VXP: Voxel-Cross-Pixel Large-Scale Camera-LiDAR Place Recognition

Yun-Jin Li[1,2⋆] Mariia Gladkova[1,2⋆] Yan Xia[1,2†] Rui Wang[3]
Daniel Cremers[1,2]
[1] TU Munich [2] Munich Center for Machine Learning [3] Microsoft
{yunjin.li, mariia.gladkova, yan.xia, cremers}.@tum.de wangr@microsoft.com

Abstract

Cross-modal place recognition methods are flexible GPS-alternatives under varying environment conditions and sensor setups. However, this task is non-trivial since extracting consistent and robust global descriptors from different modalities is challenging. To tackle this issue, we propose Voxel-Cross-Pixel (VXP), a novel camera-to-LiDAR place recognition framework that enforces local similarities in a self-supervised manner and effectively brings global context from images and LiDAR scans into a shared feature space. Specifically, VXP is trained in three stages: first, we deploy a visual transformer to compactly represent input images. Secondly, we establish local correspondences between image-based and point cloud-based feature spaces using our novel geometric alignment module. We then aggregate local similarities into an expressive shared latent space. Extensive experiments on the three benchmarks (Oxford RobotCar, ViViD++ and KITTI) demonstrate that our method surpasses the state-of-the-art cross-modal retrieval by a large margin. Our evaluations show that the proposed method is accurate, efficient and light-weight. Our project page is available at: https://yunjinli.github.io/projects-vxp/.

1. Introduction

Since the emergence of autonomous systems, global place recognition has become essential for mobile robotics. Despite the widespread availability of the Global Navigation Satellite System (GNSS), signal outages remain inevitable, particularly in parking spaces or urban areas where buildings or tunnels can block satellite signals [38]. These disruptions are critical challenges for achieving autonomous driving on a city-wide scale and must be managed using onboard devices like cameras [4], LiDARs [41], or radars [36]. The Autonomous Vehicle (AV) sensor suite provides various strategies for data recording and, thus, enables alternative ways for global localization in GNSS-denied areas. Although numerous solutions have been proposed within the computer vision and robotics communities, most still rely on the same type of data during both map acquisition and operation. This dependence on a single data source may limit the applicability of these solutions in cases of sensor malfunctions or variations in sensor setups. Consequently, there is a need for more flexible localization methods that can take advantage of different sensor modalities under varying environmental conditions. This presents significant potential for cross-modal place recognition techniques. While multi-modal approaches require data to be available from all sensors, cross-modal methods are intended to be more flexible and seamlessly switch between the map and query sources. For instance, camera-to-LiDAR method would support querying a database of encoded LiDAR scans with RGB images (2D-3D localization). In terms of practical value, it would save the on-board computational load of processing large point clouds and guarantee global localization even in cases of LiDAR malfunctioning using image data.

Although cross-modal place recognition offers significant potential, it also presents challenges due to substantial differences between observations from various sensors. Specifically, in camera-to-LiDAR localization, images and point clouds exhibit a clear gap in both raw data (2D images vs 3D scans) and extracted features. The lack of explicit correlation between these two data modalities complicates the development of cross-modal global localization solutions. Due to this, only a few approaches have been proposed to tackle the task so far. Cattaneo et al. [9] first introduce 2D and 3D feature extraction networks to create a shared embedding space between images and point clouds. LC^2 [21] proposes to transform image and point clouds into the same 2.5D space for reducing the domain gap. LIP-Loc [33] advocates usage of multi-class N-pair batched loss in the contrastive learning regime to boost cross-modal retrieval. While these methods focus on designing powerful networks to encode data into robust global descriptors, they ignore geometric relation between local structures captured

†Corresponding author. * Equal contribution.

by both modalities. Local consistency not only provides additional constraints in order to effectively bridge the domain gap during the cross-modal training, but also enhances the representative power of the shared latent space.

In light of this, we introduce a novel method *Voxel-Cross-Pixel (VXP)* for camera-LiDAR place recognition. Our pipeline is three-fold: firstly, we leverage the power of visual transformers to obtain an expressive feature map and compact global embedding for an input image. Secondly, we choose sparse voxelized representation of a corresponding LiDAR scan and hierarchically aggregate features by utilizing sparse 3D convolutions. By means of projective geometry we establish local feature correspondences between image- and voxel-based feature maps and enforce their similarity during training. Lastly, we enforce similarities between global descriptors of the cross-modal matches. This comprehensive training paradigm enables the network to effectively capture both fine-grained local details and broader global context, facilitating successful cross-modal learning. We evaluate our model on three real-world datasets, achieving state-of-the-art cross-modal retrieval.

To summarize, the main contributions of the paper are:

- We propose a novel framework for the cross-modal place recognition, *Voxel-Cross-Pixel (VXP)*, which effectively encodes images and LiDAR scans in a shared latent space.
- We demonstrate the effectiveness of local similarity constraints in learning robust global descriptors for the cross-modal place recognition task.
- We establish state-of-the-art performance in cross-modal retrieval on the Oxford RobotCar, ViViD++ datasets and KITTI benchmark, while maintaining high uni-modal global localization accuracy.
- We publicly release our code along with implemented baselines at: https://github.com/yunjinli/vxp.

2. Related Work

In this section, we first review uni-modal place recognition techniques. We then introduce some fusion-based approaches. Finally, the existing cross-modal methods are presented.

Visual and point cloud-based retrieval. Uni-modal place recognition methods operate within one sensor type and aim to find the closest query match in a database. Most widely researched modalities are visual and LiDAR-based, while other types such as radar recently have received attention from the community [5]. Traditional image-based approaches, such as bag-of-words [13], represent different places with a visual vocabulary of quantized local descriptors [28] and they are widely used in the SLAM community for re-localization and loop closure tasks [7, 15]. In recent years, Convolutional Neural Network (CNN)-based methods have gained popularity for their expressiveness and enhanced robustness. Arandjelović et al. introduced NetVLAD [4], a CNN-based approach that encodes RGB images into dense feature maps and learns to effectively aggregate these features into a global descriptor. CosPlace [6] explored to perform the retrieval as a classification task. Recent works [1, 2] proposed to process the features extracted by a CNN with a Conv-AP layer or a Feature-Mixer. AnyLoc [17] utilizes the features generated from off-the-shelf self-supervised model (DINOv2 [26]) to achieve SOTA performance in many VPR benchmarks.

As for LiDAR-based place recognition, Uy et al. proposed PointNetVLAD [34], in which they employed PointNet [29] to extract features from a point cloud map and then aggregate them into a global descriptor using a subsequent NetVLAD layer. LPD-Net was introduced by Liu et al. [24], in which an adaptive local feature extraction module is proposed to extract local features along with the graph-based aggregation module to effectively combine them. SOE-Net [39] first introduces orientation encoding into PointNet and a self-attention unit to generate a robust 3D global descriptor. Furthermore, various methods [11, 45] explored the integration of different transformer networks to learn long-range contextual relationships. In contrast, Minkloc3D [18] employed a voxel-based strategy to generate a compact global descriptor. However, the voxelization methods inevitably suffer from information loss due to the quantization. Recent CASSPR [40] thus introduced a hierarchical cross attention transformer, combining both the advantages of voxel-based strategies with the point-based strategies. Text2Loc [42] achieved the 3D localization based on textual descriptions. In this paper, our work brings the best practices of 2D image and 3D point cloud communities together into a coherent framework that can achieve state-of-the-art performance in cross-modal retrieval.

Fused-Modal Place Recognition. LiDAR-based methods are more robust to variations in illumination and appearance when compared to the vision-based approaches. However, obtained scans are limited in capturing fine details of the observed scenes, while image data offers rich and dense scene capture. To this end, researchers have started exploring the possibility of fusing image and LiDAR data for the place recognition task. Pan et al. proposed a method called CORAL [27], in which point cloud data is converted into an elevation image in order to perform further fusion. MinkLoc++ [19], on the other hand, employed a late fusion technique, processing point cloud and image data separately and performing fusion at the final stage. While our approach relies on having both image and LiDAR data available during training, due to the chosen architecture with two independent branches we are capable of dealing with a single stream data during inference, which enables cross-modal retrieval.

Cross-Modal Place Recognition. Cattaneo et al. [8]

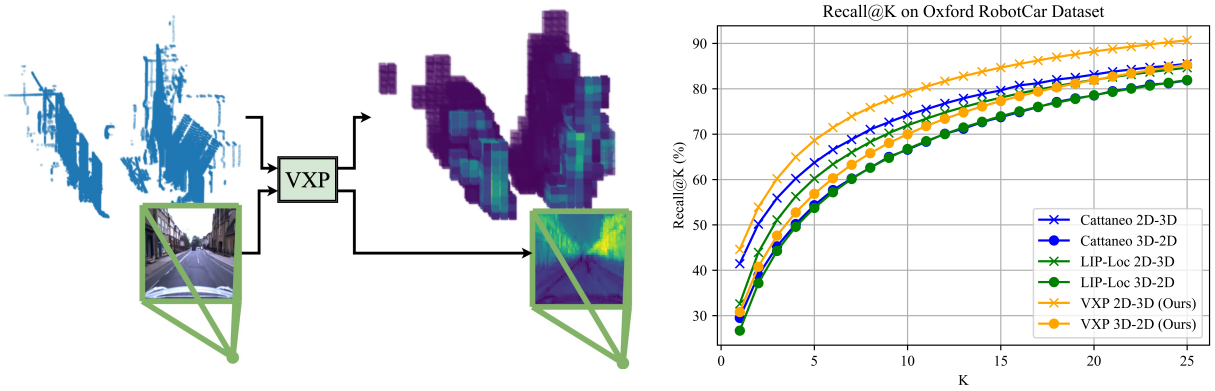

Figure 1. (**Left**) Voxel-Cross-Pixel (VXP) can effectively map data from different modalities (2D images and 3D LiDAR scans) into the shared latent space, which exhibits local similarities and captures global context. (**Right**) Recall for up-to K = 25 retrieved places on Oxford RobotCar benchmark. VXP consistently demonstrates superior cross-modal large-scale global retrieval preformance.

were the first to introduce this task, proposing a data-driven method where two networks were trained to encode images and point cloud maps separately using a teacher-student training approach. Initially, the image network (teacher) was trained using the triplet loss function [32], and then the point cloud network (student) was trained to align point embeddings within the shared latent space. In our work, we build on this paradigm with a stronger image backbone and enhance global descriptors by incorporating local feature constraints. The LC^2 approach, proposed by Lee et al. [21], presented an alternative method for cross-modal retrieval, where the domain gap was bridged by pre-processing sensor data and transforming it into the same data representations. Specifically, they converted both types of data into the 2.5D space, where RGB images were turned into disparity maps using depth network [37] and LiDAR point clouds were transformed into range images. A self-supervised pre-training scheme [22] was employed on the encoders, enabling the networks convergence. Similar method such as Lip-Loc [33] also proposed to process LiDAR-scans into range images and optimize their encoders by contrastive learning. In comparison, our method directly handles input raw data and does not require computationally demanding pre-processing steps such as generation of range images or depth maps, which would be more favorable for on-board devices.

A few studies have been proposed to tackle cross-modal registration such as 2D-3D re-localization [12, 23, 31, 35]. These methods primarily concentrate on accurately aligning a given camera view with a corresponding point cloud map and estimating relative 6-DoF transformation between them. In our work, we propose a solution for finding the cross-modal pairs, which are often unavailable in a real-world scenario, and advocate usefulness of local constraints in achieving this goal.

3. Problem Statement

We begin by defining the task of cross-modal place recognition. In particular, we are interested in camera-to-LiDAR retrieval, however the definition can be naturally extended to other modalities such as radars.

Given a reference map M_{ref}, where each element (a 2D image I or a 3D point cloud P) is tagged with a GPS coordinate, we aim to retrieve the geographically closest match to a query Q from a different sensor modality, such as LiDAR scanner or camera respectively. With this, the cross-modal place recognition can be defined formally for 3D-2D as

$$I^* = \text{argmin}\{d(g(Q), f(I))\}$$

or for 2D-3D as

$$P^* = \text{argmin}\{d(f(Q), g(P))\},$$

where $d(\cdot)$ is a distance metric (e.g. L1 norm), f is an image network, g is a point cloud network and $I, P \in M_{\text{ref}}$. This step can be efficiently done using a KD-tree (e.g. from FAISS library [16]).

4. Method

In this section, we introduce our cross-modal place recognition approach in detail. We design two separate networks that map image and point cloud into the shared latent space. Practically, dealing with raw point cloud data, which typically consists of thousands of points, can pose a significant computational challenge. To tackle this problem, we downsample each input scan before feeding it to a network. To this end, we leverage point cloud grouping techniques, which has also been shown to effectively capture local structures [30]. Consequently, we deploy voxelization method [44] to transform the raw point cloud data $\mathbf{P} \in$

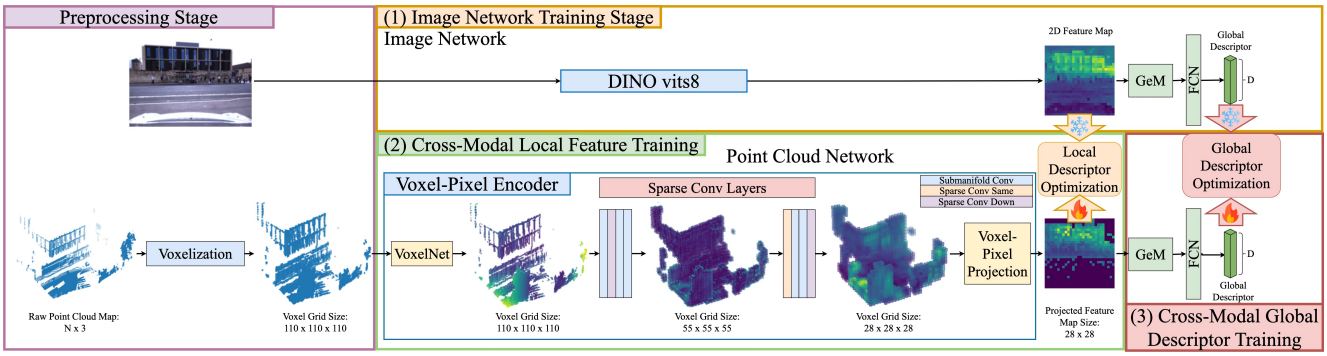

Figure 2. VXP pipeline comprises three steps: (1) image network training (Sec. 4.1), (2) cross-modal local feature training (Sec. 4.2), and cross-modal global descriptor training (Sec. 4.3). Starting from step (2) image features are frozen (❄), while the point cloud features are trained (🔥). The two networks operate independently during inference, so queries and database samples can be processed separately. The objective is to map different data into a shared latent space and minimize the distance (e.g. L2 norm) between global descriptors of different modalities taken from the same space.

$\mathbb{R}^{N\times 3}$ into a voxel grid $\mathbf{V} = \{\mathbf{v}_i \in \mathbb{R}^{M\times 3}, \mathbf{c}_i \in \mathbb{R}^3\}_{1,2,...,T}$, where T is the number of non-empty voxels and M represents the maximal number of points within a voxel. If the number of points in a voxel is lower than M, we do zero-padding. From this point, the framework employs a voxel-based representation of LiDAR scans.

Our Voxel-Cross-Pixel (VXP) pipeline comprises three steps as demonstrated in Fig. 2. Firstly, we train an image network to learn distinctive global descriptors based on positive and negative image pairs (Sec. 4.1). The learned feature space guides optimization in the second stage, where we enforce local correspondences by deploying the Voxel-Pixel Projection module in the point cloud branch (Sec. 4.2). Lastly, we optimize for the similarity between global descriptors to ensure consistency (Sec. 4.3).

4.1. Image Network

The image network architecture comprises two components: (1) the DINO ViTs-8 encoder and (2) a global pooling layer (GeM + FCN) as illustrated in Fig. 2. In the initial phase, an RGB image, denoted as $I \in R^{H\times W\times 3}$, is processed by the DINO ViTs-8 encoder $f^{enc} : R^{H\times W\times 3} \rightarrow R^{H^*\times W^*\times D}$, where $H^* = H//8$ and $W^* = W//8$. This operation yields 2D features, which are also recognized as local image feature descriptors. Subsequently, these generated image features are passed through the global pooling layer $f^{pool} : R^{H^*\times W^*\times D} \rightarrow R^D$, resulting in the creation of a global image descriptor.

We train the image network in a contrastive learning regime using a triplet loss function as per Eq. (1), where an anchor image denoted as I_i^a, a positive image I_i^p closely related to the anchor image's location, and a negative image I_i^n positioned far away from the anchor image.

$$\mathcal{L}_{\text{img}} = \sum_{I_i^{a,p,n}\in\mathcal{B}} [d(f(I_i^a), f(I_i^p)) - d(f(I_i^a), f(I_i^n)) + m]_+ \tag{1}$$

Note that $d(\cdot)$ is the distance function, $f(\cdot)$ is the image branch model, m is the margin, and $[\cdot]_+$ means $\max\{\ 0, [\cdot]\ \}$. In order to train more efficiently, we find the hardest positive sample with maximal distance and the hardest negative sample with minimal distance to the anchor within the mini-batch $\mathcal{B}$.

4.2. Cross-modal Local Feature Training

In this section we describe the second stage of our pipeline, where we pre-train point cloud-based branch using local feature correspondences. The overview can be seen in Fig. 2.

Voxel Feature Encoding. The initial voxel feature $\mathbf{v} \in \mathbb{R}^{M\times 3}$ aggregates information from M raw point coordinates contained within the voxel boundaries. We use VoxelNet [44] to extract more detailed descriptor for each voxel $\mathbf{v} \in \mathbb{R}^{M\times 3} \rightarrow \mathbb{R}^{D^*}$. Finally, we perform a series of sparse 3D convolutions [43] to generate a sparse 3D feature map of grid size $28 \times 28 \times 28$, namely $\mathbf{V}_{out}$, as formulated in Eq. (2).

$$\mathbf{V}_{\text{out}} = \{\mathbf{v}_i^{\text{out}} \in \mathbb{R}^D, \mathbf{c}_i^{\text{out}} \in \mathbb{R}^3\}_{1,2,...,T^*} \tag{2}$$

The $\mathbf{v}_i^{out} \in \mathbb{R}^D$ represents a local descriptor of a single voxel in the output voxel grid, which is a D-dimensional vector corresponding to the channel size of the 2D feature from f^{enc}, while $\mathbf{c}_i^{out}$ denotes the coordinate of this voxel within the voxel grid. Here, $\mathbf{c}_i^{out}$ is defined with respect to the voxel grid coordinate frame $\{\mathcal{V}\}$. Note that T^* represents the number of non-empty voxel local descriptors. Sparse convolutions allow us to aggregate spatial in-

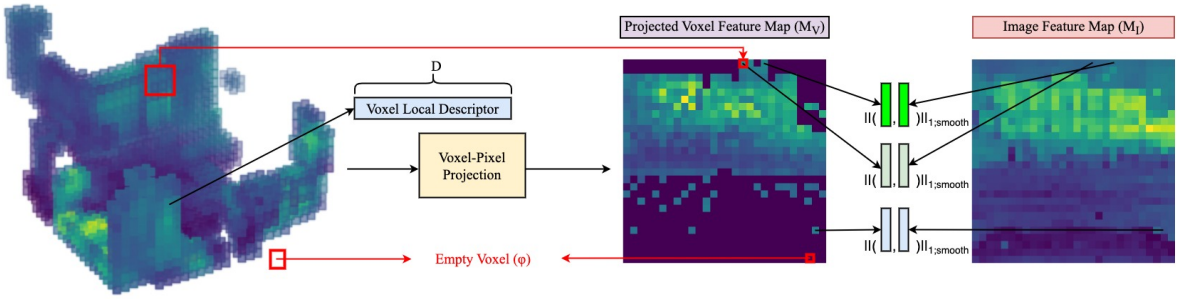

Figure 3. Illustration of our proposed local feature optimization between projected voxel- and image-based feature maps. ϕ represents "empty" as the 3D feature maps are sparse. Note that the voxel local descriptor is the $\mathbf{v}_i^{out}$ introduced in Eq. (2). After the projection, multiple $\mathbf{v}_i^{out}$ could be projected as per Eq. (4).

formation from neighboring voxels in a hierarchical fashion, which allows to capture long-distance relations.

Voxel-Pixel Projection. In order to bridge the domain gap between point cloud and image, we introduce simple yet effective *Voxel-Pixel Projection* module. This module projects voxels onto the image plane using the pinhole camera model. However, it's important to note that the voxel coordinates are defined within the voxel grid coordinate system denoted as $\{\mathcal{V}\}$. As per Eq. (3), we first transform the voxels into the point cloud (LiDAR) coordinate frame and apply projection matrix $\mathbf{M}$ to transform points onto the image plane. This way, we can obtain the voxel-based feature map and establish local descriptor constraints with the image-based features. Projection matrix is assumed to be provided and comprises intrinsic camera parameters and extrinsic LiDAR-camera calibration transformation.

$$\lambda \begin{bmatrix} u_i \\ v_i \\ 1 \end{bmatrix} = \mathbf{M} \cdot \left(\begin{bmatrix} v_x & 0 & 0 \\ 0 & v_y & 0 \\ 0 & 0 & v_z \end{bmatrix} \mathbf{c}_i^{out} + \begin{bmatrix} v_x/2 + x_{min} \\ v_y/2 + y_{min} \\ v_z/2 + z_{min} \end{bmatrix} \right) \tag{3}$$

Note that (v_x, v_y, v_z) is the dimension of the output voxel grid and the lower bound of the point cloud range is represented as $(x_{\text{min}}, y_{\text{min}}, z_{\text{min}})$.

Local Feature Optimization. During the local descriptor optimization phase shown in Fig. 3, we utilize the projected voxel coordinates (u_i, v_i) as indices to retrieve the corresponding local descriptors from the image feature map. Once retrieved, we can apply the local descriptor loss as

$$\mathcal{L}_{\text{local}} = \sum_{(u_i,v_i)\in\mathcal{M}_V} ||d_i \cdot \mathcal{M}_V(u_i, v_i) - \mathcal{M}_I(u_i, v_i)||_{1;\,\text{smooth}}. \tag{4}$$

The projected voxel feature map is denoted as $\mathcal{M}_V$, the image feature map is $\mathcal{M}_I$. We also take care of collisions, when multiple voxels are projected to the same pixel, by weighting descriptors with their voxels' inverse depths d_i. This way we give preference to the voxels that are closer to the camera, however propagate gradients to all voxels. This strategy allowed more stable training over the z-buffering.

4.3. Cross-Modal Global Descriptor Training

In the last stage, we fine-tune the Voxel-Pixel Encoder and train pooling layers together with the subsequent FCN to bring global embeddings closer to their image-based matches with

$$\mathcal{L}_{\text{global}} = \sum_i ||f(I_i) - g(P_i)||_{1;\,\text{smooth}}, \tag{5}$$

where I_i and P_i are an image and a point cloud corresponding to the same location, $f(.)$ and $g(.)$ refer to the corresponding networks. This allows us to ensure global consistency of aggregated descriptors in addition to local similarities enforced in the previous stage.

5. Experiments and Results

5.1. Implementation Details

Image Network Training: We resize the image to 224×224. During training of the image network, positive pairs are chosen from images that are within 10 meters, while the negative pairs are defined from samples that are more than 25 meters away as [21]. We set the margin of the triplet loss function to 0.3. To handle zero-triplets, i.e. anchor-positive-negative tuples with zero triplet loss, we employ a strategy of gradually increasing the batch size if the proportion of zero-triplets exceeds 30% of the original batch size. The training with branch expansion rate is adopted from [18] and configured to 1.4, while the maximum batch size is set to 256. We use pre-trained model (dino-vits8) and finetune all its parameters together with our GeM + FCN block using Eq. (1). A custom batch sampler with at least one positive pair within each batch $\{\mathcal{B}\}$ is implemented. For each sample in $\{\mathcal{B}\}$, its hard positive / negative sample is the farthest / closest sample in $\{\mathcal{B}\}$ based on the L2 distance between the global descriptors.

Point Cloud Network Training. We take the fine-tuned

image network and freeze all its parameter during the training of the point cloud network. We adopt the following voxelization parameters: point cloud boundaries range is $x : [0, 44], y : [-22, 22], z : [-4, 18]$, voxel dimensions are set to $[v_x, v_y, v_z] = [0.4, 0.4, 0.2]$. This would allow us to have a final voxel grid with size (110, 110, 110). Both the cost functions in $\mathcal{L}_{\text{local}}$ and $\mathcal{L}_{\text{global}}$ are chosen as smooth L1 loss $||\cdot||_{1;\,\text{smooth}}$ to ensure robustness to outliers. Adam optimizer and LambdaLR learning rate scheduler are utilized in our training pipeline.

5.2. Datasets

Oxford RobotCar Dataset. We utilize the Oxford RobotCar benchmark [25] for evaluation, where the same trajectory was traveled over a year in different times of the day and seasonal conditions. We generate data samples following the same protocol as conducted by Cattaneo et al. [9], where image is recorded every five meters and the corresponding point cloud map is constructed by concatenating the subsequent 2D LiDAR scans. The four test regions are excluded from the training dataset as per [34].

ViViD++ Dataset. Additionally, we assess the performance of our model on the ViViD++ dataset [20], which consists of driving and handheld sequences and offers 3D LiDAR, visual and GPS data. In the scope of our work, we are mainly interested in the urban data, which contains sensor measurements recorded during a day, evening and night. We follow the training procedures proposed by Lee et al. [21] where only the *day1* sequences are used for training, while performing evaluation with *day2* and *night* sequences.

KITTI Odometry Dataset. We further test the generalization capability of our VXP on the KITTI Odometry benchmark [14], which contains sequences with LiDAR scans, images, and ground-truth poses.

5.3. Results

Across various datasets we evaluate different combinations of modalities for query and database: 2D-3D (image query and point cloud database), 3D-2D (point cloud query and image database) and their uni-modal variations, i.e. 2D-2D (image-only) and 3D-3D (point cloud-only).

Oxford RobotCar. We adhere to the evaluation metric employed by Cattaneo [9], in which we select each pair of distinct runs from 23 sequences as query and database. The query contains samples only from the four excluded regions as per [34], while database consists of samples from the entire trajectory. Finally, the average of the recall is computed for all the pairs. In Tab. 1, we compare our model with the existing cross-modal retrieval approaches, such as the method by Cattaneo et al. [9], LC^2 [21] and LIP-Loc [33]. As the code from Cattaneo et al. [9] is not publicly released, we have implemented the approach with the authors' help to the best of our abilities. We report performance on different modality configurations, namely database and query combinations of 2D images and 3D point clouds.

Recall@1%	2D-3D	3D-2D	2D-2D	3D-3D
Cattaneo's [9]	77.3	70.4	96.6	98.4
LC^2 [21]	81.2	73.8	84.1	83.0
LIP-Loc [33]	77.8	73.6	90.2	92.3
VXP (Ours)	**84.4**	**76.9**	**98.8**	**98.8**

Table 1. Retrieval performance compared with existing cross-modal methods on Oxford dataset. Our model consistently outperforms other baselines on both cross- and uni-modal settings.

	2D-2D		3D-3D	
	1	1%	1	1%
AnyLoc [17]	93.5	98.9	–	–
MixVPR [2]	92.8	97.7	–	–
MinkLoc3D-S [46]	–	–	95.8	99.0
CASSPR [40]	–	–	94.7	98.4
VXP (Ours)	92.0	98.8	94.7	98.8

Table 2. Retrieval performance compared with existing uni-modal methods on Oxford dataset. Provided values correspond to Recall@1 and 1%. Our model has comparable performance with the uni-modal state-of-the-art approaches.

Our method outperforms other baselines on 2D-3D place recognition by a significant margin due to the proposed local constraints. We also demonstrate the best performance in the uni-modal retrieval. Fig. 1 shows average recall up to $K = 25$ nearest neighbors for cross-modal place recognition on the Oxford dataset. Our method is the most accurate and precise with respect to all the baselines from Cattaneo et al. [9] and LIP-Loc [33] and across the whole K-range, which validates consistency of our method.

We also compare the performance of our method with the state-of-the-art uni-modal approaches for visual place recognition methods AnyLoc [17] and MixVPR [2], and LiDAR-based retrieval, such as MinkLoc3D-S [46] and CASSPR [40]. Tab. 2 shows our method performs on-par with the uni-modal baselines, while additionally offering cross-modal capabilities that are practical for multi-sensor on-board suites.

ViViD++. We further evaluate our model on the ViViD++ dataset and compare the results of different approaches on *day1*$-$*day2* sequences in Tab. 3. Note that *day1*$-$*day2* represents query from *day1* sequences and database using *day2* sequences. Overall, we outperform the other baselines [9, 21, 33] on the cross-modal place recognition and perform on par with [9] on uni-modal retrieval task.

We also evaluate our method on the night-day retrieval, where database map is recorded in the day and queries are obtained at night. We report average performance computed

	2D-3D		3D-2D		2D-2D		3D-3D	
	1	1%	1	1%	1	1%	1	1%
LC^2 [21]	60.9	96.0	51.8	94.6	69.2	96.9	58.1	96.1
Cattaneo's [9]	87.6	99.6	78.6	98.6	93.4	99.8	91.0	**99.9**
LIP-Loc [33]	73.7	98.4	54.9	93.0	61.1	94.0	78.8	97.4
VXP (Ours)	**96.8**	**99.6**	**94.7**	**99.8**	**96.7**	**99.9**	**97.0**	99.7

Table 3. Retrieval performance (average recall) for top 1 and 1% retrieved places on the ViViD++ dataset (*day1−day2* sequences). Our model outperforms the other baselines on both uni- and cross-modal experiments.

	2D-2D		3D-2D	
	1	1%	1	1%
LC^2 [21]	0.8	5.5	49.4	93.4
Cattaneo's [9]	2.2	10.1	56.9	94.9
LIP-Loc [33]	2.7	12.0	45.5	90.0
VXP (Ours)	**10.2**	**21.7**	**82.0**	**97.5**

Table 4. Retrieval performance (average recall) for top 1 and 1% retrieved places on ViViD++ dataset (*night−day2* sequences). We can observe the advantage of deploying LiDAR scans as a query, which significantly boosts performance for all baselines. Due to the proposed architectural design, our VXP performs the best in both settings.

on the *city night−city day2* and *campus night−campus day2* sequences from the dataset. Despite significant appearance differences between night queries and map samples recorded during the day, our VXP is able to tackle this challenge by incorporating information from the LiDAR scans that are not affected by insufficient lighting conditions. As shown in Tab. 4, image retrieval (2D-2D) struggles in the challenging scenarios of the night-day retrieval, while cross-modal recognition is capable to offer more accurate place recognition performance across all baselines. Moreover, our approach outperforms other methods such as LC^2, [9], and [33] on the 3D-2D place recognition task and shows highly accurate results based on the top selected retrieval candidate. Specifically, on Recall@1 we achieve a boost in performance by a large margin ($\sim 25\%$ improvement), which demonstrates the effectiveness of our pipeline for this challenging scenario.

KITTI Odometry Benchmark. The results are shown in Tab. 5. Different to the evaluation procedure followed by [10, 40] for LiDAR-based place recognition, we propose our own evaluation protocol on the dataset. Specifically, we train the model on sequences 03, 04, 05, 06, 07, 08, 09, 10. For testing we select 4 regions from sequences 00 and 02 and include the remaining parts of the trajectory into the training data. Notably, none of the sequences traverses the same place, so we test our model on completely unseen regions to demonstrate generalisation capability of our method. Further training details are provided in the supplementary.

	2D-3D		3D-2D		2D-2D		3D-3D	
	1	1%	1	1%	1	1%	1	1%
Cattaneo's [9]	15.9	23.4	12.8	28.7	95.7	97.8	58.6	71.3
LIP-Loc [33]	20.0	**40.9**	**21.9**	29.3	29.3	44.0	27.2	37.8
VXP (Ours)	**32.1**	38.6	**36.1**	**38.3**	**97.8**	**100.0**	**86.3**	**89.4**

Table 5. Retrieval performance (average recall) for top 1 and 1% retrieved places on KITTI Odometry dataset (00, 02 sequences). Our model shows competitive performance among all baselines.

	2D-3D		3D-2D	
	1	1%	1	1%
Global-only	41.3	81.5	30.2	74.7
Local + Global (Ours)	**44.6**	**84.4**	**30.8**	**76.9**

Table 6. Ablation study of the local feature optimization (Eq. (4)) for cross-modal retrieval on the Oxford RobotCar benchmark. Introducing local constraints significantly improves retrieval accuracy over global-only baseline (Eq. (5)), which validates our architectural design.

As shown in Tab. 5 our method demonstrates competitive performance on all configurations. Since the full code for the LC^2 was not publicly available at the submission time, we could not provide comparison on this benchmark. While LIP-Loc [33] achieves the best performance on Recall@1% 2D-3D setting, it is more sensitive to the sampling range of the database samples and queries. We provide details of the experiment in the supplementary.

6. Ablation Studies

Local Descriptor Loss Analysis. We evaluate the impact of the Local Descriptor Optimization (Sec. 4.2) on the cross-modal place recognition. As shown in Tab. 6, the proposed combination of local and global optimizations allows the model to effectively bridge the domain gap between image and point cloud and achieve higher cross-modal retrieval performance.

Fine-tuning Image Backbone. Foundation models such as DINO [26] have demonstrated capability of addressing a wide range of tasks [3]. However, we have noticed that their off-the-shelf performance on the visual (2D-2D) place recognition task is quite poor and fine-tuning is necessary to reach a competitive accuracy. Specifically, we scored only 59.5% on the 2D-2D Recall@1 with the pre-trained DINO ViTs-8 model, while with additional fine-tuning we achieved 2D-2D accuracy of 92.0% (Tab. 2). We can also observe the effect of fine-tuning the model on the attention maps. An example from the Oxford benchmark is shown in Fig. 4. Specifically, buildings, road markings and traffic lights receive higher attention scores after fine-tuning, while the car hood is ignored.

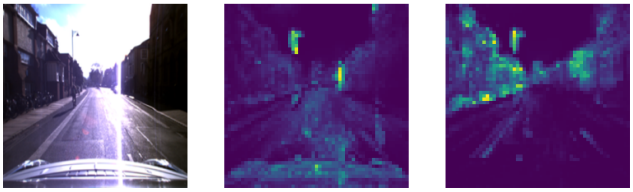

Figure 4. DINO fine-tuning effects on attention maps. From left to right: an input image, an attention map generated by pretrained DINO's ViTs-8 without fine-tuning and a map produced after fine-tuning. Due to the latter, important scene structures such as buildings and traffic poles receive higher attention.

Recall@1%	2D-3D	3D-2D	2D-2D	3D-3D
Ortho-VXP	78.1	71.9	**98.8**	**98.9**
VXP (Ours)	**84.4**	**76.9**	**98.8**	98.8

Table 7. Ablation study of projection module on the Oxford RobotCar dataset. Perspective projection with VXP benefits localization when compared with its orthographic analog, Ortho-VXP.

Voxel-Pixel Projection Module Analysis. We compare our VXP model against a simple baseline, Ortho-VXP, which transforms $\mathbf{V}_{\text{out}}$ to a dense form and performs an orthographic projection of the features to obtain an analogue in the image plane. As shown in Tab. 7, we achieve a boost on the cross-modal localization due to the perspective nature of the VXP module, which associates voxels to corresponding pixels considering the former depth and provides stronger place recognition cues than maintaining original distances and size as per orthographic projection.

Qualitative Evaluation for VXP. As we have shown in Sec. 5, VXP achieves state-of-the-art cross-modal retrieval performance and maintains high uni-modal global localization accuracy. At the same time, we are capable of mitigating the domain gap between different modalities and learning expressive shared latent space. We demonstrate the correlation between the attention map of an RGB image and the feature map of the projected voxels in Fig. 5. Notably, the projected voxels exhibit a similar pattern with the image-based attention map. Since our focus is on place recognition, structures such as buildings carry greater significance, resulting in higher attention scores in those regions for feature maps from both modalities. With this, global descriptors are learned based on consistent information across modalities and we are capable of effectively bridging the domain gap.

Training and Inference Efficiency. We evaluate model inference time using a single RTX3080 and pre-processing time with Intel i7-12700. Depth image generation for LC^2 [21] baseline is done on GPU. Our VXP takes 7 ms to obtain a global descriptor for an image and 18 ms for a point cloud respectively, while LC^2 [21] encodes input image in 17 ms and LiDAR scan in 53 ms due to expensive pre-processing step of depth image generation and point cloud-to-range image conversion. In terms of model parameters and memory footprint, 2D and 3D networks of VXP have 21.7M (87.2MB) and 5.9M (23.6MB) parameters respectively. With this, our model is fast and lightweight to run as part of a real-time system. Notably, the reference map can be encoded offline.

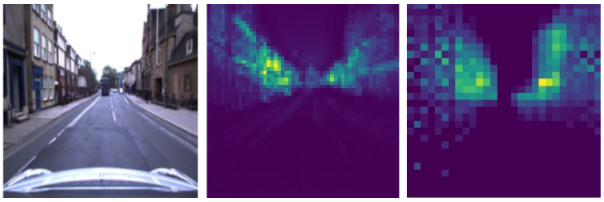

Figure 5. From left to right: an input image, its attention map and projected feature map generated from the respective point cloud.

7. Limitations and Future Work

Our VXP pipeline comprises three steps as described in Sec. 4. Although this multi-stage design showcases the best performance based on our ablation studies (Sec. 6), end-to-end training requires less engineering effort and opens a possibility for generalization when training on larger or multi-source datasets, which is desirable for the autonomous driving applications. In addition, our model is specific for every dataset. While it achieves good performance on the unseen views from the training in-domain dataset, it does not work on different, out-domain sequences. As VXP needs a dataset-specific calibration matrix to establish local descriptor consistency, it remains a limitation towards multi-dataset generalization. Learning calibration on diverse input images and point clouds is a straightforward extension of the VXP pipeline, which is part of the future work.

8. Conclusion

We have presented a new framework, Voxel-Cross-Pixel (VXP), for camera-LiDAR place recognition. VXP makes use of a novel 3D-to-2D projection module specifically designed to establish local feature correspondences and facilitate bridging the domain gap between LiDAR scans and images. To this end, we proposed a cross-modal pipeline, which captures both fine-grained local details and broader global context. Notably, our approach directly works on raw data without any pre-processing steps. Experimental evaluations demonstrate that VXP provides a new state-of-the-art performance on cross-modal image-LiDAR retrieval and offers competitive performance against uni-modal baselines. It shows real-time capability and low memory footprint, which makes it an excellent candidate for deployment on the embedded systems.

References

[1] Amar Ali-bey, Brahim Chaib-draa, and Philippe Giguère. Gsv-cities: Toward appropriate supervised visual place recognition. *Neurocomputing*, 513:194–203, 2022. 2

[2] Amar Ali-Bey, Brahim Chaib-Draa, and Philippe Giguere. Mixvpr: Feature mixing for visual place recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2998–3007, 2023. 2, 6

[3] Shir Amir, Yossi Gandelsman, Shai Bagon, and Tali Dekel. Deep vit features as dense visual descriptors. *arXiv preprint arXiv:2112.05814*, 2(3):4, 2021. 7

[4] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5297–5307, 2016. 1, 2

[5] Dan Barnes, Matthew Gadd, Paul Murcutt, Paul Newman, and Ingmar Posner. The oxford radar robotcar dataset: A radar extension to the oxford robotcar dataset. *arXiv preprint arXiv: 1909.01300*, 2019. 2

[6] Gabriele Berton, Carlo Masone, and Barbara Caputo. Rethinking visual geo-localization for large-scale applications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4878–4888, 2022. 2

[7] Carlos Campos, Richard Elvira, Juan J Gómez Rodríguez, José MM Montiel, and Juan D Tardós. Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam. *IEEE Transactions on Robotics*, 37(6):1874–1890, 2021. 2

[8] Daniele Cattaneo, Matteo Vaghi, Augusto Luis Ballardini, Simone Fontana, Domenico G Sorrenti, and Wolfram Burgard. Cmrnet: Camera to lidar-map registration. In *2019 IEEE intelligent transportation systems conference (ITSC)*, pages 1283–1289. IEEE, 2019. 2

[9] Daniele Cattaneo, Matteo Vaghi, Simone Fontana, Augusto Luis Ballardini, and Domenico G Sorrenti. Global visual localization in lidar-maps through shared 2d-3d embedding space. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4365–4371. IEEE, 2020. 1, 6, 7

[10] Daniele Cattaneo, Matteo Vaghi, and Abhinav Valada. Lcdnet: Deep loop closure detection and point cloud registration for lidar slam. *IEEE Transactions on Robotics*, 38(4):2074–2093, 2022. 7

[11] Zhaoxin Fan, Zhenbo Song, Hongyan Liu, Zhiwu Lu, Jun He, and Xiaoyong Du. Svt-net: Super light-weight sparse voxel transformer for large scale place recognition. AAAI, 2022. 2

[12] Mengdan Feng, Sixing Hu, Marcelo H Ang, and Gim Hee Lee. 2d3d-matchnet: Learning to match keypoints across 2d image and 3d point cloud. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 4790–4796. IEEE, 2019. 3

[13] Dorian Gálvez-López and Juan D Tardos. Bags of binary words for fast place recognition in image sequences. *IEEE Transactions on Robotics*, 28(5):1188–1197, 2012. 2

[14] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012. 6

[15] Mariia Gladkova, Rui Wang, Niclas Zeller, and Daniel Cremers. Tight integration of feature-based relocalization in monocular direct visual odometry. In *2021 IEEE international conference on Robotics and automation (ICRA)*, pages 9608–9614. IEEE, 2021. 2

[16] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2019. 3

[17] Nikhil Keetha, Avneesh Mishra, Jay Karhade, Krishna Murthy Jatavallabhula, Sebastian Scherer, Madhava Krishna, and Sourav Garg. Anyloc: Towards universal visual place recognition. *IEEE Robotics and Automation Letters*, 2023. 2, 6

[18] Jacek Komorowski. Minkloc3d: Point cloud based large-scale place recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1790–1799, 2021. 2, 5

[19] Jacek Komorowski, Monika Wysoczańska, and Tomasz Trzcinski. Minkloc++: lidar and monocular image fusion for place recognition. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2021. 2

[20] Alex Junho Lee, Younggun Cho, Young-sik Shin, Ayoung Kim, and Hyun Myung. Vivid++: Vision for visibility dataset. *IEEE Robotics and Automation Letters*, 7(3):6282–6289, 2022. 6

[21] Alex Junho Lee, Seungwon Song, Hyungtae Lim, Woojoo Lee, and Hyun Myung. Lc^2: Lidar-camera loop constraints for cross-modal place recognition. *IEEE Robotics and Automation Letters*, 8:3589–3596, 2023. 1, 3, 5, 6, 7, 8

[22] María Leyva-Vallina, Nicola Strisciuglio, and Nicolai Petkov. Data-efficient large scale place recognition with graded similarity supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23487–23496, 2023. 3

[23] Jiaxin Li and Gim Hee Lee. Deepi2p: Image-to-point cloud registration via deep classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15960–15969, 2021. 3

[24] Zhe Liu, Shunbo Zhou, Chuanzhe Suo, Peng Yin, Wen Chen, Hesheng Wang, Haoang Li, and Yun-Hui Liu. Lpd-net: 3d point cloud learning for large-scale place recognition and environment analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2831–2840, 2019. 2

[25] Will Maddern, Geoffrey Pascoe, Chris Linegar, and Paul Newman. 1 year, 1000 km: The oxford robotcar dataset. *The International Journal of Robotics Research*, 36(1):3–15, 2017. 6

[26] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2, 7

[27] Yiyuan Pan, Xuecheng Xu, Weijie Li, Yunxiang Cui, Yue Wang, and Rong Xiong. Coral: Colored structural representation for bi-modal place recognition. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2084–2091. IEEE, 2021. 2

[28] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *2007 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2007. 2

[29] 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. 2

[30] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017. 3

[31] Siyu Ren, Yiming Zeng, Junhui Hou, and Xiaodong Chen. Corri2p: Deep image-to-point cloud registration via dense correspondence. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(3):1198–1208, 2022. 3

[32] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 3

[33] Sai Shubodh, Mohammad Omama, Husain Zaidi, Udit Singh Parihar, and Madhava Krishna. Lip-loc: Lidar image pretraining for cross-modal localization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 948–957, 2024. 1, 3, 6, 7

[34] Mikaela Angelina Uy and Gim Hee Lee. Pointnetvlad: Deep point cloud based retrieval for large-scale place recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4470–4479, 2018. 2, 6

[35] Guangming Wang, Yu Zheng, Yanfeng Guo, Zhe Liu, Yixiang Zhu, Wolfram Burgard, and Hesheng Wang. End-to-end 2d-3d registration between image and lidar point cloud for vehicle localization. *arXiv preprint arXiv:2306.11346*, 2023. 3

[36] Yizhou Wang, Zhongyu Jiang, Xiangyu Gao, Jenq-Neng Hwang, Guanbin Xing, and Hui Liu. Rodnet: Radar object detection using cross-modal supervision. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 504–513, 2021. 1

[37] Jamie Watson, Oisin Mac Aodha, Victor Prisacariu, Gabriel Brostow, and Michael Firman. The temporal opportunist: Self-supervised multi-frame monocular depth. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1164–1174, 2021. 3

[38] Chenglu Wen, Yudi Dai, Yan Xia, Yuhan Lian, Jinbin Tan, Cheng Wang, and Jonathan Li. Toward efficient 3-d colored mapping in gps-/gnss-denied environments. *IEEE Geoscience and Remote Sensing Letters*, 17(1):147–151, 2019. 1

[39] Yan Xia, Yusheng Xu, Shuang Li, Rui Wang, Juan Du, Daniel Cremers, and Uwe Stilla. Soe-net: A self-attention and orientation encoding network for point cloud based place recognition. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 11348–11357, 2021. 2

[40] Yan Xia, Mariia Gladkova, Rui Wang, Qianyun Li, Uwe Stilla, João F Henriques, and Daniel Cremers. Casspr: Cross attention single scan place recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8461–8472, 2023. 2, 6, 7

[41] Yan Xia, Qiangqiang Wu, Wei Li, Antoni B Chan, and Uwe Stilla. A lightweight and detector-free 3d single object tracker on point clouds. *IEEE Transactions on Intelligent Transportation Systems*, 24(5):5543–5554, 2023. 1

[42] Yan Xia, Letian Shi, Zifeng Ding, João F Henriques, and Daniel Cremers. Text2loc: 3d point cloud localization from natural language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 2

[43] Yan Yan, Yuxing Mao, and Bo Li. Second: Sparsely embedded convolutional detection. *Sensors*, 18(10):3337, 2018. 4

[44] 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. 3, 4

[45] Zhicheng Zhou, Cheng Zhao, Daniel Adolfsson, Songzhi Su, Yang Gao, Tom Duckett, and Li Sun. Ndt-transformer: Large-scale 3d point cloud localisation using the normal distribution transform representation. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5654–5660. IEEE, 2021. 2

[46] Kamil Żywanowski, Adam Banaszczyk, Michał R Nowicki, and Jacek Komorowski. Minkloc3d-si: 3d lidar place recognition with sparse convolutions, spherical coordinates, and intensity. *IEEE Robotics and Automation Letters*, 7(2):1079–1086, 2021. 6