Ground Camera Image and Large-Scale 3-D Image-Based Point Cloud Registration Based on Learning Domain Invariant Feature Descriptors

Weiquan Liu, Baiqi Lai, Cheng Wang, *Senior Member, IEEE*, Guorong Cai, Yanfei Su, Xuesheng Bian, Yongchuan Li, Shuting Chen, and Jonathan Li, *Senior Member, IEEE*

***Abstract*—Multisource data are captured from different sensors or generated with different generation mechanisms. Ground camera images (images taken from ground-based camera) and rendered images (synthesized by the position information from 3-D image-based point cloud) are different-source geospatial data, called cross-domain images. Particularly, in outdoor environments, the registration relationship between the above cross-domain images is available to establish the spatial relationship between 2-D and 3-D space, which is an indirect solution for virtual–real registration of augmented reality (AR). However, the traditional handcrafted feature descriptors cannot match the above cross-domain images because of the low quality of rendered images and the domain gap between cross-domain images. In this article, inspired by the success achieved by deep learning in computer vision, we first propose an end-to-end network, DIFD-Net, to learn domain invariant feature descriptors (DIFDs) for cross-domain image patches. The DIFDs are used for cross-domain image patch retrieval to the registration of ground camera and rendered images. Second, we construct a domain-kept consistent loss function, which balances the feature descriptors for narrowing the gap in different domains, to optimize DIFD-Net. Specially, the negative samples are generated from positive during training, and the introduced constraint of intermediate feature maps increases extra supervision information to learn feature descriptors. Finally, experiments show the superiority of DIFDs for the retrieval of cross-domain image patches, which achieves state-of-the-art retrieval performance. Additionally, we use DIFDs to match ground camera images and rendered images, and verify the feasibility of the derived AR virtual–real registration in open outdoor environments.**

Manuscript received July 30, 2020; revised October 5, 2020; accepted October 25, 2020. Date of publication November 3, 2020; date of current version January 6, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant U1605254, Grant 41971424, and Grant 41871380, in part by the Key Technical Project of Fujian Province under Grant 2017H6015, and in part by the Projects of Educational Project Foundation of Young and Middleaged Teacher of Fujian Province under Grant JT180884. *(W. Liu and B. Lai contributed equally to this work.) (Corresponding author: Cheng Wang.)*

Weiquan Liu, Baiqi Lai, Cheng Wang, Yanfei Su, Xuesheng Bian, and Yongchuan Li are with the Fujian Key Laboratory of Sensing and Computing for Smart City, School of Informatics,Xiamen University, Xiamen 361005, China (e-mail: wqliu@stu.xmu.edu.cn; laibaiqi@stu.xmu.edu.cn; cwang@xmu.edu.cn; suyanfei@stu.xmu.edu.cn; xbc0809@gmail.com; liyongchuan@xmu.edu.cn).

Guorong Cai is with Computer Engineering College, Jimei University, Xiamen 361021, China (e-mail: guorongcai.jmu@gmail.com).

Shuting Chen is with Chengyi University College, Jimei University, Xiamen 361021, China (e-mail: chenst2016@jmu.edu.cn).

Jonathan Li is with GeoSTARS Laboratory, Department of Geography and Environmental Management, University of Waterloo, Waterloo, ON N2L 3G1, Canada (e-mail: junli@uwaterloo.ca).

Digital Object Identifier 10.1109/JSTARS.2020.3035359

***Index Terms*—Augmented reality (AR), cross-domain image, domain invariant feature descriptor (DIFD), image patch matching, multisource remote sensing data, virtual–real registration.**

I. Introduction

MULTISOURCE remote sensing data, generated with different sensors or different generation mechanisms, have become the most popular geospatial data. Especially, the coupling of different source data in the same scene can better perceive the world. The different source data captured from the same scene is called cross-domain data or heterogeneous data. For example, the images of the same scene acquired by different imaging mechanism provide cross-domain images [1].

Recently, unmanned aerial vehicles (UAVs) have gradually become an essential role in acquiring near-ground remote sensing images, because of the advantages of high efficiency and low cost to capture vertical and oblique aerial photography. Through the structure-from-motion (SfM) algorithm [2], [3], such aerial photography images can be used for three-dimensional 3-D reconstruction of large-scale outdoor scenes to obtain 3-D image-based point clouds, as shown in the upper half part of Fig. 1. Based on the 3-D image-based point cloud, a 3-D image-based point cloud rendered image can be synthesized by the position information (GPS and orientation) of camera image captured from ground. The rendering process is shown in Fig. 1. For convenience, in the following, we will refer to them as ground camera image and rendered image, respectively. The image mechanisms of these two kinds of images are different and from different sources, thus, we call them cross-domain images.

Essentially, suppose the ground camera images and the 3-D image-based point cloud are registered. In that case, the 2-D and 3-D spatial relationship will be established, then the virtual–real registration of augmented reality (AR) will be calculated [4]. However, directly registering images and a large-scale 3-D image-based point cloud is extremely challenging, because the data representation of 2-D images and 3-D point clouds is cross-dimensionally inconsistent. Thus, in this article, we consider inferring the spatial relationship between 2-D and 3-D space by indirectly matching ground camera images and rendered images. The derivation schematic is shown in Fig. 1. The 3-D image-based point cloud, ground camera image, and rendered image are denoted as M, C_I, R_I, respectively, and

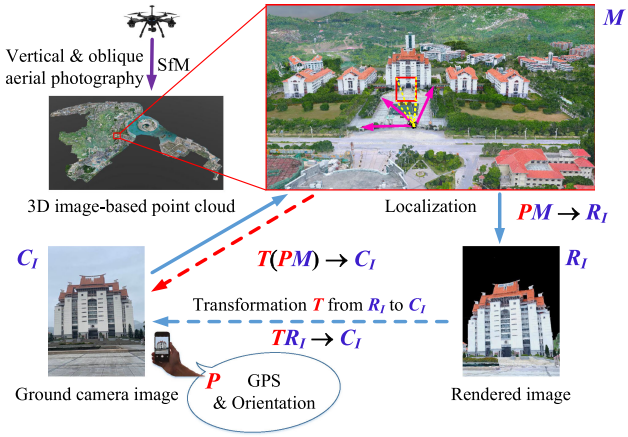


Fig. 1. Different source geospatial data: ground camera image, 3-D image-based point cloud and its rendered image. The pipeline schematic shows the process of the synthetic rendered image and the spatial relationship between the 2-D and 3-D spaces based on the above three geospatial data.

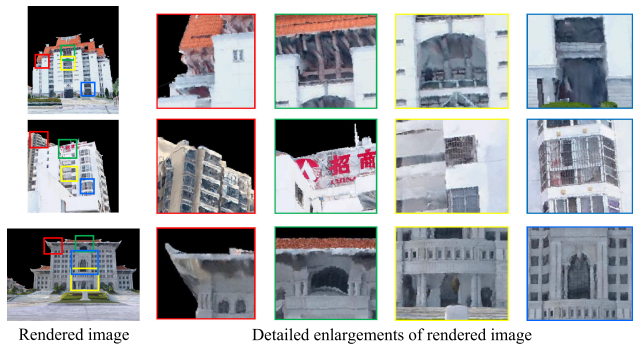


Fig. 2. Visualization of the low-quality details of rendered images.

the projection matrix is denoted as P, which obtained from the position information. So, the transformation relationship from M to R_I is $PM \rightarrow R_I$. By supposing that the transformation matrix from the rendered image to ground camera image is T, the transformation relationship from R_I to C_I is $TR_I \rightarrow C_I$. Combining the above derivation, the indirectly inferred spatial relationship between 3-D image-based point cloud and ground camera image is $T(PM) \rightarrow C_I$. Thus, the core problem of AR virtual–real registration in outdoor environments, which based on the multisource remote sensing data in this article, becomes the matching problem of ground camera images and rendered images.

It should be noted that the estimation of projection matrix P is essential. However, due to the limitations of mobile devices and experimental environments, it is difficult to estimate the projection matrix P accurately. Essentially, the projection matrix P, the spatial relationship from 3-D image-based point cloud to rendered image, is obtained from the location information of mobile devices. However, the location information from mobile devices is coarse positioning, usually with deterioration in GPS precision and distortion in the output of IMU. These errors result in the rendered images not having the exact position and orientation of the corresponding ground camera images. Thus, there is location drift between corresponding ground camera images and rendered images.

Furthermore, the motivation of this article is to explore a promising solution to virtual–real registration in outdoor AR. For the projection matrix P, we do not require it to be completely accurate. We only need a rendered image, which the viewpoint is roughly the same as the corresponding ground camera image, is synthesized from the 3-D image-based point cloud by the projection matrix P. Based on the above premises, we assume that the projection matrix P is accurate, then according to the above formula described in the third paragraph of Section I, our work can be regarded as the matching problem between the rendered images and the ground camera images, that is, estimating the transformation matrix T. So, that the 2-D and 3-D spatial relationship between the ground camera image and the 3-D image-based point cloud can be calculated through the T and P. Thus, the projection matrix P with a slight error can be tolerated to obtain cross-domain image pairs in this article.

In addition, it is challenging to match ground camera images and rendered images. First, the domain gap between ground camera images and rendered images is difficult to eliminate. Second, the rendered images are usually of low quality (such as blur, occlusion, and distortion, etc.), as shown in Fig. 2, which is caused by the following reasons: 1) Due to the inevitable occlusion of ground objects, the aerial images captured by UAVs cannot cover all the details of the terrain scene; 2) The camera lens has distorted, and the aerial images are not captured densely enough. Fig. 3 shows several failed matching results of corresponding ground camera images and rendered images by traditional handcrafted feature descriptors, such as SIFT [5], SURF [6], DAISY [7], ORB [8], BRIEF [9]. Thus, the matching problem of ground camera images and rendered images is beyond the capability of the handcrafted feature descriptors.

In this article, because of the inferior quality of the rendered image, the traditional keypoint detectors cannot detect the robust keypoints, so we consider using the image patch matching strategy to match ground camera images and rendered images. Besides, with the deep neural networks (DNNs) have achieved success in computer vision and have also become very attractive to the geoscience and remote sensing communities, we adopt the deep leaning strategy to learn invariant feature descriptors for ground camera image patches and rendered image patches.

In detail, we first propose an end-to-end network, DIFD-Net, to learn the 128-D domain invariant feature descriptors (DIFDs) for ground camera image patches and rendered image patches. The DIFD-Net is a Siamese network structure containing two autoencoders, one of which is embedded with a spatial transformer network (STN) module [10]. The STN module makes rendered images and ground camera images learning to adjust similar postures adaptively, and facilitates the feature descriptor extraction intuitively. Second, the DIFD-Net is optimized by the constructed domain-kept consistent loss function, which contains content loss, hard triplet margin loss, and feature map-based consistency loss. The negative sample sampling strategy based on hard triplet margin loss solves the interference between similar samples on feature descriptor extraction, i.e., the positive

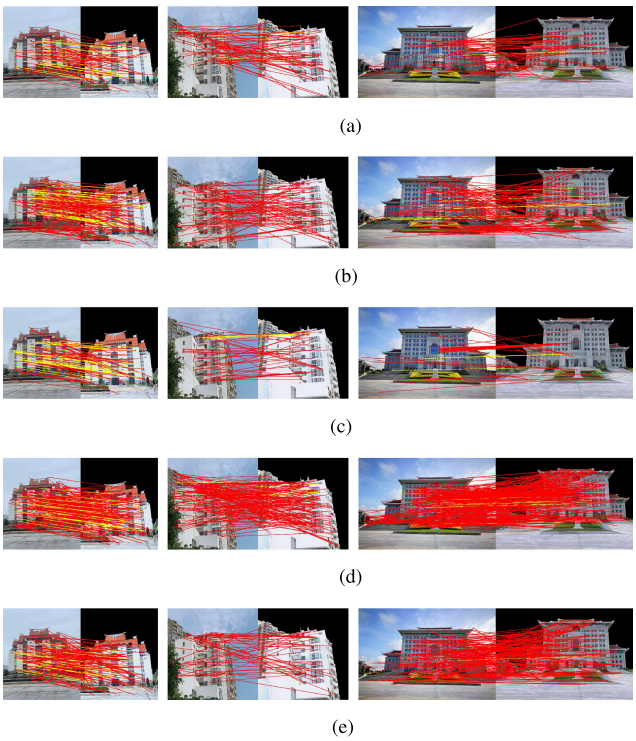


Fig. 3. Failed matching results between the cross-domain images by using the traditional handcrafted feature descriptors of SIFT [5], SURF [6], DAISY [7], ORB [8] and BRIEF [9]. The left and right sides of each pair are ground camera image and rendered image, respectively. The yellow lines represent the right match and the red lines represent the wrong match.

samples and the closest negative samples can be distinguished. The feature map-based consistency loss constraint the intermediate feature map to increase the extra supervision information for learning feature descriptors. Finally, we match the corresponding ground camera images and rendered images by using the learned DIFDs, so that the spatial relationship of ground camera images and 3-D image-based point cloud are indirectly established. Then, we verify the virtual–real registration of AR in outdoor environments by several applications.

In summary, our proposed DIFD-Net aims to learn the DIFDs for cross-domain image patches. As described in Section I, there are errors in the projection matrix P obtained by the mobile device, which result in the corresponding ground camera images and rendered images has bias and viewpoint errors. So that our collected cross-domain image patch dataset also with bias and viewpoint errors. The experimental results show that based on the above collected cross-domain image patch dataset for training, DIFD-Net still extracts the DIFDs of the cross-domain image patch pairs with bias and viewpoint errors. Therefore, DIFD-Net has generalization for cross-domain image patch pairs acquired by the projection matrix P with errors.

The main contributions of this article are as follows.

1) We propose a novel end-to-end network, DIFD-Net, to learn the DIFDs for multisource remote sensing data of ground camera image patches and rendered image patches. The ground camera images and rendered images are matched by retrieval strategy according to the learned DIFDs, then, indirectly establishing the 2-D and 3-D space relationship for virtual–real registration of AR in outdoor environments.
2) The constructed domain-kept consistent loss balances the image feature descriptors between two different domains, and narrows the domain gap between ground camera images and rendered images.
3) The learned DIFDs achieve the state-of-art retrieval performance on the retrieval benchmarks of ground camera image patches and rendered image patches.

II. Related Work

The traditional handcrafted feature descriptors cannot reach the matching between ground camera images and rendered images, and in this article, our work is to learn the DIFDs for cross-domain image matching by using the DNNs. Therefore, in this section, we only review the related feature descriptors of image patches and the image patch matching learned by DNNs. The reviews of handcrafted feature descriptors please refer to [11] and [12], which are also the comparison of handcrafted feature descriptors and learning feature descriptors.

Most image patch matching based on DNNs are retrieved according to the learned feature descriptors. Siamese networks and triplet networks are the mainstream DNNs for learning image patch feature descriptors.

Siamese network is composed of two branches of convolutional neural networks, which are divided into two categories according to whether there is a metric network. Siamese networks with or without the metric networks are usually used for binary judgment and feature descriptor extraction, respectively.

Since the high computational complexity of the metric network, the Siamese networks with the metric network consume a lot of computation, and the binary judgment output cannot provide feature descriptors that can be retrieved. Therefore, it is difficult for Siamese networks with metric network to be applied to image retrieval in real time, such as MatchNet [13] and Deepcompare [14]. On the contrary, the Siamese networks without the metric network learn the feature descriptors which can be retrieve by nearest neighbor search, such as DeepDesc [15], DeepCD [16], L2-Net [17], etc. However, the margin in the loss function that used to optimize these network is usually determined by a lot of experiments.

Triplet network, which is the improved form of the Siamese network, are optimized by the triplet loss to learn more robust feature descriptors [18]–[22]. However, the triplet networks converge slowly, or even do not converge, and the margin in triplet loss also usually empirically set by a large experiments.

It should be noted that the above Siamese networks and triplet networks have achieved excellent image patch matching performance on the Brown [23], Oxford [24], and Hpatches [25] datasets, whereas have the unsatisfactory results on the cross-domain image patching of ground camera images and rendered images (as shown in Section IV-B).

The most relevant works with cross-domain image (ground camera image and rendered image) patch matching is H-Net [26], H-Net++ [26], SiamAM-Net [1], and AE-GAN-Net [27]. H-Net only performs the binary matching judgments of cross-domain image patches. The structure of H-Net++ and

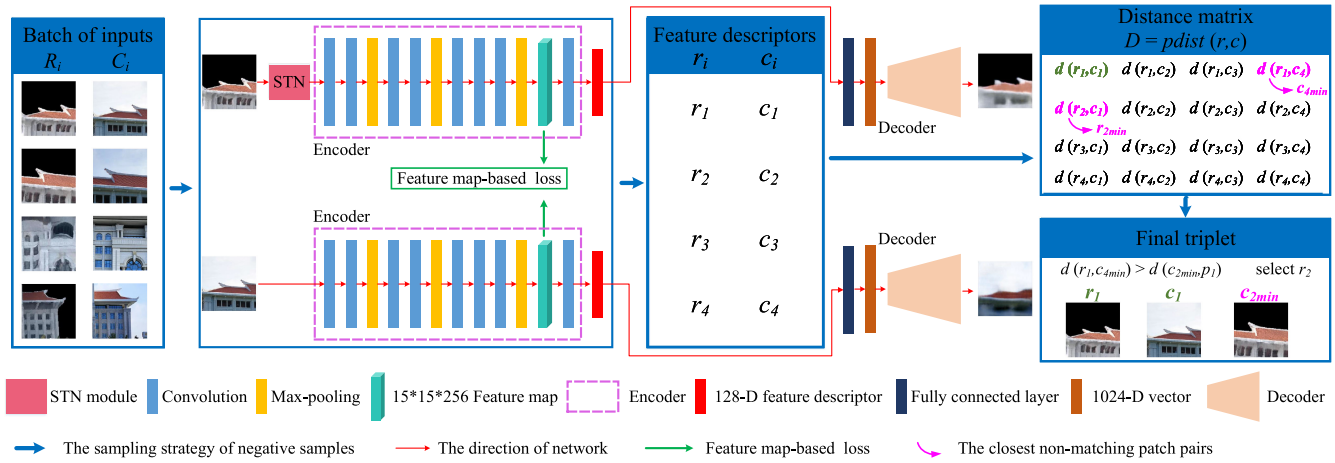


Fig. 4. Network structure of DIFD-Net and the sampling strategy of negative samples.

SiamAM-Net are similar, which embeds autoencoder into the Siamese network. However, the accuracy of image patch retrieval through the feature descriptors learned by H-Net++ and SiamAM-Net is not enough, resulting in mismatches of cross-domain image matching. AE-GAN-Net combine the generative adversarial network (GAN) [28] and autoencoder to learn the invariant feature descriptors of cross-domain image patches. Although the feature descriptors learned by AE-GAN-Net are robust, the existence of GAN makes the training process of AE-AN-Net unstable and slow convergence. Therefore, AE-GAN-Net is difficult to train and requires multiple attempts to determine suitable parameters. In addition, the negative samples of H-Net, H-Net++, SiamAM-Net, and AE-GAN-Net are randomly constructed, and the margins in the loss also are empirically set, which makes them difficult to separate the positive samples and the closest negative samples. To overcome this drawback, we adaptively obtain negative samples, which the distance is the closest, from positive samples through hard triplet loss during training.

III. Methodology

To match ground camera images and rendered images, we use the proposed DIDF-Net to learn the DIFDs of image patches, then use the image patch retrieval strategy. In this section, we introduce the proposed DIFD-Net, the constructed domain-kept consistent loss function and the training strategy.

A. Difd-Net

The framework of proposed DIFD-Net is shown in Fig. 4. DIFD-Net consists of two identical autoencoder branches, one is used to learn the feature descriptors of ground camera image patches, and the other one is used to learn the feature descriptors of rendered image patches. Specially, the branch whose input is rendered image patches has an embedded STN module [10] before autoencoder. It should be noted that the inputs of DIFD-Net are the matching patch pairs of ground camera image and rendered image, whereas the nonmatching image patch pairs used in training step are generated during the training process. The size of the input image patches is resized to $256 \times 256 \times 3$. The outputs of the DIFD-Net are the learned 128-D feature descriptors. The STN module is used to adaptively learn to adjust the spatial transformation of two inputs, so that their postures are as similar as possible. The details of encoder and decoder are described as follows:

1) Encoder: The components of encoder are the convolution layers with zero padding and maxpooling layer without zero padding. The batch normalization (BN) [29] and nonlinear activate function SeLU [30] follow each convolution layer successively, i.e., Conv-BN-SeLU. The detailed architecture of encoder is shown in Table I. The input and the output of the encoder are image patches of size $256 \times 256 \times 3$ and 128-D feature descriptors, respectively.

2) Decoder: The 128-D feature descriptor outputed by encoder is first mapped to a 1024-D vector, which is used as the input of decoder, through a fully connected layer. The output of decoder is the reconstructed image patches with size of $256 \times 256 \times 3$. Decoder is composed of the deconvolution layers. The nonlinear activate function is SeLU for each deconvolution layers except the last one, and the nonlinear activate function for the last deconvolution layer is Sigmoid. The detailed architecture of decoder is shown in Table II.

B. Loss Function

To optimize the proposed DIFD-Net, we construct a domain-kept consistent loss function, which is composed of content loss, hard triplet margin loss, and feature map-based loss, to balance the image feature descriptors between two different domains. In detail, 1) The content loss retains the image domain information into the feature while learning the feature descriptors of cross-domain images through the constraint of autoencoder; 2) The hard triplet loss constructs the nonmatching cross-domain image patche pairs from the matching cross-domain image patch pairs during training, which makes it possible to distinguish the matching cross-domain image patch pairs and the closest nonmatching cross-domain image patch pairs; 3) The feature

TABLE I
ARCHITECTURE OF ENCODER OF DIFD-NET

Layer	1	2	3	4	5	6	7	8	9	10	11
Type	Conv	Conv	MaxPool	Conv	Conv	MaxPool	Conv	Conv	Conv	MaxPool	Conv
Filters	32	64	-	96	256	-	384	384	256	-	128
Kernel size	5	5	3	3	3	3	3	3	3	3	7
Stride	2	2	2	1	1	2	1	1	1	2	1

TABLE II
ARCHITECTURE OF DECODER OF DIFD-NET

Layer	1	2	3	4	5	6	7
Type	DeConv	DeConv	DeConv	DeConv	DeConv	DeConv	DeConv
Filters	128	64	32	16	8	4	3
Kernel size	4	4	4	4	4	4	4
Stride	2	2	2	2	2	2	2

map-based loss introduces additional supervision for learning feature descriptors.

1) Content Loss: To make the learned feature descriptors contain both the essential features and the domain information of the cross-domain images, the pixel-wise mean squared error is used to conduct the content loss for the two branches of DIFD-Net. On the one hand, the branch, which is to learn the feature descriptors of ground camera image patches, is a traditional autoencoder. On the other hand, the STN and encoder module can be regarded as a combined encoder, so that the branch of learning rendered image patch feature descriptors is also a traditional autoencoder. Thus, the content loss is defined as follows:

$$L_{C-\text{Content}} = \frac{1}{NWH}\sum_{n=1}^{N}\sum_{x=1}^{W}\sum_{y=1}^{H}\left(C_{n,x,y} - C'_{n,x,y}\right)^2 \quad (1)$$

$$L_{R-\text{Content}} = \frac{1}{NWH}\sum_{n=1}^{N}\sum_{x=1}^{W}\sum_{y=1}^{H}\left(R_{n,x,y} - R'_{n,x,y}\right)^2 \quad (2)$$

$$L_{\text{Content}} = \alpha_1 L_{C-\text{Content}} + \alpha_2 L_{R-\text{Content}} \quad (3)$$

where α_1 and α_2 are the weights, C and R are the ground camera image patch and rendered image patch, respectively; C' and R' are the reconstructed image patches of C and R, respectively; N and $W \times H$ are the channels of image and the size of image patch, respectively.

In fact, the extracted feature descriptors embed the domain information of the image, and we expect the following hard triplet margin loss constrain the original domain information between cross-domain images, so that reducing the domain gap between cross-domain images. From the experiments, we found that $\alpha_1 : \alpha_2 = 1 : 1$ are the most suitable weights for the content loss.

2) Hard Triplet Margin Loss: Inspired by the Hard-Net [31], to overcome the problem of poor setting of margin in the loss function of Siamese and triplet networks, we embed the sampling strategy of nonmatching image patches into the training process of DIFD-Net. The sampling strategy of nonmatching image patches and the construction of hard triplet are detailed described as follows:

As shown in Fig. 4, denoting a batch of the training data as B

$$B = (R_i, C_i)_{i=1}^{n} \quad (4)$$

where n is the number of paired samples in the batch; R represents the anchor sample of rendered image patches, and C is the positive sample of the ground camera image patches, R_i and C_i are matching cross-domain image patch pairs.

Then, the $2n$ image patches in B are fed into the proposed DIFD-Net (see Fig. 4) to extract the feature descriptors (128-D vectors). So that a L2 pairwise distance matrix D is constructed by the above $2n$ calculated feature descriptors

$$D = \begin{bmatrix} d_{1,1} & d_{1,2} & \cdots & d_{1,j} & \cdots & d_{1,n} \\ d_{2,1} & d_{2,2} & \cdots & d_{2,j} & \cdots & d_{2,n} \\ \vdots & \vdots & & \vdots & & \vdots \\ d_{i,1} & d_{i,2} & \cdots & d_{i,j} & \cdots & d_{i,n} \\ \vdots & \vdots & & \vdots & & \vdots \\ d_{n,1} & d_{n,2} & \cdots & d_{n,j} & \cdots & d_{n,n} \end{bmatrix} \quad (5)$$

where $d_{ij} = d(r_i, c_j) = \sqrt{2 - 2r_i c_j}, i = 1, \ldots, n, j = 1, \ldots, n$; r_i and c_j are the learned feature descriptors of R_i and C_i.

With the distance matrix D, n matching cross-domain image pairs ((R_i, C_j), $i = j$) and $n^2 - n$ nonmatching cross-domain image pairs are constructed ($(R_i, C_j), i \neq j$), respectively. Next, we aim to construct n nonmatching cross-domain image pairs, which are closest to the n matching cross-domain image pairs, from the $n^2 - n$ nonmatching cross-domain image pairs.

For each matching cross-domain image patch pairs (R_i, C_i), the distance of feature descriptors r_i and c_i are the closest. Denoting r_i as anchor feature descriptors and c_i as positive feature descriptors. Then, the closest nonmatching feature descriptors, i.e., the second nearest neighbor for r_i and c_i are defined, respectively, as follows.

1) Supposing $c_{j_{\min}}$ as the closest nonmatching descriptor to r_i, where $j_{\min} = \arg\min_{j=1,\ldots,n,j\neq i} d(r_i, c_j)$.
2) Supposing $r_{k_{\min}}$ as the closest non-matching descriptor to c_i, where $k_{min} = \arg\min_{k=1,\ldots,n,k\neq i} d(r_k, c_i)$.

Then, for each quadruplet of the feature descriptors set $(r_i, c_i, p_{j_{\min}}, a_{k_{\min}})$, the triplet, i.e., the matching cross-domain image patch pair with the second nearest neighbor for r_i or c_i is

formed as follows:

$$(r_i, c_i, c_{j_{\min}}),\ \text{if}\ \mathrm{d}(\mathrm{r_i}, \mathrm{c_{j_{min}}}) < \mathrm{d}(\mathrm{r_{k_{min}}}, \mathrm{c_i})) \tag{6}$$

$$(c_i, r_i, r_{k_{\min}}),\ \text{if}\ \mathrm{d}(\mathrm{c_i}, \mathrm{r_{k_{min}}}) < \mathrm{d}(\mathrm{c_{j_{min}}}, \mathrm{r_i})). \tag{7}$$

Thus, from the n paired matching cross-domain image patches in each batch B, we construct the corresponding n paired matching cross-domain image patches with the closest distance.

Finally, we aim to exactly distinguish the matching and nonmatching cross-domain image patches. So, if the distance between the matching feature descriptors and the closest nonmatching feature descriptors of cross-domain image patches can be maximized, the matching and nonmatching cross-domain image patches will be distinguished. Thus, the hard triplet margin loss is defined as

$$L_{\text{Hard}} = \frac{1}{n}\sum_{i=1}^{n} \max\{0, 1 + d(r_i, c_i) - \min[d(r_i, c_{j_{\min}}), d(r_{k_{\min}}, c_i)]\}. \tag{8}$$

In detail, the blue bounding boxes and blue arrow flows in Fig. 4 show the process of negative sampling strategy. It takes four pairs of matching cross-domain image patches in the Batch B as the example, including the extraction of feature descriptor, the construction of distance matrix, the selection of the closest nonmatching cross-domain image patch pairs, and the construction of final triplet samples.

3) Feature Map-Based Consistency Loss: Most of the existing Siamese and triplet networks only focus on the constraints of the final output feature descriptors, and ignore the influence of the intermediate feature maps on the final output feature descriptors. In fact, the information of the intermediate feature maps is richer than the final output feature descriptors, which also improves the receptive field of the final output feature descriptors. Thus, the extra supervision of intermediate feature maps increases the constraints to the final output feature descriptors, so that increases the performance of DIFD-Net.

In the proposed DIFD-Net, the last feature maps of the encoder in the two branches are used as the constrained intermediate feature maps, as shown by the green cuboid in the encoder in Fig. 4. In detail, first, the last feature map of the encoder in DIFD-Net, which size is $15 \times 15 \times 256$, is resized to a 1-D feature vector. Second, the two resized feature vectors are constrained with the margin-based contrastive loss, as follows:

$$L_{\text{FeatureMap}} = \frac{1}{2} l D_f^2 + \frac{1}{2}(1-l)\{\max(0, m - D_f)\}^2 \tag{9}$$

where l is the label of the original inputs, $l = 1$ denotes the matching inputs, and $l = 0$ denotes the nonmatching inputs; $D_f = \|f_C - f_R\|$ is the Euclidean distance between the two resized feature vectors f_C and f_R. The role of margin m is to encourages the feature maps of matching pairs to be close and nonmatching pairs to be separated by a distance of at least m. In this article, the margin m is set as 0.2.

4) Domain-Kept Consistent Loss Function: The domain-kept consistent loss function is defined as the weighted sum of the above three sublosses

$$\mathcal{L} = \lambda_1 L_{\text{Content}} + \lambda_2 L_{\text{Hard}} + \lambda_3 L_{\text{FeatureMap}} \tag{10}$$

where λ_1, λ_2, and λ_3 are the weights of content, hard triplet margin, and feature map-based consistency losses, respectively. From the experiments, $\lambda_1 : \lambda_2 : \lambda_3 = 1 : 1 : 1$ are the most suitable weights of DIFD-Net.

C. Training Strategy

The proposed DIFD-Net, trained with a Nvidia 2080 Ti GPU, is implemented by PyTorch framework. The RMSprop optimizer is used to optimize the DIFD-Net. The initial learning rates of the DIFD-Net are set as 0.001, and then decreases by 0.99 for every 4 epochs. The batch size is set as 50, DIFD-Net converges after 70 epochs, and the training time of each epoch is about 15 min. All the weights of DIFD-Net is initialized by the standard normal distribution.

IV. Experiments And Analysis

In this section, we first introduce the cross-domain image patch dataset (ground camera and render image patches) used by the proposed DIFD-Net and the comparative networks. Second, we demonstrate the performance of the feature descriptors learned by DIFD-Net. Third, we further explore ablation studies to show the superiority and generalization of DIFD-Net. Fourth, based on the DIFDs learned by DIFD-Net, we match the corresponding ground camera images and rendered images, and verify the capability of outdoor AR virtual–real registration based on the cross-domain image matching results. Finally, we conduct analysis and discussion.

A. Dataset

The cross-domain image patch dataset, which located in Xiangan campus, Xiamen University, China, used in this article is the same as the dataset used in SiamAM-Net [1]. Fig. 5 shows the examples of 3-D image-based point cloud, corresponding cross-domain images, and matching cross-domain image patches. In the cross-domain image patch dataset of SiamAM-Net, there are 45 000 pairs of matching cross-domain image patches, 45000 pairs of randomly generated nonmatching cross-domain image patches, and extra 2000 matching cross-domain image patch pairs as the testing data (retrieval benchmark). The size of the cross-domain image patches in the dataset is between 256×256 and 512×512 pixels.

It should be noted that to obtain the better corresponding rendered image and ground image pair through the projection matrix P, we choose the open outdoor scene as the experimental scene to reduce the positioning error caused by the obstruction to the mobile device. When capturing camera images, we set up the mobile phone on a handheld gimbal to acquire a more accurate camera pose, which reduces the camera jitter, to capture images. In addition, the same as described in the literature [1] and [27], to better obtain the cross-domain image patch dataset, we have carried out manual supervision, which the ground camera image and rendered image pairs with obvious deviations are discarded.

Fig. 5. Examples of cross-domain image patch dataset used in this article. (a) 3D image-based point cloud of Xiangan campus of Xiamen University. (b) Corresponding ground camera images (top) and rendered images (bottom). (c) Matching ground camera image patches (top) and rendered imagepatches(bottom).

In addition, because of the low quality of the rendered images (mentioned in Fig. 2), it is challenging to extract meaningful handcrafted keypoints. Therefore, it is impossible to use the existing handcrafted keypoints to create the cross-domain image patch dataset automatically. Thus, when creating the cross-domain image patch dataset, the manual intervention is inevitably required, which is the same as the dataset collection method described by [1].

In detail, on the one hand, the training data of the proposed DIFD-Net is only the above 45 000 pairs of matching cross-domain image patches, whereas the nonmatching cross-domain image patches are constructed by the sampling strategy described in Section III-B2. On the other hand, the training data of the comparative networks are the above 45 000 paired matching cross-domain image patches and 45 000 paired randomly generated nonmatching cross-domain image patches. The testing data of the proposed DIFD-Net and the comparative networks are the above extra 2000 matching cross-domain image patch pairs.

B. Performances of DIFDs

1) Retrieval Performance of DIFDs: The TOP1 and TOP5 retrieval accuracy on the 2000 pairs retrieval benchmark by the learned feature descriptors is used as the measure of DIFD-Net and all comparative networks. Considering using the feature descriptor of each query ground camera image patch to retrieve in the repository of rendered image patch feature descriptors. The successful TOP1 retrieval is defined as the ground camera image patch correctly matches the corresponding rendered image patch. Similarly, the successful TOP5 retrieval is defined as one of the top five retrieved results is correct. It should be noted that, for a fair comparison, all the comparative networks are strictly reimplemented according to their respective paper, including the size of the input image patches, the details of network and the optimization methods, etc. All the comparative networks are retrained on the cross-domain image patch dataset of this article until converge.

On the extra collected 2000 pairs of cross-domain image patch retrieval benchmark, the TOP1 and TOP5 retrieval accuracy results of DIFD-Net and comparative networks are shown in Table III, which shows that our proposed DIFD-Net achieves the state-of-art retrieval performance. The structure of AE-GAN-Net [27], SiamAM-Net [1], and H-Net++ [26] are similar to the proposed DIFD-Net, but the sampling strategy of nonmatching cross-domain image patches is different. Thus, demonstrating the superiority of the embedded sampling strategy, which is to find the closest nonmatching cross-domain image patch pairs. DeepDesc [15] and Siam_l2 [14] are the traditional Siamese network without a metric network; DeepCD [16] is a Siamese network with two asymmetric branches without the metric network; L2-Net [17] is the Siamese network with a novel data sampling strategy without the metric network. DOAP [21], DDSAT [20], and DescNet [22] are the triplet networks to learn feature descriptors. However, the feature descriptors of cross-domain image patches learned by such these Siamese networks without metric network and triplet networks are not robust, because the low TOP1 and TOP5 retrieval accuracy results.

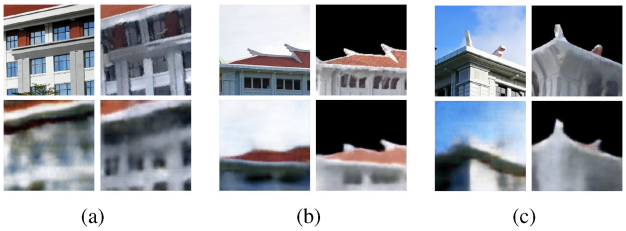
In addition, we also show the processing speed time of the DIFD-Net and the compared networks, as shown in Table III. In fact, a comparison of the training time is meaningless, because the convergence time is different for different network models. Thus, our comparison is that of the time to extract feature descriptors from the trained networks. As is seen, because the computational time of the feature extraction of the networks is related to the depth, width, and complexity of the networks, the computational time of the feature extraction on our proposed DIFD-Net is not the fastest. Although the computational time of the feature extraction of our proposed DIFD-Net is not the fastest, the performance of DIFDs are better than the learned feature descriptors learned by other compared networks. This article aims to learn robust and invariant feature descriptors of cross-domain image patches without considering the processing speed time. In future work, we plan to accelerate the computational time of feature extraction under the premise that the learned feature descriptors are robust and invariant.

2) Visualization of the Generated Image Patches: Fig. 6 shows the visualization of the paired input cross-domain image patches and the corresponding image patches generated by the autoencoder through the learned DIFDs. It can be observed that the generated image patches have the same domain information as the corresponding input image patches, which demonstrates that the DIFDs learned by DIFD-Net embedded the domain information of the image. This conclusion is the result of the interaction of the three subitems [see (3), (8), and (9)] of the proposed domain-kept consistent loss function (see (10)]. The reason is that 1) the hard triplet margin loss [see (8)] and feature map-based loss [see (9)] make the learned feature descriptors

TABLE III
TOP1, TOP5 RETRIEVAL ACCURACY AND PROCESSING SPEED TIME RESULTS OF LEARNED FEATURE DESCRIPTORS BY DIFD-NET AND COMPARATIVE NETWORKS

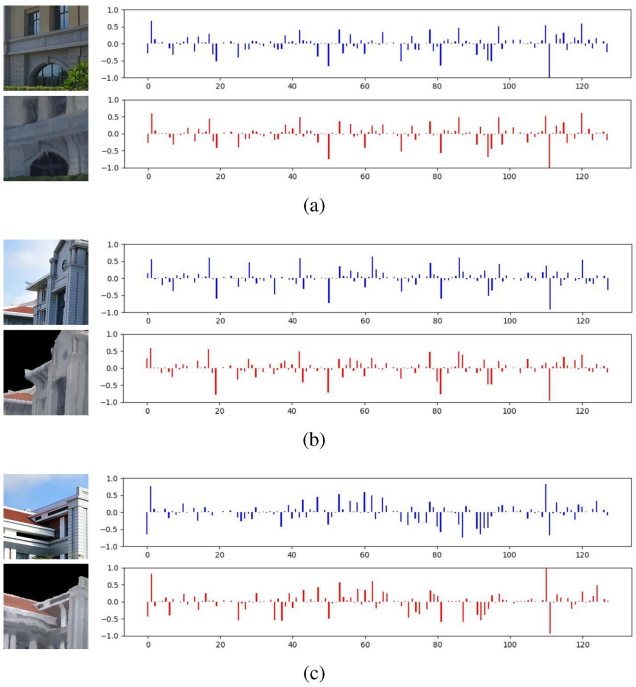
	DIFD-Net	AE-GAN-Net	SiamAM-Net	H-Net++	DeepCD	L2-Net	DeepDesc	Siam_l2	DOAP	DDSAT	DescNet
TOP1	**0.9200**	0.9025	0.8150	0.7075	0.5775	0.4695	0.5360	0.3895	0.6255	0.6125	0.6120
TOP5	**0.9875**	0.9340	0.9250	0.8590	0.6485	0.5045	0.6680	0.4475	0.6890	0.6805	0.6915
Time/(s)	**0.1937**	0.2117	0.2618	0.1931	0.1138	0.1089	0.1293	0.0649	0.1821	0.1821	0.1776

The bold values represent the result of our method and highlights the superiority of our method.



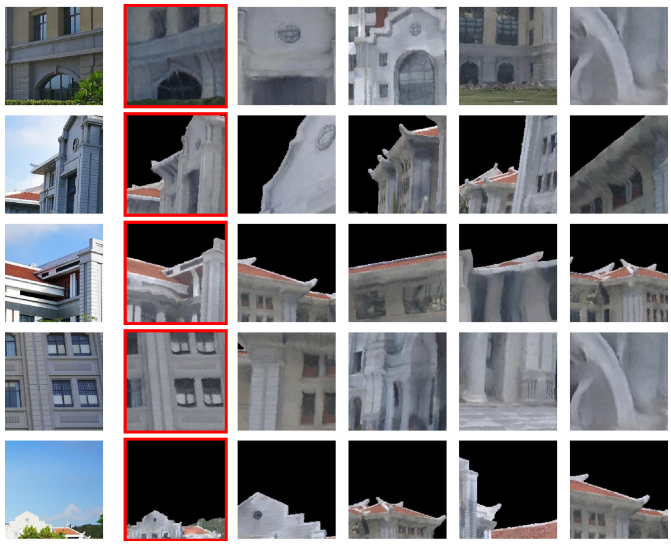
(a) (b) (c)

Fig. 6. Visualization of the generated image patches. Top: original input image patches; Bottom: generated image patches; Left: ground camera image patches; right: rendered image patches.



(a) (b) (c)

Fig. 7. Histogram visualization of the learned DIFDs of cross-domain image patch pairs.



Query TOP5 ranking results

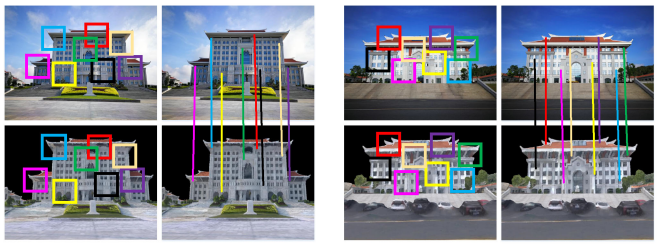
Fig. 8. TOP5 ranking results. The query image patches are ground camera image patches, and the ground truths and correct retrieval results of rendered image patches are labeled with the red bounding boxes.

invariant; 2) the content loss [see (3)] makes the learned feature descriptors not only possess the essential feature of the image patches, but also possess the domain information to restore the image patches with different domains. In addition, the generated image patches are blur, and the reason is the balanced result between the feature consistency constrained and the similarity of generated image patches.

3) Histogram Visualization of DIFDs: Fig. 7 shows the histogram visualization of the 128-D DIFDs learned by DIFD-Net. The *x*-axis and the *y*-axis are the dimension and the value of the learned DIFDs, respectively. For the matching cross-domain image patch pairs, it can be viewed that the distribution of the DIFDs is similar, and most of the values of each dimension of DIFDs are extremely similar. Thus, the histogram visualization of the matching cross-domain image patch pairs demonstrates the invariance of learned DIFDs.

4) TOP5 Retrieval Results of DIFDs: Based on the DIFDs learned by DIFD-Net, on the retrieval cross-domain image patch benchmark, Fig. 8 shows the TOP5 retrieval results of rendered image patches by using the ground image patches as the query image patches. The ground truths and correct retrieval results of rendered image patches are labeled with the red bounding boxes. It can be viewed that the most of TOP5 retrieved rendered image patches have the similar structure to the corresponding query ground camera image patches, which demonstrate the invariance of DIFDs.

5) Cross-Domain Image Patch Matching: To match the ground camera images and rendered images, we first randomly select 2000 points on the corresponding ground camera images and rendered images, which are used as the center point to collect the image patches, respectively. Second, we use the trained DIFD-Net to extract the DIFDs of the above selected cross-domain image patches. Finally, we only retain the TOP1 results retrieved based on DIFDs and use RANSAC to filter out the mismatches. Fig. 9 shows the cross-domain image patch matching results of two pairs of ground camera images and rendered images on Xiangan campus of Xiamen University.



(a) (b)

Fig. 9. Cross-domain image patch matching and center points connection results between ground camera image patches and rendered image patches on Xiangan campus of Xiamen University.

TABLE IV
ABLATION STUDIES OF DIFD-NET WITHOUT KEY MODULES

	DIFD-Net	w/o STN	w/o FeaMap	w/o HardTrip
TOP1	**0.9200**	0.9115	0.9085	0.7835
TOP5	**0.9875**	0.9675	0.9650	0.8995

The bold values represent the result of our method and highlight the superiority of our method.

The results of connecting the center points of the matched cross-domain image patches are also shown in Fig. 9.

C. Ablation Study

To demonstrate the superiority and generalization of the proposed DIFD-Net, we conduct several ablation studies.

1) DIFD-Net Without Key Modules: To quantify the benefits of the introduced STN module, feature map constraint, and the sampling strategy of constructing the closest nonmatching cross-domain image patches, we conduct several experiments to demonstrate. The experimental network structure for comparison contains DIFD-Net without STN module (w/o STN), without feature map constraint (w/o FeaMap), without hard triplet margin loss (w/o HardTrip). The TOP1 and TOP5 retrieval accuracy of above three ablation studies are shown in Table IV.

With the results of Table IV, we can conclude that the above key modules all play a role in improving the invariance and robustness for DIFD-Net to learn feature descriptors. In particular, it is obvious that the sampling strategy of constructing the closest nonmatching cross-domain image patches, that is, hard triplet margin loss, has played an extremely important role in improving the performance of DIFD-Net.

In addition, we also use the McNemar Test to demonstrate the performance of DIFD-Net with or without key modules: STN module, FeaMap module, and hard triplet margin loss. The details are as follows:

We first randomly sampled 200 pairs of cross-domain image patch pairs from the 2000 testing dataset as the validation data of the McNemar Test. Second, we performed the McNemar Test for the DIFD-Net with or without key modules: STN module, FeaMap module, and hard triplet margin loss. The results are shown in Table V– VII.

In fact, the χ^2 value follows a chi-squared distribution with one degree of freedom in our McNemar Test. We choose the confidence coefficient of 0.95, and the distribution table of the chi-square test shows that the corresponding critical value is 3.84. The χ^2 values calculated in Table V–VII are 4.00, 5.83, and 25.00, respectively, which are all greater than 3.84. Therefore, we conclude that STN module, FeaMap module, and hard triplet margin loss have a significant improvement in the performance of DIFD-Net.

TABLE V
MCNEMAR TEST OF DIFD-NET WITH OR WITHOUT STN MODULE

	After: STN Successful matching	After: STN Failed matching	Total
Before: w/o STN Succesful matching	172	2	176
Before: w/o STN Failed matching	12	12	24
Total	184	16	200

TABLE VI
MCNEMAR TEST OF DIFD-NET WITH OR WITHOUT FEAMAP MODULE

	After: FeaMap Successful matching	After: FeaMap Failed matching	Total
Before: w/o FeaMap Succesful matching	163	8	171
Before: w/o FeaMap Failed matching	21	8	29
Total	184	16	200

TABLE VII
MCNEMAR TEST OF DIFD-NET WITH OR WITHOUT HARD TRIPLET MARGIN LOSS

	After: HardTrip Successful matching	After: HardTrip Failed matching	Total
Before: w/o HardTrip Succesful matching	151	3	154
Before: w/o HardTrip Failed matching	33	13	46
Total	184	16	200

2) Generalization on Brown Dataset: To verify the generalization of DIFD-Net, we tested it on the Brown dataset [23], which is a more than 400 000 grayscale image patches with size of 64×64 pixels. The Brown dataset is similar to our dataset, both are image patches, but the image patches of the Brown dataset are in the same domain. Brown dataset contains three subsets of different scenarios, which are the Statue of Liberty (New York), Notre Dame (Paris), and Half Dome (Yosemite). The proposed DIFD-Net and all comparative approaches are trained on one subset and tested on the other two. Following with the standard evaluation protocol of [23], the false positive rate at 95% recall (FPR95) is used to measure of matching performance. In fact, the smaller the value of FPR95, the better the performance of the network in image patch matching. The comparative results are shown in Table VIII.

It should be noted that, for the fairness of comparison, the size of the input image patch by DIFD-Net is resized to be consistent with the methods of comparison, which is 64×64 pixels. In addition, in order to adapt to the input of DIFD-Net, we superimpose the input grayimage patches into three channels, i.e., $64 \times 64 \times 3$. At the same time, we also fine-tune the network details of DIFD-Net according to the input changes. Then, the encoder of the modified DIFD-Net is as follows: $C(32, 4, 2) - BN - \text{ReLU} -$

TABLE VIII
PATCH VERIFICATION PERFORMANCE ON THE BROWN DATASET [23] WITH PROPOSED DIFD-NET AND COMPARATIVE NETWORKS

Train	Notredame	Yosemite	Liberty	Yosemite	Liberty	Notredame	
Test	Liberty		Notredame		Yosemite		Mean
DeepDesc [15]	10.9		4.40		5.69		6.99
Siam l2 [14]	13.24	17.25	6.01	8.38	19.91	12.64	13.45
MatchNet [13]	7.04	11.47	3.06	3.80	11.6	8.70	8.05
L2-Net [17]	3.64	5.29	1.15	1.62	4.43	3.30	3.24
DeepCD [16]	2.87	7.78	6.08	6.65	7.76	3.02	5.69
TNet [32]	9.91	13.45	3.91	5.43	10.65	9.47	8.8
DOAP [21]	1.54	2.62	0.43	0.87	2.00	1.21	1.45
DDSAT [20]	1.79	2.96	0.68	1.02	2.51	1.64	1.76
TFeat [33]	7.39	10.13	3.06	3.80	8.06	7.24	6.64
DIFD-Net(Ours)	**3.17**	**5.46**	**5.06**	**4.88**	**5.70**	**3.19**	**4.57**

The bold values represent the result of our method and highlight the superiority of our method.

Fig. 10. The 3-D image-based point cloud of Zhangzhou Port.

$C(64,4,2)-BN-\text{ReLU}-C(128,4,2)-BN-\text{ReLU}-C(256,4,2)-BN-\text{ReLU}-C(256,128,4)-$ 128-D feature descriptor, where $C(n,k,s)$ denote a convolution layer with n filters of kernel size $k \times k$ having stride s, ReLU is the nonlinear activate function. The decoder of the modified DIFD-Net is as follows: 128-D vector $-TC(256,4,4)-BN-\text{ReLU}-TC(128,4,2)-BN-\text{ReLU}-TC(64,4,2)-BN-\text{ReLU}-TC(32,4,2)-BN-\text{ReLU}-TC(3,4,2)-\text{Sigmoid}$, where $TC(n,k,s)$ is the deconvolution with n output channels of size $k \times k$ and stride s, Sigmoid is the nonlinear activate function. The loss function and training strategy are the same to the original DIFD-Net.

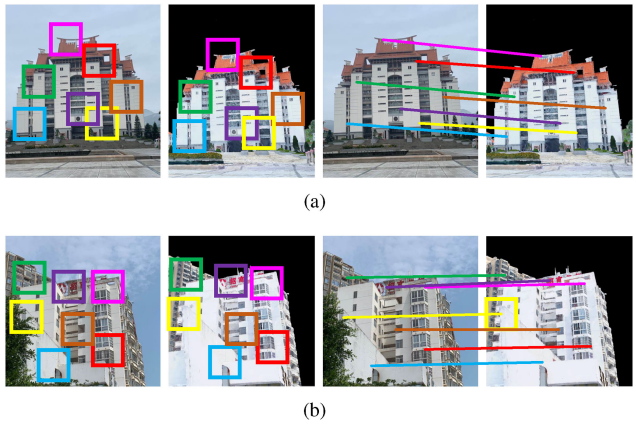
From the FPR95 results in Table VIII, compared with the existing image patch matching networks, our proposed DIFD-Net does not achieve state-of-the-art performance on the brown dataset, but it also reaches a top-ranked result, which also demonstrates the generalization of the DIFD-Net.

3) Generalization on Other Scene: To further verify the practicability and generalization of DIFD-Net, we choose another scene, Zhangzhou Port, for verification. The Zhangzhou Port is located in Fujian Province, China, about 40 square kilometers, which is reconstructed as the 3-D image-based point cloud by SfM algorithm through aerial images, as shown in Fig. 10.

In detail, we first collect 5000 paired cross-domain image patches (ground camera image patches and rendered image patches from the Zhangzhou Port) to fine tune the DIFD-Net, which has been trained on the cross-domain image patch dataset (see section IV-A). Second, we collected an additional 1000 paired cross-domain image patches from the Zhangzhou Port as testing dataset, the results are shown in Table IX. It can be viewed that the performance of DIFDs, learned by the fine-tuned DIFD-NET, has a significant improvement in the new environment.

TABLE IX
TOP1 AND TOP5 RETRIEVAL ACCURACY RESULTS OF LEARNED DIFDS BY DIFD-NET WITH OR WITHOUT FINE-TUNE IN ZHANGZHOU PORT

	W/o fine-tune	After fine-tune
TOP1	0.8115	0.9085
TOP5	0.8705	0.9690



(a)

(b)

Fig. 11. Cross-domain image patch matching and center points connection results between ground camera image patches and rendered image patches on Zhangzhou Port.

Based on the latest learned DIFDs, we match the corresponding ground camera images and rendered image in the new scene (Zhangzhou Port), the matching strategy is the same to the matching strategy in Section IV-B5, the matching results are shown in Fig. 11.

D. AR Applications

The ground camera images and rendered image are matched by the learned DIFDs with retrieval strategy. Then, the spatial relationship between 2-D and 3-D space are indirectly established according to the derivation process in Section I, i.e., the virtual–real registration of AR is established.

Fig. 12. AR applications applied by the virtual–real registration which indirectly established through the cross-domain image patch matching.

In the open outdoor environments, we implement several AR applications based on the established virtual and real registration relationship, as shown in Fig. 12, including registering the *Welcome label*, *real-time bank, and library information* to the building facade. First, we use the proposed DIFD-Net to learn the DIFDs of cross-domain image patches to match ground camera images and rendered images. Second, we use the derived spatial relationship (see the formula derivation in the third paragraph of Section I) to establish the spatial relationship between the ground camera image and the 3-D image-based point cloud, that is, the spatial relationship between 2-D and 3-D space. Third, when performing AR applications, we first specify the location of the virtual label in the 3-D space (3-D image-based point cloud), then project the above specified 3-D position to the 2-D ground camera image through the calculated spatial relationship to obtain the virtual label position in the 2-D ground camera image. Finally, the virtual labels are registered to the specified position of the ground camera images to realize the application of AR.

Such these AR applications demonstrate the feasibility of the virtual–real registration, which derived in Section I.

E. Discussions and Analysis

The benefits and limitations of the proposed DIFD-Net and virtual–real registration of AR in open outdoor environments are discussed as follows:

1) Benefit of the Embedded STN Module: There is a posture deviation between the paired cross-domain image patches. By introducing the STN module into the branch of the rendered image, it can adaptively learn to adjust the spatial transformation of the input paired cross-domain image patches, so that their postures are adjusted as similar as possible. Thus, after such prelearning posture adjustment, the similar postures of the input paired cross-domain image patches are more conducive to DIFD-Net to learn their consistent features.

2) Benefit of the Intermediate Feature Map Constraint: The intermediate feature maps have richer information than the final output feature descriptors, so that the constraint on the intermediate feature maps increases the receptive field of the final output feature descriptors, i.e., the more information, the easier to obtain deeper connections between the cross-domain image patches. Thus, the constraint on the intermediate feature maps is beneficial for DIFD-Net to learn more invariant feature descriptors for cross-domain image patch pairs.

3) Benefit of the Hard Triplet Margin Loss: The hard triplet margin loss embeds the sampling strategy of paired nonmatching cross-domain image patches, which constructs the closest nonmatching cross-domain image patch pairs from the matching cross-domain image patch pairs during training. This sampling strategy avoids the poor margin settings due to the fact that paired non-matching cross-domain image patches are randomly generated.

In fact, if the matching cross-domain image patch pairs and the closest nonmatching cross-domain image patch pairs can be distinguished by using the hard triplet margin loss, and then obviously, all nonmatching cross-domain image patch pairs also can be distinguished from matching cross-domain image patch pairs. Thus, the constraint of hard triplet margin loss makes the DIFDs learned by DIFD-Net have the ability to distinguish similar cross-domain image patch pairs.

4) Limitations: Although the DIFDs learned by proposed DIFD-Net are invariant, it is also affected by occlusion and distortion. When the occlusion and distortion are too large, the information contained in the image patch will be severely disturbed, and the learned DIFDs will be invalid. In fact, extremely serious occlusion and distortion are also beyond the scope of human visual recognition.

Meanwhile, there is a bias between the cross-domain image matching based on the image patch matching strategy, because of the center points of the two image patches may not be selected to be the same. So, the final cross-domain image matching will be biased, for example, the lines in Figs. 9 and 11 are not entirely parallel. In turn, the accuracy of virtual–real registration of AR is insufficient based on the bias of the cross-domain image matching.

In addition, it should be noted that due to the limited positioning accuracy of mobile devices, the virtual–real registration of AR in this article has only been verified in the open outdoor environments.

5) Applicability of the Scene: In fact, the proposed virtual–real registration of AR in outdoor environments relies on the image patch matching between ground camera image and rendered images, i.e., the performance of DIFDs, which is learned by our proposed DIFD-Net. If the learned DIFDs are robust, then the cross-domain images can be better matched based on the DIFDs to establish the spatial relationship of 2-D and 3-D space. In this article, the training dataset of matching cross-domain image patch pairs is collected from buildings as much as possible. The proposed DIFD-Net is trained with cross-domain image patch dataset collected on the Xiangan campus of Xiamen University, and the performance of DIFD-Net is verified in this environment. In addition, we have also tested the DIFD-Net in another scenario of Zhangzhou Port in Figs. 10 and 11. However, because the building styles of the two scenes are different, if the DIFD-Net, which trained only at Xiangan campus of Xiamen University, is applied to the Zhangzhou Port, the performance of cross-domain image patch matching is limited (see the Table IX). So that, we selected some cross-domain image patch pairs in Zhangzhou Port to fine tune the trained DIFD-Net. Therefore,

if the cross-domain image patch training dataset of DIFD-Net contains enough building data of various styles, then DIFD-Net will have stronger generalization in a new environment, and the extracted DIFDs will be more robust. On the contrary, if the amount of training data is limited, for the application of DIFD-Net in a new scene, it is necessary to collect additional data in the new environment to fine-tune the trained DIFD-Net.

V. Conclusion

In this article, we explored the matching problem of multi-source geospatial data, which are ground camera images and 3-D image-based point cloud rendered images (called as cross-domain images). Meanwhile, based on the cross-domain image matching, an virtual–real registration scheme of AR is indirectly derived in outdoor environments.

The image patch matching strategy based on deep learning is used for cross-domain image matching. Therefore, we propose an end-to-end network, DIFD-Net, to learn the DIFDs for ground camera image patches and rendered image patches. The STN module embedded in a branch of DIFD is used to adaptively learn to adjust the spatial transformation of the input paired cross-domain image patches. Then, we construct a domain-kept consistent loss function, contained content, hard triplet margin, and feature map-based consistency losses, to optimize the DIFD-Net. Essentially, the learned image feature descriptors between two different domains are balanced by the proposed domain-kept consistent loss function. The constraint on the intermediate feature maps increases the receptive field of the final output feature descriptors. Specifically, the negative sample sampling strategy through hard triplet margin loss solves the interference of similar samples on feature descriptor extraction, i.e., the positive samples and the closet negative samples can be distinguished.

Experiments show that the learned DIFDs achieve the state-of-art retrieval performance on the retrieval benchmarks of ground camera image patches and rendered image patches. Finally, several AR applications demonstrate the feasibility of the virtual–real registration which indirectly derived by the cross-domain image matching.

In future work, we will focus on applying DIFD-Net to other cross-domain data matching problems and exploring the cross-domain image matching based on point-to-point matching.

References

[1] W. Liu *et al.*, "Learning to match ground camera image and UAV 3D model-rendered image based on siamese network with attention mechanism," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 9, pp. 1608–1612, Sep. 2020.

[2] J. L. Schonberger and J.-M. Frahm, "Structure-from-motion revisited," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 4104–4113.

[3] W. Li *et al.*, "A volumetric fusing method for TLS and SFM point clouds," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 9, pp. 3349–3357, Sep. 2018.

[4] W. Liu *et al.*, "Ground camera images and UAV 3D model registration for outdoor augmented reality," in *Proc. IEEE Conf. Virt. Reality 3D User Interfaces.*, 2019, pp. 1050–1051.

[5] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[6] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," in *Proc. IEEE Eur. Conf. Comput. Vision*, 2006, pp. 404–417.

[7] E. Tola, V. Lepetit, and P. Fua, "Daisy: An efficient dense descriptor applied to wide-baseline stereo," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 5, pp. 815–830, May 2009.

[8] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to sift or surf," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2012, pp. 2564–2571.

[9] M. Calonder, V. Lepetit, C. Strecha, and P. Fua, "Brief: Binary robust independent elementary features," in *Proc. IEEE Eur. Conf. Comput. Vision*. Vienna, Austria: Springer, 2010, pp. 778–792.

[10] M. Jaderberg *et al.*, "Spatial transformer networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 2017–2025.

[11] J. L. Schönberger, H. Hardmeier, T. Sattler, and M. Pollefeys, "Comparative evaluation of hand-crafted and learned local features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6959–6968.

[12] L. Zheng, Y. Yang, and Q. Tian, "Sift meets CNN: A decade survey of instance retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 5, pp. 1224–1244, May. 2018.

[13] X. Han, T. Leung, Y. Jia, R. Sukthankar, and A. C. Berg, "MatchNet: Unifying feature and metric learning for patch-based matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3279–3286.

[14] S. Zagoruyko and N. Komodakis, "Learning to compare image patches via convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 4353–4361.

[15] E. Simo-Serra, E. Trulls, L. Ferraz, I. Kokkinos, P. Fua, and F. Moreno-Noguer, "Discriminative learning of deep convolutional feature point descriptors," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 118–126.

[16] T.-Y. Yang, J.-H. Hsu, Y.-Y. Lin, and Y.-Y. Chuang, "DeepCD: Learning deep complementary descriptors for patch representations," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 3314–3322.

[17] Y. Tian, B. Fan, and F. Wu, "L2-Net: Deep learning of discriminative patch descriptor in Euclidean space," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 661–669.

[18] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 815–823.

[19] M. Sabri and T. Kurita, "Facial expression intensity estimation using siamese and triplet networks," *Neurocomputing*, vol. 313, pp. 143–154, 2018.

[20] M. Keller, Z. Chen, F. Maffra, P. Schmuck, and M. Chli, "Learning deep descriptors with scale-aware triplet networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2762-2770.

[21] K. He, Y. Lu, and S. Sclaroff, "Local descriptors optimized for average precision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 596–605.

[22] Y. Dong *et al.*, "Local deep descriptor for remote sensing image feature matching," *Remote Sens.*, vol. 11, no. 4, 2019, Art. no. 430.

[23] M. Brown, G. Hua, and S. Winder, "Discriminative learning of local image descriptors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 1, pp. 43–57, Jan. 2011.

[24] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 10, pp. 1615–1630, Oct. 2005.

[25] V. Balntas, K. Lenc, A. Vedaldi, and K. Mikolajczyk, "HPatches: A benchmark and evaluation of handcrafted and learned local descriptors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* , vol. 4, no. 5, 2017, Art. no. 6.

[26] W. Liu, X. Shen, C. Wang, Z. Zhang, C. Wen, and J. Li, "H-Net: Neural network for cross-domain image patch matching," in *Int. Join. Conf. Arti. Intel.*, 2018, pp. 856–863.

[27] W. Liu *et al.*, "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.*, vol. 11, no. 19, 2019, Art. no. 2243.

[28] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[29] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015, *arXiv:1502.03167*.

[30] G. Klambauer, T. Unterthiner, A. Mayr, and S. Hochreiter, "Self-normalizing neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 971–980.

[31] A. Mishchuk, D. Mishkin, F. Radenovic, and J. Matas, "Working hard to know your neighbor's margins: Local descriptor learning loss," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 4826–4837.

[32] B. Kumar *et al.*, "Learning local image descriptors with deep siamese and triplet convolutional networks by minimising global loss functions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 5385–5394.

[33] V. Balntas, E. Riba, D. Ponsa, and K. Mikolajczyk, "Learning local feature descriptors with triplets and shallow convolutional neural networks," in *Proc. Brit. Mach. Vis. Conf.* , vol. 1, no. 2, 2016, Art. no. 3.

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 computer vision, machine learning, mobile laser scanning point cloud data processing, and augmented reality.

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.

Cheng Wang (Senior Member, IEEE) received the Ph.D. degree in signal and information processing from the National University of Defense Technology, Changsha, China, in 2002.

He is currently a Professor with the School of Informatics, and the Executive Director with the Fujian Key Laboratory of Sensing and Computing for Smart Cities, Xiamen University, Xiamen, China. He has coauthored more than 150 papers in referred journals and top conferences including IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, PR, IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, *Association for the Advancement of Artificial Intelligence* (AAAI), and *International Society for Photogrammetry and Remote Sensing* (ISPRS) *Journal of Photogrammetry and Remote Sensing*. His current research interests include point cloud analysis, multisensor fusion, mobile mapping, and geospatial big data.

Prof. Wang is a Fellow of the Institution of Engineering and Technology. He is also the Chair of the Working Group I/6 on Multisensor Integration and Fusion of the International Society of Remote Sensing.

Guorong Cai received the Ph.D degree in artificial intelligence from Xiamen University, Fujian, China, in 2013.

He is currently a Professor with Computer Engineering College, Jimei University, Xiamen, China. His research interests include 3-D reconstruction, point cloud processing, machine learning, object detection/recognition, and image/ video retrieval.

Yanfei Su received the B.S. and M.S. degrees in computer science and technology from Information Engineering School, Nanchang University, Nanchang, China, in 2018. 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, remote sensing image registration, mobile laser scanning point cloud data processing, and 3-D reconstruction.

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, and medical image analysis.

Yongchuan Li received the MA.Eng. degree in computer science and technology from the National University of Defense Technology, Changsha, China, in 2012.

He is currently a Research Enginner with the Fujian Key Laboratory of Sensing and Computing for Smart Cities and the School of Informatics, Xiamen University, Xiamen, China. His current research interests include mobile laser scanning data analysis and 3-D point cloud visualization.

Shuting Chen received the M.S. degrees in center for discrete mathematics and theoretical computer science from Fuzhou University, Fuzhou, China, in 2016.

She is currently an Assistant Professor with Chengyi College, Jimei University, Xiamen, China. Her current research interests include fuzzy mathematics, topology and image analysis.

Jonathan Li (Senior Member, IEEE) 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 Department of Systems Design Engineering, University of Waterloo, Waterloo, ON, Canada, and is also with the Fujian Key Laboratory of Sensing and Computing for Smart Cities, School of Informatics, Xiamen University, Xiamen, China. His research has been funded by the Natural Sciences and Engineering Research Council (NSERC) of Canada, Canada Foundation for Innovation (CFI), Agriculture and Agri-Food Canada (AAFC), Canadian Space Agency (CSA), Ministry of Transportation of Ontario (MTO), and industries. He has coauthored more than 400 publications, including more than 160 refereed journal papers. He has supervised more than 10 Ph.D. and 50 master's students to completion. His main research interests include light detection and ranging and synthetic aperture radar data processing, machine learning, and remote sensing applications.

He is currently serving as the Associate Editor for the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING, IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, and *Canadian Journal of Remote Sensing*.