Learning to Match Ground Camera Image and UAV 3-D Model-Rendered Image Based on Siamese Network With Attention Mechanism

Weiquan Liu, Cheng Wang, *Senior Member, IEEE*, Xuesheng Bian, Shuting Chen, Shangshu Yu, Xiuhong Lin, Shang-Hong Lai, Dongdong Weng, and Jonathan Li, *Senior Member, IEEE*

***Abstract*—Different domain image sensors or imaging mechanisms provide cross-domain images when sensing the same scene. There is a domain shift between cross-domain images so that the image gap between different domains is the major challenge for measuring the similarity of the feature descriptors extracted from different domain images. Specifically, matching ground camera images and unmanned aerial vehicle (UAV) 3-D model-rendered images, which are two kinds of extremely challenging cross-domain images, is a way to establish indirectly the spatial relationship between 2-D and 3-D spaces. This provides a solution for the virtual-real registration of augmented reality (AR) in outdoor environments. However, during matching, handcrafted descriptors and existing learning-based feature descriptors limit the rendered images. In this letter, first, to learn robust and invariant 128-D local feature descriptors for ground camera and rendered images, we present a novel network structure, SiamAM-Net, which embeds the autoencoders with an attention mechanism into the Siamese network. Then, to narrow the gap between the cross-domain images during the optimizing of SiamAM-Net, we design an adaptive margin for the loss function. Finally, we match the ground camera-rendered images by using the learned local feature descriptors and explore the outdoor AR virtual-real registration. Experiments show that the local feature descriptors, learned by SiamAM-Net, are robust and achieve state-of-the-art retrieval performance on the cross-domain image data set of ground camera and rendered images. In addition, several outdoor AR applications also demonstrate the usefulness of the proposed outdoor AR virtual-real registration.**

***Index Terms*—Attention mechanism, augmented reality (AR), cross-domain image patch matching, Siamese network, virtual-real registration.**

I. Introduction

IMAGES of the same scene can be acquired from different sensors or imaging mechanisms, thereby providing cross-domain images that are defined by the sensors and imaging mechanisms. Recently, two kinds of cross-domain images have become readily available: 1) camera images taken from ground with mobile devices (ground camera images) and 2) synthetic images rendered from 3-D models recovered by unmanned aerial vehicle (UAV) image sequences via the Structure-from-Motion (SfM) technology [1] (UAV 3-D model rendered images, “rendered image” for short). Essentially, the above cross images are from the same viewpoints, which are the same scene with different representations and have domain shift.

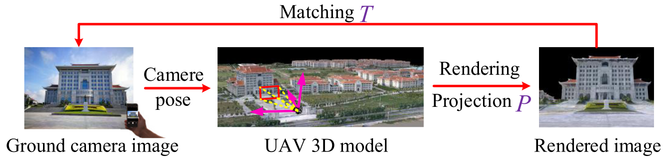


Fig. 1. Rendering schematic of UAV 3-D model rendered image and the spatial relationship between the data used in this letter.

In fact, by matching ground camera images and rendered images, the spatial relationship between ground camera images and UAV 3-D models (i.e., 2-D and 3-D space) can be indirectly established. A mechanism, such as this, provides a method for the virtual-real registration of augmented reality (AR) in outdoor environments, and the details of which are as follows.

1) A camera pose is first acquired from sensors [global position system (GPS) and inertial measurement unit (IMU)] as an initial estimate, which is used to render (synthesize) an image from the same viewpoint with the prereconstructed UAV 3-D model. Thus, the projection relationship (projection matrix P shown in Fig. 1) between the UAV 3-D model and the rendered image is established.
2) If the matching relationship (T) of the rendered image with the ground camera image can be established, the spatial relationship of the UAV 3-D model with the ground camera image can be established ($P \cdot T$, as shown in Fig. 1).

Many current outdoor AR methods in uncontrolled outdoor environments (e.g., with a huge number of objects, dramatic changes in illumination, etc.) suffer many problems, such as the impracticality of preplacing visual fiducial markers, deviation of the multisensors, sensitivity to motion blur, changes in lighting conditions, and occlusion [2]. Compared with the above-mentioned outdoor AR approaches, the proposed solution of outdoor AR virtual-real registration is intuitive and portable.

The key to performing the proposed outdoor AR virtual-real registration is to match the rendered and ground camera

Manuscript received July 22, 2019; revised September 25, 2019; accepted October 21, 2019. Date of publication November 8, 2019; date of current version August 28, 2020. This work was supported by the National Natural Science Foundation of China under Grant U1605254, Grant 41471379, and Grant 61371144. *(Corresponding author: Cheng Wang.)*

W. Liu, C. Wang, X. Bian, S. Yu, X. Lin, and J. Li are with the Fujian Key Laboratory of Sensing and Computing for Smart City, School of Informatics, Xiamen University, Xiamen 361005, China (e-mail: wqliu1026@163.com; cwang@xmu.edu.cn; junli@xmu.edu.cn).

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

S.-H. Lai is with the Department of Computer Science, National Tsing Hua University, Hsinchu 30013, Taiwan (e-mail: lai@cs.nthu.edu.tw).

D. Weng is with the Beijing Engineering Research Center of Mixed Reality and Advanced Display, School of Optics and Photonics, Beijing Institute of Technology, Beijing 100081, China (e-mail: crgj@bit.edu.cn).

Color versions of one or more of the figures in this letter are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LGRS.2019.2949351

TABLE I

TOP1, TOP5 RETRIEVAL ACCURACY, AND COMPUTATIONAL TIME OF FEATURE EXTRACTION FOR SIAMAM-NET AND COMPARATIVE NETWORKS

	SiamAM-Net	SiamAM-Net w/o AM	H-Net++ [15]	Siam_l2 [4]	DeepCD [5]	L2-Net [6]	DOAP [8]	DDSAT [7]
TOP1	**0.8150**	0.7550	0.7075	0.3895	0.5775	0.4695	0.6255	0.6125
TOP5	**0.9250**	0.8990	0.8590	0.4475	0.6485	0.5045	0.6890	0.6805
Time/(s)	**0.2618**	0.2304	0.1931	0.0649	0.1138	0.1089	0.1821	0.1803

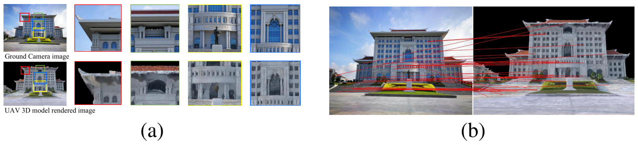


Fig. 2. Visualization of detailed enlargements and a failed matching example by handcrafted descriptor of cross-domain images. (a) Detailed enlargements. (b) Failed matching result by SIFT.

images. However, UAV 3-D models are reconstructed with the aerial images captured by vertical and slope photography. Thus, it is hard to ensure the quality of 3-D models reconstructed from SfM because the aerial images, usually occluded close to the ground, are noisy. Thus, as shown in Fig. 2(a), rendered images are of poor quality, low resolution, occluded, and hugely distorted. Therefore, handcrafted features cannot match the rendered images and ground camera images. A failed example of scale invariant feature transform (SIFT) [3] is shown in Fig. 2(b). In addition, as shown in Table I of Section III, the recent Siamese networks [4]–[6] and triplet networks [7], [8] also cannot learn the invariant feature descriptors of the rendered images and ground camera images.

Recently, deep learning is increasingly applied to the remote sensing image processing. Liu *et al.* [9], [10] use the Siamese networks to learn robust features for the remote sensing scene classification and image classification, respectively. Wang *et al.* [11] and Haut *et al.* [12] embed the attention mechanism [13] to adaptively select attention regions for the remote sensing scene classification and image superresolution, respectively.

In this letter, to match the rendered images and ground camera images, we propose a novel network, SiamAM-Net, consisting of two autoencoders, to learn the invariant local feature descriptors for above cross-domain images. First, by imitating the habit of people always observing the salient regions of two cross-domain images, the attention mechanism [13] was embedded into the encoder of the autoencoder to assist the network in feature extraction. Second, to optimize SiamAM-Net, we design an adaptive margin for margin-based contrastive loss. The adaptive margin has the advantage of automatically generating a soft-margin based on changes in the cross-domain image data. Essentially, the goal of SiamAM-Net is to map raw inputs (cross-domain image patches) with a 128-D feature vector so that the distance between representations is small for matching patches and large otherwise. Experimental results show that the invariant feature descriptors learned from SiamAM-Net achieve state-of-the-art retrieval performance on rendered images and ground camera images. Several outdoor AR applications demonstrate the promising performance of the proposed virtual-real registration.

II. NETWORK ARCHITECTURE

A. SiamAM-Net

Siamese networks with metric networks have excellent performance on image patch matching, such as MatchNet [14],

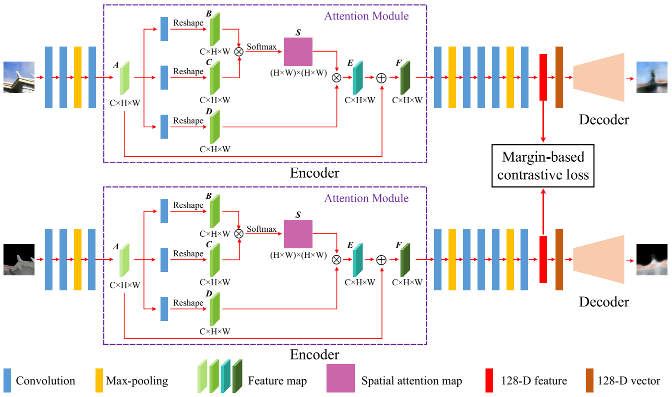


Fig. 3. Structure of SiamAM-Net. Encoder: C(32,5,2)-BN-SeLU-C(64,5, 2)-BN-SeLU-P(3,2)-C(96,3,1)-BN-SeLU-Attention Module-C(256,3, 1)-BN-SeLU-P(3,2)-C(384,3,1)-BN-SeLU-C(384,3,1)-BN-SeLU-C(256,3,1)-BN-SeLU-P(3,2)-C(128,7,1)-BN-SeLU. C(n,k,s): convolution layer with n filters of kernel size $k \times k$ with stride s; P(k,s): max-pooling layer of size $k \times k$ with strides. Decoder: FC(128,1024)-TC(128,4,2)-SeLU-TC(64,4,2)-SeLU-TC(32,4,2)-SeLU-TC(16,4,2)-SeLU-TC(8,4,2)-SeLU-TC(4,4,2)-SeLU-TC(3, 4,2)-Sigmoid. FC(p,q): the input p-dimensional feature vector is mapped to a q-dimensional feature vector through a fully connected layer; TC(n,k,s): transposed convolution with n output channels of size $k \times k$ and stride s.

H-Net [15], [16], and so on. They usually output with a binary judgment and judge only whether the input patches are similar. However, they cannot learn invariant feature descriptors of patches for retrieval and usually with extensive computation.

To learn invariant feature descriptors for ground camera images and rendered images, we use a Siamese network without a metric network. We use two autoencoders as the two branches of the Siamese network. It should be noted that, although the characteristics of rendered images and ground camera images are particularly different, it is easier for humans to judge if they are similar. Because when we observe these two kinds of cross-domain images, we often focus on their salient regions, such as outlines, special structures, and so on. Thus, to assist in extracting features, we embed the attention mechanism [13] into the encoder of the autoencoder. The details of the architecture of the proposed SiamAM-Net, as shown in Fig. 3, are as follows.

1) Encoder: The encoder consists of convolution layers with zero padding, max-pooling layers, and an attention module. Batch normalization (BN) and scaled exponential linear unit (SeLU) are used after each convolution layer. After the third convolution layer, the attention module is embedded. The inputs of the two encoders are the paired cross-domain image patches that are resized to 256 × 256 × 3. The output of the encoder is a 128-D feature vector. Details of the encoder are shown in Fig. 3.

2) Attention Module: The input of the attention module is the feature map $\boldsymbol{A} \in R^{C \times H \times W}$, which is calculated from the third convolution layer of the encoder (Fig. 3). Then, $\boldsymbol{A}$ is fed into three convolution layers (with C filters of kernel size 3×3 at stride 1) with BN and ReLU layers to generate three feature maps $\boldsymbol{B}$, $\boldsymbol{C}$, and $\boldsymbol{D} \in R^{C \times H \times W}$, respectively. Next, the spatial attention map $\boldsymbol{S} \in R^{(H\times W)\times(H\times W)}$ is calculated with a softmax layer by performing a matrix multiplication between the transpose of $\boldsymbol{B}$ and $\boldsymbol{C}$

$$s_{ji} = \frac{\exp(B_i \cdot C_j)}{\sum_{i=1}^{N} \exp(B_i \cdot C_j)} \tag{1}$$

where s_{ji} represents the impact of the ith position on the jth position. The more similar feature representations of the two positions contribute to the greater correlation between them.

Meanwhile, the feature map $\boldsymbol{D} \in R^{C\times H\times W}$ is performed a matrix multiplication with the transpose of $\boldsymbol{S}$ to obtain a feature map $\boldsymbol{E} \in R^{C\times H\times W}$. Then $\boldsymbol{E}$ is multiplied by a scale parameter α and summed with the input feature map $\boldsymbol{A}$ to obtain the final output feature map $\boldsymbol{F} \in R^{C\times H\times W}$ as follows:

$$F_j = E_j + A_j = \alpha \sum_{i=1}^{N} (s_{ji} D_i) + A_j \tag{2}$$

where α is initialized as 0 and gradually learns to assign more weight. Therefore, the feature map $\boldsymbol{F}$ at each position consists of the weighted sum of the features at all positions and the original features. Thus, $\boldsymbol{F}$ has a global contextual view and selectively aggregates contexts according to the spatial attention map [17].

3) Decoder: The 128-D features learned from the encoder are first mapped to 1024-D vectors by the fully connected network. Then, the 1024-D vectors are fed into the transpose convolution layers (with SeLU and Sigmoid) to reconstruct the input images of the encoder. Details of the decoder structure are shown in Fig. 3.

B. Loss Function

To optimize SiamAM-Net, we design a loss function that is capable of distinguishing the similar and dissimilar patch pairs of ground camera and rendered images. In detail, for the two branches, we use mean-squared error (MSE) loss, as follows:

$$L_{Ae1}(G, G') = \frac{1}{NWH} \sum_{n=1}^{N} \sum_{x=1}^{W} \sum_{y=1}^{H} \left(G_{n,x,y} - {G'}_{n,x,y}\right)^2 \tag{3}$$

$$L_{Ae2}(R, R') = \frac{1}{NWH} \sum_{n=1}^{N} \sum_{x=1}^{W} \sum_{y=1}^{H} \left(R_{n,x,y} - {R'}_{n,x,y}\right)^2 \tag{4}$$

where G and R are the patches of ground camera image and rendered image, respectively; N is the channel of the image, and $W \times H$ is the size of the image patch.

To constrain the feature descriptors learned from the two branches, we use margin-based contrastive loss, which encourages similar cross-domain image patch pairs to be close and dissimilar ones to have a Euclidean distance between them larger or equal to a margin m ($m > 0$), defined as

$$L_{\text{margin}}(G, R, l) = \frac{1}{2} l D^2 + \frac{1}{2}(1 - l)\{\max(0, m - D)\}^2 \tag{5}$$

where l is a binary label. If G and R are matched, $l = 1$; otherwise, $l = 0$. $D = \|f(G) - f(R)\|_2$ is the Euclidean distance between the learned features $f(G)$ and $f(R)$.

However, it should be noted that due to the image gap of the cross-domain images, the value of margin m is very difficult to determine. Essentially, too small a value for the margin will lead to optimizing the margin-based contrastive loss function only over the set of matching image patch pairs; otherwise, too large a value for the margin will hamper learning.

To set a logical margin, we design an adaptive strategy to obtain an adaptive margin. For each batch of the training data $\{G_k, R_k, l\}$, $k = 1, 2, \ldots, K$, where K is the number (batch size) of samples in a batch. Then, the adaptive margin m is

$$\begin{cases} m = d + \ln(d + e) \\ d = \max\{\|f(G_k) - f(R_k)\|_2 \cdot l\}. \end{cases} \tag{6}$$

Thus, based on each batch of training data, m is changed. Therefore, with the adaptive m in each batch of training data, the distance of nonmatching image paired patch feature descriptors is guaranteed to be at least greater than the sum of the maximal distance of the matching image paired patch feature descriptors (d) and an increment ($\ln(d + e)$). This mechanism better distinguishes between positive and negative samples.

Specifically, it is noteworthy that the increment $\ln(d+e)$ is important because as the training of the network converges, d will become smaller and may approach 0. If we only set m equal to d, when the network converges, m is insufficient to punish the negative samples. Thus, increment $\ln(d + e)$ guarantees that the distance between nonmatching image paired patch feature descriptors is at least 1 greater than the distance between the matching image paired patch feature descriptors. During training, at the beginning, the value of the adaptive margin m is large, which is conducive for better punishing of the negative samples. Then, as the network converges, m gradually decreases and stabilizes at about 1.

Finally, the total loss is defined as follows:

$$\mathcal{L} = \lambda_1 L_{Ae1} + \lambda_2 L_{Ae2} + \lambda_3 L_{\text{margin}} \tag{7}$$

where λ_1, λ_2, and λ_3 are the weights of content and contrastive losses, respectively. From our experiments, $\lambda_1 : \lambda_2 : \lambda_3 = 1 : 1 : 1$ are the most suitable weights of SiamAM-Net.

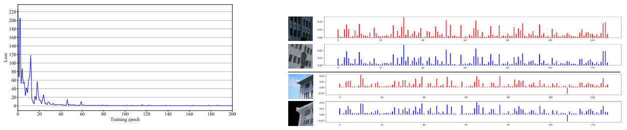
C. Training Strategy

The proposed SiamAM-Net is implemented by the Pytorch framework with an Nvidia 2080 Ti GPU. The network is optimized with the RMSprop optimizer. The learning rate, initially 0.001, decays with a factor of 0.99 in every 4 epochs. Standard normal distribution is used to initialize the weights.

III. Experiments and Results

A. Data Set

The cross-domain image patches adopted in this letter were collected from Xiang'an campus, Xiamen University, China, which covers about 3 km^2 and contains more than 100 buildings. We collected 5000 paired corresponding ground camera and rendered images (like the image pair shown in Fig. 1). To select matching cross-domain image pairs, we design a semiautomated method. First, we manually select at least four points to calculate the perspective transformations of the corresponding cross-domain images. Second, we use the detector of SIFT to extract the keypoints from the ground camera images and patches at these keypoints. Finally, to obtain matching rendered image patches, we use



(a) (b)

Fig. 4. (a) Evolution of training loss. (b) Histogram visualization of the feature descriptors learned by SiamAM-Net. First and third rows: ground camera image patches. Second and fourth rows: matching-rendered image patches.

the above-calculated perspective transformations to map the ground camera image patches to the corresponding rendered images. In addition, the nonmatching cross-domain image pairs are randomly selected from image patches in different buildings. The sizes of the collected patches are between 256 × 256 and 512 × 512 pixels.

For training data, we collected 45 000 matching and 45 000 nonmatching cross-domain image patch pairs from the 4500 paired corresponding cross-domain images; for testing data (retrieval data set), we collected 2000 matching cross-domain image patch pairs from the remaining 500 paired corresponding cross-domain images. The matching cross-domain image patch pairs are like the image patch pairs shown in Fig. 2. Also, the buildings in the testing data are not seen in the training data.

B. Experimental Results

We evaluate the performance of SiamAM-Net and comparative networks using the TOP1 and TOP5 retrieval accuracies, and the results are listed in Table I. During training, the batch size is set as 50. The evolution of the training loss is shown in Fig. 4(a), and the training loss of SiamAM-Net converges after 60 epochs. Compared with the following competing networks in Table I, SiamAM-Net shows great improvement: 1) H-Net++, which embeds the autoencoder into the Siamese network [15]; 2) Siamese network with simple Euclidean constraint [4]; 3) asymmetric Siamese network DeepCD [5]; 4) L2-Net with improved data sampling strategy [6]; and 5) triplet network DOAP [8] and DDSAT [7]. Thus, the feature descriptors learned by SiamAM-Net are robust and easily retrieved.

Also, the feature extraction computational time from the trained networks is given in Table I. In fact, the computational time of the networks is related to the depth, width, and complexity of the networks. Although the feature computational time of the compared networks is faster than SiamAM-Net, their performance of the learned feature descriptors is worse than the feature descriptors learned by SiamAM-Net. Our goal is to learn robust feature descriptors without considering the influence of time.

In addition, Fig. 4(b) shows the histogram of the cross-domain image local feature descriptors learned from SiamAM-Net. As shown, the distribution of the two matching feature descriptors is consistent, and the values of each dimension are similar, demonstrating that the feature descriptors of the matching image patches learned by SiamAM-Net are invariant.

C. Ablation Study

We conducted several experiments to quantify the introduced attention mechanism and the proposed adaptive margin.

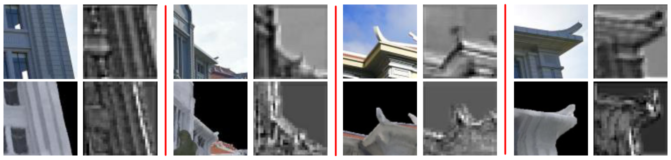


Fig. 5. Visualization of the generated image patches based on the attention mechanism. (Top) Ground camera images and their generated images. (Bottom) Corresponding rendered images of the Top row.

First, after the removal of the attention mechanism, SiamAM-Net degenerated into a specific form of H-Net++ (with the adaptive margin). Table I shows that TOP1 and TOP5 retrieval accuracies of the SiamAM-Net w/o the attention mechanism are 0.7550 and 0.8990, respectively, which is worse than the SiamAM-Net with the attention mechanism. In addition, the generated images based on the attention mechanism are visualized in Fig. 5. The highlighted pixels of the images represent salient regions, which are viewed as contours of the building and are consistent with the judgment position when humans observe whether the two cross-domain images match.

Second, the effects of multiple fixed values of the margin on SiamAM-Net were explored (Table II). Results show that our proposed adaptive margin outperforms the fixed margins. Thus, this experiment proves that it is very challenging to find a suitable fixed value for the margin and also demonstrates the superiority of our proposed adaptive margin, which can be adaptively defined as based on a change in the data.

Third, we tried using Gaussian distribution with different means and variances to initialize the weights for SiamAM-Net. We found that the performance of SiamAM-Net whether weights initialized with the above Gaussian distributions or weights initialized with the standard normal distribution is similar. Thus, SiamAM-Net is robust for the initial weights, which are initialized with Gaussian distributions.

Fourth, we tried to use triplet loss to optimize our network; however, due to the following three reasons, it did not perform well.

1) The network optimized with triplet loss converge slowly, or sometimes did not converge, and were susceptible overfitting.
2) The negative samples (nonmatched image pairs), which are randomly selected in this letter, are difficult to satisfy the hard samples. Thus, there is uncertainty in these negative samples. It is possible that the distance between positive samples is smaller than the distance between negative samples. Therefore, this situation resulted in the networks, which are optimized by triplet loss and are difficult to perform on our data set.
3) The margin in the triplet loss is difficult to define.

Finally, several failed TOP 1 retrieval results are shown in Fig. 6. The first two columns are the patches with large occlusions; the middle columns are the rendered image patches with severe distortions; the last two columns are the patches with very similar structures. It is observed that if the occlusions and distortions of the patches are very severe, it is difficult to ensure the quality of the feature descriptors learned by SiamAM-Net. In addition, the patches with very similar structures sometimes mislead the results

TABLE II
MARGIN ANALYSIS OF SIAMAM-NET

Margin	**Adaptive Margin**	0.001	0.005	0.01	0.05	0	0.1	0.5	1	5	10	50	100
TOP1	**0.8150**	0.6300	0.6355	0.6255	0.5855	0.5785	0.5900	0.6705	0.7750	0.3650	0.3755	0.1155	0.0450
TOP5	**0.9250**	0.8545	0.8610	0.8435	0.8105	0.8435	0.8655	0.8750	0.8995	0.6505	0.6845	0.3150	0.2005

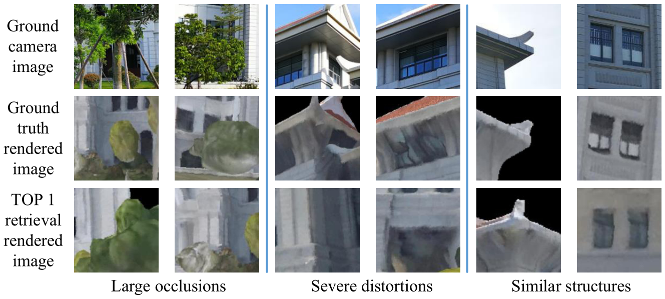


Fig. 6. Failed TOP 1 retrieval results with severe situations.

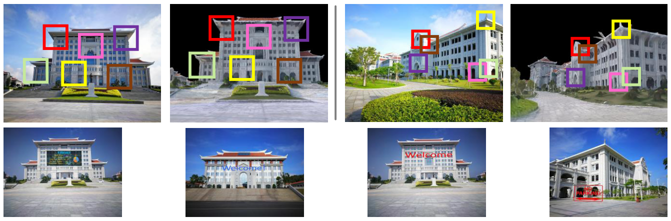


Fig. 7. (Top) Matching results. (Bottom) AR applications.

of the retrieval. In fact, these three situations sometimes are beyond the capability of humans.

D. Matching Result and AR Application

The matching results of ground camera and rendered images are computed from the matching local image patch feature descriptors (the center point of the image patch is used as keypoint). First, 2000 points in each image were selected at random, and the local patches were cropped. Second, based on the trained SiamAM-Net, feature descriptors of the cross-domain image patches were computed. Third, only matching pairs of TOP1 retrieval and cosine similarity greater than 0.9 were retained. Then, the random sample consensus (RANSAC) was used to filter mismatched pairs.

Two cross-domain image matching results are shown in the first row of Fig. 7. There are two matching results: one is the cross-domain image pairs that have a similar viewpoint and the other is the cross-domain image pairs that have a larger viewpoint bias. Combined with the experimental results shown in Fig. 4(b), we conclude that the feature descriptors learned by our SiamAM-Net are invariant against changes in distortion and viewpoints.

Based on cross-domain image matching results, several virtual objects (real-time information) were registered (second row in Fig. 7). AR applications, such as these, demonstrate the capability of the proposed outdoor AR virtual-real registration.

IV. CONCLUSION

In this letter, we proposed a network, SiamAM-Net, to learn local feature descriptors for ground camera and UAV 3-D model rendered images. First, to narrow the domain gap between the cross-domain images and assist learning features, we introduced an attention mechanism, which is beneficial for the network to focus on salient regions. Second, we proposed an adaptive strategy to design and obtain an adaptive margin (soft margin) for the margin-based contrastive loss. Experiments show that the feature descriptors learned by SiamAM-Net are robust and invariant. Finally, we performed several AR applications to demonstrate the possibility of using the proposed virtual-real registration in outdoor environments. In future work, we intend to accelerate the computational time of feature extraction and improve retrieval accuracy of features.

REFERENCES

[1] J. L. Schonberger and J.-M. Frahm, “Structure-from-motion revisited,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4104–4113.

[2] J. Rao, Y. Qiao, F. Ren, J. Wang, and Q. Du, “A mobile outdoor augmented reality method combining deep learning object detection and spatial relationships for geovisualization,” *Sensors*, vol. 17, no. 9, p. 1951, 2017.

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

[4] S. Zagoruyko and N. Komodakis, “Learning to compare image patches via convolutional neural networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4353–4361.

[5] 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. (ICCV)*, Oct. 2017, pp. 3314–3322.

[6] 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. (CVPR)*, Jul. 2017, pp. 661–669.

[7] 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. (CVPR)*, Jun. 2018, pp. 2762–2770.

[8] K. He, Y. Lu, and S. Sclaroff, “Local descriptors optimized for average precision,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 596–605.

[9] X. Liu, Y. Zhou, J. Zhao, R. Yao, B. Liu, and Y. Zheng, “Siamese convolutional neural networks for remote sensing scene classification,” *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 8, pp. 1200–1204, Aug. 2019.

[10] B. Liu, X. Yu, P. Zhang, A. Yu, Q. Fu, and X. Wei, “Supervised deep feature extraction for hyperspectral image classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 4, pp. 1909–1921, Apr. 2018.

[11] Q. Wang, S. Liu, J. Chanussot, and X. Li, “Scene classification with recurrent attention of VHR remote sensing images,” *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 1155–1167, Feb. 2019.

[12] J. M. Haut, R. Fernandez-Beltran, M. E. Paoletti, J. Plaza, and A. Plaza, “Remote sensing image superresolution using deep residual channel attention,” *IEEE Trans. Geosci. Remote Sens.*, to be published.

[13] A. Vaswani *et al.*, “Attention is all you need,” in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2017, pp. 5998–6008.

[14] 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. (CVPR)*, Jun. 2015, pp. 3279–3286.

[15] W. Liu, X. Shen, C. Wang, Z. Zhang, C. Wen, and J. Li, “H-Net: Neural network for cross-domain image patch matching,” in *Proc. Int. Join. Conf. Artif. Intell. (IJCAI)*, 2018, pp. 856–863.

[16] L. H. Hughes, M. Schmitt, L. Mou, Y. Wang, and X. X. Zhu, “Identifying corresponding patches in SAR and optical images with a pseudo-siamese CNN,” *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 5, pp. 784–788, May 2018.

[17] J. Fu *et al.*, “Dual attention network for scene segmentation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3146–3154.