Signal, Image and Video Processing (2023) 17:865–872
https://doi.org/10.1007/s11760-021-01914-5

ORIGINAL PAPER

VEDesc: vertex-edge constraint on local learned descriptors

Jianhua Yin[1] · Longzhen Zhu[1] · Yang Bai[1] · Zhenyu He[1,2]

Received: 30 December 2020 / Revised: 13 March 2021 / Accepted: 12 April 2021 / Published online: 17 May 2021

Abstract
To improve the performance of local learned descriptors, many researchers pay primary attention to the triplet loss network. As expected, it is useful to achieve state-of-the-art performance on various datasets. However, these local learned descriptors suffer from the inconsistency problem without considering the relationship between two descriptors in a patch. Consequently, the problem causes the irregular spatial distribution of local learned descriptors. In this paper, we propose a neat method to overcome the above inconsistency problem. The core idea is to design a triplet loss function of vertex-edge constraint (VEC), which takes the correlation between two descriptors of a patch into account. Furthermore, to minimize the non-matching descriptors' influence, we propose an exponential algorithm to reduce the difference between the long and short sides. The competitive performance against state-of-the-art methods on various datasets demonstrates the effectiveness of the proposed method.

Keywords Local learned descriptors · Inconsistency issue · The triplet loss function

1 Introduction

Image matching, which is a fundamental problem in computer vision, has been widely used in numerous fields, including structure from motion [1], wide-baseline stereo [2,3], 3D reconstruction [4] and simultaneous localization and mapping (SLAM) [5,6]. Generally, a patch-based matching method includes extracting the feature points and matching their feature descriptors. The stable descriptors with properties of scale and rotation invariance are crucial for correcting image matching. Traditional methods, including SIFT [7] and SURF [8], have been proven to be effective in various applications. Meanwhile, to accelerate matching and reduce storage, several algorithms [9–11] adopt the binary descriptors instead of the float descriptors. Nora Al-Garaawi et al. [12] extract some binary face descriptors with the BRIEF algorithm to achieve well performance on automatic facial expression recognition.

However, handcrafted descriptors are limited in terms of robustness and precision due to the lack of high-level structural information. Recently, with the development of convolutional neural networks (CNNs), the research hotspot focuses on the learned descriptors [13–17] instead of the handcrafted descriptors. These learned descriptors achieve much higher performance than the handcrafted ones in various areas, including image matching and patch verification. Specifically, To improve the performance of learned descriptors, the triplet loss [14] encourages that the distance between matching descriptors is less than the distance between non-matching ones. Inspired by the triplet loss, recent works [16, 17] try to minimize the distance between matching descriptors and maximize the distance between non-matching descriptors. These experimental results demonstrate the triplet loss is effective on various datasets, improving the robustness of learned descriptors in extreme conditions, including weather, season, illumination, and distortion.

Although these triplet loss methods achieve impressive performance, they suffer from an inconsistency issue, as shown in Fig. 1a. Specifically, Fig. 1 shows the Euclidean distance of four image patches involving two matching pairs. The symbols a_i and p_i $(i = 1,2)$ refer to descriptors of anchor

✉ Zhenyu He
zhenyuhe@hit.edu.cn

Jianhua Yin
yinjianhua@hit.edu.cn

Longzhen Zhu
19S051027@stu.hit.edu.cn

Yang Bai
19s151126@stu.hit.edu.cn

[1] School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, Shenzhen, China

[2] Peng Cheng Laboratory, Shenzhen, China

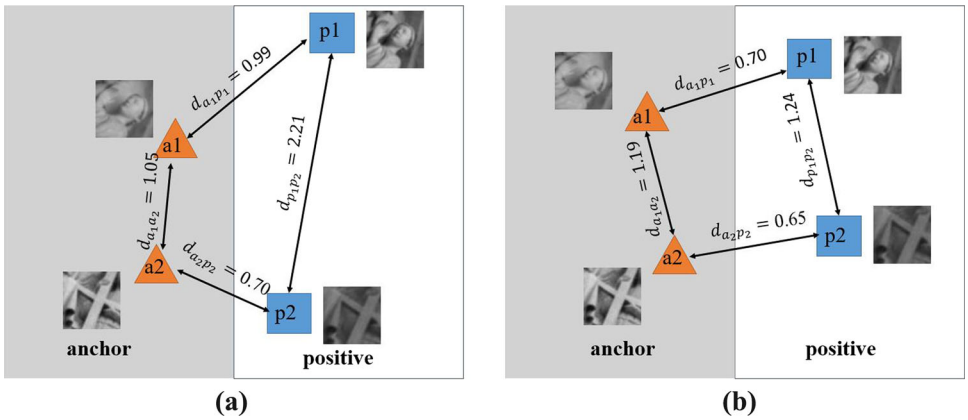

Fig. 1 Comparison between the descriptors w/wo the edge constraint. **a** A previous method with vertex constraint and **b** our method with VEC. The descriptors a_i and p_i $(i = 1, 2)$ are detected from anchor patches and positive patches. The parameter d denotes the Euclidean distance between two descriptors. The different value between $d_{a_1a_2}$ and $d_{p_1p_2}$ from 1.16 (**a**) to 0.05 (**b**). Furthermore, our method minimizes the difference between $d_{a_1p_1}$ and $d_{a_2p_2}$ from 0.29 to 0.05)

patches and positive patches, and d represents the Euclidean distance between two descriptors. If a_1-p_1 and a_2-p_2 is two matching descriptors pairs, $d_{a_1a_2}$ is equal to $d_{p_1p_2}$. However, the difference between $d_{a_1a_2}$ and $d_{p_1p_2}$ obtained in the traditional method is relatively large, which leads to the inconsistency issue. As presented in Fig. 1a, the distance ($d_{a_1a_2} = 1.05$) between descriptors a_1 and a_2 in anchor patches is significantly smaller than that ($d_{p_1p_2} = 2.21$) between their descriptors p_1 and p_2 in positive patches. The learned descriptors suffer from the inconsistency issue in terms of representing the difference between two image patches. This issue would cause irregular spatial relationships of descriptors and further reduce matching accuracy, even though the distance between matching descriptors is smaller than the distance between non-matching ones. The reason for the issue is that these works are less effective in exploiting the interrelation between the two descriptors of anchor (or positive) patches in the traditional triplet loss function.

Generally, vertex information contains several properties, and an edge represents the relationship between two vertexes. Descriptors and the distance between two descriptors can be naturally illustrated as a graph, in which the vertex represents the descriptor, and the edge denotes the relationship between two descriptors. To solve this inconsistency problem, we introduce the vertex-edge constraint (VEC) to the triplet loss function: the vertex constraint denotes the distance between two vertexes in different images; the edge constraint represents the distance between two descriptors in a patch. To exploit the vertex constraint, we propose a strategy to increase the distance between two non-matching descriptors and decrease the distance between matching ones. Furthermore, we encourage reducing the difference between the longer and shorter sides in the edge constraint. Fig. 1b shows the VEC algorithm's result, which demonstrates that our method effectively reduces the difference between opposite sides and further promotes matching accuracy of descriptors.

We summarize the main contributions of the work as follows:

(1) We introduce VEC to the triplet loss function (denoted as VEC-loss), containing vertex constraint and edge constraint.
(2) To facilitate regular distribution of descriptors in high dimensional Euclidean space, we design an exponential algorithm to reduce the difference between the long and short sides in the image matching pair.
(3) To avoid the influence of the distance between non-matching descriptors, we design an strategy that enlarges the distance of two descriptors.

2 Related work

Descriptors' detection is a critical technique for image pair matching. We briefly review descriptors detected by the handcrafted method and the learned method in section 2.1. Furthermore, we discuss some practical algorithms with metric learning to improve the performance of learned descriptors in section 2.2.

2.1 Descriptors detecting

The handcrafted method and the learned descriptor method are two main methods to detect local feature descriptors. Early works on the local descriptor mainly focused on the handcrafted method with gradient filters and intensity comparisons. SIFT [7], which computes histograms with gradient ltering approach, possibly is the most classical and popular handcrafted method. These handcrafted methods, including SIFT [7], SURF [8], BRIEF [9], DSP-SIFT [18] and PCA-SIFT [19], have limited in several aspects, such as unable to get semantic segmentation information and weak robustness in image classification.

With the emergence of the CNNs, more researches focus on the learned method instead of the traditional handcrafted method in recent years. The MatchNet [20] method proposes a Siamese network, containing feature network and metric-

learning loss function, to train samples. Compared with the traditional handcrafted method, MatchNet [20] significantly improves images matching performance and demonstrates the extraordinary potential of learned descriptors. The recent methods [13,21–23] study detecting learned descriptor on the Siamese network. Kumar et al. [21] explore a central-surround two-stream network to improve image matching performance. DeepDesc [22] introduces a stochastic sampling strategy into the Siamese network to distinguish the positive and negative samples. TFeat [23] demonstrates that the triplets of training samples, with in-triplet mining of hard negatives, can improve the performance of learned descriptors. L2Net [13] designs a deeper network and produces descriptor normalized to the unit norm by L_2 distance. The architecture of L2Net is an application and foundation of learned descriptors in the later works [14,16,17]. However, the metric-learning loss function of L2Net is less effective in finding hard samples from negative and anchor samples. Unlike the above-mention methods, we propose a new triplet loss function to train the metric network.

2.2 Metric learning

Metric learning, which is a classical method depending on learned distance function, has been successfully applied to various fields including face action detection [24], tracking [25], classification [26], and information retrieval [27]. Yu et al. [28] propose a triplet loss as metric learning to significantly improve the face verification performance. Yoshida et al. [29] defines an interpretable graph metric learning (IGML) method for graph classification. RFNet [30] extracts learned descriptors from corresponding patches by computing Euclidean distance. Kumar et al. [21] propose a global loss to minimize the classification error and produce the best performance in patch matching.

In summary, metric learning is a foundation in various fields. Many previous works of metric learning focused on learning Mahalanobis distance [31,32], while much more efforts are spent on the learning vectors with Euclidean distance metrics [33,34]. Particularly, several main studies [14,16,17] , which adopt L2Net structure, redefine the improved triplet loss function with Euclidean distance metric to improve performance of learned descriptors. HardNet [14] introduces a new triplet loss into L2Net, which can minimize the distance between the matching descriptors and closest non-matching descriptors. DOAP [15] proposes optimizing average precision (OAP) to L2Net and achieves more competitive results on many datasets. ExpTLoss [16] introduces the exponential Siamese and triplet loss function into L2Net, and achieve better performance on many tasks. CDF [17] assigns a nonparametric soft margin instead of a hard margin in L2Net and improves the performance of learned descriptors. However, learned descriptors with these above methods suffer from an inconsistency issue, which can cause irregular spatial relationship of descriptors and further reduce matching accuracy. Unlike the above methods, we propose VEC to the triplet loss function, an effective approach in exploiting the mutual relationship between adjacent descriptors of a patch.

3 Methodology

In this section, we firstly review the traditional triplet loss function and then propose VEC-loss to detect learned descriptors. Furthermore, to achieve regular distribution of descriptors in Euclidean space, we exploit a new strategy to reduce the opposite sides' difference in image matching pair.

3.1 Graph information

Generally, a graph is composed of vertexes and edges. The edge indicates the relationship between the two vertexes. Descriptors and the distance between descriptors can be naturally illustrated as a graph, in which the vertex represents the descriptor, and the edge denotes the distance between two descriptors. Supposedly, anchor sample set $\mathbf{G_A} = (\mathbf{V_A}, \mathbf{E_A})$ and positive sample set $\mathbf{G_P} = (\mathbf{V_P}, \mathbf{E_P})$ are generated, where $\mathbf{V}$ stands for the descriptor and $\mathbf{E}$ for the distance.

3.2 Problem formulation

To improve the performance of local learned descriptors, recent researchers focus on the triplet loss function. Supposedly, the spatial relationship between two adjacent descriptors is linearly invariant and rigid, preserving the distance between corresponding descriptors. Thus, the triplet loss function S can be formulated as:

$$S = \frac{1}{N}\sum_{i=0}^{N}\max(0, \alpha + F_i^+(a_i, p_i) - F_i^-(a_i, p_j)), \quad (1)$$

where α is the margin, $F_i^+(a_i, p_i)$ represents the distance between matching descriptors in positive sample and $F_i^-(a_i, p_j)$ denotes the distance between non-matching descriptors in negative sample. In our implementation, the margin parameter α equals to 1.0, which has been demostrated to achieve good performance in HardNet [14].

3.3 Vertex constraint

Due to the promising performance of Euclidean distance in images matching, this metric has been widely used for local learned descriptors, including L2Net [13], RFNet [30], HardNet [14]. In this section, we introduce L_2 pairwise distance

into the vertex-constraint function. The loss function evaluates the distance between pair descriptors from the anchor and positive patches, which are as the input of L2Net [13].

Supposedly, we extract a batch $\mathbf{X} = \{\mathbf{V_A}, \mathbf{V_P}\}$ of matching local patches, where $\mathbf{V_A} = \{a_1, a_2, ..., a_k\}$ stands for the descriptors of the anchor patches and $\mathbf{V_P} = \{p_1, p_2, ..., p_k\}$ represents the descriptors of the positive patches, and k is the batch size. We adopt the Euclidean distance to evaluate the distance between two descriptors a_i and p_j. Furthermore, the descriptor vectors a_i and p_j should be unit-length ($||a_i||_2 = 1$) to reduce the computational cost. As such, the Euclidean distance between the unit-length descriptor vectors can be computed as following:

$$d(a_i, p_j) = \sqrt{2 - 2a_i p_j}, \tag{2}$$

where $d(a_i, p_j)$ refers to the vertex-constraint function F_{vertex}.

To reduce the distance between matching descriptors and enlarge the distance between the closest non-matching descriptors, we propose the vertex constraint to find hard negatives. It can be formulated as:

$$F_i^-(a_i, p_j) = \min\left(d(a_i, p_{j_{min}}), d(a_{i_{min}}, p_j)\right), \tag{3}$$

where d is the Euclidean distance between descriptor pair, $p_{j_{min}}$ is the closest non-matching positive descriptor to a_i, and $a_{i_{min}}$ is the closest non-matching anchor descriptor to p_j, respectively.

3.4 Edge constraint

In addition to the vertex information, a graph contains the edge information indicating the relationship between two vertexes. Similar to the graph information, an image patch has descriptors and their adjacent edge. In this section, we propose an edge constraint to reduce the difference between the long and short sides.

We calculate the Euclidean distance between two different unit-length descriptor vectors in a patch to construct the edge equation. Furthermore, to promote the spatial stability of descriptors, we design an exponential algorithm to lengthen the shorter side and shorten the longer side. To reduce the computational cost, we propose a normalized method to keep its value within the range between 0 and 1. In this way, the edge constraint function can be formulated as:

$$F_{edge}(i, j) = 1 - \exp\left(-\left(\frac{d(a_i, a_j) - d(p_i, p_j)}{(d(a_i, a_j) + d(p_i, p_j))/2}\right)^2\right), \tag{4}$$

where $d(a_i, a_j)$ is the Euclidean distance between two descriptors of anchor patch and $d(p_i, p_j)$ is the Euclidean distance between two descriptors of positive patch.

3.5 Dynamic for loss

This section proposes a strategy to find the positives where the diagonal elements correspond to the matching descriptors pair distance. In summary, we redefine the positive loss function as following:

$$F_i^+(a_i, p_i) = \lambda F_{vertex}(i, i) + (1 - \lambda) F_{edge}(i, i), \tag{5}$$

where λ is a weight parameter used to balance the vertex-constraint and the edge-constraint loss functions. The parameter $\lambda = 1$ (or $\lambda = 0$) indicates that the positive loss function only consider the vertex constraint (or the edge constraint). When λ is greater than 0 and less than 1, the positive loss function takes the vertex constraint and the edge constraint into account.

4 Experiments

The main contribution of this paper is to introduce VEC into the triplet loss function. To measure performance of our method, we experiment with four benchmarks: UBC PhotoTourism [35], HPatches [36], W1BS benchmark [37] and Oxford Affine benchmark [38]. As a classical and popular patch-based benchmark, UBC PhotoTourism effectively evaluates descriptor performance on patch verification tasks. As a comprehensive and complicated dataset, HPatches is used to evaluate learned descriptors' performance by three different tasks (Patch Verification, Image Matching, and Patch Retrieval). As a novel dataset of ground-truthed image pairs, W1BS is used to evaluate two images' correspondences in a wide baseline stereo. The Oxford Affine benchmark, containing various distorted images, is helping to improve the stability of descriptors.

4.1 Implementations

Similar to the previous study, we adopt the training settings to ensure that the VEC-loss function is the main factor for better performance in the following datasets. For these experiments, we introduce VEC-loss function into HardNet [14] and CDF [17], denoted as HN+-VEC and CDF-VEC. The HN+-VEC or CDF-VEC method adopts the algorithm VEC to replace the triplet loss function in HardNet or CDF, and the other parts remain unchanged. The network consists of seven convolutional layers and is regularized with Batch Normalization and Dropout. We adopt the UBC PhotoTourism dataset [40] as training data. The dataset has three subsets, known as Liberty, Yosemite, and Notredame. We propose one subset for training data and the other two subsets for testing data. HardNet [14] and CDF [17] adopt L2Net as the descriptors extractor, where the input image size is required

Table 1 Evaluation on the UBC PhotoTourism benchmark, demonstrating that the learned descriptors with our method outperform state of the art ($\lambda = 0.85$)

Test	Liberty		Notredame		Yosemite		Mean
Train	Notredame	Yosemite	Liberty	Yosemite	Liberty	Notredame	FPR95
SIFT(2004)	29.84		22.53		27.29		26.55
DeepDesc(2015)	10.9		4.4		5.69		7.00
L2-Net+(2017)	2.36	4.70	0.72	1.29	2.57	1.71	2.23
CSL2-Net+(2017)	1.71	3.87	0.56	1.09	2.07	1.30	1.76
HardNet(2018)	1.47	2.67	0.62	0.88	2.14	1.65	1.57
HardNet+(2018)	1.49	2.51	0.53	0.78	1.96	1.84	1.52
DOAP+ (2018)	1.54	2.62	0.43	0.87	2.00	1.21	1.45
DOAP-ST+(2018)	1.47	2.29	0.39	0.78	1.98	1.35	1.38
ExpTLoss(2019)	1.16	2.01	0.47	0.67	1.32	1.10	1.12
CDF(2019)	1.21	2.01	0.39	0.68	1.51	1.29	1.18
HN+-VEC(ours)	1.31	1.93	0.38	0.68	1.50	1.17	1.16
CDF-VEC(ours)	**1.13**	**1.86**	**0.31**	**0.64**	**1.25**	**1.08**	**1.05**

Numbers shown are FPR95(%). The lower FPR95 indicates the better performance of learned descriptors. "+" denotes training with data augmentation. The best results are in **bold**

to be 32x32. Because the size of image patches in the UBC PhotoTourism is 64x64, we adopt a downsample method to obtain the patch size of 32x32 as the input of the L2Net [13]. As the input data augmentation, the image patches are flipped randomly and rotated by 90,180 or 270 degrees. Similar to HardNet [14] and CDF [17], we set momentum and weight of stochastic gradient descent (SGD) to 0.9 and 10^{-4}. Furthermore, the learning rate decays linearly from 0.1 to 0. To match the result of HardNet and CDF, we set the batch size to 1024.

4.2 UBC PhotoTourism

As one of the Brown databases, the UBC PhotoTourism dataset has been widely used to detect local learned descriptors. It contains three subsets: Liberty, Notredame, and Yosemite. Each subgroup contains more than 400k normalized 64x64 patches. These patches are reoriented with Difference-of-Gaussians (DOG) keypoints, which are extracted from 3D reconstruction scenes. The dataset comprises 100k matching and non-matching pairs. Furthermore, we assign one subset for training and the other two subsets for testing. For example, we adopt Liberty as the training dataset and the other two subsets as test datasets. For a fair comparison and without losing of generality, we keep the input approach of dataset consistent [14,17]. We propose the false positive rate of 0.95 accurate positive recall (FPR95) to evaluate the learned descriptors' performance. The lower FPR95 indicates a better understanding.

We introduce VEC to the HardNet [14] which is the first to adopt the triplet loss for learned descriptor, and CDF [17] which is the state-of-the-art method of learned descriptor. To verify performance of our method, we compare with a collection of a handcrafted method (SIFT [7]) and some learned methods (DeepDesc [22], L2-Net [13], HardNet [14], DOAP [15], ExpTLoss [16] and CDF [17]). The results of all methods are shown in Table 1.

As presented in Table 1, the mean FPR95 of HN+-VEC and CDF-VEC is less than the two original methods HardNet+ and CDF. The mean FPR95 of HN+-VEC is reduced from 1.52 to 1.16, and the mean FPR95 of CDF-VEC is decreased from 1.18 to 1.05. Furthermore, the learned descriptor with our method achieves the best performance in these subtasks. Compared with these state-of-the-art methods, the learned descriptors' version of CDF-VEC is the best (the average of FPR95 = 1.05). Unlike the other methods, our approach considers the interrelation between the adjacent descriptors in the same patch. The UBC PhotoTourism is a testament to the generalization of our algorithm.

4.3 HPatches

HPatches dataset contains more than 1.5 million patches, which are extracted from 116 sequences. The dataset consists of 59 sequences affected by geometric deformation and 57 sequences affected by illumination. In these images, the feature keypoints are detected by DoG, Hessian, and Harris detectors. HPatches dataset defines three evaluation subtasks: Patch Verification, Patch Retrieval, and Image Matching. In each subtask, the extracted patches can be divided into three geometric noise levels: Easy, Hard, and Tough. We adopt the mean average precision(mAP) to evaluate learned descriptors' performance in the dataset. The higher mAP indicates a better understanding.

For a fair comparison and without losing of generality, we use the model trained on the Liberty subset to generate learned descriptors in the HPatches. We compare our method with a collection of methods, such as SIFT [7], DOAP [15],

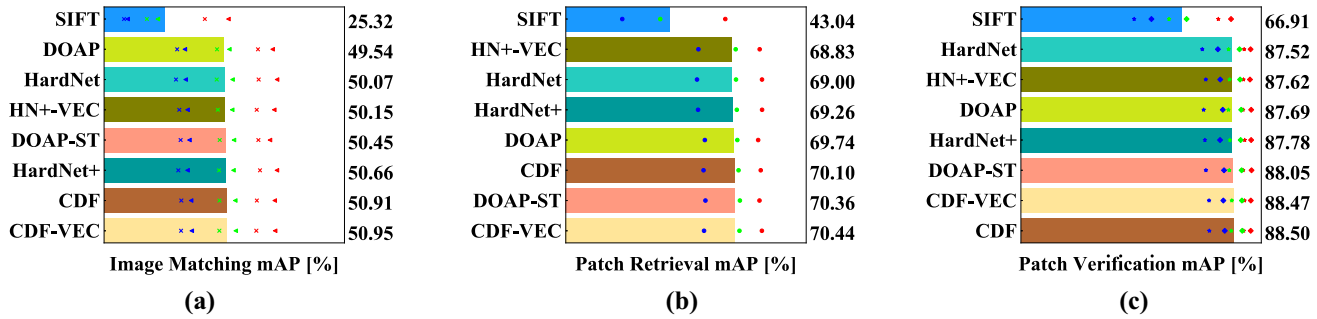

Fig. 2 Evaluating on the HPatches benchmark. In the middle figure, three colors of points indicate three levels of subtask: easy (red), hard (green), tough (blue). In the right figure, DIFFSEQ (⧫) and SAMESEQ (⋆) represent the source of negative examples in the verification task. In the left figure, ILLUM (x) and VIEWPT (◄) indicate the influence of illumination and viewpoint changes in the matching study

HardNet [14], CDF [17]. The results of all methods are shown in Fig. 2.

As presented in Fig. 2, our CDF-VEC method achieves promising performance with the best mAP score in task Image matching (mAP = 50.95) and task Patch Retrieval (mAP = 70.44). Furthermore, CDF-VEC is the second-best in task Patch Verification (mAP = 88.47), close to the primary method CDF. We attribute these gains to the regular spatial relative of learned descriptors, enabling the performance more robustly and accurately in three subtasks. HPatches is also a testament to the generalization of our method.

4.4 Wide baseline stereo (W1BS)

To verify the generalization and the application in these extreme conditions, we propose the VEC method to generate learned descriptors on W1BS dataset [42]. Wide baseline stereo matching considers pair images matching in many extreme conditions, such as weather, season, illumination, etc. W1BS dataset contains 40 image pairs, which can be divided into five groups according to the image acquisition factor.

Appearance (A): difference in appearance according to season or weather changes; Geometry (G): difference in object position, camera, and scale; Illumination (L): difference in wavelength, direction, and intensity of light source; Sensor(S): difference in sensor type (IR, MR); Map to photo (map2photo): object image and map image.

In the W1BS dataset, the local affine-covariant features are detected by FOCI [39], Hessian-Affine [40] and MSER [41] for the reference image pairs. Furthermore, W1BS normalizes the patch size into 41x41 to explore descriptor performance in extreme conditions. We adopt mean Area Under Curve (mAUC) to evaluate learned descriptors' performance. The larger mAUC means a better understanding.

Similar to the above experiments, we use the Liberty as a training dataset to generate learned descriptors in the W1BS dataset. We compare our method with a collection of methods, SIFT [7], HardNet [14], ExpTLoss [16], and CDF [17]. The results of all methods are shown in Fig. 3.

As presented in Fig. 3, it is not surprising to CDF-VEC achieves better performance than CDF in subsets, which is consistent with our observation on UBC PhotoTourism subsets. HN+-VEC produces better performance than HardNet+ in many subsets. Furthermore, Our approach in HardNet+ [14] performs better than the other methods, which achieves the best performance with the average mAUC score of 8.41(HN+-VEC), and CDF-VEC is the third-best(mAUC = 8.24). The rigid and regular relationship between adjacent descriptors is the leading factor of better performance. W1BS is also a testament to the generalization of our method.

4.5 Oxford affine dataset

To verify the learned descriptor generalization and robustness in the various distortion types, we test the VEC algorithm to evaluate image pairs matching performance on the Oxford Affine dataset [38]. These images on Oxford Affine dataset undergo some particular distortion types, such as JPEG compression, rotation, blur, viewpoint, and light. The dataset consists of 8 groups, which are bikes (blur), trees(blur), graf(viewpoint), wall(viewpoint), bark(rotation), boat(rotation), leuven (light), and ubc (JPEG compression). Every group contains six images in PPM or PGM format and homographs image pairs. The feature keypoints are extracted by the Harris-Affine detector, and the image patches are cropped with a magnification factor of 6. We adopt the matching score to evaluate learned descriptors' performance. The larger matching score means a better understanding.

We use the model trained on the Liberty subset to generate learned descriptors in the Oxford Affine dataset. We compare our method with a collection of methods, HardNet [14], CDF [17], Sosnet [42], and ExpTLoss [16]. The results of all methods are shown in Fig. 4.

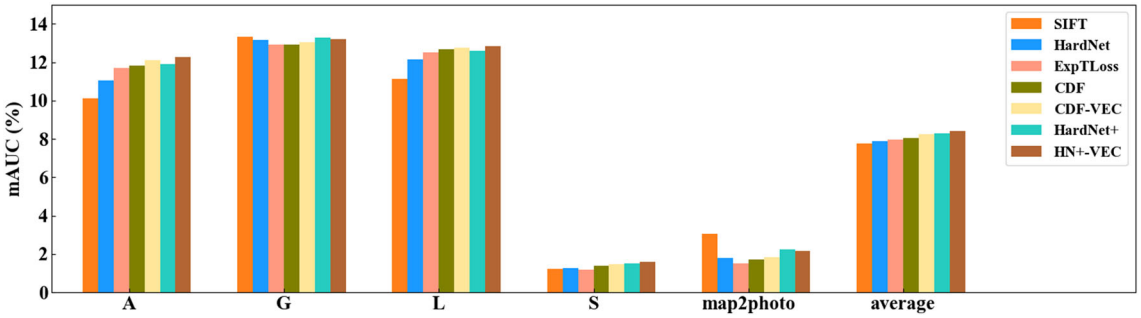

Fig. 3 Evaluating the W1BS patch dataset. W1BS dataset consists of 40 image pairs divided into 5 parts by the nuisance factor: Appearance (A), Geometry (G), Illumination (L), Sensor (S) and Map to photo. The larger mAUC indicates the better performance of learned descriptors

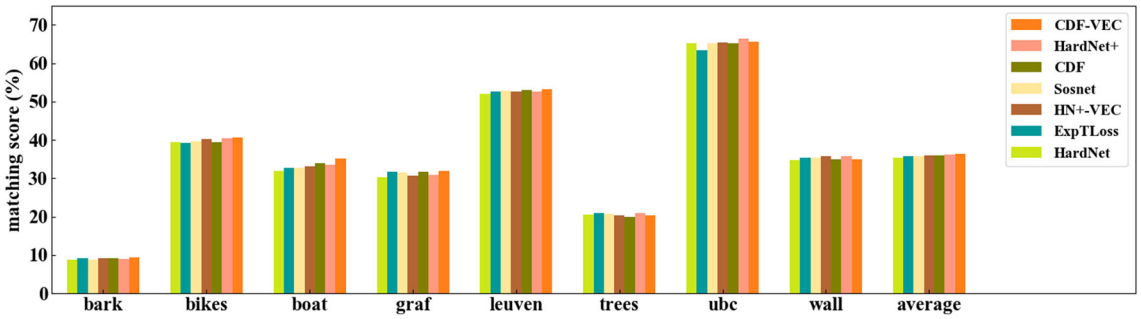

Fig. 4 Evaluating on the Oxford Affine dataset. Oxford Affine dataset consists of 8 groups: bark, bikes, boat, graf, leuven, trees, ubc, and wall. The larger matching score indicates the better performance of learned descriptors

As presented in Fig. 4, the matching performance with our method CDF-VEC achieves the best (average matching score = 36.38), compared with these state-of-the-art methods in the dataset. Furthermore, the future performance of learned descriptors with the CDF-VEC algorithm achieves state-of-art results in the other six groups, except for leuven and ubc. The better understanding shows learned descriptors with the VEC method can withstand various distortion types in the Oxford Affine dataset. We can attribute the VEC method's contribution, which considers the interrelation between the adjacent descriptors in the same patch. The Oxford Affine dataset is also a testament to the generalization of our algorithm. It is demonstrated that the performance of learned descriptors in most tasks achieves state of the art.

5 Conclusion

We introduce the VEC algorithm into the triplet loss function, inspired by learned descriptors' consistency. Furthermore, we reduce the difference between non-matching descriptors to facilitate regular distribution of learned descriptors in high-dimensional Euclidean space. To achieve better performance of stable distribution, we enlarge the distance between non-matching descriptors and reduce the distance between matching descriptors. We test the VEC method to validate the generalization and application on four datasets, including some extreme conditions. It has been demonstrated that the performance of learned descriptors in most tasks achieves state of the art.

Acknowledgements This research is supported by the National Natural Science Foundation of China (Grant No. 61672183), by the Shenzhen Research Council (Grant Nos. JCYJ20170413104556946), by Special Research project on COVID-19 Prevention and Control of Guangdong Province (2020KZDZDX1227), and by Peng Cheng Laboratory.

References

1. Zhu, S., Zhang, R., Zhou, L., Shen, T., Fang, T., Tan, P., Quan, L.: Very large-scale global sfm by distributed motion averaging. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4568–4577 (2018)
2. Azuma, R.: A survey of augmented realty, presence teleoperators and virtual. Environments **6**(4), 355–385 (1997)
3. Azuma, R.: Recent advances in augmented reality. IEEE Comput. Gr. Appl. **21**(6), 34–47 (2001)
4. Djordjevic, D., Cvetkoviç, S., Nikoliç, S.V.: An accurate method for 3D object reconstruction from unordered sparse views. Signal Image Video Process. **11**(6), 1147–1154 (2017)
5. Durrant-Whyte, H., Bailey, T.: Simultaneous localization and mapping (slam) part I. IEEE Robot. Autom. Mag. **13**(2), 99–110 (2006)
6. Bailey, T., Durrant-Whyte, H.: Simultaneous localization and mapping (slam) part II. IEEE Robot. Autom. Mag. **13**(3), 108–117 (2006)

7. David, G.L.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vision **60**(2), 91–110 (2004)
8. Bay, H., Tuytelaars, T., Van Surf, G.: Speeded up robust features. In: European Conference on Computer Vision, pp. 404–417 (2006)
9. Calonder, M., Lepetit, V., Strecha, C.: Brief: Binary robust independent elementary features. In: European Conference on Computer Vision, pp. 778–792 (2011)
10. Leutenegger, S., Siegwart, R. Y., Chli, M.: Brisk: binary robust invariant scalable keypoints. In: International Conference on Computer Vision, pp. 2548–2555 (2011)
11. Bradski, G.K., Konolige, V.R., Orb, E.R.: An efficient alternative to sift or surf. In: International Conference on Computer Vision, pp. 2564–2571 (2011)
12. Al-Garaawi, N., Wu, T., Morris, O.: Brief-based face descriptor: an application to automatic facial expression recognition (afer). Signal Image Video Process. **25**, 193 (2020)
13. Tian, Y., Fan, B., Wu, F.: L2-net: Deep learning of discriminative patch descriptor in euclidean space. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6128–6136 (2017)
14. Mishchuk, A., Mishkin, D., Radenovic, F., Matas, J.: Working hard to know your neighbor's margins: Local descriptor learning loss. In: Neural Information Processing Systems, pp. 4829–4840 (2017)
15. He, K., Yan, L., Stan, S.: Local descriptors optimized for average precision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 596–605 (2018)
16. Wang, S., Li, Y., Liang, X., Quan, D., Yang, B., Wei, S., Jiao, L.: Better and faster: exponential loss for image patch matching. In: International Conference on Computer Vision, pp. 4811–4820 (2019)
17. Zhang, L., Szymon, R.: Learning local descriptors with a cdf-based dynamic soft margin. In: International Conference on Computer Vision, pp. 2969–2978 (2019)
18. Dong, J., Soatto, S.: Domain-size pooling in local descriptors: Dsp-sift. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5097–5106 (2015)
19. Ke, Y., Sukthankar, R.: Pca-sift: a more distinctive representation for local image descriptors. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 506–513 (2004)
20. Han, X., Leung, T., Jia, Y., Rahul, S., Alexander, C.B.: Matchnet: unifying feature and metric learning for patch-based matching. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3279–3286 (2015)
21. Kumar, V.B.G., Carneiro, G.R.I.: Learning local image descriptors with deep siamese and triplet convolutional networks by minimising global loss functions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5385–5394 (2016)
22. S.-S. Edgar, T. Eduard, F. Luis, K. Iasonas, F. Pascal, M.-N. Francesc, Discriminative learning of deep convolutional feature point descriptors, in: International Conference on Computer Vision, 2015, pp. 118–126
23. Vassileios, B., Edgar, R., Daniel, P., Krystian, M.: Learning local feature descriptors with triplets and shallow convolutional neural networks. In: British Machine Vision Conference, pp. 119.1–119.11 (2016)
24. Rathee, N., Ganotra, D.: An efficient approach for facial action unit intensity detection using distance metric learning based on cosine similarity. SIViP **12**(6), 1141–1148 (2018)
25. Yuan, D., Kang, W., He, Z.: Tracking: robust visual tracking with correlation filters and metric learning. Knowl-Based Syst. **2020**(195), 137 (2020)
26. Li, D., Tian, Y.: Global and local metric learning via eigenvectors. Knowl.-Based Syst. **2017**(116), 152–162 (2017)
27. Lin, H., Fu, Y., Lu, P., Gong, S., Xue, X., Jiang, Y.: Tc-net for isbir: Triplet classifcation network for instance-level sketch based image retrieval. In: In Proceedings of the 27th ACM International Conference on Multimedia, pp. 1676–1684 (2019)
28. Yu, J., Hu, C.-H., Jing, X.-Y., Feng, Y.-J.: Deep metric learning with dynamic margin hard sampling loss for face verification. SIViP **14**(4), 791–798 (2020)
29. Yoshida, T., Takeuchi, I., Karasuyama, M.: Distance metric learning for graph structured data, arXiv:2002.00727 (2020)
30. Shen, X., Wang, C., Li, X., Yu, Z., Li, J., Wen, C., Cheng, M., He, Z.: Rf-net: An end-to-end image matching network based on receptive field. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8124–8132 (2019)
31. Bar-Hillel, A., Hertz, T., Shental, N., Weinshall, D.: Learning a mahalanobis metric from equivalence constraints. J. Mach. Learn. Res. **6**(6), 937–965 (2005)
32. Globerson, A., Roweis, S.: Metric learning by collapsing classes. Neural Inf. Process. Syst. **31**, 451–458 (2005)
33. Wen, J., Xu, Y., Liu, H.: Incomplete multiview spectral clustering with adaptive graph learning. IEEE Trans. Cybern. **20**(4), 1418–1429 (2018)
34. Wen, J., Yan, K., Zhang, Z., et al.: Adaptive graph completion based incomplete multi-view clustering. IEEE Trans. Multimed. **20**, 59 (2020)
35. Brown, M., Winder, S.J.: Learning local image descriptors. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2007)
36. Balntas, V., Lenc, K., Vedaldi, A., Mikolajczyk, K.: Hpatches: a benchmark and evaluation of handcrafted and learned local descriptors. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3852–3861 (2017)
37. Mishkin, D., Matas, J., Perdoch, M., Lenc, K.: Wxbs: wide baseline stereo generalizations. In: Proceedings of the British Machine Vision Conference, pp. 12.1–12.12 (2015)
38. Bsat, M., Sim, T., Baker, S.: A performance evaluation of local descriptors. IEEE Trans. Pattern Anal. Mach. Intell. **27**(12), 1615–1618 (2003)
39. Ramnath, K., Zitnick, C.: Edge foci interest points. In: International Conference on Machine Learning, pp. 359–366 (2011)
40. Krystian, M., Cordelia, S.: Scale and affine invariant interest point detectors. Int. J. Comput. Vision **60**(1), 63–86 (2004)
41. Matas, J., Chum, O., Urban, M., Pajdla, T.: Robust widebaseline stereo from maximally stable extremal regions. Image Vis. Comput. **22**(10), 761–767 (2004)
42. Tian, Y., Yu, X., Fan, B., Wu, F., Heijnen, H., Balntas, V.: Sosnet: second order similarity regularization for local descriptor learning. In: International Conference on Computer Vision, pp. 118–126 (2015)