

# Cross-Modal Visual Relocalization in Prior LiDAR Maps Utilizing Intensity Textures

Qiyuan Shen<sup>1</sup>, Hengwang Zhao<sup>1</sup>, Weihao Yan<sup>1</sup>, Chunxiang Wang<sup>1</sup>, Tong Qin<sup>2</sup>, Ming Yang<sup>1,\*</sup>

**Abstract**—Cross-modal localization has drawn increasing attention in recent years, while the visual relocalization in prior LiDAR maps is less studied. Related methods usually suffer from inconsistency between the 2D texture and 3D geometry, neglecting the intensity features in the LiDAR point cloud. In this paper, we propose a cross-modal visual relocalization system in prior LiDAR maps utilizing intensity textures, which consists of three main modules: map projection, coarse retrieval, and fine relocalization. In the map projection module, we construct the database of intensity channel map images leveraging the dense characteristic of panoramic projection. The coarse retrieval module retrieves the top-K most similar map images to the query image from the database, and retains the top-K' results by covisibility clustering. The fine relocalization module applies a two-stage 2D-3D association and a covisibility inlier selection method to obtain robust correspondences for 6DoF pose estimation. The experimental results on our self-collected datasets demonstrate the effectiveness in both place recognition and pose estimation tasks.

## I. INTRODUCTION

Visual localization is the problem of estimating the 6 Degree-of-Freedom (DoF) camera pose from which a given image was taken relative to a reference scene representation, also known as a type of map-based localization. It is a fundamental aspect of numerous applications, ranging from robotics and autonomous driving to augmented reality.

Typically, visual localization requires consecutive camera frames or independent images to build a scene-specific 3D map using Struct-from-Motion (SfM) or Simultaneous Localization and Mapping (SLAM) methods, with the map then used to estimate the camera pose. However, pure visual methods are challenging due to the large appearance variations caused by viewpoint changes and illumination, and the visual map itself may not be accurate enough compared to natively collected 3D points. To improve visual localization performance, researchers have proposed using cross-modal information, such as LiDAR point clouds [1], making built maps more accurate and robust.

While cross-modal visual localization has drawn increasing attention [2] [3], visual relocalization in prior LiDAR maps has been less studied. Compared to the task above, which can be considered local pose tracking for multi-frames, relocalization focuses on the single-frame global localization in a pre-built map. It is usually utilized in SLAM systems for global pose initialization [4] and loop closure [5].

This work was supported by the National Natural Science Foundation of China (U22A20100 / 62373250 / 62173228). <sup>1</sup>Department of Automation, Shanghai Jiao Tong University, Shanghai, 200240, China. <sup>2</sup>Global Institute of Future Technology, Shanghai Jiao Tong University, Shanghai, China. \*Corresponding author (e-mail: mingyang@sjtu.edu.cn).

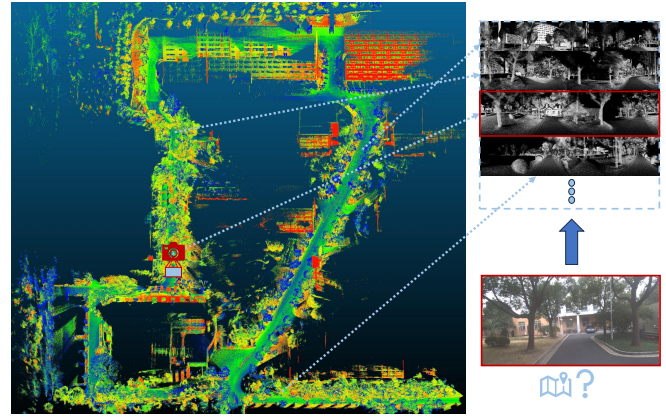


Fig. 1: The cross-modal visual relocalization in prior LiDAR maps. Given a camera image as a query, the system aims to determine its 6DoF pose in the prior LiDAR map.

A relocalization system hierarchically consists of two main components from coarse to fine [6, 7]: place recognition and pose estimation. Place recognition is responsible for finding the top-K most similar place in the map to the query image. The primary obstacle lies in pose estimation, which seeks to determine the 6DoF pose of the query image relative to multiple retrieval results. The top-1 retrieval often lacks precision, leading to significant outliers. Current work in cross-modal fields mainly focuses on either image-point cloud place recognition [8] [9] or 2D-3D registration [10], lacking a complete relocalization system contains place recognition and pose estimation together. One factor contributing to this limitation is the latter task necessitates both a pair of images and point clouds as input, rendering it unsuitable for pose estimation in relocalization scenarios.

In cross-modal association methods [11] for place recognition and 2D-3D registration, the detect-then-match strategy involves separate feature detection in images based on texture and color and point clouds based on geometry, followed by descriptor-based matching. This method encounters challenges due to the differing natures of 2D and 3D feature detection and descriptor encoding, leading to difficulties in detecting consistent features across domains and often resulting in a low inlier ratio.

Recently, intensity has emerged as a valuable source of information in point cloud feature extraction for localization tasks. As a correlated attribute of object reflectivity, intensity serves as a texture feature within the point cloud. This attribute not only aids in LiDAR-based SLAM [12, 13] but also facilitates camera-LiDAR calibration [14] by leveraging its consistency with image grayscale information. Given that

point clouds can be projected onto a range view image, cross-modal correspondence estimation can be facilitated within a single modality by using established image registration techniques, especially with the help of powerful learning-based 2D feature extractors [15] and matchers [16].

Hence, this paper delves into the cross-modal visual relocalization, propose a hierarchical system utilizing texture consistency between the grayscale and intensity values: initially, it involves creating a map image database by leveraging the intensity channel through panoramic projection. Subsequently, the top-K similar map images are retrieved for the query camera image. Following a covisibility clustering process, two-stage 2D-3D association and covisibility inlier selection are conducted to obtain robust 2D-3D correspondences, while removing non-covisible outliers. Finally, Perspective-n-Point (PnP) and Random Sample Consensus (RANSAC) techniques are employed to estimate the 6DoF pose within the prior LiDAR map. Sufficient experiments on self-collected datasets verified the effectiveness of the proposed method with a high recall of both place recognition and pose estimation.

## II. LITERATURE REVIEW

### A. Visual relocalization

Existing visual relocalization methods typically adhere to a hierarchical localization framework [6], comprising offline map construction and online localization stages. The former involves storing global and local features of keyframes and reconstructing a 3D map using Structure from Motion (SfM), while the latter conducts coarse localization via global feature matching with the database and precise localization through local feature matching, yielding 2D-3D correspondences and pose estimation.

Presently, the foremost comprehensive localization solution is HLoc [7], employing image retrieval and feature matching to achieve hierarchical localization. Moreover, another approach employs direct 2D-3D matching [17], associating each 3D point in a pre-built 3D scene model with its corresponding image descriptor. Several methods [18, 19] incorporate the fusion of global and local features from the database, falling under the umbrella of learning-based direct regression. However, these approaches are highly scene-dependent and exhibit limited generalization capabilities.

Nevertheless, purely visual methods encounter numerous challenges, including variations in camera viewpoints and lighting conditions, as well as imprecisions in pre-built visual maps. A significant challenge arises from scale ambiguity in monocular vision, leading to localization results lacking real-world scale. Thus, there is a growing consideration for the adoption of LiDAR point clouds as a scene representation, supplanting visual reconstruction maps.

Currently, there are relatively few comprehensive solutions for cross-modal relocalization, with most efforts focusing solely on cross-modal place recognition. [11] achieves hierarchical localization by simultaneously training global and local descriptors for both point clouds and images. [9] establishes a shared embedding space between images and LiDAR

maps, facilitating image-to-point cloud place recognition. [8] achieves high retrieval recall accuracy by employing a popular multimodal learning approach named CLIP. However, these methods still contend with poor modality consistency, as 2D keypoint detection relies on texture information, while 3D detection is predicated on local geometry, rendering the detection of repeatable keypoints challenging.

### B. Intensity feature extraction

The utilization of the intensity channel within LiDAR point clouds has recently garnered significant attention as a valuable information source for point cloud feature extraction, particularly in localization tasks. Serving akin to a texture-like attribute, intensity offers supplementary insights beyond the inherent 3D geometry. This additional information proves advantageous not only in LiDAR-based SLAM but also in facilitating cross-modal data association between LiDAR and camera sensors.

Previous works have explored LiDAR intensity for place recognition and SLAM. [20] developed a global descriptor using LiDAR intensity histograms, while [21] offered a full SLAM system with intensity-based processing. [12] and [13] utilized ORB features from intensity images for loop closure and addressed geometry degeneracy in SLAM, respectively. Despite the advancements, matching intensity images poses challenges due to the commonly used spherical projection model, which correlates with the spinning LiDAR's line count. This vertical distribution is unsuitable for aligning intensity images with camera images.

Additionally, [14] introduces a comprehensive camera-LiDAR calibration approach utilizing advanced learning-based 2D feature extractors and matchers to establish cross-modal correspondences. Their method leverages mutual information between intensity images and grayscale images to optimize extrinsic transformations. However, these methodologies do not fully exploit the consistency between intensity and grayscale values. The primary hurdle in intensity-based feature extraction is the sparsity of point cloud data, resulting in missing details at the pixel level.

## III. METHODOLOGY

### A. Overview

Under the hierarchical localization framework, our cross-modal relocalization system comprises three primary modules: map projection, coarse retrieval, and fine relocalization, as depicted in Fig. 2. The map projection module converts the 3D point cloud map into a panoramic image using the intensity channel, which serves as the database. Subsequently, the global and local features of each map image are stored in the database. In the coarse retrieval module, the system retrieves the top-K most similar map images to the query image from the database. To mitigate outliers among the top-K candidates, we propose a covisibility clustering method. Subsequently, the fine relocalization module estimates the 6DoF pose of the query image by establishing 2D-3D correspondences in two stages, incorporating covisibility inlier selection, and then solving the PnP problem using RANSAC.

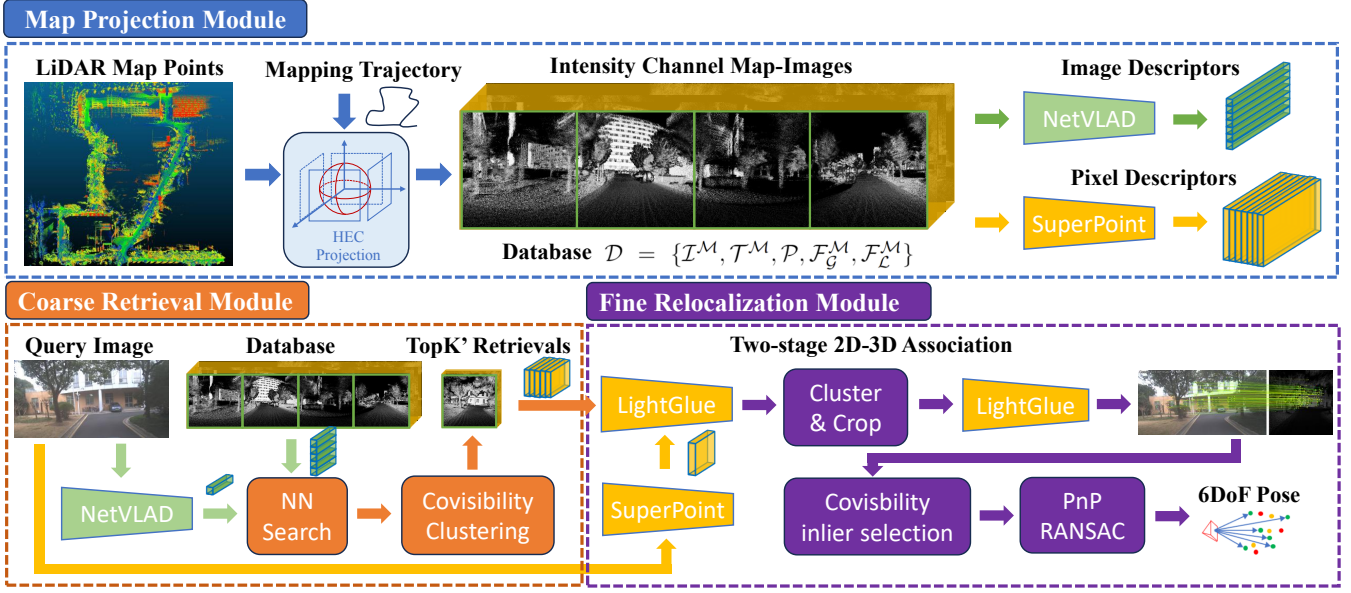


Fig. 2: The hierarchical framework of proposed cross-modal visual relocalization system in prior LiDAR maps.

To demonstrate the generalizability of our method, we employ pre-trained state-of-the-art models for both global and local feature extraction and matching.

### B. Map Projection

Imaging the point cloud plays a vital role in cross-modal registration, as it transforms the LiDAR modal into the visual-like modal for greater consistency. To achieve better texture consistency, we utilize the intensity channel of the LiDAR point cloud, a correlated attribute of reflectivity.

With the previous LiDAR map  $\mathcal{M}$ , the first step is to project the 3D point cloud map into a group of 2D images. To fully leverage the omnidirectional advantage of the 3D map over the 2D image, we choose to project the local map into panoramic images. We sampled the mapping trajectories into fixed intervals to obtain projecting poses  $\mathcal{T}^{\mathcal{M}} = \{T_1^{\mathcal{M}}, T_2^{\mathcal{M}}, \dots, T_N^{\mathcal{M}}\}$ ,  $\mathcal{T} \in \mathbb{R}^{N \times 4 \times 4}$ . For each projecting pose, we transformed the map points from the global coordinate  $\mathcal{P} \in \mathbb{R}^{N \times M \times 3}$  into the local coordinate, then projected the local maps into panoramic images  $\mathcal{I}^{\mathcal{M}} = \{I_1^{\mathcal{M}}, I_2^{\mathcal{M}}, \dots, I_N^{\mathcal{M}}\}$ , where  $\pi$  represents the projection function.

$$\mathcal{I}^{\mathcal{M}} = \pi(\mathcal{T}^{\mathcal{M}}\mathcal{P}) \quad (1)$$

Recent work often utilizes an equirectangular projection model (ERP) for panoramic images, which is simple but results in relatively high distortion [22]. Aside from the uneven distribution defects of the projection format, another challenge lies in missing details at the pixel level due to the sparsity of the point cloud. As depicted in Fig. 3, the spacing between pixels projected onto the (pseudo) camera plane from closer spatial points is larger, leading to a sparse imaging effect.

To ameliorate the uniformity and the sparsity of the projection image, we use an improved method for Cube Map Projection, namely Habrid Equiangular Cube Projection (HEC) [23], adjusting the position of the sampled pixels on

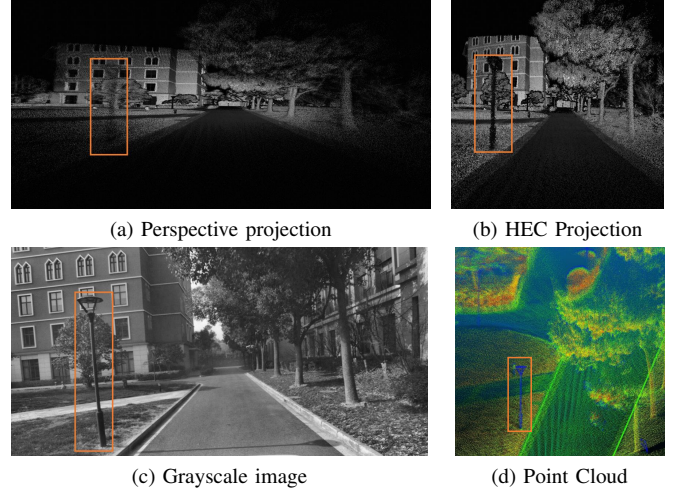


Fig. 3: The consistency comparison of different projection models between equalized map image and grayscale image. The outlined area in (a) recognized as a road lamp, is sparse while the area in (b) is much more dense which is beneficial for the following modules.

the cube corresponding to the spherical pixels, and discard the top surface and the ground surface for the sake of information validity. Given the CubeMap coordinates  $(u, v)$ , the HEC coordinates  $(u', v')$  are calculated by Equation (2) and (3), where  $t = 0.4v(u^2 - 1)$  is the adjusting parameter.

$$u' = \frac{4}{\pi} \times \arctan u \quad (2)$$

$$v' = \begin{cases} v, & t = 0 \\ \frac{1 - \sqrt{1 - 4t(v-t)}}{2t}, & t \neq 0 \end{cases} \quad (3)$$

To enhance uniformity in projected map images and grayscale images, we apply a two-stage process: 3D map

histogram equalization for initial normalization and CLAHE for image contrast enhancement, benefiting feature extraction and matching. For the following coarse-to-fine relocalization, the complete database  $\mathcal{D} = \{\mathcal{I}^M, \mathcal{T}^M, \mathcal{P}, \mathcal{F}_G^M, \mathcal{F}_L^M\}$  stores the map images, relevant projecting poses and global features which will be discussed in the following sections. To be noted, the map projection module is offline preparation.

### C. Coarse Retrieval

1) *Image retrieval*: Visual place recognition has been fully developed in the past decades, we choose the most classic learning-based methods NetVLAD [24] to extract the global feature of the query image and the map image, followed by similarity calculation.

For query image  $\mathcal{I}^Q$ , we initially apply grayscale CLAHE for consistency with intensity images, then extract its global feature  $\mathcal{F}_G^Q$ . We calculate similarity scores against map image  $\mathcal{I}^M$ 's global feature  $\mathcal{F}_G^M$ , retrieving the top-K most similar map images. Using a sliding window strategy, we maximize similarity scores for four cube patches from the panoramic map image, aligning with the resized query image shape corresponding to the CubeMap's projected side length.

2) *Covisibility clustering*: More top-K results increase the chance of accurate matches but also the presence of outliers, impacting fine relocalization. To address this, we use covisibility clustering to refine the top-K candidates. Assuming map images with similar projection poses are covisible, we employ DBSCAN clustering based on these poses, weighted by similarity scores. Candidates are reordered by cluster size and similarity score within each cluster, selecting the top-K' from the largest cluster for further relocalization.

### D. Fine Relocalization

1) *Two-stage 2D-3D association*: In the fine relocalization module, the aim is to estimate the 6DoF pose of the query image given top-K' candidates. A crucial step involves establishing 2D-3D correspondences between the query image and the map image. However, a single local feature match may not yield sufficient high-quality correspondences due to the inherent sparsity of map images and texture inconsistency between grayscale and intensity images.

The two-stage local feature matching strategy is designed to increase the number of high-quality 2D-3D correspondences by extracting the most consistent region of the query image and map image. For pairs between the query image and each map image candidate, we first conduct local feature matching and cluster the matches to obtain bounding boxes in both the query image and the map image. If the largest cluster does not contain enough matches, the pair of the query image and map image is discarded.

Then we crop the bounding boxes from the original images and extract the local features again. Second-stage matching is conducted between the cropped local features, which is expected to provide more high-quality matches. The first-stage correspondences are concatenated with the second-stage correspondences to form the final 2D-2D correspondences as shown in Fig. 4. With the correlation between the

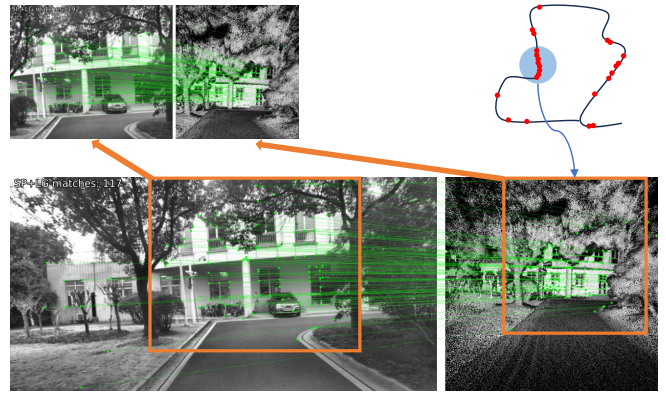


Fig. 4: The two-stage 2D-3D association results. After the covisibility clustering shown in the upper right corner, top-K' candidates are then matched with query image. The below part represents the first stage of association, while the upper left corner represents second stage results in bounding box.

map image  $\mathcal{I}^M$  and map-points  $\mathcal{P}$  in the pre-built database  $\mathcal{D}$ , the 2D-3D correspondences are expected to be easily established given the 2D-2D correspondences.

During the first stage, we use SuperPoint [15] to extract the local features of the query image  $\mathcal{F}_L^Q$  and use LightGlue [16] to match with the local feature of the retrieved map image  $\mathcal{F}_L^M$ . For the second stage, we use the same combination to extract more high-quality correspondences between the cropped query image and map image.

2) *Covisibility inlier selection*: In the 2D-3D association process, we implement covisibility inlier selection to further enhance correspondence quality, removing outliers with high texture similarity but from different scenes. The covisibility of the local feature points is defined as the number of map images in the database that contain the local feature points. It is assumed that the local feature points with higher covisibility are more likely to be in the same scene as the query image. We calculate the covisibility of the local feature points in the map image and filter the correspondences based on the covisibility. The correspondences with higher covisibility are retained, while the correspondences with lower covisibility are discarded.

3) *PnP with RANSAC*: Given the 2D-3D correspondences, which respectively denote the 2D keypoint coordinates  $\mathcal{K}^Q = [x_1^Q, x_2^Q, \dots, x_J^Q]$  and 3D coordinates  $\mathcal{P} = [p_1^M, p_2^M, \dots, p_J^M]$ , we apply PnP with RANSAC to determine the query image's pose  $\mathcal{T}_M^Q$  relative to the prior LiDAR map.

$$\mathcal{T}_M^Q = \arg \min_{\mathcal{T}_M^Q} \sum_{j=1}^J \rho(\|\pi(\mathcal{T}_M^Q p_j^M) - x_j^Q\|^2) \quad (4)$$

## IV. EXPERIMENTS

### A. Dataset

We evaluate the proposed cross-modal visual relocalization system on our self-collected campus datasets. The dataset contains two sequences, with one scene in the college buildings and the other in the dormitory buildings. Each of these consists of two parts: a prior LiDAR point cloud map

and query camera images. The point cloud data was collected using a Livox Avia, then constructed into a prior 3D map leveraging the LiDAR SLAM method. To be noted, query camera images with  $1920 \times 960$  resolution were captured in different weather conditions a week later using a standard-fov monocular camera. To evaluate the performance of the proposed method, our dataset was annotated with ground truth camera poses, which were obtained by map-based LiDAR localization within the same prior maps.

During offline map projection, the sample intervals are set to 1m, and local maps are filtered by a max distance of 50m, while the resolution of the projection image is set to  $1920 \times 480$  representing four  $480 \times 480$  cubes.

### B. Quantitative Evaluation

Since our system works hierarchically, we naturally evaluated in coarse retrieval procedure and fine relocalization procedure respectively. The quantitative evaluation for coarse retrieval was conducted following the standard place recognition whose evaluation metric is Recall@K: the ratio of the number of pairs that have at least one correct retrieval in the top-K retrieval results to the total number of pairs. Since our datasets are not large-scale, we consider the retrieval result with translation error below 5m instead of 20m as the correct retrieval result. Specifically, we choose the top-1, top-5, top-10, top-30 and top-50 retrieval results to evaluate the performance of the proposed method.

TABLE I: The Coarse Retrieval Results of Different Methods on SJTU Datasets.(R@K represents Recall@K)

Datasets	Methods	R@1	R@5	R@10	R@30	R@50
College 314m	LIP-Loc	0.20	0.28	0.48	0.95	<b>1.0</b>
	NetVLAD	0.33	0.68	0.84	0.97	<b>1.0</b>
	Patch-NetVLAD	0.65	0.91	0.97	<b>0.99</b>	0.99
	Cosplace	0.30	0.60	0.69	0.89	0.95
	Eigenplaces	0.67	0.81	0.91	0.94	0.98
	Ours	<b>0.73</b>	<b>0.94</b>	<b>0.99</b>	<b>0.99</b>	<b>1.0</b>
Dormitory 689m	LIP-Loc	0.22	0.34	0.55	0.80	0.90
	NetVLAD	0.29	0.60	0.68	0.87	<b>0.94</b>
	Patch-NetVLAD	0.37	<b>0.68</b>	0.75	0.90	0.93
	Cosplace	0.13	0.39	0.54	0.66	0.77
	Eigenplaces	0.26	0.55	0.68	0.78	0.84
	Ours	<b>0.39</b>	0.65	<b>0.82</b>	<b>0.93</b>	<b>0.94</b>

The existing place recognition methods for comparison are cross-modal based LIP-loc [8] and visual-based methods including NetVLAD [24], Patch-NetVLAD [25], Cosplace [26], Eigenplaces [27]. Aside from LIP-Loc which uses depth channel map images, the others use intensity channel map images. Considering fairness, for comparison methods, we use their pre-trained models. For each experiment, we resize the input images to 480 to get better performance.

From the results shown in TABLE I, we notice that LIP-Loc with the larger model pertained to KITTI shows bad generalization on our datasets. Our method outperforms the others including the recent state-of-the-art learning-based method Patch-NetVLAD [25], which can be seen as a reranking of NetVLAD results.

As for the fine relocalization evaluation, we present a metric named Relocalization Recall (RR): the ratio of queries

whose translation error and rotation error between the point clouds transformed by the ground truth and the predicted transformation are below certain thresholds, in our case, 0.5m and  $1^\circ$ , 1m and  $3^\circ$ , 3m and  $5^\circ$ . To some extent, Relocalization Recall is capable of representing the success rate in global pose initialization and loop closure. In addition, common localization metrics like RMSE, MSE, MAE, and Max Error are also used in the quantitative evaluation.

To show the superiority of intensity textures, we compare our method with a similar hierarchical pipeline using NetVLAD for retrieval and SuperPoint with LightGlue for pose estimation while the map images are depth channels. The top-50 retrieval candidates are extracted in the coarse-retrieval module for better performance. The TABLE III proves the effectiveness of the proposed relocalization system. Our method shows better performance with relocalization recall and RMSE on both SJTU datasets, aligning well with our expectations. The depth images, while capturing the edge details of the foreground and background, lack the texture consistency seen in grayscale images.

### C. Ablation Study

In our cross-modal relocalization system, the proposed methods share the same ultimate purpose referring to making the grayscale query image and intensity channel map image more texture-consistent, which fully utilize the potential of state-of-the-art visual pretrained models. We focus more on the applications of the proposed relocalization system and, therefore mainly show Relocalization Recall in the ablation study, as it represents the success rate to a certain extent. The results are listed in TABLE II, proving the effectiveness and contributions to the whole system of proposed methods.

TABLE II: The Ablation Study On SJTU-College Datasets.

Methods	RR $\uparrow$ (0.5m/ $1^\circ$ )	RR $\uparrow$ (1m/ $3^\circ$ )	RR $\uparrow$ (3m/ $5^\circ$ )
Ours	<b>0.25</b>	<b>0.97</b>	<b>0.98</b>
Ours w/o HEC projection	0.08	0.61	0.61
Ours w/o Two-stage equalization	0.18	0.92	0.96
Ours w/o Covisibility clustering	0.18	0.92	0.96
Ours w/o Two-stage 2D-3D association	0.06	0.92	0.94
Ours w/o Covisibility inlier selection	0.24	0.90	0.94

The HEC projection aims to enhance the density of the projected map image and two-stage equalization is applied to enhance the visualization effects of the projected map image. Covisibility clustering leverages the adjacency of projecting poses to rerank retrieval results, while two-stage 2D-3D association provides more correct 2D-3D correspondences. Before pose estimation, covisibility inlier selection rejects the local feature points that have lower observed times.

## V. CONCLUSION

In this paper, we proposed a hierarchical visual cross-modal relocalization system utilizing texture consistency between the grayscale and intensity values. The proposed method was evaluated on self-collected datasets, showing the effectiveness of the proposed method with high recall of both place recognition and pose estimation. Taking advantage of

TABLE III: The Fine Relocalization Results of Different Methods On SJTU Datasets.

Datasets	Methods <sup>1</sup>	RR $\uparrow$ (0.5m/1 $^\circ$ )	RR $\uparrow$ (1m/3 $^\circ$ )	RR $\uparrow$ (3m/5 $^\circ$ )	RMSE <sup>2</sup> $\downarrow$ (m)	MSE <sup>2</sup> $\downarrow$ (m)	MAE <sup>2</sup> $\downarrow$ (m)	Max Error <sup>2</sup> $\downarrow$ (m)
College	SP+LG(Depth)	0	0.11	0.24	1.41	1.98	0.98	<b>4.17</b>
	Ours	<b>0.25</b>	<b>0.97</b>	<b>0.98</b>	<b>1.00</b>	<b>1.00</b>	<b>0.77</b>	4.43
Dormitory	SP+LG(Depth)	0.05	0.33	0.49	2.63	6.94	2.22	4.20
	Ours	<b>0.54</b>	<b>0.79</b>	<b>0.85</b>	<b>1.18</b>	<b>1.39</b>	<b>0.98</b>	<b>3.61</b>

<sup>1</sup> SP+LG represents using SuperPoint [15] and LightGlue [16] once in fine relocalization while the coarse retrieval stays the same with the proposed method. Depth means the depth-channel map image.

<sup>2</sup> The RMSE, MSE, MAE, and Max Error are calculated based on correct relocalized queries, included in the category of RR(3m/5 $^\circ$ ).

intensity as texture, the proposed method is expected to be a promising solution for cross-modal visual relocalization in prior LiDAR maps.

In future work, we are eager to further explore the intensity's role in cross-modal relocalization. For example, re-train the global retrieval and local matching network to learn the consistency more tightly. To alleviate the influence of sparsity, endow the network with the ability to utilize both geometry and texture information.

#### REFERENCES

- [1] T. Caselitz, B. Steder, M. Ruhnke, and W. Burgard, "Monocular camera localization in 3d lidar maps," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2016, pp. 1926–1931.
- [2] C. Zhang, H. Zhao, C. Wang, X. Tang, and M. Yang, "Cross-modal monocular localization in prior lidar maps utilizing semantic consistency," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 4004–4010.
- [3] K. Chen, H. Yu, W. Yang, L. Yu, S. Scherer, and G.-S. Xia, "I2d-loc: Camera localization via image to lidar depth flow," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 194, pp. 209–221, 2022.
- [4] C. Yang, Z. Zhou, H. Zhuang, C. Wang, and M. Yang, "Global pose initialization based on gridded gaussian distribution with wasserstein distance," *IEEE Transactions on Intelligent Transportation Systems*, 2023.
- [5] Y. Cui, X. Chen, Y. Zhang, J. Dong, Q. Wu, and F. Zhu, "Bow3d: Bag of words for real-time loop closing in 3d lidar slam," *IEEE Robotics and Automation Letters*, vol. 8, no. 5, pp. 2828–2835, 2022.
- [6] M. Humenberger, Y. Cabon, N. Guerin, J. Morat, V. Leroy, J. Revaud, P. Rerole, N. Pion, C. de Souza, and G. Csurka, "Robust image retrieval-based visual localization using kapture," *arXiv preprint arXiv:2007.13867*, 2020.
- [7] P.-E. Sarlin, C. Cadena, R. Siegwart, and M. Dymczyk, "From coarse to fine: Robust hierarchical localization at large scale," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 716–12 725.
- [8] S. Shubodh, M. Omama, H. Zaidi, U. S. Parihar, and M. Krishna, "Lip-loc: Lidar image pretraining for cross-modal localization," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 948–957.
- [9] D. Cattaneo, M. Vaghi, S. Fontana, A. L. Ballardini, and D. G. Sorrenti, "Global visual localization in lidar-maps through shared 2d-3d embedding space," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 4365–4371.
- [10] M. Li, Z. Qin, Z. Gao, R. Yi, C. Zhu, Y. Guo, and K. Xu, "2d3d-matr: 2d-3d matching transformer for detection-free registration between images and point clouds," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 14 128–14 138.
- [11] M. Feng, S. Hu, M. H. Ang, and G. H. Lee, "2d3d-matchnet: Learning to match keypoints across 2d image and 3d point cloud," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 4790–4796.
- [12] T. Shan, B. Englot, F. Duarte, C. Ratti, and D. Rus, "Robust place recognition using an imaging lidar," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 5469–5475.
- [13] W. Du and G. Beltrame, "Real-time simultaneous localization and mapping with lidar intensity," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 4164–4170.
- [14] K. Koide, S. Oishi, M. Yokozuka, and A. Banno, "General, single-shot, target-less, and automatic lidar-camera extrinsic calibration toolbox," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 11 301–11 307.
- [15] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superpoint: Self-supervised interest point detection and description," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 224–236.
- [16] P. Lindenberger, P.-E. Sarlin, and M. Pollefeys, "Lightglue: Local feature matching at light speed," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 17 627–17 638.
- [17] Z. Song, C. Wang, Y. Liu, and S. Shen, "Recalling direct 2d-3d matches for large-scale visual localization," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 1191–1197.
- [18] S. Wang, Z. Laskar, I. Melekhov, X. Li, Y. Zhao, G. Toliass, and J. Kannala, "Hscnet++: Hierarchical scene coordinate classification and regression for visual localization with transformer," *International Journal of Computer Vision*, pp. 1–21, 2024.
- [19] T. Do, O. Miksik, J. DeGol, H. S. Park, and S. N. Sinha, "Learning to detect scene landmarks for camera localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 132–11 142.
- [20] K. P. Cop, P. V. Borges, and R. Dubé, "Delight: An efficient descriptor for global localisation using lidar intensities," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 3653–3660.
- [21] H. Wang, C. Wang, and L. Xie, "Intensity-slam: Intensity assisted localization and mapping for large scale environment," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 1715–1721, 2021.
- [22] H. Pi, S. Tian, M. Lu, J. Liu, Y. Guo, and S. Zhang, "A comprehensive comparison of projections in omnidirectional super-resolution," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.
- [23] J.-L. Lin, Y.-H. Lee, C.-H. Shih, S.-Y. Lin, H.-C. Lin, S.-K. Chang, P. Wang, L. Liu, and C.-C. Ju, "Efficient projection and coding tools for 360 $^\circ$  video," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 9, no. 1, pp. 84–97, 2019.
- [24] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "Netvlad: Cnn architecture for weakly supervised place recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5297–5307.
- [25] S. Hausler, S. Garg, M. Xu, M. Milford, and T. Fischer, "Patch-netvlad: Multi-scale fusion of locally-global descriptors for place recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14 141–14 152.
- [26] G. Berton, C. Masone, and B. Caputo, "Rethinking visual geo-localization for large-scale applications," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4878–4888.
- [27] G. Berton, G. Trivigno, B. Caputo, and C. Masone, "Eigenplaces: Training viewpoint robust models for visual place recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 11 080–11 090.