a novel model for point-cloud registration in real-time using deep learning. Thereby, we introduce a three stage model composed of a point cloud encoder, an attention-based graph and an optimal transport algorithm. Our model performs local and global feature matching at once using contextual aggregation. Evaluating our method on the KITTI

odometry dataset, we observe comparable results to other geometric and DNN based point cloud registration approaches but showing a significantly lower runtime. Furthermore we demonstrated our capability for robust odometry estimation by adding a subsequent mapping step on the KITTI odometry dataset where we outperformed the state-of-the-art methods regarding rotational error and showing comparable results on the translational error. Finally we proved the robustness of our approach in cases of higher speed scenarios and frame drops respectively, by providing the point clouds with various frame Δ . We showed that even by providing every fifth frame of a sequence StickyPillars is still able to predict accurate transformations thus stabilizing pose estimation when used inside LiDAR odometry and mapping approaches.

References

- [1] Yasuhiro Aoki, Hunter Goforth, Rangaprasad Arun Srivatsan, and Simon Lucey. Pointnetlk: Robust efficient point cloud registration using pointnet. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.
- [2] Xuyang Bai, Zixin Luo, Lei Zhou, Hongbo Fu, Long Quan, and Chiew-Lan Tai. D3feat: Joint learning of dense detection and description of 3d local features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6359–6367, 2020.
- [3] Jon Louis Bentley. Multidimensional binary search trees used for associative searching. *Communications of the ACM*, 18(9):509–517, sep 1975.
- [4] Paul J. Besl and Neil D. McKay. A method for registration of 3-d shapes. *IEEE Trans. Pattern Anal. Mach. Intell.*, 14(2):239–256, Feb. 1992.
- [5] JiaWang Bian, Wen-Yan Lin, Yasuyuki Matsushita, Sai-Kit Yeung, Tan-Dat Nguyen, and Ming-Ming Cheng. Gms: Grid-based motion statistics for fast, ultra-robust feature correspondence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4181–4190, 2017.
- [6] Jan Cech, Jiri Matas, and Michal Perdoch. Efficient sequential correspondence selection by cosegmentation. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1568–1581, 2010.
- [7] Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, and Tian Xia. Multi-view 3d object detection network for autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1907–1915, 2017.
- [8] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, pages 2292–2300, 2013.
- [9] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 224–236, 2018.
- [10] Ayush Dewan, Tim Caselitz, Gian Diego Tipaldi, and Wolfram Burgard. Rigid scene flow for 3d lidar scans. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1765–1770. IEEE, 2016.
- [11] Mihai Dusmanu, Ignacio Rocco, Tomas Pajdla, Marc Pollefeys, Josef Sivic, Akihiko Torii, and Torsten Sattler. D2-net: A trainable cnn for joint detection and description of local features. *arXiv preprint arXiv:1905.03561*, 2019.
- [12] Nico Engel, Stefan Hoermann, Markus Horn, Vasileios Belagiannis, and Klaus Dietmayer. Deeplocalization: Landmark-based self-localization with deep neural networks. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, pages 926–933. IEEE, 2019.
- [13] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.
- [14] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361. IEEE, 2012.
- [15] Zan Gojcic, Caifa Zhou, Jan D Wegner, Leonidas J Guibas, and Tolga Birdal. Learning multiview 3d point cloud registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1759–1769, 2020.
- [16] Xiuye Gu, Yijie Wang, Chongruo Wu, Yong Jae Lee, and Panqu Wang. Hplflownet: Hierarchical permutohedral lattice flownet for scene flow estimation on large-scale point clouds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3254–3263, 2019.
- [17] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [18] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12697–12705, 2019.
- [19] Jiaxin Li and Gim Hee Lee. Usip: Unsupervised stable interest point detection from 3d point clouds. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.
- [20] Qing Li, Shaoyang Chen, Cheng Wang, Xin Li, Chenglu Wen, Ming Cheng, and Jonathan Li. Lo-net: Deep real-time lidar odometry. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8473–8482, 2019.
- [21] Jiarong Lin and Fu Zhang. Loam.livox: A fast, robust, high-precision lidar odometry and mapping package for lidars of small fov. *arXiv preprint arXiv:1909.06700*, 2019.
- [22] Xingyu Liu, Charles R Qi, and Leonidas J Guibas. Flownet3d: Learning scene flow in 3d point clouds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 529–537, 2019.
- [23] Xingyu Liu, Mengyuan Yan, and Jeannette Bohg. Meteornet: Deep learning on dynamic 3d point cloud sequences. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9246–9255, 2019.
- [24] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.
- [25] Weixin Lu, Guowei Wan, Yao Zhou, Xiangyu Fu, Pengfei Yuan, and Shiyu Song. Deepvcv: An end-to-end deep neural network for point cloud registration. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.
- [26] A. Milioto, I. Vizzo, J. Behley, and C. Stachniss. RangeNet++: Fast and Accurate LiDAR Semantic Segmentation. In *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2019.
- [27] Peter J Mucha, Thomas Richardson, Kevin Macon, Mason A Porter, and Jukka-Pekka Onnela. Community structure in time-dependent, multiscale, and multiplex networks. *science*, 328(5980):876–878, 2010.

- [28] Marius Muja and David G Lowe. Fast approximate nearest neighbors with automatic algorithm configuration. *VISAPP (1)*, 2(331-340):2, 2009.
- [29] Andriy Myronenko and Xubo Song. Point set registration: Coherent point drift. *IEEE transactions on pattern analysis and machine intelligence*, 32(12):2262–2275, 2010.
- [30] Vincenzo Nicosia, Ginestra Bianconi, Vito Latora, and Marc Barthelemy. Growing multiplex networks. *Physical review letters*, 111(5):058701, 2013.
- [31] Yuki Ono, Eduard Trulls, Pascal Fua, and Kwang Moo Yi. Lf-net: learning local features from images. In *Advances in neural information processing systems*, pages 6234–6244, 2018.
- [32] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. *NIPS Workshops*, 2017.
- [33] Artem L Pavlov, Grigory WV Ovchinnikov, Dmitry Yu Derbyshev, Dzmitry Tsetserukou, and Ivan V Oseledets. Aa-icp: Iterative closest point with anderson acceleration. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1–6. IEEE, 2018.
- [34] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017.
- [35] 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*, pages 5099–5108, 2017.
- [36] Rahul Raguram, Jan-Michael Frahm, and Marc Pollefeys. A comparative analysis of ransac techniques leading to adaptive real-time random sample consensus. In *European Conference on Computer Vision*, pages 500–513. Springer, 2008.
- [37] Jerome Revaud, Philippe Weinzaepfel, César De Souza, Noe Pion, Gabriela Csurka, Yohann Cabon, and Martin Humenberger. R2d2: Repeatable and reliable detector and descriptor. *arXiv preprint arXiv:1906.06195*, 2019.
- [38] Szymon Rusinkiewicz and Marc Levoy. Efficient variants of the icp algorithm. In *Proceedings Third International Conference on 3-D Digital Imaging and Modeling*, pages 145–152. IEEE, 2001.
- [39] Radu Bogdan Rusu, Nico Blodow, and Michael Beetz. Fast point feature histograms (fpfh) for 3d registration. In *2009 IEEE international conference on robotics and automation*, pages 3212–3217. IEEE, 2009.
- [40] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. *arXiv preprint arXiv:1911.11763*, 2019.
- [41] Torsten Sattler, Bastian Leibe, and Leif Kobbelt. Scramsac: Improving ransac’s efficiency with a spatial consistency filter. In *2009 IEEE 12th International Conference on Computer Vision*, pages 2090–2097. IEEE, 2009.
- [42] Aleksandr Segal, Dirk Haehnel, and Sebastian Thrun. Generalized-icp. In *Robotics: science and systems*, volume 2, page 435. Seattle, WA, 2009.
- [43] Tixiao Shan and Brendan Englot. Lego-loam: Lightweight and ground-optimized lidar odometry and mapping on variable terrain. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4758–4765. IEEE, 2018.
- [44] Martin Simon, Karl Amende, Andrea Kraus, Jens Honer, Timo Samann, Hauke Kaulbersch, Stefan Milz, and Horst Michael Gross. Complexer-yolo: Real-time 3d object detection and tracking on semantic point clouds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.
- [45] Martin Simon, Stefan Milz, Karl Amende, and Horst-Michael Gross. Complex-yolo: An euler-region-proposal for real-time 3d object detection on point clouds. In *European Conference on Computer Vision*, pages 197–209. Springer, 2018.
- [46] Richard Sinkhorn and Paul Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343–348, 1967.
- [47] Todor Stoyanov, Martin Magnusson, Henrik Andreasson, and Achim J Lilienthal. Fast and accurate scan registration through minimization of the distance between compact 3d ndt representations. *The International Journal of Robotics Research*, 31(12):1377–1393, 2012.
- [48] Yanghai Tsin and Takeo Kanade. A correlation-based approach to robust point set registration. In Tomáš Pajdla and Jiří Matas, editors, *Computer Vision - ECCV 2004*, pages 558–569. Berlin, Heidelberg, 2004. Springer Berlin Heidelberg.
- [49] Tinne Tuytelaars and Luc J Van Gool. Wide baseline stereo matching based on local, affinity invariant regions. In *BMVC*, volume 412, 2000.
- [50] Arash K Ushani, Ryan W Wolcott, Jeffrey M Walls, and Ryan M Eustice. A learning approach for real-time temporal scene flow estimation from lidar data. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5666–5673. IEEE, 2017.
- [51] SS Vallender. Calculation of the wasserstein distance between probability distributions on the line. *Theory of Probability & Its Applications*, 18(4):784–786, 1974.
- [52] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.
- [53] Martin Velas, Michal Spáňal, and Adam Herout. Collar line segments for fast odometry estimation from velodyne point clouds. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4486–4495. IEEE, 2016.
- [54] Yue Wang and Justin 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.
- [55] Zi Jian Yew and Gim Hee Lee. 3dfeat-net: Weakly supervised local 3d features for point cloud registration. In *European Conference on Computer Vision*, pages 630–646. Springer, 2018.
- [56] Kwang Moo Yi, Eduard Trulls, Vincent Lepetit, and Pascal Fua. Lift: Learned invariant feature transform. In *European Conference on Computer Vision*, pages 467–483. Springer, 2016.

- [57] Ji Zhang and Sanjiv Singh. Loam: Lidar odometry and mapping in real-time. In *Proceedings of Robotics: Science and Systems Conference*, July 2014.
- [58] Ji Zhang and Sanjiv Singh. Low-drift and real-time lidar odometry and mapping. *Autonomous Robots*, 41(2):401–416, 2017.
- [59] Zhengyou Zhang. Iterative point matching for registration of free-form curves and surfaces. *International journal of computer vision*, 13(2):119–152, 1994.
- [60] Yin Zhou and Oncel Tuzel. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4490–4499, 2018.