Applied Intelligence
https://doi.org/10.1007/s10489-022-03372-z

2D3D-MVPNet: Learning cross-domain feature descriptors for 2D-3D matching based on multi-view projections of point clouds

Baiqi Lai[1] · Weiquan Liu[1] · Cheng Wang[1] · Xiaoliang Fan[1] · Yangbin Lin[2] · Xuesheng Bian[1] · Shangbin Wu[1] · Ming Cheng[1] · Jonathan Li[3]

Accepted: 9 February 2022

Abstract

Robust local cross-domain feature descriptors of 2D images and 3D point clouds play an important role in 2D and 3D vision applications, e.g. augmented Reality (AR) and robot navigation. Essentially, the robust local cross-domain feature descriptors have the potential to establish a spatial relationship between 2D space and 3D space. However, it is challenging for manual-based or traditional deep learning-based methods to represent the invariant cross-domain feature descriptors between 2D images and 3D point clouds. Specifically, the mainstream point cloud deep learning network is used to extract the global structure information of the scene. Due to the dimensional difference, there is a large gap between the two-dimensional picture and the three-dimensional structure feature in feature accommodation. In this paper, based on the 2D image patch and 3D point cloud volume dataset, a novel network, 2D3D-MVPNet, is proposed to jointly learn robust local cross-domain feature descriptors between 2D images and 3D point clouds. The 2D3D-MVPNet contains a point cloud branch and an image branch, which are optimized with triplet loss and a second-order similarity regularization. Specifically, for the point cloud branch, first, a novel point cloud feature descriptor extractor, named the image-based point cloud encoder, is introduced to learn a local 3D feature descriptor consistent with the local 2D feature descriptor, so that the local 3D feature descriptors contain both geometry and colour texture information. Second, to overcome the challenge of random order of projected image inputs, a symmetric function is introduced to deal with the feature combination of point cloud projections. Experiments show that the local cross-domain feature descriptors of 2D images and 3D point clouds learned by 2D3D-MVPNet achieve extraordinary 2D to 3D retrieval performance. In addition, several 3D point cloud registration results demonstrate the effectiveness of the image-based point cloud encoder.

Keywords Cross-domain feature descriptor · 2D-3D matching · Point cloud projection · Image patch · Point cloud volume

1 Introduction

With the development of multisource sensors, different data expressions of the same scene are captured by different sensor perceptions [1]. The 2D images captured by lightweight cameras are a set of two-dimensional grids, which is the most popular data source representing scene information. Specifically, due to the strong applicability of 2D image data formats in deep neural networks, 2D images are widely used in deep learning. However, 2D images have diffculty fully reflecting the real situation of the 3D world due to data dimensions limitations.

3D imaging techniques can be divided into two major categories: 3D based on 2D matching relationships and 3D based on time of flight. First, 3D based on a 2D matching relationship has two popular methods, structured light [2] and stereo vision [3, 4]. The multiple-shot phase shifting method is popular in structured light techniques. The method uses four phase shifts to calculate the phase map. There are also method based on one shot, such as phase pattern based one-shot methods, point-pattern based one-shot structured light methods, line-pattern based

✉ Weiquan Liu
wqliu@xmu.edu.cn

[1] Fujian Key Laboratory of Sensing and Computing for Smart Cities, School of Informatics, Xiamen University, Xiamen, 361005, China

[2] Computer Engineering College, Jimei University, Xiamen, 361021, China

[3] Departments of Geography and Environmental Management and Systems Design Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada

Published online: 03 March 2022

one-shot structured light methods and crossed-line pattern based one-shot structured light methods. For the stereo vision technique, this methods calculate the corresponding points' location between left camera and right camera to estimate the point depth. Second, the time of flight technique is widely used in radar, sonar, laser range finder and Lidar applications. The sensor notes the time of a round trip to estimate the position depth. The generation of 3D models can be reconstructed from images, or can be directly obtained by 3D sensors, such as lidar. For models directly obtained by the 3D sensor, there are no corresponding relationships with 2D images, and they need to be coupled through the matching algorithm. Moreover, a network trained from the 3D model reconstructed from the image can be used as the initial model of transfer learning adapted to match the raw point cloud with the image.

The combination of images and point clouds is in increasing demand for up-to-date spatial information of indoor environments. Recently, 2D-guided precision anchors have been applied to 3D object detection tasks [5]. In fact, by matching images and point clouds, the spatial relationship between 2D and 3D space is established [6], which provides the promotion and reference significance in the development of 2D and 3D computer vision applications, e.g.augmented Reality (AR) and robot pose estimation. The virtual and real registration problems in AR and robot pose estimation can be converted to a retrieval task [7, 8]. Pose estimation based on retrieval tasks has been efficiently used in large-scale localization [9]. Essentially, using the robust local cross-domain feature descriptors (2D and 3D feature descriptors) of images and point clouds for 2D to 3D retrieval is a solution for matching images and point clouds (2D-3D matching). The pipeline schematic of learning local cross-domain feature descriptors between images and point clouds in this paper is shown in Fig. 1. In addition, the incorrect matching problem of inaccurate image patch will affect the performance of cross-domain matching. To solve this problem, we adopt the method of system optimization. First, the image samples multiple 2D image patches, and the point cloud samples multiple 3D point cloud volumes. Then, n 2D image patches retrieve the corresponding 3D point cloud volume and construct n conditions for estimating 2D space to 3D space. Finally, we use the RANSAC algorithm to exclude abnormal conditions and estimate the optimal spatial map.

Pixel-point registration is a practical and direct method for 2D image and 3D point cloud matching. However, in the absence of any 2D images and 3D point cloud calibrations, to achieve pixel-point-based 2D images and the 3D point cloud matching, extracting cross-domain feature descriptors of 2D images and 3D point clouds is a very important basic task. In detail, feature descriptor extraction involves extracting of local information centered on 2D pixels or 3D points, that is, 2D image patches and 3D point cloud volumes.

The matching step is usually divided into three parts: keypoint detection, feature descriptor extraction, and calculation of transformation for matching. In this paper, our work focuses on learning cross-domain (2D image patches and 3D point cloud volumes) feature descriptors to serve the pixel-point-based 2D image and the 3D point cloud matching. Thus, the value of our work is to learn the invariant cross-domain feature descriptors of 2D image patches and 3D point cloud volumes, which is a basic step for 2D images and 3D point cloud registration tasks.

However, the data structure and dimensions between the images and point clouds are extremely inconsistent (shown in Fig. 2), resulting in the domain gap between the images and the point clouds. Specifically, the traditional manually designed 2D and 3D feature descriptors are essentially different, 2D feature descriptors depend on the relationship of pixel valuesz and 3D feature descriptors are calculated based on the spatial geometry of their respective data. Due to the difference in data structures, the 2D feature represents the texture feature and line feature of the scene, and the 3d feature represents the spatial structure feature of the scene. On this basis, 2D features and 3D features have difficult achieving unity in cross-domain matching, so they cannot be directly used for matching tasks. Thus, it is extremely challenging to extract the local cross-domain descriptors of images and point clouds with robust and consistent expression characteristics by using manual feature descriptors.

Recently, several neural networks have attempted to jointly learn the local feature descriptors of image patches and point cloud volumes, such as 2D3D-MatchNet [10], Siam2D3D-Net [11] and LCD [12]. The above networks use traditional 2D and 3D networks that are used to extract co-domain features, and do not consider further unifying the feature preferences between 2D and 3D networks. The local cross-domain feature descriptors learned by the above networks are not robust, which results in 2D-3D mismatching. Particularly, the following reasons make 2D-3D matching based on regular neural networks extremely challenging: 1) The data representations of images and point clouds are different, which makes it impossible to use a common coding network structure to uniformly learn the cross-domain feature descriptions. 2) Unlike images that retain the colour texture information of scene projections, the point cloud mainly retains the geometric structure information of the 3D space. The difference in information between these cross-domain data poses a great challenge to the network. 3) The large domain gap between images and point clouds makes the network difficult to converge.

Inspired by the success achieved by deep learning in computer vision, we propose using the Siamese network

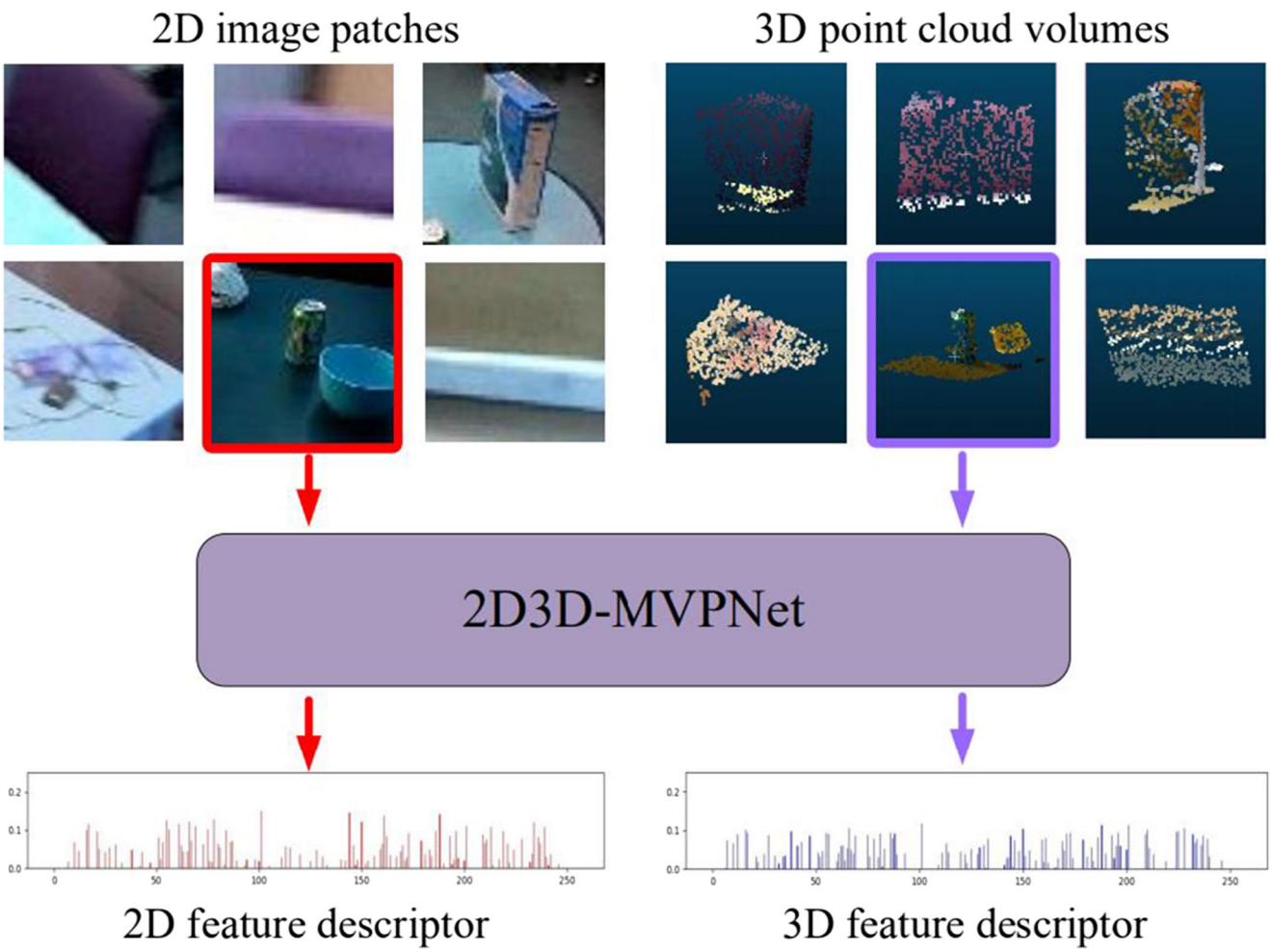


Fig. 1 The pipeline of the local cross-domain descriptors between the 2D image and 3D point cloud learned by our proposed 2D3D-MVPNet. The matching 2D image patches and 3D point cloud volumes are sampled from image and point cloud. The correspondences are fed into 2D3D-MVPNet to propose common feature descriptors

framework, which is a two branch network, to learn robust local cross-domain feature descriptors for 2D-3D matching. Specifically, one branch (point cloud branch) is used to retrieve the raw point cloud volumes and outputs 3D feature descriptors. The other branch (image branch) is used to retrieve the corresponding image patches and outputs 2D

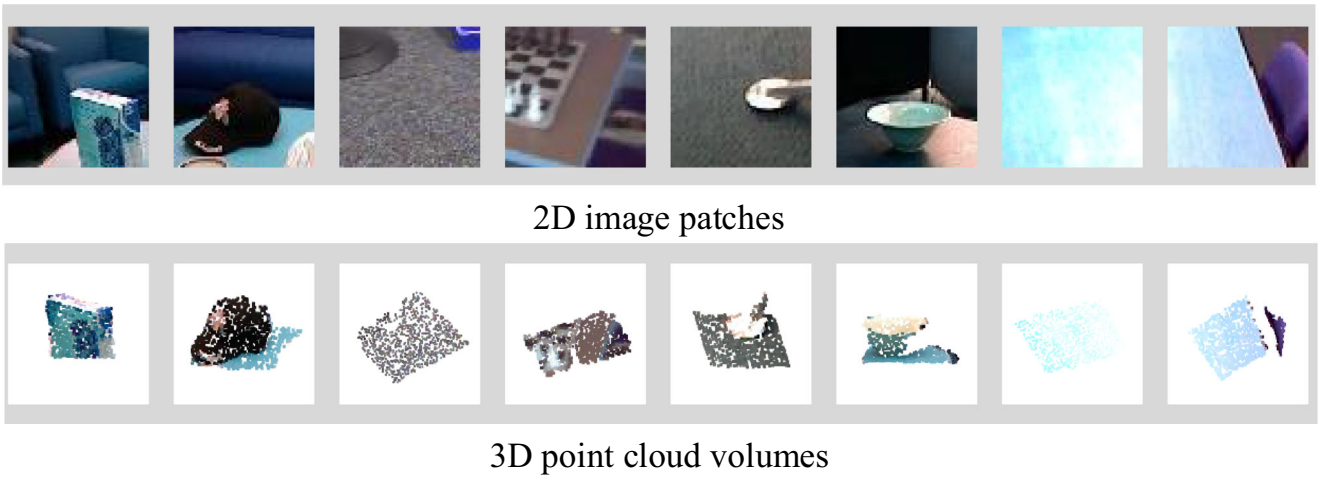


Fig. 2 2D image patches and 3D point cloud volumes data samples. Each column corresponds to matching cross-domain data

feature descriptors. Finally, the metric between the local cross-domain feature descriptors is measured by Euclidean distance.

In the point cloud deep learning network, PointNet [13] solves the point cloud disorder problem. The network uses a pooling function to extract the global features of point cloud voxels, resulting in the lack of features in the multidirectional projection. Projection-based point cloud networks, due to the voxelization and multilayer structure of the point cloud, lack detailed features. This paper proposes a combination of two methods, using the fusion network to fuse the feature information of the two methods. The advantages of using two features learned by the structure extractor and texture extractor instead of single features are as follows: (1) the projection-based network (texture extractor) provides scene colour texture and multi-view information, (2) PointNet (structure extractor) provides global information of the point cloud in space, and (3) this kind of combination of projection texture and global structure information is more similar to real application scenarios and is similar to the way humans observe point cloud information containing both texture information and structural information.

In this paper, we propose a novel network, 2D3D-MVPNet, to jointly learn the robust local cross-domain feature descriptors between images and point clouds. 2D3D-MVPNet is a Siamese framework with an image branch and a point cloud branch for learning the local cross-domain feature descriptors. Specifically, we embed a novel point cloud encoder, named the image-based point cloud encoder, to learn the 3D feature descriptors from the raw point cloud volumes. The proposed image-based point cloud encoder first performs point cloud projections to obtain multiple views from the raw point cloud. Second, features of multi-view projected images are learned by convolutional neural networks (CNNs), and a feature of raw point clouds is learned by PointNet. Then, the multi-view features and raw point cloud features are combined to generate a new point cloud feature descriptor, which contains both geometry and colour texture information. The idea of fusing both patch features and volume features is similar to 3DTNet [14]. However, 3DTNet uses camera patches instead of projections that rely on multisensor fusion technology, and 3DTNet cannot be used in 2D-3D matching tasks. In addition, to overcome the problem of random input order of the multi-view projected images, we propose to use a symmetric function to deal with the combination of point cloud projection features in the process of generating a multi-view feature. Based on the 2D image patch and 3D point cloud volume dataset established by the 3DMatch [15] dataset, experimental results show that the local cross-domain feature descriptors learned by 2D3D-MVPNet achieve state-of-the-art 2D to 3D (image patches to point cloud volumes) retrieval performance. In addition, several point cloud registration results are used to demonstrate the robustness and practicality of the local 3D feature descriptors learned by the proposed image-based point cloud encoder of 2D3D-MVPNet.

The specific contributions of this paper are as follows:

- A novel network framework, 2D3D-MVPNet, which embeds an image-based point cloud encoder, is proposed to jointly learn the robust local cross-domain feature descriptors of images and point clouds. The introduced image-based point cloud encoder assigns both texture and structure information from point cloud projections and raw point clouds, respectively.
- To avoid the interference of the random input order of multi-view projected images on the image-based point cloud encoder, we propose a multifeature fusion module that introduces a symmetric function to ensure the unity of the learned 3D feature descriptors.
- The local cross-domain feature descriptor learned by 2D3D-MVPNet is applied in 2D-3D retrieval and 3D global registration tasks, and the 2D-3D retrieval accuracy achieves state-of-the-art performance.

2 Related work

Effective and ingenious matching network frameworks and feature descriptors have been studied in previous deep learning works. These methods provide guides and references to learn the cross-domain feature descriptors of images and point clouds. In the following, we briefly introduce the deep similarity learning network, 2D image descriptors, 3D point cloud descriptor and matching networks of 2D images and 3D point clouds.

2.1 Deep similarity learning networks

Deep similarity learning networks are used to learn data features and the similarity between different data. Through the data similarity in the high-level feature space, data matching and data retrieval can be completed. Specifically, Siamese networks and triplet networks are popular deep similarity learning networks.

Siamese networks use a two-tower structure that is set to learn feature descriptors. Then, manual functions or a trained deep metric network is used to measure the similarity between the feature descriptors learned by the Siamese framework. MatchNet [16] uses two branches to extract feature descriptors with an additional metric learning network in image patch matching. The metric learning network learns a nonlinear function to measure the similarity between feature descriptors instead of Euclidean

distance. This metric learning network has achieved good performance but requires considerable time and cost to calculate the similarity between feature descriptors. DeepDesc [17] is a typical Siamese network for image patch matching based on feature retrieval. The paired image patches are fed into DeepDesc, which outputs constant dimensional feature descriptors to achieve image patch retrieval based on Euclidean distance. In addition, many image patch-based matching network frameworks have been proposed, such as SiamAM-Net [1], DeepCD [18], L2-Net [19], H-Net [20], DescNet [21], AE-GAN-Net [22] etc. The above networks have undergone different improvements to adapt to more application scenarios and have achieved good results.

Triplet networks introduce a negative sample learning strategy that accepts both positive and negative paired samples. The triplet networks ensure that the positive pairs have a high similarity while ensuring that the negative pairs have a lower similarity. Therefore, compared with the Siamese networks, the triplet networks obtain better performance. The more common triplet networks are FaceNet [23], DOAP [24], DDSAT [25], etc. However, the following difficulties exist in designing a triplet network. 1) Due to the introduction of negative samples, the triplet network optimization process is slow and difficult to converge. 2) Negative samples are difficult to select and define, and inappropriate negative samples seriously affect network performance; therefore, it is particularly important to choose a suitable negative sample construction strategy. 3) It is necessary to set a margin between feature descriptors to separate the positive pairs and the negative pairs, which requires considerable cost for tests and experiments.

2.2 2D and 3D feature descriptors

2D feature descriptors are used to describe the local features of the 2D image grid. Previously, 2D handcrafted feature descriptors have been widely used in image feature description and feature detection. With the development of deep learning, the performance of 2D feature descriptors learned from neural networks has been demonstrated to outperform 2D handcrafted feature descriptors. For example, the 2D feature descriptors learned from DeepDesc [17], DeepCD [18], L2-Net [19] (Siamese networks), FaceNet [23], DOAP [24] and DDSAT [25] (triplet networks) learn the robust 2D feature descriptors. Some novel work recently designed robust descriptors, such as Superpoint [26], R2D2 [27], D2-Net [28], ASLFeat [29].

3D feature descriptors are used to describe the features of the local 3D point cloud. Handcrafted 3D feature descriptors are defined by geometric relationships between points. 3D feature descriptors learned from deep learning networks, such as PointNet [13], PointNet++ [30], PointSIFT [31] and PointCNN [32], allow raw point clouds to be input and local 3D feature descriptors are output, and 3D local descriptors, such as PerfectMatch [33], Ppf-Net [34], FCGF [35], 3DFeat-Net [36], D3Feat [37], have good performance in 3D feature description. In addition, multi-view representations, such as MvCNN [38], GVCNN [39], take advantage of the multi-view representation of point clouds. For volumetric representation, features are passed through two 3D convolutional layers to obtain the final representation, such as 3D ShapeNet [40] and OctNet [41]. The graph-based method constructs the relationship between the point cloud structure through nodes and edges, such as Superpoint Graphs [42]. In recent years, some innovative feature methods have been proposed. PV-RCNN [43] combines both a 3D voxel convolutional neural network (CNN) and PointNet-based set abstraction. FPS-Net [44] explores the uniqueness and discrepancy among the projected image channels.

In studies of 2D feature descriptors and 3D feature descriptors, previous studies have proven their feature expression ability. The 2D feature descriptors are obtained through the pixel relationship, and the 3D feature descriptors are obtained through the geometric position relationship between the point and neighbours. Based on this, the definitions of the 2D and 3D feature descriptors are different; thus, it is difficult to use the existing feature descriptor for cross-domain tasks. Therefore, the study of extracting robust 2D-3D cross-domain descriptors is of significance.

2.3 2D-3D matching networks

2D-3D matching networks, which extract the common cross-domain descriptors of 2D images and 3D point clouds, are applied in cross-domain retrieval tasks between 2D images and 3D point clouds. 2D3D-MatchNet [10] first uses the SIFT keypoints of images and the ISS [45] keypoints of the point cloud to construct outdoor 2D-3D correspondences, and then the triplet network is used for learning local 2D and 3D feature descriptors. Siam2D3D-Net [11] constructs more refined 2D and 3D patch datasets and uses a Siamese network framework to learn local 2D and 3D feature descriptors, which are not robust to noisy datasets. LCD [12] uses an autoencoder with triplet loss to learn cross-domain feature descriptors. 2D-3D LCD correspondences are sampled from 3DMatch [15] and multitask performance has been proven by several experiments. Matching methods in recent years, 2d-3d line correspondences establish the 2D-3D spatial relationship [46]; DeepI2P [47] applies classification to estimate the relative pose; 2D-3D embedding space is used in robotic global localization [48].

3 Network architecture

In this section, we introduce the network framework, loss function and training strategy of the proposed 2D3D-MVPNet in detail.

3.1 2D3D-MVPNet framework

2D3D-MVPNet, as shown in Figs. 3 and 4, is designed to jointly learn a local cross-domain feature descriptor for image patch retrieval point cloud volumes. Because of the different image and point cloud data structures, 2D3D-MVPNet contains two encoders to learn the 2D and 3D feature descriptors. One is the patch encoder (image branch), and the other is the image-based point cloud encoder (point cloud branch). It should be noted that the 2D3D-MVPNet inputs are the matching pairs of image patches and point cloud volumes, whereas the nonmatching image patches and point cloud volumes are generated during the training process.

3.2 Patch encoder architecture

The traditional CNN architecture is introduced to learn 2D feature descriptors for image patches. The patch encoder inputs are the image patches whose size is 64 × 64 × 3. Except for the last layer, batch normalization (BN) and the nonlinear active function ReLU are added to each layer

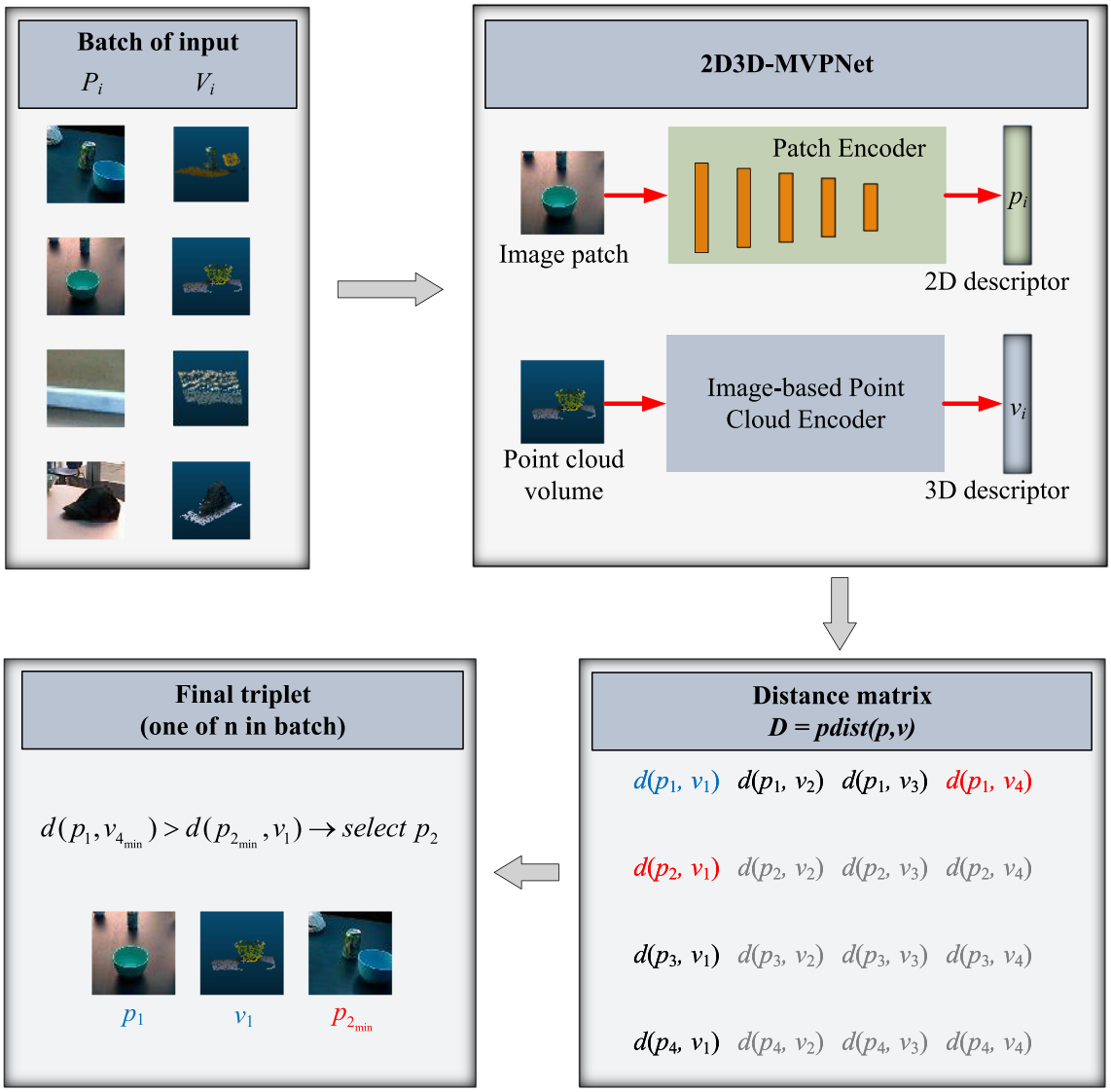


Fig. 3 The network structure of 2D3D-MVPNet and the embedded hard triplet loss schematic. Using $d(p_1, v_1)$ as an example, for the 2D feature descriptor p_1 of the image patch, assuming $d(p_1, v_4)$ is the smallest distance of nonmatching samples; for the 3D feature descriptor v1 of the point cloud volume, assuming $d(p_2, v_1)$ is the smallest distance of nonmatching samples; then, if $d(p_1, v_{4\min}) > d(p_{2\min}, v_1)$, selecting p_2 as the closest distance for v_1, the corresponding P_2 is the hardest negative sample for V_1. The detailed structure of the image-based point cloud encoder is shown in Fig. 4

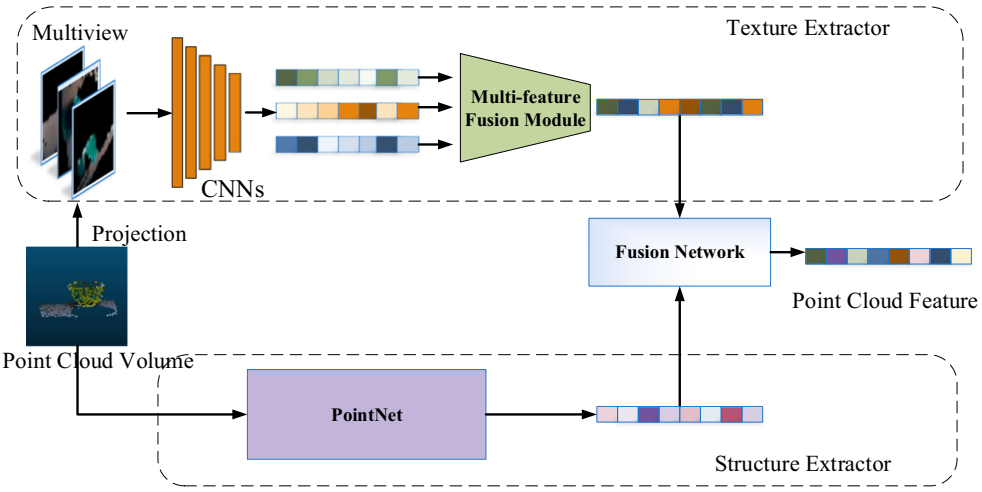


Fig. 4 The decomposed schematic of image-based point cloud encoder. A point cloud volume is input into Texture Extractor and Structure Extractor, and the output features are fused by Fusion Network to generate the final point cloud feature. Texture Extractor uses the point cloud projection basis, and Structure Extractor embeds PointNet as a feature extractor

of the convolutional layer. The specific network structure parameters are set to $C(32, 4, 2) - BN - ReLU - C(64, 4, 2) - BN - ReLU - C(128, 4, 2) - BN - ReLU - C(256, 4, 2) - BN - ReLU - C(256, 4, 4)$, finally obtain a 256-dimensional feature descriptor. $C(n, k, s)$ denotes the convolution layer with n filters of kernel size $k \times k$ with stride s.

3.3 Image-based point cloud encoder architecture

The proposed image-based point cloud encoder incorporates both structural features and texture features by a fusion network, as shown in Fig. 4. On the one hand, the structure feature is directly learned by a structure extractor, which is the PointNet [13] with a fully connected layer. The inputs and outputs of the structure extractor are the raw point cloud with 1024 points and 256-dimensional structure feature descriptors. On the other hand, the texture feature of point cloud volume is learned from the texture extractor, which contains a multi-view projection feature generator and a multifeature fusion module.

3.3.1 Multi-view projections generator

The 2D grids of images and 3D voxels of point clouds reflect the real-world state from the 2D and 3D perspectives, respectively. Thus, the 2D projection of the voxels has a certain degree of correlation with the texture information of 2D grids. Therefore, we consider embedding voxelized point cloud projections as an important part of the point cloud encoder, as shown in Fig. 4.

In our framework, we use three-view projections as multiple-view projections. Furthermore, the texture extractor is designed to extract the target's texture features, and the three-view projections are equipped with redundant texture information. Therefore, the texture feature extraction task will be completed well with three-view projections as multi-view projections.

The image-based point cloud encoder first uses point cloud voxelization to a $32 \times 32 \times 32$ voxel format; then, it is projected to three coordinate planes perpendicular to x, y, z; finally, it saves the projections as the grid with size 64×64. In detail, for a point cloud Volume $\mathrm{P} = \{p_0, p_1, \cdots, p_{1023}\}$ with 1024 points, we define a space cube circumscribed to the point cloud volume and divide it equally into $32 \times 32 \times 32$ small cubes. Each small cube is defined as $V_{i,j,k}$, where i, j, k represent the number of cubes parallel to the x, y, z coordinate axes. A zero-one matrix $M_{32\times32\times32\times1024}$ is constructed to record whether the spatial point cloud falls in $V_{i,j,k}$. The voxel value from the voxelization process is defined as:

$$V_{i,j,k} = avg\left(\sum_{v=0}^{1023} M_{i,j,k,v} \times p_v\right) \tag{1}$$

where p_v is the RGB value of point, and $V_{i,j,k}$ is the voxel value.

Finally, the multi-viewed projected images obtain their respective 256-dimensional feature descriptors by using CNNs, whose structure is the same as that of the patch encoder (Section 3.2). The CNNs used in multi-viewed projections share the same weight.

3.3.2 Multifeature fusion module

The feature fusion of multiple views plays an important role in the texture feature in image-based point cloud encoders. If n features of projected images are simply concatenated and the final 256-dimensional feature is obtained through a fully connected network, there will be $n!$ different feature combination strategies, which will lead to uncontrollable network performance. To address the confusion problem caused by the different input orders of the multi-view projection images, we design a fusion method that integrates the features of the multi-view projections. Benefitting from the uniformity of the symmetric function to the input

order, we choose an effective symmetric function, the sum function, to solve the order problem of the multi-view projections, defined as follows:

$$sum\{f_1, f_2, \cdots, f_n\} = f_{fusion} \tag{2}$$

where $f_i, i = 1, 2, \cdots, n$, is the output feature of the CNN branch with one of the projection inputs, and f_{fusion} denotes the learned texture feature of the point cloud extracted by the texture extractor.

Finally, the learned 256-dimensional structure feature and 256-dimensional temperature feature are incorporated by a fusion network. The detailed structure of the fusion network is $FC(512, 256) - ReLU - FC(256, 256)$, where $FC(p, q)$ represents the input p-dimensional feature vector map to the q-dimensional feature vector through a fully connected network. The input of the fusion network is the 512-dimensional concatenated feature of the structure feature and texture feature. The output of the fusion network is the 256-dimensional feature, i.e., the point cloud volume feature descriptor learned by the proposed image-based point cloud encoder.

3.4 Loss function

Inspired by the negative sample construction strategy of HardNet [49], the nonmatching paired image patches and point cloud volumes are generated from the matching paired image patches and point cloud volumes during training, as shown in Fig. 3. Then, we use the triplet margin loss to optimize 2D3D-MVPNet.

Assuming that there are n pairs of matching image patches and point cloud volumes for each minibatch, the patch encoder and image-based point cloud encoder will output $2n$ cross-domain feature descriptors (n 2D feature descriptors and n 3D feature descriptors). Based on the Euclidean distance of the cross-domain descriptors, the $L2$ pairwise distance matrix$D = cdist(p, v)$ of size $n \times n$ is calculated to construct nonmatching paired image patches and point cloud volumes:

$$\mathrm{D} = \begin{pmatrix} d_{1,1} & \cdots & d_{1,n} \\ \vdots & \ddots & \vdots \\ d_{n,1} & \cdots & d_{n,n} \end{pmatrix} \tag{3}$$

where p denotes the 2D feature descriptors of image patches, v denotes the 3D feature descriptors of point cloud volumes, $d_{i,j} = d\left(p_i, v_j\right) = \sqrt{2 - 2p_i v_i}, i = 1, \cdots, n$ and $j = 1, \cdots, n$. In detail, for any matching pair feature descriptor $\left(p_i, v_j\right), i = j$, the nearest nonmatching samples are measured as follows: for the p_i, the 2^{nd} nearest neighbour is defined as $v_{jmin} = argmin_{j=1,\cdots,n,i\neq j} d\left(p_i, v_j\right)$; the same for v_j, the 2^{nd} nearest neighbor is defined as $p_{kmin} = argmim_{k=1,\cdots,n,i\neq k} d\left(p_k, v_j\right)$. The visualization of the sampling strategy for hardest negative samples is shown in Fig. 3.

Finally, with the above matching cross-domain feature descriptor (p_i, v_i) and the closest nonmatching cross-domain feature descriptors $\left(p_i, v_{jmin}\right)$ and $(p_{k\,\min}, v_i)$, the triplet margin loss aims to minimize the distance between matching descriptors and maximize the distance between nonmatching descriptors. additionally, a second-order similarity regularization is attached to loss functions. Thus, the loss function is defined as follows:

$$\begin{aligned} \mathrm{L} = & \frac{1}{n}\sum_{i=1}^{n} \max\{0, 0.25 + d(p_i, v_i) - \min[d(p_i, v_{j_{\min}}), d(p_{k_{\min}}, v_i)]\} \\ & + \frac{1}{n}\sum_{i=1}^{n} \sqrt{\sum_{j\neq i}^{n} (d(p_i, p_j) - d(v_i, v_j))^2} \end{aligned} \tag{4}$$

3.5 Training strategy

During the experiments, 2D3D-MVPNet is implemented by the PyTorch framework and trained with an NVIDIA 3090 GPU. The SGD optimizer is settled for 2D3D- MVPNet. The learning rate is initially set as 0.001, and the momentum is set as 0.9. The weight decay by 0.0005 for every epoch.

4 Experiments and results

In this section, we first introduce the 2D image patch and 3D point cloud volume dataset used in this paper. Second, we demonstrate the state-of-the-art performance of the jointly local cross-domain feature descriptors learned by 2D3D-MVPNet in the 2D-3D retrieval task. Finally, we perform the learned 3D feature descriptors on the point cloud global registration task, which demonstrates the robustness of the learned local cross-domain feature descriptors.

4.1 Dataset

The 2D image patch and 3D point cloud volume dataset used in this paper is generated from the 3DMatch [15] dataset. We choose the subdataset to collect 2D-3D correspondences from 54 RGB-D scans in the 3DMatch dataset. First, for one scan, several 3D points are randomly sampled. Second, each selected 3D point is set as a centre of the sphere to generate 3D point cloud volumes. Third, to obtain 2D-3D correspondences, reprojecting the 3D points that are found in the first step to RGB-D frames. Finally, the corresponding matching image patches are generated by referring to the reprojected points, as the samples shown in Fig. 2.

Table 1 The TOP1 and TOP5 retrieval accuracy of 2D to 3D retrieval between 2D3D-MVPNet and comparative networks. 2D3D-MVPNet has better performance than other networks. The bold font indicates best retrieval performance

	TOP1	TOP5
2D3D-MVPNet (Ours)	**0.8011**	**0.9482**
LCD [12]	0.7174	0.9412
Siam2D3D-Net [11]	0.2123	0.4567
2D3D-MatchNet [10]	0.2097	0.4318

In the experiments, we use 580,000 and 20,000 pairs of corresponding image patches and point cloud volumes as the training and testing data, respectively. The training data and testing data do not intersect with each other.

4.2 2D-3D retrieval

To measure the performance of the local cross-domain feature descriptors learned by 2D3D-MVPNet, we consider using the 2D to 3D retrieval task on the testing data (20,000 pairs of matching image patches and point cloud volumes) to evaluate the learned local cross-domain feature descriptors. The TOP1 and TOP5 retrieval accuracies on the retrieval testing data are used to evaluate 2D3D-MVPNet and all comparative networks. Specifically, the 2D feature descriptor is set as a query to retrieve the 3D feature descriptor to calculate the TOP1 and TOP5 retrieval accuracies. The successful TOP1 retrieval is that the 2D feature descriptor finds the corresponding 3D feature descriptor in the nearest neighbour in cross-domain space; the successful TOP5 retrieval is that the 2D feature descriptor finds the corresponding 3D feature descriptor in the 5-nearest neighbour in cross-domain space.

The TOP1 and TOP5 retrieval accuracy results of 2D3D-MVPNet and comparative networks are shown in Table 1, which shows that our proposed 2D3D-MVPNet achieves state-of-art retrieval performance, i.e., verifying the performance of local cross-domain feature descriptors learned by 2D2D-MVPNet are superior to LCD [12], Siam2D3D-Net [11] and 2D3D-MatchNet [10]. In addition, Fig. 5 shows the TOP5 3D-2D retrieval results of point cloud volumes by using the queried image patches. The

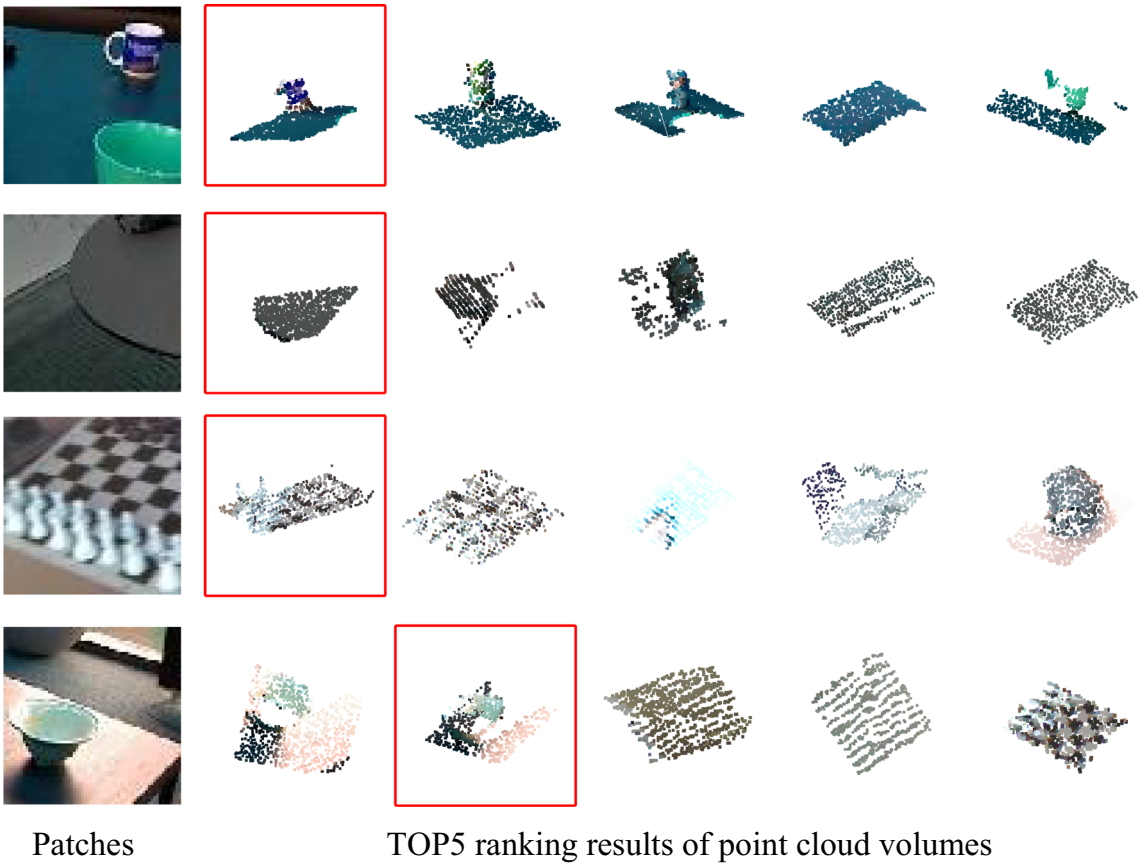


Fig. 5 The TOP5 ranking 2D-3D retrieval result by the local cross-domain feature descriptors learned by 2D3D-MVPNet. The queries are the 2D image patches, and the ground truths and correct retrieval results of 3D point cloud volumes are labeled with the red bounding boxes

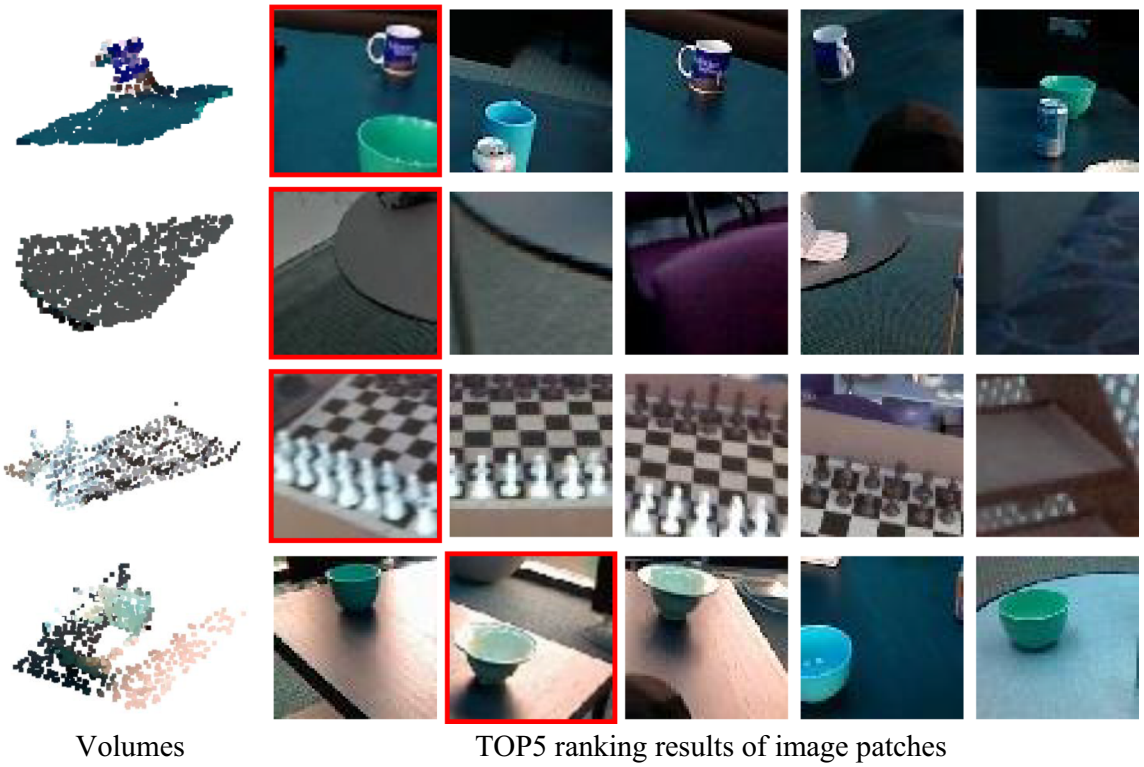
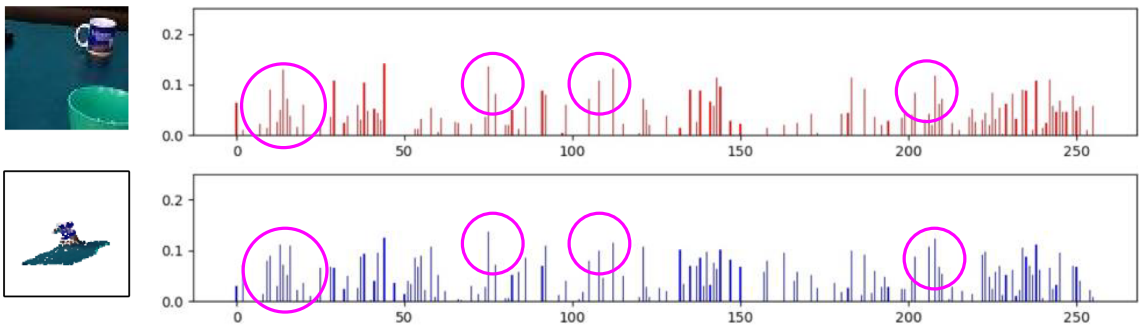
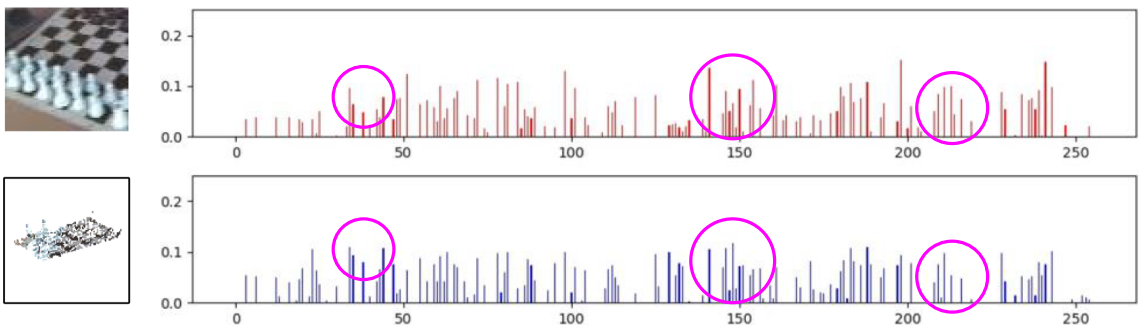


Fig. 6 The TOP5 ranking 3D-2D retrieval result by the local cross-domain feature descriptors learned by 2D3D-MVPNet. The queries are the 3D point cloud volumes, and the ground truths and correct retrieval results of 2D image patches are labelled with the red bounding boxes



(a) Top: Image patch. Bottom: Piont cloud volume.



(b) Top: Image patch. Bottom: Point cloud volume.

Fig. 7 The histogram visualization of the local cross-domain feature descriptors learned by 2D3D-MVPNet between matching 2D image patches and 3D point cloud volumes. The pink circles are the salient area. Top: 2D feature descriptor of the image patch; Bottom: 3D feature descriptor of the point cloud volume

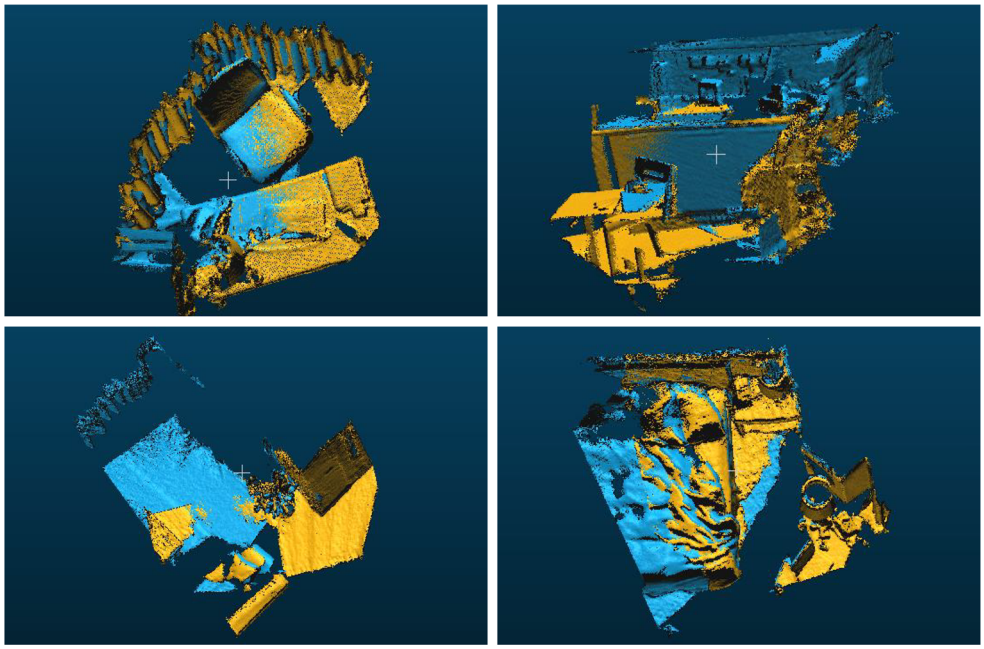


Fig. 8 Visualization of the 3D global registration based on the 3DMatch indoor dataset. Yellow point cloud set as the source and blue point cloud set as the target in our 3D registration experiments

ground truths are labelled with the red bounding boxes. The TOP5 retrieved point cloud volumes have a similar structure, which demonstrates that the local cross-domain feature descriptors learned by 2D3D-MVPNet are robust and contiguous.

Figure 6 shows the results of a 3D-2D search using the point cloud volumes in the red bounding boxes of Fig. 5. The ground truths are labelled with the red bounding boxes. The TOP5 retrieved 2D image patches have similar backgrounds and contents, which demonstrates that the local cross-domain feature descriptors learned by 2D3D-MVPNet are robust and contiguous.

In addition, the 3DMatch dataset is the RGB-D scene reconstructions. However, the RBG-D frame data do not completely cover all the details of the reconstructed scene; thus, the point cloud data in the 3DMatch dataset are inevitably occluded. For example, as shown in Fig. 5, the point cloud data with the cup background in the first and fourth rows is occluded. Therefore, some of the point cloud data of the 3DMatch dataset used in this paper are occluded. It can be seen in Fig. 5 that our proposed 2D3D-MVPNet can also solve the problem of 2D image and 3D point cloud matching with partial occlusion.

Table 2 The performance of local cross-domain feature descriptors learned by 2D3D-MVPNet with different dimensions.256-dimensional descriptors have the most superior TOP1 retrieval performance. And 64-dimensional descriptos have the most superior TOP5 retrieval performance. The bold font indicates best retrieval performance

Dimension	64	128	256	512
TOP1	0.7908	0.7913	**0.8011**	0.7750
TOP5	**0.9612**	0.9589	0.9482	0.9478

4.3 Visualization of learned cross-domain feature descriptor

To more intuitively show the relationship between the local cross-domain feature descriptors learned by 2D3D-MVPNet, we visualized the learned 2D and 3D feature descriptors of the matching image patches and point cloud volumes as a histogram, as shown in Fig. 7. The x-axis and the y-axis are the dimensions and the value of the feature descriptors, respectively. The distribution trend and the salient area of the matching feature histogram are similar (e.g., the pink labelled circled area in the histogram), which also demonstrates the similarity of the local cross-domain feature descriptors learned by 2D3D-MVPNet.

4.4 3D global registration

In addition, we perform the learned 3D feature descriptors on the point cloud global registration task. The 3D global registration works as follows: first, two fragments given in a scan are downsampled to obtain keypoints; second, the point cloud volume with a 30 cm radius is taken for each keypoint; third, each volume is fed to an image-based point cloud encoder to obtain a 3D feature descriptor; finally, all key points are matched by the descriptor nearest search, and the transformation matrix is estimated with RANSAC. Four scenes in the 3DMatch dataset are used as the testing data,

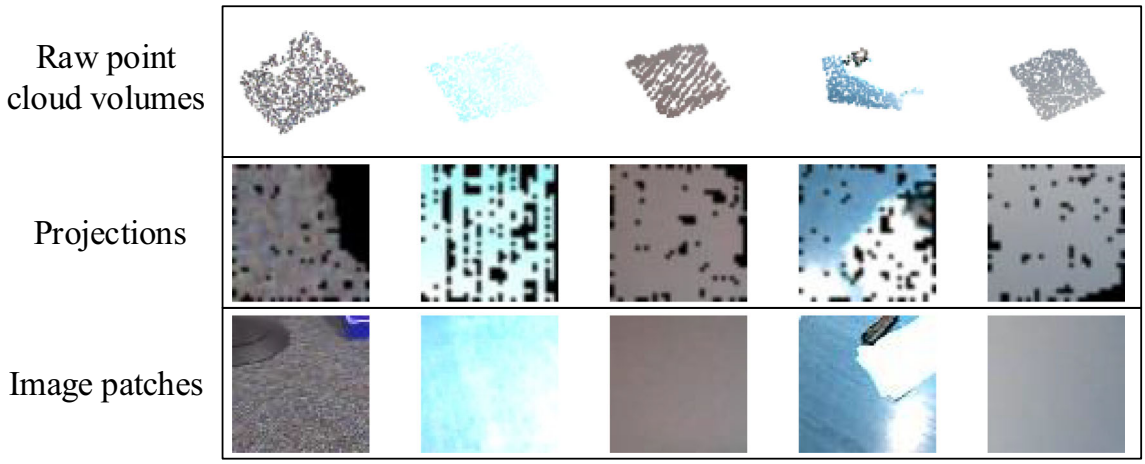


Fig. 9 Visualization of one of the multiple views generated by the point cloud projection and the matching image patches.In the plane structure, point cloud projections and image patches have a high similarity, which is conducive to the completion of 2D-3D Matching

and the 3D global registration results are shown in Fig. 8. The 3D feature descriptors learned by 2D3D-MVPNet work steadily on the point cloud registration, which demonstrates the robustness and practicality of the local 3D feature descriptors learned by the proposed image-based point cloud encoder.

5 Ablation study

To demonstrate the superiority of the proposed 2D3D-MVPNet, we conduct several ablation studies with analyses and discussions. Of note, except in Section 5.1, all the dimensions of the cross-domain feature descriptors learned from 2D3D-MVPNet in the ablation study are set as 256 dimensions.

5.1 Dimension of descriptors

To explore the impact of the dimensions on local cross-domain feature descriptors learned by 2D3D-MVPNet, we conduct experiments with output feature dimensions of 64, 128, 256, 512. The retrieval results are shown in Table 2.

When the feature dimension is 256, the local cross-domain feature descriptors learned by 2D3D-MVPNet have the best TOP1 retrieval performance. However, lower-dimensional features have better TOP5 retrieval performance than high-dimensional features, such as the 64-dimensional feature descriptor having better TOP5 retrieval performance than the 256-dimensional feature descriptor. As the feature dimension increases, its ability to distinguish hard samples and resolution ability improves; however, the overall feature quality decreases.

5.2 Effectiveness of the texture extractor

The feature descriptors learned by PointNet are not robust for plan structure; thus, we construct the texture extractor (TE) to assist 3D feature learning. The raw point cloud volumes of the planar structure have similar geometric information, and the point sets are distributed in a coplanar space. The projections generated by the point cloud volume accurately capture the colour texture similar to the corresponding patch at certain angles. Then, through the multiple views generated by the projection, a texture similar to the patch of the planar structures can be obtained, as shown in Fig. 9. Furthermore, to quantify the significance of the TE, we conduct 2D3D-MVPNet experiments with TE or without a structure extractor (SE), and the results

Table 3 The performance of 2D3D-MVPNet with and without a texture extractor (TE) and structure extractor (SE). TE and SE can be used as point cloud encoders to achieve results in cross-domain retrieval tasks, but the image-based point cloud encoder that combines the two has the best performance. The bold font indicates best retrieval performance

	TOP1	TOP5
2D3D-MVPNet	**0.8011**	**0.9482**
2D3D-MVPNet w/o TE	0.7293	0.9429
2D3D-MVPNet w/o SE	0.6496	0.9258

Table 4 The performance of 2D3D-MVPNet with or without the multifeature fusion module (MfFM). After TE replaced MfFM with FCN, it failed to converge. The reason was that FCN could not adapt to the disorder of input. The bold font indicates best retrieval performance

	2D3D-MVPNet with SE		2D3D-MVPNet w/o SE	
	with MfFM	with FCN	with MfFM	with FCN
TOP1	**0.8011**	0.7491	0.6496	0.0009
TOP5	**0.9482**	0.9403	0.9258	0.0025

are shown in Table 3. Experiments demonstrate that when 2D3D-MVPNet has both TE and SE network structures, the robustness of the learned local cross-domain feature descriptors is better. TE and SE provide richer feature information for cross-domain descriptors, greatly improving retrieval performance, especially in TOP1 retrieval results.

5.3 Effectiveness of multi-feature fusion module

To verify the role of the multifeature fusion module (MfFM) in 2D3D-MVPNet, we first replaced the MfFM with a fully connected network (FCN) (denoted as 2D3D- MVPNet with SE and FCN). All features learned from multiprojections with CNNs were concatenated and fed into the FCN to obtain a constant dimensional vector. In addition, to avoid the self-learning fusion network abandoning the TE branch in the learning process, only effective information was obtained from the SE network. Based on removing SE (denoted as 2D3D-MVPNet w/o SE and with MfFM), we also replaced MfFM with an FCN (denoted as 2D3D-MVPNet w/o SE and with FCN). The experimental results are shown in Table 4 and demonstrate the effectiveness of the symmetric function proposed for different orders of projection input. Experiments also show that the network of 2D3D-MVPNet w/o SE with FCN cannot converge due to the disordered inputs, resulting in a sharp drop in performance. The multifeature fusion module guarantees the common output of the unordered feature input.

6 Conclusion

In this paper, we proposed a novel network, 2D3D-MVPNet, to jointly learn the local cross-domain descriptor for 2D images and 3D point clouds. The proposed image-based point cloud encoder was successfully embedded into 2D3D-MVPNet to learn 3D descriptors that contain both structure and texture information, resulting in improved performance of 2D-3D retrieval. In addition, we proposed a multifeature fusion module based on a symmetric function to solve the problem of random input order of the projections in the texture extractor. Experiments showed that the local cross-domain feature descriptors learned by 2D3D-MVPNet achieved state-of-the-art results in 2D-3D retrieval tasks. Finally, the point cloud feature descriptors were successfully used in the 3D global registration task to verify the robustness and representativeness. In future work, we plan to explore more robust feature descriptors for more 2D-3D data generated from different scenes.

Acknowledgements This work is supported in part by China Postdoctoral Science Foundation (No.2021M690094), in part by National Natural Science Foundation of China (Nos. 61971363, U1605254, 61872306, 61701191, 41871380), in part by Natural Science Fund of Fujian Province (No. 2018J05108), in part by Xia-men Science and Technology Bureau (No. 3502Z20193017) and in part by the China Fundamental Research Funds for the Central Universities (No.20720210074). And we also thank Associate professor Yu Zang from the School of Informatics, Xiamen University, he helped us reorganize the logical relationship and language of this paper during rebuttal progress.

References

1. Liu W, Wang C, Bian X, Chen S, Yu S, Lin X, Lai S-H, Weng D, Li J (2019) Learning to match ground camera image and uav 3-d model-rendered image based on siamese network with attention mechanism. IEEE Geosci Remote Sens Lett 17(9):1608–1612
2. Li Y, Wang Z (2021) 3d reconstruction with single-shot structured light rgb line pattern. Sensors 21(14):4819
3. Li Y, Wang Z (2020) Rgb line pattern-based stereo vision matching for single-shot 3-d measurement. IEEE Trans Instrum Meas 70:1–13
4. Shuang YC, Wang ZZ (2021) Active stereo vision three-dimensional reconstruction by rgb dot pattern projection and ray intersection. Meas 167:108195
5. Yi Wu, Jiang X, Fang Z, Gao Y, Fujita H (2021) Multi-modal 3d object detection by 2d-guided precision anchor proposal and multi-layer fusion. Appl Soft Comput 108:107405
6. Liu W, Lai B, Wang C, Cai G, Yanfei Su, Bian X, Li Y, Chen S, Li J (2020) Ground camera image and large-scale 3-d image-based point cloud registration based on learning domain invariant feature descriptors. IEEE J Sel Top Appl Earth Obs Remote Sens 14:997–1009
7. Li Y, Snavely N, Huttenlocher D, Fua P (2012) Worldwide pose estimation using 3d point clouds. In: European conference on computer vision (ECCV), Springer, pp 15–29
8. Valgren C, Lilienthal AJ (2010) Sift, surf & seasons: Appearance-based long-term localization in outdoor environments. Robot Auton Syst 58(2):149–156
9. Sattler T, Leibe B, Kobbelt L (2016) Efficient & effective prioritized matching for large-scale image-based localization. IEEE Trans Pattern Anal Mach Intell 39(9):1744–1756
10. Feng M, Hu S, Ang MH, Lee GH (2019) 2d3d-matchnet: Learning to match keypoints across 2d image and 3d point cloud. In: 2019 International conference on robotics and automation (ICRA), IEEE, pp 4790–4796
11. Liu W, Lai B, Wang C, Bian X, Yang W, Xia Y, Lin X, Lai S-H, Weng D, Li J (2020) Learning to match 2d images and 3d lidar point clouds for outdoor augmented reality. In: 2020 IEEE Conference on virtual reality and 3d user interfaces abstracts and workshops (VRW), IEEE, pp 654–655
12. Pham Q-H, Uy MA, Hua B-S, Nguyen DT, Roig G, Yeung S-K (2020) Lcd: Learned cross-domain descriptors for 2d-3d matching. In: Proceedings of the AAAI conference on artificial intelligence (AAAI), vol 34, pp 11856–11864
13. Qi CR, Hao Su, Mo K, Guibas LJ (2017) Pointnet: Deep learning on point sets for 3d classification and segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 652–660
14. Xing X, Cai Y, Lu T, Cai S, Yang Y, Wen D (2018) 3dtnet: Learning local features using 2d and 3d cues. In: 2018 International conference on 3d vision (3DV), IEEE, pp 435–443
15. Zeng A, Song S, Nießner M, Fisher M, Xiao J, Funkhouser T (2017) 3dmatch: Learning local geometric descriptors from rgb-d reconstructions. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 1802–1811

16. Han X, Leung T, Jia Y, Sukthankar R, Berg AC (2015) Matchnet: Unifying feature and metric learning for patch-based matching. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 3279–3286
17. Simo-Serra E, Trulls E, Ferraz L, Kokkinos I, Fua P, Moreno-Noguer F (2015) Discriminative learning of deep convolutional feature point descriptors. In: Proceedings of the IEEE international conference on computer vision (ICCV) pp 118–126
18. Yang Tsun-Yi, Hsu Jo-Han, Lin Yen-Yu, Chuang Yung-Yu (2017) Deepcd: Learning deep complementary descriptors for patch representations. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 3314–3322
19. Tian Y, Fan B, Fuchao Wu (2017) L2-net: Deep learning of discriminative patch descriptor in euclidean space. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 661–669
20. Liu W, Shen X, Wang C, Zhang Z, Wen C, Li J (2018) H-net: neural network for cross-domain image patch matching. In: International joint conference on artificial intelligence (IJCAI), pp 856–863
21. Dong Y, Jiao W, Long T, Liu L, He G, Gong C, Guo Y (2019) Local deep descriptor for remote sensing image feature matching. Remote Sens 11(4):430
22. Liu W, Wang C, Bian X, Chen S, Li W, Lin X, Li Y, Weng D, Lai S-H, Li J (2019) Ae-gan-net: Learning invariant feature descriptor to match ground camera images and a large-scale 3d image-based point cloud for outdoor augmented reality. Remote Sens 11(19):2243
23. Schroff F, Kalenichenko D, Philbin J (2015) Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 815–823
24. He K, Yan Lu, Sclaroff S (2018) Local descriptors optimized for average precision. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 596–605
25. Keller M, Chen Z, Maffra F, Schmuck P, Chli M (2018) Learning deep descriptors with scale-aware triplet networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 2762–2770
26. DeTone D, Malisiewicz T, Rabinovich A (2018) Superpoint: Self-supervised interest point detection and description. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops (CVPRW), pp 224–236
27. Revaud J, Weinzaepfel P, Souza CésarD, Pion N, Csurka G, Cabon Y, Humenberger M (2019) R2d2: Repeatable and reliable detector and descriptor. CoRR, arXiv:abs/1906.06195
28. Dusmanu M, Rocco I, Pajdla T, Pollefeys M, Sivic J, Torii A, Sattler T (2019) D2-net: A trainable cnn for joint description and detection of local features. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 8092–8101
29. Luo Z, Zhou L, Bai X, Chen H, Zhang J, Yao Y, Li S, Fang T, Quan L (2020) Aslfeat: Learning local features of accurate shape and localization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 6589–6598
30. Qi CR, Li Yi, Hao Su, Guibas LJ (2017) Pointnet++: Deep hierarchical feature learning on point sets in a metric space. Adv Neural Inform Process Syst 30:5099–5108
31. Jiang M, Wu Y, Zhao T, Zhao Z, Lu C (2018) Pointsift: A sift-like network module for 3d point cloud semantic segmentation. arXiv:1807.00652
32. Li Y, Rui Bu, Sun M, Wei Wu, Di X, Chen B (2018) Pointcnn: Convolution on x-transformed points. Adv Neural Inform Process Syst 31:820–830
33. Gojcic Z, Zhou C, Wegner JD, Wieser A (2019) The perfect match: 3d point cloud matching with smoothed densities. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 5545–5554
34. Deng H, Birdal T, Ilic S (2018) Ppf-foldnet: Unsupervised learning of rotation invariant 3d local descriptors. In: Proceedings of the European conference on computer vision (ECCV), pp 602–618
35. Choy C, Park J, Koltun V (2019) Fully convolutional geometric features. In: Proceedings of the IEEE/CVF international conference on computer vision (ICCV), pp 8958–8966
36. Yew ZJ, Lee GH (2018) 3dfeat-net: Weakly supervised local 3d features for point cloud registration. In: Proceedings of the European conference on computer vision (ECCV), pp 607–623
37. Bai X, Luo Z, Zhou L, Fu H, Quan L, Tai C-L (2020) 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 (CVPR), pp 6359–6367
38. Su H, Maji S, Kalogerakis E, Learned-Miller E (2015) Multi-view convolutional neural networks for 3d shape recognition. In: Proceedings of the IEEE international conference on computer vision (ICCV), pp 945–953
39. Feng Y, Zhang Z, Zhao X, Ji R, Gao Y (2018) Gvcnn: Group-view convolutional neural networks for 3d shape recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 264–272
40. Wu Z, Song S, Khosla A, Yu F, Zhang L, Tang X, Xiao J (2015) 3d shapenets: A deep representation for volumetric shapes. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 1912–1920
41. Riegler G, Ulusoy AO, Geiger A (2017) Octnet: Learning deep 3d representations at high resolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 3577–3586
42. Landrieu L, Simonovsky M (2018) Large-scale point cloud semantic segmentation with superpoint graphs. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 4558–4567
43. Shi S, Guo C, Li J, Wang Z, Shi J, Wang X, Li H (2020) Pv-rcnn: Point-voxel feature set abstraction for 3d object detection. In: In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 10529–10538
44. Xiao A, Yang X, Lu S, Guan D, Huang J (2021) Fps-net: a convolutional fusion network for large-scale lidar point cloud segmentation. ISPRS J Photogramm Remote Sens 176:237–249
45. Zhong Yu (2009) Intrinsic shape signatures: A shape descriptor for 3d object recognition. In: IEEE International conference on computer vision workshops, ICCV workshops, IEEE, pp 689–696
46. Huai Yu, Zhen W, Yang W, Ji Z, Scherer S (2020) Monocular camera localization in prior lidar maps with 2d-3d line correspondences. In: 2020 IEEE/RSJ International conference on intelligent robots and systems (IROS), IEEE, pp 4588–4594
47. Li J, Lee GH (2021) Deepi2p: Image-to-point cloud registration via deep classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 15960–15969
48. Cattaneo D, Vaghi M, Fontana S, Ballardini AL, Sorrenti DG (2020) Global visual localization in lidar-maps through shared 2d-3d embedding space. In: IEEE international conference on robotics and automation (ICRA), IEEE, pp 4365–4371
49. Mishchuk A, Mishkin D, Radenovic F, Matas J (2017) Working hard to know your neighbor's margins: Local descriptor learning loss. In: Advances in neural information processing systems, pp 4826–4837

Baiqi Lai received the B.S degree in informatic engineer from the Sino-European School of Technology, Shanghai University, Shanghai, China, in 2019. He is currently working toward the M.S. degree in computer technology with the Department of Computer Science, Fujian Key Laboratory of Sensing and Computing for Smart Cities and School of Informatics, Xiamen University, Xiamen, China. His current research interests include computer vision, 3-D point cloud data processing, and machine learning.

Weiquan Liu received the B.S. and M.S. degrees in applied mathematics from the College of Science, Jimei University, Xiamen, China, in 2016, and recevied the Ph.D. degree in computer science and technology from the School of Informatics, Xiamen University, Xiamen, China, in 2020. He is currently a Postdoc with the Information and Communication Engineering Postdoctoral Research Station, and the Fujian Key Laboratory of Sensing and Computing for Smart Cities, School of Informatics, Xiamen University, Xiamen, China. His current research interests include 3D vision, machine learning, mobile laser scanning point cloud data processing, and augmented reality.

ChengWang is currently a Nanqiang distinguished professor at Xiamen University, the director of Fujian Key Laboratory of Sensing and Computing for Smart Cities, Xiamen University, and the director of Spatial Sensing and Computing (ASC) Lab. He has authored or coauthored more than 200 papers in referred journals and top-tier conferences including IEEE-TGRS, PR, IEEE-TITS, AAAI, CVPR, IJCAI, and ISPRS-JPRS. His research interests include the field of 3D vision, LiDAR point cloud processing, remote sensing, and spatial bigdata analysis. He was the recipient of Giuseppe Inghilleri Award from The International Society of Photogrammetry and Remote Sensing (ISPRS) in 2020; and the MMT Inovtation Prize from The 11th International Conference on Mobile Mapping Technologys in 2019. He is currently the chair of ISPRS working group on multi-sensor integration and fusion, standing council member of Chinese Society of Image and Graphics. He is also a Fellow of IET.

Xiaoliang Fan received the Ph.D. degree from University Pierre and Marie CURIE, France, in 2012. He is currently a Senior Research Specialist with Fujian Key Laboratory of Sensing and Computing for Smart Cites, Computer Science and Technology Department, Xiamen University. He has published more than 60 journals and conference papers in these areas. His research interests include spatio-temporal data mining and privacy-aware computing. He is a Senior Member of China Computer Federation (CCF).

Yangbin Lin received the B.Sc., M.Sc., and Ph.D. degrees in computer science from Xiamen University, Xiamen, China, in 2008, 2011, and 2016, respectively. He was a Research Assistant with Hong Kong City University, in 2010 and was also a Software Engineer with Google Company (Shanghai), from 2011 to 2012. He is currently a Associate professor with the Computer Engineering College, Jimei University, Xiamen, China. His current research interests include point cloud, graphics, and optimization.

Xuesheng Bian received the M.S. degree in computer technology from School of Informatics, Xiamen University, Xiamen, China, in 2017. He is currently working toward the Ph.D. degree in computer science and technology with the Department of Computer Science, Fujian Key Laboratory of Sensing and Computing for Smart Cities and School of Informatics, Xiamen University, Xiamen, China. His current research interests include computer vision,machine learning, andmedical image analysis.

Shangbin Wu received the B.E degree in computer science and technology from the School of Information Engineering, Chang'An University, Xian, China, in 2018. He is currently working toward the M.S. degree at the Department of Computer Science, Fujian Key Laboratory of Sensing and Computing for Smart Cities and School of Informatics, Xiamen University. Xiamen, China. His current research interests include Spatio-temporal Data Mining & AI algorithms, Urban Computing & Informatics applications.

Ming Cheng received the Ph.D. degree in biomedical engineering from Tsinghua University, Beijing, China, in 2004. He is currently a Professor with Fujian Key Laboratory of Sensing and Computing for Smart Cities, School of Informatics, Xiamen University, Xiamen, China. He has authored over 30 papers in refereed journals and conference proceedings, including IEEE GEOSCIENCE AND REMOTE SENSING LETTERS, Neurocomputing, and IEEE International Geoscience and Remote Sensing Symposium (IGARSS) and International Society for Photogrammetry and Remote Sensing (ISPRS) Proceedings. His research interests include remote sensing image processing, point cloud processing, computer vision, and machine learning.

Jonathan Li received the Ph.D. degree in geomatics engineering from the University of Cape Town, Cape Town, South Africa, in 2000. He is currently a Professor with the Department of Geography and Environmental Management and cross-appointed with the Department of Systems Design Engineering, University of Waterloo, Waterloo, ON, Canada. He is also a Founding Member of the Waterloo Artificial Intelligence Institute. He has coauthored more than 500 publications, over 280 of which were published in refereed journals, including IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, International Society for Photogrammetry and Remote Sensing (ISPRS) Journal of Photogrammetry and Remote Sensing, and Remote Sensing of Environment. His research interests include artificial intelligence (AI) techniques for information extraction from light detection and ranging (LiDAR) point clouds and Earth observation images and their applications in geospatial mapping, transportation, and urban digital twins. Dr. Li was a recipient of the Outstanding Achievement Award in Mobile Mapping Technology in 2019 for his pioneering contributions in developing and promoting mobile mapping technology and the ISPRS Samuel Gamble Award in 2020 for his significant contributions to point cloud analytics in mobile LiDAR mapping. He is also the Chair of the ISPRS WG I/2 on LiDAR, Air- and Space-borne Optical Sensing from 2016 to 2022 and the ICA Commission on Sensor-Driven Mapping from 2015 to 2023. He is also the Editor-in-Chief of the International Journal of Applied Earth Observation and Geoinformation, an Associate Editor of IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, and Canadian Journal of Remote Sensing.