2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

KDD-LOAM: Jointly Learned Keypoint Detector and Descriptors Assisted LiDAR Odometry and Mapping

Renlang Huang, Minglei Zhao, Jiming Chen, and Liang Li

***Abstract*— Sparse keypoint matching based on distinct 3D feature representations can improve the efficiency and robustness of point cloud registration. Existing learning-based 3D descriptors and keypoint detectors are either independent or loosely coupled, so they cannot fully adapt to each other. In this work, we propose a tightly coupled keypoint detector and descriptor (TCKDD) based on a multi-task fully convolutional network with a probabilistic detection loss. In particular, this self-supervised detection loss fully adapts the keypoint detector to any jointly learned descriptors and benefits the self-supervised learning of descriptors. Extensive experiments on both indoor and outdoor datasets show that our TCKDD achieves *state-of-the-art* performance in point cloud registration. Furthermore, we design a keypoint detector and descriptors-assisted LiDAR odometry and mapping framework (KDD-LOAM), whose real-time odometry relies on keypoint descriptor matching-based RANSAC. The sparse keypoints are further used for efficient scan-to-map registration and mapping. Experiments on KITTI dataset demonstrate that KDD-LOAM significantly surpasses LOAM and shows competitive performance in odometry.**

I. Introduction

Point cloud registration is crucial for many robotic and 3D vision applications, such as simultaneous localization and mapping (SLAM) [1] and 3D reconstruction [2]. Although the classic iterative closest point (ICP) algorithm [3] can precisely estimate the transformation, it requires an initial guess close to the ground truth and inefficient iterations to establish correct correspondences. In contrast, sparse feature matching directly establishes reliable correspondences between keypoints with similar descriptors, achieving efficient and robust point cloud registration.

Even though a few hand-crafted 3D keypoint detectors [4], [5] and descriptors [6]–[8] have been proposed over the years, the performance of 3D-3D point association remains unsatisfactory. In contrast, the learning-based descriptor is regarded as a promising approach [9], which maps the low-level geometric representations to a discriminative feature space. These descriptors are always learned in a self-supervised manner by maximizing the similarity between corresponding point features and minimizing the similarity between other point pairs. However, as it is difficult to define and label keypoints, these descriptors usually overlook keypoint detection and randomly sample points for description and matching, thus suffering from several drawbacks. First, inefficient oversampling is required to ensure a sufficient number of correspondences. Second, these poorly localized sampled points will result in inaccurate pose estimation.

This work is supported by the National Natural Science Foundation of China (62088101/62203383). The authors are with the College of Control Science and Engineering, Zhejiang University, Hangzhou, 310027, China.

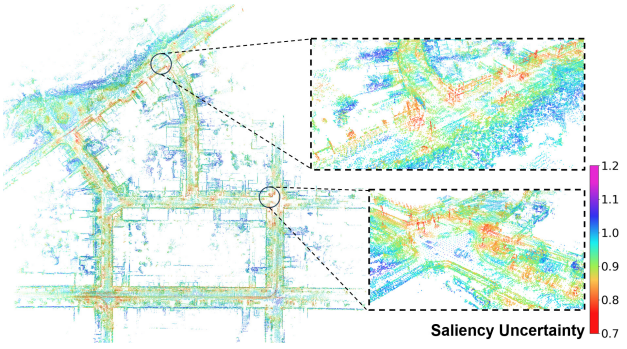

Fig. 1. The feature maps colored in saliency uncertainty built by our KDD-LOAM on KITTI sequence 07. Sharp corners and edges, distinguishable buildings, pillars, and vehicles are detected as salient regions (red), while flat surfaces, chaotic vegetation, and unstably scanned regions far from the sensor are detected as non-salient regions (blue). It is noteworthy that planar surfaces (most from roads) have been fitted as sparse surfels.

Third, non-salient points with indiscriminative descriptors degrade the inlier ratio. Similarly, existing keypoint detectors trained independently cannot fully adapt to descriptors.

To this end, we design a tightly coupled joint keypoint detector and descriptor, *i.e.*, TCKDD. Inspired by KPConv [10], we propose a fully convolutional neural network for keypoint detection and description, which utilizes a KPConv-based encoder-decoder backbone for 3D feature embedding. In addition, it densely predicts both a descriptor and the saliency uncertainty via two independent point-wise MLP heads for each point. The descriptors are learned through a quadruplet hardest contrastive loss in a self-supervised manner. To fully exert the potential of descriptors, we quantitatively define a matchability index fully based on descriptors and design a novel probabilistic detection loss based on maximum likelihood estimation. With this loss, the detection head can estimate the point-wise matchability robustly for keypoint selection and make the network concentrate on the learning of descriptors in salient regions.

As a deep front-end, TCKDD can densely predict global and local context-aware descriptors and detect keypoints with higher matchability. A series of experiments show that TCKDD achieves *state-of-the-art* performance on both the indoor RGB-D camera dataset 3DMatch [2] and the outdoor LiDAR dataset KITTI [11] for point cloud registration. With this powerful front-end, we utilize RANSAC as the middle-end to establish sparse correspondences based on the descriptors and minimize the point-to-point metric at the back-end. This pipeline can be operated in real-time for the registration between consecutive scans. Consequently, we can design a keypoint detector and descriptors assisted LiDAR odometry and mapping framework, *i.e.*, KDD-LOAM, with

this keypoint descriptor matching-based registration pipeline as the LiDAR odometry module. We propose to construct a voxel hash map consisting of only salient regions and fit the planar patches into sparse surfels, enabling a memory-efficient representation and fast nearest neighbor search. Then the keypoints from the current scan will be aligned to the map based on both point-to-point and point-to-plane metrics for more accurate localization. Experiments on the KITTI dataset demonstrate that the TCKDD-based odometry outperforms LOAM [12] which leverages the planar points and edge points for odometry by a large margin. Without loop closure detection and pose graph optimization, KDD-LOAM achieves competitive performance in odometry while maintaining real-time performance. Fig. 1 is a demonstration of our keypoint detection and mapping results. The main contributions of this work are summarized as follows:

- We design a probabilistic detection loss to tightly couple the learning of 3D keypoint detection and description.
- We propose a keypoint detector and descriptors assisted real-time LiDAR odometry and mapping framework, which achieves more accurate and robust performance in odometry in comparison to LOAM.
- A series of experiments on both indoor and outdoor datasets show that our TCKDD achieves *state-of-the-art* performance in 3D point cloud registration.

II. Related Work

Feature matching is a prominent approach in point cloud registration, efficiently establishing reliable sparse correspondences based on descriptors. Early methods use local hand-crafted descriptors based on either histograms [6], [7] or signatures [8]. Recent focus has shifted to learning-based 3D descriptors. For instance, 3DMatch [2] and PerfectMatch [13] employ 3D CNNs to learn local volumetric descriptors, converting patches into truncated distance function (TDF) or smoothed density value (SDV) representations. PPFNet [14] uses PointNet [15] for global context-aware patch descriptors, while FCGF [16] designs a sparse 3D convolutional encoder-decoder network. SpinNet [17] proposes a spatial point Transformer, converting point clouds as cylindrical volumes for transformation-invariant features. Recent methods like [18] learn point cloud interactions to enhance the inlier ratio, including coarse-to-fine registration approaches [19], [20] that achieve end-to-end correspondence learning.

Unlike the exploration of learning-based 3D descriptors, most 3D keypoint detectors are hand-crafted and target points with unique curvatures [4] or significant geometric variations in the principal direction [5] as keypoints. However, they struggle with real-world scans that are noisy, sparse, and non-uniform. Hence, researchers explore learning-based detectors for more reliable results. USIP [21] trains a feature proposal network via a probabilistic chamfer loss to predict 3D keypoints with high repeatability. To adapt the keypoint detector to the descriptors, some researchers propose to jointly learn the 3D keypoint detector and descriptors. 3DFeat-Net [22] designs a weakly supervised patch-wise network minimizing a saliency-weighted feature alignment triplet loss. However, it does not explicitly prioritize keypoint detection performance. D3Feat [9] densely predicts descriptors with a fully convolutional encoder-decoder and obtains saliency scores from descriptors using a self-supervised detection loss.

LiDAR odometry estimation involves real-time point cloud registration typically based on ICP or NDT. Nearly all modern SLAM systems are designed on top of odometry. Zhang *et al.* [12] propose LOAM that extracts and aligns planar and edge points to a sparse voxel grid-based feature map. LeGO-LOAM [23] segments the point cloud to remove unstable parts and adds ground constraints to improve accuracy. Additionally, F-LOAM [24] employs a faster non-iterative distortion compensation method to reduce the computational cost. However, these methods rely on hand-crafted feature extraction, which is only suitable for small pose derivations and requires multiple iterations for reliable correspondences.

III. Methodology

In this work, we design a **t**ightly **c**oupled **k**eypoint **d**etector and **d**escriptor for 3D point cloud registration, *i.e.*, TCKDD. The principles of *tight coupling* are three-fold: 1) the detector and the descriptor share a common feature extractor; 2) the detector fully matches the matchability of the descriptors; 3) the detector and the descriptor can enhance each other via a probabilistic detection loss. We further integrate TCKDD into a real-time LiDAR odometry and mapping system.

A. Network Architecture

We treat the joint learning of 3D keypoint detection and description as a multi-task learning paradigm and follow its fundamental neural network style, *i.e.*, plugging several separated task-specific prediction heads into a shared feature extraction backbone. Inspired by KPConv [10], we propose a fully convolutional network for 3D keypoint detection and description. KPConv directly operates on irregular point sets by interpolating point features to uniformly distributed kernel points for regular convolution, *i.e.*, linear mapping, and summation of kernel responses.

Utilizing the normalized KPConv, TCKDD can construct a fully convolutional network that directly consumes point sets, as depicted in Fig. 2. The backbone can extract multi-scale 3D features at different encoder layers consisting of a stack of residual bottleneck blocks. The locality of KPConv enables strided convolution for downsampling. The decoder can recover the resolution via nearest upsampling and aggregate the multi-scale 3D features via skip connections and 1×1 convolution (unary blocks). Different from the original KPFCNN [10], we replace the batch normalization with group normalization [25], which is robust *w.r.t.* batch size and group-wise features. Finally, TCKDD densely predicts both point-wise descriptors and saliency uncertainty based on the shared 3D features from the backbone via two independent all-MLP heads, *i.e.*, 1×1 convolutional blocks.

B. Quadruplet Contrastive Loss for Descriptor Learning

Metric learning is widely used to train descriptors. Essentially it maps low-level geometric representations to a high-dimensional feature space, where descriptors of correctly

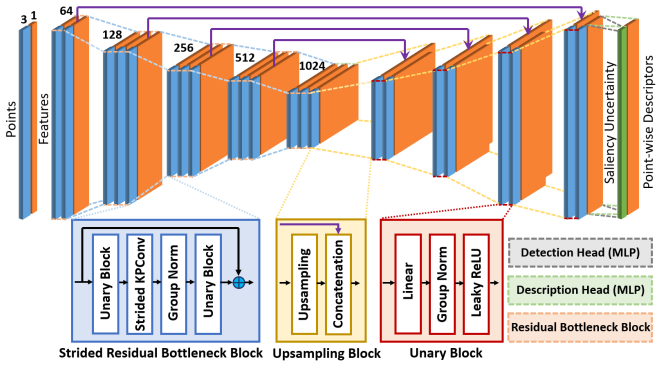

Fig. 2. The network architecture of TCKDD for jointly learning of 3D keypoint detection and description.

associated point pairs are close, while those of other pairs differ by at least a margin. The correspondence set $\mathbf{C}$ of two partially overlapped point clouds P, Q is a set of the mutually nearest points (p_i, q_j) satisfying $\|p_i - q_j\|_2 \leq R_p$. The negative point set $\mathbf{N}_i$ of a point $p_i \in P$ is a set of points $q_k \in Q$ satisfying $\|p_i - q_k\|_2 \geq R_n (R_n \geq R_p)$. Denote d_i, d_j as the descriptors of $p_i \in P, q_j \in Q$, then the self-supervised descriptor loss can be designed as a hardest quadruplet contrastive loss according to metric learning strategies:

$$\mathcal{L}_{desc} = \frac{1}{|\mathbf{C}|} \sum_{(i,j)\in\mathbf{C}} \left\{ \lambda_p \left[D(d_i, d_j) - m_p \right]^+ + \left[m_n - \min_{k\in\mathbf{N}_i} D(d_i, d_k) \right]^+ + \left[m_n - \min_{k\in\mathbf{N}_j} D(d_k, d_j) \right]^+ \right\}, \tag{1}$$

where $D(\cdot,\cdot)$ is the Euclidean distance, $\lambda_p = 2$, and m_p, m_n are positive and negative margins, respectively. This loss can maximize the similarity between descriptors of true correspondences and minimize the maximum similarity otherwise.

C. Self-supervised Probabilistic Detection Loss

The principle of joint learning of 3D keypoint detection and description is to fully adapt the detector and the descriptors to each other. Therefore, we treat the detection head of TCKDD as a saliency scoring network based on the matchability of descriptors. According to (1), we can directly design a metric named *matchability index* to characterize the matchability of a given descriptor d_i quantitatively:

$$m_i = \left[D(d_i, d_j) - m_p \right]_+ + \left[m_n - \min_{k\in\mathbf{N}_i} D(d_i, d_k) \right]^+, \tag{2}$$

where d_j is the descriptor of the correctly associated point of p_i in point cloud Q. This matchability index is actually the hardest triplet loss of a single descriptor in metric learning, which describes the distinctness of a descriptor in the feature space. Particularly, we use the hardest negative pairs for negative mining so that the matchability index directly models the decision margin of descriptor matching and optimizes the decision boundary during metric learning. A lower matchability index m_i indicates greater matchability of the descriptor d_i, *i.e.*, the 3D point p_i is more salient. Therefore, all we need is to learn a matchability index estimator as a keypoint detector. We propose to learn a probabilistic model rather than a regressive model for matchability index estimation since the descriptor is predicted from a specific point cloud sampled from the surfaces of a dense 3D scene in a specific perspective, and so as the descriptors of the associated points. Ideally, metric learning would construct a fully discriminative feature space where the matchability index of each descriptor is zero. Hence, we choose an exponential distribution to model the matchability index m_i with a parameter σ_i, or namely *saliency uncertainty*:

$$p(m_i|\sigma_i) = \frac{1}{\sigma_i} \exp\left(-\frac{m_i}{\sigma_i} \right). \tag{3}$$

It is notable that the detection head of our TCKDD directly predicts point-wise saliency uncertainty as outputs. Hence, we can design a probabilistic detection loss based on maximum likelihood estimation (MLE) to robustly fit this exponential distribution model:

$$\begin{aligned} \mathcal{L}_{det} &= -\frac{1}{|\mathbf{C}|} \sum_{(i,j)\in\mathbf{C}} \left(\ln p(m_i|\sigma_i) + \ln p(m_j|\sigma_j) \right) \\ &= \frac{1}{|\mathbf{C}|} \sum_{(i,j)\in\mathbf{C}} \left(\ln \sigma_i + \frac{m_i}{\sigma_i} + \ln \sigma_j + \frac{m_j}{\sigma_j} \right). \end{aligned} \tag{4}$$

Theoretically, the first derivative of the log-likelihood indicates that the global optimality conditions for detector learning are $\sigma_i = m_i$, making the probabilistic detection loss effective for training the detection head as a robust matchability estimator. Remarkably, this loss is also a weighted form of the hardest contrastive loss, with the descriptor losses of keypoints having higher weights than those of non-salient points. When the detector meets global optimality, this detection loss turns out to be a logarithmic contrastive loss, prioritizing the descriptors of keypoints. This approach enhances the matchability of keypoints, mitigating the negative effects of mining geometric features in non-salient areas like planar surfaces and disorganized regions. Hence, this probabilistic detection loss can not only train a keypoint detector as a robust matchability estimator that fully accommodates the jointly learned descriptors but also promote the learning of keypoint descriptors via weighted metric learning.

D. Keypoint Detector and Descriptors Assisted LOAM

As illustrated in Fig. 3, we integrate TCKDD into a LiDAR odometry and mapping system, *i.e.*, KDD-LOAM, consisting of the following key components.

Scan deskewing and scan-to-scan registration. Similar to [26], we first utilize the most generally applicable constant velocity model for scan deskewing, which requires no extra sensors involving time synchronization. As a powerful front-end of point cloud registration, TCKDD can densely predict global and local context-aware descriptors and detect keypoints by sorting the saliency uncertainty. We leverage RANSAC to establish sparse correspondences based on the descriptors and minimize the sum of point-to-point distances between inlier correspondences based on SVD. This pipeline achieves real-time scan-to-scan registration to provide a solid initial guess for incremental ego-motion estimation.

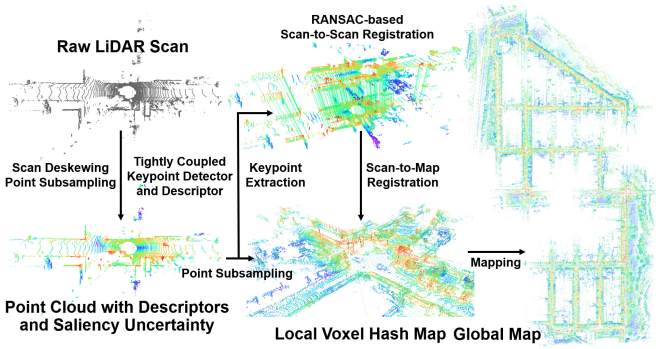

Fig. 3. The system overview of KDD-LOAM. We leverage the constant velocity model for scan deskewing and predict the point-wise descriptors and saliency uncertainty through TCKDD. Based on a reliable relative pose guess from RANSAC-based scan-to-scan registration, KDD-LOAM achieves accurate odometry by aligning the deskewed and subsampled scan with a high-resolution yet memory-efficient local map.

Keypoint subsampling and mapping. With a reliable relative pose guess from scan-to-scan registration, we refine the ego-pose estimate by aligning the deskewed and subsampled scan with the accumulated local map. Unlike methods such as [12] that create sparse feature maps of edge points or surface points with noisy directions or normals, we propose a voxel grid map capturing detailed geometry akin to surface reconstruction. We keep it simple, approximating complex 3D surfaces in any topology with ample scan points. Utilizing a voxel hash map with voxel size $v \times v \times v$ that stores up to N_{max} points per voxel, we achieve efficient insertion, indexing, deletion, and nearest neighbor search compared to 3D arrays [12] or KD-trees [24]. Practically, we set $v = 1$m, $N_{max} = 20$. Instead of high-resolution panoptic mapping [26], KDD-LOAM accumulates a voxel hash map for only salient regions with low saliency uncertainty. For better surface representation, we fit voxels with N_{max} points to planes via least square regression. Voxels meeting error criteria of regression are replaced with surfels represented by the point closest to their center, their normal vector, and radius v. Our approach, in contrast to [26], adapts voxel grids for adaptive 3D reconstruction of only salient regions, combining dense points and sparse surfels for a memory-efficient representation applicable to global mapping.

Inspired by CT-ICP [27], we adopt a two-stage voxel grid-based subsampling for sequential map update and ego-pose estimation. In the first stage, we use voxel size αv ($\alpha \in (0, 1]$) to downsample by reserving an original scan point per voxel to prevent discretization errors. After scan-to-map registration, these subsampled points are transformed using the global pose estimate and added to the voxel hash map. In the second stage, we propose a saliency-aware voxel grid subsampling using voxel size βv ($\beta \in [1, 2]$) to select keypoints for faster scan-to-map registration. Non-salient points are discarded, while salient regions accommodate more points per voxel for accurate point cloud registration.

Robust scan-to-map registration. We use scan-to-map registration for more accurate odometry as it proves more reliable and robust than scan-to-scan registration [1], [12]. Our scan-to-map registration builds on the classic ICP algorithm, which typically establishes correspondences between two point clouds via nearest neighbor search. With a reliable scan-to-scan relative pose guess grounded in geometrically consistent correspondences, it is more likely to avoid sub-optimal convergence and reduce ICP iterations. However, ICP requires a hand-crafted maximum distance threshold for outlier rejection, which depends on the expected initial error.

To this end, we treat scan-to-map registration as compensation for scan-to-scan pose estimation, determining the maximum distance threshold by assessing deviations from the scan-to-scan pose estimate over time. Denote this deviation as $\Delta T \in SE(3)$, the upper bound of the point deviation is

$$\delta(\Delta T) = \|\Delta t\|_2 + 2r \sin\left(\frac{1}{2}\arccos\frac{tr(\Delta R) - 1}{2}\right), \quad (5)$$

where r is the maximum range of LiDAR scans, $\Delta R \in SO(3)$ and $\Delta t \in \mathbb{R}^3$ refer to the rotation and translation components of ΔT, respectively. Inspired by KISS-ICP [26], we adopt a Gaussian distribution over $\delta(\Delta T)$ and compute its standard deviation σ_t to robustly set the maximum distance threshold of ICP as the three-sigma bound $\tau_t = 3\sigma_t$.

Given a point from the current scan during data association, we first search the nearest point and the nearest surfel separately in the voxel hash map based on point-to-point distances. Next, we evaluate the point-to-point distance from the nearest point and the point-to-plane distance from the nearest surfel to determine its correspondence. Finally, the maximum distance threshold τ_t determines whether to accept this correspondence as an inlier. This process establishes a set of point-to-point and point-to-plane correspondences for each ICP iteration. A robust ego-pose (*i.e.*, rotation $R \in SO(3)$ and translation $t \in \mathbb{R}^3$) is estimated by minimizing the sum of point-to-point residuals and point-to-plane residuals:

$$\min_{R,t} \sum_{p,q \in \mathcal{C}} \rho(e) = \frac{e(Rp + t, q)^2/2}{\sigma_t/3 + e(Rp + t, q)^2}, \quad (6)$$

where ρ is the Geman-McClure kernel with a strong outlier rejection property. The optimal pose can be estimated via the Gauss-Newton method. The Jacobian can be derived by applying the left perturbation model with $\delta\xi = [\delta\rho^T \ \delta\phi^T]^T \in \mathbb{R}^6$, $\delta\xi^\wedge \in \mathfrak{se}(3)$. For a point-to-point residual $e = \|Rp + t - q\|_2^2$, the Jacobian of $\mathbf{e} = Rp + t - q$ *w.r.t.* $\delta\xi$ is

$$\begin{aligned} J_p &= \frac{\partial(Rp + t - q)}{\partial\delta\xi} = \lim_{\delta\xi \to 0} \frac{exp(\delta\phi^\wedge)Rp + \delta\rho - Rp}{\delta\xi} \\ &= \begin{bmatrix} I_{3\times3} & -(Rp + t)^\wedge \end{bmatrix}_{3\times6}. \end{aligned} \quad (7)$$

For a point-to-plane residual $e = |n^T(Rp + t - q)|^2$, the Jacobian of $\mathbf{e} = nn^T(Rp + t - q)$ *w.r.t.* $\delta\xi$ is derived as

$$J_s = \frac{\partial nn^T(Rp + t - q)}{\partial\delta\xi} = \begin{bmatrix} nn^T & -nn^T(Rp + t)^\wedge \end{bmatrix}. \quad (8)$$

According to the chain rule of derivative, making the derivative of the objective function *w.r.t.* the disturbance $\delta\xi$ equal to 0 will lead to the following equation:

$$\sum_i \frac{1}{(\sigma_t/3 + e_i^2)^2} J_i^T J_i \delta\xi = -\sum_i \frac{1}{(\sigma_t/3 + e_i^2)^2} J_i^T \mathbf{e}_i. \quad (9)$$

TABLE I

EVALUATION RESULTS ON INDOOR DATASETS 3DMATCH.

Sampled Points	5000	2500	1000	500	250	5000	2500	1000	500	250
	Feature Matching Recall (%)					Registration Recall (%)				
PerfectMatch [13]	95.0	94.3	92.9	90.1	82.9	78.4	76.2	71.4	67.6	50.8
FCGF [16]	97.4	97.3	97.0	96.7	96.6	85.1	84.7	83.3	81.6	71.4
D3Feat [9]	95.6	95.4	94.5	94.1	93.1	81.6	84.5	83.4	82.4	77.9
SpinNet [17]	97.6	97.2	96.8	95.5	94.3	88.6	86.6	85.5	83.5	70.2
Predator [18]	96.6	96.6	96.5	96.3	96.5	89.0	89.9	90.6	88.5	86.6
YOHO [28]	98.2	97.6	97.5	97.7	96.0	90.8	90.3	89.1	88.6	84.5
CoFiNet [19]	98.1	**98.3**	**98.1**	**98.2**	**98.3**	89.3	88.9	88.4	87.4	87.0
GeoTrans. [20]	97.9	97.9	97.9	97.9	97.6	92.0	91.8	91.8	91.4	**91.2**
ours (rand)	**98.3**	98.0	97.9	97.7	96.8	91.9	91.6	91.3	88.4	81.5
ours (prob)	98.1	98.0	97.9	97.7	97.1	**93.0**	92.4	92.0	91.7	87.7
ours (NMS)	98.1	97.8	98.0	97.8	97.6	92.1	91.5	91.3	91.3	89.7
ours (NMS-prob)	98.1	98.0	97.8	97.8	97.3	92.4	**93.1**	**92.7**	**92.4**	89.8

The registration is performed by repeating data association and solving (9) until convergence.

IV. EXPERIMENTS

In this section, we will evaluate our TCKDD regarding point cloud registration on both indoor (3DMatch [2]) and outdoor scenes (KITTI [11]). In addition, KDD-LOAM will be evaluated on the KITTI odometry benchmark against existing LiDAR-based odometry and SLAM systems. Our source code is released at [code release].

A. Indoor Scenes: 3DMatch Benchmark

3DMatch is a widely used 3D reconstruction benchmark including 62 indoor scenes collected by RGB-D cameras. We use the training data preprocessed by [18] and evaluate our TCKDD against both hand-crafted and learning-based descriptors on the official test set including scan pairs with $> 30\%$ overlap using two metrics: feature matching recall (FMR) [14] and registration recall (RR) [2].

TCKDD is evaluated with different numbers of keypoints in Table I, compared with *state-of-the-art* learning-based descriptors PerfectMatch [13], FCGF [16], D3Feat [9], SpinNet [17], Predator [18], YOHO [28] and coarse-to-fine registration methods CoFiNet [19], GeoTransformer [20]. For TCKDD, we compare random sampling (rand) with three saliency-based keypoint selection strategies: probabilistic sampling (prob), non-maximum suppression (NMS), and probabilistic NMS (NMS-prob). In terms of FMR, TCKDD without keypoints consistently outperforms all 3D descriptors and performs on par with GeoTransformer. When sampled points are fewer than 1000, TCKDD with keypoints achieves higher FMR, showing more stable performance against existing descriptors. In terms of RR, TCKDD with probabilistic keypoints outperforms all the descriptors and CoFiNet consistently by 2$\sim$39%. With over 250 sampled points, our probabilistic keypoints even surpass GeoTransformer notably. Furthermore, the effectiveness and robustness of our keypoints are validated through consistent RR improvement, especially with fewer than 1000 sampled points.

We demonstrate the robustness of TCKDD (prob) by varying the inlier distance threshold τ_1 and the inlier ratio threshold τ_2 in FMR. As shown in Fig. 4, we report the performance of hand-crafted descriptors SpinImages [29], SHOT [8], FPFH [7] and ealier learning-based descriptors CGF [30], 3DMatch [2], PPFNet [14], PPF-FoldNet [31]. TCKDD consistently outperforms other methods with $\tau_1 \geq$ 5cm and significantly surpasses them across all inlier ratio thresholds. Under a stricter condition $\tau_2 = 0.2$, TCKDD maintains a high FMR of 89.3%, while SpinNet, D3Feat, and FCGF drop to 85.7%, 75.8%, and 67.4%, respectively, which highlights that TCKDD is more robust to maintain higher inlier ratio in challenging scenarios.

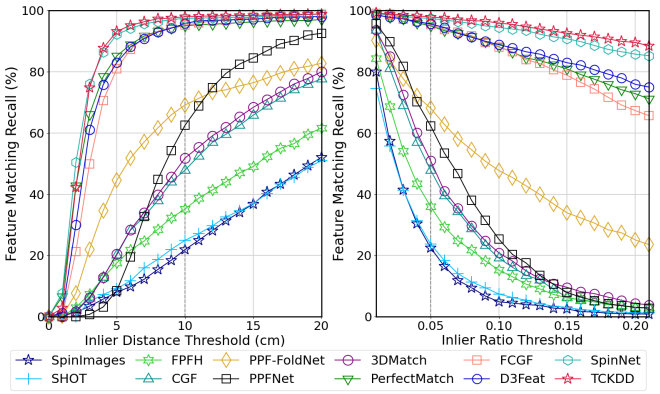

Fig. 4. Feature matching recalls on the 3DMatch dataset in relation to inlier distance threshold τ_1 (Left) and inlier ratio threshold τ_2 (Right).

TABLE II

REGISTRATION PERFORMANCE ON KITTI ODOMETRY DATASET.

Model	RTE (cm)	RRE (°)	RR (%)
3DFeat-Net [22]	25.9	0.25	96.0
FCGF [16]	9.5	0.30	96.6
D3Feat [9]	7.2	0.30	99.8
SpinNet [17]	9.9	0.47	99.1
Predator [18]	**6.8**	0.27	99.8
CoFiNet [19]	8.2	0.41	99.8
GeoTransformer (RANSAC)	7.4	0.27	99.8
GeoTransformer (LGR) [20]	**6.8**	**0.24**	99.8
TCKDD (ours, prob)	**6.8**	0.27	**100.0**

B. Outdoor Scenes: KITTI Benchmark

For the KITTI [11] dataset, we use sequences 0 to 5 for training, 6 to 7 for validation, and 8 to 10 for testing. We refine the GPS localization results via ICP [3] as ground truth. Additionally, only point cloud pairs at least 10m away from each other are selected. Following [18], we use three metrics for evaluation: *relative translation error* (RTE), *relative rotation error* (RRE) and *registration recall* (RR).

In Table II, TCKDD is compared with the *state-of-the-art* descriptors 3DFeat-Net [22], FCGF [16], D3Feat [9], SpinNet [17], Predator [18] and coarse-to-fine registration methods CoFiNet [19], GeoTransformer [20]. TCKDD achieves *state-of-the-art* RTE and the highest registration recall of 100%, which demonstrates the effectiveness and robustness. Within TCKDD, we compare random sampling (rand) with two saliency-based keypoint selection strategies: sorting (sort) and probabilistic sampling (prob). As shown in Table III, TCKDD maintains 100% of RR even with only 1000 randomly sampled points, indicating the robustness of its descriptors in establishing sparse yet reliable correspon-

TABLE III

ABLATION STUDIES OF KEYPOINT DETECTION ON KITTI DATASET.

Points	5000	2500	1000
	RTE (cm) / RRE (°) / RR (%)		
ours (rand)	7.1 / 0.26 / 100.0	8.4 / 0.30 / 100.0	14.6 / 0.50 / 100.0
ours (sort)	6.9 / 0.30 / 100.0	7.6 / 0.39 / 100.0	8.9 / 0.55 / 100.0
ours (prob)	6.8 / 0.27 / 100.0	7.6 / 0.33 / 100.0	10.3 / 0.46 / 100.0

TABLE IV

ABLATION STUDIES OF LIDAR ODOMETRY ON KITTI BENCHMARK.

Seq	A-LOAM		TCKDD + A-LOAM	
	scan-to-scan	scan-to-map	scan-to-scan	scan-to-map
00	4.13 / 1.72	0.81 / 0.31	**2.28 / 0.98**	**0.67 / 0.26**
01	3.46 / 0.95	**2.01** / 0.52	**2.89 / 0.80**	2.19 / **0.48**
02	7.47 / 2.55	4.66 / 1.46	**2.03 / 0.91**	**0.99 / 0.35**
03	4.35 / 2.08	0.92 / 0.47	**2.03 / 1.40**	**0.90 / 0.44**
04	1.65 / 0.81	0.72 / 0.36	**0.67 / 0.49**	**0.62 / 0.29**
05	4.06 / 1.66	0.51 / 0.24	**1.96 / 1.00**	**0.45 / 0.22**
06	**1.11 / 0.51**	**0.59 / 0.27**	2.01 / 1.27	**0.59 / 0.27**
07	2.84 / 1.80	0.44 / 0.24	**1.80 / 1.36**	**0.43 / 0.22**
09	5.75 / 1.88	0.70 / 0.30	**3.05 / 1.28**	**0.62 / 0.23**
10	3.60 / 1.76	0.98 / 0.38	**3.26 / 1.38**	**0.84 / 0.34**

dences. The sorted keypoints notably reduce RTE compared to random sampling, especially with fewer sampled points for registration, thus underscoring TCKDD's ability to detect keypoints with high matchability and repeatability.

C. Evaluation of LiDAR Odometry and Mapping Systems

In this subsection, we design experiments to demonstrate that 1) TCKDD effectively improves odometry accuracy against its baseline A-LOAM [12]; 2) KDD-LOAM significantly reduces cumulative error in scan-to-scan registration and outperforms classic LiDAR odometry or SLAM systems; 3) KDD-LOAM achieves more memory-efficient mapping while maintaining performance comparable to KISS-ICP. All the algorithms are evaluated with the mean relative pose error (RPE) over trajectories of 100 to 800m (relative translation error in % / relative rotational error in °/100m) [11]. Sequence 08 is excluded from evaluation due to significant errors in its ground-truth localization results.

We first demonstrate the effectiveness of TCKDD by replacing the scan-to-scan registration step of A-LOAM with TCKDD-based RANSAC. As shown in Table IV, TCKDD-based scan-to-scan registration significantly outperforms A-LOAM based on hand-crafted keypoints, providing a much more reliable and accurate pose guess for the subsequent scan-to-map registration. Furthermore, we integrate TCKDD-based scan-to-scan registration with the subsequent mapping step of A-LOAM. Apart from sequence 01 collected from a featureless environment, TCKDD-aided A-LOAM outperforms A-LOAM by a large margin, especially in sequences 00, 02, and 10 with 0.14%, 3.67% and 0.14% improvements of average RTEs, respectively. These results indicate that TCKDD-based odometry outperforms hand-crafted features and effectively complements existing mapping approaches.

Finally, we evaluate KDD-LOAM against *state-of-the-art* LiDAR odometry methods on the KITTI dataset [11], including LOAM [12], F-LOAM [24], SuMa [1], SuMa++ [32], and KISS-ICP [26]. As shown in Table V, KDD-LOAM consistently outperforms classic odometry systems, LOAM, F-LOAM, and SLAM systems SuMa, SuMa++ across nearly all the scenes. This underscores the effectiveness and robustness of KDD-LOAM. Compared with KISS-ICP using a similar voxel hash map and an ICP-based scan-to-map registration step with adaptive thresholds, KDD-LOAM achieves lower RREs in all sequences while performing on par with KISS-ICP in RTEs, showcasing its potential to achieve lower global localization drifts. The average RPEs are reported as Avg, while Avg† stands for the results without the featureless sequence 01. Except for sequence 01, KDD-LOAM even achieves a lower average RPE than KISS-ICP with a more memory-efficient map representation. We compare the average memory usage for local maps of some long sequences in Table VI, which indicates that KDD-LOAM consumes 16.6% less memory for local mapping than KISS-ICP.

TABLE V

RELATIVE POSE ERRORS OF LIDAR ODOMETRY ON KITTI DATASET.

Seq	LOAM	F-LOAM	SuMa	SuMa++	KISS-ICP	KDD-LOAM
00	0.78 / -	0.92 / 0.43	0.77 / 0.32	0.65 / 0.22	**0.52** / 0.19	**0.52 / 0.18**
01	1.43 / -	2.80 / 0.60	11.15 / 0.76	1.63 / 0.47	**0.65 / 0.14**	0.76 / **0.14**
02	0.92 / -	1.56 / 0.52	2.93 / 0.93	3.54 / 0.14	0.53 / 0.15	**0.51 / 0.14**
03	0.86 / -	1.09 / 0.66	1.25 / 0.61	0.67 / 0.47	**0.66 / 0.16**	0.67 / **0.16**
04	0.71 / -	1.43 / 0.52	0.86 / 0.27	**0.34** / 0.27	0.35 / 0.13	0.37 / **0.07**
05	0.57 / -	0.79 / 0.36	0.56 / 0.32	0.40 / 0.19	0.32 / 0.14	**0.26 / 0.12**
06	0.65 / -	0.72 / 0.39	0.64 / 0.51	0.47 / 0.27	**0.26 / 0.08**	**0.26 / 0.08**
07	0.63 / -	0.54 / 0.39	0.47 / 0.37	0.39 / 0.28	0.32 / 0.16	**0.31 / 0.15**
09	0.77 / -	1.28 / 0.55	0.79 / 0.41	0.58 / 0.20	**0.48** / 0.13	0.50 / **0.12**
10	0.79 / -	1.77 / 0.58	0.99 / 0.44	0.67 / 0.30	0.60 / 0.20	**0.53 / 0.17**
Avg	0.81 / -	1.29 / 0.50	2.04 / 0.49	0.93 / 0.28	**0.47** / 0.15	**0.47 / 0.13**
Avg†	0.74 / -	1.12 / 0.49	1.03 / 0.46	0.86 / 0.26	0.45 / 0.15	**0.44 / 0.13**

TABLE VI

MEMORY USAGE FOR LOCAL MAPPING OF DIFFERENT METHODS (KB).

Sequence	00	02	05	09	10	Avg
KISS-ICP	4685.6	3936.6	5038.9	4785.9	3836.8	4456.8
KDD-LOAM	4065.5	3091.2	4434.4	3881.8	3101.7	3714.9

V. CONCLUSIONS

This paper presents a tightly coupled 3D keypoint detector and descriptor for point cloud registration along with a keypoint detector and descriptor assisted LiDAR odometry and mapping system. We exploit the multi-task learning paradigm with a carefully designed probabilistic detection loss to learn a fully convolutional 3D descriptor and a keypoint detector fully adapted to it. Our odometry and mapping system achieves robust registration and memory-efficient mapping based on dense keypoints and sparse surfels. The evaluation results indicate that our keypoint detector and descriptor are robust to different range-sensing technologies and achieve *state-of-the-art* registration recall. The experiments on the KITTI benchmark demonstrate that our real-time and memory-efficient KDD-LOAM performs on par with *state-of-the-art* LiDAR odometry systems. In future work, we intend to investigate keypoint descriptor-based loop closure detection and pose graph optimization to extend our KDD-LOAM to a full SLAM system.

REFERENCES

[1] J. Behley and C. Stachniss, "Efficient surfel-based slam using 3d laser range data in urban environments.," in *Robotics: Science and Systems*, vol. 2018, p. 59, 2018.

[2] A. Zeng, S. Song, M. Nießner, M. Fisher, J. Xiao, and T. Funkhouser, "3dmatch: Learning local geometric descriptors from rgb-d reconstructions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1802–1811, 2017.

[3] P. Besl and N. D. McKay, "A method for registration of 3-d shapes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, no. 2, pp. 239–256, 1992.

[4] H. Chen and B. Bhanu, "3d free-form object recognition in range images using local surface patches," *Pattern Recognition Letters*, vol. 28, no. 10, pp. 1252–1262, 2007.

[5] Y. Zhong, "Intrinsic shape signatures: A shape descriptor for 3d object recognition," in *IEEE International conference on Computer Vision Workshops, ICCV workshops*, pp. 689–696, IEEE, 2009.

[6] R. B. Rusu, N. Blodow, Z. C. Marton, and M. Beetz, "Aligning point cloud views using persistent feature histograms," in *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 3384–3391, IEEE, 2008.

[7] R. B. Rusu, N. Blodow, and M. Beetz, "Fast point feature histograms (fpfh) for 3d registration," in *2009 IEEE International Conference on Robotics and Automation*, pp. 3212–3217, IEEE, 2009.

[8] S. Salti, F. Tombari, and L. Di Stefano, "Shot: Unique signatures of histograms for surface and texture description," *Computer Vision and Image Understanding*, vol. 125, pp. 251–264, 2014.

[9] X. Bai, Z. Luo, L. Zhou, H. Fu, L. Quan, and C.-L. 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*, pp. 6359–6367, 2020.

[10] H. Thomas, C. R. Qi, J.-E. Deschaud, B. Marcotegui, F. Goulette, and L. J. Guibas, "Kpconv: Flexible and deformable convolution for point clouds," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6411–6420, 2019.

[11] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3354–3361, IEEE, 2012.

[12] J. Zhang and S. Singh, "Loam: Lidar odometry and mapping in real-time.," in *Robotics: Science and systems*, 2014.

[13] Z. Gojcic, C. Zhou, J. D. Wegner, and A. Wieser, "The perfect match: 3d point cloud matching with smoothed densities," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5545–5554, 2019.

[14] H. Deng, T. Birdal, and S. Ilic, "Ppfnet: Global context aware local features for robust 3d point matching," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 195–205, 2018.

[15] C. R. Qi, H. Su, K. Mo, and L. 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*, pp. 652–660, 2017.

[16] C. Choy, J. Park, and V. Koltun, "Fully convolutional geometric features," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8958–8966, 2019.

[17] S. Ao, Q. Hu, B. Yang, A. Markham, and Y. Guo, "Spinnet: Learning a general surface descriptor for 3d point cloud registration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11753–11762, 2021.

[18] S. Huang, Z. Gojcic, M. Usvyatsov, A. Wieser, and K. Schindler, "Predator: Registration of 3d point clouds with low overlap," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4267–4276, 2021.

[19] H. Yu, F. Li, M. Saleh, B. Busam, and S. Ilic, "Cofinet: Reliable coarse-to-fine correspondences for robust point cloud registration," *Advances in Neural Information Processing Systems*, vol. 34, pp. 23872–23884, 2021.

[20] Z. Qin, H. Yu, C. Wang, Y. Guo, Y. Peng, S. Ilic, D. Hu, and K. Xu, "Geotransformer: Fast and robust point cloud registration with geometric transformer," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[21] J. Li and G. H. Lee, "Usip: Unsupervised stable interest point detection from 3d point clouds," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 361–370, 2019.

[22] Z. J. Yew and G. H. Lee, "3dfeat-net: Weakly supervised local 3d features for point cloud registration," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 607–623, 2018.

[23] T. Shan and B. Englot, "Lego-loam: Lightweight and ground-optimized lidar odometry and mapping on variable terrain," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4758–4765, IEEE, 2018.

[24] H. Wang, C. Wang, C.-L. Chen, and L. Xie, "F-loam: Fast lidar odometry and mapping," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4390–4396, IEEE, 2021.

[25] Y. Wu and K. He, "Group normalization," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 3–19, 2018.

[26] I. Vizzo, T. Guadagnino, B. Mersch, L. Wiesmann, J. Behley, and C. Stachniss, "Kiss-icp: In defense of point-to-point icp–simple, accurate, and robust registration if done the right way," *IEEE Robotics and Automation Letters*, vol. 8, no. 2, pp. 1029–1036, 2023.

[27] P. Dellenbach, J.-E. Deschaud, B. Jacquet, and F. Goulette, "Ct-icp: Real-time elastic lidar odometry with loop closure," in *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5580–5586, IEEE, 2022.

[28] H. Wang, Y. Liu, Z. Dong, and W. Wang, "You only hypothesize once: Point cloud registration with rotation-equivariant descriptors," in *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 1630–1641, 2022.

[29] A. E. Johnson and M. Hebert, "Using spin images for efficient object recognition in cluttered 3d scenes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 5, pp. 433–449, 1999.

[30] M. Khoury, Q.-Y. Zhou, and V. Koltun, "Learning compact geometric features," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 153–161, 2017.

[31] H. Deng, T. Birdal, and S. Ilic, "Ppf-foldnet: Unsupervised learning of rotation invariant 3d local descriptors," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 602–618, 2018.

[32] X. Chen, A. Milioto, E. Palazzolo, P. Giguere, J. Behley, and C. Stachniss, "Suma++: Efficient lidar-based semantic slam," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4530–4537, IEEE, 2019.