remote sensing

MDPI

Article

2D3D-DescNet: Jointly Learning 2D and 3D Local Feature Descriptors for Cross-Dimensional Matching

Shuting Chen [1], Yanfei Su [2,*], Baiqi Lai [3], Luwei Cai [4], Chengxi Hong [1], Li Li [1], Xiuliang Qiu [1], Hong Jia [3] and Weiquan Liu [3,5]

[1] Chengyi College, Jimei University, Xiamen 361021, China; chenst2016@jmu.edu.cn (S.C.); hongcx0929@jmu.edu.cn (C.H.); lilicy@jmu.edu.cn (L.L.); qiuxiuliang@jmu.edu.cn (X.Q.)
[2] School of Computer and Information Engineering, Xiamen University of Technology, Xiamen 361024, China
[3] Fujian Key Laboratory of Sensing and Computing for Smart Cities, School of Informatics, Xiamen University, Xiamen 361005, China; 23020191153173@stu.xmu.edu.cn (B.L.); jiahong1804@xmu.edu.cn (H.J.); wqliu@jmu.edu.cn (W.L.)
[4] Queen's Business School, Queen's University Belfast, Belfast BT7 1NN, UK; lcai03@qub.ac.uk
[5] College of Computer Engineering, Jimei University, Xiamen 361021, China
* Correspondence: suyanfei@xmut.edu.cn

Abstract: The cross-dimensional matching of 2D images and 3D point clouds is an effective method by which to establish the spatial relationship between 2D and 3D space, which has potential applications in remote sensing and artificial intelligence (AI). In this paper, we propose a novel multi-task network, 2D3D-DescNet, to learn 2D and 3D local feature descriptors jointly and perform cross-dimensional matching of 2D image patches and 3D point cloud volumes. The 2D3D-DescNet contains two branches with which to learn 2D and 3D feature descriptors, respectively, and utilizes a shared decoder to generate the feature maps of 2D image patches and 3D point cloud volumes. Specifically, the generative adversarial network (GAN) strategy is embedded to distinguish the source of the generated feature maps, thereby facilitating the use of the learned 2D and 3D local feature descriptors for cross-dimensional retrieval. Meanwhile, a metric network is embedded to compute the similarity between the learned 2D and 3D local feature descriptors. Finally, we construct a 2D-3D consistent loss function to optimize the 2D3D-DescNet. In this paper, the cross-dimensional matching of 2D images and 3D point clouds is explored with the small object of the 3Dmatch dataset. Experimental results demonstrate that the 2D and 3D local feature descriptors jointly learned by 2D3D-DescNet are similar. In addition, in terms of 2D and 3D cross-dimensional retrieval and matching between 2D image patches and 3D point cloud volumes, the proposed 2D3D-DescNet significantly outperforms the current state-of-the-art approaches based on jointly learning 2D and 3D feature descriptors; the cross-dimensional retrieval at TOP1 on the 3DMatch dataset is improved by over 12%.

Keywords: cross-dimensional matching; local feature descriptors; 2D image patch; 3D point cloud volume

check for updates

Citation: Chen, S.; Su, Y.; Lai, B.; Cai, L.; Hong, C.; Li, L.; Qiu, X.; Jia, H.; Liu, W. 2D3D-DescNet: Jointly Learning 2D and 3D Local Feature Descriptors for Cross-Dimensional Matching. *Remote Sens.* **2024**, *16*, 2493. https://doi.org/10.3390/rs16132493

Academic Editors: Felicia Norma Rebecca Teferle, Abdul Awal Md Nurunnabi, Meida Chen and Yan Xia

Received: 23 May 2024
Revised: 1 July 2024
Accepted: 2 July 2024
Published: 8 July 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Two-dimensional images and three-dimensional point clouds, captured by different sensors, are complementary spatial data, referred to as cross-domain data. The 2D images are represented by 2D grids in 2D space but are limited to reflecting the real scene of the 3D world fully. Meanwhile, 3D point clouds are represented by discrete points with real-point coordinates to perceive the 3D space. Essentially, if the matching relationship between 2D images and 3D point clouds is established, the spatial relationship between 2D and 3D space can be established, which has potential application prospects in AI, e.g., augmented reality [1], robot navigation [2], autonomous driving [3], etc.

In fact, there is large discrepancy between 2D images and 3D point clouds; the data modals are quite different, e.g., cross-modal, cross-dimension, cross-space, such that a

domain gap exists between the 2D images and 3D point clouds. Currently, although the 2D and 3D feature descriptors are widely available, we lack the unified feature descriptors to represent both the 2D images and the 3D point clouds.

In particular, the definitions of mature handcrafted feature descriptors for 2D images and 3D point clouds are extremely different [4], such as the 2D feature descriptors SIFT [5] and SURF [6] and the 3D feature descriptors FPFH [7], SHOT [8], and ROPS [9]. In addition, with the recent advent of deep learning, there are robust 2D and 3D feature descriptors learned automatically through the use of deep neural networks [10–12], such as the 2D learned feature descriptors in DeepDesc [13], L2-Net [14], SOSNet [15], DISK [16], RDLNet [17], and Lightglue [18] and the 3D learned feature descriptors in PointNet++ [19], PointCNN [20], PPFNet [21], D3feat [22], Spinnet [23], and Pointnext [24]. The above learned feature descriptors have demonstrated their robustness and advantages over their mature handcrafted counterparts. However, similar to the issues of 2D and 3D handcrafted feature descriptors, the 2D feature descriptors learned on the 2D domain may not be applicable in 3D space and vice versa.

In this paper, we aim to narrow the domain gap between 2D and 3D feature descriptors by using a learning strategy such that the learned unified feature descriptors can be worked on both 2D and 3D domains. In detail, we propose a novel network, 2D3D-DescNet, to learn 2D and 3D local feature descriptors jointly, which bridges the domain gap between 2D images and 3D point clouds. The 2D3D-DescNet contains three parts: feature extractor, cross-domain image-wise feature map extractor, and metric network. First, the feature extractor with two branches receives raw paired 2D image patches and 3D point cloud volumes to learn 2D and 3D local feature descriptors. Second, the cross-domain image-wise feature map extractor utilizes a shared decoder to generate the feature maps of 2D image patches and 3D point cloud volumes. Meanwhile, and most importantly, the cross-domain image-wise feature map extractor embeds a GAN [25] strategy to distinguish the source of the generated feature maps (essentially a binary classification) so that, in turn, it can promote the consistency of the learned 2D and 3D local feature descriptors. Third, the metric network is constructed to compute the similarity to determine whether the 2D image patches and 3D point cloud volumes match. Finally, the 2D3D-DescNet is optimized by a constructed 2D-3D consistent loss function, which consists of a chamfer, a hard triplet margin, and adversarial and cross-entropy losses. In this paper, the cross-dimensional matching of 2D images and 3D point clouds is explored with the small object of the 3Dmatch dataset. We conduct extensive experiments to demonstrate the invariant of 2D and 3D local feature descriptors learned by the 2D3D-DescNet on the 2D image patch and the 3D point cloud volume dataset. In addition, based on the 2D and 3D local feature descriptors learned by 2D3D-DescNet, the 2D-3D retrieval and 2D-3D matching between 2D image patches and 3D point cloud volumes achieve state-of-the-art performance.

The contributions of this paper include the following:

- A novel end-to-end network, 2D3D-DescNet, is proposed to learn 2D and 3D local feature descriptors, which can work on both 2D and 3D domains.
- The constructed 2D-3D consistent loss balances the 2D and 3D local feature descriptors between 2D and 3D domains and bridges the domain gap between 2D images and 3D point clouds.
- Compared with the current approaches based on jointly learning 2D and 3D feature descriptors, the 2D and 3D local feature descriptors learned by 2D3D-DescNet achieve state-of-the-art performance on 2D-3D retrieval and 2D-3D matching.

2. Related Works

2.1. 2D and 3D Feature Descriptor

The 2D and 3D feature descriptors are widely applied in 2D and 3D computer vision, such as in image matching, image registration, and point cloud registration, etc. The 2D and 3D feature descriptors can be divided into handcrafted and learning-based feature descriptors.

For the handcrafted 2D and 3D feature descriptors, they are often dependent on prior knowledge and constructed from low-level features. On the one hand, the description strategies of 2D handcrafted feature descriptors are usually classified into gradient statistics (SIFT [5], SURF [6]), local binary pattern statistics (LBP [26], RLBP [27]), local intensity comparison (BRIEF [28], ORB [29]), and local-intensity-order-statistics-based methods (LIOP [30,31]). On the other hand, the 3D handcrafted feature descriptors are usually divided into spatial-distribution-histogram-based and geometric-attribute-histogram-based descriptors [32,33], such as FPFH [7], SHOT [8], and ROPS [9].

As for the learning-based 2D and 3D feature descriptors, they are usually learned by the framework of the Siamese network or the triplet network, such as the 2D learned feature descriptors DeepDesc [13], L2-Net [14], SOSNet [15], etc., and the 3D learned feature descriptors in PointNet++ [19], PointCNN [20], PPFNet [21], etc. SOE-Net [34] proposed a self-attention and orientation encoding network to learn the 3D features for point-cloud-based place recognition. CASSPR [35] proposed the fuse-point-based and voxel-based approaches, using cross attention transformers to the learn 3D feature for the place recognition. Text2Loc [36] proposed a novel matching-free fine localization method to learn the 3D feature.

These learned 2D and 3D feature descriptors have demonstrated their robustness and advantages over the mature handcrafted feature descriptors; the deep learning techniques can easily distinguish the same structures, despite the fact that images or point clouds suffer from apparent variance and geometrical transformation [33]. For a comparison of handcrafted and learning feature descriptors for 2D and 3D features, please refer to [33,37–39] for details.

In addition, whether they are handcrafted or learned 2D and 3D feature descriptors, they all face the issue that the 2D feature descriptors learned on the 2D domain may not be applicable in 3D space and vice versa.

2.2. Unified 2D and 3D Feature Descriptor

More recently, there have been several networks used for jointly learning the unified 2D and 3D feature descriptors which can be used both in 2D and 3D domains. 2D3D-MatchNet [40] was the first to propose a deep learning approach to learning the descriptors that allow for direct matching of the keypoints across a 2D image and a 3D point cloud. Siam2D3D-Net [41] use a Siamese network framework to learn the 2D and 3D feature descriptors between camera image patches and LiDAR point cloud volumes which were captured by a mobile laser scanning system. Similarly, H-Net++ [42] embeds the autoencoder into the Siamese network, and LCD [43] embeds the 2D autoencoder and 3D autoencoder into the Siamese network for learning 2D and 3D feature descriptors, respectively. 2D3D-GAN-Net [44] employs the GAN strategy to distinguish the source of the learned feature 2D and 3D descriptors, facilitating the extraction of invariant local cross-domain feature descriptors. HAS-Net [45] introduces the spatial transformer network (STN) module to overcome the translation, scale, rotation, and more generic warping of 2D image patches. 2D3D-MVPNet [46] use a point cloud branch and an image branch to learn the 2D and 3D feature, and the point cloud branch uses multi-view projected images to extract the 3D feature. However, although these learned unified 2D and 3D feature descriptors can be used for retrieval, their robustness is limited and may lead to a mismatch.

3. Method

In this section, we introduce the network structure, the constructed 2D-3D consistent loss function and training strategy of the proposed 2D3D-DescNet in detail.

3.1. 2D3D-DescNet

To learn the 2D and 3D local feature descriptors for 2D images and 3D point clouds, we propose a novel network, 2D3D-DescNet (shown in Figure 1), consisting of a feature extractor, a cross-domain image-wise feature map extractor, and a metric network. It should be noted that the inputs of 2D3D-DescNet are matching 2D image patches (P) and 3D point cloud volumes (V), while the non-matching samples are generated during the training process.

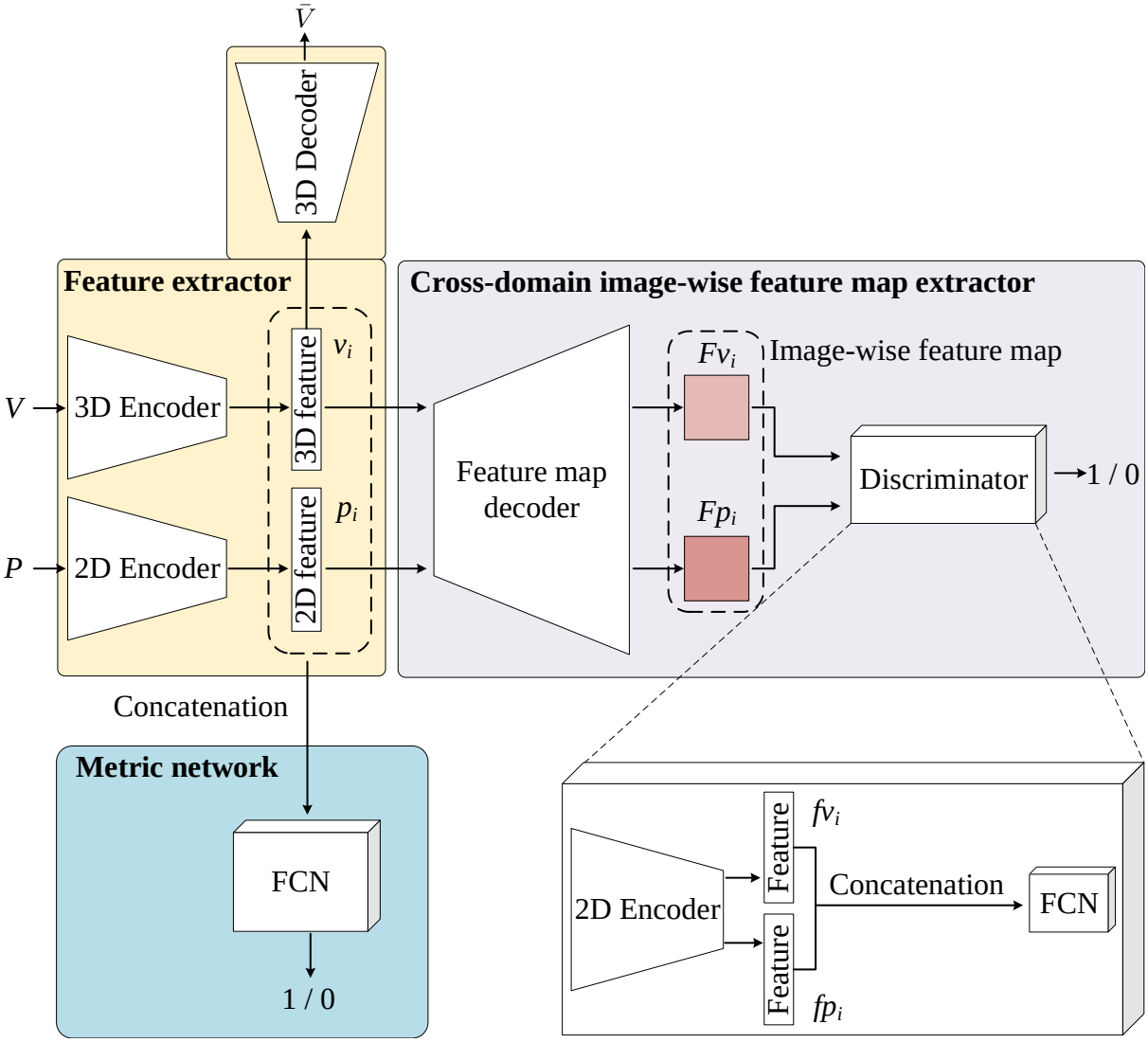


Figure 1. The network architecture of 2D3D-DescNet, which contains the feature extractor, the cross-domain image-wise feature map extractor, and the metric network. The goal of the proposed 2D3D-DescNet is to extract the 2D and 3D local feature descriptors with consistent similarity from the 2D images and the 3D point cloud.

The pipeline of the GAN strategy and the consistent loss function are as follows:

Step 1: Input 3D point cloud volume V and 2D image patch P. The matching and non-matching 2D and 3D pairs are constructed by the hard triplet margin loss $L_{HardSOS}$ (Equation (2)).

Step 2: Use the 3D encoder and 2D encoder (feature extractor module) to extract the 3D feature v_i and 2D feature p_i.

Step 3: Use the 3D decoder to reconstruct the 3D feature v_i as $\bar{V}$. The loss function is chamfer loss $L_{Chamfer}$ (Equation (1)).

Step 4: Concatenate 3D feature v_i and 2D feature p_i and then send it into the metric network to compute the similarity of the 3D point cloud volume V and the 2D image patch P. The loss function of the metric network is cross-entropy loss L_{CE} (Equation (6)).

Step 5: Use the feature map decoder to reconstruct the 3D feature v_i and 2D feature p_i as an image-wise feature map F_{vi} and F_{pi}.

Step 6: Use the discriminator to compute the similarity of the image-wise feature map F_{vi} and F_{pi}.

Step 7: Compute the adversarial loss, which consists of the discriminator L_D (Equation (4)) and the generator L_G (Equation (5)).

Step 8: Finally, obtain the cross-dimensional 2D and 3D feature descriptors f_{vi} and f_{pi}.

3.1.1. Feature Extractor

The feature extractor contains a 2D encoder and a 3D autoencoder to learn the 128-dimensional 2D and 3D feature descriptors between the input paired matching 2D image patches and the 3D point cloud volumes. The size of the input 2D image patches is $64 \times 64 \times 3$, and the input 3D point cloud volumes are sampled to 1024 points with coordinate information and color information, i.e., 1024×6.

For the 2D encoder, the detailed structure is 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(256,128,4), where C(n,k,s) denotes a convolution layer with n filters of kernel size k $\times$ k with stride s, BN is batch normalization, and ReLU is the non-linear activate function.

As for the 3D autoencoder, the 3D encoder is PointNet [47], and the last fully connected layer is changed to a 128-dimensional layer. The detailed structure of the 3D decoder is FC(128,1024)-BN-ReLU-FC(1024,1024)-BN-ReLU-FC(1024,1024$\times$6)-Tanh, where FC(p,q) represents the input p-dimensional feature vector map to the q-dimensional feature vector through a fully connected network, and Tanh is the non-linear activate function.

3.1.2. Cross-Domain Image-Wise Feature Map Extractor

To overcome the dimension differences between 2D images and 3D point clouds, we proposed to map the 2D images and 3D point clouds to a unified two-dimensional feature space, called cross-domain image-wise feature maps. (1) The 2D and 3D feature descriptors are learned by the feature extractor. (2) The cross-domain image-wise feature maps are generated by a constructed shared decoder, named the feature map decoder, as shown in Figure 1. Essentially, there are two steps here: first, the 2D image passes through a 2D encoder to extract features, then a 2D decoder is performed to obtain a 2D feature map; second, the 3D point cloud passess through a 3D encoder to extract features, then a 3D decoder is performed to obtain a 3D feature map. The 2D feature map and 3D feature map are collectively referred to as the cross-domain image-wise feature map. (3) The GAN is embedded into the 2D3D-DescNet, and the discriminator distinguishes the source of the image-wise feature maps. Thus, during training, the paired matching image-wise feature maps of the paired matching 2D image patches and 3D point cloud volumes will become similar such that through the shared feature map decoder, we can conclude that the 2D and 3D feature descriptors learned by 2D3D-DescNet are similar.

Three-dimensional point clouds encapsulate a wealth of intricate two-dimensional information, making them a valuable resource for feature extraction. Through the use of deep neural networks, 2D images can facilitate the transfer of these features, enhancing our overall understanding and analysis of the data. Specifically, the image-wise feature map derived from 3D data is capable of extracting 2D features that correspond closely to those obtained from 2D data. This ensures that the image-wise feature maps generated from both 3D and 2D datasets encapsulate the same feature information, promoting consistency and accuracy in data representation.

In the process, 2D image-wise feature maps are generated from 2D image patches. These patches undergo processing by a shared feature map decoder, which uses 2D feature descriptors as inputs. This decoder effectively translates the 2D characteristics into a

comprehensive feature map. Similarly, 3D image-wise feature maps are extracted from the volumetric data of 3D point clouds. The shared feature map decoder, utilizing 3D feature descriptors as inputs, processes these volumetric data points to produce a detailed and accurate feature map. Thus, the use of a shared feature map decoder for both 2D and 3D data ensures that the resulting image-wise feature maps are consistent and comparable, facilitating robust feature transfer and integration across different data types.

In detail, the cross-domain image-wise feature map extractor contains a shared feature map decoder and a discriminator, as follows:

For the feature map decoder, the inputs are the paired 2D and 3D feature descriptors learned from the feature extractor, and the outputs are image-wise feature maps, for which the size is $64 \times 64 \times 3$. The detailed structure is C(256,4,4)-BN-ReLU-C(128,4,2)-BN-ReLU-C(64,4,2)-BN-ReLU-C(32,4,2)-Sgimoid, where Sgimoid is the non-linear activate function.

The discriminator, consisting of a 2D encoder and a fully connected network (FCN), receives the extracted image-wise feature maps to distinguish which domain data the extracted image-wise feature maps come from. In detail, the structure of 2D encoder is the same as the 2D encoder of feature extractor, and the detailed structure of FCN is FC(128,256)-ReLU-BN-Dropout-FC(256,128)-ReLU-BN-Dropout-FC(128,64)-ReLU-BN-Dropout-FC(64,32)-ReLU-BN-Dropout-FC(32,2)-SoftMax.

3.1.3. Metric Network

The metric network receives the concatenation of the 2D and 3D feature descriptors learned by the feature extractor and computes their similarity. Specifically, the structure of metric network is the same as the FCN of the discriminator in the cross-domain image-wise feature map extractor, and the non-linear activation function is Softmax. The detailed structure of the metric network is as follows: FC(128,256)-ReLU-BN-Dropout-FC(256,128)-ReLU-BN-Dropout-FC(128,64)-ReLU-BN-Dropout-FC(64,32)-ReLU-BN-Dropout-FC(32,2)-SoftMax.

3.2. 2D-3D Consistent Loss Function

To optimize the 2D3D-DesNet, we construct a 2D-3D consistent loss function, which consists of the chamfer, hard triplet margin, adversarial and cross-entropy losses.

3.3. Chamfer Loss

To optimize the 3D autoencoder and keep the structure information of the learned 3D features, the 3D decoder is used for reconstructing the point cloud from the learned 3D feature descriptors. In detail, we measure point sets via chamfer distance to optimize the 3D autoencoder:

$$L_{\text{Chamfer}} = \max\left\{ \begin{array}{l} \frac{1}{|V|} \sum_{v \in V} \min_{q \in \bar{V}} \|v - q\|_2 \\ \frac{1}{|\bar{V}|} \sum_{q \in \bar{V}} \min_{v \in V} \|v - q\|_2 \end{array} \right\}, \tag{1}$$

where V and $\bar{V}$ are the input point cloud volumes and output reconstructed point cloud volumes with 1024 points, respectively.

3.3.1. Hard Triplet Margin Loss

To narrow the domain gap between the 2D images and 3D point clouds, we aim to make the learned 2D and 3D feature descriptors as close as possible in the Euclidean space so that they can be used for retrieval. Inspired by HardNet [48] and SOSNet [15], we use the hard triplet margin loss with a second-order similarity (SOS) regularization to constrain the 2D and 3D feature descriptors learned by 2D3D-DescNet.

Essentially, the hard triplet loss enables the matching 2D feature descriptors and 3D feature descriptors to become similar in high-dimensional space, while the non-matching pairs become alienated. Moreover, the hard triplet loss solves the problem of unstable performance caused by the randomness of negative sample sampling. In addition, the second-order similarity regularization can be used to expand the inter-class distance be-

tween descriptors of the same domain, thus improving the invariance of the jointly learned 2D and 3D local feature descriptors.

In detail, for a batch of the training data $B = (P_i, V_i)_{i=1}^{n}$, where P and V are matching 2D image patches and 3D point cloud volumes, n is the number of samples in B. Then, through the 2D3D-DescNet, the matching 2D and 3D local feature descriptors (p_i, v_i) are computed; meanwhile, the closest non-matching 2D and 3D local descriptors, namely, (p_i, v_{jmin}) and $(p_{k\min}, v_i)$, are constructed. Finally, the hard triplet margin loss with an SOS regularization is defined as follows:

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

where $d(x, y) = \sqrt{2 - 2xy}$, and x and y denote the D-dimensional feature descriptors.

3.3.2. Adversarial Loss

The 2D3D-DescNet is optimized using the mini-max two-player game framework, a foundational concept in generative adversarial networks (GANs), which is composed of two neural networks: the generator and the discriminator. These networks are trained simultaneously in a competitive setting. The generator's goal is to produce data that is as realistic as possible, while the discriminator's task is to distinguish between real data and data generated by the generator. This adversarial process drives both networks to improve continuously, leading to the generation of highly realistic data.

In the context of 2D3D-DescNet, the generator and discriminator work together to align 2D image patches with 3D point cloud volumes. The binary cross entropy (BCE) loss is used as the loss function to judge the performance of both networks. BCE is particularly suited for binary classification tasks, where the output is a probability value between 0 and 1. It measures the performance of the generator in producing realistic data and the discriminator's ability to distinguish between real and generated data. Although MSE is easy to calculate, it is very sensitive to outliers, which means that if there are outliers in the data, MSE may be dominated by these values and ignore other data points, resulting in an increase in the loss value, which affects the training of the model. In summary, in the context of this paper, if MSE loss is used, it only makes a simple pixel value comparison and cannot determine the similarity of 2D and 3D image-wise feature maps from a global perspective. We use the discriminator method to solve this problem effectively. Thus, binary cross entropy (BCE) is used for judging the generation ability of the generator and the discriminatory ability of the discriminator. The 2D and 3D image-wise feature maps are evaluated by a discriminator instead of mean square error, as follows:

$$L_D = L_{BCE}\big(D_\beta(G_\theta(P), 1)\big) + L_{BCE}\big(D_\beta(G_\theta(V), 0)\big), \tag{3}$$

where P and V are the inputs of the paired matching 2D image patches and 3D point cloud volumes; G and D are the generator and the discriminator, respectively; β denotes parameters of D, and θ denotes parameters of network framework, except D; label 1 denotes the image-wise feature of V; label 0 denotes the image-wise feature of P.

In this setting, the generator aims to create feature maps from 2D images that are so realistic that the discriminator cannot distinguish them from feature maps derived from 3D point clouds. The discriminator, on the other hand, strives to accurately identify whether the feature maps are from the 2D images or the 3D point clouds. This adversarial relationship pushes the generator to produce increasingly realistic feature maps, improving its ability to generate data that mimic the real 3D features.

The generator in 2D3D-DescNet also functions as a feature extractor. It includes a feature extractor, a feature map decoder, and a 2D encoder in the discriminator, as illustrated in Figure 1. The optimization objective for the generator is to generate feature maps that can deceive the discriminator. The loss function for the generator is as follows:

$$L_G = L_{BCE}\big(D_\beta(G_\theta(P), 0)\big) + L_{BCE}\big(D_\beta(G_\theta(V), 1)\big) + d(f_{vi}, f_{pi}), \tag{4}$$

where f_{vi} and f_{pi} are the 128-dimensional features output by the 2D encoder in the discriminator D; and the definition of $d(f_{vi}, f_{pi})$ is the same to the hard triplet margin loss. The triplet margin loss ensures that the features extracted from matching 2D and 3D data are closely aligned, while features from non-matching pairs are pushed apart. This helps in learning discriminative features for accurate matching.

The hard triplet margin loss $d(f_{vi}, f_{pi})$ is defined as follows:

$$d(f_{vi}, f_{pi}) = \frac{1}{n}\sum_{i=1}^{n} \max\{0, 0.25 + d(f_{vi}, f_{pi}) - \min[d(f_{vi}, f_{pj_{\min}}), d(f_{vk_{\min}}, f_{pi})]\}. \tag{5}$$

This loss function encourages the network to minimize the distance between the features and matching pairs while maximizing the distance from the hardest negative samples. By incorporating this loss, the generator is encouraged to produce feature maps that are not only realistic but also well-aligned with their corresponding 3D features.

To summarize, 2D3D-DescNet leverages the adversarial training strategy of GANs to optimize the feature extraction process for matching 2D images and 3D point clouds. The use of BCE loss for both the generator and discriminator ensures that the network can effectively distinguish between real and generated data. The additional triplet margin loss helps in aligning features from 2D and 3D data, improving the discriminative power of the learned features. This combination of adversarial training and feature alignment techniques allows 2D3D-DescNet to achieve robust and accurate matching between 2D images and 3D point clouds, making it a powerful tool for tasks that require cross-dimensional data integration.

3.3.3. Cross-Entropy Loss

For the embedded metric network, metric learning is used for measuring similarity between the learned local 2D and 3D feature descriptors. Specifically, since the input samples of 2D3D-DescNet are all matching 2D image patches and 3D point cloud volumes, we introduce a strategy by which to construct the positive and negative samples for the training of the metric network. For a mini-batch, when fixing a batch of 2D feature descriptors, first, the corresponding matching batch of 3D feature descriptors is constructed to provide positive samples; second, the non-matching batch of 3D feature descriptors is constructed to provide negative samples, which are sampled from the other's mini-batch. Based on this strategy, the same number of positive and negative samples are constructed for metric learning.

The cross-entropy loss is used for optimizing the metric network. In detail, n pairs of 2D-3D samples are fed into the network in one batch. y_i is the 0/1 label that indicates whether the input pair x_i, which is the concatenation of the 3D feature descriptor v_i and the 2D feature descriptor p_i, is matching or not. Label 1 means matching and label 0 means non-matching. $\widehat{y}_i$ is the output of metric network. The cross-entropy loss is defined according to the following formula:

$$L_{CE} = -\frac{1}{n}\sum_{i=1}^{n}[y_i \log(\widehat{y}_i) + (1 - y_i)\log(1 - \widehat{y}_i)]. \tag{6}$$

For the input x_i, two-dimensional vector $\begin{pmatrix} g_0(x_i) \\ g_1(x_i) \end{pmatrix}$ is computed as the similarity of the input pair x_i:

$$\widehat{y}_i = \frac{e^{g_1(x_i)}}{e^{g_0(x_i)} + e^{g_1(x_i)}}. \tag{7}$$

3.3.4. Total Loss

Finally, the 2D3D-DescNet is optimized according to the constructed 2D-3D consistent loss function, which has been described in detail above.

In detail, the 2D-3D consistent loss function is further divided into two categories during optimization. First, for 2D and 3D local feature descriptor learning, we include chamfer loss, hard triplet margin loss with SOS, cross-entropy loss, and the generator of adversarial loss; thus, these 4 losses are optimized together, as follows:

$$L_1 = \lambda_1 L_{\text{Chamfer}} + \lambda_2 L_{HardSOS} + \lambda_3 L_{CE} + \lambda_4 L_G, \tag{8}$$

where λ_1, λ_2, λ_3, and λ_4 denote the proportion of the loss functions and have been set as 1, 1, 0.5, and 0.25, respectively, in our experiments. Second, to ensure the performance of the discriminator, the parameters of the discriminator are updated in 5 steps, and other parameters of the 2D3D-DescNet are updated every step; thus, the discriminator is optimized individually, as follows:

$$L_2 = L_D. \tag{9}$$

3.4. Training Strategy

2D3D-DescNet uses the Pytorch framework to build on a cluster equipped with a NVIDIA 3090Ti and 36 GB of memory. The training process is stopped after 200 epochs. In our framework, all parameters of the network are trained with the same implementation details. The optimizer SGD is used as the optimizer, and the momentum is set as 0.9. Initially, the learning rates of the generator and discriminator are set at 0.001 and 0.001, respectively, and the decrease by 0.0005 after each step.

4. Experiments

In this section, we first introduce the 2D image patch and 3D point cloud volume dataset used in this paper. Second, we demonstrate state-of-the-art retrieval performance of the 2D and 3D local feature descriptors learned by 2D3D-DescNet. Third, we show the performance of the metric learning in 2D3D-DescNet. Fourth, we show several point cloud registration results by using the 3D feature descriptors learned from the 2D3D-DescNet. Finally, we conduct ablation studies with analyses and discussions.

4.1. Dataset

The matching 2D image patch and 3D point cloud volume dataset used in this paper is generated from the 3DMatch [49] dataset, which contains 600,000 pairs of 2D-3D correspondences. For 2D3D-DescNet and all comparative networks, 580,000 correspondence pairs are used for training, and 20,000 pairs are used to evaluate the performance of the learned 2D and 3D feature descriptors. All the training data and testing data do not intersect with each other.

4.2. Similarity of 2D and 3D Feature Descriptors

4.2.1. 2D-3D Retrieval

To demonstrate the similarity of the 2D and 3D local feature descriptors learned by the proposed 2D3D-DescNet, we use the TOP1 and TOP5 retrieval accuracy on the 2D-3D retrieval task. In detail, each 2D feature descriptor retrieves the nearest and the 5-nearest neighbors in the 3D feature descriptor. A successful TOP1 retrieval is defined as the nearest retrieval result of the 2D feature descriptor which contains its corresponding 3D feature

descriptor, and a successful TOP5 retrieval is defined as the 5-nearest retrieval result of the 2D feature descriptor which contains its corresponding 3D feature descriptor.

The retrieval performance of 2D3D-DescNet is compared with the existing state-of-the-art networks based on jointly learning 2D and 3D feature descriptors, namely, 2D3D-MatchNet [40], Siam2D3D-Net [41], 2D3D-GAN-Net [44], LCD [43], HAS-Net [45], and 2D3D-MVPNet [46], which are the state-of-art methods for jointly learning 2D and 3D feature descriptors. As shown in Table 1, compared with the comparative networks, the 2D and 3D local feature descriptors learned by 2D3D-DescNet demonstrate significant improvement in TOP1 and TOP5 retrieval accuracy and achieve state-of-the-art performance.

Table 1. TOP1 and TOP5 retrieval accuracy on 3DMatch dataset.

Method	TOP1	TOP5
2D3D-MatchNet [40]	0.2097	0.4318
Siam2D3D-Net [41]	0.2123	0.4567
2D3D-GAN-Net [44]	0.5842	0.8811
LCD [43]	0.7174	0.9412
HAS-Net [45]	0.7565	0.9623
2D3D-MVPNet [46]	0.8011	0.9482
2D3D-DescNet (Ours)	0.9271	0.9916

In addition, to show the effect of 2D-3D retrieval, based on the learned local 2D and 3D feature descriptors, we use the 2D image patches as the queries to retrieve the 5-nearest retrieval 3D point cloud volumes in the testing data. The TOP5 ranking retrieval results are shown in Figure 2, the correct TOP1 retrievals are labeled with red bounding boxes. It can be seen that all the retrieved 3D point cloud volumes have similar colors and structure information to the queried 2D image patches, which demonstrates that the local 2D and 3D feature descriptors learned by 2D3D-DescNet are similar.

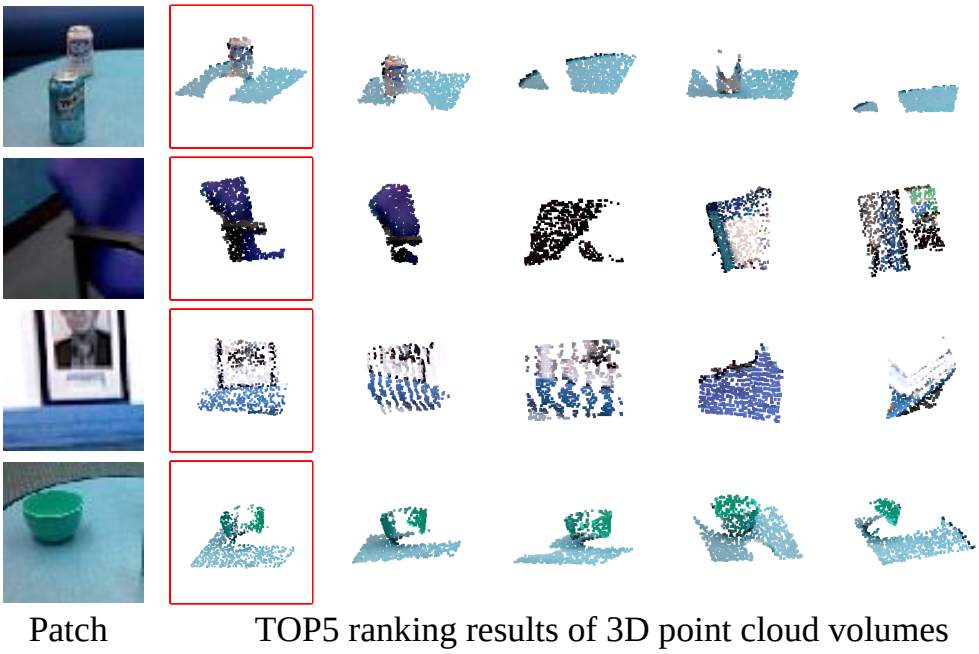
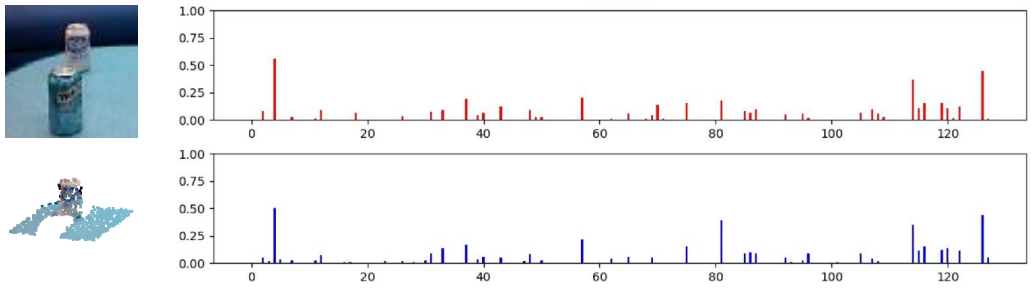


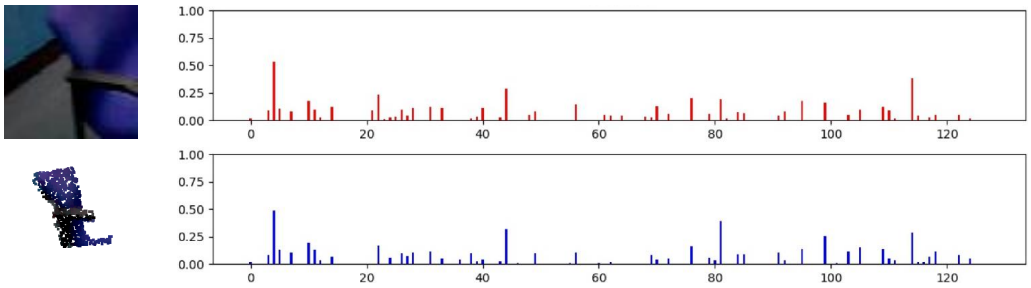
Figure 2. The TOP5 2D-3D retrieval results on the 3Dmatch dataset using the 2D and 3D local feature descriptors learned by 2D3D-DescNet. The correct TOP1 retrievals are labeled with red bounding boxes. It can be seen that all the retrieved 3D point cloud volumes have similar colors and structure information to the queried 2D image patches, which demonstrates that the local 2D and 3D feature descriptors learned by 2D3D-DescNet are similar.

4.2.2. Histogram Visualization of 2D & 3D Feature Descriptors

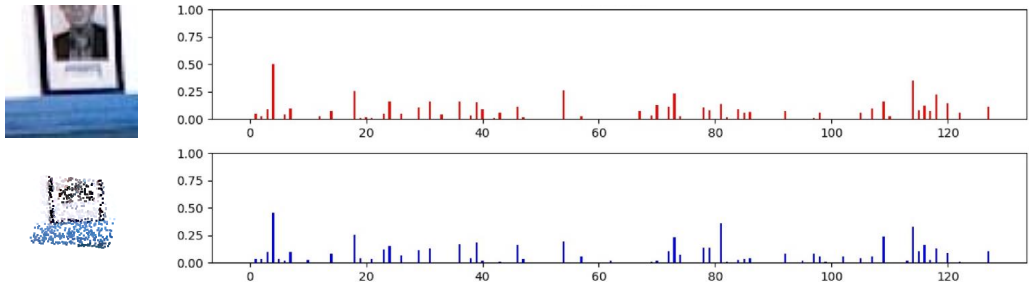
To further demonstrate that the 2D and 3D local feature descriptors learned by 2D3D-DescNet are similar, for the matching 2D image patches and 3D point cloud volumes in Figure 2, we visualize the histogram of the learned 2D and 3D feature descriptors, as shown in Figure 3. It can be observed that the distributions of the histogram in the figure are extremely similar, and the values of each dimension are also similar. Thus, the histogram visualization in Figure 3 demonstrates that the 2D and 3D local feature descriptors learned by 2D3D-DescNet are robust and similar.



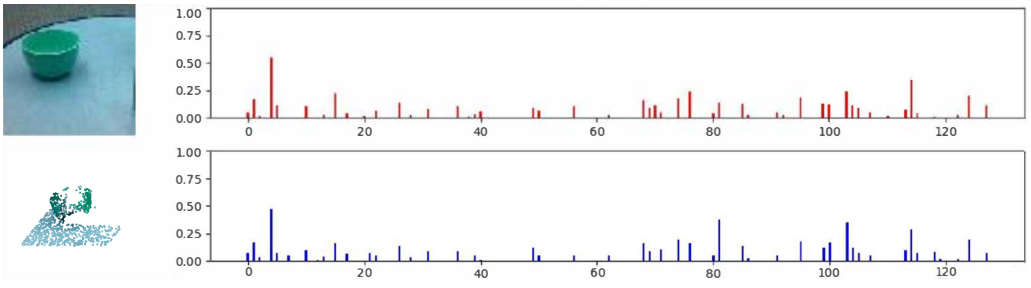
(**a**) 2D-3D pairs 1



(**b**) 2D-3D pairs 2



(**c**) 2D-3D pairs 3



(**d**) 2D-3D pairs 4

Figure 3. The histogram visualization of the 2D and 3D local feature descriptors learned by 2D3D-DescNet between the matching 2D image patches and the 3D point cloud volumes in Figure 2. It can be observed that the distributions of the histogram in the figure are extremely similar, and the values of each dimension are also similar.

4.3. 2D-3D Matching

The 2D-3D matching is implemented by the metric network, which is designed to distinguish between matching pairs and non-matching pairs. On the basis of the sampling strategy working on testing dataset, we generate 20,000 pairs of matching samples and 20,000 pairs of non-matching samples. The 2D-3D matching results are evaluated by the false positive rate at 95% recall (FPR95) and precision. Specifically, for FPR95, the lower the better, and higer precision means better performance. We report the FPR95 and precision for the evaluation of 2D-3D matching in Table 2. The high precision and low FPR95 show the excellent performance of the 2D-3D matching metric network of 2D3D-DescNet in measuring similarity.

Table 2. Evaluation of 2D-3D matching on 3DMatch dataset.

Method	FPR95	Precision
2D3D-DescNet	0.1891	99.605

In fact, the learned 2D and 3D feature descriptors can be used for both TOP1 retrieval and the measurements of the metric network. Obviously, the 2D-3D matching by the metric network (Table 1) demonstrates better performance than the 2D-3D retrieval by TOP1 retrieval (Table 2), but this comes at the cost of very expensive computation during the process of measuring the similarity of cross-domain descriptors because the network architecture of the metric network is FCN, which has high computational complexity. Such a high testing cost prevents metric-network-based methods from being employed in many applications despite their excellent performance. Thus, in practice, using the retrieved 2D and 3D local feature descriptors is more reasonable.

4.4. 3D Point Cloud Registration

Following the 3DMatch Benchmark, we have generated color 3D point cloud fragments for several scenes. These color 3D point cloud fragments are detected and sampled to generate uniform keypoints. Each feature point generates a 3D point cloud volume with a radius of 30cm, which are fed into 3D encoder of 2D3D-DescNet to extract 3D local feature descriptors. It should be noted that the 30 cm radius setting is based on the experience of previous point cloud registration research [50–53]. Finally, pairs of 3D point cloud fragments are matched through a nearest neighbor search of 3D local feature descriptors and RANSAC.

Then, five pairs of 3D point cloud fragments are input to the registration process, and the visualizations are shown in Figure 4. Registration results show that the 3D feature descriptors learned from 2D3D-DescNet not only have excellent retrieval performance but also demonstrate robustness in 3D point cloud registration.

In addition, based on the 3DMatch dataset, we calculate the point cloud registration recall of eight scenes and compare them with the existing point cloud registration methods, as shown in Table 3. The experimental results show that the performance of the 3D point cloud feature description extracted by 2D3D-DescNet is better in the registration task than some traditional methods, but there is still a certain gap compared with other deep learning methods. The reason is that the multi-task attribute of the cross-dimensional feature description makes it impossible to focus entirely on a specific task.

Table 3. The performance of the 2D and 3D local feature descriptors learned by 2D3D-DescNet with different dimensions on the 3DMatch dataset.

	2D3D-DescNet	CZK [54]	FGR [55]	3DMatch [49]	3DSmoothNet [56]	PointNetAE [43]	LCD [43]
Kitchen	0.781	0.499	0.305	0.853	0.871	0.766	0.891
Home1	0.613	0.632	0.434	0.783	0.896	0.726	0.783
Home2	0.528	0.403	0.283	0.610	0.723	0.579	0.629
Hotel1	0.714	0.643	0.401	0.786	0.791	0.786	0.808

Table 3. *Cont.*

	2D3D-DescNet	CZK [54]	FGR [55]	3DMatch [49]	3DSmoothNet [56]	PointNetAE [43]	LCD [43]
Hotel2	0.589	0.667	0.426	0.590	0.846	0.680	0.769
Hotel3	0.423	0.577	0.385	0.577	0.731	0.731	0.654
Study	0.572	0.547	0.291	0.633	0.556	0.641	0.662
MIT Lab	0.444	0.378	0.200	0.511	0.467	0.511	0.600
Average	0.583	0.543	0.342	0.688	0.735	0.677	0.725

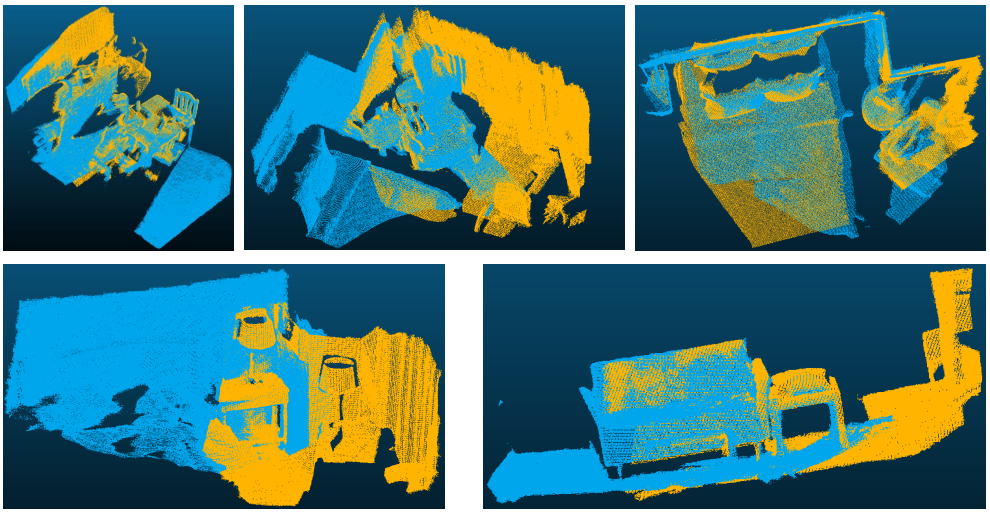


Figure 4. Three-dimensional global registration results based the three-dimensional feature descriptors learned by the proposed 2D3D-DescNet. Five pairs of 3D point cloud fragments are input to the registration process, and registration results show that the 3D feature descriptors learned by 2D3D-DescNet are robust in 3D point cloud registration. The blue point clouds are source data and the yellow point clouds are target data.

4.5. Outdoor 2D-3D Retrieval

In order to explore the reliability of our proposed method in extracting 2D and 3D cross-dimensional feature descriptors, we supplemented it with additional validation on an outdoor 2D-3D dataset. This outdoor 2D-3D dataset is obtained via a mobile LiDAR system and includes street view images and corresponding 3D LiDAR point clouds, with a total of 13,884 pairs of matching 2D image patches and 3D point cloud volumes [41]. We use 12,884 matching 2D and 3D pairs as training data and set 1000 matching 2D and 3D pairs as testing data. The experimental result is shown in Table 4. This also verifies the effectiveness of our proposed 2D3D-DescNet in extracting 2D and 3D cross-dimensional feature descriptors on an outdoor 2D-3D dataset.

Table 4. TOP1 and TOP5 retrieval accuracy on outdoor 2D-3D dataset [41].

Method	TOP1	TOP5
2D3D-MatchNet [40]	0.0081	0.0449
Siam2D3D-Net [41]	0.0084	0.0365
2D3D-GAN-Net [44]	0.0101	0.0489
LCD [43]	0.0343	0.0698
HAS-Net [45]	0.1782	0.2393
2D3D-MVPNet [46]	0.2588	0.3615
2D3D-DescNet (Ours)	0.6971	0.7563

4.6. Ablation Study

To demonstrate the superiority of the proposed 2D3D-DescNet, we conduct several ablation studies with analyses and discussions.

4.6.1. Dimensions of 2D and 3D Feature Descriptor

To explore the impact of the learned local 2D and 3D feature descriptors of different dimensions on the retrieval performance of 2D3D-DescNet, we set the embedding size of the 2D and 3D feature descriptors to 64, 128, and 256 for comparison experiments; the results are shown in Table 5. For the 2D-3D retrieval task, the 128-dimensional and 256-dimensional feature descriptors achieved the best performance in TOP1 and TOP5 retrieval accuracy, respectively. For the 2D-3D matching task, the 128-dimensional and 256-dimensional feature descriptors achieved the best performance in FPR95 and precision, respectively.

Table 5. The performance of 2D and 3D local feature descriptors learned by 2D3D-DescNet with different dimension on 3DMatch dataset.

Dimension	TOP1	TOP5	FPR95	Precision
64	0.8813	0.9844	0.2206	99.605
128	0.9271	0.9916	0.1891	99.605
256	0.9221	0.9922	0.2153	99.610

4.6.2. Cross-Domain Image-Wise Feature Map

To explore the role of image-wise feature mapping in 2D3D-DescNet, we compare the 2D3D-DescNet with the 2D3D-DescNet that removed the cross-domain image-wise feature map extractor (denoted as w/o image-wise feature map); the results are shown in Table 6. It can be seen that the 2D3D-DescNet with the cross-domain image-wise feature map extractor demonstrates better performance in 2D-3D retrieval, while the 2D3D-DescNet without the cross-domain image-wise feature map extractor demonstrates better performance in 2D-3D matching.

Table 6. The effect of the image-wise feature map extractor, metric network, and adversarial loss in 2D3D-DescNet on the 3DMatch dataset.

	TOP1	TOP5	FPR95	Precision
2D3D-DescNet	0.9271	0.9916	0.1891	99.605
2D3D-DescNet w/o image-wise feature map	0.9169	0.9900	0.1839	99.660
2D3D-DescNet w/o metric network	0.8873	0.9512	0.1855	99.391
2D3D-DescNet w/o adversarial loss	0.8452	0.9374	0.1872	98.824

4.6.3. Metric Network and Adversarial Loss

We also test whether there is a metric network in the 2D3D-DescNet and remove the adversarial loss in the 2D3D-DescNet. The results are shown in Table 6, which shows that the metric network module and adversarial loss in our proposed 2D3D-DescNet are positive factors in improving the robustness of cross-dimensional 2D and 3D local feature descriptors.

5. Conclusions

In this paper, we propose a novel network, 2D3D-DescNet, to learn the 2D and 3D local feature descriptors jointly, and the network is built based on a feature extractor and an embedded GAN strategy. The constructed 2D-3D consistent loss balances the 2D and 3D local feature descriptors between 2D and 3D domains and bridges the domain gap between 2D images and 3D point clouds. We demonstrate that the jointly learned 2D and 3D local

feature descriptors are effective in the 2D-3D retrieval and 2D-3D matching tasks, achieving state-of-the-art performance. In addition, we also demonstrate that the jointly learned 2D and 3D local feature descriptors are similar in terms of 2D-3D retrieval and feature histogram visualization. In future work, we aim to integrate our 2D and 3D local feature descriptors with keypoint detectors to develop full 2D image and 3D point cloud matching.

Author Contributions: Conceptualization, S.C., Y.S., B.L. and W.L.; methodology, S.C., B.L., Y.S. and W.L.; software, S.C., C.H., L.L. and X.Q.; validation, L.C., C.H., X.Q., L.L. and H.J.; formal analysis, S.C., B.L., L.C. and Y.S.; investigation, S.C., C.H. and L.L.; resources, S.C. and W.L.; data curation, S.C. and X.Q.; writing—original draft preparation, S.C., B.L., L.C. and Y.S.; writing—review and editing, S.C., L.L., H.J., L.C. and W.L.; visualization, S.C., C.H. and L.L.; supervision, Y.S., H.J. and W.L.; project administration, S.C., Y.S. and H.J.; funding acquisition, S.C., L.L., X.Q. and W.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Educational Project Foundation of Young and Middle-aged Teachers of Fujian Province, China, under grant numbers JAT210672, JAT220531, and JAT201032, and the China Postdoctoral Science Foundation under grant number 2021M690094.

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Liu, W.; Wang, C.; Chen, S.; Bian, X.; Lai, B.; Shen, X.; Cheng, M.; Lai, S.H.; Weng, D.; Li, J. Y-Net: Learning Domain Robust Feature Representation for Ground Camera Image and Large-scale Image-based Point Cloud Registration. *Inf. Sci.* **2021**, *581*, 655–677. [CrossRef]
2. Nadeem, U.; Bennamoun, M.; Togneri, R.; Sohel, F.; Rekavandi, A.M.; Boussaid, F. Cross domain 2D-3D descriptor matching for unconstrained 6-DOF pose estimation. *Pattern Recognit.* **2023**, *142*, 109655. [CrossRef]
3. Shi, C.; Chen, X.; Lu, H.; Deng, W.; Xiao, J.; Dai, B. RDMNet: Reliable Dense Matching Based Point Cloud Registration for Autonomous Driving. *IEEE Trans. Intell. Transp. Syst.* **2023**, *24*, 11372–11383. [CrossRef]
4. Chen, L.; Rottensteiner, F.; Heipke, C. Feature detection and description for image matching: From hand-crafted design to deep learning. *Geo-Spat. Inf. Sci.* **2021**, *24*, 58–74. [CrossRef]
5. Lowe, D.G. Distinctive Image Features from Scale-invariant Leypoints. *Int. J. Comput. Vis. (IJCV)* **2004**, *60*, 91–110. [CrossRef]
6. Bay, H.; Tuytelaars, T.; Van Gool, L. Surf: Speeded up Robust Features. In Proceedings of the European Conference on Computer Vision (ECCV), Graz, Austria, 7–13 May 2006; pp. 404–417.
7. Rusu, R.B.; Blodow, N.; Beetz, M. Fast Point Feature Histograms (FPFH) for 3D Registration. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Kobe, Japan, 12–17 May 2009; pp. 3212–3217.
8. Tombari, F.; Salti, S.; Di Stefano, L. Unique Signatures of Histograms for Local Surface Description. In Proceedings of the European Conference on Computer Vision (ECCV), Heraklion, Greece, 5–11 September 2010; pp. 356–369.
9. Guo, Y.; Sohel, F.; Bennamoun, M.; Lu, M.; Wan, J. Rotational Projection Statistics for 3D Local Surface Description and Object Recognition. *Int. J. Comput. Vis. (IJCV)* **2013**, *105*, 63–86. [CrossRef]
10. Dhal, P.; Azad, C. A Comprehensive Survey on Feature Selection in the Various Fields of Machine Learning. *Appl. Intell.* **2022**, *52*, 4543–4581. [CrossRef]
11. Bello, S.A.; Yu, S.; Wang, C.; Adam, J.M.; Li, J. Deep Learning on 3D Point Clouds. *Remote. Sens.* **2020**, *12*, 1729. [CrossRef]
12. Dubey, S.R. A Decade Survey of Content Based Image Retrieval Using Deep Learning. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *32*, 2687–2704. [CrossRef]
13. Simo-Serra, E.; Trulls, E.; Ferraz, L.; Kokkinos, I.; Fua, P.; Moreno-Noguer, F. Discriminative Learning of Deep Convolutional Feature Point Descriptors. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 118–126.
14. Tian, Y.; Fan, B.; Wu, F. L2-net: Deep Learning of Discriminative Patch Descriptor in Euclidean Space. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 661–669.
15. Tian, Y.; Yu, X.; Fan, B.; Wu, F.; Heijnen, H.; Balntas, V. Sosnet: Second Order Similarity Regularization for Local Descriptor Learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 11016–11025.
16. Tyszkiewicz, M.; Fua, P.; Trulls, E. DISK: Learning Local Features with Policy Gradient. *Adv. Neural Inf. Process. Syst. (NerulIPS)* **2020**, *33*, 14254–14265.
17. Zhang, J.; Jiao, L.; Ma, W.; Liu, F.; Liu, X.; Li, L.; Zhu, H. RDLNet: A Regularized Descriptor Learning Network. *IEEE Trans. Neural Networks Learn. Syst.* **2021**, *34*, 5669–5681. [CrossRef]

18. Lindenberger, P.; Sarlin, P.E.; Pollefeys, M. Lightglue: Local Feature Matching at Light Speed. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Paris, France, 2–3 October 2023; pp. 17627–17638.
19. Qi, C.R.; Yi, L.; Su, H.; Guibas, L.J. Pointnet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space. *Adv. Neural Inf. Process. Syst. (NerulIPS)* **2017**, *30*, 5099–5108.
20. Li, Y.; Bu, R.; Sun, M.; Wu, W.; Di, X.; Chen, B. Pointcnn: Convolution on X-transformed Points. In Proceedings of the Advances in Neural Information Processing Systems (NerulIPS), Montreal, QC, Canada, 3–8 December 2018; pp. 820–830.
21. Deng, H.; Birdal, T.; Ilic, S. Ppfnet: Global Context Aware Local Features for Robust 3D Point Matching. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 195–205.
22. Bai, X.; Luo, Z.; Zhou, L.; Fu, H.; Quan, L.; Tai, C.L. D3feat: Joint Learning of Dense Detection and Description of 3D Local Features. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 6359–6367.
23. Ao, S.; Hu, Q.; Yang, B.; Markham, A.; Guo, Y. Spinnet: Learning a General Surface Descriptor for 3D Point Cloud Registration. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 11753–11762.
24. Qian, G.; Li, Y.; Peng, H.; Mai, J.; Hammoud, H.; Elhoseiny, M.; Ghanem, B. Pointnext: Revisiting Pointnet++ with Improved Training and Scaling Strategies. *Adv. Neural Inf. Process. Syst. (NerulIPS)* **2022**, *35*, 23192–23204.
25. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Nets. *Adv. Neural Inf. Process. Syst. (NerulIPS)* **2014**, *27*, 2672–2680.
26. Ojala, T.; Pietikainen, M.; Maenpaa, T. Multiresolution Gray-scale and Rotation Invariant Texture Classification with Local Binary Patterns. *IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI)* **2002**, *24*, 971–987. [CrossRef]
27. Chen, J.; Kellokumpu, V.; Zhao, G.; Pietikäinen, M. RLBP: Robust Local Binary Pattern. In Proceedings of the British Machine Vision Conference (BMVC), Bristol, UK, 9–13 September 2013.
28. Calonder, M.; Lepetit, V.; Strecha, C.; Fua, P. Brief: Binary Robust Independent Elementary Features. In Proceedings of the European Conference on Computer Vision (ECCV), Heraklion, Greece, 5–11 September 2010; pp. 778–792.
29. Rublee, E.; Rabaud, V.; Konolige, K.; Bradski, G. ORB: An Efficient Alternative to SIFT or SURF. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Barcelona, Spain, 6–13 November 2011; pp. 2564–2571.
30. Wang, Z.; Fan, B.; Wu, F. Local Intensity Order Pattern for Feature Description. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Barcelona, Spain, 6–13 November 2011; pp. 603–610.
31. Wang, Z.; Fan, B.; Wang, G.; Wu, F. Exploring Local and Overall Ordinal Information for Robust Feature Description. *IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI)* **2015**, *38*, 2198–2211. [CrossRef] [PubMed]
32. Guo, Y.; Bennamoun, M.; Sohel, F.; Lu, M.; Wan, J.; Kwok, N.M. A Comprehensive Performance Evaluation of 3D Local Feature Descriptors. *Int. J. Comput. Vis. (IJCV)* **2016**, *116*, 66–89. [CrossRef]
33. Ma, J.; Jiang, X.; Fan, A.; Jiang, J.; Yan, J. Image Matching from Handcrafted to Deep Features: A Survey. *Int. J. Comput. Vis. (IJCV)* **2020**, *129*, 23–79. [CrossRef]
34. Xia, Y.; Xu, Y.; Li, S.; Wang, R.; Du, J.; Cremers, D.; Stilla, U. SOE-Net: A Self-attention and Orientation Encoding Network for Point Cloud Based Place Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 11348–11357.
35. Xia, Y.; Gladkova, M.; Wang, R.; Li, Q.; Stilla, U.; Henriques, J.F.; Cremers, D. CASSPR: Cross Attention Single Scan Place Recognition. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Paris, France, 2–3 October 2023; pp. 8461–8472.
36. Xia, Y.; Shi, L.; Ding, Z.; Henriques, J.F.; Cremers, D. Text2Loc: 3D Point Cloud Localization from Natural Language. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 17–21 June 2024.
37. Georgiou, T.; Liu, Y.; Chen, W.; Lew, M. A Survey of Traditional and Deep Learning-based Feature Descriptors for High Dimensional Data in Computer Vision. *Int. J. Multimed. Inf. Retr.* **2020**, *9*, 135–170. [CrossRef]
38. Jiang, X.; Ma, J.; Xiao, G.; Shao, Z.; Guo, X. A Review of Multimodal Image Matching: Methods and Applications. *Inf. Fusion* **2021**, *73*, 22–71. [CrossRef]
39. Han, X.F.; Feng, Z.A.; Sun, S.J.; Xiao, G.Q. 3D Point Cloud Descriptors: State-of-The-Art. *Artif. Intell. Rev.* **2023**, *56*, 12033–12083. [CrossRef]
40. Feng, M.; Hu, S.; Ang, M.H.; Lee, G.H. 2D3D-Matchnet: Learning to Match Keypoints Across 2D Image and 3D Point Cloud. In Proceedings of the International Conference on Robotics and Automation (ICRA), Montreal, QC, Canada, 20–24 May 2019; pp. 4790–4796.
41. Liu, W.; Lai, B.; Wang, C.; Bian, X.; Yang, W.; Xia, Y.; Lin, X.; Lai, S.H.; Weng, D.; Li, J. Learning to Match 2D Images and 3D LiDAR Point Clouds for Outdoor Augmented Reality. In Proceedings of the IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW), Atlanta, GA, USA, 22–26 March 2020; pp. 655–656.
42. Liu, W.; Shen, X.; Wang, C.; Zhang, Z.; Wen, C.; Li, J. H-Net: Neural Network for Cross-domain Image Patch Matching. In Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), Stockholm, Sweden, 13–19 July 2018; pp. 856–863.

43. Pham, Q.H.; Uy, M.A.; Hua, B.S.; Nguyen, D.T.; Roig, G.; Yeung, S.K. LCD: Learned Cross-Domain Descriptors for 2D-3D Matching. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), New York, NY, USA, 7–12 February 2020; pp. 11856–11864.
44. Liu, W.; Lai, B.; Wang, C.; Bian, X.; Wen, C.; Cheng, M.; Zang, Y.; Xia, Y.; Li, J. Matching 2D Image Patches and 3D Point Cloud Volumes by Learning Local Cross-domain Feature Descriptors. In Proceedings of the IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW), Lisbon, Portugal, 27 March–1 April 2021; pp. 516–517.
45. Lai, B.; Liu, W.; Wang, C.; Bian, X.; Su, Y.; Lin, X.; Yuan, Z.; Shen, S.; Cheng, M. Learning Cross-Domain Descriptors for 2D-3D Matching with Hard Triplet Loss and Spatial Transformer Network. In Proceedings of the Image and Graphics: 11th International Conference (ICIG), Haikou, China, 6–8 August 2021; pp. 15–27.
46. Lai, B.; Liu, W.; Wang, C.; Fan, X.; Lin, Y.; Bian, X.; Wu, S.; Cheng, M.; Li, J. 2D3D-MVPNet: Learning Cross-domain Feature Descriptors for 2D-3D Matching Based on Multi-view Projections of Point Clouds. *Appl. Intell.* **2022**, *52*, 14178–14193. [CrossRef]
47. Qi, C.R.; Su, H.; Mo, K.; Guibas, L.J. Pointnet: Deep Learning on Point Sets for 3D Classification and Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 652–660.
48. Mishchuk, A.; Mishkin, D.; Radenovic, F.; Matas, J. Working Hard to Know Your Neighbor's Margins: Local Descriptor Learning Loss. In Proceedings of the Advances in Neural Information Processing Systems (NerulIPS), Long Beach, CA, USA, 4–9 December 2017; pp. 4826–4837.
49. Zeng, A.; Song, S.; Nießner, M.; Fisher, M.; Xiao, J.; Funkhouser, T. 3Dmatch: Learning Local Geometric Descriptors from RGB-D Reconstructions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1802–1811.
50. Wu, Q.; Shen, Y.; Jiang, H.; Mei, G.; Ding, Y.; Luo, L.; Xie, J.; Yang, J. Graph Matching Optimization Network for Point Cloud Registration. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Detroit, MI, USA, 1–5 October 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 5320–5325.
51. Tamata, K.; Mashita, T. Feature Description with Feature Point Registration Error Using Local and Global Point Cloud Encoders. *IEICE Trans. Inf. Syst.* **2022**, *105*, 134–140. [CrossRef]
52. Bai, X.; Luo, Z.; Zhou, L.; Chen, H.; Li, L.; Hu, Z.; Fu, H.; Tai, C.L. Pointdsc: Robust point cloud registration using deep spatial consistency. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 15859–15869.
53. Ren, Y.; Luo, W.; Tian, X.; Shi, Q. Extract descriptors for point cloud registration by graph clustering attention network. *Electronics* **2022**, *11*, 686. [CrossRef]
54. Choi, S.; Zhou, Q.Y.; Koltun, V. Robust reconstruction of indoor scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 5556–5565.
55. Zhou, Q.Y.; Park, J.; Koltun, V. Fast global registration. In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 11–14 October 2016; Springer: Berlin/Heidelberg, Germany, 2016; pp. 766–782.
56. Gojcic, Z.; Zhou, C.; Wegner, J.D.; Wieser, A. The perfect match: 3d point cloud matching with smoothed densities. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5545–5554.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.