DOMAINDESC: LEARNING LOCAL DESCRIPTORS WITH DOMAIN ADAPTATION

Rongtao Xu [1,3,*] *Changwei Wang* [1,3,*] *Bin Fan* [5]
Yuyang Zhang [1] Shibiao Xu [2,†] Weiliang Meng [1,4,3,†] Xiaopeng Zhang [1,3]

[1] NLPR, Institute of Automation, Chinese Academy of Sciences
[2] School of Artificial Intelligence, Beijing University of Posts and Telecommunications
[3] School of Artificial Intelligence, University of Chinese Academy of Sciences [4]Zhejiang Lab
[5]School of Automation and Electrical Engineering, University of Science and Technology Beijing

ABSTRACT

Robust and efficient local descriptor is crucial in a wide range of applications. In this paper, we propose a novel descriptor **DomainDesc** which is invariant as much as possible by learning local **Desc**riptor with **Domain** adaptation. We design the feature-level domain adaptation loss to improve robustness of our **DomainDesc** by punishing inconsistent high-level feature distributions of different images, while we present the pixel-level cross-domain consistency loss to compensate for the inconsistency between the descriptors corresponding to the keypoints at the pixel level. Besides, we adopt a new architecture to make the descriptor contain as much information as possible, and combine triplet loss and cross-domain consistency loss for descriptor supervision to ensure the distinguished ability of our descriptor. Finally, we give a cross-domain dataset generation strategy to quickly construct our training dataset for diverse domains to adapt to complex application scenarios. Experiments validate that our DomainDesc achieves state-of-the-art performances on HPatches image matching benchmark and Aachen-Day-Night localization benchmark.

***Index Terms*—** local descriptors, domain adaptation, cross-domain data, consistency loss.

1. INTRODUCTION

Extracting accurate and efficient local features description is an indispensable processing step for various computer vision applications [1, 2, 3, 4]. In the traditional hand-crafted feature extraction method, SIFT [5] plays a vital role in computer vision tasks because it is scale invariant and rotation invariant. Other methods like HardNet [6], only learn to extract the descriptor for each patch with deep learning. Unlike patch-based feature descriptions, extracting dense feature descriptors for the whole image has become a trend in recent years, and various fully-convolutional neural networks are proposed including SuperPoint [7], D2Net [8], R2D2 [9], Aslfeat [10], CAPS [11], Disk [12], etc.

In general, robust descriptors should be immune from the changing of the illumination or viewpoint of the same keypoints, required to capture much invariance as possible for accurate matching. However, it is extremely challenging to extract robust descriptors for complex scenes including day-night changes and seasonal variations [2], while many methods enhance the robustness of descriptors by utilizing large training datasets. In contrast, we focus on combining domain adaptation to improve the accuracy of local feature descriptions, which refers to mapping data distributed in different domains to the same feature domain and makes the distance in the feature space as close as possible.

To accurately describe similar keypoints in images, we proposed a novel descriptor 'DomainDesc' based on our cross-domain datasets, feature-level domain adaptation loss and pixel-level cross-domain consistency loss. As shown in Fig. 1 (a), our cross-domain dataset contains more diverse domain data. We design the feature-level domain adaptation loss to make descriptors more robust by narrowing the high-level features of different images. Furthermore, we give the cross-domain consistency loss to compensate for the inconsistency between descriptors corresponding to keypoints at the pixel level.

Excellent local descriptors should also have distinguished ability, meaning that distinct keypoints from similar textures or the shapes should not be matched. For this reason, we adopt a new architecture that contains as much context information as possible to ensure distinguished ability on both image matching and visual location tasks, and combine triplet loss [6] for our descriptor supervision to distinguish outliers. Our competitive numerical results on Hpatches are shown in Fig. 1 (b).

In sum, there are three main contributions in this work:

- We propose a novel local descriptor **DomainDesc** by introducing domain adaptation in local feature learning for the first time, and design the feature-level do-

[*] Rongtao Xu and Changwei Wang contributed equally.

[†] Shibiao Xu and Weiliang Meng are the corresponding authors (shibiaoxu@bupt.edu.cn; weiliang.meng@ia.ac.cn).

978-1-6654-0540-9/22/$31.00 ©2022 IEEE

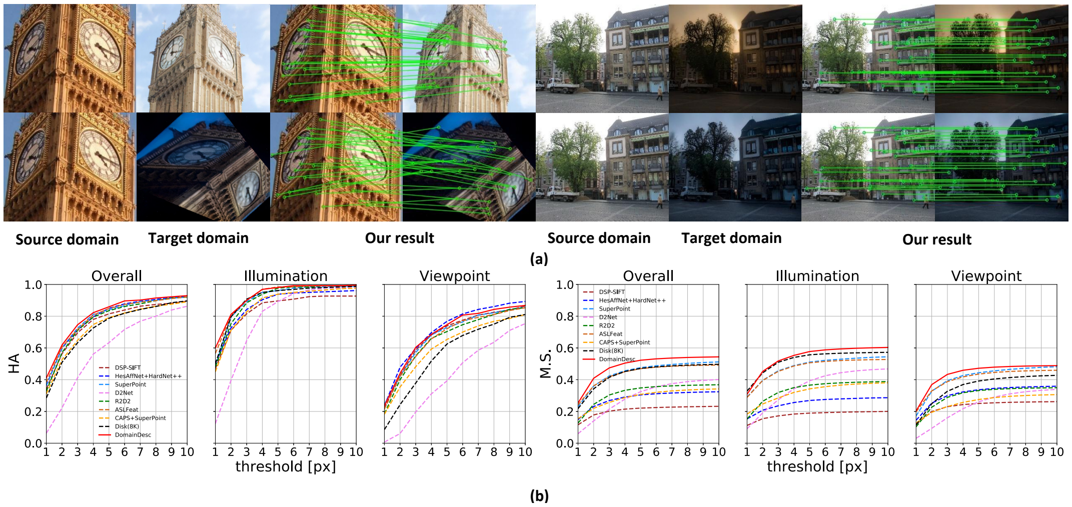


Fig. 1. Matching results of our method. (a) Two matching results of applying our DomainDesc to image pairs composed of different source domain images and target domain images. The green lines show correct correspondences. (b) The evaluation results applying our DomainDesc on the Hpatches dataset. Compared with other state-of-the-art methods, the HA and M.S. metrics of our DomainDesc are significantly leading.

main adaptation loss to reduce the difference between the feature domains of image pairs, making our descriptor more robust.

- We present the pixel-level cross-domain consistency loss to compensate for the inconsistency between the descriptors corresponding to the keypoints at the pixel, in turn to enhance the invariance of our descriptor.
- We give a cross-domain dataset generation strategy to quickly construct diversified domain dataset for our training, and implement comprehensive experiments in both image matching and visual localization tasks, validating that our method outperforms the state-of-the-art methods.

2. PROPOSED METHOD

Our **Desc**riptor with **Domain** adaptation (DomainDesc) employs a fully convolutional network, which takes Resnet18 [13] as the encoder to extract the global domain invariant features, and then uses a structure similar to FPN as the decoder to output 128-dimensional dense descriptors(Fig. 2(a)). For each input image, the encoder generates multiple scale feature maps with dimensions 64, 128, 256, and 512 respectively. To make our DomainDesc more robust, we carefully design the feature-level domain adaptation loss and the pixel-level cross-domain consistency loss, detailed in section 2.1 and section 2.2 respectively. Besides, we give the strategy of generating cross-domain datasets and correspondences supervision for the training of our DomainDesc in section 2.4.

2.1. Feature-level Domain Adaptation Supervision

As shown in Fig. 2(b), each image pair includes a source domain image I_S and a transformed target domain image I_T. As the disharmony between the feature domains of I_S and I_T can bring some interference to the descriptor generation, we design a feature-level method to align the distribution between the high-level feature maps of I_S and I_T to reduce this disharmony for improving the robustness of the descriptor.

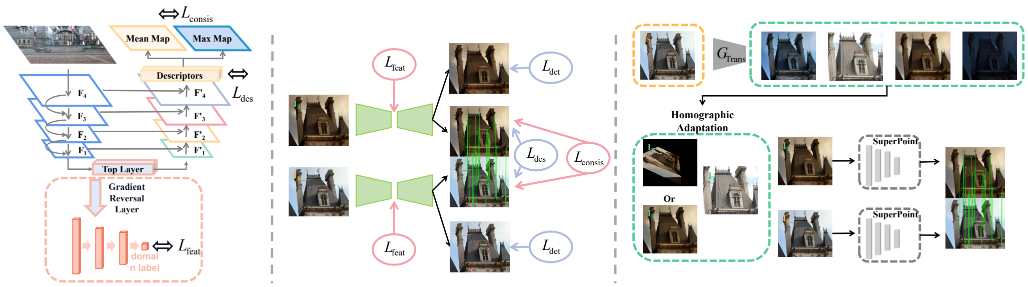
Specifically, we apply a gradient reversal layer to implement the domain adversarial learning. Unlike DANN [14], we use three fully connected layers and our feature-level domain loss function to enhance the performance of the domain classifier that distinguishing the global domain invariant features $(F(I_S), F(I_T))$ of the two images. Inspired by [15], we minimize the Shannon entropy predicted by the target sample to make the model produce high-confidence predictions:

$$H(t_i) = (-t_i log(t_i) - (1 - t_i) log(1 - t_i). \tag{1}$$

And we define our feature-level domain loss as:

$$L_{feat}(F(I_S), F(I_T)) = \\ \left[\frac{1}{N} \sum_{i=1}^{N} (-l_i log(t_i) - (1 - l_i) log(1 - t_i)) + tanh(H(t_i) \right]^2 . \tag{2}$$

where l_i is the domain category label of the image, $l_i = 1$ denotes the label of I_T, and $l_i = 0$ denotes the label of the I_S. t_i is the domain prediction score of I_T.



(a) Network architecture of our DomainDesc

(b) Training with DomainDesc

(c) Cross-domain data Labeling

Fig. 2. The overview of our DomainDesc.(a) The architecture of our DomainDesc. (b) We conduct training by applying feature-level domain adaptation loss and pixel-level cross-domain consistency loss on our DomainDesc. (c) We use the translation network and homographic adaptation to obtain a cross-domain dataset with labels.

2.2. Pixel-level Cross-domain Consistency Supervision

Although I_T and I_S may be different in appearance or style, the descriptors corresponding to the keypoints of both images should be close enough. Therefore, we present a cross-domain consistency loss to strengthen this premise and make the descriptors of two domain-specific images more consistent.

For the dense descriptors (D, D') finally obtained in Fig. 2, we calculate the mean and maximum value of each descriptor, so each image pair can get a pair of descriptor mean maps $(\bar{D}, \bar{D}')$ and max map(D^{max}, D'^{max}). We use the sampled descriptor mean map and max map as the input, and we define our pixel-level cross-domain consistency loss as:

$$L_{consis}(D,D') = \frac{1}{n^2}\sum_{i=1}^{n^2}\left[\left|\bar{d}_i - \bar{d'}_i\right| + \left|d_i^{max} - d'^{max}_i\right|\right]^2. \qquad (3)$$

where n^2 is the number of descriptors sampled during descriptor supervision, $(\bar{d}_i, \bar{d}_i)$ and (d_i^{max}, d_i^{max}) are the mean and maximum of descriptors sampled in $(\bar{D}, \bar{D}')$ and (D^{max}, D'^{max}) respectively.

2.3. Total Loss

The total loss integrates both our feature-level domain loss and our pixel-level cross-domain consistency loss, as well as the triplet loss [6] which is conducive to distinguish outliers. Given an image pair and its corresponding dense descriptor sets (D, D'), the positive distance p_i and the negative distance n_i between the descriptor $d_i \in D$ and the descriptor $d'_i \in D'$ are defined as:

$$p_i = ||d_i - d'_i||_2, \quad n_i = \underset{k\in 1\ldots N, k\neq i}{min}(||d_i - d'_k||_2). \qquad (4)$$

and the triplet descriptor loss can be defined as:

$$L_{des} = \max(0, p_i - n_i + 1). \qquad (5)$$

In sum, we define the total loss as:

$$L_{total} = L_{feat} + L_{consis} + L_{des}. \qquad (6)$$

2.4. Training via Novel Cross-domain Data

To construct a comprehensive cross-domain dataset for training, we use 4479 images of the Aachen-Day-Night dataset and 11800 image pairs selected in the MegaDepth as source domain images I_S. For generating richer domain changes and realistic images, we employ HIDM[16] to translate the I_S to the morning, noon, dusk and evening image domains respectively as shown in Fig. 2(c). Meanwhile, we perform homomorphic adaptation [7] on all translated images with a certain probability in order to increase the viewpoint change. A source domain image I_S corresponds to the result of homomorphic adaptation processing or an image with only domain transformation. If I_S comes from MegaDepth, it also corresponds to an image with ground truth correspondence. We regard all these corresponding images as target domain images I_T, and we randomly select 32558 image pairs as our cross-domain training dataset.

The corresponding relationship between them is already given for image pairs in MegaDepth. For other image pairs synthesized using random homomorphic adaptation, we divide one of the images into 20 × 20 grids, and each grid randomly samples a point uniformly, and the corresponding point is obtained based on the homography.

3. EXPERIMENTS

3.1. Evaluation on Image Matching

Evaluation Dataset. The HPatches dataset is a common evaluation dataset for image matching. Following the protocol of SuperPoint [7] and D2Net [8], we use 108 sequence scenes with viewpoint or illumination changes for a fair comparison.
Comparisons. We use three standard metrics for evaluation: Homography Accuracy (HA) for the ratio of the correct es-

HPatches dataset										
		DSP-SIFT	**HesAffNet +HardNet++**	**SuperPoint**	**D2Net**	**R2D2**	**ASLFeat**	**CAPS+SuperPoint**	**Disk(8K)**	**DomainDesc**
Illumination	HA	0.815	0.842	0.892	0.646	0.885	0.900	0.827	0.907	**0.915**
	Precision	0.559	0.505	0.694	0.527	0.715	0.775	0.691	**0.832**	0.756
	M.S.	0.172	0.236	0.456	0.274	0.320	0.452	0.285	0.510	**0.520**
Viewpoint	HA	0.586	0.596	0.571	0.196	0.550	0.546	0.482	0.378	**0.604**
	Precision	0.660	0.620	0.639	0.356	0.693	0.708	0.617	0.619	**0.706**
	M.S.	0.230	0.299	0.394	0.160	0.289	0.388	0.231	0.327	**0.435**

Table 1. Evaluation results on the HPatches dataset.

timation of homography, Precise for the average matching accuracy with a threshold of 3 pixels by default, Match Score (M.S.) following the definition in [9]. As shown in Fig. 1 and Table 1, our DomainDesc has reached the excellent performance under the overall comparison, and is significantly ahead of other SOTAs on HA and M.S., because our feature-level domain adaptation loss and pixel-level cross-domain consistency loss can compensate for the inconsistency between descriptors of corresponding keypoints to a large extent.

3.2. Evaluation on Visual Localization

Evaluation Dataset. The Aachen-Day-Night dataset [17] provides query images of the Aachen city taken during the day and night. We use the public evaluation benchmark [2] for evaluation, which generates the percentage of successfully positioned images within three tolerances.

Ablation study. Our baseline model employs a network with Resnet18 [13] as the encoder and FPN-like structure as the decoder, and uses 23600 image pairs of the MegaDepth dataset described in section 2.4 for training. To validate the advantage of our DomainDesc, we apply the same keypoints of SuperPoint to all the methods in our ablation experiment. From the last 4 rows of Table 2, we can see that our DomainDesc are superior to the baseline in all metrics.

Comparisons. We compare our DomainDesc with advanced local descriptors and other deep learning methods shown in Table 2. Our DomainDesc achieved the highest accuracy rates of 73.3 and 86.9 in the tolerance under (0.25m, 2° and $0.5m, 5^\circ$), which demonstrates that our descriptor based on domain adaptation has significantly improved the performance of the descriptor. Although CAPS [11] and D2Net [8] gain the advantage under the loosest tolerance, it benefits from the use of longer dimensional descriptors (256 and 512). In contrast, our DomainDesc has only 128-dimensional descriptors and we train our model on a dataset that is only 1/20 of D2Net. Moreover, our DomainDesc uses fewer keypoints for each image with a lightweight design structure, so it achieves a faster real-time speed (29 fps on HPatches with the image size 480×640 under TitanV). Competitive results under the three error thresholds demonstrate that our DomainDesc can achieve a balance between the descriptor invariance and distinguishability, making it effectively increase the accuracy of challenging visual localization tasks.

Table 2. Evaluation on Aachen Day-Night v1.1: + Cross-domain data: using the cross-domain dataset in section 2.4; + Feature loss: augmenting our feature-level domain adaptation loss further; + Consis loss: augmenting our pixel-level cross-domain consistency loss further (our DomainDesc).

Aachen Day-Night v1.1 dataset					
Method	Kpts	Dim	Correctly localized queries		
			0.25m,2°	*0.5m,5°*	*5m,10°*
ROOT-SIFT [18]	11K	128	53.4	62.3	72.3
DSP-SIFT [19]	11K	128	40.3	47.6	51.3
SuperPoint [7]	7K	256	68.1	85.9	94.8
D2Net [8]	14K	512	67.0	86.4	97.4
R2D2 [9]	10K	128	70.7	85.3	96.9
ASLFeat [10]	10K	128	71.2	85.9	96.9
CAPS + SuperPoint [11]	7K	256	71.2	86.4	**97.9**
DISK [12]	10K	128	72.8	86.4	97.4
Baseline	7K	128	70.2	84.3	95.3
+ Cross-domain data	7K	128	70.7	86.4	96.9
+ Feature loss	7K	128	72.8	86.4	96.3
+ Consis loss (DomainDesc)	7K	128	**73.3**	**86.9**	96.9

4. CONCLUSION

By introducing domain adaptation for local feature description for the first time, we propose a novel local descriptor **DomainDesc** to adapt to multiple scenarios. We design the feature-level domain adaptation loss in order to align the high-level features of images in different domains, and give the pixel-level cross-domain consistency loss to reduce the inconsistency between the descriptors at the pixel level. As both of our losses punish the variant features which may impair the distinguish performance, the triplet loss is combined to make our descriptor achieve a balance of distinguishability and invariance. Besides, we adopt a cross-domain dataset generation strategy to construct domain dataset for training. Experiments have demonstrated the superiority of our method in image matching and visual localization tasks.

5. ACKNOWLEDGEMENTS

This work was supported in part by the National Natural Science Foundation of China (Nos. $U21A20515$, 61971418, $U2003109$, 62171321, 62071157, 62162044), the Open Research Fund of Key Laboratory of Space Utilization, Chinese Academy of Sciences (No. LSU-KFJJ-2021-05), Open Research Projects of Zhejiang Lab (No.2021KE0AB07) and Project TC210H00L/42.

6. REFERENCES

[1] Vassileios Balntas, Karel Lenc, Andrea Vedaldi, and Krystian Mikolajczyk, "Hpatches: A benchmark and evaluation of handcrafted and learned local descriptors," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5173–5182.

[2] Torsten Sattler, Will Maddern, Carl Toft, Akihiko Torii, Lars Hammarstrand, Erik Stenborg, Daniel Safari, Masatoshi Okutomi, Marc Pollefeys, Josef Sivic, et al., "Benchmarking 6dof outdoor visual localization in changing conditions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8601–8610.

[3] Jun Wen, Risheng Liu, Nenggan Zheng, Qian Zheng, Zhefeng Gong, and Junsong Yuan, "Exploiting local feature patterns for unsupervised domain adaptation," in *Proceedings of the AAAI conference on artificial intelligence*, 2019, vol. 33, pp. 5401–5408.

[4] Zhe Xu, Jie Luo, Jiangpeng Yan, Ritvik Pulya, Xiu Li, William Wells, and Jayender Jagadeesan, "Adversarial uni-and multi-modal stream networks for multimodal image registration," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020, pp. 222–232.

[5] David G Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[6] Anastasiia Mishchuk, Dmytro Mishkin, Filip Radenovic, and Jiri Matas, "Working hard to know your neighbor's margins: Local descriptor learning loss," in *Advances in Neural Information Processing Systems*, 2017, pp. 4826–4837.

[7] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich, "Superpoint: Self-supervised interest point detection and description," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 224–236.

[8] Mihai Dusmanu, Ignacio Rocco, Tomas Pajdla, Marc Pollefeys, Josef Sivic, Akihiko Torii, and Torsten Sattler, "D2-net: A trainable cnn for joint detection and description of local features," in *CVPR 2019*, 2019.

[9] Jerome Revaud, Philippe Weinzaepfel, César De Souza, Noe Pion, Gabriela Csurka, Yohann Cabon, and Martin Humenberger, "R2d2: Repeatable and reliable detector and descriptor," *arXiv preprint arXiv:1906.06195*, 2019.

[10] Zixin Luo, Lei Zhou, Xuyang Bai, Hongkai Chen, Jiahui Zhang, Yao Yao, Shiwei Li, Tian Fang, and Long Quan, "Aslfeat: Learning local features of accurate shape and localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6589–6598.

[11] Qianqian Wang, Xiaowei Zhou, Bharath Hariharan, and Noah Snavely, "Learning feature descriptors using camera pose supervision," in *Proc. European Conference on Computer Vision (ECCV)*, 2020.

[12] Michał Tyszkiewicz, Pascal Fua, and Eduard Trulls, "Disk: Learning local features with policy gradient," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[14] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky, "Domain-adversarial training of neural networks," *The journal of machine learning research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[15] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez, "Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2517–2526.

[16] Ivan Anokhin, Pavel Solovev, Denis Korzhenkov, Alexey Kharlamov, Taras Khakhulin, Alexey Silvestrov, Sergey Nikolenko, Victor Lempitsky, and Gleb Sterkin, "High-resolution daytime translation without domain labels," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[17] Torsten Sattler, Tobias Weyand, Bastian Leibe, and Leif Kobbelt, "Image retrieval for image-based localization revisited.," in *BMVC*, 2012.

[18] Relja Arandjelović and Andrew Zisserman, "Three things everyone should know to improve object retrieval," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 2911–2918.

[19] Jingming Dong and Stefano Soatto, "Domain-size pooling in local descriptors: Dsp-sift," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 5097–5106.